

Análise de Séries Temporais

Capítulo VI. Modelos de Espaço de Estados

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Outubro, 2023

- ▶ VI.1. Modelo Gaussiano linear
- ▶ VI.2. Filtragem, alisamento e previsão
- ▶ VI.3. Estimação por máxima verossimilhança
- ▶ VI.4. Modificações de dados ausentes

Um modelo muito geral que engloba toda uma classe de casos especiais de interesse, em grande parte da mesma maneira que a regressão linear faz, é o modelo de espaço de estados ou o modelo linear dinâmico, que foi introduzido em Kalman (1960) e Kalman and Bucy (1961).

O modelo surgiu na configuração de rastreamento espacial, onde a equação de estado define as equações de movimento para a posição ou estado de uma espaçonave com localização X_t e os dados Y_t refletem informações que podem ser observadas em um dispositivo de rastreamento como velocidade e azimute.

Embora introduzido como um método principalmente para uso em pesquisas relacionadas à aeroespaciação, o modelo foi aplicado à modelagem de dados da economia, medicina e ciências do solo.

Em geral, o modelo de espaço de estados é caracterizado por dois princípios.

Primeiro, há um processo oculto ou latente X_t chamado processo de estado. O processo de estado é assumido como sendo um processo de Markov; isso significa que os futuros efeitos

$$\{X_s : s > t\} \quad \text{e os efeitos passados} \quad \{X_s : s \leq t\},$$

são condicionalmente independentes no presente, X_t .

A segunda condição é que as observações Y_t sejam independentes, dados os estados X_t . Isso significa que a dependência entre as observações é gerada pelos estados. Os princípios são exibidos na figura a seguir.

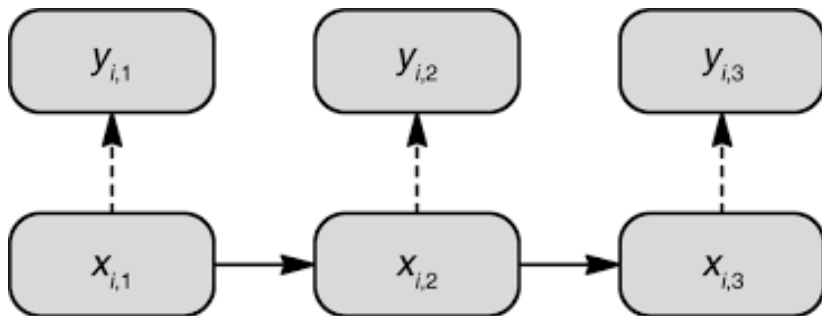


Figura VI.1: Diagrama de um modelo de espaço de estados.

O modelo de espaço de estados linear gaussiano também chamado modelo linear dinâmico (DLM), em sua forma básica, emprega um modelo autoregressivo de ordem um, de dimensão p , como equação de estado,

$$X_t = \Phi X_{t-1} + W_t,$$

com os $W_t \sim N_p(0, Q)$ iid. No DLM, assumimos que o processo começa com um vetor normal X_0 , de modo que $X_0 \sim N_p(\mu_0, \Sigma_0)$.

Não observamos diretamente o vetor de estado X_t , mas apenas uma versão linear transformada dele com ruído adicionado,

$$Y_t = A_t X_t + \nu_t,$$

onde A_t é uma matriz de medição ou observação $q \times p$; o sistema de equações acima é chamado de equações de observação. O vetor de dados observado, Y_t , é q -dimensional.

Definição VI.1 Modelo Linear Dinâmico (DLM)

Chamamos de modelo linear dinâmico ou modelo de espaço de estados linear gaussiano ao sistema de equações

$$\begin{aligned} (A) \quad X_t &= \Phi X_{t-1} + \Upsilon U_t + W_t, \\ (B) \quad Y_t &= A_t X_t + \Gamma U_t + \nu_t, \end{aligned}$$

onde a equação (A) é chamada de equações de estados e (B) de equações de observações. O vetor $r \times 1$, U_t é conhecido como vetor de entradas; Y_t , de dimensão q as respostas, X_t os estados de dimensão p e Υ é $p \times r$ e Γ é $q \times r$, qualquer uma destas matrizes pode ser a matriz zero.

Exemplo VI.1. Um exemplo biomédico.

Suponha que consideremos o problema de monitorar o nível de vários marcadores biomédicos após um paciente com câncer ser submetido a um transplante de medula óssea.

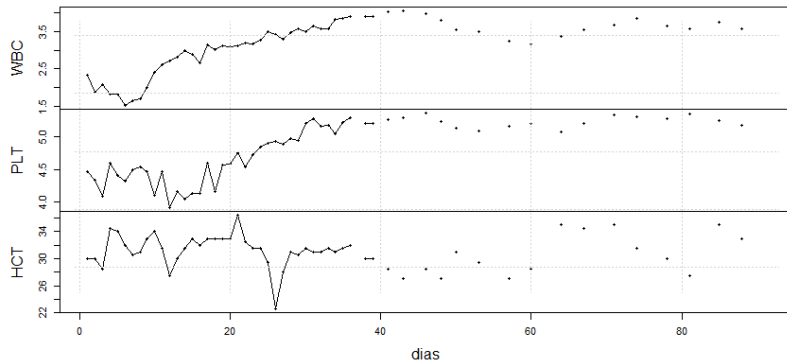
Os dados são medições feitas por 91 dias nas variáveis: WBC, logaritmo da contagem de glóbulos brancos; PLT, logaritmo do número plaquetas e HCT, hematócrito.

Aproximadamente 40% dos valores estão ausentes, com valores ausentes ocorrendo principalmente após o 35° dia. Os principais objetivos são modelar as três variáveis usando a abordagem de espaço de estados e estimar os valores ausentes. De acordo com Jones, "a contagem de plaquetas cerca de 100 dias após o transplante já demonstrou ser um bom indicador de sobrevida a longo prazo".

Para essa situação específica, modelamos as três variáveis em termos da equação de estado; isso é,

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \phi_{1,3} \\ \phi_{2,1} & \phi_{2,2} & \phi_{2,3} \\ \phi_{3,1} & \phi_{3,2} & \phi_{3,3} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \end{pmatrix} + \begin{pmatrix} W_{t,1} \\ W_{t,2} \\ W_{t,3} \end{pmatrix}.$$

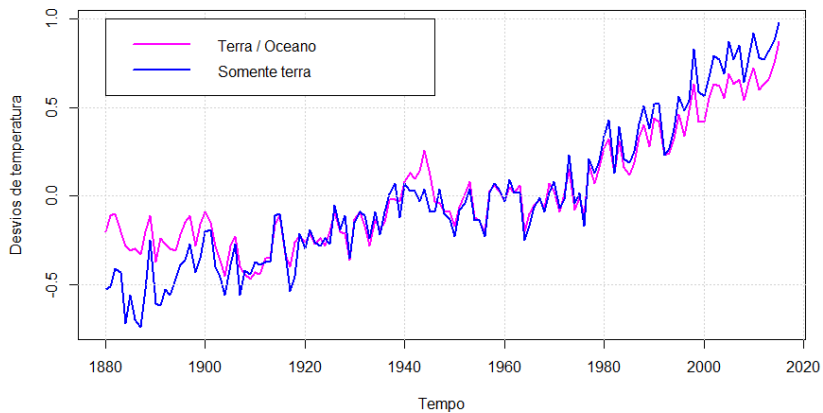
As equações de observação seriam $Y_t = A_t X_t + V_t$, onde a matriz de observação 3×3 , A_t , é a matriz de identidade ou a matriz zero, dependendo se uma amostra de sangue foi coletada naquele dia. As matrizes de covariância R e Q são, cada uma, matrizes 3×3 .



Exemplo VI.2. Aquecimento global.

A figura mostra duas formas diferentes de mensuração para as séries de temperatura global de 1880 a 2015. Uma é **globtemp**, são os dados médios globais do índice de temperatura terra-oceano. A segunda série, **globtempl**, são os dados do índice de temperatura do ar da superfície usando apenas dados da estação meteorológica.

Conceitualmente, ambas as séries devem medir o mesmo sinal climático subjacente, e podemos considerar o problema de extrair esse sinal subjacente.



A introdução da abordagem de espaço de estados como uma ferramenta para modelar dados nas ciências sociais e biológicas requer identificação do modelo e estimação de parâmetros.

As questões de interesse geral para o modelo linear dinâmico dizem respeito à estimativa dos parâmetros desconhecidos contidos em Φ , Υ , Q , Γ , A_t e R ; que definem o modelo específico e estimam ou preveem valores do processo não observado subjacente X_t .

As vantagens da formulação do espaço de estados estão na facilidade com que podemos tratar várias configurações de dados ausentes e na incrível variedade de modelos que podem ser gerados a partir deles. Podemos gerar estruturas de efeitos fixos e aleatórios que são constantes ou variam ao longo do tempo, simplesmente fazendo escolhas apropriadas para a matriz A_t e a estrutura de transição Φ .

Do ponto de vista prático, um objetivo primário de qualquer análise envolvendo o modelo de espaço de estados seria produzir estimadores para o sinal não observado subjacente X_t , dados os dados $Y_{1:s} = \{Y_1, \dots, Y_s\}$, para o tempo s . Quando $s < t$, o problema é chamado de prognóstico ou previsão. Quando $s = t$, o problema é chamado de filtragem e quando $s > t$, o problema é chamado de alisamento ou suavização.

A solução para esses problemas é realizada através do filtro e suavização de Kalman e é o foco desta seção. Ao longo deste texto, usaremos as seguintes definições:

$$X_t^s = E(X_t | Y_{1:s})$$

e

$$P_{t_1, t_2}^s = E((X_{t_1} - X_{t_1}^s)(X_{t_2} - X_{t_2}^s)^T).$$

Quando $t_1 = t_2$ escrevemos P_t^s por conveniência.

Ao obter as equações de filtragem e suavização, confiaremos fortemente na suposição gaussiana; mesmo no caso não gaussiano. Ou seja, podemos pensar em $E(X_t | Y_{1:s})$ como o operador de projeção, em vez de esperança e $Y_{1:s}$ como o espaço das combinações lineares de $\{Y_1, \dots, Y_s\}$; nesse caso, P_t^s é o correspondente ao erro quadrático médio.

Como os processos são gaussianos, P_{t_1, t_2}^s também é a covariância condicional do erro; isso é,

$$P_{t_1, t_2}^s = E((X_{t_1} - X_{t_1}^s)(X_{t_2} - X_{t_2}^s)^\top | Y_{1:s}).$$

Esse fato pode ser visto, por exemplo, observando a matriz de covariância entre $(X_t - X_t^s)$ e $Y_{1:s}$, para qualquer t e s , é zero; poderíamos dizer que eles são ortogonais. Esse resultado implica que $(X_t - X_t^s)$ e $Y_{1:s}$, são independentes, por causa da normalidade e, portanto, a distribuição condicional de $(X_t - X_t^s)$ dado $Y_{1:s}$, é a distribuição incondicional de $(X_t - X_t^s)$.

Teorema VI.1. Filtro de Kalman

Para o modelo de espaço de estados, com condições iniciais $X_0^0 = \mu_0$ e $P_0^0 = \Sigma_0$, para $t = 1, 2, \dots, n$,

$$X_t^{t-1} = \Phi X_{t-1}^{t-1} + \Upsilon U_t,$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q,$$

com

$$X_t^t = X_t^{t-1} + K_t(Y_t - A_t X_t^{t-1} - \Gamma U_t),$$

$$P_t^t = (I - K_t A_t) P_t^{t-1},$$

onde

$$K_t = P_t^{t-1} A_t' (A_t P_t^{t-1} A_t' + R)^{-1},$$

é chamado de ganho de Kalman.

Previsões para $t > n$ são realizadas avaliando X_t^{t-1} e P_t^{t-1} , nas duas primeiras equações com as condições iniciais X_n^n e P_n^n . Subprodutos do filtro são as inovações ou erros de previsão

$$\epsilon_t = Y_t - E(Y_t | Y_{1:t-1}) = Y_t - A_t X_t^{t-1} - \Gamma U_t,$$

e as matrizes de variâncias e covariâncias correspondentes

$$\Sigma_t = \text{Var}(\epsilon_t) = \text{Var}(A_t(X_t - X_t^{t-1}) + \nu_t) = A_t P_t^{t-1} A_t' + R,$$

para $t = 1, 2, \dots, n$.

Assumimos que $\Sigma_t > 0$, ou seja, Σ é uma matriz definida positiva, o que é garantido, por exemplo, se $R > 0$. Consideramos o problema de obter estimadores para X_t^n ; chamados alisadores porque um gráfico da sequência $\{X_t^n\}$ é tipicamente mais suave do que àquele das previsões $\{X_t^{t-1}\}$ ou dos filtros $\{X_t^t\}$.

A suavização implica que cada valor estimado é uma função do presente, futuro e passado, o estimador filtrado depende do presente e do passado. A previsão depende apenas do passado.

Teorema VI.2. Alisamento de Kalman

Para o modelo de espaço de estados, com condições iniciais X_n^n e P_n^n , obtidas segundo o Filtro de Kalman, para $t = n, n-1, \dots, 1$,

$$\begin{aligned}X_{t-1}^n &= X_{t-1}^{t-1} + J_{t-1}(X_t^n - X_t^{t-1}), \\P_{t-1}^n &= P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1})J_{t-1}^\top,\end{aligned}$$

onde

$$J_{t-1} = P_{t-1}^{t-1}\Phi^\top (P_t^{t-1})^{-1}.$$

Exemplo VI.5. Modelo de nível local.

Simulamos $n = 50$ observações do modelo de tendência de nível local discutido no Exemplo VI.4. Geramos um passeio aleatório

$$\mu_t = \mu_{t-1} + W_t,$$

onde $W_t \sim N(0, 1)$ independentes identicamente distribuídos e $\mu_0 \sim N(0, 1)$.

Supomos então que observamos a série Y_t consistindo da componente de tendência μ_t e do componente de ruído branco $\nu_t \sim N(0, 1)$, onde

$$Y_t = \mu_t + \nu_t.$$

As sequências $\{W_t\}$, $\{\nu_t\}$ e μ_0 foram geradas independentemente. Em seguida, executamos o filtro e o alisamento de Kalman, Teorema VI.1 e Teorema VI.2, usando os parâmetros verdadeiros.

O painel superior da Figura VI.4 mostra os valores reais de μ_t como pontos e as previsões μ_t^{t-1} , para $t = 1, 2, \dots, 50$, sobreposto no gráfico como uma linha. Além disso, exibimos

$$\mu_t^{t-1} \pm a\sqrt{P_t^{t-1}}$$

como linhas tracejadas no gráfico.

O painel do meio exhibe o filtro, μ_t^t , para $t = 1, 2, \dots, 50$, como uma linha com $\mu_t^t \pm a\sqrt{P_t^t}$ como linhas tracejadas. O painel inferior da Figura VI.4 mostra um gráfico semelhante para o alisamento μ_t^n .

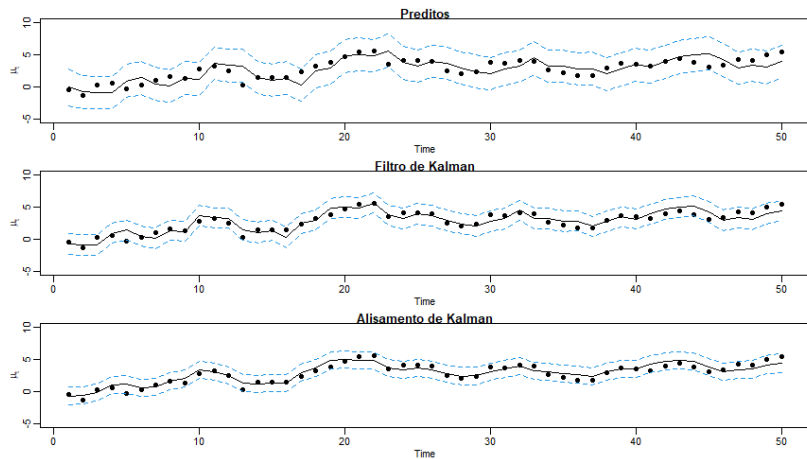


Figura VI.4:

A estimação dos parâmetros que especificam o modelo de espaço de estados, na Definição VI.1, é bastante complexa.

Usamos Θ para representar o vetor de parâmetros desconhecidos na média inicial μ_0 e covariância Σ_0 , matriz de transição Φ e as matrizes de estado e de covariância de observação Q e R e as matrizes de coeficientes de entrada Υ e Γ .

Usamos a máxima verossimilhança assumindo que o estado inicial é normal, $X_0 \sim N_p(\mu_0, \Sigma_0)$ e que os erros são normais $W_t \sim N(0, Q)$ e $\nu_t \sim N(0, R)$. Continuamos assumindo, para simplificar, que $\{W_t\}$ e $\{\nu_t\}$ não estão correlacionados.

A verossimilhança é calculada usando as inovações $\epsilon_1, \dots, \epsilon_n$, definido por

$$\epsilon_t = Y_t - A_t X_t^{t-1} - \Gamma U_t.$$

A forma da verossimilhança das inovações dos dados $Y_{1:n}$, que foi dada pela primeira vez por Schweppe (1965), é obtida usando um argumento semelhante ao que leva a

$$L(\beta, \sigma_w^2) = \frac{1}{(2\pi\sigma_w^2)^{\frac{n}{2}}} \frac{1}{\sqrt{r_1(\beta)r_2(\beta)\cdots r_n(\beta)}} \exp\left(-\frac{S(\beta)}{2\sigma_w^2}\right),$$

onde

$$S(\beta) = \sum_{t=1}^n n \frac{(X_t - X_t^{t-1}(\beta))^2}{r_t(\beta)},$$

obtida na **Seção III.5** e prossegue notando que as inovações são vetores aleatórios Gaussianos independentes com médias zero e matrizes de covariância

$$\Sigma_t = A_t P_t^{t-1} A_t^\top + R.$$

Shumway and Stoffer (1982) apresentaram um procedimento de estimação conceitualmente mais simples com base no algoritmo Baum-Welch (Baum et al., 1970), também conhecido como algoritmo EM (Dempster et al. ., 1977).

A ideia básica do algoritmo EM é que, se pudéssemos observar os estados $X_{0:n} = \{X_0, X_1, \dots, X_n\}$, além das observações $Y_{1:n} = \{Y_1, \dots, Y_n\}$, então consideraríamos $\{X_{0:n}, Y_{1:n}\}$ como o conjunto de dados completos, com densidade conjunta

$$f_{\Theta}(x_{0:n}, y_{1:n}) = f_{\mu_0, \Sigma_0}(x_0) \prod_{t=1}^n f_{\Phi, Q}(x_t | x_{t-1}) \prod_{t=1}^n f_R(y_t | x_t).$$

Sob a suposição gaussiana e ignorando constantes, a verossimilhança dos dados completos acima, pode ser escrita como

$$\begin{aligned} -2 \ln (L_{X,Y}(\Theta)) &= \ln (|\Sigma_0|) + (x_0 - \mu_0)^\top \Sigma_0^{-1} (x_0 - \mu_0) \\ &+ n \ln (|Q|) + \sum_{t=1}^n (x_t - \Phi x_{t-1})^\top Q^{-1} (x_t - \Phi x_{t-1}) \\ &+ n \ln (|R|) + \sum_{t=1}^n (y_t - A_t x_t)^\top R^{-1} (y_t - A_t x_t). \end{aligned}$$

Assim, na visão da expressão acima, se tivéssemos os dados completos, poderíamos então usar os resultados da teoria normal multivariada para obter os estimadores de máxima verossimilhança.

Embora não tenhamos os dados completos, o algoritmo EM fornece um método iterativo para encontrar os estimadores de máxima verossimilhança de Θ com base nos dados incompletos, $y_{1:n}$, maximizando sucessivamente a esperança condicional da verossimilhança dos dados completos.

Para implementar o algoritmo EM, escrevemos, na iteração j , $j = 1, 2, \dots$ como

$$Q(\Theta | \Theta^{(j-1)}) = E\left(-2 \ln(L_{X,Y}(\Theta)) | y_{1:n}, \Theta^{(j-1)}\right).$$

O cálculo da expressão acima é o passo da esperança. Obviamente, dado o valor atual dos parâmetros $\Theta^{(j-1)}$, podemos usar o Teorema VI.2 para obter as esperanças condicionais desejadas como alisamentos.

Esta propriedade rende

$$\begin{aligned} Q(\Theta | \Theta^{(j-1)}) &= \ln(|\Sigma_0|) + \text{tr}\left(\Sigma_0^{-1}(P_0^n + (x_0^n - \mu_0)(x_0^n - \mu_0)^\top)\right) \\ & n \ln(|Q|) + \text{tr}\left(Q^{-1}(S_{11} - S_{10}\Phi^\top - \Phi S_{10}^\top + \Phi S_{00}\Phi^\top)\right) \\ & n \ln(|R|) + \text{tr}\left(R^{-1} \sum_{t=1}^n ((y_t - A_t x_t^n)(y_t - A_t x_t^n)^\top + A_t P_t^n A_t^\top)\right), \end{aligned}$$

onde

$$S_{11} = \sum_{t=1}^n (x_t^n (x_t^n)^\top + P_t^n), \quad S_{10} = \sum_{t=1}^n (x_t^n (x_{t-1}^n)^\top + P_{t,t-1}^n)$$

e

$$S_{00} = \sum_{t=1}^n (x_{t-1}^n (x_{t-1}^n)^\top + P_{t-1}^n).$$

Exemplo VI.8. Algoritmo EM para o Exemplo VI.3.

Usando os mesmos dados gerados no Exemplo VI.6, realizamos a estimação dos parâmetros ϕ , σ_w^2 e σ_v^2 pelo algoritmo EM, bem como os parâmetros iniciais μ_0 e Σ_0 , usando o script EM0. A taxa de convergência do algoritmo EM em comparação com o procedimento de Newton-Raphson é lenta.

Neste exemplo, com a convergência sendo reivindicada quando a mudança relativa na log-verossimilhança é menor que 0.00001; a convergência foi alcançada após 47 iterações. As estimativas finais, junto com seus erros padrão, estão listadas e os resultados são próximos aos do Exemplo VI.6.

Um recurso atraente disponível na estrutura do espaço de estados é sua capacidade de tratar séries temporais que foram observadas irregularmente ao longo do tempo. Por exemplo, Jones (1980) usou a representação de espaço de estados para ajustar modelos ARMA a séries com observações ausentes e Palma and Chan (1997) usaram o modelo para estimação e previsão de séries ARFIMA com observações ausentes.

Shumway and Stoffer (1982) descreveram as modificações necessárias para ajustar modelos multivariados de espaço de estados através do algoritmo EM quando há dados ausentes.

Discutiremos o procedimento em detalhes nesta seção. Para simplicidade de notação, assumimos que o modelo tem a forma da Definição VI.1 quando $\Upsilon = \Gamma = 0$.

Suponha que, em um determinado momento t , definamos a partição do vetor $q \times 1$ de observações em duas partes: a primeira parte $Y_t^{(1)}$, a componente $q_{1t} \times 1$ dos valores observados e a segunda parte $Y_t^{(2)}$, a componente $q_{2t} \times 1$ dos valores não observados, onde $q_{1t} + q_{2t} = q$. Em seguida, escrevemos a equação de observações particionada como

$$\begin{pmatrix} Y_t^{(1)} \\ Y_t^{(2)} \end{pmatrix} = \begin{pmatrix} A_t^{(1)} \\ A_t^{(2)} \end{pmatrix} X_t + \begin{pmatrix} \nu_t^{(1)} \\ \nu_t^{(2)} \end{pmatrix},$$

onde $A_t^{(1)}$ e $A_t^{(2)}$ são, respectivamente, as matrizes de observações particionadas, de dimensões $q_{1t} \times p$ e $q_{2t} \times p$ e

$$\text{Cov} \begin{pmatrix} \nu_t^{(1)} \\ \nu_t^{(2)} \end{pmatrix} = \begin{pmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{pmatrix}$$

denota a matriz de covariâncias dos erros de medição entre as partes observadas e não observadas.

No caso de dados ausentes em que $Y_t^{(2)}$ não é observado, podemos modificar a equação de observação no modelo linear dinâmico, como na Definição VI.1 quando $\Upsilon = \Gamma = 0$, de modo que o modelo seja

$$X_t = \Phi X_{t-1} + W_t \quad \text{e} \quad Y_t^{(1)} = A_t^{(1)} X_t + \nu_t^{(1)},$$

agora, a equação de observação é q_{1t} -dimensional no tempo t .

Nesse caso, segue diretamente do Teorema VI.1 que as equações do filtro são válidas com as substituições notacionais apropriadas. Se não houver observações no tempo t , defina a matriz de ganho, K_t , para a matriz $p \times q$ zero no Teorema VI.1, caso em que $X_t^t = X_t^{t-1}$ e $P_t^t = P_t^{t-1}$.

Em vez de lidar com dimensões observacionais variáveis, é computacionalmente mais fácil modificar o modelo zerando certos componentes e mantendo uma equação de observação q -dimensional por toda parte.

Com as substituições sugeridas acima, os valores da inovação passarão a ser da forma

$$\epsilon_{(t)} = \begin{pmatrix} \epsilon_t^{(1)} \\ 0 \end{pmatrix}, \quad \Sigma_{(t)} = \begin{pmatrix} A_t^{(1)} P_t^{t-1} A_t^{(1)\top} + R_{11t} & 0 \\ 0 & I_{22t} \end{pmatrix},$$

de forma que a verossimilhança das inovações dada anteriormente está correta para este caso e assim com as devidas substituições, a estimação por máxima verossimilhança por meio da verossimilhança das inovações pode proceder como no caso de dados completos.

Exemplo VI.9. Dados biomédicos longitudinais.

Consideramos os dados biomédicos no Exemplo VI.1, nestes dados existem partes do vetor tridimensional ausentes após o dia 40.

O procedimento de máxima verossimilhança produziu os estimadores para a transição, matrizes de covariância do erro de estado e covariância do erro de observação, respectivamente.

O acoplamento entre a primeira e a segunda série é relativamente fraco, enquanto a terceira série HCT está fortemente relacionada às duas primeiras; isso é,

$$\hat{X}_{t,3} = -1.466X_{t-1,1} + 2.258X_{t-1,2} + 0.795X_{t-1,3}.$$

Exemplo VI.9. Dados biomédicos longitudinais. (Continuação)

Portanto, o HCT está negativamente correlacionado com o hemograma (WBC) e positivamente correlacionado com a contagem de plaquetas (PLT). Os subprodutos do procedimento são trajetórias estimadas para todas as três séries longitudinais e seus respectivos intervalos de previsão.

Em particular, a Figura VI.6 mostra os dados como pontos, os valores suavizados estimados $\hat{X}_t^{(n)}$ como linhas sólidas e limites de erro $\pm \sqrt{\hat{P}_t^{(n)}}$ como uma área cinza.

