Functional Clustering by Bayesian Wavelet methods

Shubhankar Ray and Bani Mallick

Presenter: Iulian Pruteanu-Malinici
Outline

- Motivation and introduction.
- The wavelet model.
- Shrinkage estimation.
- Regression model.
- Hierarchical wavelet model.
- Wavelet clustering model.
- Posterior inference.
- Examples.
- Discussion.
Motivation and introduction

- Functional data arise in a wide variety of applications and are often clustered to reveal differences in the sources or to provide a concise picture of the data (see clustered gene expression profiles).
- Model-based clustering relies largely on finite mixture models, assuming that the number of clusters is known in advance.
- DP-based infinite mixture models overcome this deficiency; however, they disregard the functional form of the data.
- Solution - Bayesian wavelet modeling: manages to accommodate almost all the functional forms that are observed in real life applications (see Donoho, Johnstone 1994 and 1995 or more recently Abramovich et al., 1998).
- Wavelet representations are also sparse and can be helpful in limiting the number of regressors.
- The wavelet clustering model presented here is based on the mixture of DPs (Ferguson, 1973; Antoniak, 1974).
The wavelet model

Wavelet decomposition:

A wavelet decomposition of a function $f(t)$ is an expression of the following form:

$$f(t) = \sum_{j=-\infty}^{J} \sum_{k=-\infty}^{\infty} \beta_{j,k} \psi_{j,k}(t),$$

where the wavelets $\psi_{j,k}(t)$ are generated from a single wavelet $\psi(t)$, the so-called mother wavelet, by scaling and translation:

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}} \psi(2^{-j} t - k),$$

and the wavelet coefficients $\beta_{j,k} = \langle f, \psi_{j,k} \rangle$.

One can construct wavelets $\psi(t)$ such that the dilated and translated family $\psi_{j,k}(t)$ is an orthonormal basis on $L^2(\mathbb{R})$. Of course, not every choice of $\psi(t)$ allows one to decompose general functions $f(t)$ as above; see Meyer and Daubechies orthogonal wavelets.
Wavelet system expansions:

- To decompose functions $f(t)$ defined over an interval $[0,1]$, it is necessary to modify the original wavelets $\psi_{j,k}(t)$ and introduce *scaling functions* also restricted to $[0,1]$.
- The wavelet system expansion is a tool used to decompose a function into a set of weighted basis functions and wavelets.
- The expansion of a function $f(t)$ has the following form:

$$f(t) = \sum_{k=-\infty}^{\infty} \gamma_{J,k} \phi_{J,k}(t) + \sum_{j=-\infty}^{J} \sum_{k=-\infty}^{\infty} \beta_{j,k} \psi_{j,k}(t),$$

where $\phi_{j,k}(t)$ are scaling functions, $\phi_{j,k}(t) = 2^{-j} \phi(2^{-j} t - k)$ and $\gamma_{J,k}$ are scaling coefficients, $\gamma_{J,k} = \langle f, \phi_{J,k} \rangle$.
- This expansion can be thought of as a coarse approximation of the signal $f(t)$ (low-frequency energy), plus the finer details of the signal (high-frequency energy).
Shrinkage estimation

The wavelet sparse representation implies that, when the wavelet basis is orthogonal, the *i.i.d* normal noise affects all the wavelet coefficients equally, whereas the signal information remains isolated in a few coefficients.

In shrinkage estimation, these coefficients, which are mostly noise, are discarded to retrieve an effective reconstruction of the function. Therefore, the wavelet/scaling decomposition of $f(t)$ has the form:

$$ f(t) \approx \gamma_{0,0} \phi_{0,0}(t) + \sum_{j=1}^{J} \sum_{k=0}^{2^{j-1}} \beta_{j,k} \psi_{j,k}(t), $$

where $J = \log_2(m)$ is the finest level of wavelet decomposition.

Bayesian wavelet shrinkage proceeds by imposing mixture priors over the wavelet coefficients $\beta_{j,k}$:

$$ \beta_{j,k} \sim \pi_j N(0, g_j \sigma^2) + (1 - \pi_j) \delta_0 $$

$$ g_j \sim IG(r_j, s_j) $$
Regression model

Consider a collection of unknown functions \( \{f_i\}_{i=1}^{n} \), on the unit interval, that are observed with white Gaussian noise as:

\[
y_{i,k} = f_i(k/m) + \varepsilon_{i,k}
\]

\[
\varepsilon_{i,k} \sim N(0, \sigma_i^2),
\]

where \( k \in 1, \ldots, m \) and \( m \) is a power of 2.

In a linear model notation, if \( Y_i = (y_{i,1}, \ldots, y_{i,m}) \) is the vector of \( m \) observations, the regression model is:

\[
Y_i = X_i \beta_i + \varepsilon_i \quad i = 1, \ldots, n,
\]

\[
\varepsilon_i \sim N(0, \sigma_i^2 I_m),
\]

\[
\beta_{j,k} \sim \pi_j N(0, g_j \sigma^2) + (1 - \pi_j) \delta_0.
\]
Hierarchical wavelet model

\[
\begin{align*}
Y_i | \beta_i, \sigma_i^2 & \sim N(X\beta_i, \sigma_i^2 I_m), \\
\beta_i, \sigma_i^2 | \gamma, g & \sim NIG(0, V; u, v), \\
V & = diag(\gamma)diag(g), \\
g_j & \sim IG(r_j, s_j), \\
\gamma_{j,k} & \sim Bernoulli(\pi_j),
\end{align*}
\]

for \(i \in 1, \ldots, n\), \(j \in 1, \ldots, J\) and \(k \in 0, \ldots, 2^j - 1\).

- NIG denotes the normal-inverse gamma prior, the product of the conditionals \(\beta_i|\sigma_i^2, V \sim N(0, \sigma_i^2 V)\) and \(\sigma_i^2 \sim IG(u, v)\).

- \(\gamma = (\gamma_0, 0, \gamma_1, 0, \gamma_2, 0, \gamma_2, 1)\) is a vector of latent indicator variables for selection of each coefficient \(\beta_{j,k}\) and \(g = (g_0, g_1, g_2, g_2, \ldots)\) comprise the corresponding scaling parameters.

- This hierarchical model is especially useful for modelling sparse wavelet representations.
Wavelet clustering model

In the clustering model, the parameters $\theta_i = (\beta_i, \sigma_i^2)$ for the underlying functions $f_i$ are drawn from a DP prior.

We assume we have $n$ random variables $\{\theta_i\}_{i=1}^n$ distributed according to $G$, and $G$ itself is a random measure drawn from a Dirichlet process,

$$\theta_n | G \sim G, \quad n = 1, \ldots, N,$$
$$G \sim DP(\alpha, G_0).$$

$$Y_i | \beta_i, \sigma_i^2 \sim N(X\beta_i, \sigma_i^2 I_m),$$
$$\theta_1, \theta_2, \ldots, \theta_n \sim G,$$
$$G \sim DP(\alpha, NIG(0, V; u, v)),$$
$$V = \text{diag}(\gamma) \text{diag}(g),$$
$$g_j \sim IG(r_j, s_j),$$
$$\gamma_{j,k} \sim \text{Bernoulli}(\pi_j),$$
$$\alpha \sim G(d_0, \eta_0).$$
Posterior inference

Updating equations for clustering parameters $\theta_n = (\beta, \sigma^2)_n$:

\[
(\beta, \sigma^2)_n | (\beta, \sigma^2)_{-n}, \alpha, g, \gamma, Y_n \propto q_n, 0 H_{g, \gamma, n}(\beta, \sigma^2) + \sum_{i=1}^{d_n-1} q_n, i \delta(\beta, \sigma^2)_i,
\]

\[
H_{g, \gamma, n}^* = NIG(\mu^*, V^*, u^*_n, v^*_n),
\]

\[
V^* = (V^{-1} + I_m)^{-1},
\]

\[
\mu^* = V^* X^T Y_n,
\]

\[
u_n^* = u + Y_n^T Y_n - \mu^{*T}(V^*)^{-1} \mu^*,
\]

\[
u_n^* = v + m.
\]
The posterior distribution of $\sigma^2$ is simply:

$$\sigma^2|Y, g, \gamma \sim IG(u^*, v^*),$$

where

$$u^* = u + \sum_{i=1}^{d_n} \left\{ \sum Y_j^T Y_j - \mu_i^* T (V_i^*)^{-1} \mu_i^* \right\},$$

$$v^* = v + mn,$$

$$\mu_i^* = V_i^* \sum X^T Y_j,$$

$$V_i^* = (V^{-1} + n_i I_m)^{-1}.$$
Updating equations for mixing parameters $g_j, \gamma_{j,k}$:

$$g_j | \gamma_j, C_n, \sigma_i^2, \beta_{i,j} \sim IG(r_j^*, s_j^*) ,$$

where

$$s_j^* = d_n \sum_{k=0}^{2^j-1} \gamma_{j,k} + s_j,$$

$$r_j^* = \sum_{i=1}^{d_n} \sigma_i^{-2} \sum_{k=0}^{2^j-1} \gamma_{j,k}^2 \beta_{i,j,k}^2 + r_j .$$

The indicators $\gamma_j$ are updated as follows:

$$f(\gamma_j | \gamma_{-j}, g, C_n, Y) \propto \prod_{i=1}^{d_n} f(Y_{j \in C_n(i)} | \gamma, g, C_n) \pi_j .$$
Examples

Shifted Doppler signals:

\[ f_{t_0}(t) = -0.025 + 0.6 \sqrt{t(1-t)} \sin \frac{2.10 \pi}{t - t_0}, \]

with the phase \( t_0 \) varied in eight intervals in \([0,1]\), to generate 200 signals.

- In one of these intervals, some functions were perturbed, to assess the model sensitivity to small local fluctuations. This created a total of 9 distinct classes. The functions have been equisampled over \(2^7 = 128\) points.

- For simulation, noisy data were generated by adding independent normal noise \( \varepsilon_{i,k} \sim N(0, \sigma^2) \).

- With a fixed probability \( p_m \), points were randomly selected and dropped from each function to evaluate the robustness to missing observations.
Examples (contd.)

Shifted Doppler signals (contd.):

- The estimated number of clusters \( \hat{d}_n \) and the percentage of misclassifications; \( \sigma = \text{SNR} \).

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>( \hat{d}_n ) for the following values of ( p_m ):</th>
<th>Misclassifications (%) for the following values of ( p_m ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.1</td>
<td>8.4</td>
<td>7.9</td>
</tr>
<tr>
<td>0.06</td>
<td>9.1</td>
<td>8.9</td>
</tr>
<tr>
<td>0.02</td>
<td>9.1</td>
<td>9.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>( \hat{d}_n ) for the following values of ( p_m ):</th>
<th>Misclassifications (%) for the following values of ( p_m ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.1</td>
<td>7.3</td>
<td>6.8</td>
</tr>
<tr>
<td>0.06</td>
<td>9.5</td>
<td>8.9</td>
</tr>
<tr>
<td>0.02</td>
<td>9.1</td>
<td>8.6</td>
</tr>
</tbody>
</table>

Figure 1: Performance evaluation: (a) Bayesian wavelet model, (b) mixed effects spline model. Model log-marginal likelihood for the Bayesian wavelet model (c).
Yeast cell cycle data:

- Gene expression data from DNA microarray experiments - insight about genes that behave similarly in time. There are 6178 genes from α-pheromone synchronized yeast cell cultures. With the induction of the α-pheromone, the expression levels for the pheromone response pathway (PRP) genes show a steep rise, that is of interest.

- 600 significantly expressed genes were considered, with 16 equisampled measurements each; some expression profiles were incomplete.

![Graph](a)

<table>
<thead>
<tr>
<th>$d_n$</th>
<th>Best models</th>
<th>Log-marginal-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1,2</td>
<td>$-8.697 \times 10^4$</td>
</tr>
<tr>
<td></td>
<td>3,4,5</td>
<td>$-8.699 \times 10^4$</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>$-8.700 \times 10^4$</td>
</tr>
<tr>
<td></td>
<td>2,3</td>
<td>$-8.701 \times 10^4$</td>
</tr>
<tr>
<td>6</td>
<td>1,2</td>
<td>$-8.701 \times 10^4$</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>$-8.703 \times 10^4$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$-8.704 \times 10^4$</td>
</tr>
</tbody>
</table>

**Figure 2:** (a)Preferred number of clusters over 10,000 MCMC iter.,(b)Log-marginal likelihood for the best models.
Appendix

Mixed effects spline model:

\[ Y_i = S\beta_{1,i} + S\beta_{2,i} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_i^2 I_m), \]

where \( S(m \times p) \) is a natural cubic spline design, \( \beta_{1,i} \) are cluster specific coefficients and \( \beta_{2,i} \) account for individual variations in the functions within each cluster.

The original EM implementation was replaced by Bayesian modeling as follows:

\[ \beta_{1,i} \sim G \]
\[ GD \sim DP(\alpha, N(0, \sigma_i^2 \Gamma_1)) \]
\[ \Gamma_1 \sim IW(R_1, s_1) \]
\[ \beta_{2,i} \sim N(0, \sigma_i^2 \Gamma_2) \]
\[ \Gamma_2 \sim IW(R_2, s_2). \]
Discussion

The nonparametric Bayes model offers a flexible approach to functional clustering and has been shown to perform favorably against other functional clustering methods.

This method learns about the number of clusters in an automated manner (due to the DP structure), unlike other clustering methods where a dimension change comes with a huge computational burden.

The discrete wavelet transform requires that the number of sampled points $m$ be an integer power of 2. There exist some alternatives that do not place restrictions on the discrete support. This could allow the extension of this model to unequispaced data.

There is no comment about the execution time (using an MCMC inference, we would expect a long running time)!