© Photos: DOI: 10.2516/ogst/20140013, IFPEN, X.

Dossier

This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 69, No. 2, pp. 195-372 and available online here

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 69, n°2, pp. 195-372 et téléchargeable ici

DOSSIER Edited by/Sous la direction de : **L. Duval**

Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing

Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 69 (2014), No. 2, pp. 195-372
Copyright © 2014, IFP Energies nouvelles

195 > Editorial

207 > *Multivariate Analysis for the Processing of Signals*
Traitement de signaux par analyse multivariée
J.R. Beattie

229 > *NMR Data Analysis: A Time-Domain Parametric Approach Using Adaptive Subband Decomposition*
Analyse de données RMN : une approche paramétrique basée sur une décomposition en sous-bandes adaptative
E.-H. Djermoune, M. Tomczak and D. Brie

245 > *Unsupervised Segmentation of Spectral Images with a Spatialized Gaussian Mixture Model and Model Selection*
Mélange de Gaussiennes spatialisé et sélection de modèle pour la segmentation non-supervisée d'images spectrales
S.X. Cohen and E. Le Pennec

261 > *Morphological Component Analysis for the Inpainting of Grazing Incidence X-Ray Diffraction Images Used for the Structural Characterization of Thin Films*
Analyse en composantes morphologiques pour les retouches d'images de diffraction des rayons X en incidence rasante utilisés pour la caractérisation structurale des couches minces
G. Tzagkarakis, E. Pavlopoulou, J. Fadili, G. Hadziioannou and J.-L. Starck

279 > *Inverse Problem Approach for the Alignment of Electron Tomographic Series*
Approche problème inverse pour l'alignement de séries en tomographie électronique
V.-D. Tran, M. Moreaud, É. Thiébaut, L. Denis and J.M. Becker

293 > *Design of Smart Ion-Selective Electrode Arrays Based on Source Separation through Nonlinear Independent Component Analysis*
Développement de réseaux de capteurs chimiques intelligents par des méthodes de séparation source fondée sur l'analyse de composantes indépendantes non linéaire
L.T. Duarte and C. Jutten

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 69 (2014), No. 2, pp. 245-259
Copyright © 2014, IFP Energies nouvelles
DOI: 10.2516/ogst/2014013

Dossier

Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing
Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

Unsupervised Segmentation of Spectral Images with a Spatialized Gaussian Mixture Model and Model Selection

S.X. Cohen[1]* and E. Le Pennec[2]

[1] *IPANEMA USR 3461 CNRS/MCC, Saint Aubin, BP 48, 91192 Gif-sur-Yvette - France*
[2] *École Polytechnique, CMAP UMR 7641 / Inria Saclay Idf, Select, Route de Saclay, 91128 Palaiseau Cedex - France*
e-mail: serge.cohen@synchrotron-soleil.fr - erwan.le-pennec@polytechnique.edu

* Corresponding author

Résumé — Mélange de Gaussiennes spatialisé et sélection de modèle pour la segmentation non-supervisée d'images spectrales — Dans cet article nous décrivons un nouvel algorithme de segmentation non-supervisée applicable aux images spectrales. Cet algorithme étend les techniques de classification non-supervisée fondées sur les modèles de mélange de Gaussiennes, en y incorporant les informations spatiales : les spectres sont modélisés par un mélange de K classes, chacune avec une distribution Gaussienne, dont les proportions de mélange dépendent de la position. En imposant une structure constante par morceaux aux proportions de mélange, nous construisons une procédure d'estimation, de type maximum de vraisemblance pénalisée, qui optimise simultanément la partition ainsi que les autres paramètres du modèle, en particulier le nombre de classes. Nous fournissons une garantie théorique pour cette estimation, même quand la loi génératrice ne fait pas partie des modèles envisagés, et décrivons une mise en oeuvre efficace. Finalement, nous appliquons cet algorithme à un jeu de données réel.

Abstract — Unsupervised Segmentation of Spectral Images with a Spatialized Gaussian Mixture Model and Model Selection — *In this article, we describe a novel unsupervised spectral image segmentation algorithm. This algorithm extends the classical Gaussian Mixture Model-based unsupervised classification technique by incorporating a spatial flavor into the model: the spectra are modelized by a mixture of K classes, each with a Gaussian distribution, whose mixing proportions depend on the position. Using a piecewise constant structure for those mixing proportions, we are able to construct a penalized maximum likelihood procedure that estimates the optimal partition as well as all the other parameters, including the number of classes. We provide a theoretical guarantee for this estimation, even when the generating model is not within the tested set, and describe an efficient implementation. Finally, we conduct some numerical experiments of unsupervised segmentation from a real dataset.*

INTRODUCTION

Located at the SOLEIL Synchrotron (Saint-Aubin, France), IPANEMA is a platform that is unique in the world, dedicated to the study of ancient material. It supports research projects on ancient material using the synchrotron beamlines and develops novel methodological tools to be used in these studies [1]. The high-quality light produced by SOLEIL allows, for instance, high-resolution high signal-to-noise ratio spectral image acquisition, collecting a full high-resolution spectrum for each pixel. These tools have proved to be very interesting in the ancient material study context, as shown by the conclusive studies on Stradivarius [2] varnish and microscale tissue discrimination in soft-bodied fossils from Lagerstätten [3], for instance. While studies similar to this one focus on a small set of fairly well pre-studied samples, most work in archaeology, palaeontology and cultural heritage would benefit better from the exploration of a wider set of samples that are less well pre-studied. In this context, it is beneficial to develop spectral image analysis methodologies which are both robust to low signal-to-noise ratio, enabling fast measurement of a large sample set, and require only weak prior knowledge of the samples.

Unsupervised spectral image segmentation is naturally within the scope of spectral image processing and has already been studied. While the result should be close to supervised spectral image segmentation, in which the number of classes is known and labeled examples are available for every class, unsupervised spectral image segmentation is a much harder task. Two natural approaches can be distinguished. In the first one, the spatial (or region-based) one, regions are obtained by locally grouping pixels with a similar spectrum using image segmentation techniques. In the second one, the spectral one, the spectra are clustered disregarding their spatial position using unsupervised classification techniques, and the regions are defined as the set of pixels corresponding to spectra of the same classes. The first approach yields regions that are adapted to the geometrical structures of the images but fails to detect that two disjoint regions may correspond to the same spectrum class. The second approach has exactly the inverse behavior. The first technique has been used, for instance, by Tarabalka *et al.* [4] and Bunte *et al.* [5] while the second has been used by Acito *et al.* [6] and Yang *et al.* [7].

Trying to combine these approaches to obtain a method with only the advantages is thus natural. Several directions have been explored, a review of which has been carried out by Tarabalka [8]. Amongst them the most classical are based on the hierarichal Markov field; see, for instance, the work of Farag *et al.* [9], in which spatial regularization is imposed on the clustering labels. Another direction is that of Tarabalka *et al.* [10] in which the regions are initially segmented using a spatial method and then combined according to spectral criteria.

We consider the opposite direction: extending the spectral methods to take into account the geometrical nature of images. Our proposed contribution is based on conditional density estimation by the penalized maximum likelihood technique that allows one to estimate simultaneously the number of meaningful classes and the pixel labels. Density estimation is already at the core of the most classical spectral method in which the observed spectra are modelized as a realization of a Gaussian Mixture Model (GMM). As described, for instance, by Biernacki *et al.* [11], for a given number of classes, the parameters of this mixture can be estimated and classes can be assigned by a simple maximum likelihood or maximum *a posteriori* principle. Estimating the number of classes can be performed in this setting by the penalization technique, as shown by Maugis and Michel [12]. Following ideas introduced by Kolaczyk *et al.* [13] and Antoniadis *et al.* [14], we modeled the spatial dependency through the mixing proportions of the mixture: they will depend on the pixel position to take into account the spatial inhomogeneity of spectral images. More sophisticated spatial models have been proposed, *e.g.* using a random Markov field to impose spatial constraints on the mixture proportions [15], even presenting efficient optimization algorithms for univariate or color RGB images [16].Whilst avoiding the model selection problem, these latter methods only consider the semi-unsupervised case since the proposed algorithms rely on the user to provide the number of classes, K, and typically to set the spatial regularization parameter(s).

Using the results we obtained in [17] and our extended technical report [18], we propose a true unsupervised, parameter-free methodology which includes the estimation of the number of classes, their Gaussian parameters and the spatially varying mixing proportions using a unified model selection approach. Compared with the work of Kolaczyk *et al.* [13], our proposition does not require a quantization step on the pixel-wise feature to be applicable. Furthermore, the theoretical framework we are using encompasses the frequent cases where the data is not generated by any of the tested models and the user is targeting a best approximating model within a set rather than the true model. From now on, we will call the model we are using a conditional Gaussian Mixture Model (cGMM) hereafter. The purpose of this article is to present the theoretical results, to describe an efficient numerical implementation, to discuss its calibration and to present some numerical experiments on a real dataset.

1 UNSUPERVISED SEGMENTATION BY MODEL SELECTION

Assume we observe a $n_1 \times n_2$ spectral image S, we let $n = n_1 \times n_2$ be the number of pixels, $(x_i)_{i\in\{1,\dots,n\}} = \left((x_{1,i}, x_{2,i})\right)_{i\in\{1,\dots,n\}}$ be an arbitrary ordered list of pixels and $S(x_i)$ be the observed spectrum at pixel x_i. Our goal is to assign to each pixel a class $\widehat{k}(x)$ to which the spectrum is supposed to belong. This implies estimating these classes as well as their number.

To this purpose, we use a statistical model in which each class corresponds to a Gaussian model, as in a classical GMM, but whose mixing proportions depend on the position. More precisely, we assume that the spectra $S(x_i)$ are independent realizations of law of density $s_0(\cdot|x_i)$ with respect to the Lebesgue measure that depends on the pixel position x_i. We model this conditional density by Gaussian mixtures $s(\cdots|x)$ with mixing proportions depending on the position x:

$$s(\cdot|x) = \sum_{k=1}^{K} \pi_k(x)\Phi_{\theta_k}(\cdot)$$

with K the number of mixture components, μ_k the mean of the kth component, Σ_k its covariance matrix, $\theta_k = (\mu_k, \Sigma_k)$, $\pi_k(x)$ its proportion at the position x and $\Phi_{\theta_k}(y)$ the density of a Gaussian of mean μ_k and covariance Σ_k. Each Gaussian naturally corresponds to a spectrum class. As soon as these parameters have been estimated (respectively, by $\widehat{K}$, $\widehat{\theta}_k$ and $\widehat{\pi}_k$), the spectral image segmentation is obtained by a maximum likelihood principle for the different classes:

$$\widehat{k}(y|x) = \operatorname{argmax} \widehat{\pi}_k(x)\Phi_{\widehat{\theta}_k}(y)$$

Following Kolaczyk *et al.* [13] and Antoniadis *et al.* [14], we consider mixing proportions that are piecewise constant on a hierarchical partition $\mathcal{P}$ induced by a tree structure, one of the recursive dyadic partitions of Donoho [19]. The conditional densities we consider are thus of the form:

$$s_{\mathcal{P},K,\theta,\pi}(\cdot|x) = \sum_{k=1}^{K}\left(\sum_{\mathcal{R}_l\in\mathcal{P}} \pi_k[\mathcal{R}_l]1_{\{x\in\mathcal{R}_l\}}\right)\Phi_{\theta_k}(\cdot)$$

where $\mathcal{P}$ is a partition of $\mathcal{X}$ and $\pi = (\pi[\mathcal{R}_l])_{\mathcal{R}_l\in\mathcal{P}}$, the set of proportions on each hyperrectangle $\mathcal{R}_l$, defines the function π. These parameters, as well as the number of classes K and the Gaussian parameters θ_k, will be estimated by a penalized maximum likelihood principle as described in [17, 18].

Assuming we know the number of classes K and the partition $\mathcal{P}$, as well as the *structure* of the K-uples of the Gaussian parameters (for instance, by assuming common covariance matrices or a common diagonalization basis) defined by set $\mathcal{G}$ of possible parameter K-uples, the only remaining parameters are the Gaussian parameter K-uples itself as well as the proportions $(\pi[\mathcal{R}_l])_{\mathcal{R}_l\in\mathcal{P}}$. It turns out that these parameters can be easily estimated by a maximum likelihood principle using an Expectancy Minimization (EM) type algorithm. This maximum likelihood principle is not sufficient to select the number of classes, the partition or even the structure of the K-uples: the maximum likelihood approach will overfit the data and always favors the more complex model. To avoid this issue, we will add a penalization term that should compensate for the overfit due to the model complexity.

More precisely, we define a model $S_{\mathcal{P},K,\mathcal{G}}$ by its number of classes K, a recursive dyadic partition $\mathcal{P}$ and a set $\mathcal{G}$ for the K-uples $(\Phi_{\theta_1},\dots,\Phi_{\theta_K})$ (or equivalently a set $\Theta_{\mathcal{G}}$ for $\theta = (\theta_1,\dots,\theta_K)$):

$$S_{\mathcal{P},K,\mathcal{G}} = \min\{s_{\mathcal{P},K,\theta,\pi}(\cdot|x), |(\Phi_{\theta_1},\dots,\Phi_{\theta_K}) \in \mathcal{G}, \forall\mathcal{R}_l \in \mathcal{P}, \pi[\mathcal{R}_l] \in S_{K-1}\}$$

where S_{K-1} is the $K-1$ dimensional simplex. The space $\mathcal{G}$ is chosen among the classical Gaussian K-uples described in Biernacki *et al.* [11], that is some set:

$$\mathcal{G}_{[\cdot]^K} = \left\{\left(\Phi_{\theta_1},\dots,\Phi_{\theta_K}\right)|\theta = (\theta_1,\dots,\theta_K) \in \Theta_{[\cdot]^K}\right\}$$

obtained by imposing some (mild) constraint on the means μ_k (basically that they belong to a compact set) and some (strong) constraints on the covariance matrices Σ_k. The assumptions on the covariance range from the weak assumption that the eigenvalues of the covariance matrix are within a subset $[\lambda_m, \lambda_M]$ with $\lambda_m > 0$ to the strong assumption that they are all spherical. They can further be chosen independently for all classes or assumed to share a structure; for instance, a common diagonalization basis or the same value. We refer to our technical report [18] for more details.

For a given model $S_{\mathcal{P},K,\mathcal{G}}$, we will use the maximum likelihood estimate:

$$\widehat{s}_{\mathcal{P},K,\mathcal{G}}(S_i|x_i) = \underset{s_{\mathcal{P},K,\mathcal{G}} \in S_{\mathcal{P},K,\mathcal{G}}}{\operatorname{argmin}}\left(\sum_{i=1}^{n} -\ln(s_{\mathcal{P},K,\mathcal{G}}(S_i|x_i))\right)$$

As explained below, the maximum likelihood value grows as the complexity of the models increases; in order

to select a reasonable model, we will add a penalty term $\mathrm{pen}(\mathcal{P}, K, \mathcal{G})$ that will counterbalance this effect and select the model $\widehat{\mathcal{P}, K, \mathcal{G}}$ that minimizes:

$$\left(\sum_{i=1}^{n} -\ln(\widehat{s}_{\mathcal{P},K,\mathcal{G}}(S_i|x_i))\right) + \mathrm{pen}(\mathcal{P}, K, \mathcal{G})$$

Choosing the penalty $\mathrm{pen}(\mathcal{P}, K, \mathcal{G})$ appropriately is obviously of crucial importance. A key result of our theoretical analysis, recalled in Appendix, is that the choice:

$$\mathrm{pen}(\mathcal{P}, K, \mathcal{G}) = \widetilde{k}_1 \dim\left(S_{\mathcal{P},K,\mathcal{G}}\right) + \widetilde{k}_2 ||\mathcal{P}||$$

(where $||\mathcal{P}||$ is the number of regions in the partition) is a good choice in terms of conditional density estimation. Although there is no theoretical guarantee that this is a good choice in term of unsupervised segmentation, we will nevertheless use this penalty as if our task was conditional density estimation. As:

$$\dim(S_{\mathcal{P},K,\mathcal{G}}) = ||\mathcal{P}||(K-1) + \dim(\mathcal{G})$$

this penalty can be rewritten as:

$$\mathrm{pen}(\mathcal{P}, K, \mathcal{G}) = \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\widetilde{k}_1(K-1) + \widetilde{k}_2\right) + \widetilde{k}_1 \dim(\mathcal{G})$$

which has an additive structure which is a key property to derive the efficient estimation algorithm of the next section.

2 AN EFFICIENT SEGMENTATION ALGORITHM

As described above, our procedure is based on two successive minimizations: one should first find the maximum likelihood estimate $\widehat{s}_{\mathcal{P},K,\mathcal{G}}$ for every partition $\mathcal{P}$, every number K of classes and every set of Gaussian K-uples within the collection and then only minimize the penalized criterion involving those maximum likelihood estimates. This is indeed the strategy used for the classical GMM, for which no partition is used, and thus the number of models remains $O(K_{\max})$. Such an exhaustive strategy becomes impossible when one optimizes the partition, as the number of partitions grows exponentially fast with n. To overcome this issue, we propose a minimization algorithm that simultaneously computes the best partition and the corresponding likelihood estimate given a number K of classes and a set of Gaussian K-uples. Only the optimization of the number of classes and of the set of Gaussians used is performed by an exhaustive search.

Our goal can be rewritten as the search for the minimizer in , $\mathcal{P}$, $\theta \in \mathcal{G}$ and $(\pi[\mathcal{R}_l])_{\mathcal{R}_l \in \mathcal{P}} \in S_{K-1}$ of:

$$PL(\mathcal{P}, K, \mathcal{G}, \theta, \pi) = \sum_{i=1}^{n} \left(-\ln\left(\sum_{k=1}^{K} \pi_k \left[\mathcal{R}(x_i)\right] \Phi_{\theta_k}(S_i)\right)\right) + \mathrm{pen}(\mathcal{P}, K, \mathcal{G})$$

where the penalty can be written in the following way:

$$\mathrm{pen}(\mathcal{P}, K, \mathcal{G}) = \sum_{\mathcal{R}_l \in \mathcal{P}} \mathrm{pen}_{\mathrm{spa}}(K) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G})$$

Our main concern is the case:

$$\mathrm{pen}_{\mathrm{spa}}(K) = \widetilde{k}_1(K-1) + \widetilde{k}_2$$

and

$$\mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}) = \widetilde{k}_1 \dim(\mathcal{G})$$

Denoting with a slight abuse of notation:

$$s_{K,\theta,\pi}(S) = \sum_{k=1}^{K} \pi_k \, \Phi_{\theta_k}(S)$$

this problem can thus be rewritten as the search for the minimizer of:

$$PL(\mathcal{P}, K, \mathcal{G}, \theta, \pi) = \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\left(\sum_{i|x_i \in \mathcal{R}_l} -\ln\left(s_{K,\theta,\pi[\mathcal{R}_l]}(S)\right)\right) + \mathrm{pen}_{\mathrm{spa}}(K)\right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G})$$

For a fixed number of classes K and a given structure $\mathcal{G}$ for the Gaussian parameter K-uples, we perform this minimization with an iterative scheme, very similar to the classical EM algorithm, in which one alternately modifies θ, π and $\mathcal{P}$:

- **Initialization:** let $\widehat{\mathcal{P}}^{(0)} = \{[0,1]^2\}$ be the trivial partition, let $(\widehat{\theta}^{(0)}, \widehat{\pi}^{(0)})$ be the result of a classical EM initialization (for instance, with the best K-means strategy; Biernacki *et al.* [11].)
- **Optimization:** given $(\widehat{\mathcal{P}}^{(j)}, \widehat{\theta}^{(j)}, \widehat{\pi}^{(j)})$,
  1. Majorization (expectation) step: compute, $\forall i \in \{1, \ldots, n\}, \forall k \in \{0, \ldots, K\}$,

$$\widehat{\Pi}_k^{(j)}[i] = \frac{\widehat{\pi}_k^{(j)}[\widehat{\mathcal{R}}^{(j)}(x_i)]\Phi_{\widehat{\theta}_k^{(j)}}(S_i)}{\sum_{k'=1}^{K} \widehat{\pi}_{k'}^{(j)}[\widehat{\mathcal{R}}^{(j)}(x_i)]\,\Phi_{\widehat{\theta}_{k'}^{(j)}}(S_i)}$$

2. Minimization step in θ: using the technique used in classical EM, compute $\widehat{\theta}^{(j+1)}$ the minimizer in $\mathcal{G}$ of:

$$\sum_{i=1}^{n}\left(-\sum_{k=1}^{K}\widehat{\Pi}_k^{(j)}[i]\ln\Phi_{\theta_k}(S_i)\right)$$

3. Minimization step in π: compute for all square $\mathcal{R}_l$:

$$\widehat{\pi}_k^{(j+1)}[\mathcal{R}_l]=\frac{\sum_{i|x_i\in\mathcal{R}_l}\widehat{\Pi}_k^{(j)}[i]}{\sum_{i|x_i\in\mathcal{R}_l}1}$$

4. Per square cost computation: Compute for all square $\mathcal{R}_l$:

$$C^{(j+1)}[\mathcal{R}_l]=-\sum_{i|x_i\in\mathcal{R}_l}\ln\left(\sum_{k=1}^{K}\widehat{\pi}_k^{(j+1)}[\mathcal{R}_l]\Phi_{\widehat{\theta}_k^{(j+1)}}(S_i)\right)+\mathrm{pen}_{\mathrm{spa}}(K)$$

5. Minimization step in $\mathcal{P}$: using a fast dynamic programming strategy, compute $\widehat{\mathcal{P}}^{(j+1)}$ the minimizer over all partitions of: $\sum_{\mathcal{R}_l\in\mathcal{P}}C^{(j+1)}[\mathcal{R}_l]$

- **Stopping criterion:** stop when the decrease in the cost is smaller than a prescribed precision for two consecutive steps.

This algorithm is, as hinted in its description, an example of a Majorization-Minimization (MM) algorithm, as is the Expectancy Minimization (EM) algorithm. A detailed description can be found in Appendix.

As often with the EM algorithm, initialization has to be performed carefully. We initialize our algorithm with the result of a classical GMM mixture model, one with weights that do not depend on the position, and thus partition is reduced to the unit square. This first estimate is itself obtained by the classical EM algorithm, whose initialization is obtained by selecting the parameter set yielding the largest likelihood with a set of M runs of K-means initialized by a random data subset. We used $M=10$ and ran only 10 steps of K-means.

We stress the (lack of) theoretical convergence guarantee of this algorithm. Due to the complex structure of the objective function (mainly its non-convexity), we are only able to show that the algorithm converges to a local optimum. A stochastic variant (SEM) in which the expectation step is replaced by a random draw of the unobserved label $k[x]$ according to the the current posterior law $\mathrm{s}_{K,\theta^{(i)},\pi^{(i)},\mathcal{P}^{(i)}[x]}(k[x]|S[x])$ could be used to remove this issue at a price of a slower convergence speed.

An important practical issue is the choice of the parameters $\widetilde{k}_1$ and $\widetilde{k}_2$ in the penalty:

$$\begin{aligned}\mathrm{pen}(\mathcal{P},K,\pi)&=\widetilde{k}_1\dim(S_{\mathcal{P},K,\pi})+\widetilde{k}_2||\mathcal{P}||\\&=\sum_{\mathcal{R}_l\in\mathcal{P}}(\widetilde{k}_1(K-1)+\widetilde{k}_2)+\widetilde{k}_1\dim(\mathcal{G})\end{aligned}$$

which corresponds to $\mathrm{pen}_{\mathrm{spa}}(K)=\widetilde{k}_1(K-1)+\widetilde{k}_2$ and $\mathrm{pen}_{\mathrm{par}}(K,\mathcal{G})=\widetilde{k}_1\dim(\mathcal{G})$. We propose to use here the slope heuristic introduced by Birgé and Massart [20] to calibrate these constants from the observed data, as described, for instance, by Baudry *et al.* [21] in a similar setting. For the sake of completeness, this heuristic is described in Appendix. Roughly speaking, the idea is that these parameters can be estimated from the behavior of the $-\log$-likelihood of the most complex models: it is expected that a good fit of its lower envelope can be obtained with the shape of the penalty proposed in the theorem and that using a penalty twice as larger as the estimated one yields a good selection. More precisely, we use the following procedure:

1. Compute $\widehat{\mathrm{s}}_{\mathcal{P},K,\mathcal{G}}$ and $\sum_{i=1}^{n}-\log\widehat{s}_{\mathcal{P},K,\mathcal{G}}(S_i[x_i])$ for a collection of complex models of various K, $\dim(\mathcal{G})$ and $\mathcal{P}$
2. Compute a lower envelope:

$$F(P,K,D)=\inf_{||\mathcal{P}||=P,K,\dim(\mathcal{G})=D}\sum_{i=1}^{n}-\log\widehat{s}_{\mathcal{P},K,\mathcal{G}}(S_i[x_i])$$

3. Robustly fit $F(P,K,D)$ by $\widetilde{k}_1'\,KP+\widetilde{k}_2'D+c$
4. Set $\widetilde{k}_1=2\widetilde{k}_1'$ and $\widetilde{k}_2=2\widetilde{k}_1'$

Our implementation of the algorithm is thus parameterless; the only choice left to the operator is in the definition of the collection of complex models used. Following Baudry *et al.* [21], we can increase the robustness of the selection using a stability principle: we compute the penalty and the selected model for the collection of the p most complex model for various p and choose the penalty yielding the most selected model complexity.

3 APPLICATIONS TO SPECTRAL IMAGES

To test the proposed algorithm, and more specifically the usefulness of the spatial information in the segmentation/classification process, we used it on experimental data measured in the context of a study of coating processes in lutherie [2]. The sample is observed using Fourier Transform Infrared microscopy (FTIR), producing a full infrared spectrum for each pixel of the image, aiming at the chemical characterization of the sample, and in particular of the coating layer(s).

3.1 Sample

The studied sample is a thin cross-section of maple wood with a single layer of hide glue on top of it, prepared recently using materials and processes from the *Cité de la Musique*, using materials of the same type and quality that are used for lutherie. This sample is to serve as

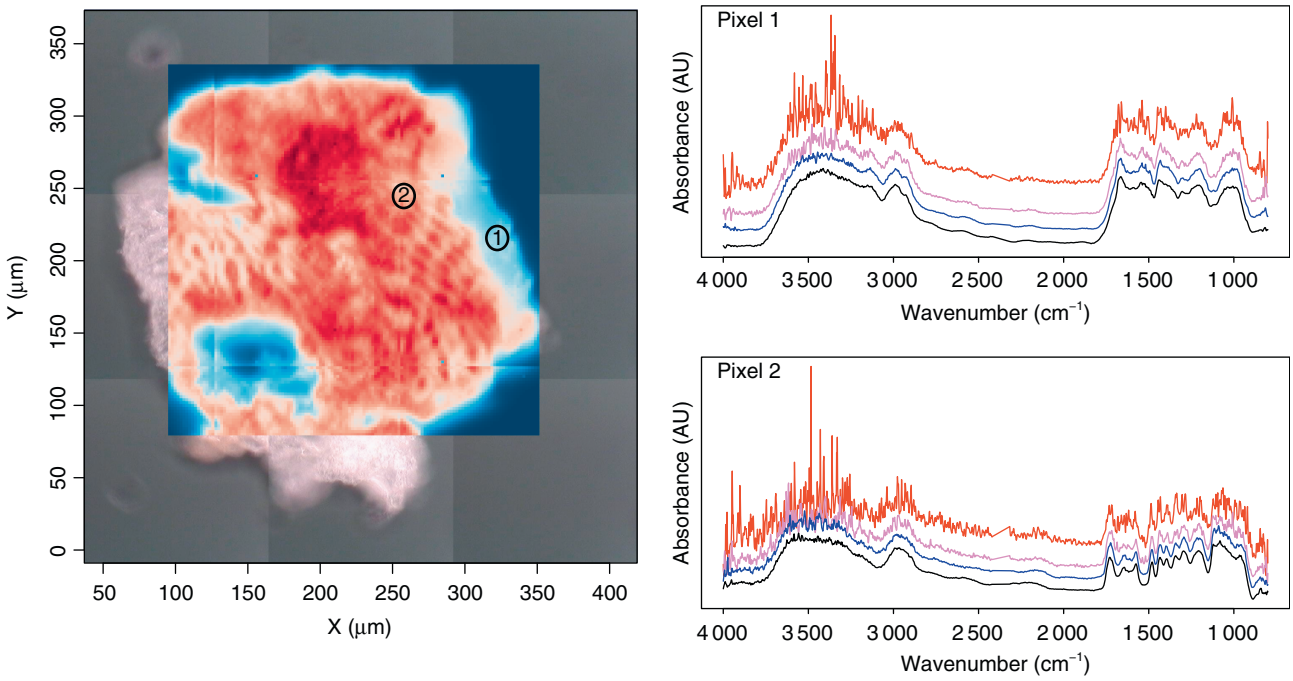

Figure 1

General view of the imaged sample. The background image is the visible color image of the section, surrounded by the MirrIR microscopy blade. The central region corresponds to the cropped image used in the spectral image analysis; the image corresponds to the total energy of the infrared absorbance spectra of each pixel in false colors, from blue for low values to red for hight values. Note that the tiling is different between the two modes of acquisition, visible light *vs* infrared. On the right, spectra obtained for two pixels (hide glue on the top, wood on the bottom) corresponding to 1, 8, 64 and 256 scan in red, violet, blue and black, respectively.

reference material to study the spectral variation of the hide glue at the various steps of the process. Infrared spectra were measured as a way to provide chemical characterization of the sample.

After application of the hide glue to a small maple wood piece, a sample was cut out of the piece and thin sections were produced using a Leica ultra-microtome fitted with a diamond knife from Diatome to produce a thin section of about 4 to 6 µm thickness. The section was then deposited on a MirrIR microscopy blade coated with gold, providing reflection in the infrared domain, resulting in the doubling of the optical path of the absorption, reaching approximately 10 µm. The spectral image was obtained using a Bruker Hyperion 3000 microscope fitted with a 20× Schwarzschild objective with a numerical aperture of 0.6, and using a focal plane array sensor made of 64 × 64 pixels with projected pixel size of 2 µm. Acquisition was carried out through a 3 × 3 tiling and the result cropped to a 128 × 128 pixel region encompassing most of the section (*Fig. 1*). Spectra were measured from 4 000 cm^{-1} to 800 cm^{-1} with a resolution of 8 cm^{-1} (1 577 samples, one every 2 cm^{-1}). To test the robustness of the proposed algorithm 4 images were collected, from a low to high signal-to-noise ratio: 1, 8 and 64 scan to serve as input for the algorithm and 256 scan to serve as ground truth. For these images the data acquisition required an measurement time of, respectively, 2, 5, 30 mn and 2 h.

The range of wavenumbers corresponding to atmospheric variation of carbon dioxide was removed from the spectra, hence removing all samples for which the wavenumber is in the range from 2 318 cm^{-1} to 2 418 cm^{-1}, hence practically reducing the dimension of the spectra from 1 577 to 1 528. Random projections were used to reduce the problem dimensions from 1 528 to 24 (*Fig. 2*). Each projection was generated using a centered uniform random number generator, then projected onto the sub-space orthogonal to the previously generated projections, and finally scaled to obtained a unitary L2 norm, hence producing an orthonormal basis of the random 24-dimensional subspace. Using a non-adaptative method for dimension reduction enabled us to use the exact same basis for all 4 datasets without privileging any particular dataset (*e.g.* the one that would have been used to optimize the basis) nor having a projection basis dependent on data

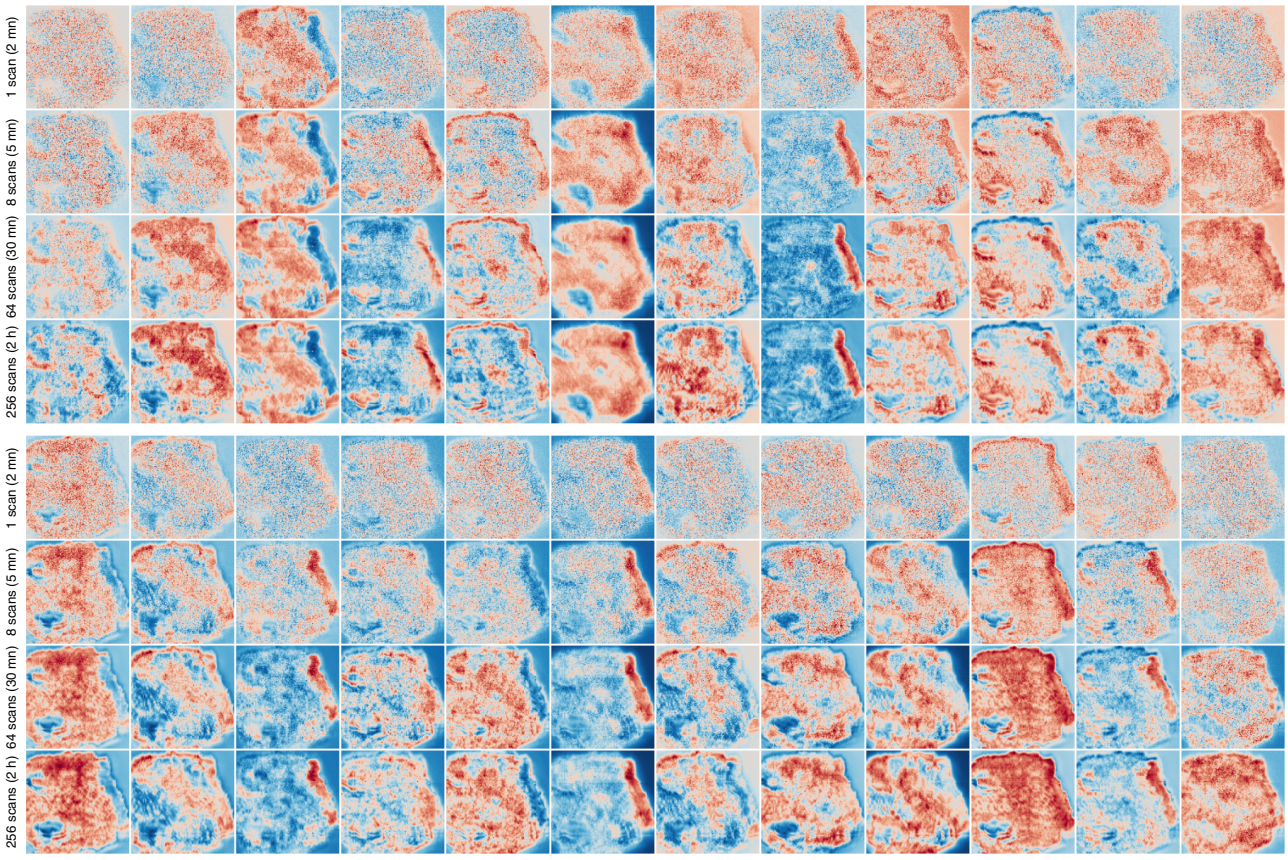

Figure 2

Projection on the 24-dimensional random orthonormal basis from the «fastest» image (top rows) to the reference image (bottom rows). To enhance the contrast, false colors from blue for low values to red for high values were used. No clear morphological differences are seen between the two extreme images, 1 *vs* 256 scans, but the effect of the reduced signal-to-noise ratio is clearly visible as a strong contrast degradation on the 1-scan projections.

from a different dataset which would have been the case if jointly adapting for all datasets at once. Within this constraint the random-projection method is advantageous as it provides an quasi isometry.

3.2 Statistical Analysis

The set of spectra of each image was submitted to both regular GMM, using the EM algorithm, as well as the spatially-aware model proposed herein. In both cases, we considered the following GMM/cGMM types:

- $p_k L_k C$ proportions and volumes of the Gaussians are not identical, but they share the same co-variance matrix;
- $p_k L_k D' A_k D$ proportions and volumes of the Gaussians are not identical; co-variance matrices are non-identical but have a joint diagonalization basis;

The number of classes was not set *a priori* but using the non-asymptotic model selection criteria described earlier: the minimization of – log-likelihood plus a penalty term proportional to the number of degrees of freedom of the model. In the regular GMM case the penalty proportionality constant was set using the maximal dimension jump slope heuristic described in Birgé and Massart [20] and further explained in Baudry *et al.* [21]. In the cGMM model, the model complexity constant and the segmentation cardinal constant were set simultaneously using the same approach but estimating the asymptotic bilinear form.

As expected, the critical slope for the number of components of the mixture increased at higher signal-to-noise ratio: 4.10, 8.60, and 12.39, respectively, for the 1-, 8- and 64-scans datasets for the GMM modelization, without spatial co-variables. The corresponding critical slope for the cGMM models are, respectively,

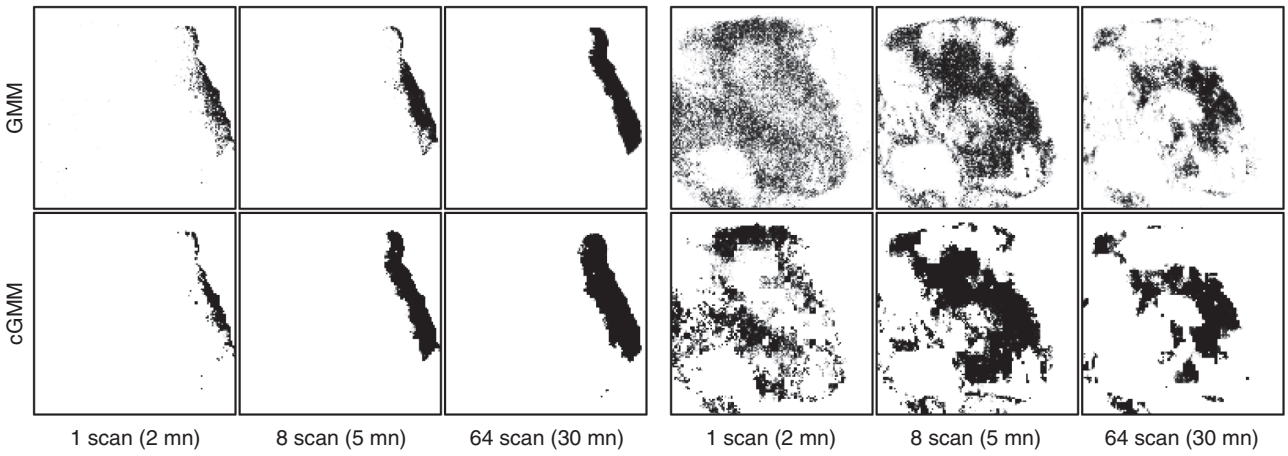

Figure 3

Map of the values of the *a posteriori* probability for a given component of the mixture for each dataset and comparing GMM and cGMM results. Maps for mixture components corresponding to pixels 1 and 2 of Figure 1 are, respectively, represented on the left and right.

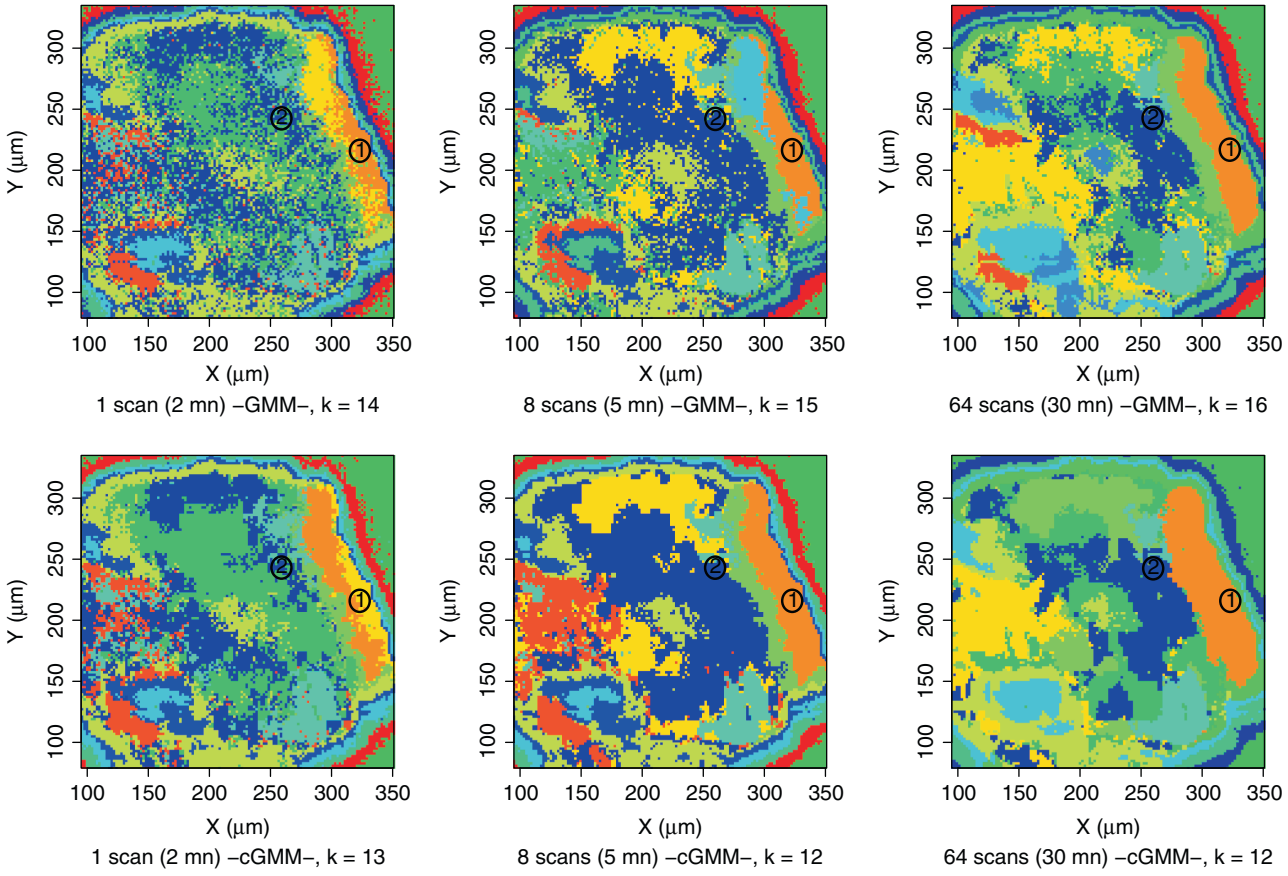

Figure 4

Result of unsupervised classification through Gaussian Mixture (conditional) density estimation and penalized likelihood model selection. Results from both the GMM and cGMM are shown, respectively, on the top and bottom rows. From left to right, an increasing level of signal-to-noise ratio; respectively, 1 scan/2 mn, 8 scan/5 mn and 64 scan/30 mn acquisitions. The optimal number of Gaussian components in each mixture ranges from 12 to 16, every time with a $p_k L_k D' A_k D$ model. To be able to visually compare the 6 images, we mapped the colors taking the 64-scan GMM case as a reference since it has the highest k, and mapping to minimize the distance between the means (μ_k) of the Gaussians from one model to the other.

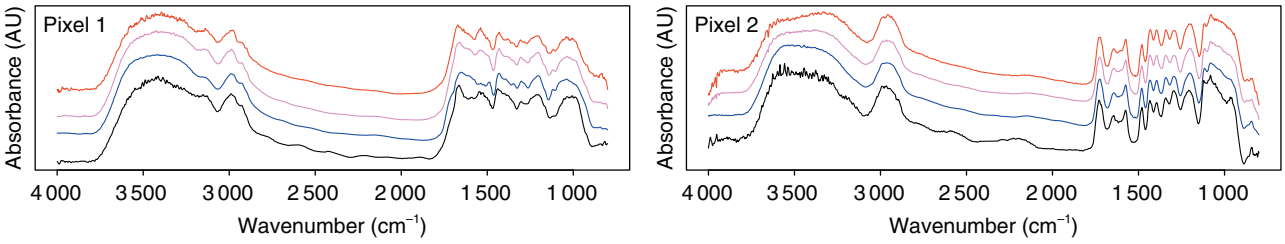

Figure 5

Spectra re-computed from the model parameters for the same two pixels as in Figure 1. Red, violet and blue correspond to reconstructed spectra, respectively, for the 1-, 8- and 64-scan datasets while the black line corresponds to the raw spectra collected in 256 scans.

6.89, 21.46 and 33.87. The number of classes is not strictly identical among the various datasets, but they are consistent: between 14 and 16 for the GMM models and between 12 and 13 for the cGMM models.

The cGMM model selection also involves a critical slope for the segmentation penalty which is consistent for the three datasets, ranging from 0.037 to 0.040; 0.0393, 0.0371 and 0.0402, respectively, for the 1-, 8- and 64-scans dataset. This is expected, since the segmentation penalty is related to the image morphology's complexity rather than to the signal-to-noise ratio of the measurements.

The result of the procedure is the estimation for each pixel of the posterior probability for it to correspond to each of the Gaussian components of the mixture. This can be visualized by plotting the "maps" for each component, displaying for each pixel the corresponding posterior probability, as in Figure 3. The effect of using the pixel's spatial coordinates as co-variables is the spatial regularization of the posterior probability maps leading to enhanced contrast compared with that obtained using the regular GMM.

Both GMM and cGMM results can also be shown in terms of classification of the pixels, as in Figure 4, where each pixel is assigned the color corresponding to the most likely component of the Gaussian mixture model. This other representation also evidences the spatial regularization effect of using the pixel's spatial coordinates as co-variables in the probability density estimation.

As for the GMM, the cGMM provides for each pixel the *posterior* probability that the spectra belong to each particular component. These probabilities can be used as weight in a weighted averaging of the measured spectra to obtain the mean spectra of each Gaussian in the full-dimensional space rather than in the 24-dimensional random space used to reduce the problem's complexity. Finally, weighted averaging can be performed for each pixel: averaging the Gaussian means using the *posterior* as weights, producing a "denoised" spectrum for each pixel. The noise reduction effect of the averaging is clear (comparing Fig. 1 and 5) in all three exposures, producing spectra of the same overall quality as the best measured spectra. Nevertheless, the produced results have to be taken qualitatively since the objective of the procedure is not to minimize quadratic error, hence the target is not to enhance the signal-to-noise ratio as such but rather to try to discriminate spectra as well as possible.

Comparing the obtained spectra with those measured using 256 scan/2 h acquisition, one sees that they are not missing a single of the small peak features which are all clearly discernible compared with the raw pixel spectra for the 1 scan/2 mn and 8 scan/5 mn datasets. The recomputed spectra from the three datasets (1-, 8- and 64-scan) are nearly identical among each other. When compared to the raw spectra obtained by a 256-scan/2 h acquisition, the exact same peaks are present in all spectra with the same positions in terms of local maxima; there are only slight variations in terms of the relative amplitudes of the various peaks. This latter difference only concerns the quantitative analysis of the chemical materials which would anyhow be better obtained using a regression method on raw data once the base constituents are identified. This identification is performed through the qualitative analysis we mentioned earlier, based on the presence and position of peaks and not relative amplitudes.

REFERENCES

1 Bertrand L., Languille M.-A., Cohen S.X., Robinet L., Gervais C., Leroy S., Bernard D., Le Pennec E., Josse W., Doucet J., Schöder S. (2011) European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials, *J. Synchrotron Radiat.* **18**, 5, 765-772. doi:10.1107/S090904951102334X.

2 Echard J.-P., Bertrand L., von Bohlen A., Le Hô A.-S., Paris C., Bellot-Gurlet L., Soulier B., Lattuati-Derieux A., Thao S., Robinet L., Lavédrine B., Vaiedelich S. (2010) The nature of the extraordinary finish of Stradivari's instruments, *Angew. Chem. Int. Ed.* **49**, 1, 197-201, ISSN 1521-3773.

3 Gueriau P., Mocuta C., Dutheil D.B., Cohen S.X., Thiaudière D., the OT1 consortium, Charbonnier S., Clément G., Bertrand L. (2014) Trace elemental imaging of rare earth elements discriminates tissues at microscale in flat fossils, *PLOS ONE* (accepted).

4 Tarabalka Y., Chanussot J., Benediktsson J. (2010) Segmentation and classification of hyperspectral data using watershed transformation, *Pattern Recognition* **43**, 7, 2367-2379.

5 Bunte M., Thompson D., Castano R., Chien S., Greeley R. (2011) Metric learning for hyperspectral image segmentation, IEEE WHISPERS, Lisbon, Portugal, 6–9 June.

6 Acito N., Corsini G., Diani M. (2003) An unsupervised algorithm for hyperspectral image segmentation based on the Gaussian mixture model, *Proc. IGARSS*, **6**, pages 3745-3747.

7 Yang J.-M., Yu P.-T., Kuo B.-C. (2010) A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data, *IEEE Trans. Geosci. Remote Sens.* **48**, 3, 1279-1293.

8 Tarabalka Y. (2010), Classification of Hyperspectral Data Using Spectral-Spatial Approaches. PhD thesis, Grenoble INP. Chapter 1.

9 Farag A., Mohamed R., El-Baz A. (2005) A unified framework for map estimation in remote sensing image segmentation, *IEEE Trans. Geosci. Remote Sens.* **43**, 7, 1617-1634.

10 Tarabalka Y., Benediktsson J.A., Chanussot J., Tilton J.C. (2010) Multiple spectral-spatial classification approach for hyperspectral data, *IEEE Trans. Geosci. Remote Sens.* **48**, 11, 4122-4132.

11 Biernacki Ch., Celeux G., Govaert G., Langrognet F. (2006) Model-based cluster and discriminant analysis with the MIXMOD software, *Comput. Statist. Data Anal.* **51**, 2, 587-600.

12 Maugis C., Michel B. (2012) A non asymptotic penalized criterion for Gaussian mixture model selection, *ESAIM Probab. Stat.* **15**, 41-68.

13 Kolaczyk E., Ju J., Gopal S. (2005) Multiscale, multigranular statistical image segmentation, *J. Amer. Statist. Assoc.* **100**, 472, 1358-1369.

14 Antoniadis A., Bigot J., von Sachs R. (2008) A multiscale approach for statistical characterization of functional images, *J. Comput. Graph. Statist.* **18**, 1, 216-237.

15 Blekas K., Likas A., Galatsanos N.P., Lagaris I.E. (2005) A spatially constrained mixture model for image segmentation, *IEEE Trans. Neural Netw.* **16**, 2, 494-498.

16 Nikou C., Likas A., Galatsanos N.P. (2010) A bayesian framework for image segmentation with spatially varying mixtures. *IEEE Transactions on Image Processing* **19**, 9, 2278-89.

17 Cohen S.X., Le Pennec E. (2012) Partition-based conditional density estimation, *ESAIM Probab. Stat.* doi:10.1051/ps/2012017.

18 Cohen S.X., Le Pennec E. (2011) *Conditional density estimation by penalized likelihood model selection and applications*, Technical report, INRIA.

19 Donoho D. (1997) CART and best-ortho-basis: a connection, *Ann. Statist.* **25**, 5, 1870-1911.

20 Birgé L., Massart P. (2007) Minimal penalties for Gaussian model selection, *Probability theory and related fields* **138**, 1-2, 33-73.

21 Baudry J.-P., Maugis C., Michel B. (2012) Slope heuristics: overview and implementation, *Stat. Comput.* **22**, 455-470.

22 Massart P. (2007). Concentration inequalities and model selection, volume **1896** of *Lecture Notes in Mathematics*, Springer, Berlin, Lectures from the *33rd Summer School on Probability Theory held in Saint-Flour*, 6-23 July, 2003, With a foreword by Jean Picard.

23 Birgé L., Massart P. (1998) Minimum contrast estimators on sieves: exponential bounds and rates of convergence, *Bernoulli* **4**, 3, 329-375.

24 van de Geer S. (1995) The method of sieves and minimum contrast estimators, *Math. Methods Statist.* **4**, 20-38.

25 Huang Y., Pollak I., Do M., Bouman C. (2006) Fast search for best representations in multitree dictionaries, *IEEE Trans. Image Process.* **15**, 7, 1779-1793.

Manuscript accepted in February 2014
Published online in March 2014

Copyright © 2014 IFP Energies nouvelles

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than IFP Energies nouvelles must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee: request permission from Information Mission, IFP Energies nouvelles, revueogst@ifpen.fr.

Appendix A : Conditional Density Estimation by Model Selection

In [17] and the corresponding extended technical report [18], we showed that the penalty choice proposed here is a good choice in terms of conditional density estimation. For the sake of completeness, we summarize here the implication of these results for spatialized Gaussian mixture models.

We should first specify our goodness criterion. The most natural quality measure in a maximum likelihood approach is the the Kullback-Leibler divergence, KL. All the conditional densities appearing here are defined with respect to the Lebesgue measure. We can thus write, with a slight abuse of notation:

$$KL(s,t) = KL(sd\lambda, td\lambda) = \begin{cases} \int_\Omega \frac{s(y)}{t(y)} \ln \frac{s(y)}{t(y)} t\mathrm{d}(y) & \text{if } s\mathrm{d}\lambda \ll t\mathrm{d}\lambda \Longleftrightarrow \forall \omega \in \Omega, s(\omega) = 0 \Rightarrow t(\omega) = 0 \\ +\infty & \text{otherwise} \end{cases}$$

This divergence is an intrinsic quality measure; it does not depend on the choice of the reference measure but only on the probability laws. This divergence should be further adapted to the conditional density setting. We are thus led to the following natural tensorized divergence:

$$KL_\lambda^{\otimes_n}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[KL_\lambda(s(\cdot|x_i), t(\cdot|x_i))]$$

Unfortunately, we will not be able to control this divergence but only a slightly smaller one. More precisely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0,1)$ defined by:

$$JKL_\rho(sd\lambda, td\lambda) = JKL_{\rho,\lambda}(s,t) = \frac{1}{\rho} KL_\lambda(s, (1-\rho)s + \rho t)$$

already used by Massart [22], Birgé and Massart [23] and van de Geer [24]. This divergence is smaller than the Kullback-Leibler one but larger than the squared Hellinger one, denoted $d_\lambda^2(s,t)$. We define their tensorized counterpart:

$$d_\lambda^{2\otimes_n}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[d_\lambda^2(s(\cdot|x_i), t(\cdot|x_i))\right]$$

and

$$JKL_{\rho,\lambda}^{\otimes_n}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[JKL_{\rho,\lambda}(s(\cdot|x_i), t(\cdot|x_i))\right]$$

In [18], we show precisely that:

Theorem 1. *Assume we observe* $(x_i, \mathrm{S_i})$ *with unknown conditional density* s_0. *Let* $\widehat{s}_m$ *be a* η*-log-likelihood minimizer in* S_m:

$$\sum_{i=1}^{n} -\ln(\widehat{s}_m(\mathrm{S}_i|x_i)) \leq \inf_{s_m \in S_m} \left(\sum_{i=1}^{n} -\ln(s_m(S_i|x_i)) \right) + \eta$$

For any $\rho \in (0,1)$ *and for any* $C_1 > 1$, *there exist a* $C_* > \pi$ *and a* $C_* > 0$, *such that the penalized estimator* $\widehat{\mathrm{s}}_{\widehat{\mathrm{m}}}$ *with* $\widehat{\mathrm{m}}$ *such that*:

$$\sum_{i=1}^{n} -\ln(\widehat{s}_{\widehat{m}}(S_i|x_i)) + \mathrm{pen}(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \left(\sum_{i=1}^{n} -\ln\,(\widehat{s}_{\widehat{m}}(S_i|x_i)) + \mathrm{pen}(m) \right) + \eta'$$

satisfies:

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{P,K,\mathcal{G}}})\right] \leq C_1 \left(\inf_{(P,K,\mathcal{G}) \in \mathcal{M}} \left(\inf_{s_{P,K,\mathcal{G}} \in S_{P,K,\mathcal{G}}} KL^{\otimes_n}(s_0, s_{P,K,\mathcal{G}}) + \frac{\mathrm{pen}(P,K,\mathcal{G})}{n} \right) + \frac{\mathcal{K}_0}{n} + \frac{\eta + \eta'}{n} \right)$$

as soon as:

$$\mathrm{pen}(P,K,\mathcal{G}) \geq \widetilde{k}_1 \ \dim\left(s_{P,K,\mathcal{G}}\right) + \widetilde{k}_2 ||\mathrm{P}||$$

with:

$$\widetilde{k}_1 \geq k\left(2C_* + \mathrm{c}_* + 1 + \left(\ln\frac{n}{eC_*}\right)_+\right)$$

and

$$\widetilde{k}_2 \geq C_* k \ln 2$$

with $k > k_0$, *where* k_0 *is a constant that depends only on* ρ *and* C_1.

The variance of the maximum likelihood in each model is asymptotically of the order $2\dim(s_{P,K,\mathcal{G}})$ and a similar bound also holds non-asymptotically [18]. The variance is thus up to a factor that may grow logarithmically with n, of the order $(s_{P,K,\mathcal{G}})$. This implies that, again up to a factor that may grow logarithmically with n, the risk of the penalized estimator is bounded by the best possible risk among the collection of models. This specific choice of penalty is thus a good choice for conditional density estimation. Although this does not imply a good classification property, this is sufficient to obtain the consistency of the number of classes and the parameters when the true conditional density is indeed a spatialized Gaussian mixture.

Appendix B: Detailed Description of the Optimization Algorithm

It is based on the construction of majorizations of PL which coincide at the current estimate and are easier to minimize. The remaining part of this section is devoted to the mathematical justification of this algorithm. To construct the majorization, we extend at each pixel the observation of the spectrum S to the observation of the couple (S,k) with $k \in \{1,\ldots,K\}$. With a slight abuse of notation, we denote:

$$s_{K,\theta,\pi}(S,k) = \pi_k \Phi_{\theta_k}(S)$$

the joint density with respect to the tensor product of the Lebesgue measure and the counting measure. This corresponds indeed to the way we assign each sample to its class through our MAP principle as:

$$\mathrm{s}_{K,\theta,\pi}(k|S) = \frac{\pi_k \Phi_{\theta_k}(S)}{\sum_{k'=1}^{K} \pi_{k'} \Phi_{\theta'_k}(S)}$$

Using this notation, the weights computed in the majorization step can be rewritten as:

$$\widehat{\Pi}_k^{(j)}[i] = \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}[\mathcal{R}_l]}(k|S_i)$$

The key property is the following majorization property:

Lemma 1. *Let* $(\widehat{\mathcal{P}}^{(j)}, \widehat{\theta}^{(j)}, \widehat{\pi}^{(j)})$ *be a current estimate,* $\forall \mathcal{P}, \theta \in \mathcal{G}, \pi \epsilon\ S_{k-1}^{||\mathcal{P}||}$

$$\begin{aligned} PL(\mathcal{P},K,\mathcal{G},\theta,\pi) \leq & \sum_{\mathcal{R}l\epsilon\mathcal{P}} \left(\left(\sum_{i|x_i\in\mathcal{R}l} - \sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i]\pi_k[\mathcal{R}_l]\right) + \mathrm{pen}_{\mathrm{spa}}(K)\right) \\ & + \sum_{i=1}^{n}\left(-\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \Phi_{\theta_k}(S_i)\right) + \mathrm{pen}_{\mathrm{par}}(K,\mathcal{G}) \\ & + \sum_{i=1}^{n}\left(\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \widehat{\Pi}_k^{(j)}[i]\right) \end{aligned}$$

with equality when $(\mathcal{P}, \pi, \theta) = (\widehat{\mathcal{P}}^{(j)}, \widehat{\pi}^{(j)}, \widehat{\theta}^{(j)})$.
proved, using a conditioning with respect to k.

The minimization step in θ corresponds exactly to the minimization of the right-hand side, a minimization that reduces to the minimization of:

$$\sum_{\mathrm{i}=1}^{n}\left(-\sum_{k=1}^{K}\widehat{\Pi}_k^{(j)}[i]\ln\Phi_{\theta_k}(S_i)\right)$$

This minimization has the exact same structure as the corresponding one in the classical Gaussian Mixture Model (GMM) case. We can thus rely on the classical optimization technique described, for instance, by Biernacki *et al* [11] This efficiently provides a new estimate $\widehat{\theta}^{(j+1)}$ for the K-uples of Gaussian parameters within the prescribed set $\mathcal{G}$.

Minimizing in $\mathcal{P}$ and π the same right-hand side is equivalent to minimizing:

$$\sum_{\mathcal{R}_l\in\mathcal{P}}\left(\sum_{i|x_i\in\mathcal{R}_l}\left(-\sum_{k=1}^{K}\widehat{\Pi}_k^{(j)}[i]\ln\pi_k[\mathcal{R}_l]\right)+\mathrm{pen}_{\mathrm{spa}}(K)\right)$$

which have an additive structure with respect to the squares of the partition. Given a square $\mathcal{R}_l$, a simple computation shows that the minimum of:

$$\sum_{i|x_i\in\mathcal{R}_l}\left(-\sum_{k=1}^{K}\widehat{\Pi}_k^{(j)}[i]\ln\pi_k[\mathcal{R}_l]\right)+\mathrm{pen}_{\mathrm{spa}}(K)$$

is attained at:

$$\widehat{\pi}_k^{(j+1)}[\mathcal{R}_l]=\frac{\sum_{i|x_i\in\mathcal{R}_l}\widehat{\Pi}_k^{(j)}[i]}{\sum_{in|x_i\in\mathcal{R}_l}1}$$

Let $C^{(j+1/2)}[\mathcal{R}_l]$ the value at this minimum:

$$C^{(j+1/2)}[\mathcal{R}_l]=\sum_{i|x_i\in\mathcal{R}_l}\left(-\sum_{k=1}^{K}\widehat{\Pi}_k^{(j)}[i]\ln\widehat{\pi}_k^{(j+1)}[\mathcal{R}_l]\right)+\ \mathrm{pen}_{\mathrm{spa}}(K)$$

the optimization in $\mathcal{P}$ becomes equivalent to the minimization of:

$$\sum_{\mathcal{R}_l\epsilon\mathcal{P}}C^{(\mathrm{j}+1)}[\mathcal{R}_l]$$

Capitalizing on the tree structure of the dyadic recursive partition, one can use the fast dynamic programming strategy of Donoho [19] and Huang *et al.* [25] described briefly in Appendix C, to obtain an optimal partition $\widehat{\mathcal{P}}^{(j+1/2)}$. Note that the algorithm could have been stopped here as, by construction:

$$PL(\widehat{\mathcal{P}}^{(j)},K,\mathcal{G},\theta^{(j)},\widehat{\pi}^{(j)})\geq PL(\widehat{\mathcal{P}}^{(j+1/2)},K,\mathcal{G},\theta^{(j+1)},\widehat{\pi}^{(j+1)})$$

A slight modification of the cost function, the one used in the description of the algorithm, yields a better partition choice. Indeed:

$$\begin{aligned}PL\left(\widehat{\mathcal{P}}^{\left(j+\frac{1}{2}\right)},K,\mathcal{G},\theta^{(j+1)},\widehat{\pi}^{(j+1)}\right)&=\sum_{\mathcal{R}_l\in\widehat{\mathcal{P}}^{\left(j+\frac{1}{2}\right)}}\left(\sum_{i|x_i\in\mathcal{R}_l}\left(-ln\sum_{k=1}^{K}\widehat{\pi}_k^{(j+1)}[\mathcal{R}_l]\Phi_{\widehat{\theta}_k^{(j+1)}}(S_i)\right)+pen_{spa}(K)\right)+\ \mathrm{pen}_{\mathrm{par}}+(\mathrm{K},\mathcal{G})\\&=\sum_{\mathcal{R}_l\in\widehat{\mathcal{P}}(j+1/2)}C^{(\mathrm{j}+1)}[\mathcal{R}_l]+\mathrm{pen}_{\mathrm{par}}(K,\mathcal{G})\end{aligned}$$

so that the optimizer $\widehat{\mathcal{P}}^{(j+1)}$ of:

$$\sum_{\mathcal{R}_l \in \mathcal{P}} C^{(j+1)}[\mathcal{R}_l]$$

the one proposed in the algorithm, can be obtained with the same dynamic programming algorithm and is such that:

$$PL(\widehat{\mathcal{P}}^{(j)}, K, \mathcal{G}, \widehat{\theta}^{(j)}, \widehat{\pi}^{(j)}) \geq PL(\widehat{\mathcal{P}}^{(j+1)}, K, \mathcal{G}, \widehat{\theta}^{(j+1)}, \widehat{\pi}^{(j+1)}) \geq PL(\widetilde{\mathcal{P}}^{(j+1)}, K, \mathcal{G}, \widehat{\theta}^{(j+1)}, \widehat{\pi}^{(j+1)})$$

Proof of Lemma 1. We will be slightly more general in the proof and assume that the per rectangle penalty may depends on $\mathcal{R}$ and π while the other may depends on θ :

$$\mathrm{pen}(\mathcal{P}, K, \mathcal{G}, \theta, \pi) = \sum_{\mathcal{R}_l \in \mathcal{P}} \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l]) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta)$$

For any probability (qk) on the classes k:

$$\begin{aligned}
\ln S_{k,\theta,\pi}(S) &= \sum_{k=1}^{K} q_k \ln S_{\mathrm{k},\theta,\pi}(S) \\
&= \sum_{k=1}^{K} q_k \ln \left(\frac{S_{k,\theta,\pi}(k,S) q_k}{S_{k,\theta,\pi}(k|S) q_k}\right) \\
&= -\sum_{k=1}^{K} q_k \ln \left(\frac{q_k}{S_{k,\theta,\pi}(k,S)}\right) + \sum_{k=1}^{K} q_k \ln \left(\frac{q_k}{S_{k,\theta,\pi}(K|S)}\right) \\
&= -KL\big(q, s_{K,\theta,\pi}(.|S)\big) + KL\big(q, s_{K,\theta,\pi}(.|S)\big)
\end{aligned}$$

Assume we have a "current" estimate $(\widehat{\pi}^{(j)}, \widehat{\theta}^{(j)})$ and let:

$$q_k = \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}}(k|S) = \frac{\widehat{\pi}_k^{(j)} \Phi_{\widehat{\theta}_k^{(j)}}(S)}{\sum_{k'=1}^{K} \widehat{\pi}_{k'}^{(j)} \Phi_{\widehat{\theta}_{k'}^{(j)}}(S)}$$

we obtain a surrogate function of –ln $\mathrm{s}_{K,\theta,\pi}$ with the help of the previous formula:

$$\begin{aligned}
-\ln \mathrm{s}_{K,\theta,\pi}(S) &= KL\left(\mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}}(\cdot|S), \mathrm{s}_{k,\theta,\pi}(\cdot|S)\right) - KL\left(\mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}}(\cdot|S), \mathrm{s}_{K,\theta,\pi}(\cdot|S)\right) \\
&\leq KL\left(\mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}}(\cdot|S), \mathrm{s}_{K,\theta,\pi}(\cdot|S)\right)
\end{aligned}$$

with equality when $(\pi, \theta = \left(\widehat{\pi}^{(j)}, \widehat{\theta}^{(j)}\right)$
This idea can be used pixelwise and thus starting with a current estimate $(\widehat{\mathcal{P}}^{(j)}, \widehat{\theta}^{(j)}, \widehat{\pi}^{(j)})$, one obtains:

$$\sum_{\mathcal{R} \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} -\ln s_{k,\theta,\pi[\mathcal{R}_l]} \right) \leq \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} KL\big(s_{K,\widehat{\theta}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right],\widehat{\theta}^{(j)}}(\cdot|S_i), s_{K,\pi[\mathcal{R}_l],\theta}(\cdot, S_i)\big) \right)$$

with equality when $(\mathcal{P}, \pi, \theta) = (\widehat{\mathcal{P}}^{(j)}, \widehat{\pi}^{(j)}, \widehat{\theta}^{(j)})$. Adding the penalties yields:

$$PL(\mathcal{P}, K, \mathcal{G}, \theta, \pi) \leq \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} KL\big(\mathrm{s}_{K,\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right],\widehat{\theta}^{(j)}}(\cdot, S_i), s_{K,\pi[\mathcal{R}_l],\theta}\big(\cdot|S_i\big) \ \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l]) \right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta)$$

still with equality when $(\mathcal{P}, \pi, \theta) = (\widehat{\mathcal{P}}^{(j)}, \widehat{\pi}^{(j)}, \widehat{\theta}^{(j)})$. This right-hand side can be rewritten:

$$\begin{aligned}
&\sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} KL\left(\mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right]}\left(\cdot|S_i\right), s_{K,\theta,\pi[\mathcal{R}_l]}(.,S_i)\right) + \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l] \right) \right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta) \\
&= \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} \left(-\sum_{k=1}^{K} \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right],}\left(k|S_i\right) \ln \mathrm{s}_{K,\theta,\pi[\mathcal{R}_l]}(k,S_i) \right) + \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l]) \right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta) \\
&\quad + \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} \sum_{k=1}^{K} \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right],}\left(k|S_i\right) \ln \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right]}\left(k|S_i\right) \right)
\end{aligned}$$

or using the notation $\widehat{\Pi}_k^{(j)}[i] = \mathrm{s}_{K,\widehat{\theta}^{(j)},\widehat{\pi}^{(j)}\left[\widehat{\mathcal{R}}_l^{(j)}(x_i)\right]}\left(k|S_i\right)$:

$$\begin{aligned}
&= \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} \left(-\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ \mathrm{ln}\ \mathrm{s}_{K,\theta,\pi[\mathcal{R}_l]}(k,S_i) + \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l]) \right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta) \right) \\
&\quad + \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \widehat{\Pi}_k^{(j)}[i] \right) \\
&= \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{i|x_i \in \mathcal{R}_l} \left(-\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \pi_k[\mathcal{R}_l] \right) + \mathrm{pen}_{\mathrm{spa}}(\mathcal{R}_l, K, \pi[\mathcal{R}_l]) \right) \\
&\quad + \sum_{i=1}^{n} \left(-\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \Phi_{\theta_k} \mathcal{S}_i \right) + \mathrm{pen}_{\mathrm{par}}(K, \mathcal{G}, \theta) + \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \widehat{\Pi}_k^{(j)}[i] \ln \widehat{\Pi}_k^{(j)}[i] \right)
\end{aligned}$$

Appendix C: Partition Optimization Algorithm

The fast linear programming strategy used to minimize over the set of dyadic partitions an additive cost:

$$\sum_{\mathcal{R}_l \in \mathcal{P}} C(\mathcal{R}_l)$$

capitalizes on the quadtree structure of those dyadic partitions. For any leaf $\mathcal{R}$, we denote $\mathcal{P}(\mathcal{R})$ a generic partition of $\mathcal{R}$ and $\widetilde{\mathcal{P}}(\mathcal{R})$ the one minimizing the local cost, $\sum_{\mathcal{R}_l \in p(\mathcal{R})} C(\mathcal{R}_l)$.

The key observation is that the best partition $\widetilde{\mathcal{P}}(\mathcal{R}_0)$ of a square $\mathcal{R}_0$ is either the whole square $\mathcal{R}_0$ or the union of the best partitions of its four subsquares $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ and $\mathcal{R}_4$. Furthermore, the decision is obtained by comparing the cost of these two possibilities. If we denote:

$$\widetilde{C}(\mathcal{R}) = \sum\nolimits_{\mathcal{R}_l \in \widetilde{\mathcal{P}}(\mathcal{R})} C(\mathcal{R}_l)$$

$$\widetilde{\mathcal{P}}(\mathcal{R}_0) = \begin{cases} \{\mathcal{R}_0\} & \text{if } C(\mathcal{R}_0) \leq \sum_{\mathrm{i}=1}^{4} \widetilde{C}(\mathcal{R}) \\ \cup_{i=1}^{4} \widetilde{\mathcal{P}}(\mathcal{R}_\mathrm{i}) & \text{otherwise} \end{cases}$$

This leads to a recursive algorithm as soon as one notes that there is a minimal size for the subsquares, for which the only possible partition is the trivial one, thus allowing the initialization of the recursion.