Quimiometria e Análise de Dados Funcionais

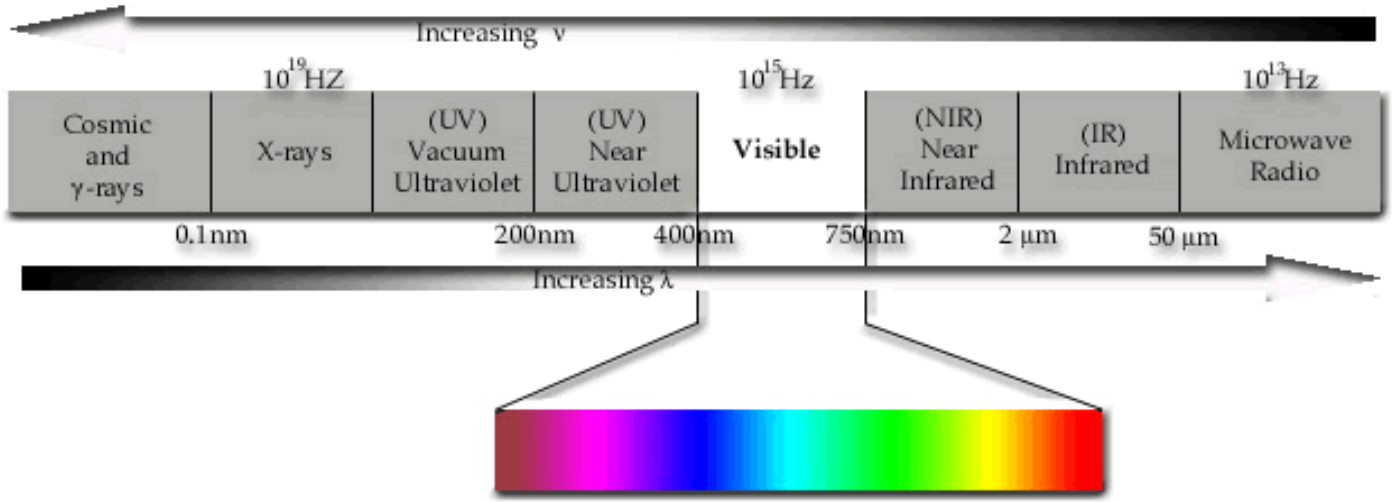
Nancy L. Garcia[1,2]

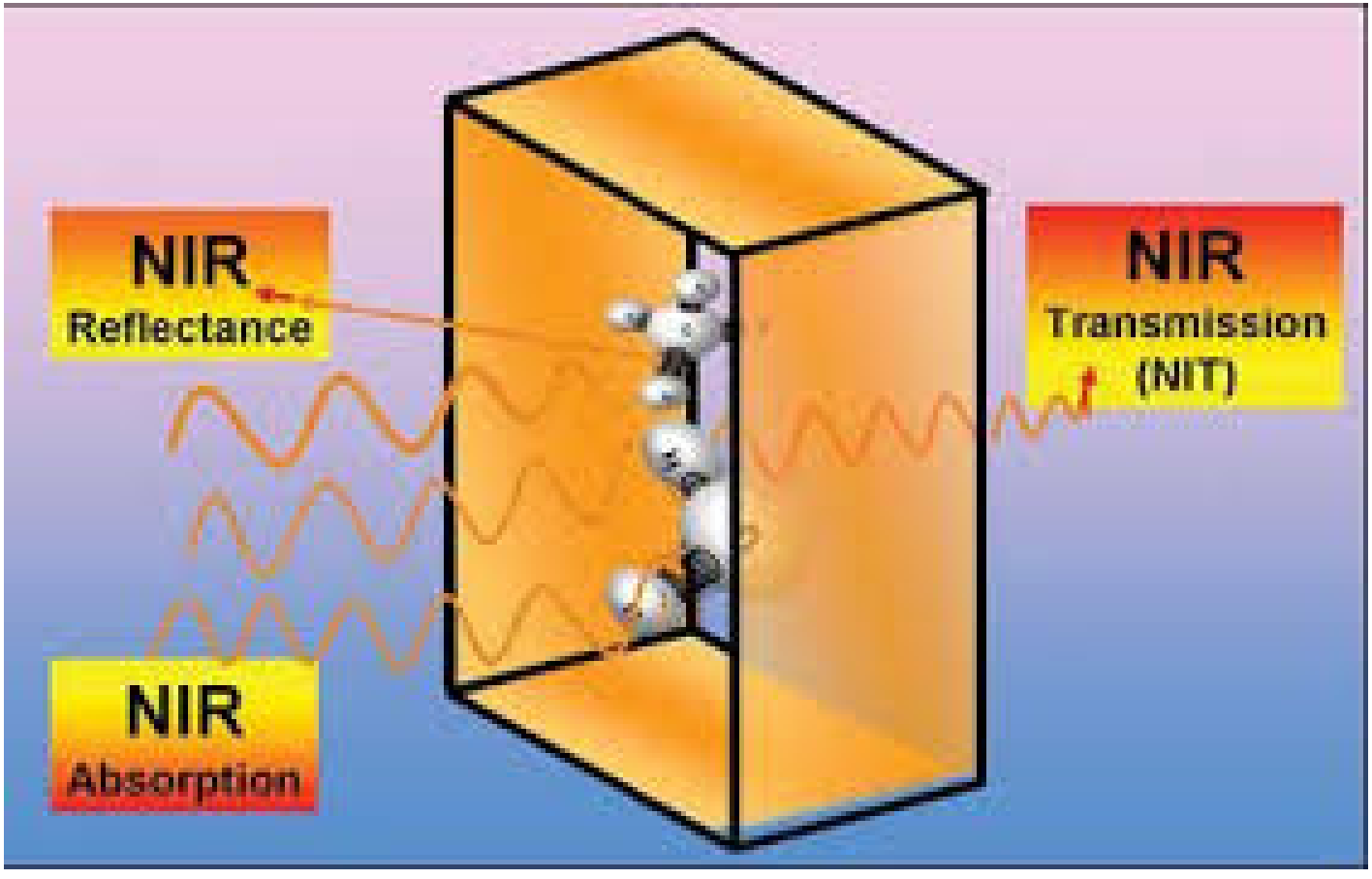
[1]IMECC/UNICAMP

[2]trabalho conjunto com Alex Rodrigo, Alexandra Schmidt, Guilherme Ludwig, Mariana Rodrigues Motta, Marley Saraiva e Ronaldo Dias

29 de Maio - Dia do Estatístico
Maringá

Comprimento de onda em nanômetros:

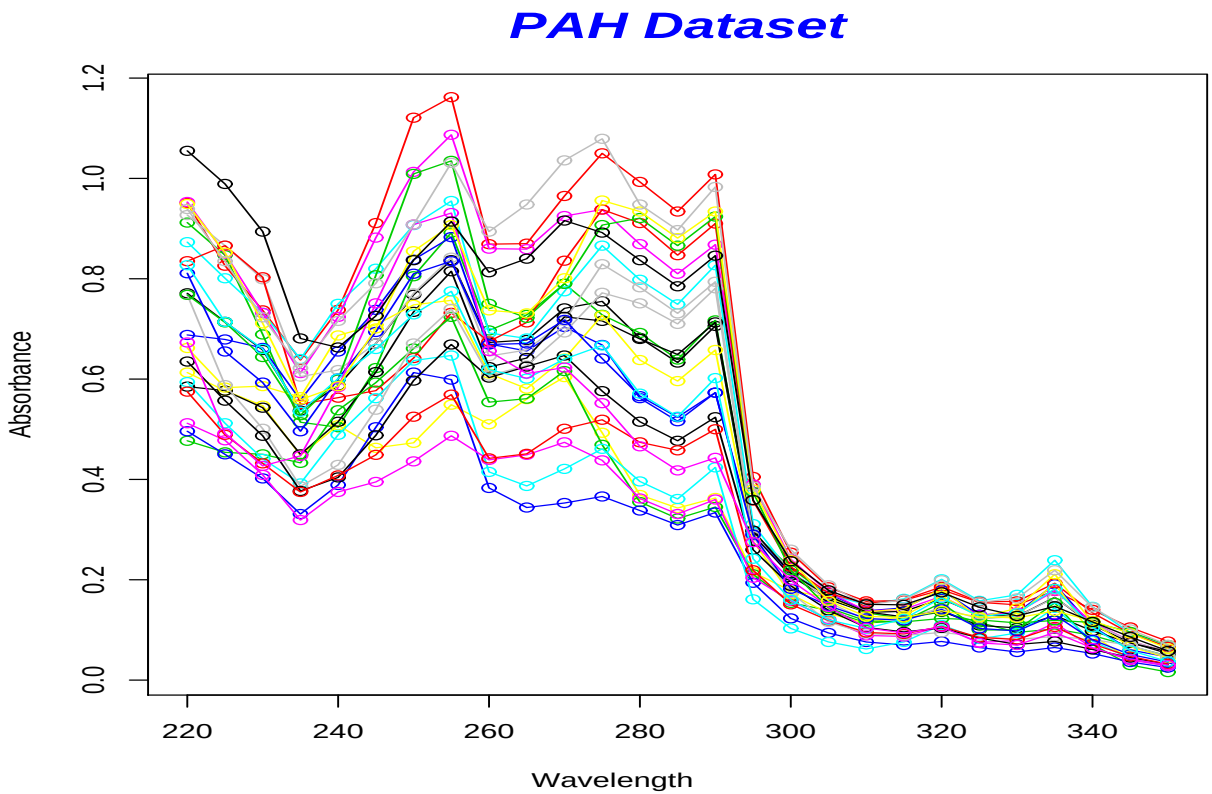



NIR
Reflectance
NIR
Transmission
(NIT)
NIR
Absorption

- Quando uma substância é submetida a diferentes comprimentos de onda, os sobretons e combinações na banda do InfraVermelho Próximo (NIR) irá produzir padrões muito complexos que caracterizam os componentes da amostra.
- As diferentes características dos constituintes da amostra se sobrepõem e dão origem a uma curva que é uma soma de várias curvas dependendo da concentração de cada constituinte.

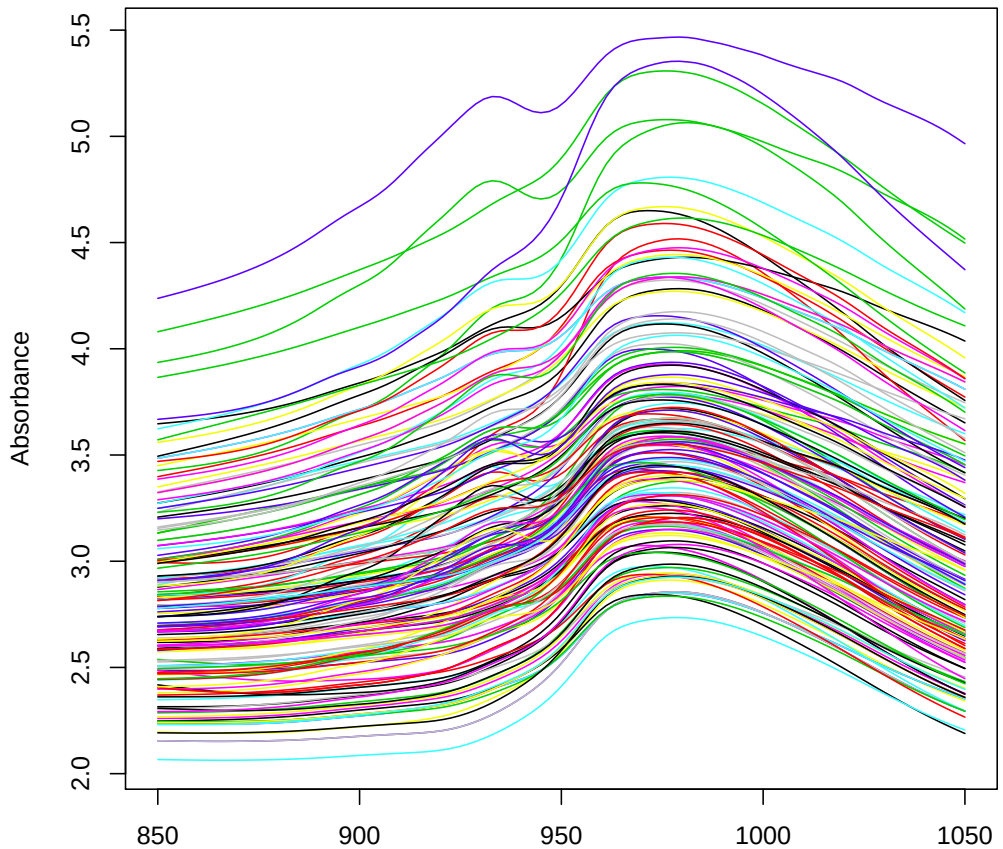
Hidrocarbonetos poliaromáticos – Polyaromatic hydrocarbons(PAH)

- EAS (Electronic Absorption Spectroscopy)
- 25 amostras químicas
- Cada amostra com 10 elementos químicos diferentes (constituintes)
- 27 comprimentos de onda (220 nm – 350 nm)


PAH Dataset
Absorbance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
220
240
260
280
300
320
340
Wavelength

Tecator

- Cada amostra contém carne moída pura, com diferentes quantidades de umidade, gordura e proteína.
- Conjunto de aprendizado: 129 + 43 amostras
- Conjunto de teste: 43 amostras.
- Cada amostra contém 3 constituientes: umidade, gordura e proteína
- 100 comprimentos de onda (850nm – 1050nm)
- Disponível em domínio público sem qualquer responsabilidade da fonte de dados original
  http://lib.stat.cmu.edu/datasets/tecator



Training + Monitoring data (129+43 observations)

Problema de calibração

- Existem várias técnicas para medir os constituintes de uma amostra.
- Alguns deles são muito precisos e caros (técnicas analíticas).
- Alguns deles não são tão precisas, mas baratas (por exemplo, NIR espectroscopia).
- Para usar técnicas de baixo custo é necessário calibrar o instrumento.
- Por razões práticas, só é possível medir os dados espectrais em um número finito de comprimentos de onda $t_1 < t_2 < \ldots < T_T$. Muitas vezes T está por volta de 100–200 ou mais.

Problema de predição

- Depois de calibrar o instrumento
- Tomar novas medidas
- Estimar as concetrações na nova amostra.

- Uma amostra química é constituída de vários constituintes ($\ell = 1, \ldots, m$).
- Cada constituinte de interesse é chamado de **analito**.
- y_ℓ: Concetração do analito ℓ
- Amostra é **fechada** se

$$y_1 + \ldots + y_m = 1$$

  i.e. quando todos os constituintes são analisados
- absorbância do analito ℓ no comprimento de onda t será denotado por

$$\alpha_\ell(t)$$

- Assuma que temos n amostra de composições variadas
- Conjunto de dados Y:
  $y_{i,\ell}$: concentração (medida através de um método de referência - padrão ouro) para analito ℓ na i-ésima amostra
- **Dados espectrais**: Absorbâncias

$$W(t_1), \ldots, W(t_T)$$

medidas através de NIR espectroscopia em T comprimentos de onda $t_1 < t_2 < \ldots < t_T$.

O modelo: Lei de Beer-Lambert

- Para a i-ésima amostra fechada composta de m constituintes e T comprimentos de onda,

$$W_i(t_j) = \sum_{\ell=1}^{K} y_{i,\ell}\alpha_\ell(t_j) + \epsilon(t_j), \text{para} j = 1, \ldots, T$$

- Geralmente somente um subconjunto de constituintes é analisado. Neste caso, não temos a restrição $y_{i,1} + \ldots + y_{i,m} = 1$ mas uma modificação da Lei de Beer-Lambert pode ser usada

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell} a_\ell(t_j) + \epsilon(t_j), \text{ para } j = 1, \ldots, T$$

onde $\epsilon(t)$ é um processo Gaussiano com função de covariância dada por $\sigma(s,t) = \mathrm{Cov}(\epsilon(s), \epsilon(t))$.

- Modelo de Calibração

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}\alpha_\ell(t_j) + \epsilon_i(t_j), \text{ para } j = 1, \ldots, T$$

- Modelo de Predição

$$W_i^*(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}^*\alpha_\ell(t_j) + \epsilon^* i(t_j), \text{ para } j = 1, \ldots, T$$

Problema de calibração Dados $y_{i,\ell}$ e $W_i(t_j)$ estimar $\alpha_\ell(t)$.
Problema de predição Dado um novo conjunto de dados $W_i^*(t_j)$, estimar $y_{i,\ell}^*$.

Expansão em bases

Assuma que existe um inteiro positivo K e uma sequência de *knots* ξ tal que

$$\alpha_\ell(t) = \sum_{k=1}^{K} \beta_{\ell,k} B_k(t),$$

onde $B_k(t)$, $k = 1, \ldots, K$ são B-splines cúbicos.

Estrutura de covariância

- Assumimos que a correlação entre os pontos $W_i(t)$ e $W_i(s)$ decai exponencialmente quando $|t - s|$ cresce

$$\sigma^2 \exp(-\phi|t - s|)$$

.

Aprendizado Supervisionado.

- Datos: $W_i(t), i = 1, \ldots, I$ observados em $t = t_1, \ldots, t_T$.
- Constantes conhecidas: matriz Y com linhas linermente independentes contendo as concentraç̧oes medidas por um método de referência $y_{i,\ell}$ para $i = 1, \ldots, I$ e $\ell = 1, \ldots, m$
- Dado o modelo:

$$W_i(t) = \sum_{\ell=1}^{m} y_{i,\ell} \sum_{k=1}^{K} \beta_{k,\ell} B_k(t) + \epsilon_i(t)$$

- Estimar: $\beta_{\ell,k}$, $k = 1, \ldots, K$ and $\ell = 1, \ldots, m$, σ^2 e ϕ.

Basis smoothing:

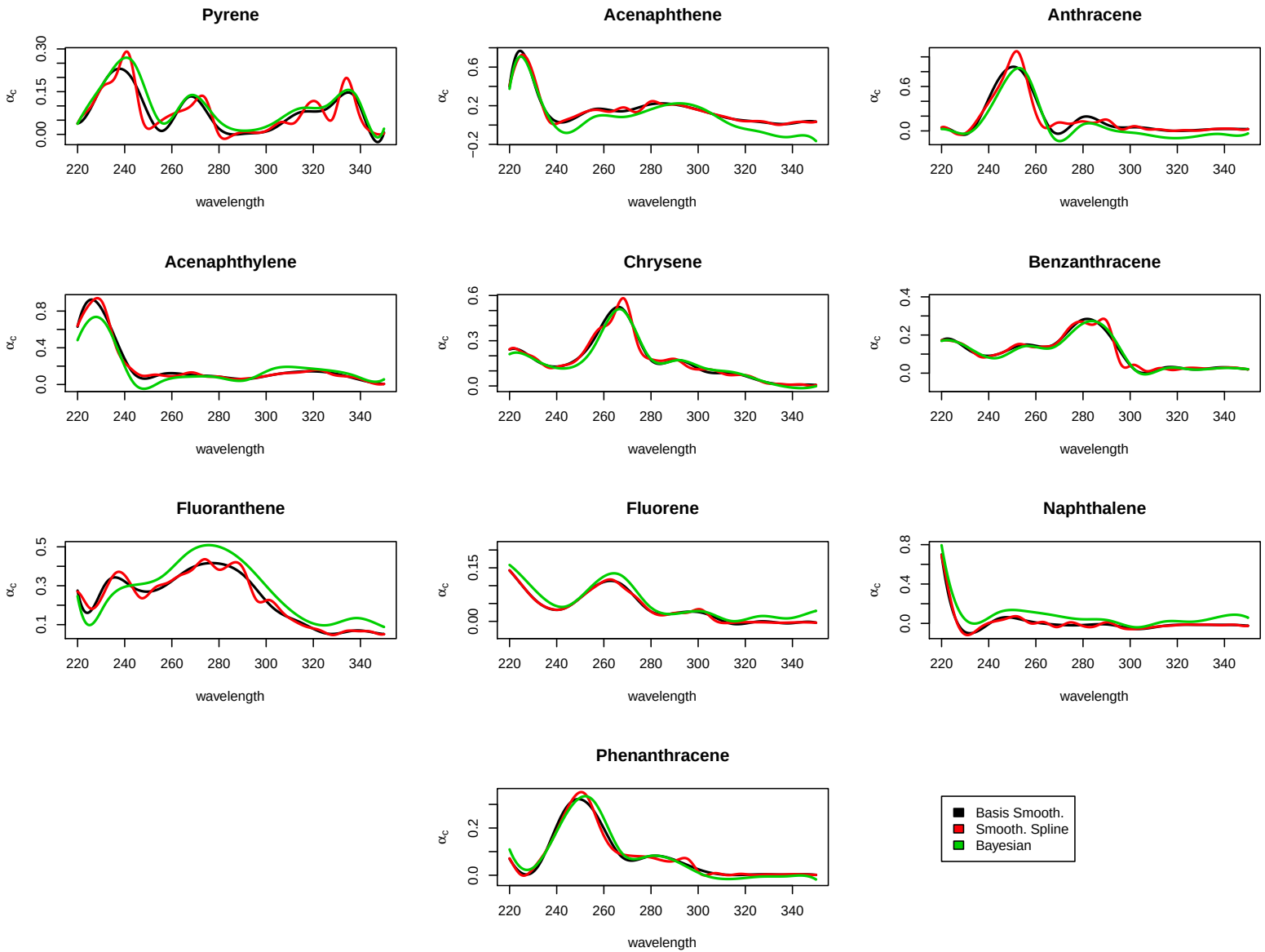
- K é escolhido de maneira *ad-hoc*
- β pode ser facilmente estimado pro mínimos quadrados, com solução explícita

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W},$$

Smoothing splines:

- Cada observação é um *knot*
- Número de coeficientes a serem estimadosé maior queo número de observações.
- Suavização: penalizar a norma quadrada da secunda derivada das bases spline no problema de mínimos quadrados:
- $$\hat{\beta} = (\mathbf{X}^{+\prime}\mathbf{X}^{+} + \lambda \mathbf{I}_{m\times m} \otimes \mathbf{R})^{-1}\mathbf{X}^{+\prime}\mathbf{W}^{+}.$$

  onde $\mathbf{R}$ é uma matriz com entradas $\mathbf{R}_{i,j} = \int_{-\infty}^{\infty} D^2 B_i(t) D^2 B_j(t) \mathrm{d}t$, e D^2 é o operador diferencial de segunda ordem $\partial^2/\partial t^2$.


Pyrene
Acenaphthene
Anthracene
Acenaphthylene
Chrysene
Benzanthracene
Fluoranthene
Fluorene
Naphthalene
Phenanthracene
wavelength
Basis Smooth.
Smooth. Spline
Bayesian

- Abordagem Bayesiana segue bastante próxima de *Basis Smoothing*
- *Smoothing Spline* estimativas são menos suaves e tem mais “bumps” e captura mais variação local dos dados
- Brereton(2003): curvas são amostradas em um conjunto de comprimentos de onda mais esparso e isto causa uma redução no ruído.
- Portanto, a maior parte da variação local nas curvas são características importantes da variação local dos dados e precisam ser identificadas.

Relembrando:

- Modelo de calibração

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}\alpha_\ell(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \ldots, T$$

- Modelo de Predição:

$$W_i^*(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}^*\alpha_\ell(t_j) + \epsilon_i^*(t_j), \text{ for } j = 1, \ldots, T$$

Problema de calibração Dados $y_{i,\ell}$ e $W_i(t_j)$ estimar $\alpha_\ell(t_j)$.
PProblema de predição Dado novo conjunto de dados $W_i^*(t_j)$, estimar $y_{i,\ell}^*$.

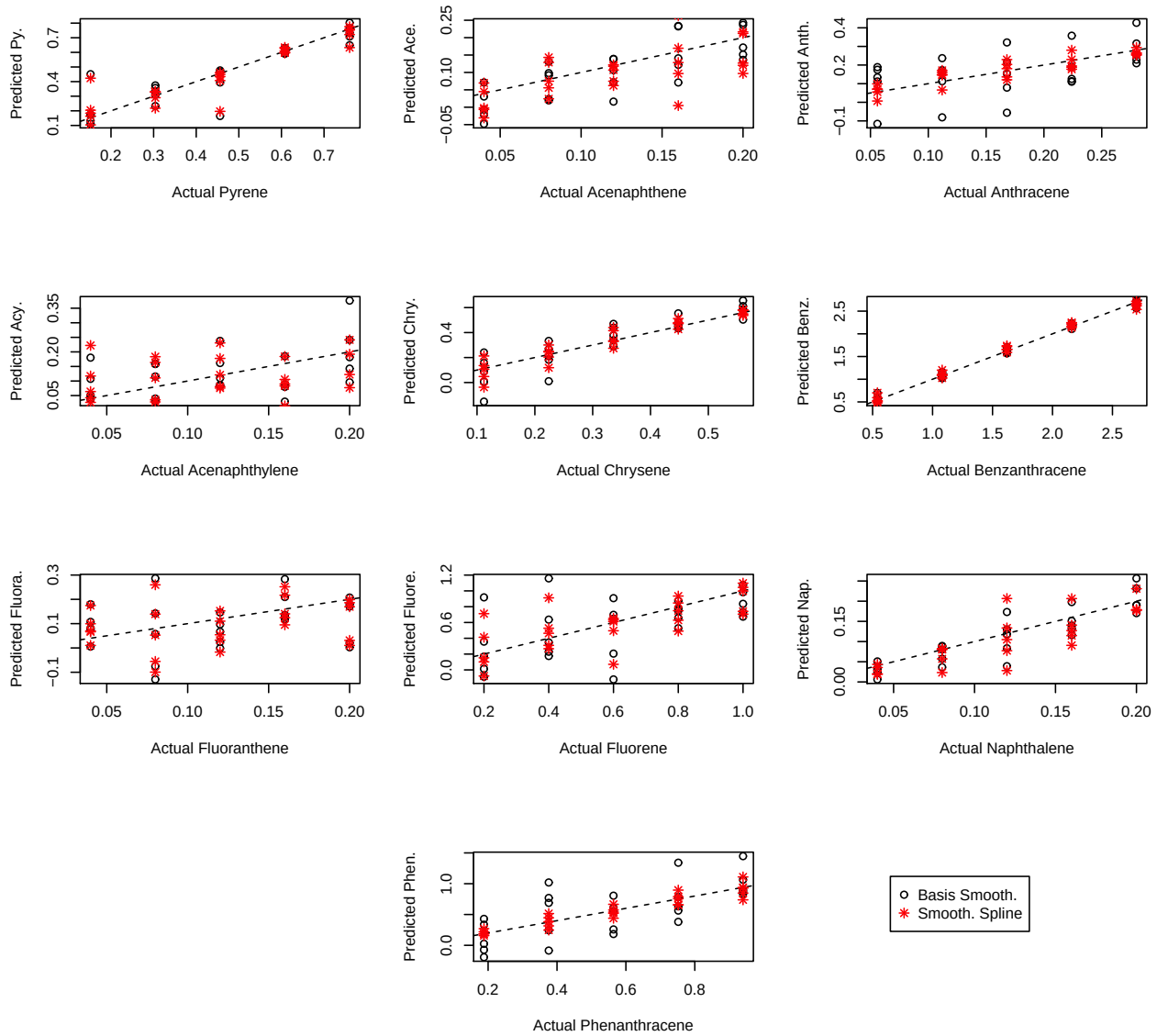
Abordagem simples:

- Useos espectros estimados $\hat{\alpha}_\ell$
- Coloque-os no modelo

$$W_i^*(t_j) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}^* \alpha_\ell(t_j) + \epsilon_i(t_j)$$

e encontre o conjunto de $\mathbf{y}_j^*$ que minimizam a soma de quadrados dos erros de predição

$$\hat{\mathbf{y}}_j^* = \arg\min_{\mathbf{y}_j^*} \sum_{r=1}^{T} \sum_{\ell=1}^{m} \left(W_j^*(t_r) - \hat{y}_{j,\ell}^* \hat{\alpha}_\ell(t_r) \right)^2 .$$


Predicted Py.
Actual Pyrene
Predicted Ace.
Actual Acenaphthene
Predicted Anth.
Actual Anthracene
Predicted Acy.
Actual Acenaphthylene
Predicted Chry.
Actual Chrysene
Predicted Benz.
Actual Benzanthracene
Predicted Fluora.
Actual Fluoranthene
Predicted Fluore.
Actual Fluorene
Predicted Nap.
Actual Naphthalene
Predicted Phen.
Actual Phenanthracene
Basis Smooth.
Smooth. Spline

Estimação simultânea

- Modelo de calibração:
  $W_i(t) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}\alpha_\ell(t) + \epsilon_i(t)$, for $i = 1, \ldots, I$
- Modelo de predição
  $W_i^*(t) = \alpha_j + \sum_{\ell=1}^{K} y_{i,\ell}^*\alpha_\ell(t) + \epsilon^*(t)$, for $i = 1, \ldots, I^*$
- Considere como dados:
  1. $W_i(t_j)$ para $i = 1, \ldots, I$, $j = 1, \ldots, T$
  2. $y_{i,\ell}$ para $i = 1, \ldots, I$, e $\ell = 1, \cdots, C$
  3. $W_i^*(s_j)$ para $i = 1, \ldots, I^*$, $j = 1, \ldots, T^*$
- Considere como parâmetros
  1. β tais que $\alpha_c(t) = \sum_{k=1}^{K} \beta_{c,k} B_k(t)$,
  2. σ^2 e ϕ parâmetros da covariância
  3. $y_{i,\ell}^*$ para $i = 1, \ldots, I^*$, e $\ell = 1, \cdots, C$

Erro padrão de predição – Standard Error of Prediction (SEP)

A contribuição do componente ℓ é dada por

$$\mathrm{SEP}_\ell = \left(\frac{1}{J-1} \sum_{j=1}^{J} \left(y^*_{j,\ell} - \hat{y}^*_{j,\ell} \right) \right)^{1/2}$$

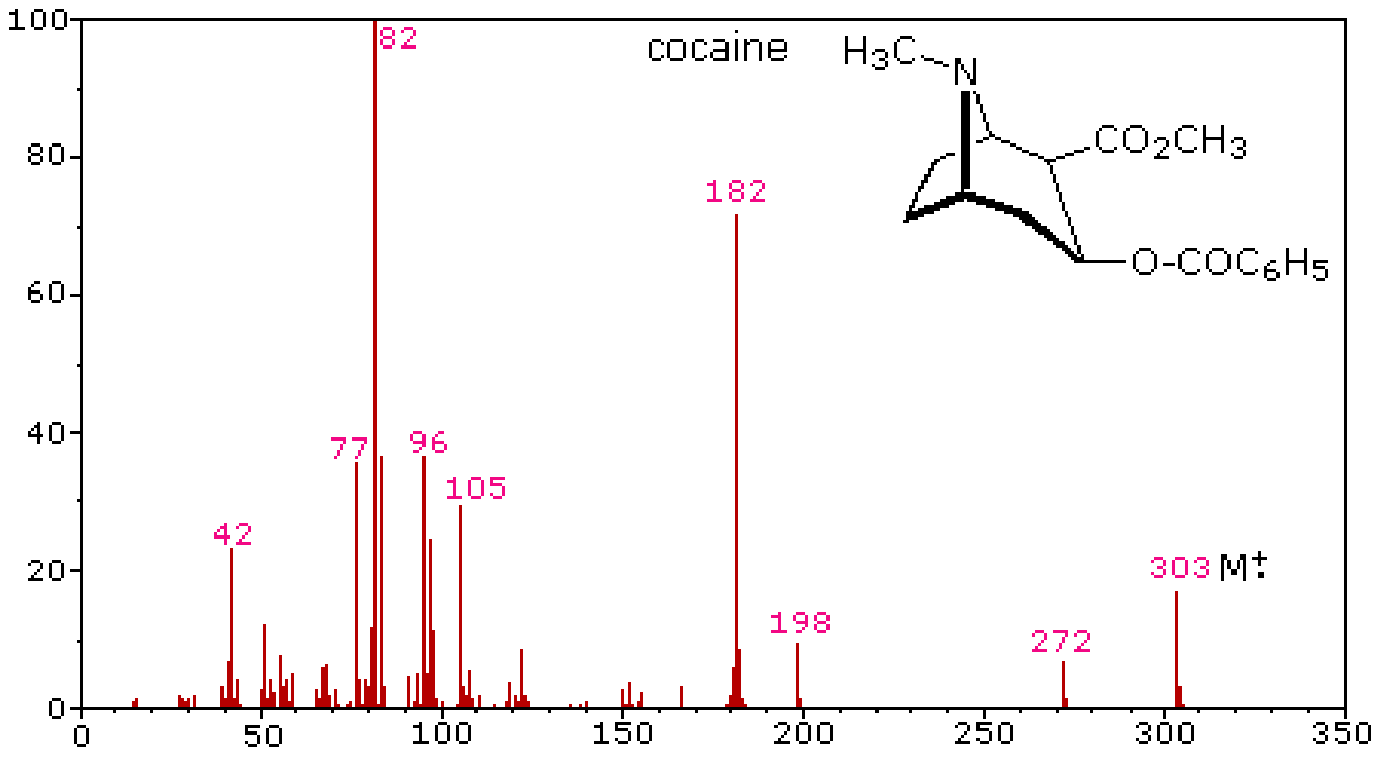
e o SEP total será dado por

$$\mathrm{SEP} = \left(\frac{1}{mJ-1} \sum_{\ell=1}^{m} \sum_{j=1}^{J} \left(y^*_{j,\ell} - \hat{y}^*_{j,\ell} \right) \right)^{1/2}$$

SEP (mg/l) para PAH.

Constituente	MLR	PCR	PLS	OLS-K	OLS-SS	ML
Pyrene	0.09	0.09	0.09	0.09	0.09	0.06
Acenaphthene	0.06	0.03	0.03	0.05	0.06	0.03
Anthracene	0.03	0.03	0.03	0.11	0.04	0.01
Acenaphthylene	0.09	0.06	0.06	0.08	0.09	0.03
Chrysene	0.06	0.05	0.04	0.10	0.06	0.02
Benzanthracene	0.07	0.07	0.06	0.06	0.07	0.04
Fluoranthene	0.09	0.08	0.08	0.10	0.09	0.04
Fluorene	0.24	0.20	0.15	0.32	0.23	0.07
Naphthalene	0.04	0.03	0.04	0.03	0.04	0.03
Phenanthracene	0.08	0.09	0.08	0.31	0.09	0.06
Total	0.11	0.09	0.08	0.16	0.10	0.05

CSI

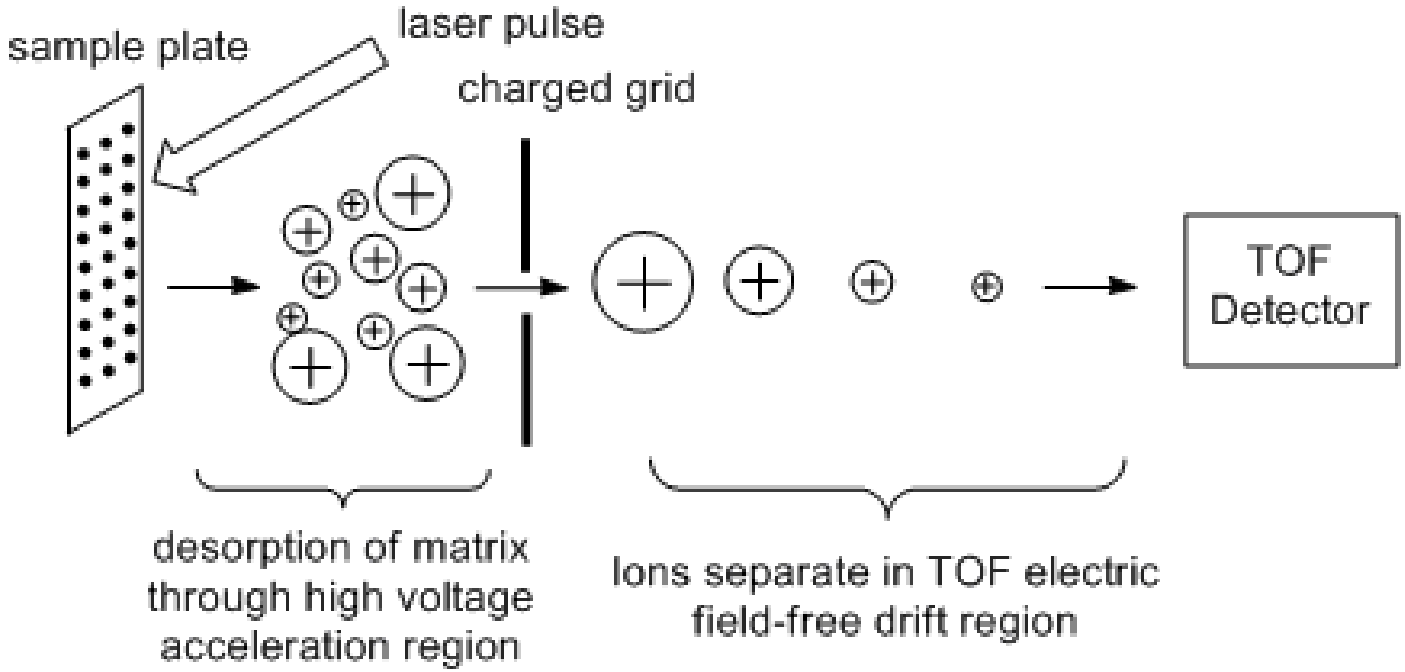

cocaine
H3C
N
CO2CH3
O-COC6H5
82
182
77
96
105
42
303
M+
198
272
100
80
60
40
20
0
0
50
100
150
200
250
300
350

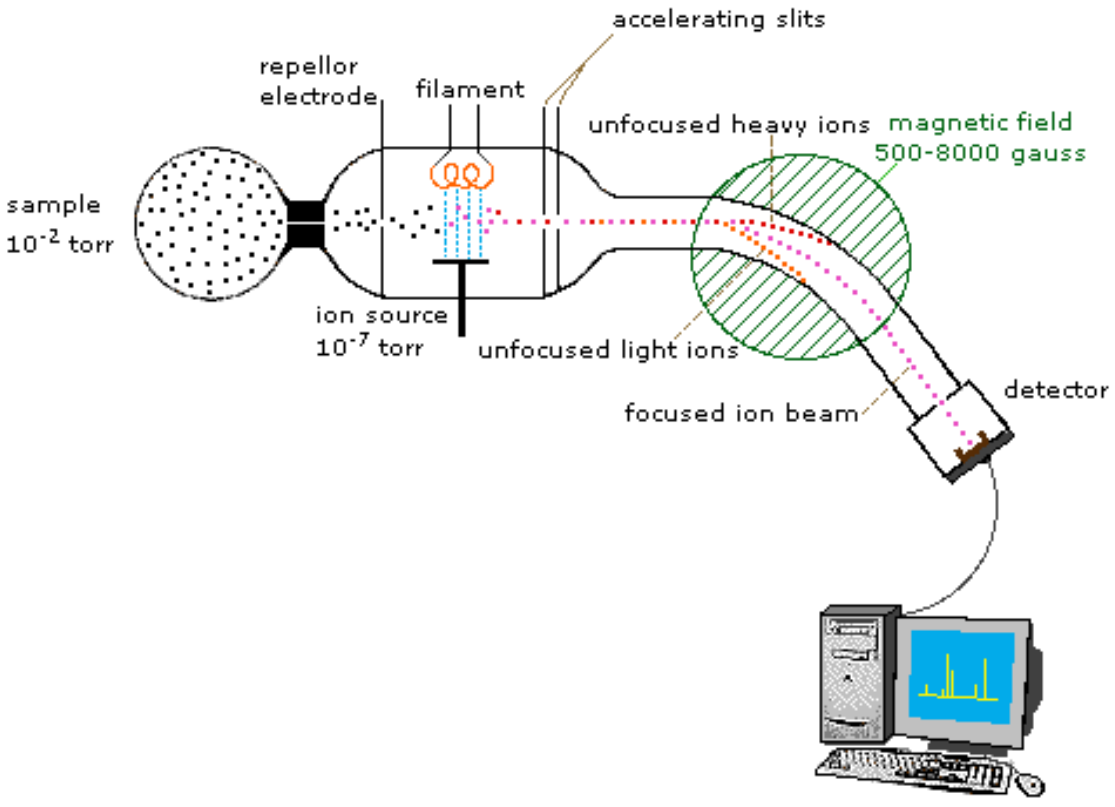
Espectrometria de massa

- A amostra, a qual pode ser sólida, líquida, ou gás, é ionizada, por exemplo, através do bombardeamento com electrons.
- Isso pode fazer com que algumas moléculas da amostra se quebrem em fragmentos carregados eletricamente.
- Estes ionss são então separados de acordo com a sua razão massa/carga, tipicamente acelerando-as e submetendo-as a um campo elétrico ou magnético: ions de mesma razão massa/carga vão sofrer a mesma quantidade de deflexão.
- Os ions são detectados por um mecanismo capaz de detectar partículas carregadas, tais como um multiplicador de electrons.

Espectro de massa

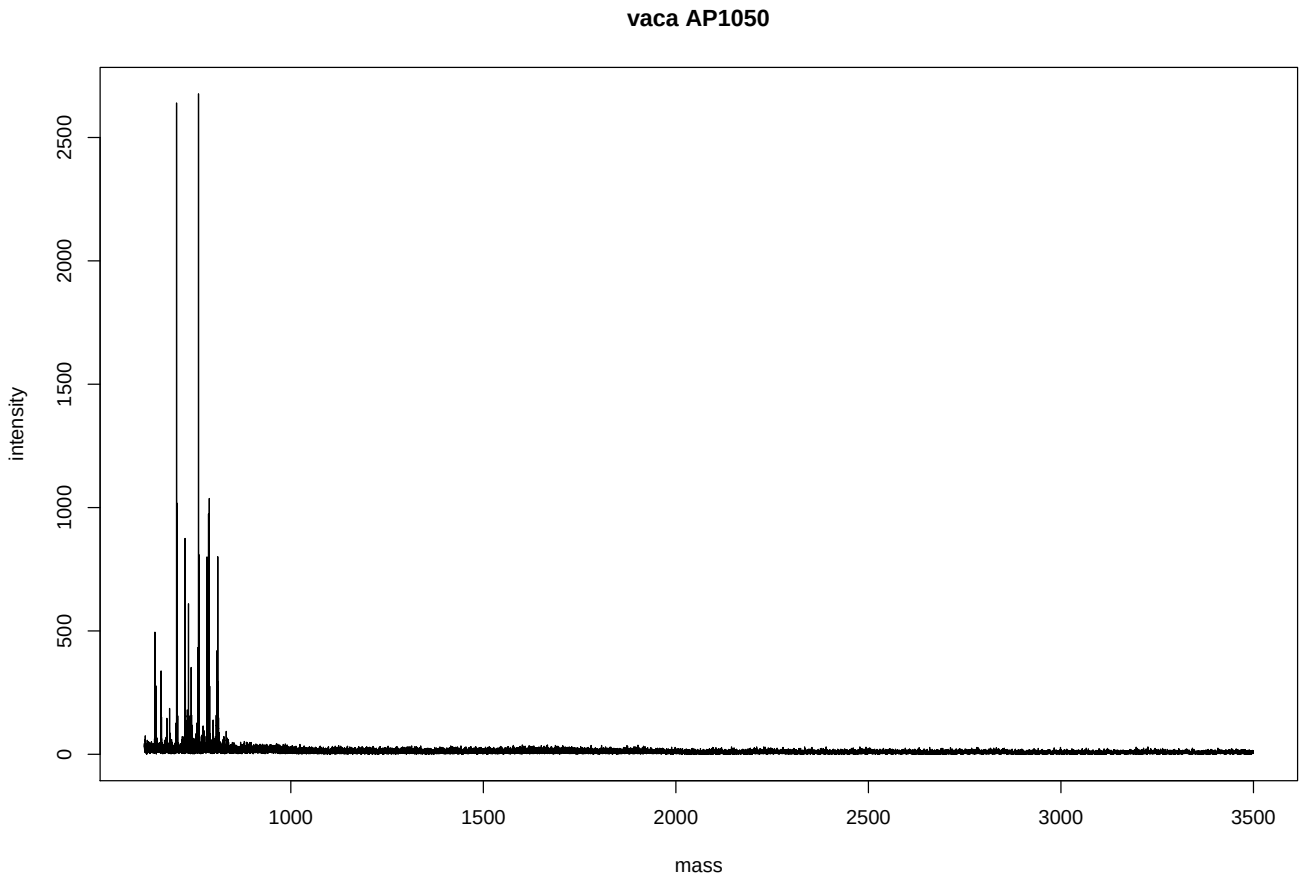
- Os resultados são apresentados como espectros da abundância relativa dos ions detectados como função da razão de massa/carga.
- Os átomos ou moléculas na amostra podem ser identificados por meio da correlação entre massas conhecidas para as massas identificadas ou por meio de um esquema de fragmentação característico.


sample plate
laser pulse
charged grid
TOF
Detector
desorption of matrix
through high voltage
acceleration region
Ions separate in TOF electric
field-free drift region

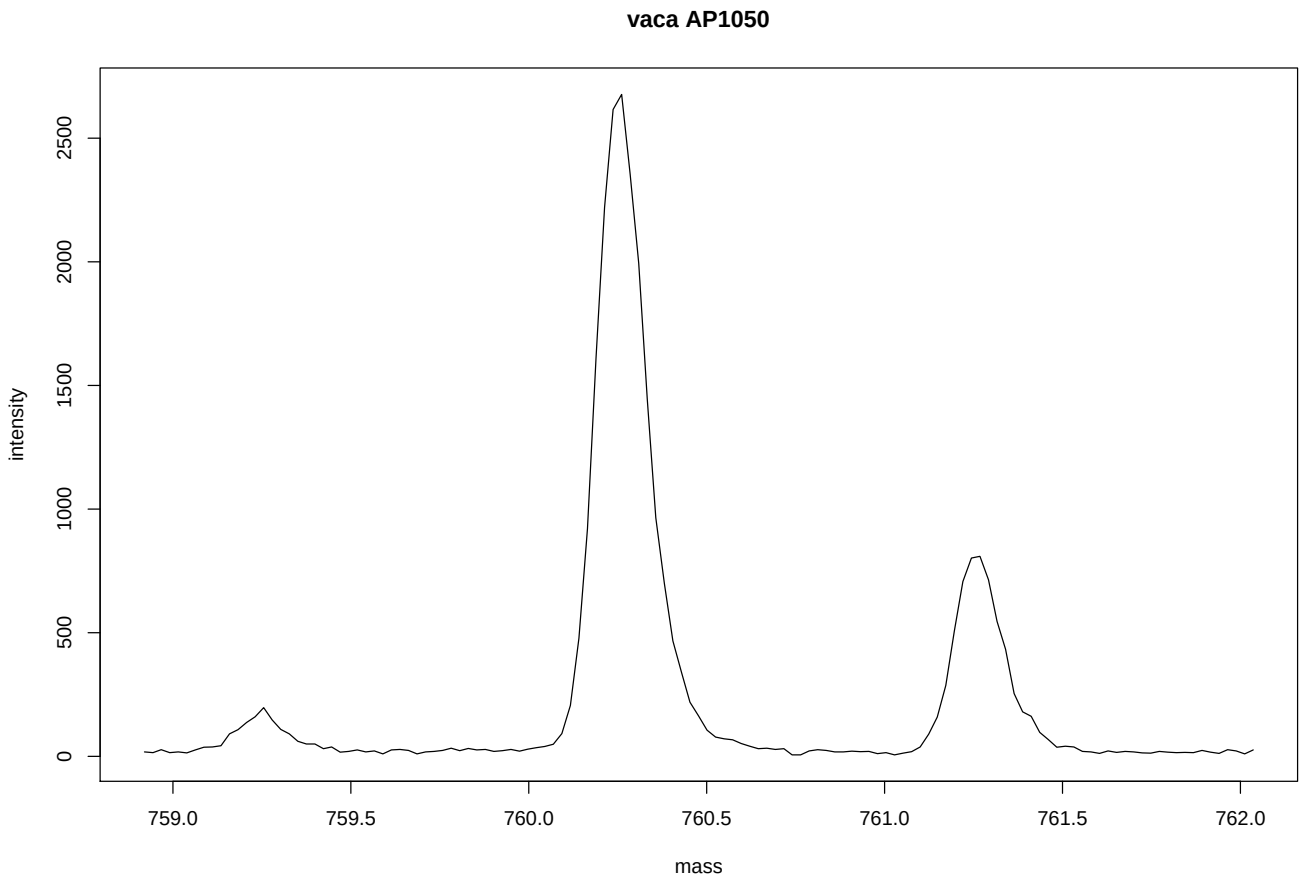

accelerating slits
repellor electrode
filament
unfocused heavy ions
magnetic field 500-8000 gauss
sample 10^{-2} torr
ion source 10^{-7} torr
unfocused light ions
detector
focused ion beam

Marcador biológico para progesterona

- 17 espectros de massa para dois grupos de vacas: AP (alta progesterona) e BP (baixa progesterona).
- Para cada vaca temos duas amostras (medidas repetidas)
- Problema muito importante: insiminação artificial
- Prof. Mário Binelli (Faculdade de Medicina Veterinária e Zootecnia da USP de Pirassununga)


vaca AP1050
intensity
0
500
1000
1500
2000
2500
1000
1500
2000
2500
3000
3500
mass

Zoom: 759 a 762 m/z



Dados de útero de vaca

- 7 (14) curvas para grupo AP $W_{i,AP}(t), i = 1, \ldots, 7$
- 10 (20) curvas for grupo BP $W_{j,BP}(t), j = 1, \ldots, 10$

O modelo: Assuma que existam conjuntos C_1 e C_2 tais que

$$W_{i,AP} = f(t) + \sum_{x \in C_1} R_x^{AP} K\left(\frac{t-x}{\sigma_x^{AP}}\right) + \epsilon_{i,AP}(t)$$

e

$$W_{j,BP} = f(t) + \sum_{y \in C_2} R_y^{BP} K\left(\frac{t-y}{\sigma_y^{BP}}\right) + \epsilon_{j,BP}(t)$$

para $t \in [619, 3500]$, onde $\epsilon_{i,AP}(.)$ e $\epsilon_{i,BP}(.)$são processos Gaussianos independentes, $f(t)$ is é uma fun¸ao comum (baseline), e $K(t)$ é uma função kernel (eg Gaussiana).

Os objetivos:

- Identificar os conjuntos C_1 e C_2
- Estimar os parâmetros f, R_x^{AP} e σ_x^{AP}, para $x \in C_1$ bem como R_y^{BP} e σ_y^{BP}, para $y \in C_2$.
- Finalmente identificar $C \subset C_1 \cup C_2$ tais que $R_x^{AP} \neq R_x^{BP}$ para $x \in C$

Muito obrigada!