Curve Clustering and Functional Mixed Models.
Modeling, variable selection and application to Genomics

Franck Picard, LBBE - Lyon

Madison Giacofci LJK (Grenoble)
Sophie Lambert-Lacroix (TIMC - Grenoble)
Guillemette Marot (Univ. Lille 2)
Carlos Correa-Shokiche, (LBBE - Lyon)
Outline

1 Introduction
2 Functional Clustering Model with random effects
3 Estimation and model selection
4 Applications
5 Dimension Reduction for FANOVA
6 Conclusions & Perspectives
The Genomic Revolution

- Genomics is the field that investigates biological processes at the scale of Genomes.
- It started in the 70s-80s with the development of Molecular Biology techniques (sequencing, transcripts quantification).
- Genomics (and Post-Genomics) exploded in the 90s-2000s thanks to the miniaturization and industrialization of quantification processes.

Sequencing the Human Genome?

took ~ 10 years and can be done within a week now.
Towards Population-Based Genomic Studies

- Quantification mainly concern: Copy Number Variations, messenger RNAs, and proteins mostly using microarrays and Mass Spectrometry
- For long the task has been to extract signal from noise for one individual experiment (sometimes with replicates!)
- Prices decreasing, these technologies are now used at the population levels: this is the rise of Population Genomics

Statistical Tasks remain standard

Differential Analysis, Clustering, Discrimination
but the dimensionality of the data is overwhelming
Example with Mass Spectrometry data

- **Aim:** characterize the content of a mixture of peptides by mass-spect
- One peak corresponds to one peptide (signature)
- Each spectra contains 15154 ionised peptides defined by a $m/z$ ratio.
- 253 ovarian cancer samples: 91 Controls, 162 Cases [10]

Figure: MALDI-TOF Spectra.

http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp
Example with array CGH data

- Aim: characterize copy number variations between 2 genomes
- Segments with positive mean corresponds to regions that are amplified (negative/deleted)
- 55 aCGH profile from Breast Cancer patients

Figure: Breast Cancer CGH profiles [8] (log scale)
Towards Functional Models

- Proteomic Data: records are sampled on a very fine grid (m/z) and spectra have long been modeled using FDA.
- Genomic Data are mapped on a reference genome and show a spatial (1D?) structure.
- Functional models can account for this kind of structure, and working on curves should be more efficient than working on peaks or segments.
Towards Functional Mixed Models

- Subject specific fluctuations are known to be the largest source of variability in Mass-Spec data [6]
- Inter-Individual variability is the “curse” of biological data! (Technical / Biological Variabilities), and often under-estimated
- Mixed Linear Models: well known in Genetics to structure the variance according to experimental design and pedigrees
- We propose to analyze genomic data using functional mixed models
Outline

1 Introduction

2 Functional Clustering Model with random effects

3 Estimation and model selection

4 Applications

5 Dimension Reduction for FANOVA

6 Conclusions & Perspectives
Functional ANOVA Model

- We observe $N$ replicates of a noisy version of function $\mu$ over a fine grid $t = \{t_1, \ldots, t_M\}$, $t_j \in [0, 1]$, such that:

$$Y_i(t_m) = \mu(t_m) + E_i(t_m), \ E_i(t) \sim \mathcal{N}(0, \sigma^2),$$

- with $i = 1, \ldots, N$, $m = 1, \ldots, M = 2^J$

- In the following we use notations

$$Y_i(t) = [Y_i(t_1), \ldots, Y_i(t_M)], \ \mu(t) = [\mu(t_1), \ldots, \mu(t_M)]$$

- We propose to use wavelets to analyse such data:
  - Modelling curves with irregularities
  - Computationally efficiency (the DWT is in $O(M)$)
  - Dimension Reduction
Definition of wavelets and wavelet coefficients

- Wavelets provide an orthonormal basis of $L^2([0, 1])$ with a scaling function $\phi$ and a mother wavelet $\psi$ such that:

  \[
  \{ \phi_{j_0 k}(t), k = 0, \ldots, 2^{j_0} - 1; \psi_{j k}(t), j \geq j_0, k = 0, \ldots, 2^j - 1 \}
  \]

- Any function $Y \in L^2([0, 1])$ is then expressed in the form:

  \[
  Y_i(t) = \sum_{k=0}^{2^{j_0} - 1} c_{i,j_0 k}^* \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j - 1} d_{i,j k}^* \psi_{j k}(t)
  \]

where $c_{i,j_0 k}^* = \langle Y_i, \phi_{j_0 k} \rangle$ and $d_{i,j k}^* = \langle Y_i, \psi_{j k} \rangle$ are the theoretical scaling and wavelet coefficients.
The DWT and empirical wavelet coefficients

- Denote by $W$ an orthogonal matrix of filters (wavelet specific),
- The Discrete Wavelet Transform is given by

$$W \begin{bmatrix} Y_i(t) \end{bmatrix}_{[M \times 1]} = \begin{bmatrix} c_i \\ d_i \end{bmatrix}_{[M \times M]}$$

- $(c_i, d_i)$ are empirical scaling and wavelet coefficients
- Once the data are in the coefficient domain we retrieve a linear model such that: 
  \[ \begin{bmatrix} c_i \\ d_i \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0_M, \sigma^2_\epsilon I_M) \]

\[ W Y_i(t) = W \mu(t) + W E_i(t) \]

F. Picard (LBBE)
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 \( \zeta_{i\ell} \sim \mathcal{M}(1, \pi = (\pi_1, \ldots, \pi_L)) \) such that given \( \{\zeta_{i\ell} = 1\} \)

\[
Y_i(t_m) = \mu_{\ell}(t_m) + E_i(t_m)
\]

In the coefficient domain, we retrieve a Multivariate Gaussian Mixture such that given \( \{\zeta_{i\ell} = 1\} \) [3]:

\[
\begin{bmatrix}
c_i \\
d_i
\end{bmatrix} = \begin{bmatrix}
\alpha_{\ell} \\
\beta_{\ell}
\end{bmatrix} + \varepsilon_i.
\]
Functional Mixed models are considered to introduce inter-individual functional variability such that given \( \{ \zeta_{i\ell} = 1 \} \):

\[
Y_i(t_m) = \mu_\ell(t_m) + U_i(t_m) + E_i(t_m)
\]

\( U_i(t) | \{ \zeta_{i\ell} = 1 \} \sim \mathcal{N}(0, K_\ell(t, t')) \), \( U_i(t) \perp E_i(t) \)

In the wavelet domain, and given \( \{ \zeta_{i\ell} = 1 \} \) the model resumes to

\[
\begin{bmatrix}
  c_i \\
  d_i
\end{bmatrix} = \begin{bmatrix}
  \alpha_\ell \\
  \beta_\ell
\end{bmatrix} + \begin{bmatrix}
  \nu_i \\
  \theta_i
\end{bmatrix} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0_M, \sigma_\varepsilon^2 I_M)
\]

\[
\begin{bmatrix}
  \nu_i \\
  \theta_i
\end{bmatrix} \sim \mathcal{N} \left( 0_M, \begin{bmatrix}
  G_\nu & 0 \\
  0 & G_\theta
\end{bmatrix} \right)
\]

\[
\begin{bmatrix}
  \nu_i \\
  \theta_i
\end{bmatrix} \perp \varepsilon_i
\]
Suppose $G_\theta$ is diagonal by the whitening property of wavelets [7]

The fixed and random effects should lie in the same Besov space. Introduce parameter $\eta$ related to the regularity of process $U_i$

Theorem Abramovich & al. [1]

Suppose $\mu(t) \in B_{p,q}^{s}$ and $\nabla(\theta_{i,jk}) = 2^{-j\eta} \gamma_{\theta}^{2}$ then

$$U_i(t) \in B_{p,q}^{s}[0,1] \text{ a.s. } \iff \begin{cases} 
\eta = 2s + 1, & \text{if } 1 \leq p < \infty \text{ and } q = \infty \\
\eta > 2s + 1, & \text{otherwise.}
\end{cases}$$

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

1. Introduction
2. Functional Clustering Model with random effects
3. Estimation and model selection
4. Applications
5. Dimension Reduction for FANOVA
6. Conclusions & Perspectives
Using the EM algorithm

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

\[
\log L (c, d, \nu, \theta, \zeta; \pi, \alpha, \beta, G, \sigma^2_\varepsilon) = \log L (c, d|\nu, \theta, \zeta; \pi, \alpha, \beta, \sigma^2_\varepsilon) + \log L (\nu, \theta|\zeta; G) + \log L (\zeta; \pi).
\]

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

\[
\begin{bmatrix}
c_i \\
d_i
\end{bmatrix} \left| \begin{bmatrix}
\nu_i \\
\theta_i
\end{bmatrix}, \{\zeta_i \ell = 1\} \sim \mathcal{N} \left( \begin{bmatrix}
\alpha_\ell + \nu_i \\
\beta_\ell + \theta_i
\end{bmatrix}, \sigma^2_\varepsilon I \right) \right.
\]
The EM algorithm provides posterior probabilities of membership:

$$\tau_{i\ell}^{[h+1]} = \frac{\pi^{[h]}_\ell f \left( c_i, d_i; \alpha^{[h]}_\ell, \beta^{[h]}_\ell, G^{[h]} + \sigma^2_\varepsilon I \right)}{\sum_p \pi^{[h]}_p f \left( c_i, d_i; \alpha^{[h]}_p, \beta^{[h]}_p, G^{[h]} + \sigma^2_\varepsilon I \right)}.$$

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

$$\hat{\nu}_{i\ell}^{[h+1]} = \left( c_i - \alpha^{[h]}_\ell \right) / \left( 1 + \lambda^{[h]}_\nu \right), \lambda_\nu = \sigma^2_\varepsilon / \gamma^2_\nu,$$

$$\hat{\theta}_{i\ell}^{[h+1]} = \left( d_i - \beta^{[h]}_\ell \right) / \left( 1 + 2i\eta \lambda^{[h]}_\theta \right), \lambda_\theta = \sigma^2_\varepsilon / \gamma^2_\theta.$$
ML estimates for fixed effects & variances

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

\[
\alpha^{[h+1]}_\ell = \sum_{i=1}^{n} \tau_{i\ell} \left( c_i - \hat{\nu}^{[h]}_{i\ell} \right) / N^{[h]}_\ell,
\]

\[
\beta^{[h+1]}_\ell = \sum_{i=1}^{n} \tau_{i\ell} \left( d_i - \hat{\theta}^{[h]}_{i\ell} \right) / N^{[h]}_\ell,
\]

\[
\gamma^{2[h+1]}_{\theta} = \frac{1}{n(M-1)} \sum_{ijk\ell} 2\eta \tau_{i\ell}^{[h]} \left( \hat{\theta}_{ijk\ell}^{2[h]} + \frac{\sigma_{\varepsilon}^{2[h]}}{1 + 2\eta \lambda^{[h]}_{\theta}} \right),
\]

\[
\gamma^{2[h+1]}_{\nu} = \frac{1}{n} \sum_{i\ell} \left( \hat{\nu}_{i00\ell}^{2[h]} + \frac{\sigma_{\varepsilon}^{2[h]}}{1 + \lambda^{[h]}_{\nu}} \right).
\]

- Parameter \( \eta \) can be estimated by numerical optimization
Model selection using a BIC

- $m_L$ stands for a clustering model with $L$ clusters.
- We select the dimension that maximizes

$$\text{BIC}(m_L) = \log \mathcal{L} \left( c, d; \hat{\pi}, \hat{\alpha}, \hat{\beta}, \hat{G}, \hat{\sigma}_\varepsilon^2, m_L \right) - \frac{|m_L|}{2} \times \log(N).$$

$$|m_L| = |\alpha| + |\beta| + |G| + |\pi| - 1 + |\sigma_\varepsilon^2|$$

$$= (M + 1)L + |G|.$$

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

1. Introduction

2. Functional Clustering Model with random effects

3. Estimation and model selection

4. Applications

5. Dimension Reduction for FANOVA

6. Conclusions & Perspectives
Back to Mass Spectrometry data

- 91 controls 162 cases [10]
- Pre-treatment (baseline correction, peak alignment)
- Results (EER %) on a window of 512

<table>
<thead>
<tr>
<th>model</th>
<th>global align.</th>
<th>group align.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_2$</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>$m_2[\gamma_2]$</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>$m_2[\gamma_\ell^2]$</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>$m_2[\gamma_{jk}^2]$</td>
<td>23</td>
<td>0.4</td>
</tr>
<tr>
<td>$m_2[\gamma_{j\ell k}^2]$</td>
<td>23</td>
<td>36</td>
</tr>
</tbody>
</table>

Inaccuracy in spectra-alignment is lethal for clustering!
Applications

Application to array CGH data

- 3 main subtypes identified [8]
- Not reproduced by others [11] but one group is associated to the best patient outcome.
- We were able to identify the 1q/16p subtype on the complete dataset (with 1 mismatch).

<table>
<thead>
<tr>
<th>cluster ID</th>
<th>$\frac{SNR^2}{\mu}$</th>
<th>$\hat{\lambda}_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.1e-4</td>
<td>3.9e-04</td>
</tr>
<tr>
<td>2</td>
<td>2.3e-3</td>
<td>3.8e-05</td>
</tr>
<tr>
<td>3</td>
<td>1.3e-3</td>
<td>6.4e-04</td>
</tr>
<tr>
<td>4 (1q/16p)</td>
<td>1.5e-3</td>
<td>1.3e-04</td>
</tr>
<tr>
<td>5</td>
<td>9.3e-4</td>
<td>4.3e-05</td>
</tr>
</tbody>
</table>

**Figure:** Array CGH profiles from [8]
Outline

1. Introduction
2. Functional Clustering Model with random effects
3. Estimation and model selection
4. Applications
5. Dimension Reduction for FANOVA
6. Conclusions & Perspectives
The Case of One Curve

- When only one curve is observed, classical procedures consist in selecting a reduced number of coefficients while controlling for reconstruction properties.
- Among classical procedures, soft thresholding is well known\cite{5}:
  \[
  \hat{\beta}_{jk} = \text{sign} (d_{jk}) (|d_{jk}| - \lambda)_{+}
  \]

- \(\lambda\) is usually chosen as the “universal” threshold \(\sigma \sqrt{2 \log M}\), \(\sigma\) is estimated by the MAD estimator.
- Soft thresholding has good reconstruction properties and attains a near-minimax rate of convergence.
The LASSO and penalized regression

- Considering a regression model \( Y = X\beta + E \), the LASSO performs shrinkage and variable selection by solving a penalized estimation problem.

- Denoting by \( J(\beta) = \frac{1}{2} ||Y - X\beta||^2 \), the LASSO consists in solving

\[
\hat{\beta} = \arg\min_{\beta} \left\{ J(\beta) + \lambda \|\beta\|_1 \right\}
\]

- It is well known that in the case of orthogonal design, the LASSO resumes to soft thresholding.

**Aim**

How to use penalization techniques to propose an estimation framework for FANOVA and perform dimension reduction simultaneously?
1 fixed effect, $N$ replicates

- The first simple model is given by: $Y_i(t_m) = \mu(t_m) + E_i(t_m)$
- First attempts [2] propose to average and shrink coefficients
- The LASSO gives the appropriate framework by solving

$$J(\beta) + \lambda \text{pen}(\beta) = \frac{1}{2} \sum_{i=1}^{n} \|d_i - \beta\|^2 + \lambda \|\beta\|_1$$

$$\hat{\beta}_{jk} (\lambda) = \text{sign} (d_{\bullet,jk}) (|d_{\bullet,jk}| - \lambda)_+$$

- $\lambda$ can be estimated using a BIC:

$$\text{BIC}(\lambda) = \log \mathcal{L}(d, \hat{\beta}(\lambda), \sigma^2) - \frac{1}{2} \log(N) \times \|\hat{\beta}(\lambda)\|_0$$
$L$ fixed effects, $N$ replicates

- The Functional Clustering Model is (given $\{\zeta_{i\ell} = 1\}$),
  \[ Y_i(t_m) = \mu_\ell(t_m) + E_i(t_m) \]

- The LASSO can be used in the context of mixtures as well
  \[
  J_L(\zeta; \beta, \pi) + \lambda \text{pen}(\beta) = \frac{1}{2} \sum_{i=1}^{n} \zeta_{i\ell} \| d_i - \beta_{\ell} \|^2 + \lambda \sum_{\ell=1}^{L} \pi_\ell \| \beta_\ell \|_1
  \]

- MLE is performed by using a penalized EM [9] algorithm with
  \[
  J_L(\beta; \pi) = -\sum_{i=1}^{n} \log \left\{ \sum_{\ell=1}^{L} \pi_\ell f(d_i; \beta_{\ell}, \sigma^2) \right\}
  \]
1 fixed effect, $N$ replicates, $N$ functional Random Effects

- The Functional Mixed Model is $Y_i(t_m) = \mu(t_m) + U_i(t_m) + E_i(t_m)$

$$
\begin{bmatrix}
  c_i \\
  d_i
\end{bmatrix} = \begin{bmatrix}
  \alpha \\
  \beta
\end{bmatrix} + \begin{bmatrix}
  \nu_i \\
  \theta_i
\end{bmatrix} + \varepsilon_i, \quad \begin{bmatrix}
  \nu_i \\
  \theta_i
\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}
  \mathbf{G}_\nu & 0 \\
  0 & \mathbf{G}_\theta
\end{bmatrix}\right).
$$

- Dimension reduction is performed
  - On Fixed effects $\beta$
  - On random effects through a spare representation of Kernel $K(t, t') = \text{cov}(U_i(t), U_i(t'))$

- $\mathbf{G}_\theta$ has general (diagonal) term $\nabla(\theta_{i,jk}) = 2^{-j\eta}\gamma_{\theta,jk}^2$

- The LASSO can be used to shrink terms $\gamma_{\theta,jk}^2$
The LASSO for Mixed Linear Models

- Perform MLE estimation using a hidden variable representation of Mixed Linear Models
- Use the EM algorithm to optimize [4]

\[ J(\beta, G, \sigma^2) + \lambda \beta \text{pen}(\beta) + \lambda \gamma \text{pen}(\gamma) = -\log \mathcal{L}(d; \beta, G, \sigma^2) \]
\[ + \lambda \| \beta \|_1 \]
\[ + \lambda \| \gamma \|_1 \]

- The Maximization is performed indirectly by using the conditional expectation of the complete-data log-likelihood

\[ \mathbb{E} \{ \log \mathcal{L}(d, \theta; \beta, G, \sigma^2) | d \} \]
Reparametrization and M-step

- For convexification, use the following reparametrization

\[ \mathbf{d}_i = \beta + \mathbf{G}_\theta^{-1/2} \theta_i^* + \varepsilon_i, \quad \theta_i^* \sim \mathcal{N}(\mathbf{0}_M, \mathbf{I}_M) \]

\[
-2\mathbb{E} \left\{ \log \mathcal{L}(\mathbf{d}, \theta; \beta, \mathbf{G}, \sigma^2)|\mathbf{d} \right\} = Mn \log \sigma_{\varepsilon}^2
\]
\[
+ \frac{1}{\sigma_{\varepsilon}^2} \| \mathbf{d} - \beta - \mathbf{G}_\theta^{-1/2} \hat{\theta}_i^* \|^2
\]
\[
+ \text{tr} \left( \mathbf{G}_\theta^{-1/2}^T \mathbb{V} \{ \theta^*|\mathbf{d} \} \mathbf{G}_\theta^{-1/2} \right)
\]
\[
+ \hat{\theta}_i^*^T \hat{\theta}_i^* + \text{cst}
\]

- For dimensionality purposes, use Conditional M-steps
The penalized estimator of fixed effects $\beta$ is based on

$$\tilde{d}_{i,jk} = d_{i,jk} - 2^{-j\eta/2}\hat{\gamma}_{jk}\hat{\theta}^*_{i,jk}$$

$$\hat{\beta}_{jk}(\lambda_\beta) = \text{sign}\left(\tilde{d}_{\bullet,jk}\right)\left(\tilde{d}_{\bullet,jk} - \frac{\lambda_\beta\sigma^2}{N}\right) +$$

The penalized estimator of $\gamma$ is based on

$$\rho_{ijk}(\lambda_\beta) = 2^{-j\eta}\hat{\theta}^*_{i,jk} \times \left(d_{i,jk} - \hat{\beta}_{jk}(\lambda_\beta)\right)$$

$$\hat{\gamma}_{jk}(\lambda_\beta, \lambda_\gamma) \propto \left(\left|\rho_{\bullet,jk}(\lambda_\beta)\right| - \frac{\lambda_\gamma\sigma^2}{N}\right) +$$
Outline

1. Introduction
2. Functional Clustering Model with random effects
3. Estimation and model selection
4. Applications
5. Dimension Reduction for FANOVA
6. Conclusions & Perspectives
We developed a model for functional clustering with random effects.

All the codes are available with the R package curvclust
http://cran.r-project.org/

Main challenge now concerns thresholding of wavelet coefficients in multiple contexts using the LASSO machinery.

Mixture Models, Mixed Models, Mixture + Mixed Models!

What are the reconstruction properties of the predicted random effects $\hat{U}_i(t) = \mathbb{E}(U_i(t)|Y_i(t))$ (functional). Do we have a Best-Predictor Property?
Wavelet thresholding via a bayesian approach.

U. Amato and T. Sapatinas.
Wavelet shrinkage approaches to baseline signal estimation from repeated noisy measurements.

A. Antoniadis, J. Bigot, and R. von Sachs.
A multiscale approach for statistical characterization of functional images.

Joint variable selection for fixed and random effects in linear mixed-effects models.

D.L. Donoho and I.M. Johnstone.
Ideal spatial adaptation by wavelet shrinkage.

An insight into high-resolution mass-spectrometry data.

Littlewood-Paley Theory and the Study of function Spaces.

J Fridlyand and al.
Breast tumor copy number aberration phenotypes and genomic instability.
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: $\log \mathcal{L}(c, d, \nu, \theta, \zeta | m_L[\gamma^2])$
- Need to derive the integrated log-likelihood of the random effects and for the label variables.

$$- \frac{2}{N} \times \text{ICL}(m_L[\gamma^2]) = M \log \text{RSS}(c, d | \hat{\nu}, \hat{\theta}, \tau)$$

$$+ \sum_\ell \hat{\pi}_\ell \left( \log \text{RSS}_\ell(\hat{\nu}, \tau) + (M - 1) \log \text{RSS}_\ell(\hat{\theta}, \tau) \right)$$

$$- \frac{2}{N} \sum_\ell \left\{ \log \Gamma \left( \frac{\hat{N}_\ell}{2} \right) + \log \Gamma \left( \frac{\hat{N}_\ell(M - 1)}{2} \right) \right\}$$

$$- 2 \sum_{\ell=1}^L \hat{\pi}_\ell \log(\hat{\pi}_\ell) + \frac{(M + 1)L}{N} \times \log(N).$$
Model selection BIC vs ICL
We properly define the power of the signal:

\[
\lim_{T \to \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \sum_{\ell} \pi_\ell \mathbb{E}[|\mu_\ell(t) + U_i(t)|^2] \, dt
\]

We need to control two terms:

\[
\begin{align*}
\text{SNR}^2_{\mu} &= \frac{1}{M \sigma^2_E} \sum_{\ell=1}^{L} \pi_\ell \left( \sum_{k=0}^{2j_0-1} \alpha^2_{j_0 k\ell} + \sum_{j \geq j_0} \sum_{k=0}^{2j-1} \beta^2_{jk\ell} \right), \\
\lambda_U &= \frac{\sigma^2_E}{\gamma^2 + \frac{\gamma^2_\theta}{1 - 2(1-\eta)}},
\end{align*}
\]
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?
- $n = 50$, $M = 512$, $L = 2$,
- $\text{SNR}_\mu \in \{0.1; 1; 3; 5; 7\}$, $\lambda_U \in \{0.25, 1, 4\}$
- Fixed effects can be Haar, Bumps, Heavisine, Doppler
- Study the Empirical Error Rate:

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

- Development of a package curvclust
Empirical Error Rates (2 clusters)
Empirical Error Rates (4 clusters)
## Union-set Dimension Reduction performance

<table>
<thead>
<tr>
<th>SNR$_\mu^2$ / $\lambda_U$</th>
<th>FPR</th>
<th></th>
<th>FNR</th>
<th></th>
<th>% of selected coef</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>1</td>
<td>4</td>
<td>0.25</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>68.7</td>
<td>81.4</td>
<td>90.3</td>
<td>2.8</td>
<td>1.4</td>
<td>1.1</td>
</tr>
<tr>
<td>1</td>
<td>68.4</td>
<td>78.1</td>
<td>82.9</td>
<td>3.8</td>
<td>2.6</td>
<td>2.2</td>
</tr>
<tr>
<td>Haar 3</td>
<td>67.8</td>
<td>75.5</td>
<td>77.2</td>
<td>7.7</td>
<td>6.8</td>
<td>6.7</td>
</tr>
<tr>
<td>5</td>
<td>69.1</td>
<td>75.0</td>
<td>75.8</td>
<td>8.6</td>
<td>7.9</td>
<td>7.8</td>
</tr>
<tr>
<td>7</td>
<td>70.0</td>
<td>75.2</td>
<td>75.7</td>
<td>8.8</td>
<td>8.2</td>
<td>8.0</td>
</tr>
<tr>
<td>0.1</td>
<td>91.3</td>
<td>94.1</td>
<td>96.7</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>1</td>
<td>88.8</td>
<td>91.8</td>
<td>92.6</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Bumps 3</td>
<td>88.6</td>
<td>89.6</td>
<td>90.5</td>
<td>1.5</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>5</td>
<td>88.8</td>
<td>89.6</td>
<td>90.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.8</td>
</tr>
<tr>
<td>7</td>
<td>88.9</td>
<td>89.2</td>
<td>89.9</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>SNR&lt;sub&gt;μ&lt;/sub&gt;</th>
<th>0.1</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>FCM</td>
<td>Haar</td>
<td>2.3</td>
<td>2.4</td>
<td>2.3</td>
<td>2.4</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Bumps</td>
<td>2.6</td>
<td>2.5</td>
<td>2.6</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>FCMunion</td>
<td>Haar</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Bumps</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>FCMM</td>
<td>Haar</td>
<td>16.0</td>
<td>16.1</td>
<td>15.6</td>
<td>15.8</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>Bumps</td>
<td>16.1</td>
<td>16.3</td>
<td>15.2</td>
<td>15.3</td>
<td>15.4</td>
</tr>
<tr>
<td>FCMunion</td>
<td>Haar</td>
<td>6.9</td>
<td>7.1</td>
<td>7.6</td>
<td>7.6</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>Bumps</td>
<td>6.7</td>
<td>6.7</td>
<td>6.8</td>
<td>6.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Spline</td>
<td>Haar</td>
<td>25.5</td>
<td>26.2</td>
<td>23.0</td>
<td>23.6</td>
<td>22.3</td>
</tr>
<tr>
<td></td>
<td>Bumps</td>
<td>23.3</td>
<td>26.6</td>
<td>22.0</td>
<td>21.2</td>
<td>21.7</td>
</tr>
</tbody>
</table>

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