# 3
# Modelo Poisson-gama semi-paramétrico

## 3.1
## Especificação teórica

Considere-se estender o modelo Poisson-gama descrito no capítulo anterior para uma especificação semi-paramétrica. Nesta especificação, o preditor linear em 2-12 é substituído por preditor híbrido – paramétrico e suavizado – que é combinado de forma multiplicativa com o nível do modelo sem variável explicativa. Seja o vetor $X_t$ particionado da forma $X = (X_t^p, X_t^s)$, tal que $X^p$ sejam as covariáveis que compõem a partição paramétrica do preditor e $X^s$ são as covariáveis que compõem a partição não-paramétrica do preditor do modelo.

Sem perda de generalidade, de forma equivalente a equação 2-11, a distribuição de $y_t$ condicional em $\mu_t$ é Poisson com média dada por

$$\mu_t^* = \mu_t \exp\left(\eta_t^+ + \eta_t^* + \text{offset}\right) \tag{3-1}$$

$$\eta_t^+ = \sum_{j=1}^{p} \beta_j X_j^p \tag{3-2}$$

$$\eta_t^* = \sum_{k=1}^{s} g_k\left(X_k^s\right) \tag{3-3}$$

em que $\mu_t$ é o nível da série temporal $y$, $\eta_t^+$ é a partição paramétrica do preditor e $\eta_t^*$ é a partição não-paramétrica do preditor do modelo. Por simplificação da notação, as partições paramétrica e não-paramétrica do preditor serão referidas como preditor paramétrico e preditor não-paramétrico, respectivamente. É importante notar que $\eta_t^+$ é uma particularização de $\eta_t^*$ na qual as funções $g(\cdot)$ são lineares. O modelo 3-1 tem $p + s$ variáveis explicativas. O termo "offset" tem a mesma função que o *offset* dos modelos lineares generalizados, isto é, representa uma covariável ou uma função de covariáveis com coeficiente linear igual a 1.

Hastie e Tibshirani (1986 e 1990) [22, 23] discutem detalhadamente

um número de opções para as funções suavizadoras $g(\cdot)$. Entretanto, devido às boas propriedades matemáticas abordadas no capítulo anterior, apenas as *splines* cúbicas naturais são utilizadas na classe de modelos Poisson-gama semi-paramétricos considerada aqui. Porém, é facultativa a implementação de outros suavizadores nesta especificação de modelo semi-paramétrico.

Tal como no modelo Poisson-gama paramétrico, a distribuição de $\mu_{t-1}$ condicionada em $Y_{t-1}$ é gama. A distribuição de $\mu_t$ condicionada em $Y_{t-1}$ também é gama, com parâmetros $a_{t|t-1}$ e $b_{t|t-1}$. Então, as equações 2-13 e 2-14 de previsão do filtro Poisson-gama para o modelo com variáveis explicativas podem ser reescritas para o modelo semi-paramétrico como

$$a_{t|t-1} = \omega a_{t-1} \tag{3-4}$$

$$b_{t|t-1} = \omega b_{t-1} \exp\left(-\eta_t^+ - \eta_t^*\right) \tag{3-5}$$

e as equações 2-15 e 2-16 de atualização do filtro para o modelo com variáveis explicativas com ajuste semi-paramétrico são

$$a_t = \omega a_{t-1} + y_t \tag{3-6}$$

$$b_t = \omega b_{t-1} + \exp\left(\eta_t^+ + \eta_t^*\right) \tag{3-7}$$

com $t = \tau + 1, \ldots, n$, em que $\tau$ é o índice da primeira observação não nula de $y$.

A média e variância da distribuição preditiva do modelo com preditor híbrido permanecem as mesmas que em 2-9 e 2-10 respectivamente. Os parâmetros da distribuição do nível do modelo semi-paramétrico condicionada em $Y_{t-1}$ agora são calculadas de acordo com 3-4 e 3-5. Os hiperparâmetros $\omega$ e $\beta_j$ são estimados por máxima verossimilhança, tal como na especificação paramétrica do modelo Poisson-gama, dada pela equação 2-8. As funções suaves $g_k\left(X_k^s\right)$ são estimadas pelo algoritmo *backfitting* abordado no capítulo anterior.

A idéia básica da estimação do Poisson-gama semi-paramétrico consiste em estimar a parte paramétrica do modelo, que depende apenas de $X^p$, por máxima verossimilhança. Dado o preditor linear, calcula-se um resíduo parcial devido ao ajuste paramétrico. Então, este resíduo parcial é usado como variável resposta para o ajuste não-paramétrico pelo algoritmo *backfitting*. O preditor não-paramétrico calculado pelo *backfitting* é agora introduzido na estimação paramétrica como um termo constante, parte do *offset*. Este processo é iterado até que a seqüência de valores da verossimilhança, $\{L(\omega, \beta_j)^i\}$, convirja para algum critério de parada do algoritmo.

Inicialmente, é necessário definir uma forma de resíduo parcial devido ao ajuste da partição paramétrica do modelo. A dificuldade reside no fato de o preditor linear e o preditor não paramétrico não se combinarem diretamente com a equação de previsão do modelo, e sim por meio de um filtro iterativo. Ainda, devido a função de ligação exponencial, o nível da série e os preditores do modelo se relacionam em escalas diferentes. A fim de construir uma proposta de resíduo parcial na mesma escala do preditor não-paramétrico, considere-se que a equação 3-1 pode ser reescrita como

$$\mu_t^* = \mu_t \exp\left(\eta_t^+ + \text{offset}\right) \exp\left(\eta_t^*\right). \tag{3-8}$$

Usando 2-11 e incluindo, sem prejuízo, o termo de *offset* na partição paramétrica do modelo, a seguinte forma também é equivalente

$$\mu_t^* = \mu_t^+ \exp\left(\eta_t^*\right) \tag{3-9}$$

em que $\mu_t^+$ é o nível do modelo Poisson-gama paramétrico.

Considere-se o logaritmo da equação 3-9. O preditor não-paramétrico $\eta_t^*$ se combina de forma aditiva com o logaritmo do nível devido a partição paramétrica para formar o logaritmo do nível do modelo semi-paramétrico. Assim, pode ser escrita a seguinte expressão

$$\log \mu_t^* - \log \mu_t^+ = \eta_t^*. \tag{3-10}$$

Então, é razoável definir o resíduo parcial devido ao ajuste paramétrico da forma

$$rp_t = \log y_t - \log \hat{y}_{t|t-1}^+ \tag{3-11}$$

em que $\hat{y}_{t|t-1}^+$ é o valor previsto pelo modelo considerando apenas a partição paramétrica do preditor, estimado de acordo com a equação 2-9.

O processo de estimação do Poisson-gama semi-paramétrico pode ser sistematizado no algoritmo 3.1.

Uma dificuldade dos modelos Poisson-gama semi-paramétricos é a falta de uma forma explícita para a associação das variáveis explicativas $X^s$, no preditor não-paramétrico $\eta_t^*$, com a variável resposta $Y$. Tal limitação é inerente aos modelos Poisson-gama.

**Algoritmo 3.1** Estimação do Poisson-gama semi-paramétrico com *backfitting*

1. Ajusta-se um modelo Poisson-gama à partição paramétrica das covariáveis $X^p$, obtendo-se as estimativas de máxima verossimilhança iniciais dos hiperparâmetros $\omega$ e $\beta_j$.

2. Dado o preditor linear $\eta_t^+ + \text{offset}$, calcula-se a previsão $\hat{y}_{t|t-1}$ devida à partição paramétrica do modelo.

3. Calcula-se o resíduo do ajuste paramétrico definido em 3-11,

$$rp_t = \log y_t - \log \hat{y}_{t|t-1}^+$$

4. Estima-se a superfície de regressão não-paramétrica das covariáveis $X^s$ sobre o resíduo parcial $rp$ via o algoritmo *backfitting*. Obtendo-se as funções $g_k\left(X_k^s\right)$.

5. Dado o preditor não-paramétrico $\eta_t^*$, faz-se

$$\text{offset}^* = \text{offset} + \eta_t^*$$

6. Reestima-se o modelo paramétrico usando os hiperparâmetros estimados $\omega$ e $\beta_j$ como valores iniciais e o novo *offset*.

7. Repete-se o processo a partir do item 2 até a convergência da seqüência $\{L\left(\omega, \beta_j\right)^i\}$.

## 3.2
## Inferência no modelo semi-paramétrico

Na maioria das aplicações, deseja-se avaliar a qualidade estatística do modelo estimado. Entretanto, não está completamente desenvolvida uma teoria distribucional exata dos estimadores para os modelos semi-paramétricos. Alguma teoria assintótica está restrita à partição paramétrica do modelo. Assim, os procedimentos heurísticos propostos para inferência sobre $\sigma^2$ e para os efeitos dos preditores são derivados da regressão linear. É importante notar que na falta de uma teoria distribucional apropriada, estes procedimentos devem ser usados com cautela em testes de significância formais. Entretanto, oferecem uma orientação adequada para a seleção de modelos.

Considere-se que a soma de funções das covariáveis no preditor não-paramétrico incorporam uma estrutura paramétrica do modelo Poisson-gama usual, o *offset*. Então, as técnicas de diagnósticos dos modelos lineares generalizados podem ser utilizadas no Poisson-gama não-paramétrico tal

como são no Poisson-gama paramétrico. Outras ferramentas de diagnóstico da partição não-paramétrica do modelo podem ser adaptadas dos modelos aditivos generalizados.

Em Fernandes (1990) [13] é discutida a qualidade da aproximação normal para os estimadores de máxima verossimilhança dos parâmetros do modelo Poisson-gama. A aproximação normal para amostras de tamanho entre 25 e 30 é considerada satisfatória. Também é válido o uso da aproximação $\chi^2$ para a estatística do teste de razão de verossimilhança. O autor mostra que à medida que o fator de desconto $\omega$ se aproxima de 1, o seu limite superior, a aproximação normal para o estimador de máxima verossimilhança de $\omega$ se torna inadequada.

Campos e colaboradores [4] apresentam uma solução analítica para o cálculo da matriz de informação assintótica dos estimadores. Opcionalmente, um método numérico pode ser usado para achar a matriz hessiana dos estimadores de máxima verossimilhança. Então, intervalos de confiança para os estimadores podem ser calculados.

O número de graus de liberdade dos resíduos do modelo semi-paramétrico é dado por $n-p-\tau-\sum_{k=1}^{s} gl_k$, em que $\tau$ é o índice da primeira observação não nula da série temporal $y$ e $gl_k$ é o número de graus de liberdade equivalentes da curva $g_k(X_k)$. Hastie e Tibshirani (1990) [23] sugerem uma correção para o número de graus de liberdade estimados de cada curva $g_k$ da partição não-paramétrica do modelo, $gle_k$, dado por $\mathrm{tr}H(\lambda_k) - 1$. Então, o número de graus de liberdade dos resíduos corrigido, $gl_r$, do modelo é dado pela quantidade

$$gl_r = n - p - \tau - \sum_{k=1}^{s} gle_k. \tag{3-12}$$

Logo, o número de graus de liberdades do modelo semi-paramétrico, denotado por $gle$, é $p + \tau + \sum_{k=1}^{s} gle_k$.

É importante notar que $\tau$ observações são perdidas devido à iniciação difusa do filtro Poisson-gama. Assim, o número de observações disponíveis para a estimação da cada *spline* de suavização é $n - \tau$, denotado por $n'$. O número de graus de liberdade dos resíduos corrigido referente a cada curva $g_k$, com $k = 1, \ldots, s$, é dado por $n' - \mathrm{tr}H(\lambda_k) - 1$. Entretanto, esta quantidade deve ser usada apenas para inferência sobre as curvas $g_k$ individuais. O número de graus de liberdade dos resíduos do modelo semi-paramétrico estimado pela equação 3-12 já contempla a perda das $\tau$ observações.

A qualidade do ajuste global pode ser estimada pela função desvio.

Suponha-se que a função desvio do modelo semi-paramétrico, estimada de acordo com a equação 2-17, tenha distribuição $\chi^2$, então $gl_r$ é o número de graus de liberdade desta distribuição. Não há garantias de que a função desvio no modelo semi-paramétrico tenha distribuição $\chi^2$, nem mesmo assintoticamente. Entretanto, é razoável utilizá-la informalmente como distribuição de referência para o teste de razão de verossimilhança na comparação de modelos aninhados [23, 18, 16].

Outra medida importante de qualidade do ajuste do modelo é a estatística generalizada de Pearson denotada por $X^2$. Utilizandos os resultados das equações 2-9 e 2-10, a estatística generalizada de Pearson para o modelo Poisson-gama semi-paramétrico pode ser escrita da forma

$$X^2 = \sum_{t=\tau+1}^{n} \frac{\left(y_t b_{t|t-1} - a_{t|t-1}\right)^2}{a_{t|t-1}\left(1 + b_{t|t-1}\right)}, \tag{3-13}$$

em que $\tau$ é o índice da primeira observação não nula na série temporal $y$.

Considere-se construir faixas de confiança para as curvas $g_k$, com $k = 1, \ldots, s$. A matriz de covariâncias do vetor ajustado $\hat{g}_k$ é dada por $\sigma_k^2 H(\lambda_k) H(\lambda_k)'$, em que $H(\lambda_k)$ é a matriz de suavização referente a covariável $X_k^s$ com parâmetro de suavização $\lambda_k$. Dada uma estimativa de $\sigma_k^2$, esta quantidade pode ser utilizada para a construção da faixa de confiança para a curva $g_k$. Sob a hipótese de normalidade dos erros e desprezando o viés, estas faixas podem ser interpretadas como intervalos de confiança para as curvas $g_k$, com $k = 1, \ldots, s$. Uma faixa calculada com um fator $\pm 2$ corresponde a um intervalo de confiança de aproximadamente 95% [23].

Wahba (1983) [35] propõe a construção de intervalos de confiança bayesianos com boas propriedades amostrais quando a *spline* de suavização é estimada por validação cruzada. A distribução a posteriori de $g_k$ é $N(\hat{g}, \sigma_k^2 H(\lambda_k))$, em que $H(\lambda_k)$ é a matriz de suavização da spline $g_k$.

Uma estimativa para $\sigma_k^2$ motivada pela regressão clássica, usando a correção para o número de graus de liberdade estimado, é dada por

$$\hat{\sigma}_k^2 = \frac{\sum_{t=\tau+1}^{n} (y - \hat{g}_k)^2}{n' - gle_k}. \tag{3-14}$$

Estudos de simulação mostram que este é um bom estimador de $\sigma_k^2$ [33].

Considere-se estimar o parâmetro de dispersão ou escala do modelo ajustado. Tomando por referência a distribuição preditiva binomial negativa no modelo Poisson-gama, é razoável aproximar o parâmetro de dispersão em função da estatística generalizada de Pearson, definida na equação 3-13, da

seguinte forma

$$\phi = \frac{X^2}{gl_r}, \tag{3-15}$$

em que $gl_r$ é estimado de acordo com a equação 3-12.

Como uma medida de parcimônia dos modelos Poisson-gama semi-paramétricos, pode-se ainda definir uma estatística $AIC$ da forma

$$AIC = \frac{D\left(y;\hat{\mu}\right) + 2gle\phi}{n'}, \tag{3-16}$$

em que $D\left(y;\hat{\mu}\right)$ é o valor da função desvio de acordo com a equação 2-17, $gle$ é o número de graus de liberdade do modelo semi-paramétrico e $\phi$ é o parâmetro de dispersão estimado de acordo com a equação 3-15. Esta quantidade tem a forma do critério de informação de Akaike [23].

Estes procedimentos são aproximados e derivados da regressão linear por analogia. Os vários estudos que foram referenciados nesta seção sugerem que estas aproximações são úteis ao menos para orientarem a seleção e comparação de modelos. Esta abordagem inferencial tem sido aplicada aos modelos aditivos generalizados na falta de uma teoria distribucional adequada e ainda em desenvolvimento.



## 3.3 Aspectos computacionais

Nesta seção são apresentados alguns aspectos referentes à implementação computacional do algoritmo de estimação dos modelos da classe Poisson-gama semi-paramétricos. Todos os algoritmos que são empregados na estimação dos modelos Poisson-gama semi-paramétricos foram implementados na forma de uma biblioteca nas linguagens *R* [31] e *C* denominada *pgam*. E, apesar da portabilidade para plataformas e sistemas operacionais suportados pelo *R*, todo o código foi otimizado para execução sobre o sistema operacional *Linux*. A escolha pelas linguagens e sistema operacional é coerente com a filosofia de *software* livre e código aberto. A estimação das *splines* cúbicas naturais é realizada pela biblioteca *modreg* integrante do ambiente *R*.

As estimativas de máxima verossimilhança dos hiperparâmetros são calculadas por meio de processos de otimização não-linear irrestrita como os algoritmos de métrica variável. A implementação do algoritmo neste trabalho empregou o BFGS (Broyden-Fletcher-Goldfarb-Shanno) [34]. A estimação de $\omega$ requer especial atenção uma vez que $0 < \omega \leq 1$, logo,

se faz necessário mapear o hiperparâmetro $\omega$ em um domínio irrestrito. A transformação utilizada é a *logit*. Seja $\alpha$, o hiperparâmetro $\omega$ transformado, dado por

$$\alpha = \log\left(\frac{\omega}{1-\omega}\right). \tag{3-17}$$

O hiperparâmetro $\alpha$ é irrestrito sobre o domínio dos números reais e a transformação garante que $\omega \in (0, 1]$. Para recuperar $\omega$ basta resolver a equação 3-17 para $\omega$. Logo, $\omega$ é dado por

$$\omega = \frac{\exp(\alpha)}{1+\exp(\alpha)}. \tag{3-18}$$

O processo de estimação dos modelos Poisson-gama semi-paramétricos consiste de três processos iterativos: o algoritmo de otimização não linear, o algoritmo *backfitting* e o algoritmo de estimação semi-paramétrica que alterna entre os dois anteriores. Em análises realizadas até agora, o algoritmo de estimação semi-paramétrica convergiu após poucas iterações.

O procedimento analítico para o cálculo da matriz de informação dos hiperparâmetros estimados apresentado em Campos e colaboradores (2003) [4] é derivado em função de $\omega$. Entretanto, opcionalmente, pode-se utilizar a matriz hessiana do algoritmo de otimização calculada na solução encontrada para estimar numericamente as covariâncias dos estimadores e, neste caso, a variância de $\omega$ pode ser aproximada aplicando a regra delta à variância estimada de $\alpha$ [24]. Assim a variância de $\omega$ é

$$\sigma_\omega^2 = \sigma_\alpha^2 \left(\frac{\exp(\alpha)}{1+\exp(\alpha)^2}\right)^2 \tag{3-19}$$

em que $\sigma_\alpha^2$ é a variância do estimador do hiperparâmetro $\alpha$.

O resíduo parcial $rp$, definido em 3-11, está na mesma escala do preditor não-paramétrico. Um problema surge nas observações nulas, porém na prática, quando $y_t = 0$ a observação nula pode ser substituída por um valor muito pequeno. Entretanto, outras formas de resíduos podem ser empregadas no algoritmo de estimação dos modelos Poisson-gama semi-paramétricos, por exemplo, o resíduo de desvio, também implementado.

A rotina principal da biblioteca oferece uma interface simples para o usuário. Por meio dos argumentos da função *pgam* é possível selecionar o algoritmo de otimização numérica, bem como controlar a sua convergência. Os valores iniciais dos hiperparâmetros também são argumentos desta função. Também é possível selecionar através dos argumentos da rotina principal o tipo de resíduo parcial $rp$ a ser utilizado na estimação dos modelos semi-

paramétricos. Ainda, com uso dos argumentos, pode-se selecionar o suavizador a ser usado para estimar as funções $g$ bem como controlar a convergência do algoritmo *backfitting*.

A inclusão de covariáveis que definem fatores sazonais na fórmula do modelo na função *pgam* deve ser realizada por meio do operador $\boldsymbol{f}$. Este operador constrói a partição da matriz desenho correspondente aos fatores sazonais a partir de seu único argumento que é a variável com os níveis de cada período sazonal, evitando assim o problema de não identificabilidade do modelo, e por conseqüência a impossibildade de estimação.

As covariáveis para o ajuste da partição não-paramétrica do preditor são também especificadas na fórmula do modelo na função *pgam*. O operador $\boldsymbol{g}$ tem dois argumentos: a covariável a ser suavizada e o número de graus de liberdade equivalentes para o ajuste não-paramétrico no algoritmo *backfitting*. Cada covariável na partição não-paramétrica deve usar uma instância diferente do operador $\boldsymbol{g}$.

Entre os elementos de saída da rotina principal cabe destacar a importância do componente estrutural de nível dos modelos Poisson-gama, as funções suaves $\hat{g}_k\left(X_k^s\right)$, os componentes $d_t$ da função desvio na observação referente ao instante $t$, os graus de liberdade dos resíduos e o parâmetro de escala dado por $\sum d_t/\mathrm{gl}_r$, em que $\mathrm{gl}_r$ é o número de graus de liberdade dos resíduos do modelo ajustado. Estas informações podem ser utilizadas na avaliação dos modelos semi-paramétricos.

Na biblioteca, está implementado o gráfico de envelope simulado dos resíduos. Apesar de empreender um grande esforço computacional, outro diagnóstico útil para avaliação dos modelos Poisson-gama semi-paramétricos é o envelope simulado dos resíduos [1]. Uma interpretação possível para este gráfico é a seguinte: em uma amostra de tamanho $r$, com $r \to \infty$, de resíduos de modelos estimados para os dados simulados com a mesma distribuição estimada da variável resposta, se até cinco por cento dos resíduos do modelo estimado com os dados reais estiverem fora da faixa de confiança definida pelos resíduos dos modelos simulados, então o modelo proposto é adequado aos dados.

Algumas rotinas utilitárias como o envelope simulado dos resíduos estão disponíveis para diagnóstico e avaliação dos modelos estimados. Outras rotinas de diagnóstico, gráficos e estatísticas de teste utilizadas para avaliar a adequação de modelos, sobretudo os modelos lineares generalizados, estão disponíveis no ambiente $R$ e são aplicáveis aos modelos Poisson-gama semi-paramétricos.