PREVISÃO DE SURTO EPIDÊMICO DE EBOLA
ABORDAGENS PROBABILÍSTICA E DETERMINÍSTICA

Camila Soares de Souza

Trabalho de Conclusão de Curso apresentado
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo

Orientador: Prof. Dr. Sergio Muniz Oliva

São Paulo, julho de 2017

Resumo

SOUZA, C. S. **Previsão Precoce de Surto Epidêmico de Ebola**. 2016. 12 f. Trabalho de Conclusão de Curso - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Com uma das mais altas taxas de caso-fatalidade, a doença do vírus Ebola ainda é vista como uma ameça à humanidade, devido à inexistência de vacinas preventivas autorizadas e à sua fácil transmissão pelo contato humano, via fluidos corpóreos. Este trabalho apresenta um algoritmo para estimar o pico de epidemia do vírus Ebola. Foi utilizado para este trabalho o modelo Suscetível-Infectado-Recuperado (SIR, Brauer and Castillo-Chavez, 1994), que consiste em um sistema fechado em que indivíduos suscetíveis podem se tornar infectados eventualmente, e posteriomente se recuperam ou morrem. Ainda, as simulações contidas aqui foram feitas a partir dos dados do surto de Ebola ocorrido em 2014, nos países Guinea, Libéria, e Serra Leoa. Modelar e simular epidemias é de extrema importância porque pode prover resultados que auxiliarão tomadores de decisão a controlar novos surtos, reduzir custos e evitar mortes.

Palavras-chave: Ebola, SIR, Filtros.

Abstract

SOUZA, C. S. **Early prediction of Epidemic Outbreak of Ebola**. 2016. 12 p. Term Paper (Graduation) - Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2016.

With one of the highest rates of case-fatality, the Ebola virus is still seen as a menace to humanity, due to non-existence of authorized preventive vaccines and its rather easy transmission through human contact, via body fluids. This work presents an algorithm to estimate the peak of the Ebola epidemic. The Susceptible-Infected-Recovered model (SIR, Brauer and Castillo-Chavez, 1994) was used for this work. SIR is a closed system model is which individuals may eventually become infected, and posteriorly recover or die. The simulations contained here were made from data collected from the Ebola outbreak occurred in 2014, in countries Guinea, Liberia, and Sierra Leone. To model and simulate epidemics is of extreme importance because the outcome may provide results that assist decision makers to control new outbreaks, reduce costs and avoid deaths.

Keywords: Ebola, SIR Model, Filters.

Conteúdo

Capítulo 1

Introdução

Além de afetarem a saúde da população em geral, epidemias podem afetar negativamente a economia pois aumentam custos de imunização, de tratamento posteriores e de acompanhamentos, geram ausência no trabalho e oportunidades de negócio perdidas, entre outros. No caso estudado neste trabalho, o do vírus Ebola, um dos principais eventos que se deseja evitar é o óbito, já que esse vírus possui uma das mais altas taxas de caso/fatalidade que se conhece. Ou seja, o risco de morte é tão eminente que consequências como ausência no trabalho e oportunidades de negócio perdidas são vistas como preocupações secundárias. Uma epidemia causada pelo vírus do Ebola, se iniciada em um local bastante populoso e povoado, poderia resultar na morte de centenas de milhares de pessoas em questão de dias.

Além de ser altamente contagiosa, a doença Ebola ainda não possui cura e nem métodos de imunização, e é por isso que a capacidade de prever comportamentos do desenvolvimento da doença na população é de extrema importância.

Vigilância sindrômica é definida por coleção, análise, e interpretação de dados de saúde pública para o propósito de detecção precoce – e muitas vezes com custo reduzido – de um surto epidêmico, seu monitoramento contínuo, e resposta oportuna de agências públicas de saúde. O racional por trás deste método está na suposição de que a propagação de uma doença infecciosa está normalmente associada com mudanças no comportamento social.

É importante frisar que apesar das medidas para vigilância sindrômica frequentemente dependerem de observações médicas - visitas a médicos, ocorrência de sintomas especiais, número de casos diagnosticados – eles não são limitados a dados puramente médicos. Estudos recentes a, b, demonstram que dados não-médicos como ausência no trabalho, vendas farmacêuticas, pesquisa na internet, e similares podem levar a conclusões sobre o estado epidêmico na comunidade.

Algoritmos para vigilância sindrômica geralmente são divididos em dois grupos, *mineração de dados* e *fusão de informação*. O método de mineração de dados se interessa na extração de padrões de quantidades massivas de dados crus, utilizando-se de modelos dinâmicos do processo em si. Algoritmos de fusão de informação, pelo contrário, dependem fortemente de modelos matemáticos: neste caso, um modelo dinâmico de um surto epidêmico e o modelo de medição para uma série de dados sindrômicos em particular. Obviamente a acurácia dos algoritmos de fusão de informação é determinada pela fidelidade dos modelos subjacentes.

Este trabalho apresenta um estudo de algoritmo recursivo de fusão de informação para vigilância sindrômica. O problema está formulado no contexto Bayesiano de fitros estocásticos não-lineares. Uma vantagem significativa do modelo epidemiológico adotado é uma rigorosa estimação de seu componente de ruído, resultando em estimação mais acuradas de seus parâmetros. O estimador sequencial ótimo - filtro - e preditor são então formulados na estrutura Bayesiana e para sua solução é utilizado um filtro de partícula. Este algoritmo faz parte do que chamamos abordagem probabilística. Após a apresentação deste, proporemos uma simplificação do modelo, cujas equações diferenciais são lineares, e a solução se dará sob o contexto determinístico.

Capítulo 2

Ebola

2.1 Vírus

Ebola é uma doença causada por um vírus cujos sintomas iniciais incluem febre, fraqueza extrema, dores musculares e dor de garganta, segundo a Organização Mundial da Saúde (OMS). À medida que a doença avança, ela pode causar vômitos, diarréias e, em alguns casos, hemorragia interna e externa. Pacientes com a doença podem morrer de falência múltipla dos órgãos ou desidratação. Variedades diferentes da doença podem matar entre 50 a 90 por cento dos infectados.
Entre humanos, o vírus pode se espalhar por meio do contato direto com sangue contaminado, fluidos corporais ou órgãos do doente, ou mesmo por meio do contato com ambientes contaminados. Até funerais de vítimas de ebola podem representar risco, se outras pessoas tiverem contato direto com o corpo do defunto. O período de incubação pode demorar de dois dias a três semanas, e o diagnóstico é difícil. Agentes de saúde pública também correm risco caso tratem pacientes sem tomar as precauções adequadas para prevenir a contaminação. As pessoas permanecem contaminadas enquanto seu sangue e suas secreções contiverem o vírus, em alguns casos, até sete semanas depois da recuperação.
Acredita-se que o vírus tenha sua origem entre morcegos.

2.2 História

A doença apareceu originalmente na República Democrática do Congo (quando se chamava Zaire), em 1976, nas proximidades do rio Ebola, daí seu nome. O primeiro caso foi um professor de 44 anos que ficou doente após comer carne de animais selvagens em uma feira local, e buscou tratamento no Hospital Yambuku Mission, onde agulhas eram utilizadas sem esterilização. O vírus se espalhou por 55 vilas nos arredores. Desde então, se espalhou para o leste, afetando países como Uganda e Sudão. O surto de 2014 tem a particularidade de ter se iniciado na Guiné, que nunca tinha registrado um caso antes, e de estar se espalhando por áreas urbanas. De Nzerekore, uma área rural no sudeste da Guiné, o vírus chegou à capital, Conakry, e aos países vizinhos Libéria e Serra Leoa. Estes países tinham um sistema de saúde muito frágil, com falta de médicos e enfermeiros, além da falta de equipamentos e recursos para combater o vírus. Nigéria e Senegal também registraram casos da doença, mas em números bem menores. Um caso também foi registrado nos Estados Unidos, onde o paciente morreu.

2.3 Epidemias

De acordo com o Centro de Controle de Doenças [**?**], desde seu descobrimento, o vírus Ebola já ocasiou as seguintes epidemias:

Data	Local	Casos	Morte	Mortalidade
1976	Zaire	318	280	88,0%
1995	Rep. Dem. Congo	315	250	39,37%
2000-2001	Uganda	425	224	52,71%
2001-2002	Rep. Dem. Congo	57	43	75,44%
2007	Uganda	149	37	24,83%
2007	Rep. Dem. Congo	264	187	70,83%
2014	África Ocidental	27.000	11.000	40,74%

Tabela 2.1: *Surtos de Ebola em ordem cronológica*

Na tabela acima, os dados de mortalidade são computados sobre os casos efetivamente registrados.

2.4 Prevenção de Infecção

As medidas de prevenção do vírus Ebola são:

- Lavar as mãos com água e sabão várias vezes ao dia;
- Ficar afastado dos doentes com Ebola e também dos mortos pelo Ebola porque eles também podem transmitir a doença;
- Não comer 'carne de caça' e morcegos porque elas podem estar contaminadas com o vírus;
- Não tocar nos fluidos corporais de um infectado, como sangue, vômito, fezes, urina, secreções da tosse e espirros e das partes íntimas;
- Usar luvas, roupa de borracha e máscara quando entrar em contato com um contaminado, não tocando nesta pessoa e desinfetar todo este material após o uso;
- Queimar todas as roupas da pessoa que morreu por causa do Ebola.

Como a infecção com o Ebola pode demorar até 21 dias para ser descoberta, durante um surto de Ebola recomenda-se evitar viajar para os locais afetados e também locais que fazem fronteiras com estes países. Uma outra medida que pode ser útil é evitar locais públicos com grandes concentrações de pessoas, porque nem sempre se sabe quem pode estar infectado e a transmissão do vírus é fácil

Capítulo 3

Abordagem Probabilística

Descreveremos primeiro um modelo matemático probabilístico para descrever a evolução do número de casos da doença ao longo do tempo. Segundo este modelo [?, ?, SIR] a população é dividida em três grupos relacionados com a doença: suscetíveis (S), infectados (I) e removidos (R). Os indivíduos contidos no primeiro grupo nunca tiveram contato com a doença e podem ser infectados, partindo para o segundo grupo. Indivíduos infectados podem, eventualmente, se recuperar ou morrer, e por isso passam a ser parte do terceiro grupo. O surto de uma epidemia pode ser mais veloz que a dinâmica vital de uma população, portanto, para este modelo, manteremos a população P constante. Além disso, normalizaremos $s = S/P$, $i = I/P$, $r = 1 - i - s$. O modelo proposto abaixo contém os seguintes parâmetros: β, , que é o tempo de recuperação da doença e é equivalente ao período contagioso da doença; $\alpha = \rho\beta$, em que ρ é parâmetro famoso em epidemiologia e refere-se a um *número básico de reprodução*, ou seja, a quantidade de pessoas infectadas por contato direto com uma pessoa doente; e finalmente ξ, que é o parâmetro que determina a homogeneidade da população.
Este parâmetro ξ pode variar, pois na existência de uma epidemia, as pessoas se comportam de maneiras diferentes, podendo entrar em pânico ou evitar o risco de contaminação (isolamento total). Para este trabalho, utilizaremos ρ sendo igual a 1, o que significa que estamos assumindo comportamento igual para todos os indivíduos da comunidade.
Abaixo está a figura do modelo.

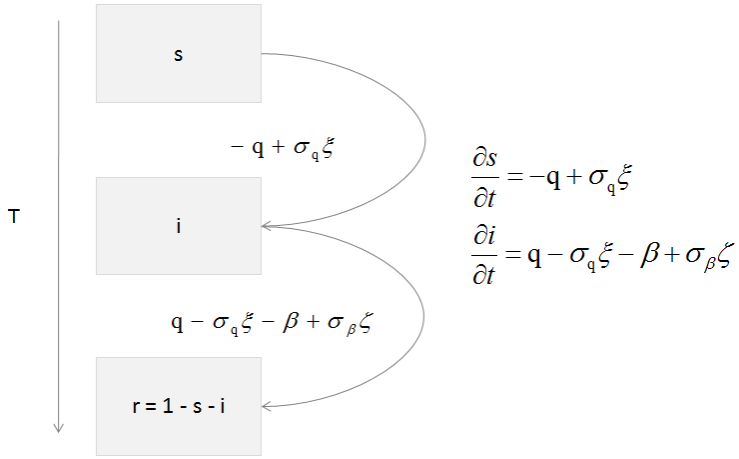


Figura 3.1: *Modelo Probabilístico*

Uma vez escolhido este modelo, consideraremos a abordagem probabilística primeiramente. Das equações do sistema acima, ξ e β são dois ruídos gaussianos brancos, ambos com média zero e variância unitária. Os termos $\sigma_q \equiv \sigma_q(s,i)$ e $\sigma_\beta \equiv \sigma_\beta(s,i)$ são introduzidos para capturar o ruído demográfico (variações aleatórias na taxa de contato α e o tempo de recuperação β entre os indivíduos).
A amplitude dos termos de ruído σ_q e σ_β podem ser estabelecidos por uma lei de escalonamento de flutuações gaussianas geradas pela taxa aleatória de contato $q = \alpha i s^\rho$ e tempo de recuperação β. Então, para um sistema dinâmico que possui um enorme número de indivíduos P podemos escrever:

$$\sigma_q \approx \sqrt{\frac{\alpha i s^\rho}{P}}, \sigma_\beta \approx \sqrt{\frac{\beta i}{P}} \tag{3.1}$$

Abaixo está uma lei de relação entre razões de chances de observações sindrômicas e números de pessoas

infectadas:

$$\frac{z_j}{1-z_j} \propto \left[\frac{i}{1-i_j}\right]^{\varsigma} \tag{3.2}$$

em que z_j é a síndrome observada com índice $j = 1, ..., N_z$ (normalizada pelo tamanho da População P), ς_j é o grau da potência (geralmente diferente para diferentes síndromes). Como no estágio inicial da epidemia, $i \ll 1, z_j \ll 1$, a equação (1.7) pode ser reduzida para:

$$z_j = b_j i_j^{\varsigma} + \tau_j \tag{3.3}$$

em que $b_j = const.$ O termo do ruído τ_j é adicionado para simular a natureza aleatória das medidas, que são supostamente não-correlacionadas com outras síndromes ou ruídos ξ e ζ. Como $z_j \geq 0$, o termo do ruído τ_j associado com a síndrome j deve ser modelado com uma variável aleatória que fornece realizações estritamente positivas. Neste trabalho, utilizaremos uma variável aleatória que obedece a distribuição log-normal, isso significa que determinaremos $\tau_j = \sigma_j \eta_j$, em que η_j é o ruído log-normal padrão $\eta_j \sim (0, 1)$.
Parâmetros $b_j, \sigma_j, \varsigma_j$ são tipicamente desconhecidos, mas com dados observados o modelo pode ser facilmente calibrado. Neste trabalho simplificaremos e assumiremos que todo $\varsigma_j = \varsigma = const.$ Como não se sabe os valores específicos dos parâmetros, usaremos valores heurísticos para $b_j e \sigma j$ em nossas simulações.
Nosso problema está formulado no contexto Bayesiano, e servirá para estimar o número de infectados i e suscetíveis s no tempo t, utilizando observações sindrômicas z_j coletadas até o tempo t. Seja x o vetor de estado a ser estimado, ele inclui i e s, mas também inclui os parâmetros conhecidos imprecisamente α, β and ν. A solucão Bayesiana formal é dada pela forma função densidade de probabilidade (FDP) $\mathrm{p}(x_t|z_{1:t})$ e pode-se prever o progresso da epidemia utilizando o modelo dinâmico apresentado.

3.1 Solução Bayesiana Ótima

Apresentaremos, brevemente, a abordagem Bayesiana do problema. O intuito é tornar o texto mais completo, porém não é o objetivo deste trabalho. A apresentação mais detalhada do que vem a seguir pode ser encontrada em [4].
O modelo apresentado é dado em tempo contínuo. Para um propósito de implementação computacional, é requerido uma aproximação discreta deste modelo.

$$x = \begin{bmatrix} i & s & \alpha & \beta & \nu \end{bmatrix}^T$$

em que ^{T} denota a matrix transposta. Ignorando os termos de ruído do processo, a evolução do estado da epidemia pode ser escrito como $\dot{x} = gx$, onde

$$g(x) = \begin{bmatrix} (\alpha s^{\nu} - \beta)i & -\alpha i s^{\nu} & 0 & 0 & 0 \end{bmatrix}^T$$

A equação não-linear que governa a evolução do estado não pode ser resolvida na forma fechada. O método Euler fornece uma simples aproximação válida para um pequeno intervalo de integração $\tau > 0$: $x(t + \tau) \approx x(t) + \tau g(x(t))$. A evolução do estado no tempo discreto t_k pode ser expressado como:

$$x_{k+1} \approx f_k(x_k) + w_k \tag{3.4}$$

onde $k = t_k/\tau$ é o índice discreto e $f_k(x_k)$ é a função de transição dada por:

$$f_k(x_k) = \begin{bmatrix} x_k[1] + \tau x_k[1](x_k[3]x_k[2]^{x_k[5]} - x_k[4] \\ x_k[2] - \tau x_k[3]x_k[1]x_k[2]^{x_k[5]} \\ x_k[3] \\ x_k[4] \\ x_k[5] \end{bmatrix}$$

Nesta notação $x_k[1]$ representa o i-ésimo componente do vetor x_k
O ruído do processo w_k em 1.9 é um ruído gaussiano branco com média zero e pode ser expressado como $Q \approx diag[(\alpha + \beta)\tau^2/P^2, \alpha\tau^2/P^2]$.
O filtro bayesiano ótimo é tipicamento apresentado em dois estágios, *previsão* e *atualização*. Suponha que a FDP no tempo t_k seja dada por $p(x_k|z_{1:k})$. Então o passo de previsão computa a FDP prevista no tempo

$t_m = t_k + \tau$.

$$p(x_m|z_{1:k}) = \int \pi(x_m|x_k)p(x_k|z_{1:k})dx_k \tag{3.5}$$

onde $\pi(x|x')$ é a densidade transacional. Se $\mathcal{N}(y;\mu,P)$ denota uma FDP gaussiana com média μ e covariância P. Daí, novamente de acordo com 1.9, a densidade transacional é dada por:

$$\pi(x|x') = \mathcal{N}(x; f_k(x'), Q) \tag{3.6}$$

O passo de previsão é feito diversas vezes com pequeninos intervalos τ até que a observação $z_{j,k+1}$ fique disponível sobre a síndrome j no tempo t_{k+1}.
A FDP prevista em t_{k+1} é $p(x_{k+1}|z_{1:k})$. No contexto de estimação bayesiana padrão, esta FDP é atualizada utilizando a medida $z_{j,k+1}$ por multiplicação com a função de medida semelhante, que, segundo 1.8, é $l(z_{j,k+1}|x_{k+1}) = \ln\mathcal{N}(z; h(x_{k+1}); \sigma_j^2)$ onde $h(x_k;\varsigma) = b_j x_{k0[1]^\varsigma}$. Agora, o problema é que $h(x_k;\varsigma)$ definido deste jeito não é uma função porque $\varsigma \in [\underline{\varsigma}, \overline{\varsigma}]$. Uma solução para a transformação de medição imprecisa está disponível no contexto da teoria dos conjuntos aleatórios. Nesta abordagem, $h(x;\sigma)$ define um conjunto fechado $\sum_x$ no espaço de medição ζ. O conjunto fechado é aleatório porque o estado x é aleatório. De fato, o conjunto fechado aleatório CFA $\sum_x$ pode ser visto como um mapeamento composto. O primeiro mapeamento mapeável que define a variável aleatória $x : \Omega \to X$ em que Ω é o espaço de amostragem e ζ é o espaço estado. O segundo mapeamento é $h(x;\sigma) : \zeta \to I\zeta$, em que $L\zeta$ é o conjunto de conjuntos fechados de $]zeta$. O CFA $\sum_x$ é portanto uma variável aletória que admite valores como intervalos fechados de ζ. O passo atualizado de Bayes utilizando a medição $z_{j,k+1}$ é definida como:

$$p(x_{k+1}|z_{1:k+1}) \tag{3.7}$$

onde

$$\underline{\sum}_x = min h(x;\underline{\sigma}), h(x;\overline{\sigma}) \tag{3.8}$$

$$\overline{\sum}_x = max h(x;\underline{\sigma}), h(x;\overline{\sigma}) \tag{3.9}$$

e $\phi(u;\mu,P) = \int_{-\infty}^{u} \ln N(y;\mu,P)dy$ é a distribuição lognormal. As recursões do filtro Bayesiano começam com a FDP incial (no tempo $t_k = 0$), denotado como $p(x_0)$, que é conhecido.Bayesiano não pode ser resolvido na forma fechada. Uma alternativa é desenvolver uma solução aproximada baseada em filtros de partícula. Na literatura, este processo preditor-corretor chama-se Filtro de Kalman [3].

Capítulo 4

Abordagem Determinística

O uso de modelos estocásticos para acomodar incertezas em sistemas dinâmicos é, sem dúvidas, bastante eficaz - principalmente em sistemas como o apresentado acima, cuja interação entre suscetíveis e infectados é premissa, resultando na não linearidade do modelo. Nem todo sistema de equações diferenciais não lineares admite solução explícita, e modelo tal como foi apresentado neste trabalho está nessa categoria. Portanto, um primeiro passo é simplificar o modelo, de forma que suas equações sejam lineares. Posteriormente, mostraremos que o filtro de Kalman admite uma formulação de mínimos quadrados bastante satisfatória. Este algoritmo utiliza a recursão e a equação de Ricatti para computar os coeficientes do filtro. É sabido que o filtro de Kalman é o estimador de mínimos quadrados em uma forma estocástica, já que ele minimiza o valor esperado do erro quadrático estimado. Assim como na abordagem estocástica, os distúrbios do modelo terão papel fundamental neste algorítmo, já que explicaremos as observações como se fossem geradas pelas entradas de distúrbios da norma de mínimos quadrados. A substituição dos distúrbios de mínimos quadrados pode gerar estimadores para qualquer variável de sistemas relacionados. Com isso é fácil ver que isso também gera as mesmas fórmulas como a estimação de verossimilhança máxima, se estes distúrbios são ditos estocásticos e distribuidos normalmente. Mesmo assim, essa interpretação para sistemas em intervalos contínuos não é muito óbvia, por conta as propriedades de ruídos brancos e do movimento Browniano. Em particular, quanto maior a norma L2 do ruído branco, maior a sua probabilidade de ação. Logo, no caso de normas L2 infinitas essa interpretação de semelhança é bastante informal.

4.1 Notação

$\Re$	Reta Real
$\Re_+$	$[0,\infty)$
$\Re^n$	Vetor real de dimensão n
$\Re^{n\times m}$	Matrizes $n \times m$, com coeficientes reais
$M = M^T \in \Re^{n\times n}$	É simétrica definida não-negativa $M \succeq 0$, se $a^T M a \geq 0$ para todo $a \in \Re^n$
$M = M^T \in \Re^{n\times n}$	É simétrica definida positiva $M \succ 0$, se $a^T M a > 0$ para todo $0 \neq a \in \Re^n$
$\parallel a \parallel^2_M$	$a^T M a$ para $M = M^T \in \Re^{n\times n}$ e $a \in \Re^n$
$\parallel a \parallel$	Norma Euclidiana de $a \in \Re^n$, $\parallel a \parallel = \sqrt{\sum_{i=1}^{N} \lvert a_i \rvert^2}$
$f : A \longrightarrow B$	Aplicação f do conjunto A ao B
$f : a \longrightarrow b$	Aplicação f leva elemento $a \in A$ ao $b \in B$
$\mathcal{L}_2(A, \Re^N)$	Conjunto de todos as aplicações quadrado-integráveis de A a $\Re^n$ e A um intervalo finito ou infinito em $\Re$
$\parallel f \parallel^2_{\mathcal{L}_2(A,\Re^N)}$	Norma $\mathcal{L}_2$ de $f \in \mathcal{L}_2(A, \Re^N)$

4.2 Estimação Estática

Assume-se que x é um vetor de dimensão n composto de variáveis reais aleatórias, cujas distribuições conjuntas são normais e têm média zero e covariância normalizadas para a identidade. Vamos estimar $z = Hx$ a partir da observação $y = Cd$, com $H \in \Re^{z\times d}$ e $C \in \Re^{y\times d}$, matrizes fixas e conhecidas.

É possível interpretar z' sem aleatoriedade, observar que y pode ser obtido por d, que tem norma euclidiana mínima. Chamaremos este d dos mínimos quadrados de d^* e $z'' = Hd^*$ será o estimador resultante. Podemos interpretar (3.1) de forma probabilística ou determinística - via mínimos quadrados. Para o caso do filtro de Kalman, em tempo contínuo, a abordagem determinística evita ruído branco, movimento Browniano e cálculos estocásticos avançados.

Consideraremos um problema estático para ilustrar o processo, conforme figura a seguir:

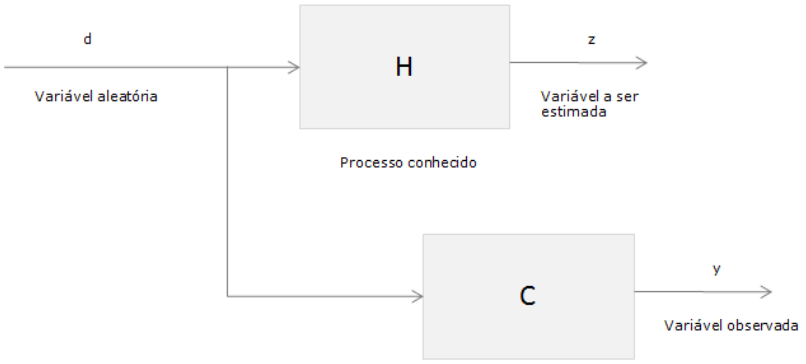


Figura 4.1: *Esquema Estático*

Esse é um problema clássico em estimador de máxima verossimilhança cuja solução é

$$z' = HC^T(CC^T)^{-1}Y \tag{4.1}$$

que pode ser interpretado da seguinte forma, dentre todas as variáves d, tais que $Cd = y$, escolhemos $d*$ que possui norma euclidiana mínima, então $z' = Hd$, segue que $z' = C^T(CC^T)^{-1}y$.
Utilizaremos essa última interpretação para deduzir o chamado Filtro de Kalman para o caso determinístico conhecido, cujo espírito é o mesmo do estático, ou seja, estimar uma variável quando se tem uma observação conhecida.
O filtro de Kalman, também conhecido como estimador linear quadrático, é um algoritmo que utiliza uma série de medidas ao longo do tempo, que contêm ruídos estatísticos e outras imprecisões, e produz estimativas de variáveis desconhcidas que tendem a ser mais precisas do que aquelas baseadas em uma simples medição apenas. Para isso, o algoritmo utiliza inferências Bayesiana e estima uma distribuição de probabilidade conjunta para as variáveis em cada tempo.

4.3 Filtros

Em problemas de filtros, temos dois tipos de sinais envolvidos: o observado e o estimado. Chamaremos, por convenção, o sinal a ser estimado de z e o observado de y. Vamos, num primeiro momento, definir que o conjunto de tempo em que esses sinais são definidos é a reta real $\Re_+$. Posteriormente veremos o que acontece quando o conjunto de tempo é um intervalo finito.
O sinal observado é $y : \Re_+ \longrightarrow \Re^O$ e $z : \Re_+ \longrightarrow \Re^E$ é o sinal a ser estimado. Deve-se achar uma aplicação $\mathcal{F} : y \longmapsto z^*$ tal que $z^* : \Re_+ \longrightarrow \Re^E$ é um bom estimador de z. O interessante desse método é que a restrição que o estimador $z^*(T)$ no tempo T depende apenas do passado de y. A aplicação de filtro $\mathcal{F}$ não permite antecipação.
$\mathcal{F}$ é chamado um $filtro$.

O próximo passo é determinar a relação entre y e z matematicamente, depois estipular um princípio de estimação e por último obter um algoritmo que compute z^* a partir de y, ou seja, um algortimo que implemente $\mathcal{F}$.

No capítulo anterior, que investiga a abordagem estocástica, vimos que a relação entre z e y assume que (z, y) é um processo estocástico com estatísticas conhecidas. O princípio de estimação é a exigência de que z^* tem que ser a esperança condicional de $z(T)$, dado $y(t)$ para $0 \leq t \leq T$. Neste capítulo veremos que as fórmulas para filtros de Kalman são obtidas a partir do modelo estocástico que compreende (z, y) é dado por um modelo de $Gauss - Markov$, através de um sistema linear guiado por inputs de distúrbios, e estes distúrbios modelados como processos de ruídos brancos.
Os ruídos brancos referem-se a modelos estatísticos para sinais e fontes de sinais. Um vetor aleatório pode ser considerado um vetor de ruído branco se os seus componentes possuem um distribuição de probabilidade com média zero e variância finita, além de serem não-correlacionados estatisticamente, ou seja, a distrribuição de probabilidade conjunta tem que ser o produto das distribuições de cada componente. Em particular, se, além de independente, cada variável no vetor tiver distribuição normal, com média zero e mesma variância, este vetor é dito ser um vetor gaussiano de ruído branco.

Com os sinais (z, y) gerados pelo sistema linear abaixo:

$$\dot{x} = Ax + Gd_1, y = Cx + d_2, z = Hx \tag{4.2}$$

Com $A \in \Re^{n\times n}$, $G \in \Re^{n\times d}$, $C \in \Re^{y\times n}$, e $H \in \Re^{z\times n}$ matrizes fixas e conhecidas que parametrizam o

sistema, e $d_1 : \Re_+ \longrightarrow \Re^d$, $d_2 : \Re_+ \longrightarrow \Re^y$, $x : \Re_+ \longrightarrow \Re^n$, $y : \Re_+ \longrightarrow \Re^O$, e $z : \Re_+ \longrightarrow \Re^E$ são os sinais mencionados acima,e conforme figura.
O vetor (d_1, d_2) é o input de distúrbio não observado. Este vetor e o estado inicial não-observado $x(0) \in \Re^n$ da equação anterior determinam o sinal observado y,e o sinal a ser estimado z. Melhor ainda, $d \in \mathcal{L}_2^{loc}(\Re_+, \Re^d)$ e $e \in \mathcal{L}_2^{loc}(\Re_+, \Re^y)$, then y e y são dados em termos de (d, e), $x(0)$, e as matrizes (A, G, C, H) que parametrizam o sistema, por:

$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Gd_1(\tau)d_\tau + d_2(t) \tag{4.3}$$

$$z(t) = He^{At}x(0) + \int_0^t He^{A(t-\tau)}Gd(\tau)d_\tau \tag{4.4}$$

para $t \geq 0$. Destas expressões $y \in \mathcal{L}_2^{loc}(\Re_+, \Re^y)$ e $z \in \mathcal{L}_2^{loc}(\Re_+, \Re^x)$. Para não haver confusão, iremos diferenciar o output y arbitrário do output realmente observado, que chamaremos de $\tilde{y}$.

O trabalho é achar um filtro $\mathcal{F} : \mathcal{L}_2^{loc}(\Re_+, \Re^y) \longrightarrow \mathcal{L}_2^{loc}(\Re_+, \Re^z)$, tal que $\mathcal{F}(\tilde{y})(T)$ é uma boa estimação de $z(T)$.

4.4 Mínimos Quadrados

Sabendo que o output $\tilde{y} : \Re \longrightarrow \Re^O$ foi observado, queremos o filtro que transforme $\tilde{y}$ no estimado $\hat{z}$.

1. Dentre todos os $d_1 \in \mathcal{L}_2([0,T], \Re^d)$ e $d_2 \in \mathcal{L}_2([0,T], \Re^y)$, e $x(0)$ que explicam o $\tilde{y} \in \mathcal{L}_2([0,T], \Re^y)$ observado, basta computar aquele que minimiza a norma quadrada abaixo:

$$\| x(0) \|_\Gamma^2 + \| d_1 \|_{\mathcal{L}_2([0,T],\Re^d)}^2 + \| d_2 \|_{\mathcal{L}_2([0,T],\Re^y)}^2 \tag{4.5}$$

A expressão acima é a medida de incerteza e γ é uma matriz definida positiva. Substituindo em (4.3), temos:

$$\tilde{y}(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Gd_1(\tau)d_\tau + d_2(t) \tag{4.6}$$

para $0 \leq t \leq T$.

2. Chamaremos os (d_1, d_2) e $x(0)$ obtidos no passo anterior de (d_1^*, d_2^*) e $x(0)^*$, e os substituiremos em (4.4). Chame o output resultante de z^*. Daí:

$$z^*(t) = He^{At}x(0) + \int_0^t He^{A(t-\tau)}Gd_1(\tau)d_\tau \tag{4.7}$$

4.5 Completando Quadrados

Lema 1. Considere o seguinte sistema de equações diferenciais envolvendo $d_1 : [0,T] \longrightarrow \Re^d$, $d_2 : [0,T] \longrightarrow \Re^y$, $x : [0,T] \longrightarrow \Re^n$, $y : [0,T] \longrightarrow \Re^y$, $\hat{x}[0,T] \longrightarrow \Re^n$ e $\Sigma : [0,T] \longrightarrow \Re^d$:

$$\dot{z} = Ax + Gd_1, y = Cx + d_2 \tag{4.8}$$

$$\dot{\hat{z}} = A\hat{x} + \Sigma C^T(y - C\hat{x}) \tag{4.9}$$

$$\dot{\Sigma} = GG^T + A\Sigma + \Sigma A^T - \Sigma C^T C\Sigma \tag{4.10}$$

Então, se $\Sigma(t) \in \Re^{nxn}$ é simétrica e não-singular para $0 \leq t \leq T$, então:

$$\| x(0) - \hat{x}(0) \|_{\Sigma(0)^{-1}}^2 + \| d_1 \|_{\mathcal{L}_2([0,T],\Re^d)}^2 + \| d_2 \|_{\mathcal{L}_2([0,T],\Re^y)}^2 \tag{4.11}$$

Igual a

$$\| x(T) - \gamma(T) \|_{\Sigma(T)^{-1}}^2 + \| d - G^T\Sigma^{-1}(x - \hat{x}) \|_{\mathcal{L}_2([0,T],\Re^d)}^2 + \| y - C\hat{x} \|_{\mathcal{L}_2([0,T],\Re^y)}^2 \tag{4.12}$$

Se utilizarmos o lema acima especificando o valor inicial da equação diferencial de Riccati e depois o $\hat{x}$. Para o valor inicial de Γ, usaremos $\Sigma(0) = \Gamma^{-1}$. Como Γ é positiva definida, ela tem solução única para o intervalo $[0,T]$, e que ela é simétrica e também positiva definida.

Lema 2. Seja $\Sigma : [0,T] \longrightarrow \Re^{nxn}$ a solução única da equação de Riccati:

$$\dot{\Sigma} = GG^T + A\Sigma + \Sigma A^T - \Sigma C^T C\Sigma, com \Sigma(0) = \Gamma^{-1} \tag{4.13}$$

Então, $\Sigma(t) = \Sigma(t)^T \succ 0$ para $0 \leq t \leq T$.

A prova do Lema 1 acima pode ser encontrado no artigo [5]

4.6 Filtro de Mínimos Quadrados

O filtro ótimo é deduzido do Lema 2 da seção anterior. De fato, (4.3) mostra que sempre que $(d_1, d2)$ levam ao sinal observado $\tilde{y}$, então é válido que:

$$\| x(0) \|^2_{\Gamma} + \| d_1 \|^2_{\mathcal{L}_2([0,T],\Re^d)} + \| d_2 \|^2_{\mathcal{L}_2([0,T],\Re^y)} \geq \| \tilde{y} - C\hat{x} \|^2_{\mathcal{L}_2([0,T],\Re^y)} \tag{4.14}$$

com $\hat{x}$ gerado de $\tilde{y}$ por

$$\dot{\hat{x}} = A\hat{x} + \Sigma C^T(\tilde{y} - C\hat{x}), \hat{x}(0) = 0 \tag{4.15}$$

No entanto, observe que $\hat{x}$ é uma função de $\tilde{y}$, mas que não depende de (d, e) e $x(0)$ que geraram $\tilde{y}$. Então o lado direito da inequação acima depende de $\tilde{y}$ somente. Então:

$$\| x(0) \|^2_{\Gamma} + \| d \|^2_{\mathcal{L}_2([0,T],\Re^d)} + \| e \|^2_{\mathcal{L}_2([0,T],\Re^y)} \tag{4.16}$$

será minimizado se a inequação se mantiver em (4.14). O lema 2 mostra que a inequação só é válida se, dentre os (d, e) e $x(0)$ que geraram $\tilde{y}$, nós pudermos escolher um que:
A.$x(T) = \hat{x}(T)$
B.$d(t) = G^T\Sigma(t)^{-1}(x(t) - \hat{x}(t)), 0 \leq t \leq T$
Tal escolha existe, e ela virá do teorema abaixo

Teorema do Filtro de Mínimos Quadrados Seja $\tilde{y}$ um output observado. Seja Σ a solução única da equação diferencial de Riccati

$$\dot{\Sigma} = GG^T + A\Sigma + \Sigma A^T - \Sigma C^T C\Sigma, com \Sigma(0) = \Gamma^{-1} \tag{4.17}$$

Então o filtro de mínimos quadrados será dado por

$$\dot{\hat{x}} = A\hat{x} + \Sigma C^T(\tilde{y} - C\hat{x}), \hat{x}(0) = 0, \Sigma(0) = \Gamma^{-1} \tag{4.18}$$

Visto como uma aplicação $\tilde{y} \in \mathcal{L}_2([0,\infty), \Re^O) \longrightarrow \hat{z} \in \mathcal{L}_2([0,\infty), \Re^E)$

Capítulo 5

Modelo e Resultados

Uma vez discutida a simplificação do modelo não-linear e a possibilidade de implementar um algoritmo que trará a solução de forma determinística, veremos a seguir o modelo linear escolhido, as origens dos parâmetros fixos, os dados y observados e o gráfico de resultados.
A modelagem inicial previa interação entre indíviduos s suscetíveis e i infectados, tornando o modelo em não-linear. O modelo simplificado desconsidera essa interação, conforme demonstrado em figura abaixo.
Em forma matricial, temos o modelo de $x = (x_s, x_i)$, o observado $y = (0, y_i)$ e o estimado $z = (0, z_i)$:
E A,C,G,H são matrizes fixas e conhecidas:

$$A = \begin{bmatrix} -\lambda & 0 \\ \lambda & \sigma \end{bmatrix}, G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 \end{bmatrix} e\ H = \begin{bmatrix} 0 & \alpha \end{bmatrix}$$

$$\dot{x} = Ax + Gd$$

$$y = Cx$$

$$z = Hx$$

Recapitulando, e, d, y, x e z são sinais que estão relacionados no sistema descrito assim. O vetor sinal (d, e) deve ser interpretado como um distúrbio não-observado, que, conjuntamente com o sinal $x(0)$ determinam o sinal observado y e o sinal a ser estimado z.
Fixaremos agora os parâmetros λ, σ, e α.
O λ é a taxa pela qual os suscetíves à doença contraem a doença. Para este trabalho, utilizaremos um valor fixo $\lambda = 0.371$, que é resultado encontrado no artigo científico publicado em 2015, por Khan, Naveed, Dur-e-Ahmad e Imram [1].
O σ neste modelo simplificado representa o percentual de infectados que não sobrevivem. Este parâmetro no modelo foi obtido de forma simplicada, ou seja, dos números de infectados reportados, calculou-se, para cada instante t, o número de óbtidos. Deste conjunto foi calculado uma mortalidade média. O resultado obtido foi $\sigma = 0.5$.Esta aproximação chega perto do número contido no mesmo artigo [?] que estimou o λ. Neste artigo, percentual de infectados que se recuperam da doença é 0.45.
O α representa um fator de correção, ou seja, ajusta a quantidade de casos reportados do total de infectados. Esse percentual é o inverso do fator de correção utilizado no mesmo artigo [1].
Importante a diferenciação destes parâmetros em relação aqueles apresentados no capítulo 3 - Abordagem Probabilísticas, que são fundamentados em interação entre suscetíveis e infectados - o que torna o modelo 1 não linear.
Substituiremos os parâmetros fixados na equação 4.13, do capítulo Abordagem Determinística.

$$A = \begin{bmatrix} -0.371 & 0 \\ 0.371 & 0.5 \end{bmatrix}, G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 \end{bmatrix} e\ H = \begin{bmatrix} 0 & 0.4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \end{bmatrix}$$

$$\dot{\Sigma} = GG^T + A\Sigma + \Sigma A^T - \Sigma C^T C\Sigma$$

$$\Sigma(0) = \Gamma^{-1}$$

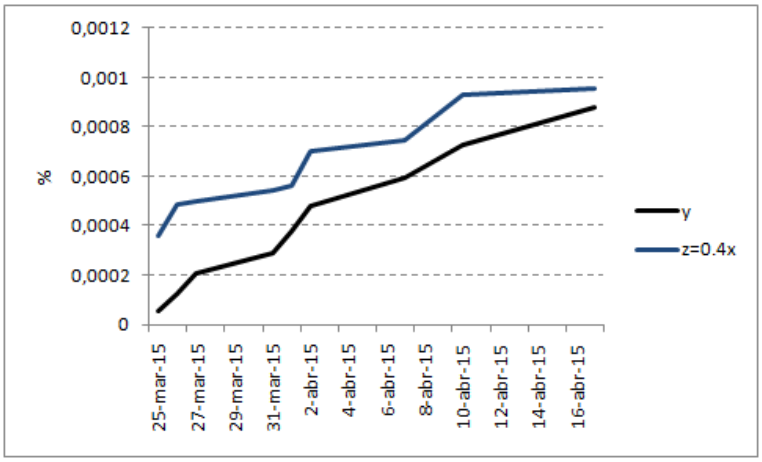
$$\begin{cases} k_1 = 1 - 0.741\ k_1 + 0.16\ k_2 k_3 \\ k_2 = 0.371\ k_1 - 0.871\ k_2 + 0.16 k_2 k_4 \\ k_3 = 0.371\ k_1 - 0.871\ k_3 + 0.16 k_4 k_4 \\ k_4 = 1 + 0.371\ k_2 + 0.371\ k_3 - k_3 + 0.16 {k_4}^2 \end{cases}$$

Para cada instante t, teremos um $\Sigma(t)$, e utilizaremos este $\Sigma(t)$ no filtro 4.18.

$$\dot{x} = Ax + \Sigma C^T (y - Cx), com\ x(0) = 0\ e\ \Sigma(0) = \Gamma^{-1}$$

$$\begin{cases} \dot{x}_1 = -0.371 x_1 + 0.4 k_2 (y(t) - 0.4 x_2) \\ \dot{x}_2 = 0.371 x_1 + 0.5 x_2 + 0.4 k_4 (y(t) - 0.4 x_2) \end{cases}$$

Com o auxílio do R Arquive e o pacote deSolve, obtivemos o seguinte resultado:



Referências

Bibliografia

[1] A. Khan, M. Naveed, M. Dur-e-Ahmad, e M. Imran *Estimating the basic reproductive ratio for the Ebola outbreak in Liberia and Sierra Leone* (Infect Dis Poverty) 4:13 (2015).

[2] A.G. McKendrick *Applications of mathematics to medical problems* (Proceedings of the Edinburgh Mathematical Society) 14, 98:140 (1926).

[3] R. E. Kalman *A New Approach to Linear Filtering and Prediction Problems* (Transactions of the ASME–Journal of Basic Engineering) 84, 35-45 (1960).

[4] A. Skvortsov, B. Ristic, C. Woodruff *Predicting an epidemic based on syndromic surveillance.*

[5] J.C. Willems *Deterministic least squares Filtering* (Journal of Econometrics) 118, 341:373 (2004). 21

15

14

20