General pairwise Markov chains for unsupervised image segmentation

Hugo Gangloff, Katherine Morales, Yohan Petetin

GENERAL PAIRWISE MARKOV CHAINS FOR UNSUPERVISED IMAGE SEGMENTATION

Hugo Gangloff, Katherine Morales, Yohan Petetin

Samovar, Telecom Sudparis, Institut Polytechnique de Paris, 91011 Évry, France

ABSTRACT

Probabilistic graphical models are popular tools in statistical signal processing. The dependencies between the random variables described by such models enable to model a large class of statistical problems. Among probabilistic graphical models, Hidden Markov models and their extensions, Pairwise Markov models, are latent variable models which have found applications in image segmentation. In this paper, we address this problem by introducing a new class of Pairwise Markov models whose parametrization allows the use of (deep) neural networks architectures, for example. We focus on the unsupervised parameters estimation in these general models and we show that the combination of our general framework with (deep) neural architectures outperforms classical Pairwise Markov models for the task of unsupervised image segmentation.

Index Terms— Pairwise Markov Chains, Image Segmentation, Deep Neural Networks, Expectation-Maximization.

1. INTRODUCTION

Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ (resp. $\boldsymbol{H} = (H_1, \cdots, H_k)$) a sequence of observed (resp. hidden) random variables (r.v.), where $X_k \in \mathbb{R}$ and $H_k \in \Omega = \{\omega_1, \cdots, \omega_C\}$, for all k, $1 \leq k \leq K$. The joint distribution of $(\boldsymbol{H}, \boldsymbol{X})$ is denoted $p(\boldsymbol{h}, \boldsymbol{x})$. In this paper, we focus on the Bayesian estimation of H_k from $\boldsymbol{X}$, for all k, $1 \leq k \leq K$. Bayesian image segmentation is an critical application of this problem. In this context, the hidden r.v. H_k is associated to the class of a given pixel of an image, while the observed r.v. X_k represents the noisy observation of this pixel. Dealing with this application requires a relevant probabilistic model $p_{\boldsymbol{\theta}}(\boldsymbol{h}, \boldsymbol{x})$ in which it is possible to compute or to approximate the marginal smoothing distribution $p_{\boldsymbol{\theta}}(h_k|\boldsymbol{x})$, for all k, $1 \leq k \leq K$, when $\boldsymbol{\theta}$ is known.

Hidden Markov Chains (HMCs) define an important family of probabilistic graphical models and have been the subject of many investigations for the Bayesian image segmentation application [**?**]. Roughly speaking, in an HMC, $\boldsymbol{H}$ is Markovian and given $\boldsymbol{H}$, the observations $\boldsymbol{X}$ are independent and only depend on the equivalent hidden r.v. These models have been generalized by the introduction of the Pairwise Markov Chains (PMCs) [1] which relax the three assumptions above while keeping the interesting computational properties of HMCs. These models have also received a particular attention for image segmentation, see e.g. [2, 3, 4]

By contrast, artificial Deep Neural Networks (DNNs) do not model the observations with a probabilistic model but have gained in popularity due to their excellent performances in many tasks, such as classification. Indeed, a DNN can be seen as an universal approximator $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ of a function $f(\boldsymbol{x})$ [5]. The parameter $\boldsymbol{\theta}$ is estimated in a supervised way by the backpropagation algorithm which requires a labeled training dataset [6]. These architectures have been recently combined with probabilistic graphical models in order to provide powerful generative models [7, 8] which aim at modeling the unknown distribution $p(\boldsymbol{x})$ of these observations.

Let us turn now to the contributions of this paper. We first remark that unsupervised image segmentation via PMCs models [9, 2] has been done for a particular subclass of PMCs. These models satisfy a stationary assumption and rely on an implicit parametric Markovian transition distribution which may be difficult to tune. In this paper, we first relax these underlying assumptions and propose a general framework for image segmentation with PMCs. This general framework enables us to robustify PMCs by parametrizing the key distributions with universal approximators based on DNNs. Next, as opposed to deep generative models which particularly focus on the distribution $p(\boldsymbol{x})$ via continuous latent variables [10], we propose an efficient unsupervised parameter estimation procedure which aims at taking into account that H_k is associated to the class of pixel k and thus to reconstruct the image from noisy observations. As shown in simulations, this procedure illustrates the great potential of Deep PMCs for unsupervised image segmentation.

The rest of the paper is organized as follows. Section 2 reviews Markovian probabilistic models and introduces our general PMC models for image segmentation. In Section 3, we explain how to estimate the parameters of our models in an unsupervised fashion, and how to estimate the discrete latent process from the observations. Finally, Section 4 provides some simulations in which our models are compared with classical probabilistic models.

2. PROBABILISTIC MODELS

In this section, we first review the PMC framework and its application for image segmentation.

2.1. PMC models

As recalled in Introduction, a PMC is a probabilistic model where the pair (H_k, X_k) is Markovian. So, the joint distribution of $(\boldsymbol{H}, \boldsymbol{X})$ reads

$$p(\boldsymbol{h}, \boldsymbol{x}) = p(h_1, x_1) \prod_{k=2}^{K} p(h_k, x_k | h_{k-1}, x_{k-1}), \quad (1)$$

where the Markovian transition distribution can be factorized as

$$p(h_k, x_k | h_{k-1}, x_{k-1}) = p(h_k | h_{k-1}, x_{k-1}) \times p(x_k | h_k, h_{k-1}, x_{k-1}). \quad (2)$$

Factorization (2) is the most general in terms of direct dependencies that can be modeled by a PMC. From now on, we will refer to this model as PMC with Correlated Noise 2 (PMC-CN2) and its graphical representation is given in Fig. 1e.

From (2), several submodels can be derived. We list them by decreasing order of complexity. The PMC-CN1 (see Fig. 1d) is a model where we have

$$p(h_k, x_k | h_{k-1}, x_{k-1}) = p(h_k | h_{k-1}, x_{k-1}) p(x_k | h_k, x_{k-1}). \quad (3)$$

In the next three following models, $\boldsymbol{H}$ becomes Markovian, whence the terminology. The factorization associated to the HMC-CN model (see Fig. 1c) reads

$$p(h_k, x_k | h_{k-1}, x_{k-1}) = p(h_k | h_{k-1}) p(x_k | h_k, x_{k-1}). \quad (4)$$

In the HMC with Independent Noise 2 (HMC-IN2) (Fig. 1b) model we have

$$p(h_k, x_k | h_{k-1}, x_{k-1}) = p(h_k | h_{k-1}, x_{k-1}) p(x_k | h_k). \quad (5)$$

Note that this model coincides with the probabilistic Recurrent Neural Networks defined in [11]. Finally, the HMC-IN (Fig. 1a) coincides with the classical HMC,

$$p(h_k, x_k | h_{k-1}, x_{k-1}) = p(h_k | h_{k-1}) p(x_k | h_k). \quad (6)$$

2.2. PMC for image segmentation

In [1], the image segmentation issue has been treated for the particular class of stationnary PMC models. These models rely on the constraint that $p(h_{k-1}, x_{k-1}, h_k, x_k)$ do not depend on k and are thus characterized by the distributions

$$p(h_{k-1}, h_k) p(x_{k-1}, x_k | h_{k-1}, h_k) = p(h_{k-1}, x_{k-1}, h_k, x_k).$$

Next, the particular stationary Gaussian PMC has been investigated: $p(x_{k-1}, x_k | h_{k-1} = \omega_i, h_k = \omega_j)$ is a bi-variate Gaussian whose parameters depend on (ω_i, ω_j), for all $(\omega_i, \omega_j) \in \Omega^2$. While these models have achieved good results in practice, the stationary assumption tends to restrict the modeling power of the general PMC (2). First, it constraints the parameters of the model since $p(h_{k-1} = \omega_i)$ has to be equal to $p(h_k = \omega_i)$, for all k, for example; next, if we take the bi-variate Gaussian case, it necessarily implies that the distribution of X_k given x_{k-1}, h_{k-1} and h_k is a Gaussian whose mean is a linear function of x_{k-1}. Finally, the estimation of the parameters of stationary PMCs has to be approximated, even in simple cases [1].

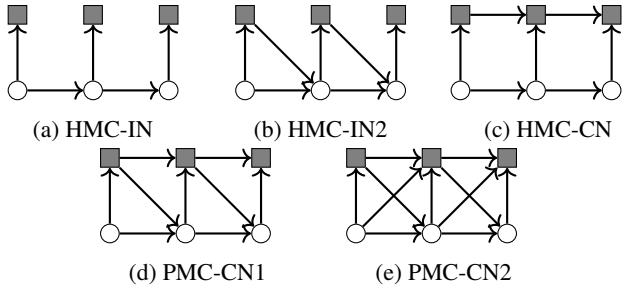

Fig. 1: Graphical representations of the directed probabilistic graphical models of the study. The transitions can themselves be parametrized by a DNN (see Section 3), leading to the *deep* models. White circles represent a hidden random variable and gray squares represent an observed random variable.

We now show that these assumptions can be relaxed via the introduction of parametric functions in the transition distribution (2).

3. DEEP PMCS

3.1. The general model

Let $f_{\boldsymbol{\theta},h}(h_{k-1}, x_{k-1})$ and $f_{\boldsymbol{\theta},x}(h_k, h_{k-1}, x_{k-1})$ be functions parametrized by an unknown parameter $\boldsymbol{\theta}$. We assume that these functions are differentiable w.r.t. $\boldsymbol{\theta}$. We parametrize the PMC in (2) via these two functions,

$$p_{\boldsymbol{\theta}}(h_k | h_{k-1}, x_{k-1}) = \lambda_h(h_k; f_{\boldsymbol{\theta},h}(h_{k-1}, x_{k-1})), \quad (7)$$
$$p_{\boldsymbol{\theta}}(x_k | h_k, h_{k-1}, x_{-1}) = \lambda_x(x_k; f_{\boldsymbol{\theta},x}(h_k, h_{k-1}, x_{k-1})), \quad (8)$$

where $\lambda_h(h; z)$ (resp. $\lambda_x(x; z')$) is a probability distribution (resp. a probability density function) on Ω (resp. on $\mathbb{R}$) w.r.t. h (resp. x) which depends on z (resp. z') and which is differentiable as a function of z (resp. z').

As an illustrative example, let us consider the case where $\Omega = \{\omega_1, \omega_2\}$ and a model where

$$f_{\boldsymbol{\theta},h}(h_{k-1}, x_{k-1}) = a_{h_{k-1}} x_{k-1} + b_{h_{k-1}}, \quad (9)$$
$$f_{\boldsymbol{\theta},x}(h_k, h_{k-1}, x_{k-1}) = \left[c_{h_k,h_{k-1}} x_{k-1} + d_{h_k,h_{k-1}} \sigma_{h_k,h_{k-1}}\right], \quad (10)$$
$$\lambda_h(h = \omega_1; z) = \mathrm{sigm}(z) = \frac{1}{1 + \exp(-z)}, \quad (11)$$
$$\lambda_x(x; z' = (z'_1, z'_2)) = \mathcal{N}(x; z'_1; (z'_2)^2), \quad (12)$$

where $\mathcal{N}(x;\mu;\sigma^2)$ is the Gaussian distribution with mean μ, variance σ^2 taken at point x. Here,

$$\boldsymbol{\theta} = (a_{\omega_i}, b_{\omega_i}, c_{\omega_j,\omega_i}, d_{\omega_j,\omega_i}, \sigma_{\omega_j,\omega_i}),$$

for all $(i,j) \in \{0,1\}^2$.

It can be shown that the transition distribution of the resulting PMC in (7)-(8) and parametrized by (9)-(12) is close to that of the bi-variate Gaussian stationary PMC model of Section 2.2, except that it does not rely on a stationarity assumption [1].

This very general formulation of PMC models enables us to model more complex dependencies through the functions $f_{\boldsymbol{\theta},h}(h_{k-1}, x_{k-1})$ and $f_{\boldsymbol{\theta},x}(h_k, h_{k-1}, x_{k-1})$. In particular, these functions can be represented by DNNs which have the capacity to approximate complex functions. In this case, the set of parameters $\boldsymbol{\theta}$ includes the weights and the biases of these architectures [12].

From now on, we will refer to the general class of PMC models (7)-(8) as Deep PMC models (denoted DPMC models). Let us now derive the Bayesian inference tools related to these PMCs.

3.2. Bayesian restoration in Deep PMCs

Since our Deep PMC models are particular PMC models, the general restoration formulae derived in [1] and which extend the Forward-Backward algorithm [13] are still valid here. We just need to redefine $\alpha_k(h_k) = p_{\boldsymbol{\theta}}(x_1, \cdots, x_k, h_k)$ and $\beta_k(h_k) = p_{\boldsymbol{\theta}}(x_{k+1}, \cdots, x_k | h_k, x_k)$, and we have for all k, $1 \leq k \leq K$,

$$\alpha_k(h_k) = \sum_{h_{k-1}} \alpha_{k-1}(h_{k-1}) p_{\boldsymbol{\theta}}(h_k, x_k | h_{k-1}, x_{k-1}), \quad (13)$$

$$\beta_k(h_k) = \sum_{h_{k+1}} p_{\boldsymbol{\theta}}(h_{k+1}, x_{t+1} | h_k, x_k) \beta_{k+1}(h_{k+1}). \quad (14)$$

3.3. Unsupervised estimation of Deep PMCs

We need to estimate the parameter $\boldsymbol{\theta}$ of $f_{\boldsymbol{\theta},h}$ and $f_{\boldsymbol{\theta},x}$ in (7)-(8) from an observation $\boldsymbol{X} = \boldsymbol{x}$. Unlike [1], we now use a maximum likelihood estimation approach via the Expectation-Maximisation (EM) algorithm [14]. For a fixed $\boldsymbol{\theta}^{(j)}$, the quantity $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \mathrm{E}_{\boldsymbol{\theta}^{(j)}}(\log(p_{\boldsymbol{\theta}}(\boldsymbol{h}, \boldsymbol{x}))|\boldsymbol{x})$ becomes

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \sum_{k=2}^{K} \sum_{h_{k:k+1}} p_{\boldsymbol{\theta}^{(j)}}(h_k, h_{k+1}|\boldsymbol{x}) \times$$
$$\left(\log(p_{\boldsymbol{\theta}}(h_{k+1}|h_k, x_k)) + \log(p_{\boldsymbol{\theta}}(x_{k+1}|h_{k+1}, h_k, x_k))\right), \quad (15)$$

up to the initial distribution of the PMC, and where $p_{\boldsymbol{\theta}^{(j)}}(h_k, h_{k+1}|\boldsymbol{x}) \propto \alpha_k(h_k)\beta_{k+1}(h_{k+1})\, p_{\boldsymbol{\theta}^{(j)}}(h_{k+1}, x_{k+1}|h_k, x_k)$.

It remains to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})$. In general PMC models, computing $\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})$ is impossible, even if we replace $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})$ by its stochastic approximation [SEM].

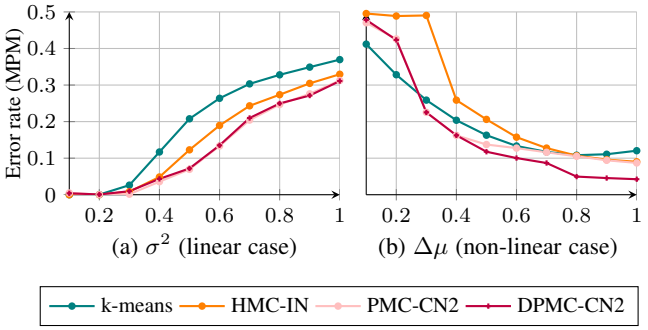

Fig. 2: Error rate in unsupervised segmentations. Section 4.1 (resp. Section 4.2) illustrated on the left (resp. right) graph. Only relative performances are comparable.

When the exact maximization is not possible, we use the Gradient EM (GEM) algorithm [15] to get a set of parameters $\boldsymbol{\theta}^{(j+1)}$ which increases the likelihood. Since $\lambda_h(h;z)$ and $\lambda_x(x;z')$ are differentiable functions of z and z', and since $f_{\boldsymbol{\theta},h}$ and $f_{\boldsymbol{\theta},x}$ are also differentiable as function of $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}(h_{k+1}|h_k, x_k)$ and $p_{\boldsymbol{\theta}}(x_{k+1}|h_{k+1}, h_k, x_k)$ are differentiable w.r.t. $\boldsymbol{\theta}$ and so is $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})$. Finally, introducing a learning rate ϵ, and according to [15], $\boldsymbol{\theta}^{(j)}$ is updated as

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \epsilon \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}, \quad (16)$$

4. EXPERIMENTS AND RESULTS

In this section, we provide experiments that illustrate the generalization offered by Deep PMC models, and we show that they lead to improved error rates in the task of unsupervised signal segmentation. All the following experiments study unsupervised binary image segmentation ($\Omega = \{\omega_1, \omega_2\}$) under the Maximum Posterior Mode criterion (MPM) [16]. When dealing with images, we work with natural images from the Binary Shape Database[1].

For all PMCs (7)-(8), we will use the probability distributions $\lambda_h(h = \omega_1; z)$ and the probability distribution function $\lambda_x(x; z' = (z'_1, z'_2))$ defined in our illustrative example, see (11)-(12). When non deep architectures (resp. deep architectures) are considered, $f_{\boldsymbol{\theta},x}$ and $f_{\boldsymbol{\theta},h}$ are defined from (9)-(10) (resp. from a DNN).

When DNNs are involved, $\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},x}$ and $\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},h}$ are computed via the backpropagation algorithm [17] for the update rule (16). In practice, convergence issues may be encountered when we try to maximize the loglikelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with randomly initialized parameters. To deal with this issue, we propose a *pretraining* step. In our case, the parameters of $f_{\boldsymbol{\theta},x}$ and $f_{\boldsymbol{\theta},h}$ are initialized with some iterations of a backpropagation algorithm which uses $\boldsymbol{x}$ as inputs and the latent variables

[1] http://vision.lems.brown.edu/content/available-software-and-databases

estimated by an easily available segmentation (such as the result of the k-means algorithm [18], or a less general model) as output. The cost function of the backpropagation procedure is chosen as the mean square error, (resp. cross-entropy), for the DNN related to $f_{\boldsymbol{\theta},x}$ (resp. $f_{\boldsymbol{\theta},h}$).

Finally, we use a Hilbert-Peano filling curve [19] to transform the images into a unidimensional data structure [20, 21].

4.1. Segmentations of linear and correlated noise

This experiment consists of the binary shapes artificially corrupted with a 0-mean, additive and correlated Gaussian noise taken as the realization of a Gaussian Markov Random Fields (GMRF) with an exponential correlation function [22]. Such a noise is then parametrized by a correlation range r (that we fix to 3) and a noise variance σ^2. The DNNs for the Deep model are set to one hidden layer with 10 neurons. Note that we also give, for comparison purposes, the results of the K-means algorithm.

Fig. 2a illustrates the experiment. The results are averaged on all the 'dude' images of the database. We can see that the deep model performs equivalently for all the noise levels, which should not be surprising since we can show that this kind of noise introduces a Gaussian linear relation between the observations. So it shows that our general PMC model is able to adapt to the linear case.

4.2. Segmentations of non-linear and correlated noise

In this section we propose to blur 1-D signals with a complex noise which exhibits non-linearities to highlight the ability of the generalized models to learn such a signal corruption. The hidden signal is taken as the realization of a two states Markov chain with symmetric transition matrix with a switching probability is fixed to 0.2. Then each observation is simulated as

$$x_k \sim \mathcal{N}\Big(a_{h_k}+\cos(x_{k-1});(c_{h_k}+\mathbb{1}_{\{x_{k-1}<0\}}[h_{k-1}\oplus h_k]b_{h_k})^2\Big),$$

where $\oplus$ denotes the logical XOR operator. Such a noise is correlated and has the property that when transitioning from an hidden state to an other, the noise distributions changes if $x_{k-1} < 0$. Set $\Delta\mu = |a_{\omega_1} - a_{\omega_2}|$, $c_{\omega_1} = c_{\omega_2} = 0.2$ and $b_{\omega_1} = b_{\omega_2} = 0.6$. All these parameters are then considered lost to meet the case of unsupervised segmentation. The DNNs for the Deep model are set to one hidden layer with 2 neurons.

Fig. 2b illustrates the error rate with the k-means, HMC-IN, PMC-CN2 and DPMC-CN2 models (other models are discarded for brevity of the presentation) for a varying $\Delta\mu$, the latter affects the noise level. It is clear that the general model performs always better or equivalently, whatever the noise model, with up to a 4%-point improvement. This fact agrees with the theory we introduced earlier where DPMC-CN2 stands out as the most general model. From a signal processing point of view, the gain in the error rate obtained with our general model tends to show that it is now able to learn such non-linear noise and is available without any further modeling effort.

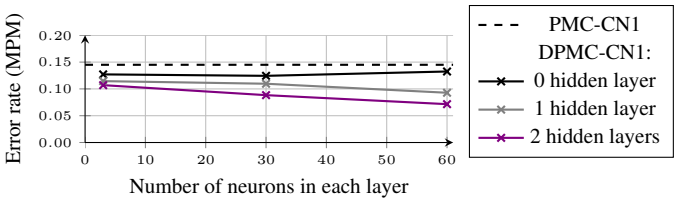

Fig. 3: Error rate of the DPMC-CN1 model as a function of the NN architectures. Note that in this graph, HMC-IN mean error rate is a constant equal to 43.9%. The results are averaged on all the 'camel' images of the database.

4.3. DNN architectures experiment

We finally turn back again to image segmentation and we propose to blur images with a causal correlated and non-linear noise defined by

$$x_k \sim \mathcal{N}\Big(a_{h_k}x_{k-1} + b_{h_k} + \cos(x_{k-1}); (c_{h_k} + \cos(x_{k-1}))^2\Big).$$

We set $a_{\omega_1} = 0.2$, $a_{\omega_2} = 0.5$, $c_{\omega_1} = 0.2$, $c_{\omega_2} = 0.3$ and $b_{\omega_1} = 0.2$, $b_{\omega_2} = 0.6$. Again, these parameters are considered lost, because we simulate the case of unsupervised segmentation. In this experiment we study the variation of the error rate with the DPMC-CN2 model when its DNN architectures evolve. The same number of hidden layers and of neurons are used for parametrizing (7)-(8)

Fig. 3 shows that we can decrease the error by using a more complex DNN architecture. It is notable that the segmentation can be greatly improved without any more modeling effort (the price to pay is purely computational). Also, notice that DPMC-CN1 performs better than PMC-CN1 even without hidden layers. It may seem counter intuitive because both model are equivalent, but it actually highlights the interest of our proposed unsupervised *pretraining* by backpropagation as a way to initialize the gradient EM procedures.

5. CONCLUSION

In this paper, we have introduced a deep probabilistic architecture based on PMC models and on DNNs adapted to the unsupervised image segmentation problem. For this general deep probabilistic model, we have proposed a parameter estimation procedure and tuned it in order to outperform classical probabilistic models for image segmentation. Our experiments have indeed shown that substantial gains can be attained over the classical models.

In a future work, we might consider studying similar generalizing approaches in the context of Triplet Markov Models [23], by introducing a third random process to complexify the distribution of the noise, for example.

6. REFERENCES

[1] W. Pieczynski, "Pairwise Markov chains," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 5, pp. 634–639, 2003.

[2] I. Gorynin, H. Gangloff, E. Monfrini, and W. Pieczynski, "Assessing the segmentation performance of pairwise and triplet Markov Models," *Signal Processing*, vol. 145, pp. 183–192, 2018.

[3] J.-B. Courbot, V. Mazet, E. Monfrini, and C. Collet, "Pairwise Markov fields for segmentation in astronomical hyperspectral images," *Signal Processing*, vol. 163, pp. 41–48, 2019.

[4] H. Gangloff, J.-B. Courbot, E. Monfrini, and C. Collet, "Unsupervised image segmentation with gaussian pairwise Markov fields," *Computational Statistics & Data Analysis*, vol. 158, pp. 107178, 2021.

[5] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[7] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams, "Composing graphical models with neural networks for structured representations and fast inference," *arXiv preprint arXiv:1603.06277*, 2016.

[8] K. Tran, Y. Bisk, A. Vaswani, D. Marcu, and Kevin Knight, "Unsupervised neural Hidden Markov Models," *arXiv preprint arXiv:1609.09007*, 2016.

[9] S. Derrode and W. Pieczynski, "Signal and image segmentation using pairwise Markov chains," *IEEE Transactions on Signal Processing*, vol. 52, no. 9, pp. 2477–89, 2004.

[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," *arXiv preprint arXiv:1411.7610*, 2014.

[12] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[13] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[15] S. Balakrishnan, M. J. Wainwright, B. Yu, et al., "Statistical guarantees for the em algorithm: From population to sample-based analysis," *Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[16] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Journal of the American statistical association*, vol. 82, no. 397, pp. 76–89, 1987.

[17] K. Gurney, *An Introduction to Neural Networks*, Taylor & Francis, Inc., USA, 1997.

[18] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the 18th ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[19] H. Sagan, *Space-filling curves*, Springer, 2012.

[20] S. Bricq, C. Collet, and J.-P. Armspach, "Unifying framework for multimodal brain MRI segmentation based on Hidden Markov chains," *Medical image analysis*, vol. 12, no. 6, pp. 639–652, 2008.

[21] M. Yahiaoui, E. Monfrini, and B. Dorizzi, "Markov chains for unsupervised segmentation of degraded NIR iris images for person recognition," *Pattern Recognition Letters*, vol. 82, pp. 116–123, 2016.

[22] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*, CRC press, 2005.

[23] W. Pieczynski, "Chaines de Markov triplet," *Comptes Rendus de l'Academie des Sciences - Mathematiques*, vol. 335, pp. 275–278, 2002, in French.