A Comparative Study of Gamma Markov Chains for Temporal Non-Negative Factorization

Louis Filstroff, Olivier Gouvert, Cédric Févotte, Olivier Cappé

A Comparative Study of Gamma Markov Chains for Temporal Non-Negative Factorization

Louis Filstroff [1] Olivier Gouvert [2] Cédric Févotte [3] Olivier Cappé [4]

[1] Department of Computer Science, School of Science, Aalto University, Finland

[2] Mila - Quebec Artificial Intelligence Institute

[3] IRIT, Université de Toulouse, CNRS, France

[4] DI ENS, CNRS, INRIA, Université PSL

February 17, 2021

Non-negative matrix factorization (NMF) has become a well-established class of methods for the analysis of non-negative data. In particular, a lot of effort has been devoted to probabilistic NMF, namely estimation or inference tasks in probabilistic models describing the data, based for example on Poisson or exponential likelihoods. When dealing with time series data, several works have proposed to model the evolution of the activation coefficients as a non-negative Markov chain, most of the time in relation with the Gamma distribution, giving rise to so-called temporal NMF models. In this paper, we review four Gamma Markov chains of the NMF literature, and show that they all share the same drawback: the absence of a well-defined stationary distribution. We then introduce a fifth process, an overlooked model of the time series literature named BGAR(1), which overcomes this limitation. These temporal NMF models are then compared in a MAP framework on a prediction task, in the context of the Poisson likelihood.

Keywords: Non-negative matrix factorization, Time series data, Gamma Markov chains, MAP estimation

1. Introduction

1.1. Non-negative matrix factorization

Non-negative matrix factorization (NMF) (Paatero and Tapper, 1994; Lee and Seung, 1999) has become a widely used class of methods for analyzing non-negative data. Let us

consider N samples in $\mathbb{R}_+^F$. We can store these samples column-wise in a matrix, which we denote by $\mathbf{V}$ (therefore of size $F \times N$). Broadly speaking, NMF aims at finding an approximation of $\mathbf{V}$ as the product of two non-negative matrices:

$$\mathbf{V} \simeq \mathbf{W}\mathbf{H}, \tag{1}$$

where $\mathbf{W}$ is of size $F \times K$, and $\mathbf{H}$ is of size $K \times N$. $\mathbf{W}$ and $\mathbf{H}$ are referred to as the dictionary and the activation matrix, respectively. The factorization rank K is usually chosen such that $K \ll \min(F, N)$, hence producing a low-rank approximation of $\mathbf{V}$. This factorization is often retrieved as the solution of an optimization problem, which we can write as:

$$\min_{\mathbf{W} \geq 0,\ \mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{W}\mathbf{H}), \tag{2}$$

where D is a measure of fit between $\mathbf{V}$ and its approximation $\mathbf{W}\mathbf{H}$, and the notation $\mathbf{A} \geq 0$ denotes the non-negativity of the entries of the matrix $\mathbf{A}$. One of the key aspects to the success of NMF is that the non-negativity of the factors $\mathbf{W}$ and $\mathbf{H}$ yields an interpretable, part-based representation of each sample: $\mathbf{v}_n \simeq \mathbf{W}\mathbf{h}_n$ (Lee and Seung, 1999).

Various measures of fit have been considered in the literature, for instance the family of β-divergences (Févotte and Idier, 2011), which includes some of the most popular cost functions in NMF, such as the squared Euclidian distance, the generalized Kullback-Leibler divergence, or the Itakura-Saito divergence. As it turns out, for many of these cost functions, the optimization problem described in Eq. (2) can be shown to be equivalent to the joint maximum likelihood estimation of the factors $\mathbf{W}$ and $\mathbf{H}$ in a statistical model, that is:

$$\max_{\mathbf{W},\mathbf{H}} p(\mathbf{V}|\mathbf{W}, \mathbf{H}). \tag{3}$$

This leads the way to so-called *probabilistic* NMF, i.e., estimation or inference tasks in probabilistic models whose observation distribution may be written as:

$$\mathbf{v}_n \sim p(\ .\ ; \mathbf{W}\mathbf{h}_n, \boldsymbol{\Theta}), \quad \mathbf{W} \geq 0, \quad \mathbf{H} \geq 0, \tag{4}$$

that is to say that the distribution of $\mathbf{v}_n$ is parametrized by the dot product of the factors $\mathbf{W}$ and $\mathbf{h}_n$. Other potential parameters of the distribution are generically denoted by $\boldsymbol{\Theta}$. Most of the time these distributions are such that $\mathbb{E}(\mathbf{v}_n) = \mathbf{W}\mathbf{h}_n$.

This large family encompasses many well-known models of the literature, for example models based on the Gaussian likelihood (Schmidt et al., 2009) or the exponential likelihood (Févotte et al., 2009; Hoffman et al., 2010). It also includes factorization models for count data, which are most of the time based on the Poisson distribution[1] (Canny, 2004; Cemgil, 2009; Zhou et al., 2012; Gopalan et al., 2015), but can also make use of distributions with a larger tail, e.g., the negative binomial distribution (Zhou, 2018). Finally, more complex models using the compound Poisson distribution have been considered

[1] These models are sometimes generically referred to as "Poisson factorization" or "Poisson factor analysis".

(Şimşekli et al., 2013; Basbug and Engelhardt, 2016; Gouvert et al., 2019), allowing to extend the use of the Poisson distribution to various supports $(\mathbb{N}, \mathbb{R}_+, \mathbb{R}, \dots)$.

In the vast majority of the aforementioned works, prior distributions are assumed on the factors $\mathbf{W}$ and $\mathbf{H}$. This is sometimes referred to as *Bayesian* NMF. In this case, the columns of $\mathbf{H}$ are most of the time assumed to be independent:

$$p(\mathbf{H}) = \prod_{n=1}^{N} p(\mathbf{h}_n). \tag{5}$$

The factors being non-negative, a standard choice is the Gamma distribution[2], which can be sparsity-inducing if the shape parameter is chosen to be lower than one. The inverse Gamma distribution has also been considered.

1.2. Temporal structure of the activation coefficients

In this work, we are interested in the analysis of specific matrices $\mathbf{V}$ whose columns cannot be treated as exchangeable, because the samples $\mathbf{v}_n$ are correlated. Such a scenario arises in particular when the columns of $\mathbf{V}$ describe the evolution of a process over time.

From a modeling perspective, this means that correlation should be introduced in the statistical model between successive columns of $\mathbf{V}$. This can be achieved by lifting the prior independence assumption of Eq. (5), thus introducing correlation between successive columns of $\mathbf{H}$. In this paper, we consider a Markov structure on the columns of $\mathbf{H}$:

$$p(\mathbf{H}) = p(\mathbf{h}_1) \prod_{n \geq 2} p(\mathbf{h}_n | \mathbf{h}_{n-1}). \tag{6}$$

We will refer to such a model as a *dynamical* NMF model. Note that recent works go beyond the Markovian assumption, i.e., assume dependency with multiple past time steps, and are labeled as "deep" (Gong and Huang, 2017; Guo et al., 2018).

Several works (Févotte et al., 2013; Schein et al., 2016, 2019) assume that the transition distribution $p(\mathbf{h}_n|\mathbf{h}_{n-1})$ makes use of a transition matrix $\mathbf{\Pi}$ of size $K \times K$ to capture relationships between the different components. In this case, the distribution of h_{kn} depends on a linear combination of all the components at the previous time step:

$$p(\mathbf{h}_n|\mathbf{h}_{n-1}) = \prod_k p(h_{kn}|\sum_l \pi_{kl} h_{l(n-1)}). \tag{7}$$

In this work, we will restrict ourselves to $\mathbf{\Pi} = \mathbf{I}_K$. Equivalently, this amounts to assuming that the K rows of $\mathbf{H}$ are a priori independent, and we have

$$p(\mathbf{H}) = \prod_k p(h_{k1}) \prod_{n \geq 2} p(h_{kn}|h_{k(n-1)}). \tag{8}$$

[2] Throughout the article, we consider the "shape and rate" parametrization of the Gamma distribution, i.e. $\text{Gamma}(x|\alpha, \beta) \propto x^{\alpha-1} \exp(-\beta x)$.

We will refer to such a model as a *temporal* NMF model.

A first way of dealing with the temporal evolution of a non-negative variable is to map it to $\mathbb{R}_+$. It is then commonly assumed that this variable evolves in Gaussian noise. This is for example exploited in the seminal work of Blei and Lafferty (2006) on the extension of latent Dirichlet allocation to allow for topic evolution[3]. A similar assumption is made in Charlin et al. (2015), which introduces dynamics in the context of a Poisson likelihood (factorizing the user-item-time tensor). Gaussian assumptions allow to use well-known computational techniques, such as Kalman filtering, but result in loss of interpretability.

We will focus in this paper on naturally non-negative Markov chains. Various non-negative Markov chains have been proposed in the NMF literature (see Section 2 and references therein). They are all built in relation with the Gamma (or inverse Gamma) distribution. As a matter of fact, these models exhibit the same drawback: the chains all have a degenerate stationary distribution. This can lead to undesirable behaviors, such as the instability or the degeneracy of realizations of the chains. We emphasize that this is problematic from the probabilistic perspective only, since these prior distributions may still represent an appropriate regularization in a MAP setting.

1.3. Contributions and organization of the paper

The contributions of this paper are 4-fold:

- We review the existing non-negative Markov chains of the NMF literature and discuss some of their limitations. In particular we show that these chains all have a degenerate stationary distribution;
- We present an overlooked non-negative Markov chain from the time series literature, the first-order autoregressive Beta-Gamma process, denoted as BGAR(1) (Lewis et al., 1989), whose stationary distribution is Gamma. To the best of our knowledge, this particular chain has never been considered to model temporal dependencies in matrix factorization problems;
- We derive majorization-minimization-based algorithms for maximum a posteriori (MAP) estimation in the NMF models (with a Poisson likelihood) with four of the presented prior structures on $\mathbf{H}$, including BGAR(1);
- We compare the performance of all these models on a prediction task on three real-world datasets.

The paper is organized as follows. Section 2 introduces and compares non-negative Markov chains from the literature. Section 3 presents MAP estimation in temporal NMF models. Experimental work is conducted in Section 4, before concluding in Section 5.

[3] Note that this particular mapping is actually slightly more complex, as the K-dimensional real vector must be mapped to the $(K-1)$ simplex due to further constraints in the model.

2. Comparative study of Gamma Markov chains

This section reviews existing models of Gamma Markov chains, i.e., Markov chains which evolve in $\mathbb{R}_+$ in relation with the Gamma distribution. We have identified four different models in the NMF literature:

1. Chaining on the rate parameter of a Gamma distribution (Section 2.1);
2. Chaining on the rate parameter of a Gamma distribution with an auxiliary variable (Section 2.2);
3. Chaining on the shape parameter of a Gamma distribution (Section 2.3);
4. Chaining on the shape parameter of a Gamma distribution with an auxiliary variable (Section 2.4).

As will be discussed in these subsections, these four models all lack a well-defined stationary distribution, which leads to the degeneracy of the realizations of the chains. A fifth model from the time series literature, called BGAR(1), is presented in Section 2.5. It is built to have a well-defined stationary distribution (it is marginally Gamma distributed). The realizations of the chain are not degenerate and exhibit some interesting properties. To the best of our knowledge, this kind of process has never been used in a probabilistic NMF problem to model temporal evolution.

Throughout the section, $(h_n)_{n\geq 1}$ denotes the (scalar) Markov chain of interest, where the index k as in Eq. (8) has been dropped for enhanced readability. It is further assumed that h_1 is set to a fixed, deterministic value.

2.1. Chaining on the rate parameter

2.1.1. Model

Let us consider a general Gamma Markov chain model with a chaining on the rate parameter:

$$h_n|h_{n-1} \sim \text{Gamma}\left(\alpha, \frac{\beta}{h_{n-1}}\right). \tag{9}$$

As it turns out, Eq. (9) can be rewritten as a multiplicative noise model:

$$h_n = h_{n-1} \times \phi_n, \tag{10}$$

where ϕ_n are i.i.d. Gamma random variables with parameters (α, β). We have

$$\mathbb{E}(h_n|h_{n-1}) = \frac{\alpha}{\beta}h_{n-1}, \quad \text{var}(h_n|h_{n-1}) = \frac{\alpha}{\beta^2}h_{n-1}^2. \tag{11}$$

This model was introduced in Févotte et al. (2009) to add smoothness to the activation coefficients in the context of audio signal processing. The parameters were set to $\alpha > 1$ and $\beta = \alpha - 1$, such that the mode would be located at $h_n = h_{n-1}$. It is also a particular case of the dynamical model of Févotte et al. (2013) and is considered in Virtanen and Girolami (2020). A similar inverse Gamma Markov chain was also considered in Févotte et al. (2009) and in Févotte (2011).

2.1.2. Analysis

From Eq. (10) we can write:

$$h_n = h_1 \prod_{i=2}^{n} \phi_i. \tag{12}$$

The independence of the ϕ_i yields:

$$\mathbb{E}(h_n) = h_1 \left(\frac{\alpha}{\beta}\right)^{n-1}, \tag{13}$$

$$\text{var}(h_n) = h_1^2 \left[\left(\frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2}\right)^{n-1} - \left(\frac{\alpha^2}{\beta^2}\right)^{n-1}\right]. \tag{14}$$

We enumerate all the possible regimes ($n \to +\infty$), which all give rise to degenerate stationary distributions for different reasons:

- $\beta > \sqrt{\alpha(\alpha+1)}$: both mean and variance go to zero;
- $\beta = \sqrt{\alpha(\alpha+1)}$: variance converges to 1, however the mean goes to zero;
- $\beta \in \left]\alpha; \sqrt{\alpha(\alpha+1)}\right[$: variance goes to infinity, mean goes to zero;
- $\beta = \alpha$: mean is equal to 1, but the variance goes to infinity;
- $\beta < \alpha$: both mean and variance go to infinity.

Each subplot of Figure 1 displays ten independent realizations of the chain, for a different set of parameters (α, β). As we can see, the realizations of the chain either collapse to 0, or diverge.

2.2. Hierarchical chaining on the rate parameter

2.2.1. Model

Let us consider the following Gamma Markov chain model introduced in Cemgil and Dikmen (2007):

$$z_n | h_{n-1} \sim \text{Gamma}(\alpha_z, \beta_z h_{n-1}), \tag{15}$$

$$h_n | z_n \sim \text{Gamma}(\alpha_h, \beta_h z_n). \tag{16}$$

As it turns out, this model can also be rewritten as a multiplicative noise model:

$$h_n = h_{n-1} \times \tilde{\phi}_n, \tag{17}$$

where $\tilde{\phi}_n$ are i.i.d. random variables defined as the ratio of two independent Gamma random variables with parameters (α_h, β_h) and (α_z, β_z). The distribution of $\tilde{\phi}_n$ is actually known in closed form, namely

$$\tilde{\phi}_n \sim \text{BetaPrime}\left(\alpha_h, \alpha_z, 1, \tilde{\beta}\right), \tag{18}$$

with $\tilde{\beta} = \frac{\beta_z}{\beta_h}$ (see Appendix A for a definition). We have

$$\mathbb{E}(h_n|h_{n-1}) = \tilde{\beta}\frac{\alpha_h}{\alpha_z - 1}h_{n-1} \qquad \text{for } \alpha_z > 1, \tag{19}$$

$$\text{var}(h_n|h_{n-1}) = \tilde{\beta}^2\frac{\alpha_h(\alpha_h + \alpha_z - 1)}{(\alpha_z - 1)^2(\alpha_z - 2)}h_{n-1}^2 \qquad \text{for } \alpha_z > 2. \tag{20}$$

This model is less straightforward in its construction than the previous one, as it makes use of an auxiliary variable z_n (note that a similar inverse Gamma construction was proposed as well in Cemgil and Dikmen (2007)). There are two motivations behind the introduction of this auxiliary variable:

1. Firstly, it ensures what is referred to as "positive correlation" in Cemgil and Dikmen (2007), i.e., $\mathbb{E}(h_n|h_{n-1}) \propto h_{n-1}$ (something the model described by Eq. (9) does as well).

2. Secondly, it ensures the so-called conjugacy of the model, i.e., the conditional distributions $p(z_n|h_{n-1}, h_n)$ and $p(h_n|z_n, z_{n+1})$ remain Gamma distributions. Indeed, these are the distributions of interest when considering Gibbs sampling or variational inference. This property is not satistfied by the model described by Eq. (9) (i.e., $p(h_n|h_{n-1}, h_{n+1})$ is neither Gamma, nor a known distribution).

This particular chain has been used in the context of audio signal processing in Virtanen et al. (2008) (under the assumption of a Poisson likelihood, which does not fit the nature of the data), and also to model the evolution of user and item preferences in the context of recommender systems (Jerfel et al., 2017; Do and Cao, 2018).

2.2.2. Analysis

From Eq. (17), we can write:

$$h_n = h_1\prod_{i=2}^{n}\tilde{\phi}_i. \tag{21}$$

We have by independence of the $\tilde{\phi}_i$:

$$\mathbb{E}(h_n) = h_1\left(\tilde{\beta}\frac{\alpha_h}{\alpha_z - 1}\right)^{n-1} \qquad \text{for } \alpha_z > 1, \tag{22}$$

$$\text{var}(h_n) = h_1^2\tilde{\beta}^{2(n-1)}\left[\left(\frac{\alpha_h^2}{(\alpha_z - 1)^2} + \frac{\alpha_h(\alpha_h + \alpha_z - 1)}{(\alpha_z - 1)^2(\alpha_z - 2)}\right)^{n-1} - \left(\frac{\alpha_h^2}{(\alpha_z - 1)^2}\right)^{n-1}\right] \quad \text{for } \alpha_z > 2. \tag{23}$$

As in the previous model, we can show that either the expectation or the variance diverges or collapses as $n \to \infty$ for every possible choice of parameters, which means

that they all give rise to a degenerate stationary distribution of the chain. Each subplot of Figure 2 displays ten independent realizations of the chain, for a different set of parameters $(\alpha_z, \beta_z, \alpha_h, \beta_h)$. As we can see, the realizations of the chain either collapse to 0 or diverge.

2.3. Chaining on the shape parameter

2.3.1. Model

Let us consider a general Gamma Markov chain model with a chaining on the shape parameter:

$$h_n | h_{n-1} \sim \text{Gamma}(\alpha h_{n-1}, \beta). \tag{24}$$

We have

$$\mathbb{E}(h_n | h_{n-1}) = \frac{\alpha}{\beta} h_{n-1}, \quad \text{var}(h_n | h_{n-1}) = \frac{\alpha}{\beta^2} h_{n-1}. \tag{25}$$

In contrast with the two models presented previously, this model cannot be rewritten as a multiplicative noise model. This model is therefore more intricate to interpret. It was introduced in Acharya et al. (2015) in the context of Poisson factorization. It is mainly motivated by a data augmentation trick that can be used when working with a Poisson likelihood, which enables a Gibbs sampling procedure. The authors set the value of α to 1 (although the same trick can be applied for any value of α). This model is also a particular case of the dynamical model of Schein et al. (2016). It has since been used in the context of topic modeling (Acharya et al., 2018).

2.3.2. Analysis

Using the law of total expectation and total variance, it can be shown that

$$\mathbb{E}(h_n) = h_1 \left(\frac{\alpha}{\beta}\right)^{n-1}, \ \text{var}(h_n) = h_1 \frac{1}{\beta} \left(\frac{\alpha}{\beta}\right)^{n-1} \sum_{i=0}^{n-2} \left(\frac{\alpha}{\beta}\right)^{i}. \tag{26}$$

The discussion is hence driven by the value of $r = \alpha/\beta$.

- If $r < 1$, mean and variance go to zero;
- If $r = 1$, mean is fixed but variance goes to infinity (linearly);
- If $r > 1$, mean and variance go to infinity.

This chain only exhibits degenerate stationary distributions. Each subplot of Figure 3 displays ten independent realizations of the chain, for a different set of parameters (α, β). As we can see, the realizations of the chain either collapse to 0, or diverge.

2.4. Hierarchical chaining on the shape parameter

2.4.1. Model

Let us consider the following Gamma Markov chain model

$$z_n|h_{n-1} \sim \text{Poisson}(\beta h_{n-1}), \tag{27}$$
$$h_n|z_n \sim \text{Gamma}(\alpha + z_n, \beta). \tag{28}$$

This model is a particular case of the dynamical model firstly introduced in Schein et al. (2019). It cannot be rewritten as a multiplicative noise model. Using the law of total expectation and total variance, we obtain

$$\mathbb{E}(h_n|h_{n-1}) = h_{n-1} + \frac{\alpha}{\beta}, \tag{29}$$
$$\text{var}(h_n|h_{n-1}) = \frac{2}{\beta}h_{n-1} + \frac{\alpha}{\beta^2}. \tag{30}$$

The motivation behind the introduction of this model is once again computational: it leads to closed-form conditional distributions when considering a Poisson likelihood. As stated in Schein et al. (2019), the auxiliary variable z_n can actually be marginalized out, leading to the so-called randomized Gamma distribution of the first type (RG1), whose analytical expression makes use of modified Bessel functions.

The authors also consider the particular limit case $\alpha = 0$, which leads here to a chain which will only take value 0 after obtaining $z_n = 0$.

2.4.2. Analysis

Using the law of total expectation and total variance, it can be shown that

$$\mathbb{E}(h_n) = h_1 + (n-1)\frac{\alpha}{\beta}, \tag{31}$$
$$\text{var}(h_n) = (n-1)\frac{2}{\beta}h_1 + (n-1)^2\frac{\alpha}{\beta^2}. \tag{32}$$

As such, for $\alpha, \beta > 0$, when $n \to +\infty$, both the expectation and variance of h_n diverge, leading to a degenerate stationary distribution of the chain. Each subplot of Figure 4 displays ten independent realizations of the chain, for a different set of parameters (α, β).

2.5. BGAR(1)

We now discuss the first order autoregressive Beta-Gamma process of Lewis et al. (1989), a stochastic process which is marginally Gamma distributed. The authors referred to the process as "BGAR(1)". However, to the best of our knowledge, no extension to higher-order autoregressive processes exists in the time series literature. As such, from now on, we will simply refer to it as "BGAR".

2.5.1. Model

Consider $\alpha > 0$, $\beta > 0$, $\rho \in [0, 1[$. The BGAR process is defined as:

$$h_1 \sim \text{Gamma}(\alpha, \beta), \tag{33}$$
$$h_n = b_n h_{n-1} + \epsilon_n \qquad \text{for } n \geq 2, \tag{34}$$

where b_n and ϵ_n are i.i.d. random variables distributed as:

$$b_n \sim \text{Beta}(\alpha\rho, \alpha(1-\rho)), \tag{35}$$
$$\epsilon_n \sim \text{Gamma}(\alpha(1-\rho), \beta). \tag{36}$$

The sequence $(h_n)_{n\geq 1}$ is called the BGAR process. It is parametrized by α, β and ρ. We emphasize that the distribution $p(h_n|h_{n-1})$ is not known in closed form. Only $p(h_n|h_{n-1}, b_n)$ is known; it is a shifted Gamma distribution. The generative model may therefore be rewritten as

$$h_1 \sim \text{Gamma}(\alpha, \beta), \tag{37}$$
$$b_n \sim \text{Beta}(\alpha\rho, \alpha(1-\rho)) \quad \text{for } n \geq 2, \tag{38}$$
$$h_n|b_n, h_{n-1} \sim \text{Gamma}(\alpha(1-\rho), \beta, \text{loc} = b_n h_{n-1}) \quad \text{for } n \geq 2, \tag{39}$$

where the distribution in Eq. (39) is a shifted Gamma distribution with a location parameter "loc".

We have

$$\mathbb{E}(h_n|h_{n-1}) = \rho h_{n-1} + \frac{\alpha(1-\rho)}{\beta}, \tag{40}$$
$$\text{var}(h_n|h_{n-1}) = \frac{\rho(1-\rho)}{\alpha+1} h_{n-1}^2 + \frac{\alpha(1-\rho)}{\beta^2}. \tag{41}$$

2.5.2. Analysis

To study the marginal distribution of the process, we recall the following lemma.

Lemma 1. If $X \sim \text{Beta}(a, b)$ and $Y \sim \text{Gamma}(a+b, c)$ are independent random variables, then $Z = XY$ is $\text{Gamma}(a, c)$ distributed.

Proposition 1. Let $(h_n)_{n\geq 1}$ be a BGAR process. Then h_n is marginally $\text{Gamma}(\alpha, \beta)$ distributed.

Proof. Follows by induction. Consider n such that h_n is $\text{Gamma}(\alpha, \beta)$ distributed. Then, $\epsilon_{n+1} h_n$ is $\text{Gamma}(\alpha\rho, \beta)$ distributed (Lemma 1). Finally, $h_{n+1} = \epsilon_{n+1} h_n + b_{n+1}$ is $\text{Gamma}(\alpha, \beta)$ distributed (sum of independent Gamma random variables), which concludes the proof. □

Therefore the parameters α and β control the marginal distribution. The parameter ρ controls the correlation between successive values, as discussed in the following proposition.

Proposition 2. Let $(h_n)_{n\geq 1}$ be a BGAR process. Let n and r be two integers such that $r > 1$. We have $\text{corr}(h_n, h_{n+r}) = \rho^r$.

Proof. See Appendix B for $r = 1$. □

Proposition 2 implies that the BGAR(1) process admits a (second order) AR(1) representation. Two limit cases of BGAR can be exhibited:

- When $\rho = 0$, the h_n are i.i.d. random variables;
- When $\rho \to 1$, the process is not random anymore, and $h_n = h_1$ for all n (note that $\rho = 1$ is not an admissible value).

Finally, from Eq. (40), we have

$$\left(\mathbb{E}(h_n|h_{n-1}) > h_{n-1}\right) \Leftrightarrow \left(h_{n-1} < \frac{\alpha}{\beta}\right). \tag{42}$$

If h_{n-1} is below the mean of the marginal distribution $\mathbb{E}(h_n) = \frac{\alpha}{\beta}$, then h_n will be in expectation above h_{n-1}, and vice-versa.

Note that BGAR is not the only Markovian process with a marginal Gamma distribution considered in the literature. We mention the GAR(1) process (first-order autoregressive Gamma process) of Gaver and Lewis (1980), which is also marginally Gamma distributed. However, this particular process is piecewise deterministic, and its parameters are "coupled": the parameters of the marginal distribution also have an influence on other properties of the model. As such, it is less suited to our problem, and will not be considered here.

Figure 5 displays three realizations of the BGAR process, with parameters fixed to $\alpha = 2$ and $\beta = 1$, and a different parameter ρ in each subplot. The mean of the marginal distribution is displayed in red. When $\rho = 0.5$, the correlation is weak, and no particular structure is observed. However, as ρ goes to 1, the correlation becomes stronger, and we typically observe piecewise constant trajectories.

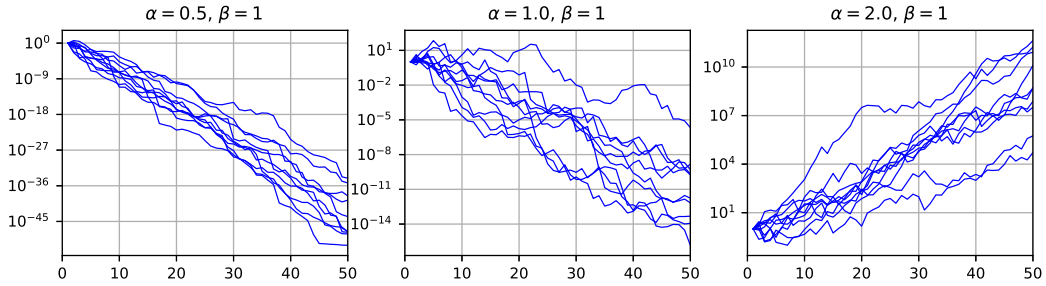

Figure 1: Realizations of the Markov chain defined in Eq. (9). The initial value h_1 is set to 1, and chains were simulated until $n = 50$. Each subplot contains ten independent realizations, with the value of the parameters (α, β) given at the top of the subplot. $\log_{10}(h_n)$ is displayed.

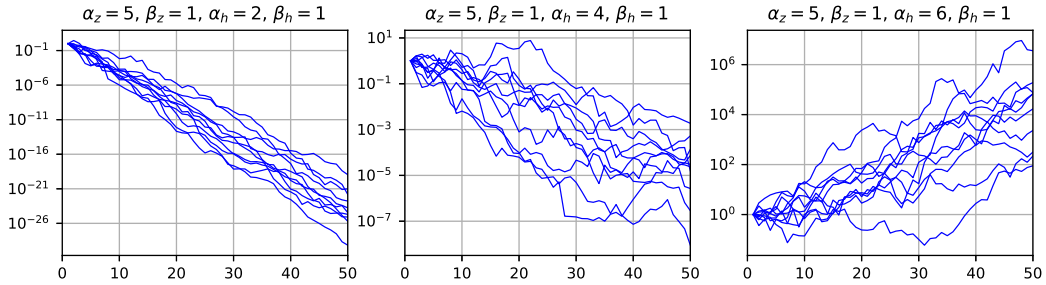

Figure 2: Realizations of the Markov chain defined in Eq. (15)-(16). The initial value h_1 is set to 1, and chains were simulated until $n = 50$. Each subplot contains ten independent realizations, with the value of the parameters $(\alpha_z, \beta_z, \alpha_h, \beta_h)$ given at the top of the subplot. $\log_{10}(h_n)$ is displayed.

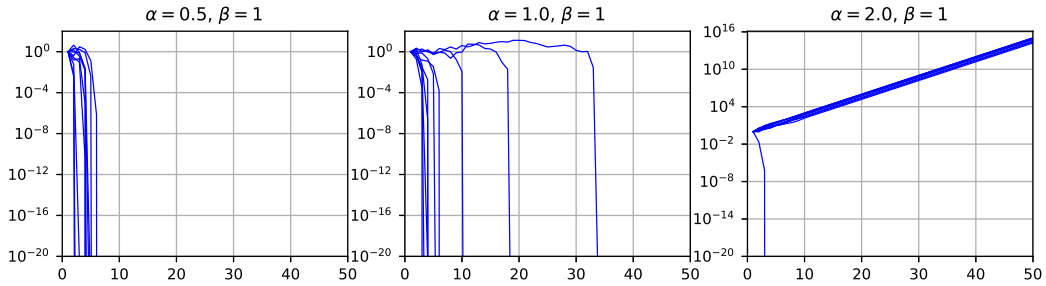

Figure 3: Realizations of the Markov chain defined in Eq. (24). The initial value h_1 is set to 1, and chains were simulated until $n = 50$. Each subplot contains ten independent realizations, with the value of the parameters (α, β) given at the top of the subplot. $\log_{10}(h_n)$ is displayed.

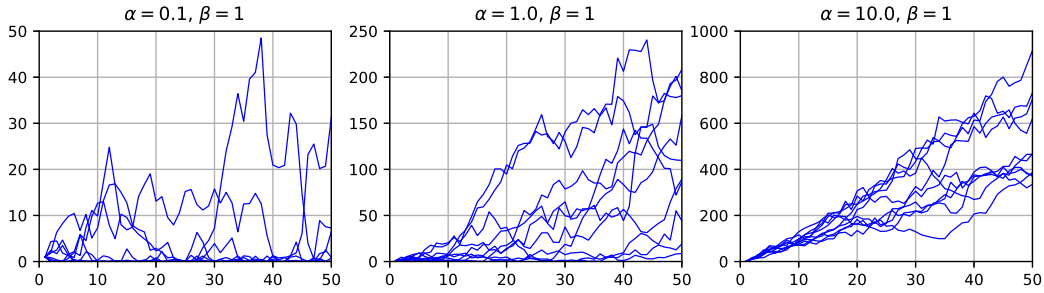

Figure 4: Realizations of the Markov chain defined in Eq. (27)-(28). The initial value h_1 is set to 1, and chains were simulated until $n = 50$. Each subplot contains ten independent realizations, with the value of the parameters (α, β) given at the top of the subplot. $\log_{10}(h_n)$ is displayed.

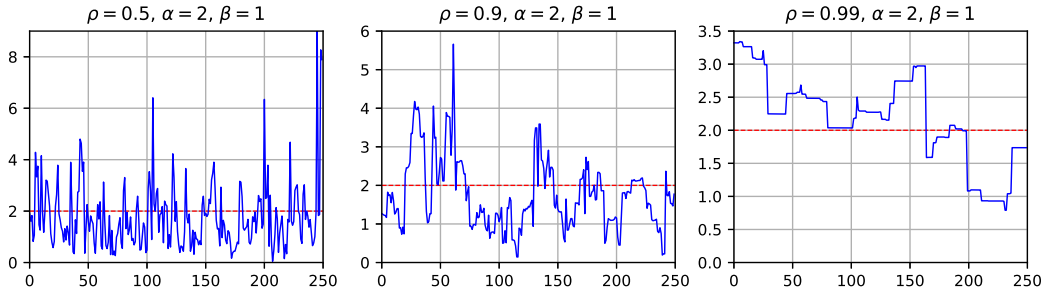

Figure 5: Three realizations of the BGAR(1) process, with parameters fixed to $\alpha = 2$ and $\beta = 1$, and a different parameter ρ in each subplot. The mean of the process is displayed by a dashed red line.

3. MAP estimation in temporal NMF models

We now turn to the problem of maximum a posteriori (MAP) estimation in temporal NMF models. More precisely, we assume a Poisson likelihood, that is

$$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn}), \tag{43}$$

and we also assume that $\mathbf{W}$ is a deterministic variable. The variables $\mathbf{V}$ and $\mathbf{H}$ then define a hidden Markov model, as displayed on Figure 6.

We consider four different models corresponding to the temporal structures on $\mathbf{H}$ presented in subsections 2.1, 2.2, 2.3, and 2.5. Only the temporal structure presented in 2.4 is left out. Indeed, deriving a MAP algorithm in this model using the auxiliary variables $\mathbf{Z}$ (similar to the one of Section 3.3) would involve integer programming. This leads to technical developments which are out-of-scope of our current study.

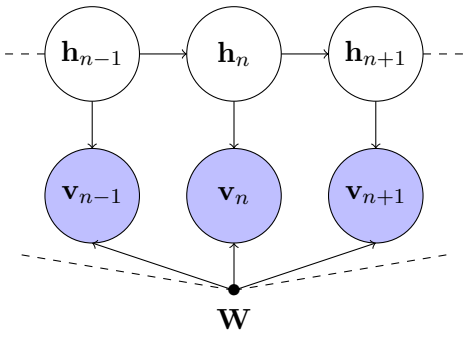

Figure 6: Hidden Markov model arising in temporal NMF models. $\mathbf{v}_n$ is of dimension F, while $\mathbf{h}_n$ is of dimension K. Observed variables are in blue.

$$C(\mathbf{W}, \mathbf{H}) = -\log p(\mathbf{V}, \mathbf{H}; \mathbf{W}) \tag{44}$$

$$= -\log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) - \sum_k \left[\log p(h_{k1}) + \sum_{n \geq 2} \log p(h_{kn}|h_{k(n-1)}) \right], \tag{45}$$

that is to say that the factors $\mathbf{W}$ and $\mathbf{H}$ are going to be estimated. Both shape hyperparameters (α_k or ρ_k) and scale hyperparameters (β_k) will be treated as fixed and selected using a validation set. However, note that deriving the maximum likelihood estimate of β_k is feasible in closed form for all the models presented below. Unfortunately, estimating β_k this way led to overly flat priors in our experience (likely due to the MAP estimation setting).

The optimization of the function C is carried out with a block coordinate descent scheme over the variables $\mathbf{W}$ and $\mathbf{H}$. We resort to a majorization-minimization (MM) scheme, which consists in iteratively majorizing the function C (by a so-called auxiliary function, tight for some $\tilde{\mathbf{W}}$ or $\tilde{\mathbf{H}}$), and minimizing this auxiliary function instead.

We refer the reader to Hunter and Lange (2004) for a detailed tutorial. Under this scheme, the function C is non-increasing. As it turns out, only the Poisson likelihood term $-\log p(\mathbf{V}|\mathbf{H};\mathbf{W})$ needs to be majorized. This is a well-studied issue in the NMF literature. As stated in Lee and Seung (2000); Févotte and Idier (2011), the function

$$G_1(\mathbf{H};\tilde{\mathbf{H}}) = -\sum_{k,n} p_{kn}\log(h_{kn}) + \sum_{k,n} q_k h_{kn}, \tag{46}$$

with the notations

$$p_{kn} = \tilde{h}_{kn}\sum_f w_{fk}\frac{v_{fn}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}}, \quad q_k = \sum_f w_{fk}, \tag{47}$$

is a tight auxiliary function of $-\log p(\mathbf{V}|\mathbf{H};\mathbf{W})$ at $\mathbf{H}=\tilde{\mathbf{H}}$. Similarly the function

$$G_2(\mathbf{W};\tilde{\mathbf{W}}) = -\sum_{f,k} p'_{fk}\log(w_{fk}) + \sum_{f,k} q'_k w_{fk}, \tag{48}$$

with the notations

$$p'_{fk} = \tilde{w}_{fk}\sum_n h_{kn}\frac{v_{fn}}{[\tilde{\mathbf{W}}\mathbf{H}]_{fn}}, \quad q'_k = \sum_n h_{kn}, \tag{49}$$

is a tight auxiliary function of $-\log p(\mathbf{V}|\mathbf{H};\mathbf{W})$ at $\mathbf{W}=\tilde{\mathbf{W}}$.

3.1. Minimization w.r.t. W

The optimization w.r.t. $\mathbf{W}$ is common to all algorithms, and amounts to minimizing $G_2(\mathbf{W};\tilde{\mathbf{W}})$ only. The scale of $\mathbf{W}$ must be however be fixed in order to prevent potential degenerate solutions such that $\mathbf{W}\to+\infty$ and $\mathbf{H}\to 0$. Indeed, consider $\mathbf{W}^\star$ and $\mathbf{H}^\star$ minimizers of Eq. (44), and let $\mathbf{\Lambda}$ be a diagonal matrix with non-negative entries. Then

$$\begin{aligned} C(\mathbf{W}^\star\mathbf{\Lambda}^{-1},\mathbf{\Lambda}\mathbf{H}^\star) &= -\log p(\mathbf{V}|\mathbf{\Lambda}\mathbf{H}^\star;\mathbf{W}^\star\mathbf{\Lambda}^{-1}) - \log p(\mathbf{\Lambda}\mathbf{H}^\star) &(50)\\ &= -\log p(\mathbf{V}|\mathbf{H}^\star;\mathbf{W}^\star) - \log p(\mathbf{\Lambda}\mathbf{H}^\star), &(51)\end{aligned}$$

and depending on the choice of the prior distribution $p(\mathbf{H})$, we may obtain $C(\mathbf{W}^\star\mathbf{\Lambda}^{-1},\mathbf{\Lambda}\mathbf{H}^\star) < C(\mathbf{W}^\star,\mathbf{H}^\star)$, i.e., a contradiction. Therefore, in the following we impose that $||\mathbf{w}_k||_1 = 1$.

The constrained optimization is performed with the following update rule

$$w_{fk} = \frac{p'_{fk}}{\sum_f p'_{fk}}, \tag{52}$$

see Appendix C for the proof.

The following subsections detail the optimization w.r.t. $\mathbf{H}$ (and other variables when necessary) in the four considered models, which amounts to the minimization of $G_1(\mathbf{H};\tilde{\mathbf{H}}) - \log p(\mathbf{H})$.

Table 1: Coefficients of the polynomial equation Eq. (53)

n	$a_{2,kn}$	$a_{1,kn}$	$a_{0,kn}$
1	q_k	$\alpha_k - p_{1k}$	$-\beta_k h_{k2}$
$2, \dots, N-1$	$q_k + \frac{\beta_k}{h_{k(n-1)}}$	$1 - p_{kn}$	$-\beta_k h_{k(n+1)}$
N	0	$q_k + \frac{\beta}{h_{k(N-1)}}$	$1 - \alpha_k - p_N$

3.2. Chaining on the rate parameter

The transition distribution $p(h_{kn}|h_{k(n-1)})$ is given by Eq. (9). The optimization w.r.t. h_{kn} amounts to solving an order-2 polynomial equation on $\mathbb{R}_+$

$$a_{2,kn} h_{kn}^2 + a_{1,kn} h_{kn} + a_{0,kn} = 0. \tag{53}$$

As it turns out, there is always exactly one non-negative root. The coefficients of the polynomial equation are given in Table 1. This bears resemblance with the methodology described in Févotte et al. (2009), where the authors aimed at retrieving MAP estimates with a EM-like algorithm (with an exponential likelihood).

3.3. Hierarchical chaining on the rate parameter

In this case, we resort to using the auxiliary variables $\mathbf{Z}$, which results in the the slightly more involved following criterion

$$\begin{aligned} C(\mathbf{W}, \mathbf{H}, \mathbf{Z}) = &- \log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) \\ &- \sum_k \left[\log p(h_{k1}) + \sum_{n \geq 2} \left(\log p(z_{kn}|h_{k(n-1)}) + \log p(h_{kn}|z_{kn}) \right) \right]. \end{aligned} \tag{54}$$

We recall that $p(z_{kn}|h_{k(n-1)})$ and $p(h_{kn}|z_{kn})$ are given by Eq. (15) and Eq. (16), respectively. Note that Cemgil and Dikmen (2007) proposed a Gibbs sampler and variational inference, and as such the development of the MAP algorithm is novel.

Imposing $\alpha_{h,k} \geq 1$, we obtain the following update for z_{kn}

$$z_{kn} = \frac{\alpha_{z,k} + \alpha_{h,k} - 1}{\beta_{z,k} h_{k(n-1)} + \beta_{h,k} h_{kn}}, \tag{55}$$

and the following updates for h_{kn}

$$h_{k1} = \frac{p_{k1} + \alpha_{z,k}}{q_k + \beta_{z,k} z_{k2}}, \tag{56}$$

$$h_{kn} = \frac{p_{kn} + \alpha_{h,k} + \alpha_{z,k} - 1}{q_k + \beta_{h,k} z_{kn} + \beta_{z,k} z_{k(n+1)}}, n \in \{2, \dots, N-1\}, \tag{57}$$

$$h_{kN} = \frac{p_{kN} + \alpha_{h,k} - 1}{q_k + \beta_{h,k} z_{kN}}. \tag{58}$$

3.4. Chaining on the shape parameter

The transition distribution $p(h_{kn}|h_{k(n-1)})$ is given by Eq. (24). The optimization w.r.t. h_{kn} amounts to solving the following equations on $\mathbb{R}_+$

$$-p_{k1} + (q_k - \alpha_k \log(\beta_k h_{k2}) + \alpha_k \Psi(\alpha h_{k1}))h_{k1} = 0, \tag{59}$$

$$(1 - \alpha_k h_{k(n-1)} - p_{kn}) + (q_k + \beta_k - \alpha_k \log(\beta_k h_{k(n+1)}))h_{kn} + \alpha_k \Psi(\alpha_k h_{kn}) h_{kn} = 0, \tag{60}$$
for $n \in \{2, \dots, N-1\}$,

where Ψ denotes the digamma function. Solving such equations can be done numerically with Newton's method. Finally the update for h_{kN} is given by

$$h_{kN} = \frac{p_{kn} + \alpha_k h_{k(N-1)} - 1}{q_k + \beta_k}. \tag{61}$$

Note that a Gibbs sampling procedure is proposed in Acharya et al. (2015); Schein et al. (2016), and as such the development of the MAP algorithm is novel.

3.5. BGAR(1)

In this case, since the transition distribution $p(h_{kn}|h_{k(n-1)})$ is not known in closed form, we resort to optimizing the slightly more involved following criterion

$$C(\mathbf{W}, \mathbf{H}, \mathbf{B}) = -\log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) - \sum_k \left[\log p(h_{k1}) + \sum_{n \geq 2} \left(\log p(h_{kn}|h_{k(n-1)}, b_{kn}) + \log p(b_{kn}) \right) \right]. \tag{62}$$

In the following, we will use the notations $\gamma_k = \alpha_k(1 - \rho_k)$ and $\eta_k = \alpha_k \rho_k$.

3.5.1. Constraints

By construction, the variables h_{kn} and b_{kn} must lie in a specific interval given the values of all the other variables. Indeed, as $h_{kn} = b_{kn} h_{k(n-1)} + \epsilon_{kn}$ (see Eq. (34)), where ϵ_{kn} is a non-negative random variable, we obtain $h_{kn} \geq b_{kn} h_{k(n-1)}$, $b_{kn} \leq \frac{h_{kn}}{h_{k(n-1)}}$, and $h_{kn} \leq \frac{h_{k(n+1)}}{b_{k(n+1)}}$.

Table 2: Coefficients of the polynomial equation Eq. (68). Def. int. = Definition interval.

n	Def. interval	$a_{3,kn}$	$a_{2,kn}$	$a_{1,kn}$	$a_{0,kn}$
1	$[0, d_{k1}]$	0	$-(q_k + \beta_k(1 - b_{k2}))$	$-(1 - \alpha_k - p_{k1}) + (q_k + \beta_k(1 - b_{k2}))d_{k1} - (1 - \gamma_k)$	$(1 - \alpha_k - p_{k1})d_{k1}$
$2, \dots, N-1$	$[c_{kn}, d_{kn}]$	$-(q_k + \beta_k(1 - b_{k(n+1)}))$	$p_{kn} - 2(1 - \gamma_k) + (q_k + \beta_k(1 - b_{k(n+1)}))(c_{kn} + d_{kn})$	$-p_{kn}(c_{kn} + d_{kn}) + (1 - \gamma_k)(c_{kn} + d_{kn}) - (q_k + \beta_k(1 - b_{k(n+1)}))c_{kn}d_{kn}$	$p_{kn}c_{kn}d_{kn}$
N	$[c_{kN}, +\infty[$	0	$q_k + \beta_k$	$-p_{kN} - c_{kN}(q_k + \beta_k) + (1 - \gamma_k)$	$c_{kN}p_{kN}$

This leads to the following constraints

$$0 \leq h_{k1} \leq \frac{h_{k2}}{b_{k2}}, \tag{63}$$

$$b_{kn}h_{k(n-1)} \leq h_{kn} \leq \frac{h_{k(n+1)}}{b_{k(n+1)}} \quad n \in \{2, \dots, N-1\}, \tag{64}$$

$$b_{kN}h_{k(N-1)} \leq h_{kN}, \tag{65}$$

and

$$0 \leq b_{kn} \leq \min\left(1, \frac{h_{kn}}{h_{k(n-1)}}\right). \tag{66}$$

We therefore introduce the notations

$$c_{kn} = b_{kn}h_{k(n-1)}, \quad d_{kn} = \frac{h_{k(n+1)}}{b_{k(n+1)}}, \quad x_{kn} = \frac{h_{kn}}{h_{k(n-1)}}, \tag{67}$$

as these quantities arise naturally in our derivations.

3.5.2. Minimization w.r.t. h_{kn}

The optimization of Eq. (62) w.r.t. h_{kn} may give rise to intractable problems, due to the logarithmic terms in the objective function. To alleviate this issue, we propose to control the limit values of the auxiliary function, by restricting ourselves to certain values of the hyperparameters. In particular, choosing $(1-\gamma_k) < 0$ ensures the existence of at least one minimizer.

For all n, the optimization w.r.t. h_{kn} amounts to solving an order-3 polynomial equation

$$a_{3,kn}h_{kn}^3 + a_{2,kn}h_{kn}^2 + a_{1,kn}h_{kn} + a_{0,kn} = 0. \tag{68}$$

The coefficients of the equation and definition intervals are given in Table 2. If several roots belong to the definition interval, we simply choose the root which gives the lowest objective value.

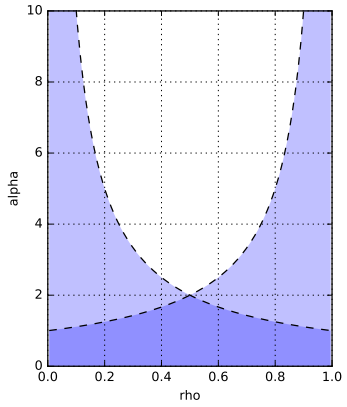

Figure 7: Admissible values of the hyperparameters in the MAP algorithm presented in Section 3.5. Admissible values are in white.

3.5.3. Minimization w.r.t. b_{kn}

Similarly, logarithmic terms of the objective function may give rise to degenerate solutions. Using the same reasoning, we choose to impose $(1-\gamma_k) < 0$ and $(1-\eta_k) < 0$ to ensure the existence of at least one minimizer.

The minimization of the auxiliary function w.r.t. b_{kn} amounts to solving the following order 3 polynomial over the interval $[0, \min(1, x_{kn})]$

$$a_{3,kn} b_{kn}^3 + a_{2,kn} b_{kn}^2 + a_{1,kn} b_{kn} + a_{0,kn} d_{kn} = 0, \tag{69}$$

where

$$\begin{aligned}
a_{3,kn} &= -\beta_k h_{k(n-1)}, && (70)\\
a_{2,kn} &= 2(1-\gamma_k) + (1-\eta_k) + \beta_k h_{k(n-1)}(x_{kn}+1), && (71)\\
a_{1,kn} &= -(1-\gamma_k)(x_{kn}+1) - (1-\eta_k)(x_{kn}+1) && (72)\\
&\quad - \beta_k h_{k(n-1)} x_{kn}, \\
a_{0,kn} &= (1-\eta_k) x_{kn}. && (73)
\end{aligned}$$

3.5.4. Admissible values of hyperparameters

To recap the discussion on admissible values of hyperparameters, to ensure the existence of minimizers of the auxiliary function, we have restricted ourselves to

$$\begin{cases} \alpha_k(1-\rho_k) > 1, \\ \alpha_k \rho_k > 1. \end{cases} \tag{74}$$

This set is graphically displayed on Figure 7. As we can see, choosing the value of ρ_k to be close to one (to ensure correlation) leads to high values of α_k.

4. Experimental work

We now compare the performance of all considered temporal NMF models on a prediction task on three real datasets. This task will consist in hiding random columns of the considered datasets and estimating those missing values. We will also include the performance of a naive baseline, which we detail in the following subsection. Adapting the MAP algorithms presented in Section 2 in a setting with a mask of missing values only consist in a slight modification, presented in Appendix D. Python code is available online[4].

4.1. Experimental protocol

For each considered dataset, the experimental protocol is as follows.

First of all, a value of the factorization rank K (which will be used for all considered methods) must be selected. To do so, we apply the standard KL-NMF algorithm (Lee and Seung, 2000; Févotte and Idier, 2011) on 10 random training sets, which consist of 80% of the original data, with a pre-defined grid of values for K. We then select the value of K which yields the lowest generalized Kullback-Leibler error (KLE) (see definition below) on the remaining 20% of the data, on average.

For the prediction experiment itself, we create 5 random splits of the data matrix, where 80% corresponds to the training set, 10% to the validation set, and the remaining 10% to the test set. To do so, we randomly select non-adjacent columns of the data matrix (excluding the first one and always including the last one), half of which will make up the validation set and the other half the test set (the last column is always included in the test set). We also consider 5 different random initializations.

Then, for each split-initialization pair, all the algorithms are run from this initialization point on the training set until convergence (the algorithms are stopped when the relative decrease of the objection function falls under 10^{-5}). For each method, a grid of hyperparameters is considered, and their selection is based on the lowest KLE on the validation set. Details of the grids used for each method can be found in Appendix E. The predictive performance of each method is then computed on the test set by comparing the original value v_{fn} and its associated estimate $\hat{v}_{fn} = [\mathbf{WH}]_{fn}$ with the following metric. Denoting by $\mathcal{T}$ the test set, we compute the generalized Kullback-Leibler error (KLE), which is defined as

$$\mathrm{KLE} = \sum_{(f,n)\in\mathcal{T}} \left[v_{fn} \log\left(\frac{v_{fn}}{\hat{v}_{fn}}\right) - v_{fn} + \hat{v}_{fn} \right]. \tag{75}$$

Finally, we construct a baseline based on the Gamma-Poisson (GaP) model of Canny (2004). The GaP model is based on independent Gamma priors on $\mathbf{H}$, i.e., a non-temporal prior. However, it is unable to estimate columns $\mathbf{h}_n$ associated with missing columns $\mathbf{v}_n$. We propose to set $\mathbf{h}_n = \frac{1}{2}(\mathbf{h}_{n-1} + \mathbf{h}_{n+1})$ and $\mathbf{h}_N = \mathbf{h}_{N-1}$ for these columns. MAP estimation in the GaP model is described in Dikmen and Févotte (2012) and is recalled in Appendix F.

[4] https://github.com/lfilstro/TemporalNMF

Model	KLE-S	KLE-F
GaP (App. F)	$6.19 \times 10^4 \pm 9.69 \times 10^3$	$1.08 \times 10^5 \pm 2.67 \times 10^3$
Rate (3.2)	$6.07 \times 10^4 \pm 8.97 \times 10^3$	$1.03 \times 10^5 \pm 3.53 \times 10^3$
Hier (3.3)	$6.06 \times 10^4 \pm 9.18 \times 10^3$	$3.24 \times 10^5 \pm 2.09 \times 10^5$
Shape (3.4)	$9.37 \times 10^4 \pm 2.99 \times 10^4$	$1.30 \times 10^5 \pm 2.35 \times 10^4$
BGAR (3.5)	$6.17 \times 10^4 \pm 8.24 \times 10^3$	$1.36 \times 10^5 \pm 2.00 \times 10^3$

Table 3: Prediction results on the `NIPS` dataset. Lower values are better. The mean and standard deviation of each metric are reported over 25 runs.

4.2. Datasets

The following datasets are considered

- The `NIPS` dataset[5], which contains word counts (with stop words removed) of all the articles published at the NIPS[6] conference between 1987 and 2015. We grouped the articles per publication year, yielding an observation matrix of size 11463×29. We obtained $K = 3$.
- The `last.fm` dataset, based on the so-called "last.fm 1K" users[7], which contains the listening history with timestamps information of users of the music website last.fm. We preprocessed this dataset to obtain the monthly evolution of the listening counts of artists with at least 20 different listeners. This yields a dataset of size 7017×53 (i.e., we have the listening history of 7017 artists over 53 months). We obtained $K = 5$.
- The `ICEWS` dataset[8], an international relations dataset, which contains the number of interactions between two countries for each day of the year 2003. The matrix is of size 6197×365. We obtained $K = 5$.

4.3. Experimental results

As previously mentioned, the test set consists of 10 % of the columns of the data matrix $\mathbf{V}$, always including the last one. The KLE will be computed separately on all the columns minus the last one (denoted by "S" for smoothing), and on the last one (denoted by "F" for forecasting). Their averaged values over the 25 split-initialization pairs are reported on Table 3 for the `NIPS` dataset, on Table 4 for the `last.fm` dataset, and on Table 5 for the `ICEWS` dataset.

All considered temporal models achieve comparable predictive performance, both on smoothing and forecasting tasks, and the slight advantage of one method over the others

[5] https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
[6] Now called NeurIPS.
[7] http://ocelma.net/MusicRecommendationDataset/
[8] https://github.com/aschein/pgds

Model	KLE-S	KLE-F
GaP (App. F)	$1.30 \times 10^4 \pm 3.35 \times 10^2$	$6.89 \times 10^3 \pm 5.84 \times 10^1$
Rate (3.2)	$1.23 \times 10^4 \pm 2.35 \times 10^2$	$7.76 \times 10^3 \pm 4.13 \times 10^1$
Hier (3.3)	$1.23 \times 10^4 \pm 2.95 \times 10^2$	$6.35 \times 10^3 \pm 1.57 \times 10^3$
Shape (3.4)	$1.58 \times 10^4 \pm 2.45 \times 10^3$	$2.04 \times 10^4 \pm 9.92 \times 10^3$
BGAR (3.5)	$1.24 \times 10^4 \pm 3.00 \times 10^2$	$9.65 \times 10^3 \pm 2.91 \times 10^3$

Table 4: Prediction results on the `last.fm` dataset. Lower values are better. The mean and standard deviation of each metric are reported over 25 runs.

Model	KLE-S	KLE-F
GaP (App. F)	$8.91 \times 10^4 \pm 2.82 \times 10^3$	$1.59 \times 10^3 \pm 6.34 \times 10^1$
Rate (3.2)	$9.17 \times 10^4 \pm 2.99 \times 10^3$	$1.62 \times 10^3 \pm 7.57 \times 10^1$
Hier (3.3)	$9.11 \times 10^4 \pm 2.81 \times 10^3$	$1.62 \times 10^3 \pm 9.02 \times 10^1$
Shape (3.4)	$9.95 \times 10^4 \pm 3.82 \times 10^3$	$1.84 \times 10^3 \pm 1.37 \times 10^2$
BGAR (3.5)	$8.99 \times 10^4 \pm 2.80 \times 10^3$	$1.78 \times 10^3 \pm 8.62 \times 10^1$

Table 5: Prediction results on the `ICEWS` dataset. Lower values are better. The mean and standard deviation of each metric are reported over 25 runs.

seems to be data-dependent. The methods "Rate" and "Hier" rank first or second most of time but not always significantly so. The method "Shape" tends to achieve worse results than others, but not consistently. This suggests that in the MAP estimation framework, prior distributions do not act as strong regularization terms, and are outweighed by the likelihood term. Moreover, the baseline based on the GaP model also achieves good performance. This might be attributed to the high correlation between successive columns on the datasets, and as such, using adjacent columns for estimation is reasonable.

We conclude this section by saying a few words about computational complexity, which can act as a differentiating criterion, as all models achieve similar predictive performance. The algorithms for chains involving the rate parameters of the Gamma distribution, described in Sections 3.2 and 3.3 have closed-form update rules for all their variables. This leads to efficient block-descent algorithms. This is in contrast with the algorithms for the model based on the chaining on the shape parameter (Section 3.4), which involves solving $K(N-1)$ equations numerically at each iteration, and the algorithm for BGAR (Section 3.5), which involves serially solving $2K(N-1)$ order-3 polynomials at each iteration.

5. Conclusion

In this paper, we have reviewed existing temporal NMF models in a unified MAP framework and introduced a new one. These models differ by the choice of the Markov chain structure used on the activation coefficients to induce temporal correlation. We began by studying the previously proposed Gamma Markov chains of the NMF literature, only to find that they all share the same drawback, namely the absence of a well-defined stationary distribution. This leads to problematic behaviors from the generative perspective, because the realizations of the chains are degenerate (although this is not necessarily a problem in MAP estimation). We then introduced a Markovian process from the time series literature, called BGAR(1), which overcomes this limitation, and which, to the best of our knowledge, had never been exploited for learning tasks.

We then derived MAP estimation algorithms in the context of a Poisson likelihood, which allowed for a comprehensive comparison on a prediction task on real datasets. As it turns out, we cannot claim that there is a single model which outperforms all the others. It seems that in our framework, MAP estimation will tend to homogenize the performance of all the models.

Future work will focus on finding a way to perform inference with the BGAR prior for a less restrictive set of hyperparameters, which might increase the performance of this particular model. Moreover, it should be noted that this work can easily be extended to other likelihoods than Poisson thanks to the MM framework. To illustrate this, we present in Appendix G the derivation of an algorithm for MAP estimation in a model consisting in an exponential likelihood and BGAR(1) temporal prior. Finally, it would be interesting to carry out similar experimental work within a fully Bayesian estimation paradigm, which might make the differences between the models more striking.

Acknowledgments

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No 681839 (project FACTORY). Louis Filstroff and Olivier Gouvert were with IRIT, Univ. Toulouse, CNRS, France at the time this research was conducted.

Appendix A The Beta-Prime distribution

Distribution for a continuous random variable in $[0, +\infty[$, with parameters $\alpha > 0$, $\beta > 0$, $p > 0$ and $q > 0$. Its probability density function writes, for $x \geq 0$:

$$f(x; \alpha, \beta, p, q) = \frac{p \left(\frac{x}{q}\right)^{\alpha p - 1} \left(1 + \left(\frac{x}{q}\right)^{p}\right)^{-\alpha - \beta}}{q \mathrm{B}(\alpha, \beta)}. \tag{76}$$

Appendix B BGAR(1) linear correlation

We have between two successive values h_n and h_{n+1}:

$$\begin{aligned}
&\mathrm{corr}(h_n, h_{n+1}) \\
&= \frac{\mathbb{E}(h_n h_{n+1}) - \mathbb{E}(h_n)\mathbb{E}(h_{n+1})}{\sigma(h_n)\sigma(h_{n+1})} && (77)\\
&= \frac{\mathbb{E}(h_n(b_{n+1} h_n + \epsilon_{n+1})) - \mathbb{E}(h_n)\mathbb{E}(h_{n+1})}{\sigma(h_n)\sigma(h_{n+1})} && (78)\\
&= \frac{\mathbb{E}(b_{n+1})\mathbb{E}(h_n^2) + \mathbb{E}(h_n)\mathbb{E}(\epsilon_{n+1}) - \mathbb{E}(h_n)\mathbb{E}(h_{n+1})}{\sigma(h_n)\sigma(h_{n+1})} && (79)\\
&= \frac{\frac{\alpha\rho}{\alpha\rho + \alpha(1-\rho)} \frac{\alpha(\alpha+1)}{\beta^2} + \frac{\alpha}{\beta}\frac{\alpha(1-\rho)}{\beta} - \frac{\alpha}{\beta}\frac{\alpha}{\beta}}{\frac{\alpha}{\beta^2}} && (80)\\
&= \rho. && (81)
\end{aligned}$$

Appendix C Constrained optimization

We want to optimize $G_2(\mathbf{W}; \tilde{\mathbf{W}})$ w.r.t. $\mathbf{W}$ s.t. $\sum_f w_{fk} = 1$. Rewriting this with Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^{\mathrm{T}}$, this is tantamount to

$$\min_{\mathbf{W}, \boldsymbol{\lambda}} G_2(\mathbf{W}; \tilde{\mathbf{W}}) + \sum_k \lambda_k (||\mathbf{w}_k||_1 - 1). \tag{82}$$

Deriving w.r.t w_{fk} yields

$$w_{fk} = \frac{p'_{fk}}{q'_k + \lambda_k}. \tag{83}$$

We retrieve the constraint by summing this expression over f. This gives the expression of the Lagrange multiplier: $\lambda_k = \sum_f p'_{fk} - q'_k$. Substituting this expression into Eq. (83), we obtain the following update rule

$$w_{fk} = \frac{p'_{fk}}{\sum_f p'_{fk}}. \tag{84}$$

Appendix D Algorithms with missing values

In the context of missing values, let us consider a mask matrix $\mathbf{M}$ of size $F \times N$ such that $m_{fn} = 1$ if the entry v_{fn} is observed and 0 otherwise. The likelihood term can then be written as

$$-\log p(\mathbf{V}|\mathbf{H};\mathbf{W}) = -\sum_{f,n} m_{fn} \log p(v_{fn}|[\mathbf{WH}]_{fn}). \tag{85}$$

The auxiliary function G_1 of Eq. (46) and G_2 of Eq. (48) can then be written is the same way, with

$$p_{kn} = \tilde{h}_{kn} \sum_f w_{fk} \frac{m_{fn} v_{fn}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}}, \quad q_{kn} = \sum_f m_{fn} w_{fk}, \tag{86}$$

for G_1, and

$$p'_{fk} = \tilde{w}_{fk} \sum_n h_{kn} \frac{m_{fn} v_{fn}}{[\tilde{\mathbf{W}}\mathbf{H}]_{fn}}, \quad q'_{kn} = \sum_n m_{fn} h_{kn}, \tag{87}$$

for G_2.

Appendix E Hyperparameter grids

For all methods, we have considered constant hyperparameters w.r.t. k (for example $\alpha_k = \alpha$ for all k). Additional details regarding each method can be found in the list below.

- For GaP, we have considered a two-dimensional grid for the parameters α and β. Values were $\alpha = \{0.1, 1, 10\}$ and $\beta = \{0.1, 1, 10\}$.
- For "Rate", we have set $\alpha = \beta$, which implies that $\mathbb{E}(h_{kn}|h_{k(n-1)}) = h_{k(n-1)}$. We considered a one-dimensional grid with values $\{1.5, 10, 100\}$.
- For "Hier", we have set $\alpha_h = \beta_h$, and $\alpha_z = \beta_z$, which implies $\mathbb{E}(z_{kn}|h_{k(n-1)}) = h_{k(n-1)}$ and $\mathbb{E}(h_{kn}|z_{kn}) = z_{kn}$. We considered a two-dimensional grid with values $\alpha_h = \{1.5, 10, 100\}$ and $\alpha_z = \{1.5, 10, 100\}$.
- For "Shape", we have set $\alpha = \beta$, which implies that $\mathbb{E}(h_{kn}|h_{k(n-1)}) = h_{k(n-1)}$. We considered a one-dimensional grid with values $\{0.1, 1, 10\}$.
- For "BGAR", we have set $\rho = 0.9$, and considered a two-dimensional grid for the parameters α and β (note that setting $\rho = 0.9$ implies $\alpha > 10$ in our MAP framework). Values were $\alpha = \{11, 110, 1100\}$ and $\beta = \{0.1, 1, 10\}$.

Appendix F MAP estimation in the GaP model

The prior distribution on $\mathbf{H}$ is such that

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k). \tag{88}$$

MAP estimation amounts to minimizing

$$C(\mathbf{W},\mathbf{H}) = -\log p(\mathbf{V}|\mathbf{H};\mathbf{W}) + \sum_{k,n}\left((1-\alpha_k)h_{kn} + \beta_k h_{kn}\right), \tag{89}$$

which leads to the following MM update rule (Dikmen and Févotte, 2012)

$$h_{kn} = \begin{cases} 0 & \text{if } p_{kn}+\alpha_k - 1 \leq 0, \\ \frac{p_{kn}+\alpha_k-1}{q_{kn}+\beta_k} & \text{else.} \end{cases} \tag{90}$$

Appendix G BGAR with an exponential likelihood

Another popular likelihood used in probabilistic NMF models is the Exponential likelihood (Févotte et al., 2009; Hoffman et al., 2010) which writes

$$v_{fn} \sim \text{Exp}\left(\frac{1}{[\mathbf{WH}]_{fn}}\right), \tag{91}$$

where $\text{Exp}(\beta) = \text{Gamma}(1,\beta)$ refers to the exponential distribution with mean $1/\beta$. This model underlies so-called Itakura-Saito NMF and has most notably been used in audio signal processing applications. We consider MAP estimation in this model with BGAR(1) prior on $\mathbf{H}$. To do so, we resort to the same MM scheme than what was presented in the beginning of Section 3. In this case, the majorization of the likelihood term is known from (Cao et al., 1999; Févotte and Idier, 2011). In particular, the function

$$G(\mathbf{H};\tilde{\mathbf{H}}) = \sum_{k,n}\left(\frac{p_{kn}}{h_{kn}} + q_{kn}h_{kn}\right), \tag{92}$$

with the notations

$$p_{kn} = \tilde{h}_{kn}^2 \sum_f \frac{w_{fk}v_{fn}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}^2}, \quad q_{kn} = \sum_f \frac{w_{fk}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}}, \tag{93}$$

is a tight auxiliary function of $-\log p(\mathbf{V};\mathbf{W},\mathbf{H})$ at $\mathbf{H} = \tilde{\mathbf{H}}$ (up to irrelevant constants).

The exact same constraints on h_{kn} and admissible values of hyperparameters detailed in Section 3.5 apply. The minimization w.r.t. h_{kn} then amounts to solving an order-4 polynomial equation

$$a_{4,kn}h_{kn}^4 + a_{3,kn}h_{kn}^3 + a_{2,kn}h_{kn}^2 + a_{1,kn}h_{kn}^1 + a_{0,kn} = 0, \tag{94}$$

whose coefficients are detailed below.

For h_{k1}:

$$a_{4,kn} = 0, \tag{95}$$
$$a_{3,kn} = -(q_{k1} + \beta_k(1 - b_{k2})), \tag{96}$$
$$a_{2,kn} = (q_{k1} + \beta_k(1 - b_{k2}))d_{k1} - (1 - \alpha_k) - (1 - \gamma_k), \tag{97}$$
$$a_{1,kn} = (1 - \alpha_k)d_{k1} + p_{k1}, \tag{98}$$
$$a_{0,kn} = -p_{k1}d_{k1}. \tag{99}$$

For $h_{kn}, n \in \{2, \ldots, N-1\}$:

$$a_{4,kn} = -(q_{kn} + \beta_k(1 - b_{k(n+1)})), \tag{100}$$
$$a_{3,kn} = (q_{kn} + \beta_k(1 - b_{k(n+1)}))(c_{kn} + d_{kn}) - 2(1 - \gamma_k), \tag{101}$$
$$a_{2,kn} = -c_{kn}d_{kn}(q_{kn} + \beta_k(1 - b_{k(n+1)})) + p_{kn} + (1 - \gamma_k)(c_{kn} + d_{kn}), \tag{102}$$
$$a_{1,kn} = -p_{kn}(c_{kn} + d_{kn}), \tag{103}$$
$$a_{0,kn} = p_{kn}c_{kn}d_{kn}. \tag{104}$$

For h_{kN}:

$$a_{4,kn} = 0, \tag{105}$$
$$a_{3,kn} = q_{kN} + \beta_k, \tag{106}$$
$$a_{2,kn} = -(q_{kN} + \beta_k)c_{kN} + (1 - \gamma_k), \tag{107}$$
$$a_{1,kn} = -p_{kN}, \tag{108}$$
$$a_{0,kn} = c_{kN}p_{kN}. \tag{109}$$

References

Acharya, A., Ghosh, J., and Zhou, M. (2015). Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9.

Acharya, A., Ghosh, J., and Zhou, M. (2018). A dual markov chain topic model for dynamic environments. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1099–1108.

Basbug, M. E. and Engelhardt, B. E. (2016). Hierarchical Compound Poisson Factorization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1795–1803.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 113–120.

Canny, J. (2004). GaP: A Factor Model for Discrete Data. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129.

Cao, Y., Eggermont, P. P., and Terebey, S. (1999). Cross burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2):286–292.

Cemgil, A. T. (2009). Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience*, (Article ID 785152).

Cemgil, A. T. and Dikmen, O. (2007). Conjugate Gamma Markov random fields for modelling nonstationary sources. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 697–705.

Charlin, L., Ranganath, R., McInerney, J., and Blei, D. M. (2015). Dynamic Poisson Factorization. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, pages 155–162.

Dikmen, O. and Févotte, C. (2012). Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10):5163–5175.

Do, T. D. T. and Cao, L. (2018). Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5829–5840.

Févotte, C. (2011). Majorization-Minimization Algorithm for Smooth Itakura-Saito Nonnegative Matrix Factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1980–1983.

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830.

Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456.

Févotte, C., Le Roux, J., and Hershey, J. R. (2013). Non-negative Dynamical System with Application to Speech and Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3158–3162.

Gaver, D. and Lewis, P. (1980). First-order autoregressive gamma sequences and point processes. *Advances in Applied Probability*, 12(3):727–745.

Gong, C. and Huang, W.-B. (2017). Deep Dynamic Poisson Factorization Model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1666–1674.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable Recommendation with Hierarchical Poisson Factorization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 326–335.

Gouvert, O., Oberlin, T., and Févotte, C. (2019). Recommendation from Raw Data with Adaptive Compound Poisson Factorization. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

Guo, D., Chen, B., Zhang, H., and Zhou, M. (2018). Deep Poisson Gamma Dynamical Systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8451–8461.

Hoffman, M. D., Blei, D. M., and Cook, P. R. (2010). Bayesian Nonparametric Matrix Factorization for Recorded Music. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 439–446.

Hunter, D. R. and Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician*, 58(1):30–37.

Jerfel, G., Basbug, M. E., and Engelhardt, B. E. (2017). Dynamic Collaborative Filtering With Compound Poisson Factorization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 738–747.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lee, D. D. and Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562.

Lewis, P., McKenzie, E., and Hugus, D. K. (1989). Gamma processes. *Communications in Statistics. Stochastic Models*, 5(1):1–30.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Schein, A., Linderman, S., Zhou, M., Blei, D., and Wallach, H. (2019). Poisson-Randomized Gamma Dynamical Systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 782–793.

Schein, A., Wallach, H. M., and Zhou, M. (2016). Poisson-Gamma Dynamical Systems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5005–5013.

Schmidt, M. N., Winther, O., and Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 540–547.

Virtanen, S. and Girolami, M. (2020). Dynamic content based ranking. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2315–2324.

Virtanen, T., Cemgil, A. T., and Godsill, S. (2008). Bayesian Extensions to Non-negative Matrix factorisation for Audio Signal Modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1825–1828.

Zhou, M. (2018). Nonparametric Bayesian Negative Binomial Factor Analysis. *Bayesian Analysis*, 13(4):1065–1093.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-Negative Binomial Process and Poisson Factor Analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1462–1471.

Şimşekli, U., Cemgil, A. T., and Yılmaz, Y. K. (2013). Learning the beta-divergence in Tweedie compound Poisson matrix factorization models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1409–1417.