Mixtures and Hidden Markov Models for analyzing genomic data

Marie-Laure Martin-Magniette

UMR AgroParisTech/INRA Mathématique et Informatique Appliquées, Paris
UMR INRA/UEVE ERL CNRS Unité de Recherche en Génomique Végétale, Evry
Responsable de l’équipe Bioinformatique pour la génomique prédictive

4 février 2014
Plan

1. Biological context
2. Mixture model / Gene expression
3. Hidden Markov Model / Whole-genome analysis
Transcriptome = set of transcribed genes in a sample.

Transcriptome data are the main source of ’omic information available for living organisms

- Microarrays (∼1995 - )
- High-throughput sequencing (HTS) : RNA-seq (∼2008 - )
Microarray principle

It requires a support and biological knowledge on the organism.

Data are hybridization signal (continuous data) for one or two samples
Transcriptome-based analysis

Goal: deciphering the functions of the genes

**Differential analysis**
- Provides a list of genes behaving differentially across several biological conditions
- Based on hypothesis tests with a false-positive control
- Probe density increases -> genome-wide study is achievable

**Co-expression study**
- Eisen *et al.* (1998, PNAS) showed that co-expressed genes may be involved in the same biological process(es).
- Co-expression means similar expression profiles across many experiments
- Based on clustering methods
- A first step towards the functional annotation of an organism
Each biological question is specific but a set of questions can be reformulated into a same methodological framework.

For example, one question is to reveal an underlying structure in transcriptomic data.

Models with latent variables define a natural framework:

1. Model: analysis of genomic data
2. High dimension: co-expression study
3. Spatial dependences: whole-genome analysis
Plan

1. **Biological context**

2. **Mixture model / Gene expression**
   - Introduction on model-based clustering
   - Application in genomics
     - MixThres: Truncated Gaussian mixtures
     - ChIPmix: Mixtures of linear regressions

3. **Hidden Markov Model / Whole-genome analysis**
Independent mixture models

what we observe | the model | the expected results
---|---|---

\[ Z = ? \]

\[ Z : 1 = \circ, 2 = +, 3 = * \]

- Introduction of a latent variable: \( Z_i \sim M(1; \pi_1, \ldots, \pi_D) \)
- Let \( y = (y_1, \ldots, y_n) \) the observations with \( y_i \in \mathbb{R}^Q \)
- \( y_i \)'s are independent conditionally to \( Z \)

\[
(y_i | Z_i = d) \sim \psi(\cdot; \gamma_d)
\]

\[
f(y) = \prod_{i=1}^{n} \sum_{d=1}^{D} \pi_d \varphi(y_i; \gamma_d), \text{ where } \alpha = (\pi_1, \ldots, \pi_{D-1}, \gamma_1, \ldots, \gamma_D)
\]
Bayesian Information Criterion : Schwarz (1978)

- $P(y|m)$ is typically difficult to calculate
- An asymptotic approximation of $2 \ln \{ P(y|m) \}$ is generally used.
- This approximation is the Bayesian Information Criterion (BIC)

$$BIC(m) = \log P(y|m, \hat{\theta}) - \frac{\nu_m}{2} \log(n).$$

where

- $\nu_m$ is the number of free parameters of the model $m$
- $P(y|m, \hat{\theta})$ is the maximum likelihood under this model.

- The selected model $\hat{m}$ maximizes the BIC criterion.
- BIC consistently estimates the number of components (Keribin C., 2000).
- BIC is expected to mostly select the dimension according to the global fit of the model.
Integrated Information Criterion (ICL)

- Biernacki et al. (2000) proposed a criterion based on the integrated complete likelihood

\[ P(y, Z|m) = \int P(y, Z|m, \theta) \pi(\theta|m) d\theta \]

- Using a BIC-like approximation of this integral

\[ ICL(m) = \log P(y, \hat{Z}|m, \hat{\theta}) - \frac{\nu m}{2} \log(n), \]

where \( \hat{Z} \) stands for posterior mode of \( Z \).

- McLachlan & Peel (2000) proposed to replace \( \hat{Z} \) with the conditional expectation of \( Z \) given the observation

\[ ICL(m) = \log P(y|m, \hat{\theta}) - \mathcal{H}_Y(Z) - \frac{\nu m}{2} \log(n), \]

where

\[ \mathcal{H}_Y(Z) = -\mathbb{E} \left[ \log P(Z|y, m, \hat{\theta}) \right] = -\sum_{i=1}^{n} \sum_{d=1}^{D} \tau_{id} \log \tau_{id} \]
Conclusions for the model selection

- Several asymptotic criteria exist
  - BIC aims at finding a good number of components to a global fit of the data distribution
  - ICL is dedicated to a classification purpose. The penalty has an entropy term that penalizes stronger models for which the classification is uncertain.
- Recent works have been done in a non-asymptotic context (see Maugis & Bertrand, 2010)
Classification rules (McLachlan & Peel, 2000)

Always based on the conditional probability:

\[
\tau_k(y) = P(Z = k | Y = y) = \frac{\pi_k \psi(y; \gamma_k)}{\sum_{\ell=1}^{K} \pi_{\ell} \psi(y; \gamma_{\ell})}
\]

Maximum A Posteriori rule:

\[
\psi_{\text{MAP}}(y) = \arg \max_k \tau_k(y)
\]

- optimality in the sense of prediction error
- but optimality does not prevent against misclassification
- all components are of interest

Thresholded MAP rule:

\[
\psi_{T}(y) = \begin{cases} 
    p & \text{if } p = \arg \max_k \tau_k(y) \text{ and } \tau_p(y) > 1 - \alpha, \\
    0 & \text{otherwise,}
\end{cases}
\]

where \(0 < \alpha < 1\) is a parameter to be chosen by the experimenter.
MixThres : Truncated Gaussian mixtures

Collaboration with F. Picard

Where does the background stop and where does the signal begin?

MixThres method

- A mixture of truncated Gaussian mixture to provide a good fit of the intensity histogram
- An adapted EM algorithm to estimate the parameters
- A model selection according to BIC
- An hybridization threshold based on the conditional probabilities

→ bounded signal
→ left peak.
Where does the background stop and where does the signal begin?

**MixThres method**

- A mixture of truncated Gaussian mixture to provide a good fit of the intensity histogram
- An adapted EM algorithm to estimate the parameters
- A model selection according to BIC
- An hybridization threshold based on the conditional probabilities

→ bounded signal
→ left peak.
selected mixture model
4 Gaussian components truncated on the right at 15.73
Discovery of new genes

- 677 probes of CATMA array were designed outside the official genes
- Analysis of 522 transcriptome samples
- 465 new genes identified

Aubourg et al. (2007), BMC Genomics
ChIP-chip experiments

Collaboration with C. Bérard, T. Mary-Huard and S. Robin

A relationship exists between IP and Input and can be modeled
MultiChIPmix: Mixture of two linear regressions

- Let $Z_i$ the status of the probe $i$: $P(Z_i = 1) = \pi$
- The linear relation between IP and Input depends on the probe status

$$IP_i = \begin{cases} 
  a_0 + b_0 \text{Input}_i + E_i & \text{if } Z_i = 0 \text{ (normal)} \\
  a_1 + b_1 \text{Input}_i + E_i & \text{if } Z_i = 1 \text{ (enriched)} 
\end{cases}$$

$$V(IP_i) = \sigma^2$$

Martin-Magniette et al. (2008), Bioinformatics
Let $Z_i$ the status of the probe $i$: $P(Z_i = 1) = \pi$

The linear relation between IP and Input depends on the probe status

$$IP_{ir} = \begin{cases} 
    a_0r + b_0r Input_{ir} + E_{ir} & \text{if } Z_i = 0 \text{ (normal)} \\
    a_1r + b_1r Input_{ir} + E_{ir} & \text{if } Z_i = 1 \text{ (enriched)}
\end{cases}$$

$V(IP_{ir}) = \sigma_r^2$
Roudier et al. (2011), EMBO Journal
Moghaddam et al. (2011), Plant Journal
Bérard et al. (2013), BMC Bioinformatics
1. Biological context

2. Mixture model / Gene expression

3. Hidden Markov Model / Whole-genome analysis
   - Introduction on HMM
   - Application in genomics
   - Mixture for estimating emission distributions
   - Conclusions
- Probes are (almost) equally spaced along the genome
- Probes tend to be clustered
The $y_t$'s are still independent conditionally to the $Z_t$'s:

$$(y_t | Z_t = d) \sim \psi(\cdot; \gamma_d)$$

But the latent variables are (Markov-)dependent:

$$\{Z_t\} \sim MC(\Pi)$$

$$\pi_{k\ell} = P(Z_t = k | Z_{t-1} = \ell)$$
Hidden Markov Model with biological knowledge

- $C_t$ = annotation of the probe $t$ (intron, exon, intergenic, ...)
- $Z_t$ (status of the probe) $\sim$ Markov chain
  
  \[ P(Z_t = d | Z_{t-1} = l, C_t = p) = \pi_{ld}^{C_t} \rightarrow \text{one transition matrix for each annotation category} \]

- $Y$ : observed signal (Int$_1$, Int$_2$)
  
  \[ (Y_t | Z_t = d) \sim \mathcal{N}(\mu_d, \Sigma_d) \text{ for } d = 1, ..., 4 \]
Hidden Markov Model with biological knowledge

- $C_t =$ annotation of the probe $t$ (intron, exon, intergenic, ...)
- $Z_t$ (status of the probe) $\sim$ Markov chain
- $P(Z_t = d | Z_{t-1} = l, C_t = p) = \pi_{ld}^{C_t}$ → one transition matrix for each annotation category
- $Y:$ observed signal ($\text{Int}_1, \text{Int}_2$)
- $(Y_t | Z_t = d) \sim \mathcal{N}(\mu_d, \Sigma_d)$ for $d = 1, ..., 4$

$\mu =$ mean vector $\sim$ center of the ellipse
$\Sigma =$ variance matrix $\sim$ shape, orientation, volume of the ellipse
Hidden Markov Model with biological knowledge

- \( C_t \) = annotation of the probe \( t \) (intron, exon, intergenic, ...)
- \( Z_t \) (status of the probe) \( \sim \) Markov chain
- \( P(Z_t = d | Z_{t-1} = l, C_t = p) = \pi_{ld}^{C_t} \rightarrow \) one transition matrix for each annotation category
- \( Y \) : observed signal (Int\(_1\), Int\(_2\))
  \( (Y_t | Z_t = d) \sim \mathcal{N}(\mu_d, \Sigma_d) \) for \( d = 1, ..., 4 \)
  - Constraints on the variance matrices
- Maximum likelihood estimation under constraints using the EM algorithm
- Posterior probability : \( \tau_{td} = \Pr\{Z_t = d | Y\} \) (Forward-Backward)
Sub-models

- without annotation $\Rightarrow$ Hidden Markov Model with 4 hidden states
- without probe dependence $\Rightarrow$ Mixture Model of 4 components per annotation category
- without probe dependence and annotation $\Rightarrow$ Mixture Model of 4 components
- For some dataset, the number of hidden states could be lower
- All of these models are compared with model selection criterion (BIC or ICL)

Bérard et al. (2011), SAGMB
Application on *Arabidopsis thaliana* transcriptomic dataset

- **Experiment**: Seed and leaf samples co-hybridized on a same tiling array
- **Two biological replicats performed in dye-swap**

<table>
<thead>
<tr>
<th></th>
<th>Mixture</th>
<th>HMM</th>
<th>Mixture + annot.</th>
<th>HMM + annot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>param. nb</td>
<td>18</td>
<td>30</td>
<td>24</td>
<td>60</td>
</tr>
<tr>
<td>BIC</td>
<td>406457</td>
<td>371656</td>
<td>373561</td>
<td>357312</td>
</tr>
<tr>
<td>ICL</td>
<td>436185</td>
<td>416529</td>
<td><strong>399975</strong></td>
<td>400330</td>
</tr>
</tbody>
</table>

According to BIC, the HMM with annot. cat. is the best

According to ICL, mixture model per annot. cat. should be chosen, meaning that annotation information contains information about spatial dependence.
In all cases, emission distributions aren’t well fitted.
Gaussian mixture as emission distribution

PhD of C. Bérard and S. Volant, co-supervised with S. Robin

\[ (y_t | Z_t = d) \sim \sum_{k=1}^{K_d} \lambda_{dk} \phi(\cdot; \mu_{dk}, \Sigma_{dk}) \]

State diagram:

\[ Z_{t-1} \rightarrow Z_t \rightarrow Z_{t+1} \]

Observation diagram:

\[ y_{t-1} \rightarrow y_t \rightarrow y_{t+1} \]

Model collection:

\[ \mathcal{M} = \left\{ m := (D, K_1, \ldots, K_D); 1 \leq D \ \forall d \ K_d \geq 1 \ \text{with} \ \sum_{d=1}^{D} K_d = K \right\} \]
Gaussian mixture as emission distribution

\[ (y_t | Z_t = d) \sim \sum_{k=1}^{K_d} \lambda_{dk} \phi(\cdot; \mu_{dk}, \Sigma_{dk}) \]

state \( \ldots \rightarrow Z_{t-1} \rightarrow Z_t \rightarrow Z_{t+1} \rightarrow \ldots \)

component \( \ldots \rightarrow W_{t-1} \rightarrow W_t \rightarrow W_{t+1} \rightarrow \ldots \)

observation \( \ldots \rightarrow y_{t-1} \rightarrow y_t \rightarrow y_{t+1} \rightarrow \ldots \)

Model collection

\[ \mathcal{M} = \left\{ m := (D, K_1, \ldots, K_D); 1 \leq D \ \forall d \ K_d \geq 1 \text{ with } \sum_{d=1}^{D} K_d = K \right\}, \]
Inference and model selection

Inference for \( m := (D, K_1, \ldots, K_D) \) still possible with an EM algorithm

**E-step**  Forward/Backward algorithm to estimate \( P(Z, W|Y; \theta^h) \)

**M-step**  Maximization of the completed likelihood

\[ \mathbb{E}_Y[\log P(Y, Z, W)] \]

Overview of HMM\textsubscript{MIX}

1. Fit an HMM with \( K \) components.
2. From \( G = K, K - 1, \ldots, 1 \)
   - Select the clusters \( k \) and \( l \) to be combined based on the maximization of \( \mathbb{E}_Y[\log P(Y; G'_k \cup l)] \)
   - Update the parameters with a few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of states

\[ \hat{D} = \arg\max_{k \in \{K, \ldots, 1\}} ICL_Z(k) \]
Inference and model selection

Inference for $m := (D, K_1, \ldots, K_D)$ still possible with an EM algorithm

**E-step** Forward/Backward algorithm to estimate $P(Z, W| Y; \theta^h)$

**M-step** Maximization of the completed likelihood $\mathbb{E}_Y[\log P(Y, Z, W)]$

**Overview of HMMMix**

1. Fit an HMM with $K$ components.
2. From $G = K, K - 1, \ldots, 1$
   - Select the clusters $k$ and $l$ to be combined based on the maximization of $\mathbb{E}_Y[\log P(Y; G'_{k,l})]$
   - Update the parameters with a few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of states

$$\hat{D} = \arg \max_{k \in \{K, \ldots, 1\}} ICL_Z(k)$$
Inference and model selection

Inference for $m := (D, K_1, \ldots, K_D)$ still possible with an EM algorithm

**E-step** Forward/Backward algorithm to estimate $P(Z, W|Y; \theta^h)$

**M-step** Maximization of the completed likelihood $\mathbb{E}_Y[\log P(Y, Z, W)]$

Overview of HMM\textsubscript{MIX}

1. Fit an HMM with $K$ components.
2. From $G = K, K - 1, \ldots, 1$
   - Select the clusters $k$ and $l$ to be combined based on the maximization of $\mathbb{E}_Y[\log P(Y; G'_{k\cup l})]$
   - Update the parameters with a few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of states

$$\hat{D} = \arg\max_{k \in \{K, \ldots, 1\}} \text{ICL}_Z(k)$$
Inference and model selection

Inference for $m := (D, K_1, \ldots, K_D)$ still possible with an EM algorithm

**E-step** Forward/Backward algorithm to estimate $P(Z, W | Y; \theta^h)$

**M-step** Maximization of the completed likelihood $\mathbb{E}_Y[\log P(Y, Z, W)]$

Overview of HMM_{MIX}

1. Fit an HMM with $K$ components.
2. From $G = K, K - 1, \ldots, 1$
   - Select the clusters $k$ and $l$ to be combined based on the maximization of $\mathbb{E}_Y[\log P(Y; G_k^l)]$
   - Update the parameters with a few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of states

$$\hat{D} = \text{argmax } ICL_Z(k)$$

$k \in \{K, \ldots, 1\}$
Comparisons of two IP samples

5000 probes of Chromosome 4 of *Arabidopsis thaliana*

HMMMix with $K = 40 \rightarrow \hat{D} = 8$

Volant *et al.* (2013), Stat & Computing
Conclusions

- Mixture models and HMM are relevant for identifying latent structures
- Available softwares: mixmod and Rmixmod
- Methodological questions on modeling, parameter estimation and model selection
- Numerous applications of mixture and HMM for transcriptomic data
- New challenge is to develop Mixture models and HMM for High-throughput sequencing (HTS) data
High-throughput sequencing (HTS) data

Gene 1
Gene 2
Gene 1
Gene 2
Sample 1
Sample 2

M.L Martin-Magniette (INRA)

Mixture models and HMM for genomic
High-throughput sequencing data, continued

**Advantage of HTS data**
- Does not require prior genome knowledge
- Best tool for allele specific or gene family expression analysis

**Statistical challenges of HTS data**
- Discrete, non-negative, and skewed data with very large dynamic range (up to 5 orders of magnitude)
- Counts are correlated with gene length
- Sequencing depth (= “library size”) varies among experiments. It is the major technical bias.

Evaluation of normalization methods in Dillies et al. (2012, Briefings in bioinformatics).