Variational Bayesian inference for pairwise Markov models

Katherine Morales, Yohan Petetin

VARIATIONAL BAYESIAN INFERENCE FOR PAIRWISE MARKOV MODELS

Katherine Morales, Yohan Petetin

Samovar, Telecom Sudparis, Institut Polyechnique de Paris, 91011 Evry, France

ABSTRACT

Generative models based on latent random variables are a popular tool for time series forecasting. Generative models include the Hidden Markov Model, the Recurrent Neural Network and the Stochastic Recurrent Neural Network. In this paper, we exploit the Pairwise Markov Models, a generalization of Hidden Markov models, as generative models. We first show that the previous generative models are a particular instance of Pairwise Markov models. Next, we also show that they can potentially model a large class of distributions for given observations. In particular, we analyze the particular linear and Gaussian case, where it is possible to characterize the modeling power of these generative models. Finally, we present a parameter estimation algorithm for general Pairwise Markov Models based on Bayesian variational approaches. Simulations are presented and support our statements.

***Index Terms*—** Generative Models; Variational Inference; Time series; Recurrent Neural Networks; Pairwise Markov Models.

1. INTRODUCTION

Let $x \in \mathbb{R}^d$ be a random variable (r.v.) and $\mathbf{x}_T = (x_0, x_1, \ldots, x_T)$ a sequence of r.v. of length $T+1$, for all $T \geq 0$. The (unknown) distribution of $\mathbf{x}_T$ is noted $p(\mathbf{x}_T)$. As far as notations are concerned, we do not distinguish random variables and their realizations.

1.1. Time series modelling

Time series modelling appears in numerous problems in statistical signal processing [1, 2] such as prediction [3], econometrics [4] or speech recognition [5]. In this paper, we focus on a generative model approach which consists of different steps. First, the observations $\mathbf{x}_T$ are described by a generative models, i.e. a distribution $p_\theta(\mathbf{x}_T)$ which models the unknown distribution $p(\mathbf{x}_T)$; next the parameter θ is estimated from a realization $\mathbf{x}_T$. In particular, the maximum likelihood estimate is a popular estimate due to its asymptotic properties [6, 7]. Finally, when the parameter θ is known, the prediction of future observations can be deduced from the distribution $p_\theta(x_{T+1}, \ldots, x_{T+\tau}|\mathbf{x}_T)$. The practical choice of a model p_θ should take into account these modelling and computational constraints.

Among popular generative models, we are interested in those based on latent r.v. $(h_t)_{t\geq 0}$, where $h_t \in \mathbb{R}^m$. These models are defined by a joint distribution $p_\theta(\mathbf{x}_T, \mathbf{h}_T)$ from which the distribution of the observation reads $p_\theta(\mathbf{x}_T) = \int p_\theta(\mathbf{x}_T, \mathbf{h}_T)\mathrm{d}\mathbf{h}_T$. These models include the Hidden Markov Model (HMM) [5, 8], the Recurrent Neural Network (RNN) [9, 10], and the Stochastic RNN (SRNN) [11, 12]. In this paper, we start by including all these models into a common probabilistic model called the Pairwise Markov Model (PMM) [13]. This model has been introduced in a Bayesian framework where the objective is to estimate the latent process from the observed one [14, 15, 16, 17]. Next, i) we focus on the generative aspect of the PMM, and show that this model $p_\theta(.)$ is potentially relevant to take into account a (unkonwn) complex distribution of the observations $p(.)$; ii) we show that the parameters of PMMs can be estimated with a Bayesian variational approach [18]; iii) we compare the PMM with classical generative models on simulations.

1.2. Generative models based on latent variables

Before giving technical details about our work, we first review the popular generative models introduced in this section.

1.2.1. HMM

A (continuous state) HMM is a model where the latent process $(h_t)_{t\geq 0}$ is Markovian; moreover, given $\mathbf{h}_T$, the observations $\mathbf{x}_T$ are independent and x_t only depends on h_t. For all T, the distribution of $(\mathbf{h}_T, \mathbf{x}_T)$ reads

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

Bayesian inference algorithms in such models are based on the Kalman filter and its extensions, and also on sequential Monte Carlo methods [19].

1.2.2. RNN

An RNN is a particular neural network which takes into account the sequential aspect of the data. Contrary to the HMM, the latent variable h_t is deterministically obtained given the previous observation x_{t-1} and the previous latent variable h_{t-1}. Its expression relies on an activation function f_θ, which

can be parametrized by a feed-forward neural network, for example. As in the HMM, given $\mathbf{h}_T$, the observations $\mathbf{x}_T$ are independent and x_t only depends on h_t. The generative models is given by

$$h_t = f_\theta(h_{t-1}, x_{t-1}), \quad (2)$$
$$p_\theta(x_t|x_{0:t-1}) = p_\theta(x_t|h_t). \quad (3)$$

1.2.3. SRNN

Finally, the SRNN is an extension of the RNN where the latent variable h_t becomes random, as in the HMM, but can also depend on x_{t-1} given the past observations and the latent variables. In other words,

$$p_\theta(\mathbf{x}_T, \mathbf{h}_T) = \prod_{t=1}^{T} p_\theta(x_t|h_t)p_\theta(h_t|x_{t-1}, h_{t-1}). \quad (4)$$

In fact, this general stochastic model includes both the HMM and the RNN and has been studied in [20] in the linear and Gaussian framework. Neural network architectures such as the Stochastic Recurent network (STORN) [11] or the Variational RNN (VRNN) [12] are also particular SRNNs.

1.3. Scope of the paper

Even if the SRNN includes the HMM and the RNN, observe that $(h_t)_{t\geq 0}$ remains a Markovian process. Indeed, starting from (4), it is easy to show that $p_\theta(h_T|\mathbf{h}_{T-1}) \propto \int p_\theta(\mathbf{h}_T, \mathbf{x}_T)\mathrm{d}\mathbf{x}_T$ coincides with $p_\theta(h_T|h_{T-1})$.

Thus, the scope of this paper is to propose a generative model in which the latent process in not necessarily Markovian. Our model is based on the PMM which only relies on the assumption that the pair $(h_t, x_t)_{t\geq 0}$ is Markovian, with transition $p(h_t, x_t|h_{t-1}, x_{t-1})$. The distribution of $(\mathbf{h}_T, \mathbf{x}_T)$ associated to a PMM reads

$$p_\theta(h_0, x_0)\prod_{t=1}^{T} p_\theta(h_t|h_{t-1}, x_{t-1})p_\theta(x_t|h_{t-1}, x_{t-1}, h_t). \quad (5)$$

The non-Markoviannity of $(h_t)_{t\geq 0}$ is explained via the introduction of new dependencies between x_t and (h_{t-1}, x_{t-1}). Fig. 1 summarizes the different models presented until now.

The rest of this paper is organized as follows. In section 2, we first measure the impact of model (5) as a generative model w.r.t. the SRNN (4). To that end, we study the theoretical distributions $p_\theta(\mathbf{x}_T)$ in the linear and Gaussian framework and show that the PMM can model a larger class of generative distributions. Next, in section 3, we propose a variational Bayesian approach to estimate the parameter θ of general PMMs with a Maximum-Likelihood approach. This approach is particularly suitable for high dimensional PMMs. Finally, in section 4 we describe an example of generative PMM and we compare it with other popular models on simulations.

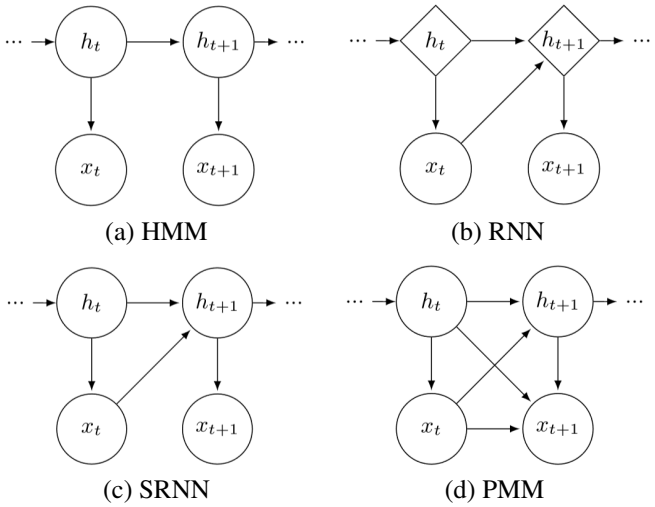

Fig. 1. Generative models based on latent r.v.

2. THEORETICAL MOTIVATIONS

In this section, our objective is to build generative models $p_\theta(\mathbf{x}_T)$ for scalar observations such that it coincides with a Gaussian distribution $p(\mathbf{x}_T)$, for all $T \geq 0$, which satisfies

$$p(x_t) = \mathcal{N}(x_t; 0; 1), \text{ for all } 0 \leq t \leq T \quad (6)$$

($\mathcal{N}(x; \mu, \sigma^2)$ denotes the Gaussian distribution with mean μ, variance σ^2 taken at point x). To that end, we focus on linear and Gaussian PMM with $h_t \in \mathbb{R}$. Model (5) satisfies

$$p_\theta(h_0, x_0) = \mathcal{N}\left(\begin{pmatrix} h_0 \\ x_0 \end{pmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix}\right) \quad (7)$$
$$p_\theta(h_t|h_{t-1}, x_{t-1}) = \mathcal{N}(h_t; ah_{t-1} + cx_{t-1}, \alpha), \quad (8)$$
$$p_\theta(x_t|h_{t-1:t}, x_{t-1}) = \mathcal{N}(x_t; bh_t + eh_{t-1} + fx_{t-1}, \beta), \quad (9)$$

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 HMM also satisfies $c = 0$.

A study in [20] has made a comparison between the linear and Gaussian HMM, RNN and SRNN which satisfy constraint (6). In particular, it has been shown that the linear and Gaussian SRNN ($e = f = 0$, $\gamma = b$ in (7)-(9)) is able to model any centered Gaussian distribution with a geometric Toeplitz covariance matrix,

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}; \Sigma), \quad (10)$$
$$\Sigma(t, t+k) = \mathrm{cov}(x_t, x_{t+k}) = A^{k-1}B, \quad (11)$$

for all $T \geq 0$, $t \geq 0$ and $k \geq 0$. More precisely, $A^{k-1}B$ defines valid covariance sequence if $-1 \leq A \leq 1$ and $\frac{A-1}{2} \leq B \leq \frac{A+1}{2}$, and it has been shown that for a given (A, B) satisfying this condition, it is always possible to find a set of parameters $\theta = (a, b, c, \alpha, \beta, \eta)$ such that $p_\theta(\mathbf{x}_T)$ satisfies (10)-(11).

Now, we study the impact of the new parameters e, f and γ which describe the PMM (7)-(9) w.r.t. the SRNN. We start by adding a transition between h_{t-1} and x_t, i.e. $e \neq 0$. We then have the following original result.

Proposition 1 *Let $p(\mathbf{x}_T)$ be a Gaussian distribution satisfying for all positive integers T, t, k*

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}; \tilde{\Sigma}), \quad (12)$$

$$\text{cov}(x_t, x_{t+k}) = \begin{cases} \tilde{A}^k & \text{if } k \text{ is even} \\ \tilde{A}^{k-1}\tilde{B} & \text{otherwise.} \end{cases}, \quad (13)$$

such that $\tilde{A}$ and $\tilde{B}$ defines a valid covariance matrix. Then for such $(\tilde{A}, \tilde{B})$, it exists a set of parameters $\theta = (a, b, c, e \neq 0, \alpha, \beta, \eta)$ such that $p_\theta(\mathbf{x}_T) = p(\mathbf{x}_T)$.

In other words, this proposition shows that the linear and Gaussian PMM can model some Gaussian distributions which cannot be modeled by the previous linear and Gaussian SRNN. The proof is omitted due to lack of space but proceeds as follows. First, the covariance matrix of $p_\theta(\mathbf{x}_T)$ satisfies (13) with $\tilde{A} = \sqrt{ce}$ and $\tilde{B} = b\Big(c(1-b^2\eta) + e\eta\Big)$. Next, we identify all the valid covariance matrices satisfying (13). This step relies on the Caratheodory theorem [21] from which we deduce the constraints $-1 \leq \tilde{A} \leq 1$, $-\frac{\tilde{A}^2+1}{2} \leq \tilde{B} \leq \frac{\tilde{A}^2+1}{2}$. Finally, we show that for any $\tilde{A}$ and $\tilde{B}$ satisfying these constraints, it is possible to identify a set of parameters θ such that $p_\theta(\mathbf{x}_T) = p(\mathbf{x}_T)$.

Let us finally consider the full PMM case, where all the parameters are considered. This case is more difficult to analyze, but we have the following intermediate result

Proposition 2 *Let $p_\theta(h_t, x_t|h_{t-1}, x_{t-1})$ be a linear and Gaussian PMM described by* (7)-(9) *and satisfying* (6)*. Then the associated generative distribution reads, for all positive integers T, t, k,*

$$p_\theta(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}; \overline{\Sigma}), \quad (14)$$

$$cov(x_t, x_{t+k}) = \overline{A}^k(\overline{B} + \frac{1}{2}) - \overline{C}^k(\overline{B} - \frac{1}{2}), \quad (15)$$

where

$$\overline{A} = \frac{a + bc + f - K}{2}, \quad (16)$$

$$\overline{B} = \frac{a - bc - f - 2\gamma\eta(ab + e)}{2K}, \quad (17)$$

$$\overline{C} = \frac{a + bc + f + K}{2}, \quad (18)$$

$$K = \sqrt{(a + bc + f)^2 - 4(af - ce)} \quad (19)$$

and where the following constraints are satisfied :

$$\gamma\eta = b\eta + (ae + af\gamma + ce\gamma) + fc, \quad (20)$$

$$0 \leq (1 - a^2 - 2ac\gamma)\eta - c^2, \quad (21)$$

$$0 \leq 1 - b^2\eta - 2b\eta(\gamma - b) - e\eta(e + 2f\gamma) - f^2. \quad (22)$$

The proof is omitted due to lack of space. This results generalizes the form of the previous covariance matrices as it can be checked by setting $e = f = 0$ or $e = 0$, for example. At this point of our work, we have not identified if the full linear and Gaussian PMM $p_\theta(\mathbf{x})$ can model any Gaussian distribution with a covariance matrix satisfying (15), except in some particular cases (see e.g. Proposition 1).

3. VARIATIONAL INFERENCE FOR PMMS

The previous section has illustrated the modelling power of PMM in a linear and Gaussian framework. We now focus on general PMMs (i.e. non linear and/or non Gaussian PMMs) and we look for estimating the parameter θ from a realization $\mathbf{x}_T$. Here, we focus on variational Bayesian approaches which are particularly suitable for high dimensional models [18]. Let us first review the rationale of the variational Bayesian estimation.

3.1. Variational Bayesian estimation

If $p_\theta(\mathbf{x}_T)$ aims at the distribution of the observations, maximum likelihood estimation consists in computing the estimate $\hat{\theta}$ which maximizes $p_\theta(\mathbf{x}_T)$. A direct maximization of $p_\theta(\mathbf{x}_T)$ is not always possible, particularly in models with latent variables where the likelihood $p_\theta(\mathbf{x}_T) = \int p_\theta(\mathbf{x}_T, \mathbf{h}_T)\mathrm{d}\mathbf{h}_T$ may be not computable . However, let us remark that for any "variational" distribution $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$,

$$\log(p_\theta(\mathbf{x}_T)) \geq Q(\theta, q_\phi), \quad (23)$$

$$Q(\theta, q_\phi) = -\int \log\left(\frac{q_\phi(\mathbf{h}_T|\mathbf{x}_T)}{p_\theta(\mathbf{x}_T, \mathbf{h}_T)}\right) q_\phi(\mathbf{h}_T|\mathbf{x}_T)\mathrm{d}\mathbf{h}_T. \quad (24)$$

The exact alternate maximization of $Q(\theta, q_\phi)$ w.r.t. θ and q_ϕ coincides with the Expectation-Maximisation (EM) algorithm [22] but relies on the computation of $p_\theta(\mathbf{h}_T|\mathbf{x}_T)$. In the general case, a class of variational distributions $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$ parametrized by ϕ is introduced such that the Evidencial Lower Bound (ELBO) $Q(\theta, \phi)$ is computable or can be approximately and efficiently maximized [23]. A simple way to approximate $Q(\theta, \phi)$ is to choose a parametric distribution $q_\phi(\mathbf{h}_T|\mathbf{x}_T)$ such that a sample $\mathbf{h}_T^{(i)} \sim q(\mathbf{h}_T|\mathbf{x}_T)$ can be written as a differentiable function of ϕ [24]. This technique is called the reparametrization trick.

3.2. Variational inference for PMMs

In the case of the PMM, remember that $p(\mathbf{x}_T, \mathbf{h}_T)$ coincides with (5). Thus, the ELBO in (24) reads

$$Q(\theta, \phi) = -\int \log\left(\frac{q_\phi(h_0|\mathbf{x}_T)}{p(x_0, h_0)}\right) q_\phi(h_0|\mathbf{x}_T)\mathrm{d}\mathbf{h}_T - \sum_{t=1}^{T} \int \log\left(\frac{q_\phi(h_t|h_{t-1}, \mathbf{x}_T)}{p_\theta(h_t, x_t|h_{t-1}, x_{t-1})}\right) q_\phi(\mathbf{h}_t|\mathbf{x}_T)\mathrm{d}\mathbf{h}_t \quad (25)$$

Since $p_\theta(h_t|h_{t-1}, \mathbf{x}_T)$ is generally not computable in PMM models (except in the linear and Gaussian case (7)-(9) where

it is possible to derive a Kalman smoother) we choose a variational distribution which satisfies

$$q_\phi(h_t|h_{t-1}, \mathbf{x}_T) = q_\phi(h_t|h_{t-1}, \mathbf{x}_t), \tag{26}$$

and from which a sample can be obtained with the reparametrization trick. An example of such a variational distribution is

$$q_\phi(h_t|h_{t-1}, \mathbf{x}_t) = \mathcal{N}(h_t; f_\phi(h_{t-1}, \mathbf{x}_t); \mathrm{diag}(g_\phi(h_{t-1}, \mathbf{x}_t))), \tag{27}$$

where f_ϕ and g_ϕ are parametrized and differentiable functions of ϕ, $\mathrm{diag}(.)$ 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

$$h_t^{(i)} = f_\phi(h_{t-1}, \mathbf{x}_t) + (\mathrm{diag}(g_\phi(h_{t-1}, \mathbf{x}_t))^{\frac{1}{2}} \times \epsilon^{(i)}, \tag{28}$$

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 (25) can be approximated by (up to the term associated to $t = 0$ that we omit for clarity)

$$\widehat{Q}(\theta, \phi) = -\sum_{i=1}^N \sum_{t=1}^T \log \left(\frac{q_\phi(h_t^{(i)}|h_{t-1}^{(i)}, \mathbf{x}_t)}{p_\theta(h_t^{(i)}, x_t|h_{t-1}^{(i)}, \mathbf{x}_{t-1})} \right) \tag{29}$$

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

4. SIMULATIONS

4.1. Deep generative PMM model

In this section, we use a similar model as the VRNN [12], which is a particular instance of SRNN. The model is as follows, we set $h_t = (z_t, \overline{h}_t)$ and $p(h_t, x_t|h_{t-1}, x_{t-1})$ in (5) is described with the following set of equations:

$$p_\theta(h_t, x_t|h_{t-1}, x_{t-1}) = p_\theta(x_t|h_{t-1:t}, x_{t-1}) p_\theta(h_t|h_{t-1}, x_{t-1}),$$

$$z_t = f(\psi_x(x_{t-1}), \psi_{\overline{h}}(\overline{h}_{t-1}), z_{t-1}), \tag{30}$$

$$p_\theta(\overline{h}_t|z_{t-1:t}, \overline{h}_{t-1}, x_{t-1}) = \mathcal{N}(\overline{h}_t; \mu_{p\overline{h},t}; \mathrm{diag}(\sigma_{p\overline{h},t})), \tag{31}$$

$$p_\theta(x_t|\overline{h}_{t-1:t}, x_{t-1}, z_{t-1:t}) = \mathrm{Ber}(x_t; \rho_{x,t}), \tag{32}$$

where f is a deterministic non-linear function describring a RNN cell, $[\mu_{.,t}, \sigma_{.,t}]$ and $\rho_{x,t}$ denote the parameters of the Gaussian and Bernoulli distributions respectively, which can be produced by any highly flexible function $\psi_{(.)}$ such as neural networks,

$$[\mu_{p\overline{h},t}, \sigma_{p\overline{h},t}] = \psi_{p\overline{h}}(z_t), \tag{33}$$

$$\rho_{x,t} = \psi_{px}(\psi_{\overline{h}}(\overline{h}_t), \psi_{\overline{h}_1}(\overline{h}_{t-1}), \psi_{x_p}(x_{t-1}), z_t, z_{t-1}). \tag{34}$$

The variational distribution q_ϕ is given by

$$q_\phi(\overline{h}_t|\overline{h}_{t-1}, \mathbf{x}_t) = \mathcal{N}(\overline{h}_t; \mu_{q\overline{h},t}; \mathrm{diag}(\sigma_{q\overline{h},t})), \tag{35}$$

$$[\mu_{q\overline{h},t}, \sigma_{q\overline{h},t}] = \psi_{q\overline{h}}(\psi_x(x_t), z_t). \tag{36}$$

4.2. Results

In the experiments, we used the MNIST data set [25] which contains 60000 (resp. 10000) train (resp. test) 28×28 binary images. An observation x_t consists of a column of the image ($\dim(x_t) = 28$), and the length of a sequence is $T = 28$. Each model was trained with stochastic gradient descend on the negative evidence lower bound using the Adam optimizer [26] with a learning rate of 0.001 and a batch size of 512.

We compare the VRNN with different sub classes of PMM models. The PMM-III is the most general model defined by (30)-(36), where $\psi_{\overline{h}_1}$, $\psi_{\overline{h}}$, ψ_x, ψ_{x_p} are taken as the identity function. For the PMM-II, we have $\psi_{\overline{h}_1} = 0$; for the PMM-I, $\psi_{\overline{h}_1} = 0$ and (34) does not depend on z_{t-1}; finally, the VRNN satisfies $\psi_{\overline{h}_1} = \psi_{x_p} = 0$ and (34) does not depend on z_{t-1}. For each model, $\psi_{p\overline{h}}$, $\psi_{q\overline{h}}$, ψ_{px} have two hidden layers using rectified linear units, with appropriate outputs (linear, softplus and sigmoid). We consider two configurations, the first one has 100 hidden units for each hidden layer. The second one has 100 (resp. 95, 79, 78) hidden units for the VRNN (resp. PMM-I, PMM-II, PMM-III) in order to have a similar number of parameters.

Model	Config. 1		Config. 2	
	ELBO	approx. LL	ELBO	approx. LL
VRNN	-67,248	-64,760	-67,222	-64,762
PMM-I	-66,544	-64,076	-67,322	-64,698
PMM-II	-66,784	-64,201	**-66,815**	**-64,255**
PMM-III	**-66,518**	**-63,876**	-67,513	-64,876

Table 1. Average evidence lower bound (ELBO) and approximated log-likelihood (approx. LL) of the observations on the test set with two different configurations.

The performance of the models is evaluated in terms of the ELBO and approximated log-likelihood of the observations of the test data set based on a particle filtering using 100 samples [27]. In Table 1, we report the average ELBO and the average approximated log-likelihood (approx. LL) on the test set assigned by our models. The results with the Config.1 (resp. Config. 2) show that PMM-III (resp. PMM-II) has the higher average ELBO and average approx LL, in general, higher is better. As we see, the PMM performs better than VRNN. Due to lack of space, additional results on other data set are not presented here.

5. CONCLUSIONS

In this paper, we have included popular generative models for times series modelling into a common model, the PMM. We have shown that this model may be more relevant to model complex distribution w.r.t. the SRNN, and the increase of modelling power has been measured in the linear and Gaussian case. For general PMM models, a parameter estimation algorithm has been provided.

6. REFERENCES

[1] R.-H. Shumway and D.-S. Stoffer, *Time Series Analysis and its Applications*, Springer-Verlag, Berlin, Heidelberg, 2005.

[2] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[3] A. C. Harvey, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, 1989.

[4] L. C. Thomas, D. E. Allen, and N. Morkel-Kingsbury, "A Hidden Markov chain model for the term structure of bond credit risk spreads," *International Review of Financial Analysis*, vol. 11, no. 3, pp. 311–29, 2002.

[5] 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.

[6] H. White, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, vol. 50, no. 1, pp. 1–25, January 1982.

[7] R. Douc, E. Moulines, and T. Ryden, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime," *Annals of Statistics*, vol. 32, no. 5, pp. 2254–2304, 2004.

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

[9] 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.

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

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

[12] 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.

[13] W. Pieczynski, "Pairwise Markov chains," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 634–39, May 2003.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] O. Cappé, S. J. Godsill, and É. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proc. of the IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.

[20] 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.

[21] N.-I. Akhiezer and N Kemmer, *The classical moment problem and some related questions in analysis*, vol. 5, Oliver & Boyd Edinburgh, 1965.

[22] 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.

[23] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[24] 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*, 2014.

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

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

[27] F. Desbouvries and W. Pieczynski, "Particle filtering in pairwise and triplet Markov chains," in *Proc. IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing*, Grado-Gorizia, Italy, June 8-11 2003.