A Bayesian model for joint unmixing, clustering and classification of hyperspectral data

Adrien Lagrange
Ph.D. student at IRIT/INP-ENSEEIHT

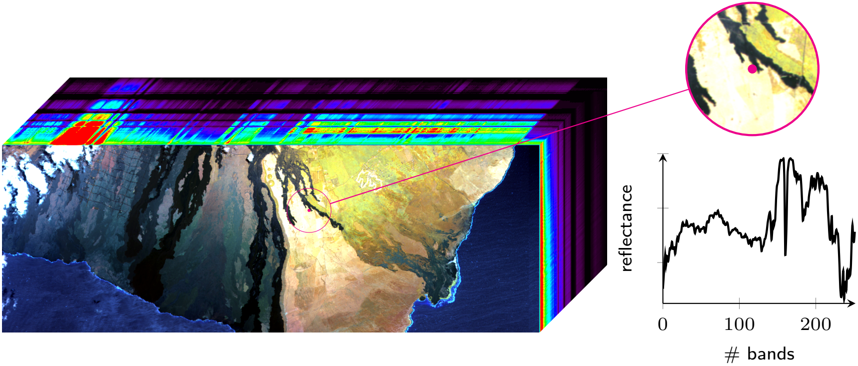
Supervisors: Nicolas Dobigeon (IRIT/INP-ENSEEIHT),
Mathieu Fauvel (DYNAFOR - INRA/INP-ENSAT) and Stéphane May (CNES)

Séminaire RESSTE – May 16th, 2017

Nature of an hyperspectral image

A remote sensing hyperspectral image is:

- same area at different wavelength → **hundreds of measurements** per pixel,
- poor spatial resolution due to sensor limitations, e.g., resolution around 10x10m per pixel for aerial applications

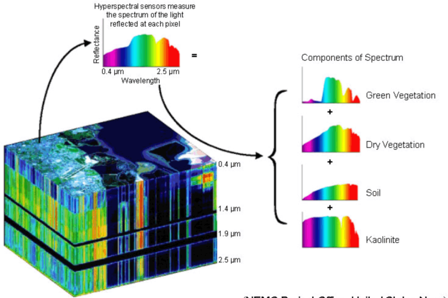


Hyperspectral image interpretation

Spectral unmixing

$$\mathbf{y}_p \approx \mathbf{M}\mathbf{a}_p$$

- $\mathbf{y}_p$: p-th observation
- $\mathbf{M}$: endmember matrix (spectra of elementary components)
- $\mathbf{a}_p$: p-th abundance vector



(NEMO Project Office, United States Navy)

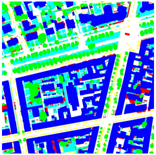
Classification

Maximum a posteriori (MAP) rule:

$$\mathbf{y}_p \text{ belongs to } j \Leftrightarrow j = \arg\max_{j \in \mathcal{J}} p(j|\mathbf{y}_p),$$
$$\Leftrightarrow j = \arg\max_{j \in \mathcal{J}} p(j)p(\mathbf{y}_p|j).$$

$\mathbf{y}_p$? buildings / water / vegetation

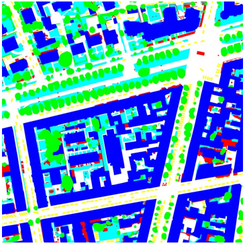
$\mathbf{y}_p$ buildings / water / vegetation



Classification

Some key issues in classification:

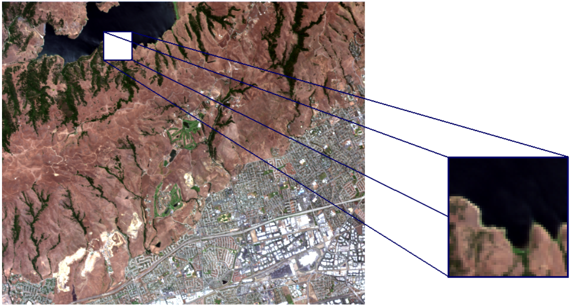
- Curse of dimensionality
- Cost of expert groundtruth
- Label noise in training set
- Multimodal classes (intraclass variability)



Spectral unmixing

One illustrative example

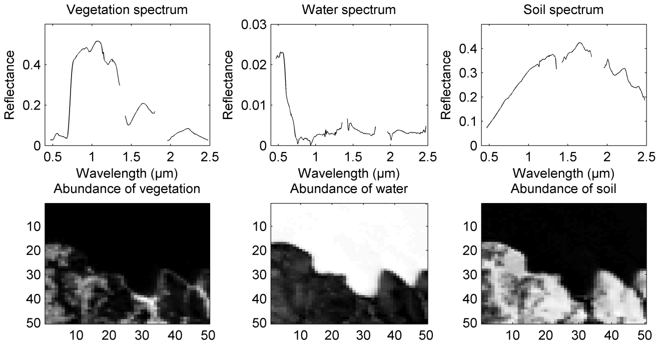
- Image: 50×50 pixels (Moffett field), $L = 224$ bands,
- 3 materials: vegetation, water, soil.



Spectral unmixing

One illustrative example

- Image: 50×50 pixels (Moffett field), $L = 224$ bands,
- 3 materials: vegetation, water, soil.



Spectral unmixing

A *matrix factorization*, *latent factor modeling* or *blind source separation* problem: $\mathbf{Y} \approx \mathbf{MA}$

1. Principal Component Analysis (PCA)
   - Searching for orthogonal "principal components" (PCs) $\mathbf{m}_r$,
   - PCs = directions with maximal variance in the data,
   - Generally used as a dimension reduction procedure.
2. Independent Component Analysis (ICA) (of $\mathbf{Y}^T$)
   - Maximizing the statistical independence between the sources $\mathbf{m}_r$,
   - Several measures of independence $\Rightarrow$ several algorithms.
3. Nonnegative Matrix Factorization (NMF)
   - Searching for $\mathbf{M}$ et $\mathbf{A}$ with positive entries,
   - Several measures of divergence $d\left(\cdot|\cdot\right) \Rightarrow$ several algorithms.
4. (Fully Constrained) Spectral Mixture Analysis (SMA)
   - Positivity constraints on $\mathbf{m}_r \Rightarrow$ positive "sources"
   - Positivity and sum-to-one constraints on $\mathbf{a}_r$
     $\Rightarrow$ mixing coefficients = proportions/concentrations/probabilities.

Objective

Spectral unmixing	Classification
Low-level biophysical information	High-level semantic information
Abundance vector per pixel	Unique label per pixel
Unsupervised	Supervised

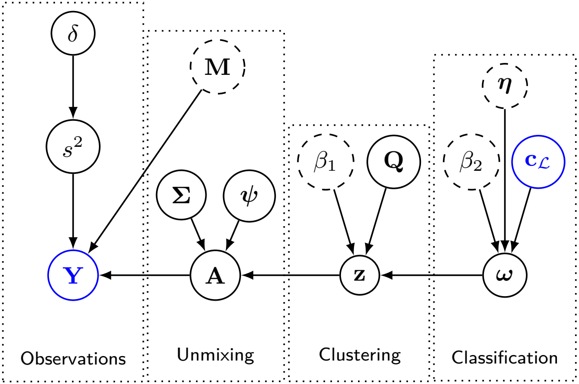
$\Longrightarrow$ **Scarcely considered jointly.**

Objective

Propose a **unified framework** to estimate jointly a classification map and a spectral unmixing from an hyperspectral image.

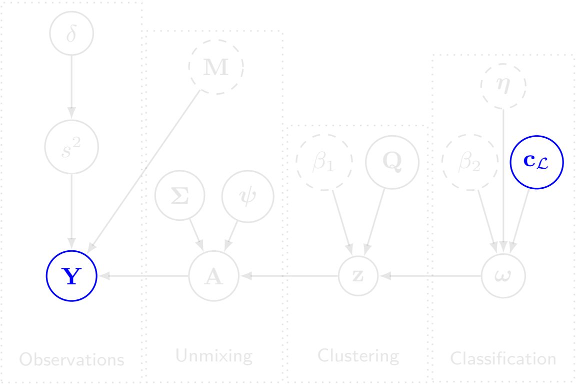
Bayesian model

- conventional linear mixing model;
- clustering of homogeneous abundance vectors;
- classification with a non-homogeneous Markov random field (MRF) to promote coherence between cluster and class labels.



Bayesian model

- conventional linear mixing model;
- clustering of homogeneous abundance vectors;
- classification with a non-homogeneous Markov random field (MRF) to promote coherence between cluster and class labels.



Linear Mixture Model (1)

Linear combination of elementary signatures corrupted by an additive Gaussian noise

$$\mathbf{y}_p = \mathbf{M}\mathbf{a}_p + \mathbf{n}_p$$

with

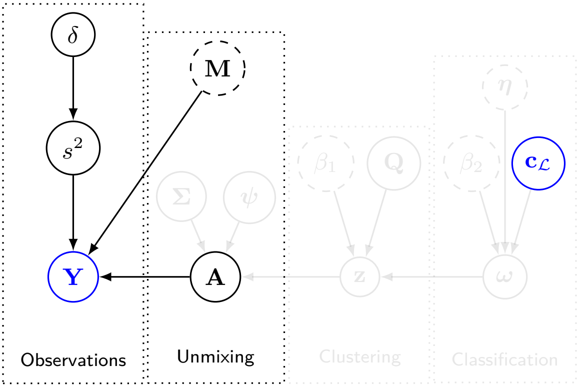
- $\mathbf{y}_p$: observation
- $\mathbf{M}$: endmember matrix (spectra of elementary components)
- $\mathbf{a}_p$: abundance vector
- $\mathbf{n}_p$: noise

Linear Mixture Model (2)

Noise prior modeling

$$\mathbf{n}_p | s^2 \sim \mathcal{N}(\mathbf{0}_D, s^2 \mathbf{I}_D),$$
$$s^2 | \delta \sim \mathcal{IG}(1, \delta), \quad p(\delta | s^2) \propto \frac{1}{\delta} \mathbb{1}_{\mathbb{R}+}(\delta).$$

Hierarchical model



Clustering (1)

- Assumption: several unknown **spectrally coherent clusters** with statistically homogeneous abundance vectors, $\forall k \in \{1, ..., K\}$,

$$\mathbf{a}_p | z_p = k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \text{ with } \boldsymbol{\Sigma}_k = \text{diag}(\sigma_{k,1}, \ldots, \sigma_{k,R})$$

where $z_1, \ldots, z_p$ are discrete labels identifying the belonging to the clusters.

- Vague priors for cluster parameters:
  - $\psi_k \sim \text{Dir}(\mathbf{1})$
    $\rightarrow$ ensures nonnegativity and sum-to-one constraints of $\mathrm{E}[\mathbf{a}_p | z_p = k]$
    (*soft constraints on* $\mathbf{a}_p$)
  - $\sigma_{k,r} \sim \mathcal{IG}(1, 0.1)$

Clustering (1)

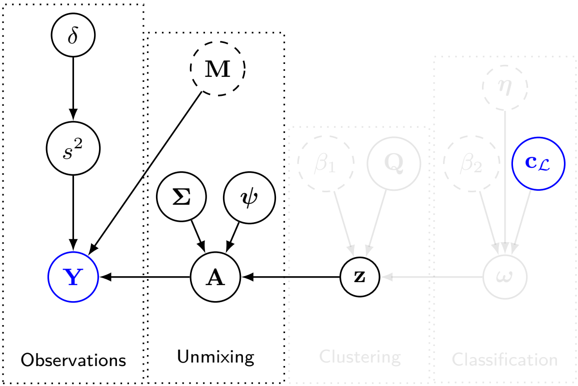
- Assumption: several unknown **spectrally coherent clusters** with statistically homogeneous abundance vectors, $\forall k \in \{1, ..., K\}$,

$$\mathbf{a}_p | z_p = k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \text{ with } \boldsymbol{\Sigma}_k = \text{diag}(\sigma_{k,1}, \ldots, \sigma_{k,R})$$

  where $z_1, \ldots, z_p$ are discrete labels identifying the belonging to the clusters.

- Vague priors for cluster parameters:
  - $\boldsymbol{\psi}_k \sim \text{Dir}(\mathbf{1})$
    $\rightarrow$ ensures nonnegativity and sum-to-one constraints of $\mathrm{E}[\mathbf{a}_p | z_p = k]$ (*soft constraints on* $\mathbf{a}_p$)
  - $\sigma_{k,r} \sim \mathcal{IG}(1, 0.1)$

Hierarchical model



Clustering (2)

Clustering with a **non-homegeneous Markov random field**

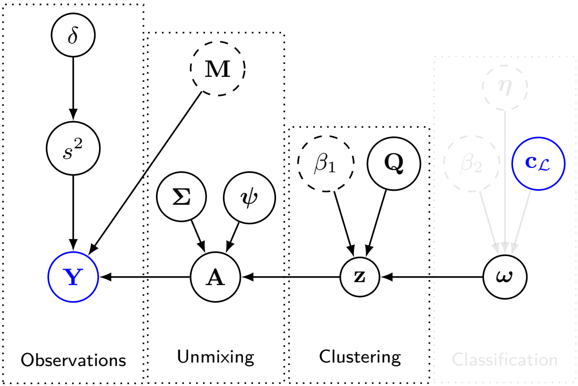
$$\mathrm{P}[z_p = k | \mathbf{z}_{\mathcal{V}(p)}, \omega_p, q_{k,\omega_p}] \propto \exp\left(V_1(k, \omega_p, q_{k,\omega_p}) + \sum_{p' \in \mathcal{V}(p)} V_2(k, z_{p'})\right)$$

with $\mathcal{V}(p)$ neighborhood of p, ω_p classification label of p.

Two potentials:

- To promote coherence with classification $\rightarrow V_1(k, j, q_{k,j}) = \log(q_{k,j})$;
- To promote spatial coherence (Potts-Markov potential) $\rightarrow V_2(k, z_{p'}) = \beta_1 \delta(k, z_{p'})$ with $\delta(\cdot, \cdot)$ Kronecker function.

Hierarchical model



Clustering (3)

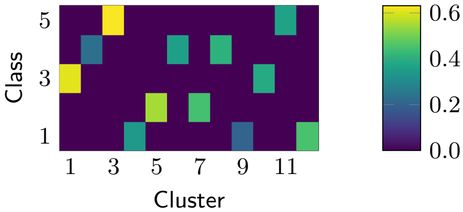
Estimation of coefficients of **interaction between high-level and low-level information**:

$$\mathbf{q}_j \sim \text{Dir}(\mathbf{1}) \rightarrow \mathbf{q}_j|\mathbf{z},\boldsymbol{\omega} \sim \text{Dir}(n_{1,j},\ldots,n_{K,j}) \quad \text{with} \quad n_{k,j} = \#\{p|z_p = k, \omega_p = j\}$$

In particular:

$$\mathrm{E}\left[q_{k,j}|\mathbf{z},\boldsymbol{\omega}\right] = \frac{n_{k,j}}{\sum_{i=1}^{K} n_{i,k}}$$
$$\approx \mathrm{P}\left[z_p = k|\omega_p = j\right]$$

Example of estimated $\mathbf{Q}$ matrix



Classification (1)

Classification rule with a Markov random field

$$\mathrm{P}[\omega_p = j | \boldsymbol{\omega}_{\mathcal{V}(p)}, c_p, \eta_p] \propto \exp\left(W_1(j, c_p, \eta_p) + \sum_{p' \in \mathcal{V}(p)} W_2(j, \omega_{p'}) \right)$$

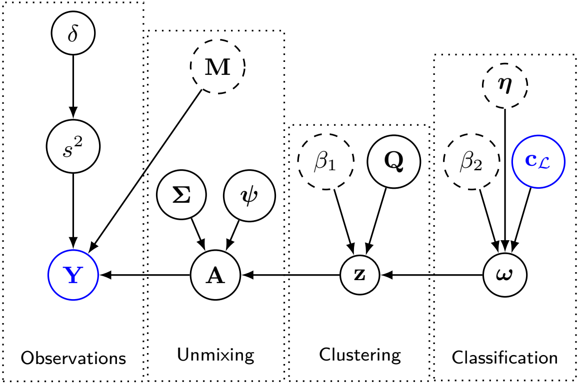
Two potentials:

- To promote coherence with labeled data

$$W_1(j, c_p, \eta_p) = \begin{cases} \begin{cases} \log(\eta_p), & \text{if } j = c_p \\ \log(\frac{1-\eta_p}{J-1}), & \text{otherwise} \end{cases} & \text{if } p \in \mathcal{L} \\ -\log(J) & \text{otherwise} \end{cases}$$

- To promote spatial coherence $\rightarrow W_2(j, \omega_{p'}) = \beta_2 \delta(j, \omega_{p'})$.

Hierarchical model



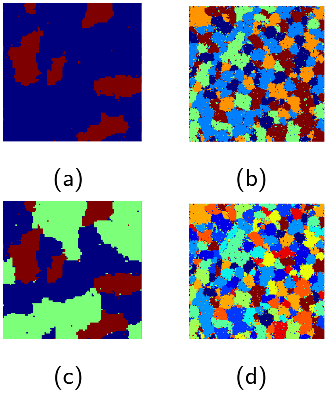
Classification (2)

Robust classification:

- $\eta_p \in (0, 1)$ the confidence in label c_p provided by user
- Possibility to correct labeled data when $\eta_p < 1$

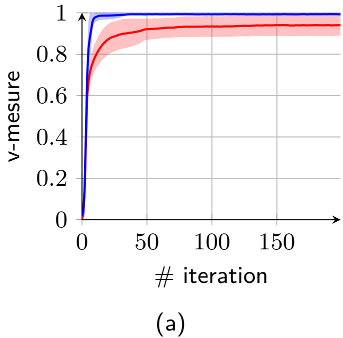
Dataset

- 413 spectral bands
- SNR = 30dB
- Clustering generated with Potts-Markov MRF
- Classes created by aggregating several clusters
- Image 1: 3 clusters, 2 classes, 3 endmenbers, 100x100px
- Image 2: 12 clusters, 5 classes, 9 endmembers, 200x200px



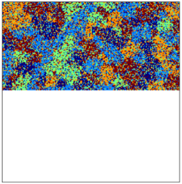
Classification map: (a) image 1, (b) image 2;
Clusters: (c) image 1, (d) image 2

Results

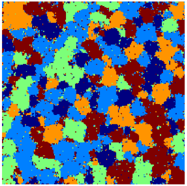


Proposed model in blue, model without classification stage (Eches *et al.*).(a) Clustering convergence for image 1, (b) Clustering convergence for image 2

Results



Provided labeled data



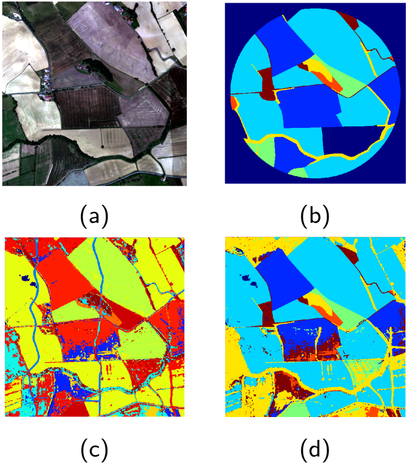
Classification obtained

- Deterioration of labeled data (40% of error)
- Confidence set to 60%

⇒ Correction of mislabeled pixels

Dataset

- 349 spectral bands
- 10 endmembers extracted with VCA
- 6 classes (straw cereal, summer crop, wooded area, artificial surfaces, bare soil, pasture)
- Top half of groundtruth provided as labeled data



Muesli dataset: (a) colored composition of data, (b) groundtruth, (c) obtained clustering and (d) obtained classification

Conclusions

- A new hierarchical Bayesian model
- Multiple outputs: abundance maps, clusters (means and variances), classification map
- Interesting byproduct for interpretation: interaction matrix describing data structure
- Robustness to labeling error and correction of errors