SpaCEM$^3$: a software for the spatial clustering of incomplete, high dimensional data

Damien Leroux$^1$, Matthieu Vignes$^1$, Juliette Blanchet$^2$, Florence Forbes$^3$

http://spacem3.gforge.inria.fr/

$^1$BIA Unit, INRA Toulouse

$^2$WSL - SLF, Davos

$^3$Mistis Project, INRIA Rhône-Alpes

ModGraphII - 6 Sept. 2010
Outline

Introduction

Models included in the software for classifying objects
- Hidden Markov Random Fields
- Gaussian model for high-dimensional data
- Supervised classification with Triplet Markov fields

Algorithms implemented in the software
- Standard algorithms
- Variational (mean field-like) EM approximations for the Markovian modelling

Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion

Summary and perspectives

Some reading
Outline

Introduction

Models included in the software for classifying objects
- Hidden Markov Random Fields
- Gaussian model for high-dimensional data
- Supervised classification with Triplet Markov fields

Algorithms implemented in the software
- Standard algorithms
- Variational (mean field-like) EM approximations for the Markovian modelling

Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion

Summary and perspectives

Some reading
Outline

Introduction

Models included in the software for classifying objects
   Hidden Markov Random Fields
   Gaussian model for high-dimensional data
   Supervised classification with Triplet Markov fields

Algorithms implemented in the software
   Standard algorithms
   Variational (mean field-like) EM approximations for the Markovian modelling
   Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion

Summary and perspectives

Some reading
Outline

Introduction

Models included in the software for classifying objects
  - Hidden Markov Random Fields
  - Gaussian model for high-dimensional data
  - Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  - Standard algorithms
  - Variational (mean field-like) EM approximations for the Markovian modelling
  - Practical use of the algorithms

Model selection
Outline

Introduction

Models included in the software for classifying objects
  - Hidden Markov Random Fields
  - Gaussian model for high-dimensional data
  - Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  - Standard algorithms
  - Variational (mean field-like) EM approximations for the Markovian modelling
  - Practical use of the algorithms

Model selection

Example of use on biological data
Outline

Introduction

Models included in the software for classifying objects
  Hidden Markov Random Fields
  Gaussian model for high-dimensional data
  Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  Standard algorithms
  Variational (mean field-like) EM approximations for the
  Markovian modelling
  Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion
  Summary and perspectives
  Some reading
Outline

Introduction

Models included in the software for classifying objects
  Hidden Markov Random Fields
  Gaussian model for high-dimensional data
  Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  Standard algorithms
  Variational (mean field-like) EM approximations for the
  Markovian modelling
  Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion

  Summary and perspectives
  Some reading
Introduction

- Goal: classify objects of interest (image pixels, genes . . . ) from *complex* datasets.
Introduction

• Goal: classify objects of interest (image pixels, genes . . . ) from *complex* datasets.

• Peculiar features of data we focused on:
  • there are *dependencies* between objects,
  • data are *high-dimensional* and
  • some measures can be *missing*.
Introduction

• Goal: classify objects of interest (image pixels, genes . . . ) from complex datasets.

• Peculiar features of data we focused on:
  • there are dependencies between objects,
  • data are high-dimensional and
  • some measures can be missing.

• Markovian setting.
Introduction

• Goal: classify objects of interest (image pixels, genes…) from complex datasets.

• Peculiar features of data we focused on:
  • there are dependencies between objects,
  • data are high-dimensional and
  • some measures can be missing.

• Markovian setting.

• Applications: Image analysis (biomedical, satellite surveys….) More generally computer vision), genomics datasets….
Included functionalities

- Unsupervised clustering based on Hidden Markov Random Fields (HMRF).
Included functionalities

- **Unsupervised clustering** based on Hidden Markov Random Fields (HMRF).
- **Supervised classification** based on Triplet Markov models.
Included functionalities

- Unsupervised clustering based on Hidden Markov Random Fields (HMRF).
- Supervised classification based on Triplet Markov models.
- Criteria for Model selection.
Included functionalities

- Unsupervised clustering based on Hidden Markov Random Fields (HMRF).
- Supervised classification based on Triplet Markov models.
- Criteria for Model selection.
- Data simulation to generate either images or classical graphs.
Outline

Introduction

Models included in the software for classifying objects
  - Hidden Markov Random Fields
  - Gaussian model for high-dimensional data
  - Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  - Standard algorithms
  - Variational (mean field-like) EM approximations for the Markovian modelling
  - Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion
  - Summary and perspectives
  - Some reading
Markov Random Field (MRF)

Definition

\( Z = (Z_1 \ldots Z_n) \) is a Markov Random Field iff:

(i) \( P(Z_i|Z) = P(Z_i|Z_{N_i}) \) and

(ii) \( P(Z = z) > 0. \)
Markov Random Field (MRF)

Definition
\( \mathbf{Z} = (Z_1 \ldots Z_n) \) is a Markov Random Field iff:
(i) \( P(Z_i|\mathbf{Z}) = P(Z_i|\mathbf{Z}_{\bar{N}_i}) \) and
(ii) \( P(\mathbf{Z} = \mathbf{z}) > 0. \)

Consequence: (Hamersley-Clifford Theorem) \( \mathbf{Z} \) has a Gibbs distribution: \( \frac{\exp(-H(\mathbf{z}))}{\mathcal{W}} \) where the energy function is decomposed on clique potentials: \( H(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z}_c). \)
Markov Random Field (MRF)

Definition

\( \mathbf{Z} = (Z_1 \ldots Z_n) \) is a Markov Random Field iff:

(i) \( P(Z_i|\mathbf{Z}) = P(Z_i|\mathbf{Z}_{N_i}) \) and 
(ii) \( P(\mathbf{Z} = \mathbf{z}) > 0. \)

Consequence: (Hamersley-Clifford Theorem) \( \mathbf{Z} \) has a Gibbs distribution: \( \frac{\exp(-H(\mathbf{z}))}{W} \) where the energy function is decomposed on clique potentials: \( H(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z}_c) \).

Potts model and extensions

Potentials on singletons (external field) & pairs (dependencies):

\[
H(\mathbf{z}) = \sum_i V_i(z_i) + \sum_{j \in N_i} V_{ij}(z_i, z_j)
\]

\( = (\text{if not dep. site } i) - z_i^\alpha \)

\( = (\text{if not dep. sites } i, j) - z_i^\beta z_j \)
Hidden Markov Random Fields (HMRF)

...With independent noise (seen as a generalization of mixture models):

\[
Z \text{ MRF } + P(X|Z) = \prod_i P(X_i|Z_i) \Rightarrow (X, Z) \text{ MRF}.
\]
Hidden Markov Random Fields (HMRF)

...With independent noise (seen as a generalization of mixture models):

\[ Z \text{ MRF } + P(X|Z) = \prod_i P(X_i|Z_i) \Rightarrow (X, Z) \text{ MRF). } \]

Hence (but not equivalent to) \( Z|x \), \textit{a posteriori} distribution is a MRF as well with energy function: \( H(z, \alpha, \beta) = \sum_i \log f(x_i|\theta_{z_i}); \) classical Bayesian methods for parameter estimation and clustering can be used.

Extension to pairwise and Triplet Markov fields...See slides to come.
Gaussian model for high-dimensional data

Idea from 14 models in Banfield & Raftery, 1993 (orientation, size and shape of the distribution around the mean).
Gaussian model for high-dimensional data

Idea from 14 models in Banfield & Raftery, 1993 (orientation, size and shape of the distribution around the mean).

Models from Bouveyron et al. 2007: Spectral decomposition of the covariance matrix $\Sigma_k = Q_k \Delta_k Q_k'$:

$$\Delta_k = \begin{pmatrix} a_{k1} & 0 & \cdots & 0 \\ 0 & a_k D_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_k \end{pmatrix} \begin{pmatrix} (0) \\ (0) \\ \vdots \\ (0) \end{pmatrix} \begin{pmatrix} (0) \\ (0) \\ \vdots \\ (0) \end{pmatrix} = D_k \begin{pmatrix} (D - D_k) \\ (D - D_k) \\ \vdots \\ (D - D_k) \end{pmatrix}$$
High-D segmentation of an image of Mars.

(a) Image to be clustered, (b) A pixel spectrum, (c) Segmented image and (d) average spectra for the 4 classes.
Triplet Markov model for supervised classification

\{ Independent/unimodal \} noise hypothesis too restrictive (e.g. textures).

\[ P_G(x, y, z) \propto \exp\left[ -\sum_{i \sim j} V_{ij}(y_i, z_i, y_j, z_j) + \sum_i \log f(x_i | \theta_{y_i, z_i}) \right] \]
Triplet Markov model for supervised classification

\[ P_G(x, y, z) \propto \exp[-\sum_{i \sim j} V_{ij}(y_i, z_i, y_j, z_j) + \sum_i \log f(x_i | \theta_{y_i, z_i})] \]
Triplet Markov model for supervised classification

\{ Independent/unimodal \} noise hypothesis too restrictive (e.g. textures).

\[
P_G(x, y, z) \propto \exp\left[ - \sum_{i \sim j} V_{ij}(y_i, z_i, y_j, z_j) + \sum_i \log f(x_i | \theta_{y_i, z_i}) \right]
\]

**Learning**: estimate \((X, Y|Z) \rightarrow \theta_{lk}\) and \(B_{kk'}\).

**Testing**: estimate \((X, (Y, Z)) \rightarrow C\) (given fixed \(\theta\) and \(B\)).

Then perform clustering.
Triplet Markov model simulation

Simulations with $L=K=2$; each of the 4 different $(y_i, z_i)$ is associated to a different gray level.

(a) $(Y, Z)$ realization, (b) $Z$ realization, (c) $X$ realization and (d) realization of an HMRF with added independent noise $\mathcal{N}(0, 0.3)$. 

$\begin{align*}
  b &= -2 \\
  c &= 2
\end{align*}$

$\begin{align*}
  b &= 2 \\
  c &= -2
\end{align*}$
**Triplet Markov model simulation**

Simulations with $L=K=2$; each of the 4 different $(y_i, z_i)$ is associated to a different gray level.

(a) $(Y, Z)$ realization, (b) $Z$ realization, (c) $X$ realization and (d) realization of an HMRF with added independent noise $\mathcal{N}(0, 0.3)$.

**Drawback:** supervised framework needed (to achieve identifiability).
Outline

**Introduction**

**Models included in the software for classifying objects**
- Hidden Markov Random Fields
- Gaussian model for high-dimensional data
- Supervised classification with Triplet Markov fields

**Algorithms implemented in the software**
- Standard algorithms
- Variational (mean field-like) EM approximations for the Markovian modelling
- Practical use of the algorithms

**Model selection**

**Example of use on biological data**

**Conclusion**
- Summary and perspectives
- Some reading
Standard algorithms

• Iterated Conditional Modes,
• k-means,
• ...and extensions (Clustering EM, Neighbour EM and NCEM).
EM with spatial dependencies and missing observations?

\[ x = (x^o, x^m) \]

\[ x^o = \{x^o_i\} \]

\[ x^m = \{x^m_i\} \]

MAR hypothesis \( P(m|x, z) = P(m|x^o) \).
EM with spatial dependencies and missing observations?

MAR hypothesis ($P(m|x, z) = P(m|x^o)$).

EM aims to maximize the completed likelihood:

$$\psi^{(q+1)} = \arg \max E \left[ \log P(x^o, X^m, Z | \psi) | x^o, \psi^{(q)} \right]$$

...Intractable with a Z MRF!
Neighbour Recovery EM (NREM) with missing observations

... but OK with a factorized distribution $\rightarrow$ Celeux et al., 2003.

$$P_G(Z) \approx \prod_i Q_i(Z_i)$$
Neighbour Recovery EM (NREM) with missing observations

...but OK with a factorized distribution $\rightarrow$ Celeux et al., 2003.

$$P_G(Z) \approx \prod_i P(Z_i|\tilde{Z}_{Ni})$$

(MF-like approximation)
Neighbour Recovery EM (NREM) with missing observations

...but OK with a factorized distribution → Celeux et al., 2003.

\[ P_G(Z) \approx \prod_i P(Z_i|\tilde{Z}_{Ni}) \]

(MF-like approximation)

Iterative EM-like procedure:

**NR** Fix a \( \tilde{z} \) configuration from \( x^o \) and \( \psi^{(q)} \). In particular \( \tilde{z} \) can be simulated according to \( P(Z|x^o, \psi^{(q)}) \) (Gibbs sampler): Simulated Field (SF) algorithm.

**EM** Apply EM on factorized model to update \( \psi^{(q+1)} \).

Finally MAP (or MPM) to reconstruct \( z \). But also \( x^m \).
Practical use of the algorithms

- **Data loading**
  - **Image** (regular 2-D grid): raw pixel values; specify size, number of dimensions, neighbouring type (4/8)
  - **Graph** (user-defined structure): specify neighbours list file instead of size
Practical use of the algorithms

- **Data loading**
  - **Image** (regular 2-D grid): raw pixel values; specify size, number of dimensions, neighbouring type (4/8)
  - **Graph** (user-defined structure): specify neighbours list file instead of size

- **Choose number of classes**: fixed or range
Practical use of the algorithms

- **Data loading**
  - **Image** (regular 2-D grid): raw pixel values; specify size, number of dimensions, neighbouring type (4/8)
  - **Graph** (user-defined structure): specify neighbours list file instead of size

- **Choose number of classes**: fixed or range

- **Initialization**: random, k-means or user-defined.
Practical use of the algorithms

- **Data loading**
  - **Image** (regular 2-D grid): raw pixel values; specify size, number of dimensions, neighbouring type (4/8)
  - **Graph** (user-defined structure): specify neighbours list file instead of size

- **Choose number of classes**: fixed or range

- **Initialization**: random, k-means or user-defined.

- **Stopping criterion**:
  - Likelihood convergence
  - Fuzzy classification
  - Hard classification
  - Number of iterations.
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Classification models</th>
<th>Algorithms</th>
<th>Model selection</th>
<th>Example of use on biological data</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

**Outline**

**Introduction**

**Models included in the software for classifying objects**
- Hidden Markov Random Fields
- Gaussian model for high-dimensional data
- Supervised classification with Triplet Markov fields

**Algorithms implemented in the software**
- Standard algorithms
- Variational (mean field-like) EM approximations for the Markovian modelling
- Practical use of the algorithms

**Model selection**

**Example of use on biological data**

**Conclusion**

- Summary and perspectives
- Some reading
Model selection: selection criteria

The "best" model = compromise between fitting the data (adequacy to what is observed) and allowed complexity (!!Overfitting!!). Amongst the many existing criteria we use the Bayes Information Criterion (BIC, Schwarz, Ann. Stat. 1978) and the Integrated Completed Likelihood (ICL, Biernacki et al., IEEE PAMI 2000).
Model selection: selection criteria

The "best" model = compromise between fitting the data (adequacy to what is observed) and allowed complexity (!!Overfitting!!). Amongst the many existing criteria we use the Bayes Information Criterion (BIC, Schwarz, Ann. Stat. 1978) and the Integrated Completed Likelihood (ICL, Biernacki et al., IEEE PAMI 2000).

Approximations are needed when the model is Markovian: BIC$^p$ that approximates $P_G$ with a mean field approach while BIC$^w$ approximates the partition function $W$. Proves $BIC^p \leq BIC^w \leq BIC^{true}$ in theory (Forbes and Peyrard, 2003). Verified in practice.
Introduction

Models included in the software for classifying objects
- Hidden Markov Random Fields
- Gaussian model for high-dimensional data
- Supervised classification with Triplet Markov fields

Algorithms implemented in the software
- Standard algorithms
- Variational (mean field-like) EM approximations for the Markovian modelling
- Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion
- Summary and perspectives
- Some reading
Summary of the data analysis workflow

(Blanchet and Vignes, 2009)
Outline

Introduction

Models included in the software for classifying objects
  Hidden Markov Random Fields
  Gaussian model for high-dimensional data
  Supervised classification with Triplet Markov fields

Algorithms implemented in the software
  Standard algorithms
  Variational (mean field-like) EM approximations for the
  Markovian modelling
  Practical use of the algorithms

Model selection

Example of use on biological data

Conclusion

  Summary and perspectives
  Some reading
Technical characteristics

- Written in C++: 52 classes, 30,000 lines of code.

- Present version (2.0) includes a GUI (QT library; + 20,000 lines of code) in addition to the commandline software.

- Freely downloadable (CeciLL-B licence) at http://spacem3.gforge.inria.fr/. Works on Linux (Fedora, Debian/Ubuntu packages, as well as SFX archive) and Windows environments.

- Data in either text or binary formats. Program I/O in XML format.

- Extensive documentation.
Summary and perspectives

Wrapping up

SpaCEM$^3$ is wonderful ;). Did I tell you Spatial Clustering with EM Markov Models ??
Summary and perspectives

Wrapping up
SpaCEM³ is wonderful ;). Did I tell you Spatial Clustering with EM Markov Models ??

Prospects

- Promote the use of the software (e.g. on varied molecular biology datasets): Present collaborations at the INRA in Toulouse and Application Note in Bioinformatics to be submitted soon.
- Graph not totally fixed ? incomplete ? Treat edges as missing in a similar manner to observations (theoretical work needed).
- Include different distribution: multinomial useful for ecological data (theoretical work needed).
- Triplet models for unsupervised clustering (theoretical work needed).
Some references

Jeffrey D. Banfield and Adrian E. Raftery
Model-based Gaussian and non-Gaussian clustering.

Gilles Celeux, Florence Forbes and Nathalie Peyrard.
EM procedures using mean field-like approximations for Markov model-based image segmentation.

Florence Forbes and Nathalie Peyrard.
Hidden Markov random field model selection criteria based on mean field-like approximations.

Charles Bouveyron, Stéphane Girard and Cordelai Schmidt.
High dimensional data clustering.

Juliette Blanchet and Florence Forbes.
Triplet Markov fields for the supervised classification of complex structure data.

Matthieu Vignes and Florence Forbes.
Gene clustering via integrated Markov models combining individual and pairwise features.

Juliette Blanchet and Matthieu Vignes.
A model-based approach to gene clustering with missing observations reconstruction in a Markov random field framework.
Thanks to our colleagues

Nathalie Peyrard, Lemine Abdallah, Sophie Choppart, Lamiae Azizi, Senan Doyle, Soraya Arias...
Thanks to our colleagues

Nathalie Peyrard, Lemine Abdallah, Sophie Choppart, Lamiae Azizi, Senan Doyle, Soraya Arias . . .

And to you

for your attention.

Any (welcome) question, remark, criticism ?