Aplicação do método de séries temporais funcionais em linguagem R

Vitória Maria Martini Wendt [1]

Eduardo de Oliveira Horta[2]

Resumo: Dados funcionais estão cada vez mais em evidência principalmente no âmbito científico. Nesse sentido, propomos a implementação de uma ferramenta que padronize computacionalmente o uso de um método importante na área introduzido por Bathia et al (2010). Desta forma, foi desenvolvido o pacote em linguagem R ftsa2 que almeja tornar análises de séries temporais funcionais mais rápidas e universais, automatizando e melhorando processos.

Palavras-chave: *Dados funcionais, Série Temporal, Métodos Numéricos*.

1 Introdução

Séries temporais funcionais são sequências de dados funcionais ordenadas no tempo. Em muitos casos, uma série temporal funcional é obtida a partir de um processo estocástico a tempo contínuo, mediante uma quebra do processo original em uma sequência de processos concatenados. Por exemplo, o monitoramento contínuo da temperatura em uma estação metereológica induz uma sequência de gráficos anuais de trajetórias dessa variável. Em outras situações, uma série temporal funcional pode ser constituída de funções cujo domínio não é o tempo contínuo. Esse é o caso, por exemplo, de uma sequência de *kernel density estimates*, onde a dimensão temporal reside tão somente no ordenamento dessa sequência.

Esta nova forma de visualizar e analisar os dados está ganhando espaço nas mais diversas áreas da ciência, especialmente pelo fato de que sua aplicabilidade está diretamente relacionada ao aumento da capacidade computacional de processamento de dados. De fato, apenas recentemente encontra-se a implementação destes métodos em pacotes estatísticos como o R (pacote ftsa – functional time series analysis). Ainda assim, tal pacote não engloba uma importante contribuição metodológica feita por Bathia et al (2010). Nesse contexto, o presente trabalho tem como objetivo implementar e automatizar o uso computacional desta metodologia em linguagem R.

O principal método utilizado no contexto de análise de dados funcionais e que está presente no pacote ftsa é o método de componentes principais funcionais, que consiste em expandir cada curva observada em uma base associada à função de covariância correspondente. Este método apenas consegue fazer

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *vitoriawendt@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *eduardohorta@ufrgs.br*

inferência sobre os dados com certas restrições, a saber, de que não há presença de erros de medida. O modelo desenvolvido por Bathia et al (2010) vem justamente suprir esta necessidade de se trabalhar com dados funcionais realistas e que apresentam dependência entre suas observações e erros de medida.

Nota-se que é latente a necessidade de implementação computacional do modelo de séries temporais funcionais citado. Porém, é necessário que esta implementação ocorra de forma padronizada e respeitando boas práticas de programação já que o modelo prevê o uso de métodos computacionalmente custosos como o Bootstrap, além de envolver bases de dados usualmente grandes.

2 Metodologia

O método de análise de componentes principais funcionais é central no contexto de dados funcionais e séries temporais funcionais. Em suma, dado um conjunto de observações de dados funcionais (que podem ser uma série temporal) $x_1, x_2, \ldots, x_n$, a representação de Karhunen–Loève garante que, sob certas condições de regularidade, a seguinte representação é válida:

$$x_t(u) = \mathbb{E}x_1(u) + \sum_{j=1}^{d} Z_{t,j}\varphi_j(u), \tag{1}$$

onde $Z_{t,j}$ são variáveis aleatórias reais de média zero e variância λ_j, e onde φ_j são funções determinísticas que satisfazem a equação

$$\varphi_j(u) = (1/\lambda_j) \int \mathrm{Cov}(y_0(u),\, y_0(v))\, \varphi_j(v)\, dv.$$

Assim, por exemplo, a dinâmica dos dados funcionais $x_1, \ldots, x_n$ se resume à dinâmica do vetor aleatório $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{t,d})$. Esse fato é importante pois permite ao analista modelar e predizer os dados funcionais através de métodos usuais de séries temporais multivariadas, como o modelo VAR.

Nesse contexto, uma importante contribuição foi dada por Bathia et al. (2010). Em muitos casos, os dados funcionais de interesse, x_t, são mensurados na presença de ruído ε_t, de forma que o estatístico tem acesso somente aos dados (y_t), onde

$$y_t(u) = x_t(u) + \varepsilon_t(u), \qquad \mathbb{E}(\varepsilon_t(u)) = 0. \tag{2}$$

Em um cenário desse tipo, torna-se impossível estimar a estrutura de componentes principais funcionais dada em (1). Os autores propõem uma representação alternativa,

$$x_t(u) = \mathbb{E}x_1(u) + \sum_{j=1}^{d} W_{t,j}\psi_j(u), \tag{3}$$

a qual pode ser recuperada a partir dos dados.

O procedimento de estimação baseia-se no cômputo das quantidades $\int y_t(u),\, y_s(u)\, du$, com $t, s = 1, \ldots, n$, a partir das quais é possível recuperar os demais estimadores da teoria. Esse fato mostra que a metodologia pode rapidamente demandar um elevado custo computacional.

3 Implementação computacional

Uma das principais ferramentas atualmente disponíveis para padronização do uso computacional de uma metodologia estatística é a linguagem R com sua rica estrutura de pacotes. Esta ferramenta possibilita o desenvolvimento e compartilhamento de funções de forma compacta e informativa, podendo conter exemplos e tutoriais explicativos para o usuário. Existem múltiplas formas de desenvolvimento de pacotes em liguagem R. No presente projeto, utilizou-se funcionalidades disponíveis na plataforma RStudio, as quais permitem a criação de pacotes sem a necessidade do uso de mais intervenientes.

A versão Beta do pacote ftsa2 está em uso atualmente para que suas características sejam testadas. Estas características referem-se principalmente a aplicabilidade do método quanto a armazenagem de dados e tempo de execução de funções. Em paralelo, técnicas estatísticas estão sendo testadas e desenvolvidas visando uma melhor aderência do pacote às premissas do modelo.

3.1 Armazenamento

A primeira função criada neste pacote permite que o usuário carregue seus dados sem precisar manipulá-los dentro de objetos. Esta é uma facilidade que se contrapõe à abordagem presente no pacote ftsa, onde o usuário deve utilizar como *input* um objeto de classe fts. Tal objeto trata-se de um array de dimensões (1,t_u,u), aonde t_u é o número de séries temporais de comprimento u.

No pacote ftsa2, o usuário carrega seus dados utilizando arquivos de extensão .txt ou .csv e uma função específica compacta estas informação dentro de uma matriz. A escolha por se trabalhar com matrizes ao invés de arrays ou outros tipos de objetos está ligada ao evidente ganho computacional na manipulação de dados utilizando este tipo em linguagem R. Comparando tempos de desempenho, o pacote ftsa2 reduz em quase 80% os tempos de execução quando comparado com a função do método de componentes principais funcionais do pacote ftsa.

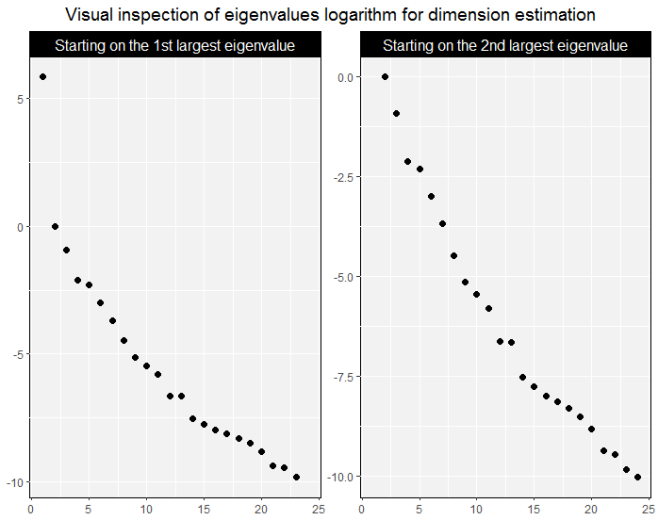
3.2 Estimação

Uma das principais premissas que o pacote ftsa2 carrega é o de automatizar o uso de séries temporais funcionais para usuários em geral, dentro e fora da academia. Seguindo a metodologia proposta em Bathia et al. (2010), o pacote apresenta a funcionalidade de estimação dos parâmetros do modelo especi-

ficado em (2) e (3), a saber, a função-média $\mathbb{E}x_1(u)$, a dimensão d e as auto-funções $\psi_j(u)$, $j = 1, \ldots, d$, além de retornar estimativas para as séries temporais latentes $W_{t,j}$, $j = 1, \ldots, d$.

3.3 Seleção de modelos

O parâmetro d representa a dimensão do modelo de séries temporais funcionais e, como citado anteriormente, também deve ser estimado. A forma mais intuitiva e que também está contemplada no pacote ftsa2 é estimação por inspeção visual dos autovalores.



Todavia, essa abordagem pode ser considerada *ad hoc*. Nesse sentido, o pacote oferece ao usuário um critério de seleção de d via método de bootstrap conforme apresentado em Bathia et al. (2010).

Outro método para a seleção de d que pode ser utilizado é o de validação cruzada por erro de previsão fora da amostra. Esta técnica ainda encontra-se em período de testes no pacote, pois requer um banco de dados significativamente grande, além de ser extremamente custosa do ponto de vista computacional.

3.4 Predição

O principal objetivo do método de séries temporaris funcionais introduzido por Bathia et al (2010) é fazer predições acuradas sobre os dados. Deste modo, é crucial a presença de uma função que a partir dos dados seja capaz de fazer previsões respeitando as premissas de aplicabilidade de métodos de *forecasting*.

A primeira premissa que deve ser cumprida é a de que a série temporal Y trata-se de um processo estacionário. Mas testar estacionariedade de Y é o mesmo que testar a estacionariedade dos coeficientes latentes $\hat{W}_{tj}$. Deste modo, foi desenvolvido uma função no ftsa2 que utiliza testes de estacionariedade já implementados em outros pacotes de R, como o ADF, para averiguar a adequação das suposições acima mencionadas.

Na função de predição, o usuário pode escolher se deseja conduzir a modelagem das séries temporais latentes de forma autônoma ou automática. No segundo caso, o pacote gerará predições a partir de modelos ARIMA se $d = 1$ ou de modelos VAR no caso em que $d > 1$.

3.5 Visualização de dados funcionais

Uma questão inerente ao modelo de dados funcionais utilizado é: como melhor visualizar um conjunto de dados funcionais? Esta é uma questão ainda não explorada em linguagem R de forma geral. Pensando neste problema, o pacote ftsa2 trará em sua composição um conjunto de funções que permite esta visualização utilizando métodos diferentes.

Um destes métodos já implementados é o Waterfall Graphs e é baseado em uma publicação da área de econometria sobre a visualização de múltiplas densidades. Este método permite que o usuário veja o comportamento da série temporal Y, podendo até avaliar de forma visual a estacionariedade do processo.

4 Conclusão

Conforme mencionado, a inferência sobre conjuntos de dados de natureza funcional está a cada dia se difundindo mais na literatura estatística e em áreas afins. O paradigma teórico neste campo é o método de componentes principais. Tal abordagem apresenta um importante inconveniente no caso em que os dados funcionais são observados com erro de medida. No contexto de séries temporais funcionais, a metodologia proposta em Bathia et al. (2010) representa uma importante contribuição, a qual ainda não havia sido implementada em pacotes estatísticos de amplo uso. O pacote ftsa2 surge para preencher essa lacuna.

Os resultados computacionais obtidos pelo pacote são extremamente satisfatórios quando comparado aos métodos já implementados no pacote ftsa. Isto se dá principalmente pela aplicação de boas técnicas de programação como o uso de matrizes para armazenar os dados.

O método desenvolvido por Bathia et al (2010) prevê a estimação de parâmetros que nem sempre podem ser obtidos de forma analítica. Deste modo, o pacote ftsa2 procurou também aplicar métodos computacionais para certas resoluções, não se privando de utilizar técnicas já desenvolvidas em outro pacotes, aumentando assim também o ganho computacional.

O pacote ftsa2 for construído sob a plataforma RStudio utilizando a versão 3.4 da linguagem R e ainda está sendo testado e desenvolvido. Por enquanto, apenas o caso univariado do modelo foi aplicado. Portanto, faz parte das próximas fases de desenvolvimento implementar também o modelo para casos multivariados em que o custo computacional será potencialmente maior. Porém, as técnicas desenvolvidas até o momento pelos autores neste projeto de Iniciação Científica Voluntária, à qual este trabalho está vinculado, mostraram-se extremamente importantes para uma futura propagação computacional do

modelo como principal forma de análise de séries temporais funcionais.

Referências

[1] SMART, Francis. *Waterfall and 3D plotting exploration*. Disponível em: <https://github.com/EconometricsBySimulation/BivariateSlicer/blob/master/slicedens.R>. Acesso em: 27 jul. 2017.

[2] BATHIA, Neil; YAO, Qiwei; ZIELGELMANN, Flavio *IDENTIFYING THE FINITE DIMENSIONALITY OF CURVE TIME SERIES*. Disponível em: <https://projecteuclid.org/euclid.aos/1291126960>. Acesso em: 15 mai. 2017.