CE017 - Análise de Séries Temporais
Departamento de Estatística
Universidade Federal do Paraná

Trabalho No.3

Redigir de maneira individual e entregar na área correspondente no sistema **Microsoft Teans** um relatório eletrônico até o dia **9 de junho de 2025**.

SSA - Análise de Espectro Singular

A série de dados Deaths contêm o número de mortes acidentais mensais nos EUA entre 1973 e 1978. Esses dados foram usados por diversos autores e podem ser encontrados no arquivo de dados R **USAccDeaths**. Aplique a técnica SSA a esse conjunto de dados para ilustrar a capacidade da técnica de extrair tendência, oscilação, ruído e previsão.

O comprimento da janela L é o único parâmetro no estágio de decomposição. A seleção do comprimento da janela adequado depende do problema em questão e de informações preliminares sobre a série temporal. Os resultados teóricos nos dizem que L deve ser grande o suficiente, mas não maior que $T/2$, T é o comprimento da série. Além disso, se sabemos que a série temporal pode ter um componente periódico com um período inteiro, por exemplo, se esse componente for um componente sazonal, então para obter melhor separabilidade desse componente periódico é aconselhável tomar o comprimento da janela proporcional a esse período. Usando essas recomendações, escolher $L = 24$. Então, com base nesse comprimento de janela e no SVD da matriz de trajetória (24×24), temos 24 autotriplos, ordenados por sua contribuição (participação) na decomposição.

A disponibilidade de informações auxiliares em muitas situações práticas aumenta a capacidade de construir o modelo adequado. Certamente, informações auxiliares sobre a série inicial sempre tornam a situação mais clara e ajudam na escolha dos parâmetros dos modelos. Essas informações não apenas podem nos ajudar a selecionar o grupo adequado, mas também são úteis para previsão e detecção de ponto de mudança com base na técnica SSA. Por exemplo, a suposição de que há uma periodicidade anual na série Death sugere que devemos prestar atenção à frequência $k/12$, $k = 1, \cdots, 12$. Obviamente, podemos usar as informações auxiliares para selecionar o comprimento de janela adequado também.

(a) Mostre o gráfico dos logaritmos dos 24 valores singulares da série Death.

Cinco pares evidentes com valores singulares principais quase iguais correspondem a cinco componentes (quase) harmônicos da série: os pares autotriplos 2-3, 4-5, 7-8, 9-10 e 11-12 devem estar relacionados a harmônicos com períodos específicos.

(b) A mostre a tendência extraída no fundo da série original que é obtida do primeiro autotriplo.

Note que podemos construir uma aproximação mais complicada da tendência se usarmos alguns outros autotriplos, este seria o caso de utilizarmos o primeiro e sexto autotriplos. No entanto, a precisão que ganharíamos seria muito pequena e o modelo da tendência se tornaria muito mais complicado.

Tendência é o componente de variação lenta de uma série temporal que não contém componentes oscilatórios. Suponha que a série temporal em si seja um componente desse tipo. A prática mostra que, neste caso, um ou mais dos autovetores principais também variarão lentamente. Sabemos que os autovetores têm (em geral) a mesma forma que os componentes correspondentes da série temporal inicial. Portanto, devemos encontrar autovetores de variação lenta. No nosso caso, o autovetor líder é definitivamente da forma necessária.

(c) Identificação harmônica: mostre o gráfico da extração da oscilação obtida pelos autotriplos de 2 a 12.

Ao comparar a figura obtida desta maneira com os dados originais, fica claro que os autotriplos selecionados para identificar os componentes harmônicos foram feitos corretamente. Como um trabalho alternativo, mostre a figura da oscilação de nossa série obtida pelos autotriplos 2–5 e 7–12. Neste caso, consideramos o sexto autotriplo como um componente de tendência. Parece não haver grande discrepância entre a seleção do sexto autotriplo nos componentes de tendência ou oscilação, como aparece nas figuras aqui solicitadas.

O problema geral aqui é a identificação e separação dos componentes oscilatórios da série que não constituem partes da tendência. A declaração do problema em SSA é especificada principalmente pela natureza livre de modelo do método. A escolha $L = 24$ nos permite extrair simultaneamente todos os componentes sazonais: 12, 6, 4, 3 e 2.5 meses, bem como a tendência.

(d) Previsão: Os valores para os primeiros seis meses de 1979 são 7798 7406 8363 8460 9217 9316.

Utilize os autotriplos identificados, ou seja, os autotriplos entre 1 e 12, para fazer a previsão dos seis primeiros meses de 1979. Mostre a série reconstruída, que é obtida dos autotriplos 1-12. Observe que e a série original e a série reconstruída estão próximas, indicando que os valores previstos são razoavelmente precisos.

(e) Utilize o erro absoluto médio (MAE), que é uma medida de erros entre observações pareadas que expressam o mesmo fenômeno. O MAE é calculado como a soma dos erros absolutos, ou seja, a distância de Manhattan dividida pelo tamanho da amostra:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{T} |y_i - \widehat{y}_i|,$$

onde y_i são os valores observados da série, $\widehat{y}_i$ os valores estimados e T é o número de pontos estimados.

Calcule também o erro percentual absoluto médio (MAPE), também conhecido como desvio percentual absoluto médio (MAPD), é uma medida de precisão de previsão de um método de previsão em estatística. Ele geralmente expressa a precisão como uma razão definida pela fórmula:

$$\mathrm{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{T} \left|\frac{y_i - \widehat{y}_i}{y_i}\right|$$

onde y_i é o valor real da série e $\widehat{y}_i$ é o valor previsto. A diferença deles é dividida pelo valor real y_i. O valor absoluto dessa razão é somado para cada ponto previsto no tempo e dividido pelo número de pontos previstos T.

Modelos dinâmicos

As Empresas de Rede de Transporte (TNCs) fornecem serviços de transporte por aplicativo, permitindo que os passageiros usem seus aplicativos de smartphone para se conectar com motoristas próximos que normalmente dirigem meio período usando seu próprio carro.

Consideramos dados semanais de uso de transporte por aplicativo de $g = 105$ diferentes zonas de táxi em Nova York de janeiro de 2015 a junho de 2017, fornecendo dados para $n = 129$ semanas. Os dados consistem no uso de TNC (tnc), táxi (taxi) e metrô (subway) por zonas de táxi (zoneid) para cada semana (date).

Dados disponíveis em

`http://leg.ufpr.br/~lucambio/CE017/20251S/tnc_weekly_data.csv`

O uso de TNCs agrega três modos de viagem: Uber, Lyft e Via. O uso de táxi agrega táxis amarelos e verdes. Na maioria das zonas, o uso de TNCs mostra um aumento contínuo ao longo do tempo, enquanto o uso de táxi exibe uma tendência decrescente. O uso de metrô, como esperado para Nova York, domina os outros dois na maioria das zonas. O conjunto de dados também inclui dados sobre potenciais preditores, obtidos de várias fontes, como o Portal de Dados Abertos de Nova York (NYC Open Data Portal), a Administração Oceânica e Atmosférica Nacional (National Oceanic and Atmospheric Administration - NOAA), o Departamento do Censo dos EUA (U.S. Census Bureau), etc..

Eles incluem:

(a) dados que variam por semana, mas permanecem constantes em todas as zonas de táxi para qualquer semana, como feriados (holidays) um indicador e precipitação (precip) em polegadas, e

(b) dados sobre variáveis socioeconômicas e de uso do solo que presumimos variarem por zona de táxi e consistem em População Total/Número de Edifícios, Empregados em Tempo Integral/Número de Edifícios, Idade Mediana e Rendimento Mediano.

Como os padrões estocásticos parecem variar ao longo do tempo, o uso de modelos dinâmicos é uma escolha natural para esses dados. Modelos dinâmicos nos permitem reunir informações de g grupos ou indivíduos (zonas de táxi em nosso exemplo) para estimar coeficientes comuns, bem como estimar coeficientes específicos de grupo/zona.

No conjunto de dados, dividimos cada contagem observada por $k = 10.000$ e dividimos os dados da série temporal do painel em partes de treinamento (ou calibração) e de retenção (ou teste). Especificamente, reservamos $n_h = 5$ semanas de cada uma das zonas de táxi $g = 105$.

(a) Faça um estudo descritivo mostrando o uso semanal em escala de TNCs e táxis em duas zonas de táxis selecionadas aleatoriamente em quatro distritos ou zonas de táxis da cidade de Nova York. Comente se as observações acerca do comportamento do uso do TNCs e do táxi se observa nos gráficos.

(b) As séries temporais de resposta Y_t escalonada representam o uso de TNCs na semana t na zona i, com $n = 129$ e $g = 105$. A resposta pode ser escrita como uma matriz de dados Y com g linhas (sujeitos/locais) e n colunas (pontos no tempo).

(c) Consideramos três tipos de preditores: preditores que variam por semana t e zona i: estes incluem o uso do metrô e o uso de táxi. Preditores que variam apenas na semana t: essas variáveis são constantes em todas as zonas de táxi i para qualquer t dado. Eles incluem feriados (holidays) - um indicador - e precipitação (precip) em polegadas. Preditores que variam apenas pela zona i: presume-se que permaneçam constantes semana após semana. Incluem população escalonada (scaled.population), em-

prego escalonado (scaled.employment), idade mediana (median.age) e rendimentos medianos (media.earnings).

(d) Considere o seguinte modelo:

$$\begin{aligned} Y_{it} &= \alpha + \beta_{it,0} + \boldsymbol{b}'_{it}\boldsymbol{\beta}_{it} + \boldsymbol{d}'_t\boldsymbol{\gamma}_t + \boldsymbol{s}'_i\boldsymbol{\eta}_i + \nu_{it} \\ \beta_{it,0} &= \phi_0\beta_{i,t-1,0} + \omega_{it,0}, \\ \beta_{it,h} &= \phi_h^{(\beta)}\beta_{i,t-1,h} + \omega_{it,h}^{(\beta)},\ h = 1, 2, \\ \gamma_{t,\ell} &= \phi_\ell^{(\gamma)}\gamma_{t-1,\ell} + \omega_{t,\ell}^{(\gamma)},\ \ell = 1, 2. \end{aligned}$$

(e) Incluímos um nível α e um intercepto variável no tempo $\beta_{it,0}$ que é modelado como um processo $AR(1)$ latente de média zero com coeficiente ϕ_0. Os coeficientes correspondentes ao uso do metrô e do táxi são $\beta_{it,h}$, $h = 1, 2$, cada um dos quais evolui como um processo $AR(1)$ gaussiano latente com coeficientes $\phi_1^{(\beta)}$ e $\phi_2^{(\beta)}$, respectivamente. Os coeficientes correspondentes a feriados e precipitação são $\gamma_{t,\ell}$, $\ell = 1, 2$ e evoluem como processos $AR(1)$ gaussianos latentes com coeficientes $\phi_1^{(\gamma)}$ e $\phi_2^{(\gamma)}$, respectivamente. Os coeficientes $\boldsymbol{\eta}_i$ correspondentes aos preditores demográficos e de uso do solo são específicos apenas para zonas de táxi e não seguem nenhum modelo dinâmico. Assumimos que todos os coeficientes $AR(1)$ estão entre -1 e 1. Os erros de observação ν_{it} são assumidos como $N(0, \sigma_\nu^2)$. Os erros de estado $\omega_{it,0}$ são $N(0, \sigma_{\omega,0}^2)$, os erros $\omega_{it,h}^{(\beta)}$ são $N(0, \sigma_{\omega,\beta_h}^2)$, $h = 1, 2$ e os erros $\omega_{t,\ell}^{(\gamma)}$ são $N(0, \sigma_{\omega,\gamma_\ell}^2)$, $\ell = 1, 2$. Os erros são não correlacionados. Para elementos de $\boldsymbol{\eta}_i$ no modelo, assumimos a priori normais independentes.

(f) Estime o modelo dinâmico proposto e verifique a precisão do ajuste nos dados reservados como de retenção. Apresente o esudo dos resíduos e, se conveniente, faça sugestões outros possíveis trabalhos (modelos).