BMM5729 Análise sistêmica e engenharia do metabolismo microbiano

Galo A.C. Le Roux

Experimentos: medidas e modelos

- Probabilidade e estatística, um sobrevoo
- Estimação
- Inferências
- Complexidade e Predição
- Conclusões

Probabilidade e Estatística

- Origem:
  - Contagem de estoques, de eventos
  - Teoria dos jogos: baralho, dados
- Essência axiomática:

1) $\forall A, \; 0 =< P(A)$

2) $P(\Omega) = 1$

3) se $A \cap B = \varnothing$ então $P(A \cup B) = P(A) + P(B)$

Probabilidade e Estatística

- Situações práticas:
  - probabilidade enquanto desconhecimento ou renúncia
  - probabilidade enquanto elemento inerente ao processo
  - Demônio de Laplace

Probabilidade e Estatística

- Erros comuns:
  - Probabilidade de caras = ½
  - Definição errada: $P(\text{cara}) = \lim_{N \to \infty} \frac{N_{cara}}{N}$

Probabilidade e Estatística

- Variável aleatória contínua: X
- Não tem valor, tem função distribuição de probabilidade:

(1) $p(x) \geq 0$

(2) $\int_{-\infty}^{+\infty} p(x)\,dx = 1$

(3) $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(u)\,du$

Probabilidade e Estatística

$$P(x_1 \le X \le x_2) = P(x_1 < X \le x_2) = P(x_1 \le X < x_2) = P(x_1 < X < x_2)$$

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(u)\,du$$

a média:	$\mu = E(X) = \int_{-\infty}^{+\infty} x p(x)\,dx$
A variância:	$\sigma^2 = E((X - E(X))^2) = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\,dx$
o desvio padrão:	$\sigma = \sqrt{\sigma^2}$

Revisão de Probabilidades

- Distribuição uniforme

$p(x) = 1/(b-a),$	$a \le x \le b$

$\mu = E(x) = (a + b) / 2$
$\sigma^2 = V(X) = (b - a)^2 / 12$

Revisão de Probabilidades

- Distribuição normal:

$$p(x) = N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$E(X) = \mu$ e $V(X) = \sigma^2$

$$\boxed{Z = (X - \mu) / \sigma}$$

Estimação

- Um estimador é uma função qualquer de uma variável aleatória que fornece como resultado uma variável aleatória e que pode ser usada para a estimativa de uma grandeza não aleatória qualquer.

Estimação

- Utilidade:
  - para descrever um conjunto finito de elementos contáveis a partir de uma amostra reduzida
  - para se fazer inferências sobre processos inerentemente aleatórios abstratos. Por exemplo, a probabilidade de obter cara ao se jogar uma moeda

Estimação

- Média de uma variável aleatória

$$\mu = E(X)$$

- Estimativa da média:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Nunca um chapéu fez tanta diferença:
  - μ não é uma variável aleatória
  - $\hat{\mu}$ é uma variável aleatória, e tem uma p.d.f.

Estimação

- Como o estimador é uma variável aleatória, um estimador tem uma função distribuição de probabilidade, chamada de distribuição de amostragem

Estimação

- Vamos supor que existe um verdadeiro valor dos parâmetros θ^*
- Propriedades do estimador:
  - Bias: $\mathbf{b} = \mathrm{E}\left(\hat{\theta}\right) - \theta^*$
  - Variância Mínima: $\mathbf{V}_{\hat{\theta}} = \mathbf{R}^{-1}$

- Impossível ter variância menor que R^{-1} e não ter bias

Estimação

- Só a estimação pode permitir recuperar alguma realidade determinística ... Através de medidas

- Famílias de estimadores:
  - Mínimos quadrados
  - Mínimos quadrados ponderados
  - Máxima verossimilhança
  - Bayesianos ...

Estimação – Modelos Lineares em Relação aos Parâmetros

- Um exemplo prático:

$$y = a + b\,x$$

- Outro exemplo:

$$y = \theta_1 + \theta_2 \sin(2\pi t) + \theta_3 \sin(\pi t) + \theta_4 \sin(\pi t)$$

- Podemos agrupar os parâmetros em um vetor $\boldsymbol{\theta}$ e as variáveis independentes em um vetor $\mathbf{x}$:

$$y = f(\mathbf{x}, \theta)$$

- Expressão geral:

$$y = \theta_1 r_1(t) + \theta_2 r_2(t) + \theta_3 r_3(t) + \theta_4 r_4(t)$$

- Ou:

$$y = \mathbf{r}^T(\mathbf{x})\,\theta$$

Estimação – Modelos Lineares em Relação aos Parâmetros

- Onde:

$$\mathbf{r}(\mathbf{x}) = \begin{bmatrix} r_1(\mathbf{x}) \\ r_2(\mathbf{x}) \\ \\ r_p(\mathbf{x}) \end{bmatrix}$$

ou

$$y = \sum_{i=1}^{p} r_i(\mathbf{x})\,\theta_i$$

Estimação – Modelos Lineares em Relação aos parâmetros

- Seja um conjunto de medidas das variáveis dependentes:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- e :

$$\mathbf{X} = \begin{bmatrix} \mathbf{r}^T(\mathbf{x}_1) \\ \mathbf{r}^T(\mathbf{x}_2) \\ \vdots \\ \mathbf{r}^T(\mathbf{x}_n) \end{bmatrix}$$

Estimação – Modelos Lineares em Relação aos parâmetros

- Resíduo: $\varepsilon(\theta) = \mathbf{y} - \mathbf{X}\,\theta$
- Solução: $\hat{\theta} = \arg \min_{\theta \in \Re^p} \Phi(\theta)$
- Minimizar: $\Phi(\theta) = \sum_{i=1}^{n} \left(y_i - \mathbf{r}^T(\mathbf{x}_i)\,\theta\right)^2 = \varepsilon(\theta)\varepsilon^T(\theta)$
- Solução: $\mathbf{X}^T\mathbf{X}\,\hat{\theta} = \mathbf{X}^T\,\mathbf{y}$

Inferências

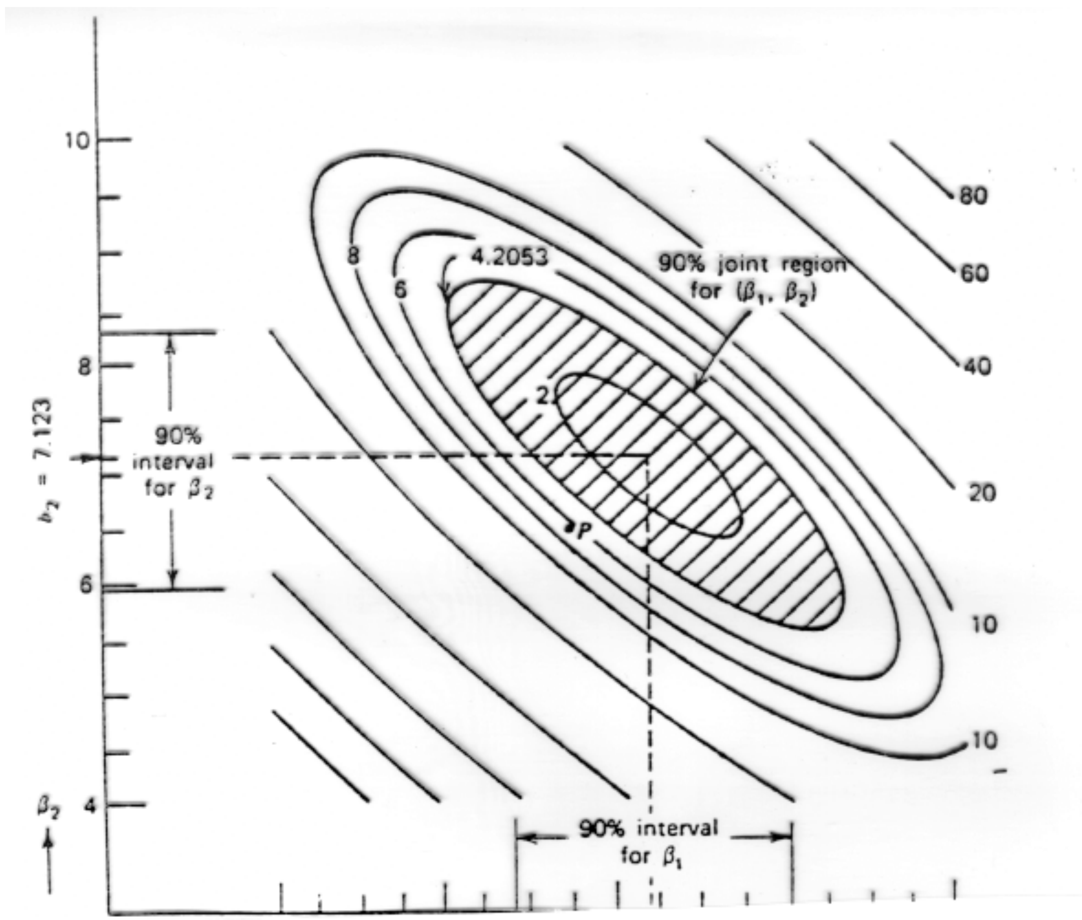
- Uma estimativa é uma realização de uma variável aleatória
- Região Conjunta de Confiança:

$$\left(\theta - \hat{\theta}\right)^T \hat{\mathbf{V}}_\theta^{-1} \left(\theta - \hat{\theta}\right) \leq p\, F\left(p, n-p; 1-\alpha\right)$$

- Intervalo Marginal de Confiança:

$$\theta_i = \hat{\theta}_i \pm \sqrt{\hat{\sigma}^2_{\theta_i}}\; t(n-p, \alpha/2)$$

Inferências



Extraído de Box H. & H.

Inferências

- Um conjunto de confiança é um variável aleatória
- A cada conjunto de experimentos corresponde um conjunto de confiança
- Quanto maior a confiança, maior o conjunto

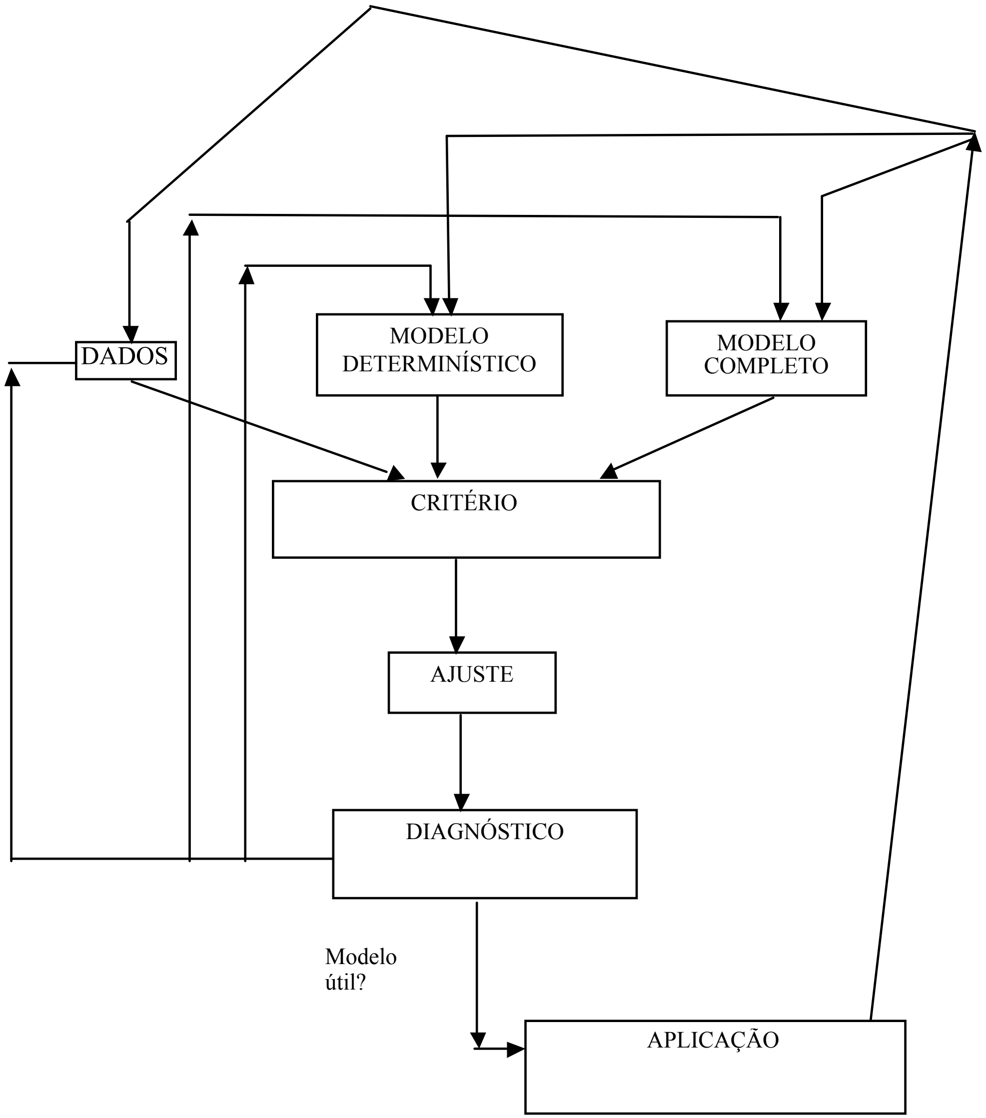
Inferências

- Como faço para validar um modelo?
- Karl Popper, em interpretação livre: "não existem meios de se provar a validade de um modelo, existem apenas meios para provar a sua não validade. Quando da proposição de um modelo científico o único mecanismo para prová-lo é tentar de todo modo invalidá-lo". Neste sentido, basta uma evidência física, e apenas uma, para pôr em cheque toda uma teoria.
- Box: "Nenhum modelo é bom, o importante é que ele seja útil”

Inferências

- Weisberg (1983): "O comportamento de um procedimento de diagnóstico tem que ser conhecido, ao menos aproxidamente, tanto para o modelo correto quanto para o modelo incorreto com alguma hipótese adicional modificada. Os métodos de diagnóstico não devem ser computacionalmente intensivos e de preferência ter sempre um equivalente gráfico e sugerir alguma ação alternativa".

Inferências

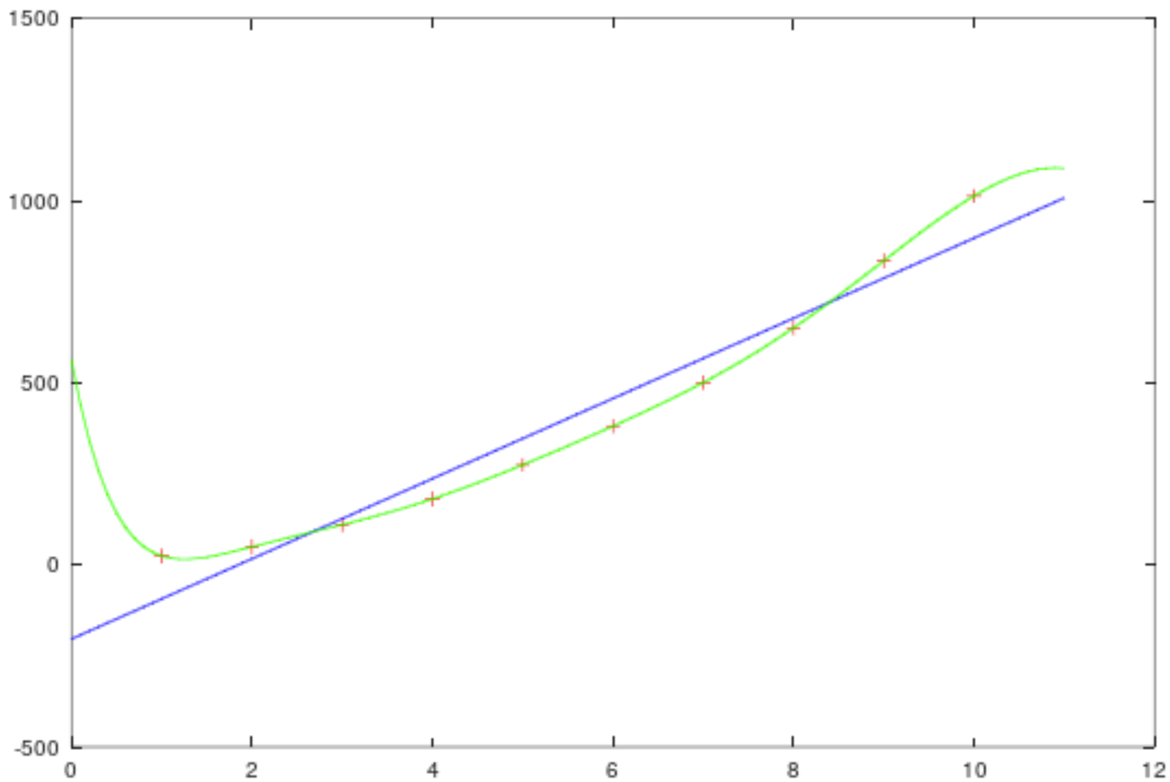


Complexidade e Predição

- Complexidade de um modelo e métodos de ajuste convencionais (mínimos quadrados, máxima verossimilhança ...)
  - Se muitos parâmetros: sobreajuste
  - Se poucos parâmetros: falta de ajuste

Complexidade e Predição

- Verde = grau 9
- Azul = grau 2
- Dados = grau 3



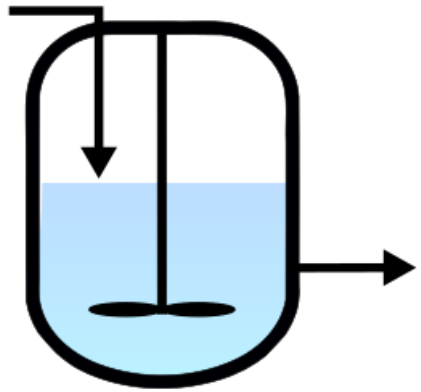
Complexidade e Predição

- Falta de ajuste: não consegue representar o sistema
- Sobre ajuste:
  - parâmetros têm valores esquisitos e variam muito
  - Extrapolação absolutamente errática
  - Aparecem auto-valores muito pequenos na matriz $\mathbf{X}^T\mathbf{X}$

Complexidade e Predição

- Reação Enzimática em CSTR:

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{EP} \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{E} + \mathrm{P}.$$



Complexidade e Predição

- Reação Enzimática em CSTR:

$$
\begin{aligned}
-k_1 x_E x_S + k_{-1} x_{ES} + F(x_S^{in} - x_S) &= 0 \\
-k_1 x_E x_S + k_{-1} x_{ES} + k_3 x_{EP} - k_{-3} x_E x_P + F(x_E^{in} - x_E) &= 0 \\
k_1 x_E x_S - k_{-1} x_{ES} - k_2 x_{ES} + F(x_{ES}^{in} - x_{ES}) &= 0 \\
k_2 x_{ES} - k_3 x_{EP} + k_{-3} x_E x_P + F(x_{EP}^{in} - x_{EP}) &= 0 \\
k_3 x_{EP} - k_{-3} x_E x_P + F(x_P^{in} - x_P) &= 0
\end{aligned}
$$

Complexidade e Predição

- Caso I

  $k_{-1} = k_2 = k_3 = k_{-3} = 1$ $\qquad$ $k_1 = 1.2$

  Vazões de alimentação: 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1 e 2.5

  Matriz X e Y no arquivo caso1.m

Complexidade e Predição

- Caso II

  $k_1 = 1.2\ 10^9$

  $k_{-1} = 10^9$

  $k_2 = k_3 = k_{-3} = 1$

  Matriz X e Y no arquivo caso2.m

  Impossível obter uma estimativa sem nenhum tratamento ou consideração.

Complexidade e Predição

- Diagnóstico: IDENTIFICABILIDADE
- Classifica-se na literatura:
  – Identificabilidade estrutural: inerente ao modelo
  – Identificabilidade qualitativa: depende do valor dos parâmetros e da qualidade das medições

Complexidade e Predição

- Não identificabilidade estrutural
- $\mathbf{X}^T\mathbf{X}$ singular
- Exemplo: $y = a\, x_1 + b\, x_2$

  mas $x_2 = 1 - x_1$ sempre

Complexidade e Predição

- Modelo é estruturalmente identificável
- $\mathbf{X}^T\mathbf{X}$ mal condicionada (autovalores pequenos)
- Exemplo: $y = a\,(x_1 + b\,x_2)$

dependendo da ordem de grandeza de b, x_1 e x_2

o sistema fica muito mal condicionado

Complexidade e Predição

- O que fazer?
  - Simplificar o modelo
  - Medir mais coisas (por exemplo 13C)
  - Usar métodos automáticos que permitem predições adequadas

Complexidade e Predição

- Um dos problemas é que quanto mais complexo o modelo, mais ele pode ajustar os dados
- Por exemplo, no caso do polinômio, qual o grau do polinômio a escolher?

Complexidade e Predição

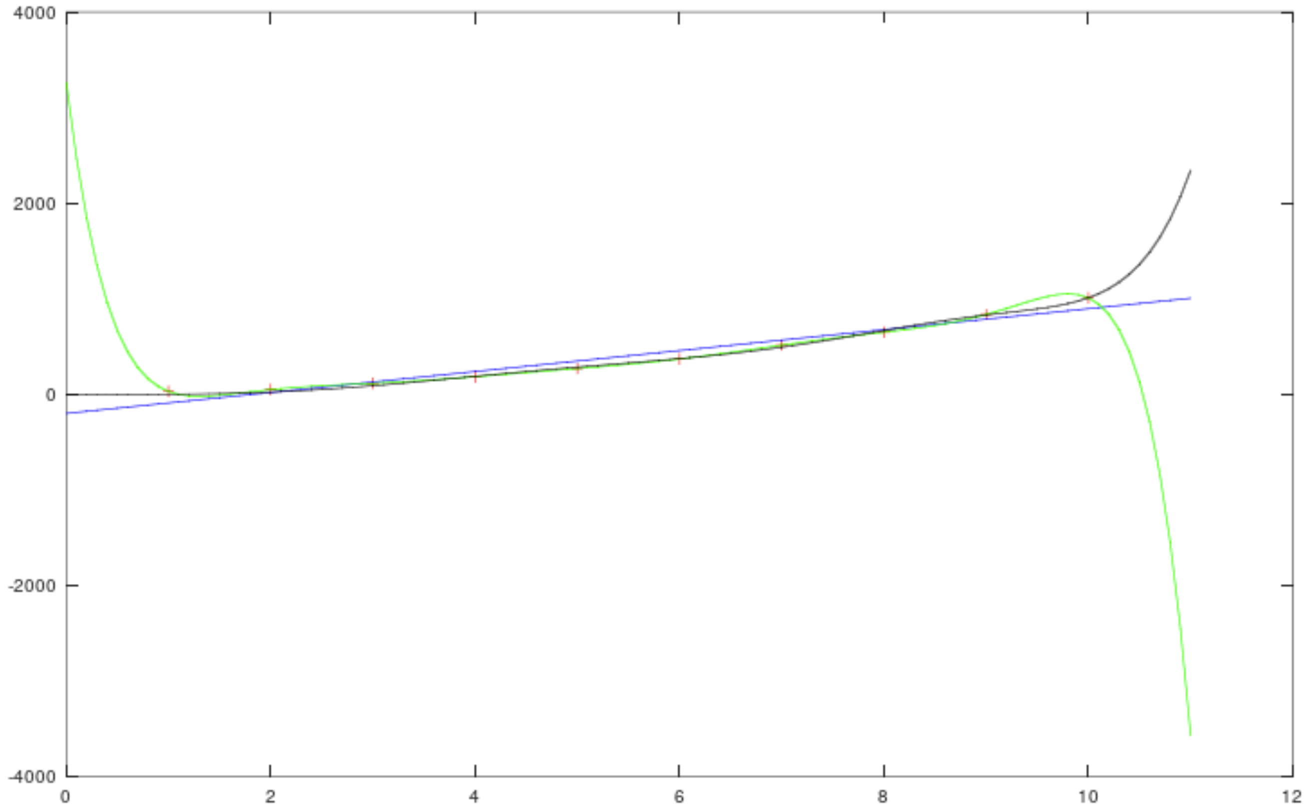
- No aprendizado de máquina, assim como já foi com a quimiometria, e outras técnicas preditivas em que se usam modelos qualitativamente complexos
- A técnica mais comum é utilizar a validação cruzada:
  - Um conjunto de dados é utilizado para ajustar (treinar, aprender) o modelo e outro para testar (validar)

Complexidade e Predição

- Exemplo 1
- técnica para evitar sobreajuste
- Ridge-regression: $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathrm{k}\,\mathbf{I}\right)\hat{\theta}^{\mathrm{RR}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$
- k é um parâmetro chamado antigamente de parâmetro de bias, hoje em dia é chamado de "hiperparâmetro"

Complexidade e Predição

- Preto: ridge com k = 0.1



Complexidade e Predição

- Exemplo 2: método automático para descobrir aonde podem se aplicar simplificações em modelos complexos
- Baseado em decomposição da matriz $\mathbf{X}^T\mathbf{X}$
- A análise diz que as constantes k_1 e k_{-1} estão envolvidas em um equilíbrio

	$\tilde{V}_1$ (EN-OC)			
	$\tilde{v}_1$	$\tilde{v}_2$	$\tilde{v}_3$	$\tilde{v}_4$
k_1	0	-0.640	0	0
k_{-1}	0	0.768	0	0
k_2	0	0	-1	0
k_3	-1	0	0	0
k_{-3}	0	0	0	1

Colocações Finais

- A dialética modelos x experimentos ainda continua tendo uma importância central no desenvolvimento do conhecimento
- Não existe solução universal. Este problema já gerou polêmicas enormes na história
- A construção de modelos é uma tarefa complexa e continuará sendo