Wavelet-based clustering for mixed-effects functional models in high dimension.

Franck Picard, LBBE - Lyon

Madison Giacofci LJK (Grenoble)
Sophie Lambert-Lacroix (TIMC - Grenoble)
Guillemette Marot (Univ. Lille 2)

Outline

Functional data analysis

- More and more fields collect curve-like data (growth curves, mass spectrometry, ...)
- *Functional data* refers to observations that are curves sampled on a fine grid
- The usual statistical framework used to analyze such data is nonparametric regression: $(m = 1 \ldots, M)$

$$Y(t_m) = \mu(t_m) + E(t_m), \;\; E(t) \sim \mathcal{N}(0, \sigma^2).$$

- Goal: recover function $\mu(t)$ from noisy observations

Choosing wavelets when dealing with high dimensional data

- Traditional approaches when dealing with functional data is to use a functional basis (polynomial, splines, wavelets)
- Splines have been long studied in longitudinal data analysis for instance

Wavelets offer 3 main advantages:

$\rightarrow$ The fine modeling of curves with irregularities

$\rightarrow$ sparse representations

$\rightarrow$ Computationally efficiency (the DWT is in $\mathcal{O}(M)$)

Definition of wavelets and wavelet coefficients

- Wavelets provide an orthonormal basis of $L^2(\mathbb{R})$ with a scaling function ϕ and a mother wavelet ψ such that:

$$\left\{\phi_{j_0 k}(t), k = 0, \ldots, 2^{j_0} - 1; \psi_{jk}(t), j \geq j_0, k = 0, \ldots, 2^j - 1\right\}$$

- Any function $Y \in L^2(\mathbb{R})$ is then expressed in the form:

$$Y(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{j_0 k} \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j - 1} d^*_{jk} \psi_{jk}(t)$$

where $c^*_{j_0 k} = \langle Y, \phi_{j_0 k} \rangle$ and $d^*_{jk} = \langle Y, \phi_{jk} \rangle$ are the theorical scaling and wavelet coefficients.

DWT and empirical wavelet coefficients

- We observe function $Y(t)$ on discrete sample points (t_m),

$$\mathbf{Y}(\mathbf{t}) = [Y(t_1), \ldots, Y(t_M)]$$

- The Discrete Wavelet Transform is given by

$$\underset{[M\times M]}{\mathbf{W}} \underset{[M\times 1]}{\mathbf{Y}} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

- $\mathbf{W}$ is an orthogonal matrix of filter (wavelet specific),
- $(\mathbf{c}, \mathbf{d})$ are empirical wavelet coefficients such that:

$$\begin{aligned} \mathbf{c} &\simeq \sqrt{M} \times \mathbf{c}^* \\ \mathbf{d} &\simeq \sqrt{M} \times \mathbf{d}^* \end{aligned}$$

From non parametric to parametric linear models

- Once the data have been projected in the functional domain we retrieve a linear model such that:

$$\begin{aligned} \mathbf{WY(t)} &= \mathbf{W}\boldsymbol{\mu}\mathbf{(t)} + \mathbf{WE(t)} \\ \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon} \end{aligned}$$

- The next step is often to threshold wavelet coefficients for reconstruction purposes
- Many strategies have been proposed among which the standard hard thresholding rule [3] which sets to zero (d_{jk})s whose absolute value is lower than $\widehat{\sigma}\sqrt{2\log(M)}$

Functional ANOVA

- Experiments are now designed to collect **sets of curves** on different individuals
- We now observe many realizations of the same function which can be modeled by functional models: $i = 1, \ldots, N$, $m = 1 \ldots, M$

$$Y_i(t_m) = \mathbf{X}_i \boldsymbol{\mu}(t_m) + E_i(t_m), \;\; E_i(t) \sim \mathcal{N}(0, \sigma^2).$$

- $\boldsymbol{\mu}(\mathbf{t})$ becomes a fixed functional effect, and $\mathbf{X}$ is its design matrix
- Standard statistical questions can be assessed in the functional setting: test of a functional effect, comparison of treatments...

Functional Clustering Model (FCM)

- Among "classical" questions, clustering has focused much attention
- The idea is to cluster individuals based on functional observations
- We suppose that the cluster structure concerns the fixed effects of the model
- When using a mixture model we introduce the label variable $\zeta_{i\ell} \sim \mathcal{M}\left(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)\right)$ such that given $\{\zeta_{i\ell} = 1\}$

$$Y_i(t_m) \quad = \quad \mathbf{X}_i \boldsymbol{\mu}_\ell(t_m) + E_i(t_m)$$

- In the coefficient domain, a standard EM algorithm can be used to estimate the parameters (case $\mathbf{X} = \mathbf{I}$) [2]:

$$\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ \boldsymbol{\beta}_\ell \end{bmatrix} + \boldsymbol{\varepsilon}_i.$$

Application of Mass Spectrometry data

- Each spectra contains 15154 ionised peptides defined by a m/z ratio.
- Available at `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`
- Samples from 253 women: 91 Controls, 162 Cases (ovarian cancer) [6]

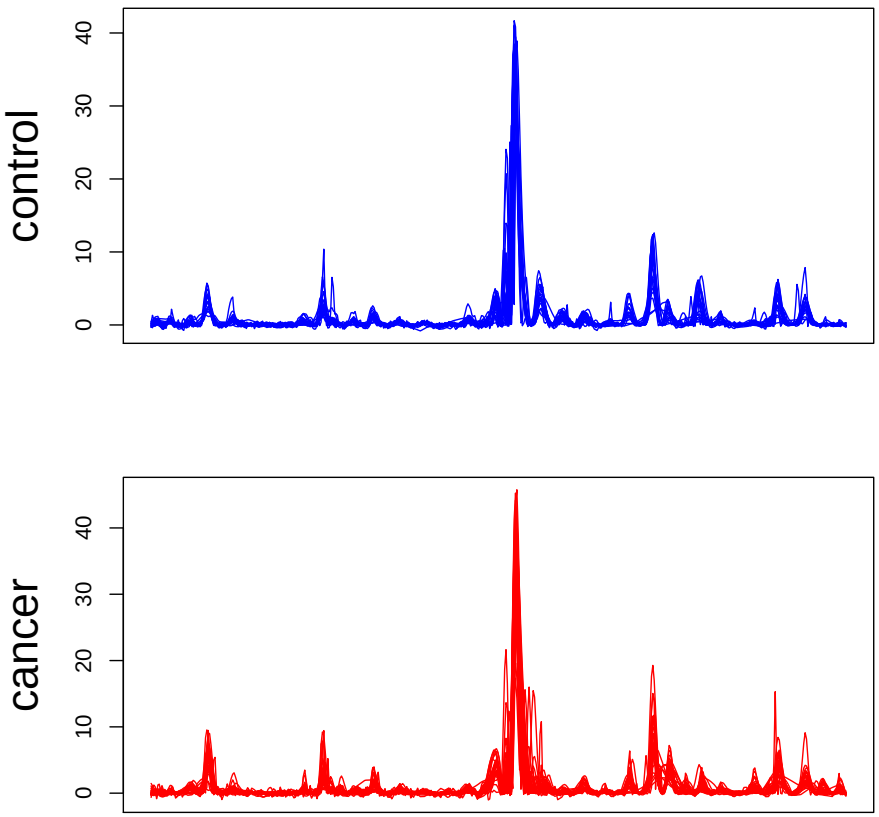


Figure: MALDI-TOF Spectra (window of 512).

Application to array CGH data

- Each profile is CGH profile from Breast Cancer patients
- Samples from 55 profiles with clinical informations [5]
- Subgroup discovery (1q16q)
- Super high inter-individual variability

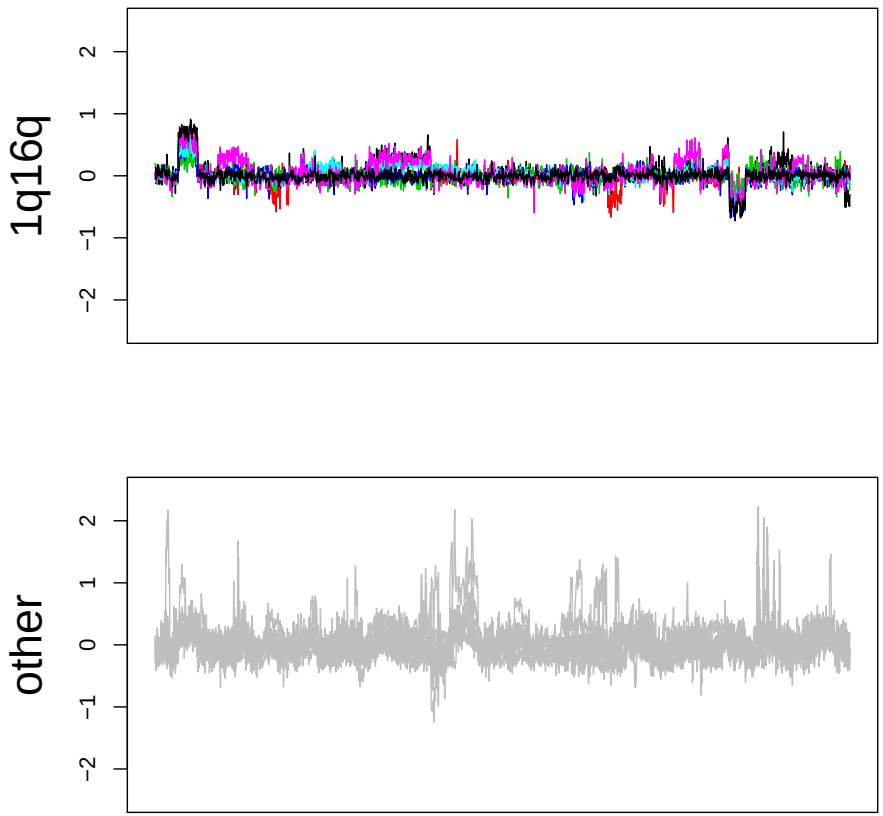


Figure: Array CGH profiles from [5]

Outline

Functional Clustering Mixed Models

- Mixed models are used to account for some structure in the variability of the observations
- Functional Mixed models are considered to introduce inter-individual functional variability such that given $\{\zeta_{i\ell} = 1\}$:

$$Y_i(t_m) = \mathbf{X}_i \boldsymbol{\mu}_\ell(t_m) + \mathbf{Z}_i \mathbf{U}_i(t_m) + E_i(t_m)$$

- $U_i(t) \sim \mathcal{N}(0, K_\ell(s,t))$ are random functions independent of $E(t)$
- In the wavelet domain, the model resumes to (case $\mathbf{X} = \mathbf{Z} = \mathbf{I}$):

$$\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ \boldsymbol{\beta}_\ell \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} + \boldsymbol{\varepsilon}_i,$$

$$\begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{G}_\nu & 0 \\ 0 & \mathbf{G}_\theta \end{bmatrix}\right).$$

Specification of the covariance of random effects

- We suppose that $\mathbf{G}$ is diagonal [4]
- Then the fixed and random effects should lie in the same Besov space. Introduce parameter η related to the regularity of process $\mathbf{U}_i(\mathbf{t})$

Theorem Abramovich & al. [1]

Suppose $\mu(t) \in B^s_{p,q}$ and $\mathbb{V}(\theta^i_{jk}) = 2^{-j\eta}\gamma^2_\theta$ then

$$U_i(t) \in B^s_{p,q}[0,1] \text{ a.s.} \quad \leftrightarrow \begin{cases} s + 1/2 - \eta/2 = 0, \text{ if } 1 \leq p < \infty \text{ and } q = \infty \\ s + 1/2 - \eta/2 < 0, \text{ otherwise.} \end{cases}$$

- The structure of the random effect can also vary wrt position and scale ($\gamma^2_{\theta,jk}$), and/or group membership ($\gamma^2_{\theta,jk\ell}$)

Dimensionality reduction step

- Inspired by a strategy proposed in Antoniadis & al. [2] in two steps
- Individual hard thresholding with the universal threshold $\widehat{\sigma}_\varepsilon \sqrt{2 \log M}$.
- Use the average of the MAD estimators computed on each indidivual
- This strategy seems reasonnable since:

$$\mathbb{V}(d^i_{Jk}) = \sigma^2_\varepsilon + 2^{-J\eta}\gamma^2_\theta \simeq \sigma^2_\varepsilon$$

- Take union of selected coefficients
- Removes positions that are non informative wrt to the clustering goal (i.e positions that are zero for all individuals)

Outline

Using the EM algorithm

- In the wavelet domain, the model is a Gaussian mixture model with a structured variance
- Both label variables $\boldsymbol{\zeta}$ and random effects $(\boldsymbol{\nu}, \boldsymbol{\theta})$ are unobserved
- The complete data log-likelihood can be written such that:

$$\begin{aligned}\log \mathcal{L}\left(\mathbf{c}, \mathbf{d}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{G}, \sigma_\varepsilon^2\right) &= \log \mathcal{L}\left(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2\right) \\ &+ \log \mathcal{L}\left(\boldsymbol{\nu}, \boldsymbol{\theta} | \boldsymbol{\zeta}; \mathbf{G}\right) \\ &+ \log \mathcal{L}\left(\boldsymbol{\zeta}; \boldsymbol{\pi}\right).\end{aligned}$$

- This likelihood can be easily computed thanks to the properties of mixed linear models such that:

$$\left[\begin{array}{c}\mathbf{c}_i \\ \mathbf{d}_i\end{array}\right] \Bigg| \left[\begin{array}{c}\boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i\end{array}\right], \{\zeta_{i\ell} = 1\} \sim \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{\alpha}_\ell + \boldsymbol{\nu}_i \\ \boldsymbol{\beta}_\ell + \boldsymbol{\theta}_i\end{array}\right], \sigma_\varepsilon^2 \mathbf{I}\right).$$

Predictions of hidden variables

- The EM algorithm provides the *posterior* probability of membership to cluster ℓ,

$$\tau_{i\ell}^{[h+1]} = \frac{\pi_\ell^{[h]} f\left(\mathbf{c}_i, \mathbf{d}_i; \boldsymbol{\alpha}_\ell^{[h]}, \boldsymbol{\beta}_\ell^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2[h]}\mathbf{I}\right)}{\sum_p \pi_p^{[h]} f\left(\mathbf{c}_i, \mathbf{d}_i; \boldsymbol{\alpha}_p^{[h]}, \boldsymbol{\beta}_p^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2[h]}\mathbf{I}\right)}.$$

- The E-step also provides the BLUP of random effects:

$$\begin{aligned}
\widehat{\boldsymbol{\nu}}_{i\ell}^{[h+1]} &= \left(\mathbf{c}_i - \boldsymbol{\alpha}_\ell^{[h]}\right) / \left(1 + \lambda_\nu^{[h]}\right), \lambda_\nu = \sigma_\varepsilon^2/\gamma_\nu^2, \\
\widehat{\boldsymbol{\theta}}_{i\ell}^{[h+1]} &= \left(\mathbf{d}_i - \boldsymbol{\beta}_\ell^{[h]}\right) / \left(1 + 2^{j\eta}\lambda_\theta^{[h]}\right), \lambda_\theta = \sigma_\varepsilon^2/\gamma_\theta^2.
\end{aligned}$$

ML estimates for fixed effects & variance terms

- the M-step provides the estimators of the mean curve coefficients and and of the variance of the random effects

$$
\begin{aligned}
\boldsymbol{\alpha}_\ell^{[h+1]} &= \sum_{i=1}^{n} \tau_{i\ell}^{[h]} \left(\mathbf{c}_i - \widehat{\boldsymbol{\nu}}_{i\ell}^{[h]} \right) / N_\ell^{[h]}, \\
\boldsymbol{\beta}_\ell^{[h+1]} &= \sum_{i=1}^{n} \tau_{i\ell}^{[h]} \left(\mathbf{d}_i - \widehat{\boldsymbol{\theta}}_{i\ell}^{[h]} \right) / N_\ell^{[h]}, \\
\gamma_\theta^{2[h+1]} &= \frac{1}{n(M-1)} \sum_{ijk\ell} 2^{j\eta} \tau_{i\ell}^{[h]} \left(\widehat{\theta}_{ijk\ell}^{2\ [h]} + \frac{\sigma_\varepsilon^{2[h]}}{1 + 2^{j\eta} \lambda_\theta^{[h]}} \right), \\
\gamma_\nu^{2[h+1]} &= \frac{1}{n} \sum_{i\ell} \left(\widehat{\nu}_{i00\ell}^{2\ [h]} + \frac{\sigma_\varepsilon^{2[h]}}{1 + \lambda_\nu^{[h]}} \right).
\end{aligned}
$$

- Parameter η can be estimated using a golden search section algorithm

Model selection using a BIC

- $\mathbf{m}_L$ stands for a clustering model with L clusters
- We select the dimension that maximizes

$$\begin{aligned} \text{BIC}(\mathbf{m}_L) &= \log \mathcal{L}\left(\mathbf{c}, \mathbf{d}; \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{G}}, \widehat{\sigma}^2_\varepsilon, \mathbf{m}_L\right) - \frac{|\mathbf{m}_L|}{2} \times \log(N). \\ |\mathbf{m}_L| &= |\boldsymbol{\alpha}| + |\boldsymbol{\beta}| + |\mathbf{G}| + |\boldsymbol{\pi}| - 1 + |\sigma^2_\varepsilon| \\ &= (M+1)L + |\mathbf{G}|. \end{aligned}$$

- The dimension of $\mathbf{G}$ depends on the variance structure of the random effects.
- $|\mathbf{G}| = 2$ is the case of constant variances $(\gamma^2_\nu, \gamma^2_\theta)$, and $|\mathbf{G}| = ML$ when variances depend on group, scale and position.

Model selection using a ICL

- It is likely that predictions of random effects provide information regarding L.
- The ICL criterion is based on the integrated likelihood of the complete data: $\log \mathcal{L}(\mathbf{c}, \mathbf{d}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta} | \mathbf{m}_L[\gamma_\ell^2])$
- Need to derive the integrated log-likelihood of the random effects and for the label variables.

$$
\begin{aligned}
-\frac{2}{N} \times \mathrm{ICL}(\mathbf{m}_L[\gamma_\ell^2]) &= M \log \mathrm{RSS}(\mathbf{c}, \mathbf{d} | \widehat{\boldsymbol{\nu}}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\tau}) \\
&+ \sum_\ell \widehat{\pi}_\ell \left(\log \mathrm{RSS}_\ell(\widehat{\boldsymbol{\nu}}, \boldsymbol{\tau}) + (M-1) \log \mathrm{RSS}_\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\tau}) \right) \\
&- \frac{2}{N} \sum_\ell \left\{ \log \Gamma\left(\frac{\widehat{N}_\ell}{2} \right) + \log \Gamma\left(\frac{\widehat{N}_\ell (M-1)}{2} \right) \right\} \\
&- 2 \sum_{\ell=1}^{L} \widehat{\pi}_\ell \log(\widehat{\pi}_\ell) + \frac{(M+1)L}{N} \times \log(N).
\end{aligned}
$$

Outline

Fine Definition of a simulation framework

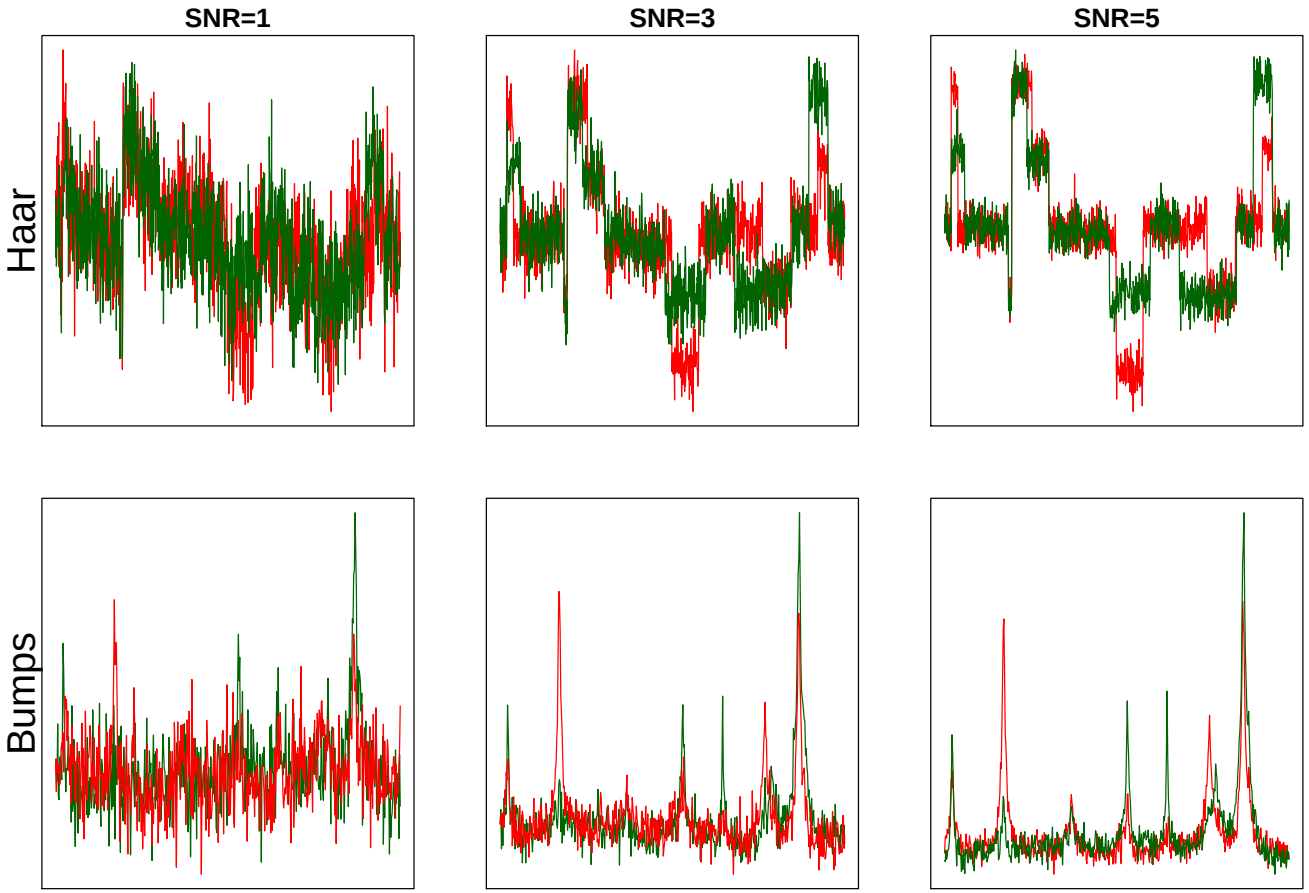
- We properly define the power of the signal:

$$\lim_{T\to\infty} \frac{1}{T} \int_{\frac{T}{2}}^{-\frac{T}{2}} \sum_{\ell} \pi_\ell \mathbb{E}\left[|\mu_\ell(t) + U_i(t)|\right]^2 \mathrm{d}t$$

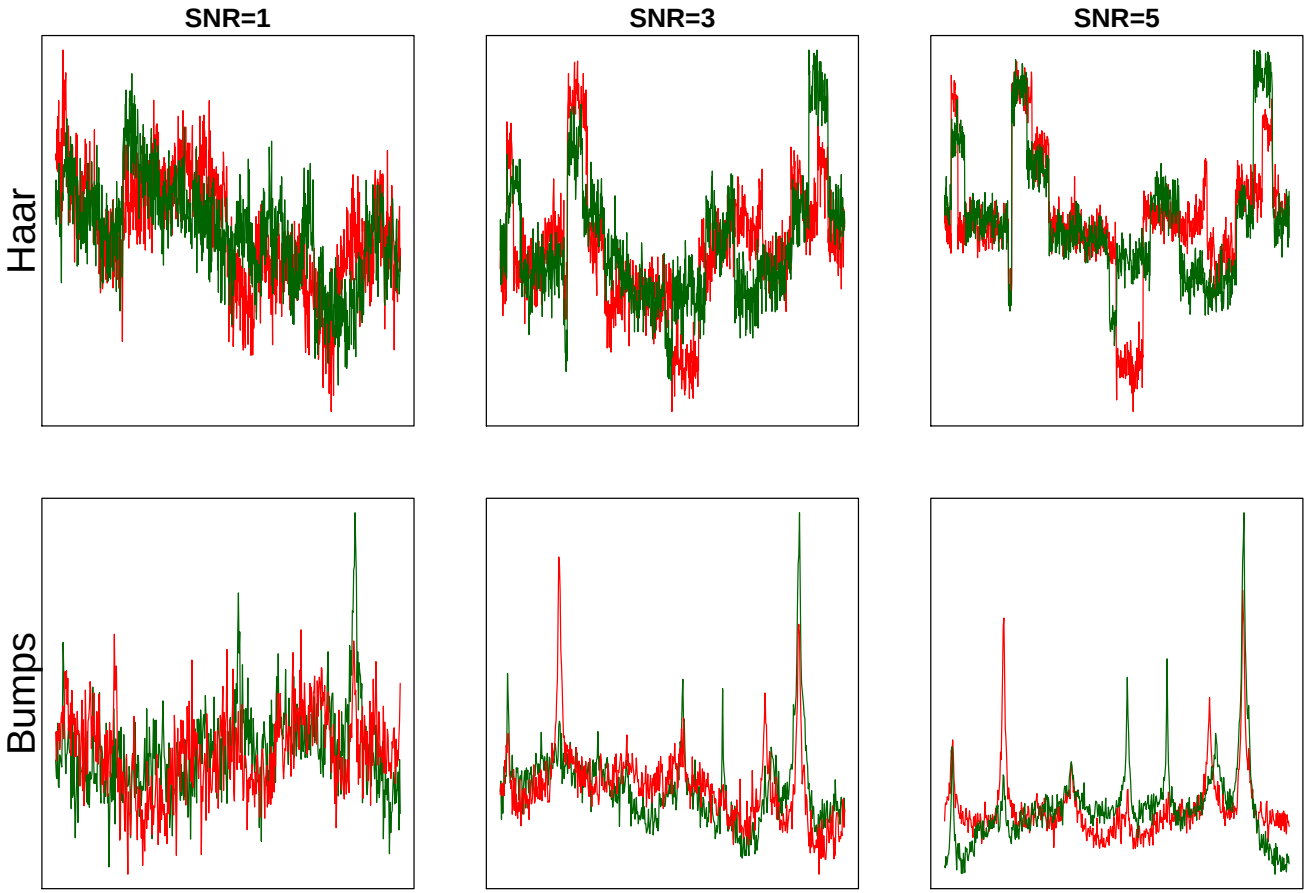
- We need to control two terms:

$$\begin{aligned} \mathrm{SNR}_\mu^2 &= \frac{1}{M\sigma_E^2} \sum_{\ell=1}^{L} \pi_\ell \left(\sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k \ell}^2 + \sum_{j\geq j_0} \sum_{k=0}^{2^j-1} \beta_{jk\ell}^2 \right), \\ \lambda_{\mathsf{U}} &= \sigma_E^2 / \left(\gamma_\nu^2 + \frac{\gamma_\theta^2}{1-2^{(1-\eta)}} \right), \end{aligned}$$

Simulated data with a low random effect $\lambda_U = 4$



Simulated data with a strong random effect $\lambda_U = 1/4$



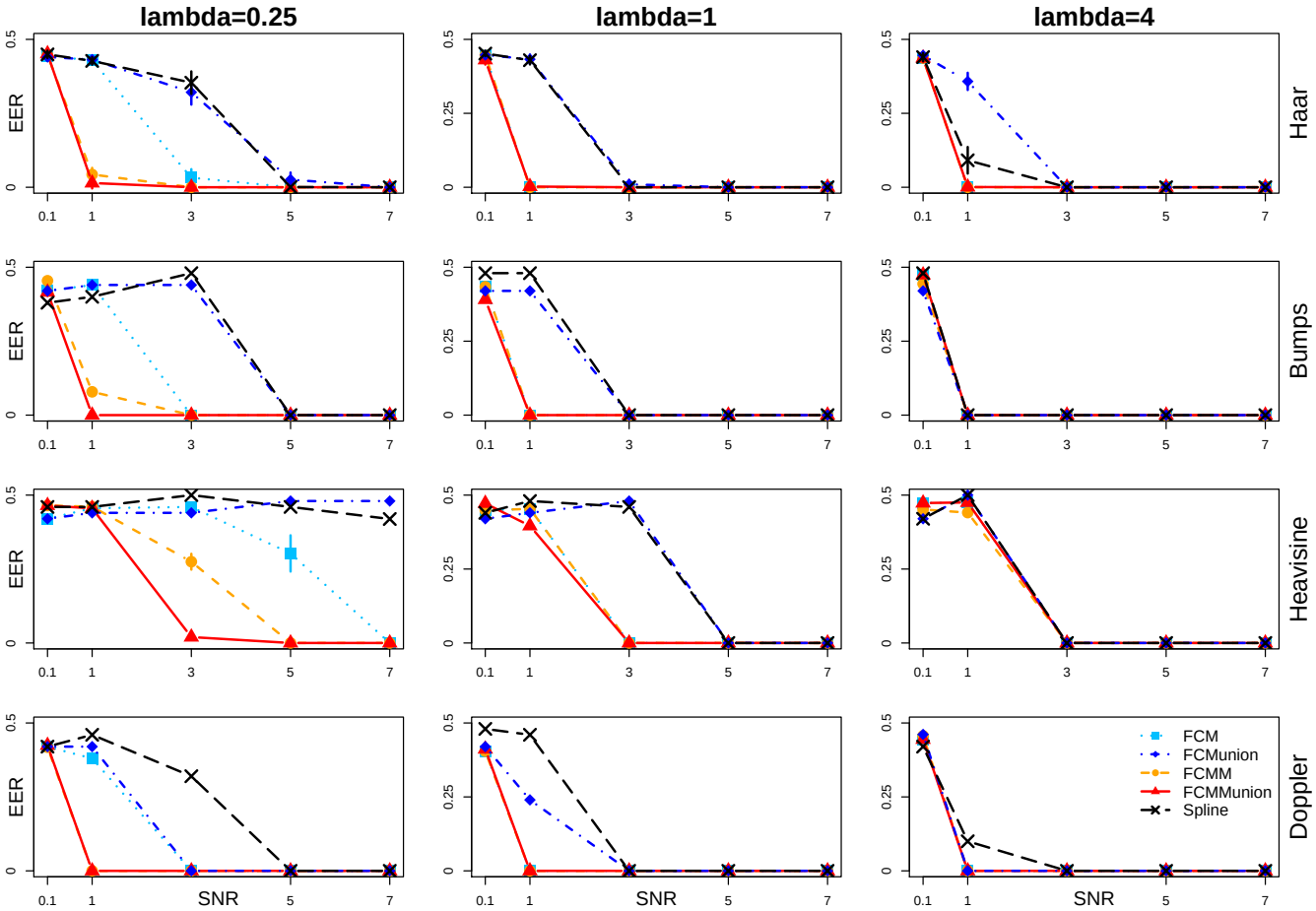
Aim & design of the simulation study

- What is the gain when using a functional random effect in terms of clustering (FCM/FCMM)?
- What is the performance of splines ?
- Is dimension reduction appropriate ?
- $n = 50$, $M = 512$, $L = 2$,
- $\text{SNR}_\mu \in \{0.1; 1; 3; 5; 7\}$, $\lambda_{\text{U}} \in \{0.25, 1, 4\}$
- Fixed effects can be Haar, Bumps, Heavisine, Doppler
- Study the Empirical Error Rate:

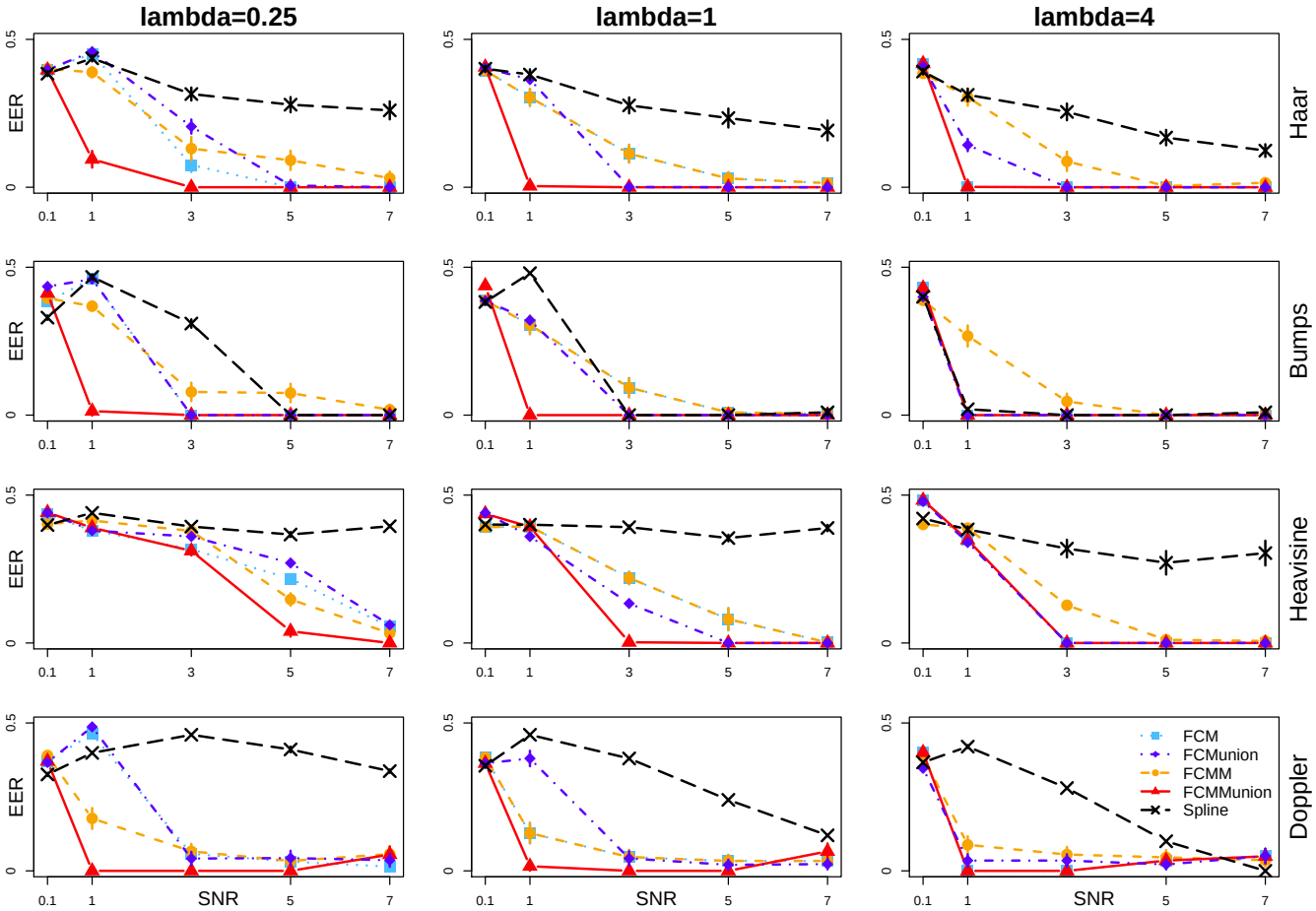
$$EER = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{\widehat{\zeta}_{i\ell} \neq \zeta_{i\ell}\}$$

- Development of a package `curvclust`

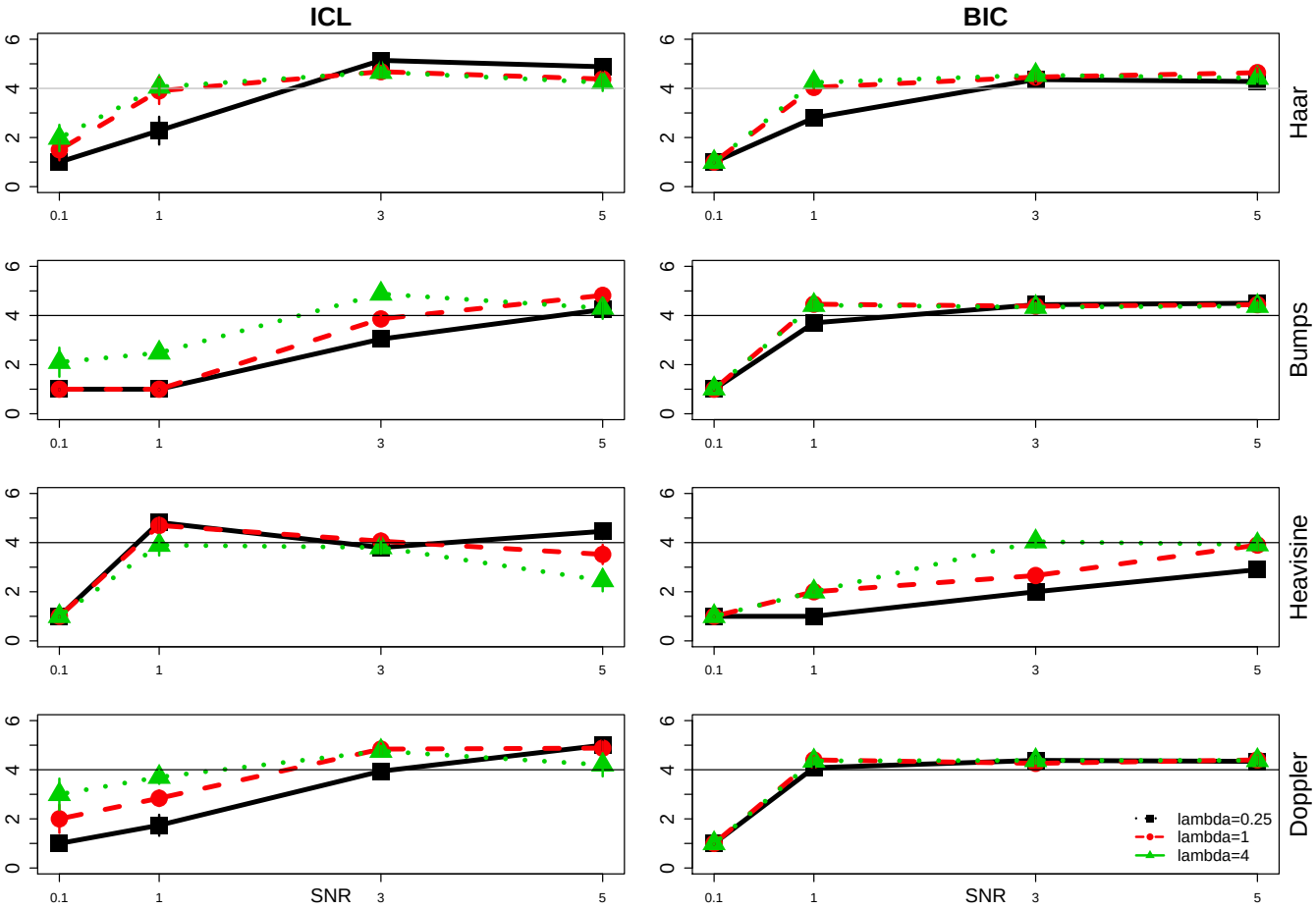
Empirical Error Rates (2 clusters)



Empirical Error Rates (4 clusters)



Model selection BIC vs ICL



Union-set Dimension Reduction performance

	FPR			FNR			% of selected coef		
SNR^2_μ / λ_U	0.25	1	4	0.25	1	4	0.25	1	4
0.1	68.7	81.4	90.3	2.8	1.4	1.1	7.5	4.2	2.5
1	68.4	78.1	82.9	3.8	2.6	2.2	8.4	5.8	4.6
Haar 3	67.8	75.5	77.2	7.7	6.8	6.7	11.7	9.7	9.4
5	69.1	75.0	75.8	8.6	7.9	7.8	12.3	10.7	10.5
7	70.0	75.2	75.7	8.8	8.2	8.0	12.3	10.9	10.7
0.1	91.3	94.1	96.7	2.3	2.3	2.3	7.0	4.9	3.1
1	88.8	91.8	92.6	2.3	2.3	2.3	8.9	6.7	6.1
Bumps 3	88.6	89.6	90.5	1.5	2.3	2.3	8.9	8.3	7.7
5	88.8	89.6	90.5	1.5	1.5	1.8	8.7	8.1	7.6
7	88.9	89.2	89.9	1.5	1.5	1.5	8.7	8.4	7.9

Table: FPR (non-thresholded among true null coefficients), FNR (thresholded among non null coefficients) and percentage of selected wavelet coefficients

Time of execution

		SNR_μ				
		0.1	1	3	5	7
FCM	Haar	2.3	2.4	2.3	2.4	2.3
	Bumps	2.6	2.5	2.6	2.5	2.5
FCMunion	Haar	0.4	0.4	0.5	0.5	0.5
	Bumps	0.5	0.5	0.5	0.5	0.5
FCMM	Haar	16.0	16.1	15.6	15.8	16.0
	Bumps	16.1	16.3	15.2	15.3	15.4
FCMMunion	Haar	6.9	7.1	7.6	7.6	7.6
	Bumps	6.7	6.7	6.8	6.7	6.7
Spline	Haar	25.5	26.2	23.0	23.6	22.3
	Bumps	23.3	26.6	22.0	21.2	21.7

Table: Average time of execution in minutes for different models on simulated data ($n = 50$ individuals, $M = 512$ positions).

Back to spectrometry data

- Samples from 91 controls 162 cases [6]
- We compete wavelet-based FCM on these data considering different random effect structures.
- Pre-treatment (baseline correction, peak alignment)
- Results on a window of 512

	$\mathbf{m}_2$	$\mathbf{m}_2[\gamma^2]$	$\mathbf{m}_2[\gamma^2_\ell]$	$\mathbf{m}_2[\gamma^2_{jk}]$	$\mathbf{m}_2[\gamma^2_{jk\ell}]$
global alignment	38	24	24	23	23
group alignment	20	21	22	0.4	36

Inaccuracy in spectra-alignment is lethal for clustering !

Conclusions & perspectives

- We developed a model for functional clustering with random effects
- All the codes are available with the R package `curvclust`
- Perspectives will mainly concern dimension reduction, supervised classification and model selection
- Perspectives in terms of application to piece-wise constant data like array CGH data.

References

F. Abramovich, T. Sapatinas, and B.W. Silverman.
Wavelet thresholding via a bayesian approach.
Journal of the Royal Statistical Society Series B Stat Methodol, 60:725–749, 1998.

A. Antoniadis, J. Bigot, and R. von Sachs.
A multiscale approach for statistical characterization of functional images.
Journal of Computational and Graphical Statistics, 18(1):216–237, 2008.

D.L. Donoho and I.M. Johnstone.
Ideal spatial adaptation by wavelet shrinkage.
Biometrika, 81(3):425–455, 1994.

M. Frazier, B. Jawerth, and G. Weiss.
Littlewood-Paley Theory and the Study of function Spaces.
Number 79. American Mathematical Society, 1991.

J. Fridlyand, A. M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. M. Ljung, A. N. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. W. Gray, F. Waldman, D. Pinkel, and D. G. Albertson.
Breast tumor copy number aberration phenotypes and genomic instability.
BMC Cancer, 6:96, 2006.

E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta.
Use of proteomic patterns in serum to identify ovarian cancer.
Lancet, 359:572–577, Feb 2002.