Pairwise Markov Chains as Generative Models

Katherine MORALES[1], Yohan PETETIN[1]

[1]Samovar, Telecom Sudparis, Institut Polytechnique de Paris, 91011 Évry, France
katherine.morales_quinga@telecom-sudparis.eu, yohan.petetin@telecom-sudparis.eu

Résumé – Les modèles génératifs tels que la chaîne de Markov cachée, le réseau de neurones récurrents et le réseau de neurones récurrents stochastique ont de nombreuses applications avec des données séquentielles. Dans cet article, nous nous concentrons sur une extension de ces modèles, les chaînes de Markov Couple. Nous proposons ce modèle comme un modèle génératif et un algorithme d'estimation des paramètres basé sur une approche bayésienne variationnelle. Nous analysons également le cas linéaire et Gaussien, où nous pouvons caractériser la distribution des observations. Les résultats d'expériences montrent que ce nouveau modèle donne de meilleurs résultats que d'autres modèles génératifs.

Abstract – Generative models such as the Hidden Markov Chain, the Recurrent Neural Network, and the Stochastic Recurrent Neural Network have found many applications with sequential data. In this paper, we focus on a particular extension of these models, the Pairwise Markov Chain. We propose this model as a generative model with a parameter estimation algorithm based on a variational Bayesian approach. We also analyze the particular linear and Gaussian case, where it is possible to characterize the generative distribution. Finally, we show that our model with its associated algorithm outperforms other generative models.

1 Introduction

Let $\mathbf{x}_T = (x_0, \dots,\ x_T)$, $x_t \in \mathbb{R}^{d_x}$ and $\mathbf{h}_T = (h_0, \dots,\ h_T)$, $h_t \in \mathbb{R}^{d_h}$ be two sequences of observed and hidden latent random variables (r.v.) of length $T+1$, respectively. As far as notations are concerned, we do not distinguish r.v. and their realizations.

In this article, our interest is to present a generative model approach based on latent r.v. which is defined by a joint distribution $p_\theta(\mathbf{x}_T, \mathbf{h}_T)$ and provides learning from the observations $\mathbf{x}_T$ since its distribution reads $p_\theta(\mathbf{x}_T) = \int p_\theta(\mathbf{x}_T, \mathbf{h}_T)\mathrm{d}\mathbf{h}_T$. First, $(\mathbf{x}_T, \mathbf{h}_T)$ are described by a parameterized distribution $p_\theta(\mathbf{x}_T, \mathbf{h}_T)$ which allows to model the unknown distribution $p(\mathbf{x}_T, \mathbf{h}_T)$. Next, the set of parameters θ is estimated from the realizations $\mathbf{x}_T$. Finally, for a set of known parameters, we can generate observations from $p_\theta(\mathbf{x}_T)$.

A popular generative model is the Hidden Markov Chain (HMC) [1] which has been used for sequential data modeling problems (*e.g* music, images, text). In a HMC, the sequence $\mathbf{h}_T$ is a Markov chain and given $\mathbf{h}_T$, the observations x_t are independent and only depend on the corresponding h_t. The joint distribution $p_\theta(\mathbf{h}_T, \mathbf{x}_T)$ is factorized as

$$p_\theta(\mathbf{x}_T, \mathbf{h}_T) = p_\theta(h_0) \prod_{t=1}^{T} p_\theta(h_t|h_{t-1}) \prod_{t=0}^{T} p_\theta(x_t|h_t), \qquad (1)$$

where $p_\theta(h_t|h_{t-1})$ and $p_\theta(x_t|h_t)$ are the distributions representing the transitions of the Markov chain $\mathbf{h}_T$ and the relations between the observation and the hidden variable, respectively. This model has been generalized by the introduction of the Pairwise Markov Chain (PMC)[2] which only satisfies the assumption that the pair $(\mathbf{x}_T, \mathbf{h}_T)$ is a Markov chain. PMC incorporates more complex relationships between the observed and latent variables. However, PMC has been introduced in a Bayesian framework where the objective is to estimate the latent process from the observed one [3, 4, 5]. The aim of this paper is to present the Pairwise Markov Chain (PMC) as a (general) generative model, which includes some generative models such as the Hidden Markov Chain (HMC) [6, 7], the Recurrent Neural Network (RNN) [8, 9], and the Stochastic RNN (SRNN) [10, 11] and to extend some of the models proposed in [12]. On the other hand, a maximum likelihood estimation can be proposed for the estimation of θ. However, a direct maximization of $p_\theta(\mathbf{x}_T) = \int p_\theta(\mathbf{x}_T, \mathbf{h}_T)\mathrm{d}\mathbf{h}_T$ is not always possible due to the fact that if a model comprises many unobservable variables the integration can become analytically burdensome or even intractable. Here, we focus on variational Bayesian approaches, which are particularly suitable for high dimensional models [13]. The paper is organized as follows. In Section 2, we introduce the PMC and a variational Bayesian approach to estimate the set of parameters θ of general PMCs. Next, in Section 3, we present the connection between the PMC and the popular generative models presented before. Finally, we describe examples of a generative PMCs and we compare them on simulations in Section 4.

2 Generative PMC

2.1 Definition

The PMC is a direct generalization of HMC which has received a particular attention for image segmentation, see e.g. [14, 15, 16]. In this article, we present the PMC as a generative model which aim at modeling an unknown distribution $p_\theta(\mathbf{x})$ of observations.

The PMC only assumes that the pair $(\mathbf{x}_T, \mathbf{h}_T)$ is markovian, with transition $p(h_t, x_t|h_{t-1}, x_{t-1})$. The distribution $p_\theta(\mathbf{h}_T, \mathbf{x}_T)$ reads

$$p_\theta(h_0, x_0) \prod_{t=1}^{T} p_\theta(h_t|h_{t-1}, x_{t-1}) p_\theta(x_t|h_{t-1}, x_{t-1}, h_t). \quad (2)$$

2.2 Variational Inference

We propose a variational Bayesian approach for estimating the set of parameters θ from a realization $\mathbf{x}_T$. In the variational inference framework, it is not the log model $\log(p_\theta(\mathbf{x}_T))$ itself which is evaluated, but rather a lower bound approximation to it, called the Evidence Lower Bound (ELBO). This ELBO $Q(\theta, q_\phi)$ is derived from the negative (exclusive) Kullback-Leiber (KL) divergence between a variational distribution $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$ and the posterior distribution $p_\theta(\mathbf{h}_T|\mathbf{x}_T)$, which is convenient because of the complexity of $p_\theta(\mathbf{h}_T|\mathbf{x}_T)$ [13]. The following inequality holds for any variational distribution $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$,

$$\begin{aligned} \log(p_\theta(\mathbf{x}_T)) &\geq Q(\theta, q_\phi) & (3) \\ &= -\int \log\left(\frac{q_\phi(\mathbf{h}_T|\mathbf{x}_T)}{p_\theta(\mathbf{x}_T, \mathbf{h}_T)}\right) q_\phi(\mathbf{h}_T|\mathbf{x}_T) \mathrm{d}\mathbf{h}_T. & (4) \end{aligned}$$

Our objective is to maximize the ELBO w.r.t θ and q_ϕ which can be done with the Expectation-Maximization (EM) algorithm [17]. However, it relies on the computation of $p_\theta(\mathbf{h}_T|\mathbf{x}_T)$ and since $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$ can be chosen, in general, as a parametric function [18, 19], it is possible to approximate $Q(\theta, q_\phi)$ by using the reparametrization trick [20] which allows to obtain samples $\mathbf{h}_T^{(i)} \sim q(\mathbf{h}_T|\mathbf{x}_T)$ that can be written as a differentiable function of ϕ. The choice of the variational distribution $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$ is important, we have to consider that it should be close to $p_\theta(\mathbf{h}_T, \mathbf{x}_T)$ but, at the same time, the associated ELBO should be calculable or easily approximated while remaining differentiable w.r.t. (θ, ϕ). In the case of the PMC, remember that $p(\mathbf{x}_T, \mathbf{h}_T)$ coincides with (2). Thus, the ELBO in (4) reads

$$\begin{aligned} Q(\theta, \phi) = &-\int \log\left(\frac{q_\phi(h_0|\mathbf{x}_T)}{p(x_0, h_0)}\right) q_\phi(h_0|\mathbf{x}_T) \mathrm{d}\mathbf{h}_T \\ &-\sum_{t=1}^{T} \int \log\left(\frac{q_\phi(h_t|\mathbf{h}_{t-1}, \mathbf{x}_T)}{p_\theta(h_t, x_t|h_{t-1}, x_{t-1})}\right) q_\phi(\mathbf{h}_t|, \mathbf{x}_T) \mathrm{d}\mathbf{h}_t. \quad (5) \end{aligned}$$

Since $p_\theta(h_t|h_{t-1}, \mathbf{x}_T)$ is generally not computable in PMC models (except in the linear and Gaussian case) we can choose a variational distribution from which a sample can be obtained with the reparametrization trick, and which satisfies

$$q_\phi(h_t|\mathbf{h}_{t-1}, \mathbf{x}_T) = q_\phi(h_t|h_{t-1}, \mathbf{x}_t). \quad (6)$$

Example 1. *We choose a variational distribution as follows*

$$q_\phi(h_t|h_{t-1}, \mathbf{x}_t) = \mathcal{N}(h_t; f_\phi(h_{t-1}, \mathbf{x}_t); \mathrm{diag}(g_\phi(h_{t-1}, \mathbf{x}_t))),$$

where $\mathcal{N}(h_t; \mu; \Sigma)$ denotes the Gaussian distribution with mean μ and variance Σ taken at point h_t; and f_ϕ and g_ϕ are parameterized and differentiable functions of ϕ, $\mathrm{diag}(\cdot)$ denotes the diagonal matrix deduced from the values of g_ϕ and where a sample $h_t^{(i)} \sim q_\phi(h_t|h_{t-1}, \mathbf{x}_t)$ can be obtained as

$$h_t^{(i)} = f_\phi(h_{t-1}, \mathbf{x}_t) + (\mathrm{diag}(g_\phi(h_{t-1}, \mathbf{x}_t)))^{\frac{1}{2}} \times \epsilon^{(i)}, \quad (7)$$

with $\epsilon^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that for this choice of variational distribution, the components h_t are independently given $(h_{t-1}, \mathbf{x}_t)$ in the regard of the variational distribution q_ϕ.

Thus, by sampling $\mathbf{h}_T^{(i)} \sim q(h_0|x_0) \times \prod_{t=1}^{T} q_\phi(h_t|h_{t-1}, \mathbf{x}_t)$, for all i, $1 \leq i \leq N$, $Q(\theta, \phi)$ in (5) *can be approximated by (up to the term associated to $t = 0$ that we omit for clarity)*

$$\widehat{Q}(\theta, \phi) = -\sum_{i=1}^{N} \sum_{t=1}^{T} \log\left(\frac{q_\phi(h_t^{(i)}|h_{t-1}^{(i)}, \mathbf{x}_t)}{p_\theta(h_t^{(i)}, x_t|h_{t-1}^{(i)}, \mathbf{x}_{t-1})}\right) \quad (8)$$

and optimized with a gradient ascent algorithm w.r.t. (θ, ϕ).

3 Relation with other generative models

3.1 Generative models

HMC. As recalled in the Introduction, the HMC is a popular model which satisfies

$$p_\theta(\mathbf{x}_T, \mathbf{h}_T) = p_\theta(h_0) \prod_{t=1}^{T} p_\theta(h_t|h_{t-1}) \prod_{t=0}^{T} p_\theta(x_t|h_t). \quad (9)$$

where $\mathbf{h}_T$ is a Markov chain, and $p(\mathbf{x}_T|\mathbf{h}_T) = \prod_{t=0}^{T} p(x_t|h_t)$.

RNN. An RNN is a particular neural network where the latent variable h_t is deterministically obtained given the previous observation x_{t-1} and the previous latent variable h_{t-1} (so $p_\theta(h_t|h_{t-1}, x_{t-1})$ becomes a Dirac measure). Its expression relies on an activation function f_θ. As in the HMC, given $\mathbf{h}_T$, the observations $\mathbf{x}_T$ are independent and x_t only depends on h_t. The distribution of $(\mathbf{x}_T, \mathbf{h}_T)$ reads

$$\begin{aligned} h_t &= f_\theta(h_{t-1}, x_{t-1}), & (10) \\ p_\theta(x_t|\mathbf{x}_{t-1}) &= p_\theta(x_t|h_t). & (11) \end{aligned}$$

SRNN. The SRNN is an extension of the RNN where the latent variable h_t becomes random and also depends on x_{t-1} given the past observations and the latent variables. The generative model is given by,

$$p_\theta(\mathbf{x}_T, \mathbf{h}_T) = p_\theta(h_0) \prod_{t=0}^{T} p_\theta(x_t|h_t) \prod_{t=1}^{T} p_\theta(h_t|x_{t-1}, h_{t-1}). \quad (12)$$

This model includes the Variational RNN (VRNN) [11] or the Stochastic Recurent network (STORN) [10].

3.2 Theoretical comparison

In this section, we focus on linear and Gaussian PMC with $h_t \in \mathbb{R}$ and scalar observations. Model (2) satisfies

$$p_\theta(h_0, x_0) = \mathcal{N}\left(\begin{pmatrix} h_0 \\ x_0 \end{pmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix}\right) \quad (13)$$

$$p_\theta(h_t|h_{t-1}, x_{t-1}) = \mathcal{N}(h_t; a h_{t-1} + c x_{t-1}, \alpha), \quad (14)$$

$$p_\theta(x_t|h_{t-1:t}, x_{t-1}) = \mathcal{N}(x_t; b h_t + e h_{t-1} + f x_{t-1}, \beta), \quad (15)$$

where $\theta = (a, b, c, e, f, \alpha, \beta, \eta, \gamma)$. The linear and Gaussian SRNN coincides with $e = f = 0$, $\gamma = b$, while the linear and Gaussian HMC also satisfies $c = 0$.

Our objective is to build a generative model $p_\theta(\mathbf{x}_T)$ such that it coincides with a Gaussian distribution $p(\mathbf{x}_T)$, for all $T \geq 0$, which satisfies

$$p(x_t) = \mathcal{N}(x_t; 0; 1), \text{ for all } 0 \leq t \leq T. \tag{16}$$

We have shown in [12] that the associated generative distribution reads, for all positive integers T, t, k,

$$\begin{aligned} p_\theta(\mathbf{x}_T) &= \mathcal{N}(\mathbf{x}_T; \mathbf{0}; \overline{\Sigma}), & (17)\\ \text{cov}(x_t, x_{t+k}) &= \overline{A}^k(\overline{B} + \frac{1}{2}) - \overline{C}^k(\overline{B} - \frac{1}{2}), & (18)\end{aligned}$$

where $\overline{A}, \overline{B}, \overline{C}$ and K depend on θ, and the following constraints are satisfied :

$$\gamma\eta = b\eta + (ae + af\gamma + ce\gamma) + fc, \tag{19}$$

$$0 \leq (1 - a^2 - 2ac\gamma)\eta - c^2, \tag{20}$$

$$0 \leq 1 - b^2\eta - 2b\eta(\gamma - b) - e\eta(e + 2f\gamma) - f^2. \tag{21}$$

The proof of this result is presented in [12]. This results shows that the linear and Gaussian PMC can model some Gaussian distributions which cannot be modeled by the linear and Gaussian SRNN [21] since the linear and Gaussian SRNN is able to model any centered Gaussian distribution with a covariance matrix $\text{cov}(x_t, x_{t+k}) = A^{k-1}B$, for all $T \geq 0$, $t \geq 0$ and $k \geq 0$.

4 Simulations

4.1 Deep generative PMCs

In this section, we present our deep generative PMCs (DPMCs), a generalization of the VRNN [11]. DPMCs a particular instance of the PMC where the parameters can be produced by any function $\psi(\cdot)$, in particular, by (deep) neural networks. Additionally, we set $h_t = (h'_t, z_t)$ with the latent variable h'_t deterministic. The transition (2) of this model is described with the following set of equations :

$$h'_t = f(x_{t-1}, z_{t-1}, h'_{t-1}), \tag{22}$$

$$p_\theta(z_t | z_{t-1}, h'_{t-1:t}, x_{t-1}) = \mathcal{N}(z_t; \mu_{pz,t}; \text{diag}(\sigma_{pz,t})), \tag{23}$$

$$p_\theta(x_t | z_{t-1:t}, h'_{t-1:t}, x_{t-1}) = \text{Ber}(x_t; \rho_{x,t}). \tag{24}$$

We denote $\mathcal{N}(z; \mu; \Sigma)$ the Gaussian distribution, $\text{Ber}(x; \rho)$ the Bernoulli distribution with parameter ρ; and f is a deterministic non-linear function describring a RNN cell. On the other hand, the variational distribution q_ϕ is given by

$$q_\phi(z_t | z_{t-1}, \mathbf{x}_t) = \mathcal{N}(z_t; \mu_{qz,t}; \text{diag}(\sigma_{qz,t})). \tag{25}$$

The parameters $\theta = \{\mu_{pz,t}, \sigma_{pz,t}, \rho_{x,t}\}$ and $\phi = \{\mu_{qz,t}, \sigma_{qz,t}\}$ can be derived with

$$[\mu_{qz,t}, \sigma_{qz,t}] = \psi_{qz}\big(x_t, h'_t\big), \tag{26}$$

$$[\mu_{pz,t}, \sigma_{pz,t}] = \psi_{pz}(h'_t), \tag{27}$$

$$\rho_{x,t} = \psi_{px}\big(z_{t:-1:t}, h'_{t-1:t}, x_{t-1}\big). \tag{28}$$

In the VRNN [11], Eq. (24) does not depend on h'_{t-1}, z_{t-1} and x_{t-1} . We also introduce three particular instances of this model, DPMC-I, DPMC-II and DPMC-III whose simulation results will be presented in the next section. Note that DPMC is the model defined by Eqs. (22)-(28). For the DPMC-I (resp. DPMC-II), Eq. (24) does not depend on h'_{t-1} and z_{t-1} (resp. z_{t-1}). In the DPMC-III, ψ_{pz} in (27) also depends on x_{t-1} and satisfies the conditions of DPMC-II.

4.2 Results

In this section, we compare the VRNN with the classical RNN (10)-(11), the DPMC and its instances. We use the set of MIDI music [22] and the MNIST [23] data set.

MNIST data set contains 60000 (resp. 10000) train (resp. test) 28 × 28 binary images. An observation x_t consists of a column of the image, and the length of a sequence is $T = 28$. In the MIDI music set, three polyphonic music data sets are used, the classical piano music (Piano), the folk tunes (Nottingham) and the four-part chorales by J.S. Bach (JSB). In this case, we use an input of 88 binary visible units that span the whole range of piano from A0 to C8.

Each model was trained with stochastic gradient descent on the negative evidence lower bound using the Adam optimizer [24]. $\psi_{pz}, \psi_{qz}, \psi_{px}$ in Eqs. (26)-(28) are the outputs of two hidden layers using rectified linear units. Note that the standard RNN model only has ψ_{x_p}. Additionally, we match the total number of parameters of all models to be equal or close between them, so the number of hidden units is different for each model.

For MIDI data sets, we set the dimension of z to be 300 and we use 300 (resp. 260, 272, 278, 294, 562) hidden units for the VRNN (resp. DPMC, DPMC-III, DPMC-II, DPMC-I, RNN). On the another hand, for MNIST we set 100 (resp. 78, 74, 79, 95, 162) hidden units for the VRNN (resp. DPMC, DPMC-III, DPMC-II, DPMC-I, RNN) and the dimension of z is three.

In Table 1, we report the log-likelihoods $\log p_\theta(\mathbf{x}_T)$ on the test datasets. Except for the RNN, these likelihoods are approximated by an importance sampler [25] using 100 samples.

TABLE 1 – Results on the MIDI and MNIST data sets.

	MNIST	Piano	Nottingham	JSB
RNN	-65,70	-10,52	-23,89	-10,77
VRNN	≈ -64,76	≈ -9,40	≈ -13,30	≈ -10,27
DPMC	≈ -64,88	≈ -9,23	≈ -13,39	≈ -10,11
DPMC-I	≈ -64,70	≈ -9,31	≈ -11,39	≈ -10,31
DPMC-II	≈ **-64,26**	≈ **-8,83**	≈ -14,85	≈ -10,24
DPMC-III	≈ -64,92	≈ -9,41	≈ **-10,60**	≈ **-9,23**

In general, higher numbers are better. Our results show that DPMC-II (resp. DPMC-III) has the higher average approximated log-likelihood with the MNIST and Piano (resp. Nottingham and JSB) data sets. In other words, the results show that we have an improvement in terms of likelihood by incorporating more complex relationships between the observed and latent variables. As we see, the particular instances of DPMC perform better than the VRNN and the classical RNN.

5 Conclusion

In this paper, we have included popular generative models into a common model. We have shown that the PMC allows to model complex distribution w.r.t. the SRNN, in the linear and Gaussian case. We have proposed a parameter estimation algorithm for PMCs and our experiments have indeed shown that a better performance can be attained over the classical models.

Références

[1] 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.

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

[3] W. Pieczynski, “Pairwise Markov chains and Bayesian unsupervised fusion,” in *Fusion 2000*, Paris, France, July 2000, vol. 1.

[4] S. Derrode and W. Pieczynski, “SAR image segmentation using generalized pairwise Markov chains,” in *Proceedings of SPIE International Symposium on Remote Sensing*, Crete, Greece, September 22-27, 2002.

[5] N. Brunel and W. Pieczynski, “Unsupervised signal restoration using copulas and pairwise Markov chains,” in *Proceedings of the 2003 IEEE Workshop on Statistical Signal Processing*, St. Louis, MI, September 2003.

[6] 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, February 1989.

[7] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, Springer-Verlag, 2005.

[8] J.-T. Connor, R. Martin, Douglas, and L.-E Atlas, “Recurrent Neural Networks and robust time series prediction,” *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 240–254, 1994.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of Gated Recurrent Neural Networks on sequence modeling,” *NeurIPS*, 2014.

[10] J. Bayer and C. Osendorfer, “Learning stochastic recurrent networks,” *arXiv preprint arXiv :1411.7610*, 2014.

[11] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C Courville, and Y. Bengio, “A recurrent latent variable model for sequential data,” in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[12] K. Morales and Y. Petetin, “Variational bayesian inference for pairwise markov models,” in *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2021, pp. 251–255.

[13] D.-M. Blei, A. Kucukelbir, and J.-D. McAuliffe, “Variational Inference : A review for statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Apr 2017.

[14] 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.

[15] 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.

[16] 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.

[17] 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 (B)*, vol. 39, no. 1, pp. 1–38, 1977.

[18] 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.

[19] Diederik P Kingma and Max Welling, “An introduction to variational autoencoders,” *arXiv preprint arXiv :1906.02691*, 2019.

[20] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *2nd International Conference on Learning Representations, ICLR*, 2014.

[21] A. Salaün, Y. Petetin, and F. Desbouvries, “Comparing the modeling powers of RNN and HMM,” in *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 1496–1499.

[22] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu, “Advances in optimizing recurrent networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8624–8628.

[23] Yann LeCun, “The mnist database of handwritten digits,” *http ://yann. lecun. com/exdb/mnist/*, 1998.

[24] D. Kingma and J. Ba, “Adam : A method for stochastic optimization,” *International conference on learning representations*, 12 2014.

[25] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, “Stochastic backpropagation and approximate inference in Deep Generative Models,” in *International Conference on Machine Learning*, 2014, vol. 2.