*mathematics*

MDPI

*Article*

# Variational Inference over Nonstationary Data Streams for Exponential Family Models †

**Andrés R. Masegosa [1,*], Darío Ramos-López [2], Antonio Salmerón [1], Helge Langseth [3] and Thomas D. Nielsen [4]**

1 Department of Mathematics and Center for Development and Transfer of Mathematical Research to Industry (CDTIME), University of Almería, 04120 Almería, Spain; antonio.salmeron@ual.es
2 Department of Applied Mathematics, Materials Science and Engineering, and Electronic Technology, Rey Juan Carlos University, 28933 Móstoles, Spain; dario.ramos.lopez@urjc.es
3 Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway; helge.langseth@ntnu.no
4 Department of Computer Science, Aalborg University, 9220 Aalborg, Denmark; tdn@cs.aau.dk
* Correspondence: andresmasegosa@ual.es
† This paper is an extended version of our previous work "Bayesian models of data streams with hierarchical power priors", presented at the International Conference on Machine Learning 2017 Conference.

Received: 8 October 2020; Accepted: 27 October 2020; Published: 3 November 2020


**Abstract:** In many modern data analysis problems, the available data is not static but, instead, comes in a streaming fashion. Performing Bayesian inference on a data stream is challenging for several reasons. First, it requires continuous model updating and the ability to handle a posterior distribution conditioned on an unbounded data set. Secondly, the underlying data distribution may drift from one time step to another, and the classic i.i.d. (independent and identically distributed), or data exchangeability assumption does not hold anymore. In this paper, we present an approximate Bayesian inference approach using variational methods that addresses these issues for conjugate exponential family models with latent variables. Our proposal makes use of a novel scheme based on hierarchical priors to explicitly model temporal changes of the model parameters. We show how this approach induces an exponential forgetting mechanism with adaptive forgetting rates. The method is able to capture the smoothness of the concept drift, ranging from no drift to abrupt drift. The proposed variational inference scheme maintains the computational efficiency of variational methods over conjugate models, which is critical in streaming settings. The approach is validated on four different domains (energy, finance, geolocation, and text) using four real-world data sets.

**Keywords:** latent variable models; nonstationary data streams; concept drift; variational inference; power priors; exponential forgetting

## 1. Introduction

One core problem in Bayesian statistics is the computation of posterior probability over the parameters (and the latent variables) of a model given a data set. In most relevant cases, the computation of this posterior probability is not feasible and requires the use of approximate inference algorithms [1,2]. Moreover, in many real-life settings, the data set is not static but arrives sequentially, in a streaming fashion. Computing the Bayesian posterior probability over the parameters (and the latent variables) of a model, in this case, is even more challenging. First, it requires the ability to handle a posterior distribution conditioned on an unbounded data set. Secondly, the posterior probability has to be updated frequently, at every time step, imposing restrictions over the speed of computation of the posterior probability. Finally, the classical i.i.d. (independent and

identically distributed, or data exchangeability) assumption does not hold because the underlying data distribution may drift/change from one time step to the next. This last issue is known in the machine learning literature as learning from data streams that exhibit *concept drift* [3] in the sense that the underlying generative process may contain both gradual and abrupt changes.

In this paper, we look at the common setting in many real-world problems, where data arrives as a sequence of (potentially large) batches. Specifically, each batch is associated with a new time step and the data points within a batch are assumed to be i.i.d. However, the underlying distribution generating the batches may *change* from one time step to the next. For illustration, in Section 6, we will apply a probabilistic topic model [4] to identify the main topics of research papers published over a sequence of years. New papers are submitted yearly; hence, every year, the probabilistic model must be updated to infer the topics of that year's papers. One approach could be to discard all data from previous years, but that would result in a loss of relevant data that could potentially have been used to draw more accurate inferences about the current year's topic distribution (research topics typically span multiple years). On the other hand, we also have to take into account that the underlying distribution of the papers may change from one year to the next following the development of the research field. As another example domain, we will model the financial profile of customers asking for a loan/mortgage from a financial institution [5]. In this example, we update the financial profile of the customers on a monthly basis while simultaneously keeping in mind that the occurrence of exogenous events (like an economic crisis) may strongly affect the customers and therefore induce a drift in the underlying data distribution. Other situations similar to the ones above arise in many different real-life settings.

Standard temporal models [6] have the potential to capture the underlying temporal dynamics of the data set. However, for the problems addressed in this paper, temporal models are not readily applicable. The reason for this is two-fold. First, the set of *objects* differs from one time step to the next. For example, a new set of papers is submitted to a conference every year. Thus, temporal modeling needs to be applied at the level of the *parameters* in the model, not at the objects themselves. Secondly, drift (both gradual and abrupt) in the underlying distributions is not easily modelled with a stationary transition model as would typically be the case in temporal models. Instead, we will employ an *implicit transition model* to model the temporal dynamics, thus sidestepping the problem of specifying an explicit stationary transition distribution.

The approach presented in this paper is applicable to domains that, at each time point, can be described by a conjugate exponential family model. These models are widely used in probabilistic modeling [7,8] and include popular models like *latent Dirichlet allocation* (LDA) models [4] to uncover the hidden topics in a text corpora, a mixture of (multivariate) Gaussian models to discover hidden clusters in data [9], and probabilistic principal component analysis for revealing a low-dimensional representation of the data [10]. See, for instance, [9,11,12] for detailed reviews and applications of these models in data mining and machine learning settings.

Exact Bayesian inference is intractable for the model-class proposed in this paper, partly due to the implicit transition model and partly due to the high dimensionality of the parameters/latent variables in the model. We therefore resort to approximate Bayesian inference methods, and owing to the potential real-time requirements imposed by the streaming domains, we focus on variational methods (see [13] for an introduction). Markov Chain Monte Carlo methods (MCMC) [1,2], although widely used in Bayesian statistics, are not as computationally efficient as variational methods, especially in the presence of large datasets or complex models [13]. Moreover, variational inference can be performed very efficiently in conjugate exponential family models by exploiting properties of this model family.

The rest of the paper is organized as follows. Section 2 discusses related works, and Section 3 introduces preliminaries relevant for the remainder of the paper. In Section 4, we introduce an *implicit transition model* that works as an *off-the-self* parameter transition scheme to model both gradual and abrupt changes in the data distribution. Section 5 positions the transition model in a hierarchical Bayesian setting to model the rate of change of the data stream as an unobserved mechanism. Section 5

also includes the derivation of an ad hoc variational inference algorithm [13], based on a novel lower-bound of the data log-likelihood function for the proposed model family (i.e., a conjugate exponential model with an implicit transition scheme). As a consequence, the proposed approximate inference scheme is deterministic and scales to large data streams. Section 6 empirically evaluates the appropriateness of our approach using both synthetic and real-life data (covering energy, finance, geolocation, and text data), showing promising results. We conclude in Section 7, where the main conclusions and future works are discussed.

## 2. Related Work

Making inferences from *nonstationary* data streams has been extensively studied in statistics. Time series models [14] is a classic example of models that could be assumed relevant for this problem, but as we are not dealing with temporal data (where the the same set of objects are observed at each time step), this model class is not applicable in our situation. Work by [15,16] is closer to ours, but these contributions focus on simpler model classes (like generalized linear models without latent variables), and due to this simplicity, they do not consider variational inference.

Methods for change point detection [17,18] are also related to our situation. However, where these methods consider *abrupt* changes in the data stream, our method is able to deal with both abrupt and smooth changes in the distribution. Additionally, the main focus of change point detection methods is the detection of points where the changes occur, while our goal is to have a statistical model that accurately models the data at each time step by continuously adapting the model parameters.

Learning from nonstationary data streams has also been extensively studied in the machine learning literature, especially in the context of classification and clustering [3,19–21]. One of the main techniques employed in this area has been *exponential forgetting* [22,23]. Here, each new observation is initially assigned a weight (or "importance") equal to 1, followed by an exponential decrease of the weight at each subsequent time step. In this way, older observations are less relevant than newer ones when the model is learned, thereby accounting for potential drift in the data stream. The main problem with this approach is to quantify the so-called *exponential forgetting rate*, which is usually determined by the analyst on a trial-and-error basis. In the present work, we provide a sound approach based on Bayesian methods to automatically adjust the exponential forgetting rate and show the advantages of this strategy empirically.

Bayesian modeling of nonstationary data streams for general probabilistic inference has not been extensively studied so far. An online variational inference method, which exponentially forgets old data, was proposed by Honkela and Valpola [24], but similarly to [22,23], this approach suffers from the problem of setting the exponential forgetting rate. Another proposal, called population variational Bayes (PVB), was introduced by [25]. It builds directly on the stochastic variational inference (SVI) algorithm [26]. SVI assumes the existence of a fixed data set observed in a sequential manner and, in particular, that this data set has a known finite size. This is unrealistic when modeling data streams. PVB addresses the problem by using the frequentist notion of a *population distribution*, $\mathbf{F}$, which is assumed to generate the data stream by repeatedly sampling $M$ data points at a time; here, $M$ parameterizes the size of the population and helps control the variance of the population posterior. By artificially having high variance in the posterior (i.e., by choosing a small value for $M$), PVB is able to accommodate drift in the data set. Unfortunately, $M$ must be specified by the user, and no clear rule exists for specifying it. Furthermore, McInerney et al., [25] shows that the optimal value for $M$ may differ between data stream. The streaming variational Bayes (SVB) algorithm by [27] also tries to address Bayesian inference in data streams. SVB builds on a Bayesian recursive updating approach but does not provide a mechanism for dealing with concept drift. In Section 6, we will show that our proposed method, which does not rely on hyperparameters that are hard to tune, outperforms these closely related approaches on several real-world data sets.

The so-called *power prior* approach [28] has also been studied in the context of data aggregation for Bayesian modeling. Power priors provide a sound mechanism for Bayesian updating in light of

new data and introduces partial forgetting of older observations. The approach enjoys nice theoretical properties [29] but depends on a hyperparameter (set by the analyst) to control the forgetting rate. Our work builds partially on the power prior model but extends this model with a Bayesian model of the forgetting mechanism, thereby dispensing of the analyst-specified hyperparameter.

A modeling approach based on time-series models for concept drift using *implicit transition models* was pursued by [30,31]. Unfortunately, the implicit transition model also depends on a hyperparameter determining the forgetting factor, which has to be manually set. Our work also partially builds on this approach and establishes novel connections between the *implicit transition models* [30,31], *power priors* [28], and exponential forgetting [22,23].

Many other contributions have proposed an ad hoc extension to specific statistical models in order to deal with nonstationary data streams [32,33], including dynamic extensions of *latent Dirichlet allocation* (LDA) models [34–36]. However, none are so far applicable to general conjugate exponential family models. Moreover, most of these contributions rely on complex and tailor-made inference mechanisms, instead of variational inference (which is well understood for conjugate exponential models).

## 3. Preliminaries

### 3.1. Conjugate Exponential Models

Let $\boldsymbol{x} = \boldsymbol{x}_{1:N}$ denote a set of observed variables, $\boldsymbol{z} = \boldsymbol{z}_{1:N}$ denote a set of latent variables, and $\boldsymbol{\beta} = \boldsymbol{\beta}_{1:M}$ denote the parameters of the model. Let $\boldsymbol{\alpha}$ also denote a hyperparameter vector.

We assume that the joint distribution of our statistical model factorizes into a product of local terms and a global term:

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}|\boldsymbol{\alpha}) = p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta}).$$

This kind of factorization is quite usual in many statistical models [26], for example, the case of the Bayesian mixture of Gaussian, where each latent variable is a categorical variable defining the component to which each data point belongs. We also consider the special case when there are no latent variables; this case would cover models like Bayesian linear regression. Figure 1a provides a graphical description of these model family in terms of probabilistic graphical models with plateau notation [11].

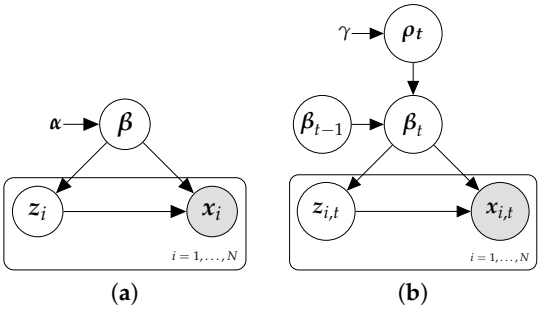



**Figure 1.** The left figure (**a**) displays a graphical representation of the probabilistic model examined in this paper (see Section 3.1). The right figure (**b**) includes a temporal evolution model for $\boldsymbol{\beta}_t$ as described in Section 5. The graphical notation corresponds to probabilistic graphical models with plateau notation [11]. Under this notation, a random variable is represented by a node, and it is conditionally dependent on those random variables with a node pointing to it. The nodes within the box denote random variables that are replicated for every data sample.

We further assume that the (conditional) distributions defining the statistical model belong to the *conjugate exponential family* [37]. This model family has been largely studied and covers a wide range of probability distributions such as multinomial, normal, gamma, Dirichlet, beta, etc. According to this

assumption, the functional form of the conditional distribution of the local variables $(\boldsymbol{x}_n, \boldsymbol{z}_n)$ given the parameters $\boldsymbol{\beta}$ has the well-known *exponential family* form [37]:

$$\ln p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta}) = \ln h(\boldsymbol{x}_n, \boldsymbol{z}_n) + \boldsymbol{\beta}^T t(\boldsymbol{x}_n, \boldsymbol{z}_n) - a_l(\boldsymbol{\beta}), \tag{1}$$

where the scalar functions $h(\cdot)$ and $a_l(\cdot)$ are the base measure and the log-normalizer, respectively, and the vector function $\boldsymbol{t}(\cdot)$ is the *sufficient statistics* vector. Moreover, the prior distribution $p(\boldsymbol{\beta})$ also belongs to the exponential family and has the following structure:

$$\ln p(\boldsymbol{\beta}) = \ln h(\boldsymbol{\beta}) + \boldsymbol{\alpha}^T t(\boldsymbol{\beta}) - a_g(\boldsymbol{\alpha}) \tag{2}$$

where the sufficient statistics are $t(\boldsymbol{\beta}) = (\boldsymbol{\beta}, -a_l(\boldsymbol{\beta}))$ and the hyperparameter $\boldsymbol{\alpha}$ has two components $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^*, \xi)$, where the first component $\boldsymbol{\alpha}^*$ has the same dimension as $\boldsymbol{\beta}$ and encodes the prior belief about the distribution over $\boldsymbol{\beta}$ and the second component $\xi > 0$ is a scalar and encodes the strength in our prior belief [38]. This second parameter is also known in the literature [39] as the *equivalent sample size of the prior distribution* ($ESS_{prior}$).

Our inference goal is to approximate the posterior distribution over the parameters and latent variables given the following observations:

$$p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{\beta}) \prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta})}{\int_{\boldsymbol{\beta}} p(\boldsymbol{\beta}) \prod_n \int_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta}) d\boldsymbol{z}_n d\boldsymbol{\beta}}.$$

However, this posterior is intractable in general due to the integral in the denominator (i.e., *the evidence integral*), and we therefore have to resort to approximate inference algorithms such as variational inference. These (and other) inference algorithms benefit enormously when the *complete conditional distribution* of the parameters and the latent variables belongs to the exponential family [26]. Therefore, we make this extra assumption about the model class (Note however, that the presented approach also applies to the more general *conjugate exponential family*.), which states that the conditional distribution over $\boldsymbol{\beta}$ and $\boldsymbol{z}$ given the rest of the variables has the same functional form as the priors:

$$\begin{aligned} \ln p(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha}) &= \ln h(\boldsymbol{\beta}) + \eta_g(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha})^T t(\boldsymbol{\beta}) - a_g(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha}) \\ \ln p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\beta}) &= h(\boldsymbol{z}_n) + \eta_l(\boldsymbol{x}_n, \boldsymbol{\beta})^T t(\boldsymbol{z}_n) - a_l(\eta_l(\boldsymbol{x}_n, \boldsymbol{\beta})), \end{aligned}$$

where the vector function $\eta_\cdot(\cdot)$ denotes the *natural parameter vectors* of the conditional probability distributions.

By Equations (1) and (2), the natural parameter vector of $p(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha})$ can be expressed as

$$\eta_g(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\alpha}) = \left( \boldsymbol{\alpha}^* + \sum_{n=1}^{N} t(\boldsymbol{x}_n, \boldsymbol{z}_n), \xi + N \right). \tag{3}$$

Therefore, computing the full posterior reduces to updating the natural parameters of the prior. Moreover, the *equivalent sample size of the posterior* ($ESS_{post}$) is equal to the *equivalent sample size of the prior* ($ESS_{prior}$) plus the size of the observations.

### 3.2. Variational Inference

Variational inference is a deterministic technique for finding a tractable posterior distribution, denoted by $q$, which approximates the true posterior, $p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{x})$, that is often intractable to compute. More specifically, by letting $\mathcal{Q}$ be a set of possible approximations of this posterior, variational inference solves the following optimization problem for any model in the conjugate exponential family:

$$\min_{q(\boldsymbol{\beta}, \boldsymbol{z}) \in \mathcal{Q}} KL\left(q(\boldsymbol{\beta}, \boldsymbol{z}) \,||\, p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{x})\right), \tag{4}$$

where *KL* denotes the Kullback–Leibler divergence between two probability distributions.

In the *mean field variational* approach, the approximation family $\mathcal{Q}$ is assumed to fully factorize. Following the notation of Hoffman et al. [26], we have that

$$q(\boldsymbol{\beta}, z|\boldsymbol{\lambda}, \boldsymbol{\phi}) = q(\boldsymbol{\beta}|\boldsymbol{\lambda}) \prod_{n=1}^{N} q(z_n|\boldsymbol{\phi}_n).$$

Furthermore, each factor of the variational distribution is assumed to belong to the same family as the model's complete conditionals:

$$\begin{aligned} \ln q(\boldsymbol{\beta}|\boldsymbol{\lambda}) &= h(\boldsymbol{\beta}) + \boldsymbol{\lambda}^T t(\boldsymbol{\beta}) - a_g(\boldsymbol{\lambda}) \\ \ln q(z_n|\boldsymbol{\phi}_n) &= h(z_n) + \boldsymbol{\phi}_n^T t(z_n) - a_l(\boldsymbol{\phi}_n). \end{aligned}$$

It can be seen that $\boldsymbol{\lambda}$ parameterizes the variational distribution of $\boldsymbol{\beta}$ while $\boldsymbol{\phi}$ plays the same role for the variational distribution of $z$.

To solve the minimization problem in Equation (4), the variational approach exploits the transformation:

$$\ln p(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{\alpha}) + KL\left(q(\boldsymbol{\beta}, z|\boldsymbol{\lambda}, \boldsymbol{\phi}) \,||\, p(\boldsymbol{\beta}, z|\boldsymbol{x})\right), \tag{5}$$

where $\mathcal{L}(\cdot|\cdot)$ is a *lower bound* of $\ln p(\boldsymbol{x})$ since *KL* is nonnegative. This lower bound can be written as

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{\alpha}) = \mathbb{E}_q[\ln p(\boldsymbol{x}|z, \boldsymbol{\beta})] - \mathbb{E}_q[KL\left(q(z|\boldsymbol{\phi}) \,||\, p(z|\boldsymbol{\beta})\right)] - KL\left(q(\boldsymbol{\beta}|\boldsymbol{\lambda}) \,||\, p(\boldsymbol{\beta}|\boldsymbol{\alpha})\right)]. \tag{6}$$

We introduce $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ in $\mathcal{L}$'s notation to make explicit the function's dependency on $\boldsymbol{x}$, the data sample, and $\boldsymbol{\alpha}$, the natural parameters of the prior over $\boldsymbol{\beta}$. As $\ln p(\boldsymbol{x})$ is constant, and minimizing the *KL* term is equivalent to maximizing the lower bound. Equation (6) shows the trade-off involved in the lower-bound. The first term measures the model's fit to the data and favors variational posterior mass concentrated around the maximum likelihood estimate. The second and third terms are regularization terms and favor variational posteriors close to their respective prior distributions.

This lower bound can be maximized, for example, by a coordinate ascent method that iteratively updates each individual variational distribution while holding the others fixed. As shown in [26], these iterative updating equations have the following closed-form solutions:

$$\boldsymbol{\lambda} = \boldsymbol{\alpha} + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\phi}_n}[(t(\boldsymbol{x}_n, z_n), 1)], \tag{7}$$

$$\boldsymbol{\phi}_n = \mathbb{E}_{\boldsymbol{\lambda}}[\eta_l(\boldsymbol{x}_n, \boldsymbol{\beta})], \tag{8}$$

where $\mathbb{E}_{\boldsymbol{\lambda}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\phi}_n}[\cdot]$ denote the expected value according to $q(\boldsymbol{\beta}|\boldsymbol{\lambda})$ and $q(z_n|\boldsymbol{\phi}_n)$, respectively. If the number of data points is large, alternative scalable methods can also be used [26,40].

### *3.3. Variational Inference over Data Streams*

As commented in the previous section, we envision a situation where the data stream is defined by a sequence of data batches generated at discrete points in time $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ and where each batch is composed by a set of data samples, $\boldsymbol{x}_t = \boldsymbol{x}_{t,i=1:N_t}$. As new batches arrive, we want to update the posterior distribution over the parameters of the model. This can be addressed by applying a Bayesian recursive approach:

$$p(\boldsymbol{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) \propto p(\boldsymbol{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}) \prod_{i=1}^{N_t} \int p(\boldsymbol{x}_{t,i}, z_{t,i}|\boldsymbol{\beta}) dz_{t,i}.$$

Hence, updating the posterior at time $t$ reduces to a problem of computing a posterior over $\boldsymbol{\beta}$ conditional on the data $\boldsymbol{x}_t$ given that $\boldsymbol{\beta}$ gets a prior equal to $p(\boldsymbol{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$, i.e., the posterior in the previous time step.

When the above posterior is intractable to compute, we can use the streaming variational Bayes (SVB) algorithm [27]. This algorithm translates the above recursive updating approach to the variational settings described in the previous sections. Firstly, it approximates the posterior in the previous time step with a variational approximation, $p(\boldsymbol{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}) \approx q(\boldsymbol{\beta}|\boldsymbol{\lambda}_{t-1})$, so that the new posterior is expressed as

$$p(\boldsymbol{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) \propto q(\boldsymbol{\beta}|\boldsymbol{\lambda}_{t-1}) \prod_{i=1}^{N_t} \int p(\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,i}|\boldsymbol{\beta}) d\boldsymbol{z}_{t,i}.$$

This posterior is still intractable due to integration. We therefore use variational inference to compute a new approximation $q(\boldsymbol{\beta}|\boldsymbol{\lambda}_t)$ to the posterior at time $t$. The variational parameters are given as the solution to the optimization problem:

$$(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t) = \arg\min_{\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t} \mathcal{L}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}),$$

which, similarly to the static case (see Equation (6)), can be solved by a coordinate ascent method that iteratively updates each individual variational distribution while holding the others fixed. The end result is defined by the following closed-form solutions:

$$\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \sum_{i=1}^{N_t} \mathbb{E}_{\boldsymbol{\phi}_{t,i}}[(t(\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,i}), 1)], \tag{9}$$

$$\boldsymbol{\phi}_{t,i} = \mathbb{E}_{\lambda}[\eta_l(\boldsymbol{x}_{t,i}, \boldsymbol{\beta})], \tag{10}$$

where $\mathbb{E}_{\lambda}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\phi}_{t,i}}[\cdot]$ denote the expected value according to $q(\boldsymbol{\beta}|\boldsymbol{\lambda}_t)$ and $q(\boldsymbol{z}_{i,t}|\boldsymbol{\phi}_{i,t})$, respectively.

### *3.4. Exponential Forgetting in Variational Inference*

Exponential forgetting is a classic technique [19] that allows to gradually *forget* past data and to put more focus on more recent data samples when learning from a nonstationary data stream. This idea is usually implemented by exponentially down-weighting the loss function term associated with each data sample so that data samples closer in time have more impact on the model that older data samples.

In probabilistic terms, exponential forgetting is achieved by using a log-likelihood function of the following form:

$$\ln p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t|\boldsymbol{\beta}) = \sum_{i=1}^{t} \rho^{t-i} \ln p(\boldsymbol{x}_i|\boldsymbol{\beta}) + cte,$$

where $\rho \in [0, 1]$ is the exponential decay weight and $cte$ is a constant term. By using a small $\rho$ value, we aggressively *forget* old data samples. However, $\rho$ values close to 1 tend to mildly *forget* previous data samples.

Similarly, in Bayesian learning settings [24], we can use this scheme to compute the posterior:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \rho) &\propto p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t|\boldsymbol{\beta}, \rho) p(\boldsymbol{\beta}) \\ &= p(\boldsymbol{x}_t|\boldsymbol{\beta}) p(\boldsymbol{x}_{t-1}|\boldsymbol{\beta})^{\rho} \cdots p(\boldsymbol{x}_1|\boldsymbol{\beta})^{\rho^{t-1}} p(\boldsymbol{\beta}). \end{aligned}$$

This scheme also applies to variational learning by considering this exponential down-weighted log-likelihood instead of the standard data log-likelihood, as used by [24]. Then, the *lower bound* function has the following form:

$$\mathcal{L}_\rho(\boldsymbol{\lambda}, \boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{\alpha}_u) = \mathbb{E}_q \left[ \sum_{i=1}^{t} \rho^{t-i} \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\beta}) \right] - KL\left(q(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{\lambda}, \boldsymbol{\phi}) \,||\, p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{\alpha}_u)\right), \tag{11}$$

where $\boldsymbol{\alpha}_u$ denotes the natural parameters of a non-informative prior, $p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{\alpha}_u)$.

The updating equation of the coordinate gradient ascent algorithm described in Equation (7) can now be expressed as follows [41]:

$$\boldsymbol{\lambda} = \boldsymbol{\alpha}_u + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E}_{\boldsymbol{\phi}_i}[(t(\boldsymbol{x}_i, \boldsymbol{z}_i), 1)]. \tag{12}$$

The main point here is to highlight how, at the convergence point, the variational solution $\boldsymbol{\lambda}$ exponentially down-weights the contribution (i.e., the expected sufficient statistics) of old data samples.

Exponential forgetting also addresses the problem of Bayesian learning over unbounded data streams. According to Equation (7), one of the components of the $\boldsymbol{\lambda}$ parameter corresponds to the *equivalent sample size* of the variational posterior ($ESS_{post}$). In this case, this value can be computed as

$$ESS_{post,t} = ESS_{prior} + \sum_{i=1}^{t} \rho^{(i-1)}.$$

If $\rho < 1$, then $ESS_{post,t}$ converges to a finite number:

$$\lim_{t\to\infty} ESS_{post,t} = ESS_{prior} + \frac{1}{1-\rho}, \tag{13}$$

avoiding the problem of having a degenerated Bayesian posterior distribution in the presence of an unbounded data stream. As noted in [30,42], this schema approximates a posterior distribution conditioned on the last $\frac{1}{1-\rho}$ data samples of the stream.

Stochastic varitional inference (SVI) [26] is a widely used variational learning algorithm for dealing with large data sets. Population variational Bayes [25] is a simple modification of SVI used when the total size of the data set is unknown. When these algorithms are applied in data streaming settings, they use a constant learning rate $\nu$ (It is usually set to small values like 0.1 or 0.01.), and the sequential updating equation of the global variational parameters $\boldsymbol{\lambda}$ can be written as

$$\boldsymbol{\lambda}_t = (1-\nu)\boldsymbol{\lambda}_{t-1} + \nu(\boldsymbol{\alpha}_u + S\mathbb{E}_{\boldsymbol{\phi}_t}[(t(\boldsymbol{x}_t, \boldsymbol{z}_t), 1)]), \tag{14}$$

where $S$ is equal the total size of the data set. This size is equal to $N$ in the case of SVI or to the size of the population $M$ in the case of PVB. By expanding this equation, we find that

$$\boldsymbol{\lambda}_t = (1-(1-\nu)^t)\boldsymbol{\alpha}_u + N\nu \sum_{i=1}^{t} (1-\nu)^{t-i} \mathbb{E}_{\boldsymbol{\phi}_i}[(t(\boldsymbol{x}_i, \boldsymbol{z}_i), 1)]. \tag{15}$$

The above equation highlights that SVI and PVB also exponentially down-weight old data samples, with a forgetting rate of $\rho = 1-\nu$ (compare the above equation with Equation (12)). Therefore, this is one of the mechanisms that these two methods use to adapt to drifts in the nonstationary data stream.

In the case of the PVB method, the parameter $M$ helps to adapt to drifts in the data set through the effect it has on computing $\boldsymbol{\phi}_i$, as discussed by [25]. However, when the model does not contain local random variables, the variational parameters $\boldsymbol{\phi}_i$ do not exist and, then, the size of the population does not play any role in adapting to drifts in the data stream.

## 4. Implicit Transition Models

In order to extend the model in Figure 1a to data streams, we may introduce a transition model $p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})$ to explicitly model the evolution of the parameters over time, enabling estimation of the predictive density at time $t$:

$$p(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}) = \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})p(\boldsymbol{\beta}_{t-1}|\boldsymbol{x}_{1:t-1})d\boldsymbol{\beta}_{t-1}. \tag{16}$$

However, this approach introduces two problems. First, in nonstationary domains, we may not have a single transition model or the transition model may be unknown. Secondly, if we seek to position the model within the conjugate exponential family in order to be able to compute the gradients of $\mathcal{L}$ in closed form, we need to ensure that the distribution family for $\boldsymbol{\beta}_t$ is its own conjugate distribution, thereby severely limiting the model's expressive power (e.g., we cannot assign a Dirichlet distribution to $\boldsymbol{\beta}_t$).

Rather than explicitly modeling the evolution of the $\boldsymbol{\beta}_t$ parameters as in Equation (16), we instead follow a similar approach to Kárnỳ [31] and Ozkan *et al.* [30] by defining the time evolution model *implicitly* by constraining the maximum *KL* divergence over consecutive parameter distributions. Specifically, by defining

$$p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}) = \int \delta(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})p(\boldsymbol{\beta}_{t-1}|\boldsymbol{x}_{1:t-1})d\boldsymbol{\beta}_{t-1} \tag{17}$$

one can restrict the space of possible distributions $p(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})$, supported by an unknown transition model, by the constraint

$$KL\left(p(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) \leq \kappa. \tag{18}$$

We then propose to approximate $p(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})$ by the distribution $\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})$ having minimum Kullback–Leibler divergence w.r.t. (with respect to) a prior density $p_u(\boldsymbol{\beta}_t)$. This approach ensures that we will not underestimate the uncertainty in the parameter distribution. The following result shows how this constraining optimization problem has an amenable solution.

**Theorem 1 (Implicit Transition Models).** *The density* $\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}, \rho_t)$ *which has minimum Kullback–Leibler divergence w.r.t. a prior density* $p_u(\boldsymbol{\beta}_t)$*,*

$$\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}, \rho_t) = \arg\min_q KL\left(q(\boldsymbol{\beta}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right)$$

*and which satisfies the constrain imposed in Equation* (18) *takes the form*

$$\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}, \rho_t) \propto p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})^{\rho_t} p_u(\boldsymbol{\beta}_t)^{(1-\rho_t)}, \tag{19}$$

*where* $0 \leq \rho_t \leq 1$ *is indirectly defined by Equation (18) and therefore depends on the user-defined parameter* $\kappa$*.*

**Proof.** The proof can be found in Appendix A. □

This approach defines transition models without having to make explicit assumptions about the parametric family. In consequence, it provides a generic off-the-self mechanism for defining transition models. In subsequent subsections, we provide an intuitive interpretation of this approach by establishing a direct relationship with exponential forgetting and power prior [28] approaches, which has been previously use to deal with nonstationary data streams. We deviate from the original formulation given in Kárnỳ [31] and Ozkan et al. [30], which proposes to approximate $p(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})$ by the distribution $\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})$ having maximum entropy under the constraint in Equation (18). However, in this case, they require that $p_u(\boldsymbol{\beta}_t)$ is defined as an *invariant measure* (i.e., $\mathbb{E}_q[\ln p_u(\boldsymbol{\beta}_t)]$ has to be

constant w.r.t. any density $q(\boldsymbol{\beta}_t)$). Note that, in case we employ a prior $p_u(\boldsymbol{\beta}_t)$ satisfying this property, Theorem 1 also chooses the maximum entropy distribution.

In our streaming data setting, we follow the *assumed density filtering* approach [43] and the SVB approach [27] and use the approximation $p(\boldsymbol{\beta}_{t-1}|\boldsymbol{x}_{1:t-1}) \approx q(\boldsymbol{\beta}_{t-1}|\boldsymbol{\lambda}_{t-1})$, where $q(\boldsymbol{\beta}_{t-1}|\boldsymbol{\lambda}_{t-1})$ is the variational distribution calculated in the previous time step. Using this approximation in Equations (16) and (17), we can express $p_\delta$ in terms of $\boldsymbol{\lambda}_{t-1}$, in which case, Equation (19) becomes

$$\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t) \propto p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})^{\rho_t} p_u(\boldsymbol{\beta}_t)^{(1-\rho_t)}, \tag{20}$$

which we use as the prior density for time step $t$. Now, if $p_u(\boldsymbol{\beta}_t)$ belongs to the same family as $q(\boldsymbol{\beta}_{t-1}|\boldsymbol{\lambda}_{t-1})$, then $\hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)$ will stay within the same family and have natural parameters $\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u$, where $\boldsymbol{\alpha}_u$ are the natural parameters of $p_u(\boldsymbol{\beta}_t)$. Therefore, we can write

$$\ln \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t) = \ln h(\boldsymbol{\beta}_t) + (\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u)\boldsymbol{t}(\boldsymbol{\beta}_t) - a_g(\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u) \tag{21}$$

where $h(\boldsymbol{\beta}_t)$ is the base measure, which does not depend on any parameter. Thus, under this approach, the transitioned posterior remains within the same exponential family, so we can enjoy the full flexibility of the conjugate exponential family (i.e., computing gradients of the $\mathcal{L}$ function in closed form), an option that would not be available if one were to explicitly specify a transition model as in Equation (16).

Therefore, at each time step, we simply have to solve the following variational problem, where only the prior changes with respect to the original SVB approach:

$$\arg\max_{\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t} \mathcal{L}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t|\boldsymbol{x}_t, \rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u). \tag{22}$$

We shall refer to the method outlined in this section as SVB with *power priors* (SVB-PP). The term *power prior* [28] will be explained in Section 4.2.

### *4.1. Exponential Forgetting as Implicit Transition Models*

In this section, we show that the exponential forgetting mechanism used in variational inference, as described in Section 3.4, is an implicit transition model with constant forgetting rate $\rho$.

The updating equation detailed in Equation (7) to optimize the lower-bound function described in Equation (6) can be easily adapted to optimize the lower-bound associated to the implicit transition models given in Equation (22). This new updating equation for implicit transition models can be expressed as follows:

$$\boldsymbol{\lambda}_t = \mathbb{E}_{\boldsymbol{\phi}_t}[(t(\boldsymbol{x}_t, \boldsymbol{z}_t), 1)] + \rho\boldsymbol{\lambda}_{t-1} + (1-\rho)\boldsymbol{\alpha}_u. \tag{23}$$

Expanding the above equation, we have

$$\boldsymbol{\lambda}_t = \sum_{i=1}^{t} \rho^{t-i}\mathbb{E}_{\boldsymbol{\phi}_i}[(t(\boldsymbol{x}_i, \boldsymbol{z}_i), 1)] + \boldsymbol{\alpha}_u, \tag{24}$$

which exactly matches exponentially down-weighting scheme of old data samples given in Equation (12). Therefore, it is clear that the classic technique of exponential forgetting, which was usually supported by heuristic arguments, has a sound interpretation in terms of implicit transition models.

### *4.2. Power Priors as Implicit Transition Models*

*Power priors* [28] is a widely used class of informative priors for dealing with situations in which historical data are available. Let $\boldsymbol{x}_0$ denote a previously obtained data set, and let $\boldsymbol{x}_1$ be our current data set. According to the *power priors* scheme [28], the posterior probability over the model parameters should be computed as

$$p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_0, \rho) \propto p(\boldsymbol{x}_1|\boldsymbol{\beta}) p(\boldsymbol{x}_0|\boldsymbol{\beta})^{\rho} p(\boldsymbol{\beta}), \tag{25}$$

where $\rho \in [0, 1]$ is a scalar parameter down-weighting the likelihood of historical data relative to the likelihood of the current data.

As stated in the following lemma, power priors can also be interpreted as implicit transition models.

**Lemma 1.** *The Bayesian updating scheme described by Figure 1b and Equation (19) but with $\rho_t$ fixed to a constant value is equivalent to the recursive application of the Bayesian updating scheme of power priors given in Equation (25).*

**Proof.** The proof can be found in Appendix A. □

This connection allows us to introduce well-known results of power priors [29], finding that

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_0, \rho) \quad = \quad & \arg\min_{r \in \mathcal{P}} \{\rho \; KL\left(r \,||\, p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_0, \rho = 1)\right) \\ & + (1 - \rho) \; KL\left(r \,||\, p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_0, \rho = 0)\right)\} \end{aligned}$$

where $\mathcal{P}$ denotes the set of all possible densities over $\boldsymbol{\beta}$. Citing [29], "*power priors minimize the convex combination of KL (Kullback–Leibler) divergences between two extremes: one in which no historical data is used and the other in which the historical data and current data are given equal weight.*"

## 5. Hierarchical Power Priors

In the approach taken by Ozkan et al. [30] (and, by extension, SVB-PP), the forgetting factor $\rho_t$ is user-defined. In this paper, we instead pursue a (hierarchical) Bayesian approach and introduce a prior distribution over $\rho_t$, allowing the distribution over $\rho_t$ (and thereby the forgetting mechanism) to adapt to the data stream.

### *5.1. A Hierarchical Prior over the Forgetting Rate $\rho$*

In this section, we extend the model in Figure 1a to also account for the dynamics of the data stream being modeled. We shall assume here that only the parameters $\boldsymbol{\beta}$ in Figure 1a are time-varying, which we will indicate with the subscript $t$, i.e., $\boldsymbol{\beta}_t$. The resulting model can be illustrated as in Figure 1b. We shall refer to models of this type as *hierarchical power prior* (HPP) models.

We will show in Section 5.3 that the exponential and normal distributions, both of which truncated to the interval [0,1], are valid alternatives as prior distributions, $p(\rho_t|\gamma)$. The densities of these distributions have the following forms:

$$p(\rho_t|\gamma) = \frac{\gamma \exp(-\gamma \rho_t)}{1 - \exp(-\gamma)}, \quad 0 \leq \rho_t \leq 1, \tag{26}$$

$$p(\rho_t|\mu, \sigma) = \frac{\exp\left(-(\rho_t - \mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}\left(\Phi(\frac{1-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma})\right)}, \quad 0 \leq \rho_t \leq 1, \tag{27}$$

where $\Phi$ represents the standard normal cumulative distribution function, $\mu \in \mathbb{R}$ (can be outside the interval $[0, 1]$), and $\sigma > 0$. Since the natural parameters of the normal distribution are $\left(\mu/\sigma^2, -1/(2\sigma^2)\right)$, it is sometimes convenient to parameterize in terms of the precision $\eta = 1/\sigma^2$ (the

reciprocal of the variance). Using the precision, the natural parameters become $(\mu\eta, -\eta/2)$. Notice that precision $\eta$ appears in both components.

Figure 2 plots examples of both densities for different values of the parameters. The truncated exponential allows to model a uniform prior and priors that either favors $\rho_t$ values close to 1 (i.e., non-forgetting past data) or $\rho_t$ values close to 0 (i.e., forgetting past data). The truncated normal distribution when using a mean equal to 0.5 tends to favor non-extreme $\rho_t$ values (i.e., partial forgetting of past data), where the variance parameter defines the strength of this belief.

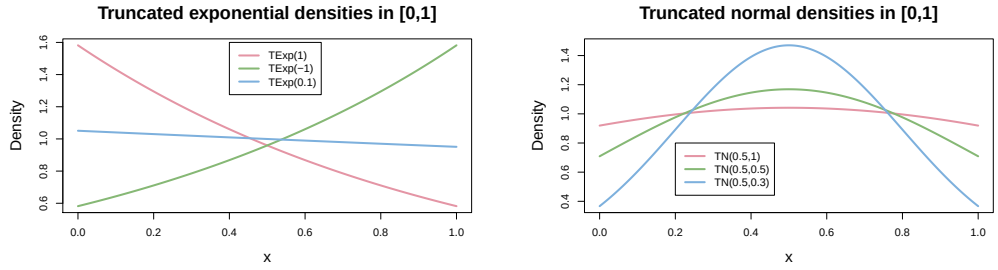



**Figure 2.** Density functions of the truncated exponential and the truncated normal distributions, respectively, for different values of their parameters.

For later use, we also detail here the equation for computing the expected value of $\rho_t$ for both distributions:

$$\mathbb{E}[\rho_t|\gamma] = \frac{1}{(1-e^{-\gamma})} - \frac{1}{\gamma}, \tag{28}$$

$$\mathbb{E}_q[\rho_t|\mu,\sigma] = \mu + \sigma\frac{\phi(\frac{-\mu}{\sigma}) - \phi(\frac{1-\mu}{\sigma})}{\Phi(\frac{1-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma})}, \tag{29}$$

where $\gamma$ is the mean parameter parameter of the truncated exponential, $\mu$ and $\sigma$ are the parameters of the truncated normal distribution in $[0,1]$, and $\phi$ and $\Phi$ are respectively the probability density function and the cumulative distribution function of the standard normal distribution.

### *5.2. The Double Lower Bound*

For updating the model distributions, we pursue a variational approach where we seek to maximize the evidence lower bound $\mathcal{L}$ in Equation (5) for time step $t$. However, since the model in Figure 1b does not define a conjugate exponential distribution due to the introduction of $p(\rho_t|\gamma)$, we cannot maximize $\mathcal{L}$ directly. Instead, we will derive a (double) lower bound $\hat{\mathcal{L}}$ (with $\hat{\mathcal{L}} \leq \mathcal{L}$) and use this lower bound as a proxy for updating the rules of the variational posteriors.

First, by instantiating the lower bound $\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1})$ in Equation (5) for the HPP model, we obtain

$$\begin{aligned}\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) = \mathbb{E}_q[\ln p(\boldsymbol{x}_t|\boldsymbol{Z}_t, \boldsymbol{\beta}_t)] - \mathbb{E}_q[KL\left(q(\boldsymbol{Z}_t|\boldsymbol{\phi}_t) \,||\, p(\boldsymbol{Z}_t|\boldsymbol{\beta}_t)\right)] \\ - \mathbb{E}_q[KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)\right)] \\ - KL\left(q(\rho_t|\boldsymbol{\omega}_t) \,||\, p(\rho_t|\gamma)\right)\end{aligned} \tag{30}$$

where $\boldsymbol{\omega}_t$ is the variational parameter of the variational distribution of $\rho_t$. For ease of presentation, we shall sometimes drop from $\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1})$ the subscript and the explicit specification of the parameters when this is otherwise clear from the context.

We now define the *double lower bound* $\hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1})$ as

$$\begin{aligned}\hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) = \mathbb{E}_q[\ln p(\boldsymbol{x}_t|\boldsymbol{Z}_t, \boldsymbol{\beta}_t)] - \mathbb{E}_q[KL\left(q(\boldsymbol{Z}_t|\boldsymbol{\phi}_t) \,||\, p(\boldsymbol{Z}_t|\boldsymbol{\beta}_t)\right)] \\ - \mathbb{E}_q[\rho_t] KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})\right) \\ - (1 - \mathbb{E}_q[\rho_t]) KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p(\boldsymbol{\beta}_t|\boldsymbol{\alpha}_u)\right) \\ - KL\left(q(\rho_t|\boldsymbol{\omega}_t) \,||\, p(\rho_t|\gamma)\right) \end{aligned} \tag{31}$$

which, according to Theorem 2, provides a lower bound for $\mathcal{L}$.

**Theorem 2.** *Let $\hat{\mathcal{L}}_{HPP}$ be as defined in Equation (31). Then,*

$$\hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) \leq \mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}).$$

**Proof.** The proof can be found in Appendix A. □

Even though Equation (31) defines an alternative objective function, when we compare this double lower bound with Equation (6), we can observe that the double lower bound still has the intuitive interpretation of the standard lower bound in terms of data fitting and Kullback–Leibler (KL) regularization. The only difference is that the KL regularization term associated to $q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)$ appears now as a convex combination of two KL terms, one regularizing with respect to $p(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})$ and the other regularizing with respect to $p(\boldsymbol{\beta}_t|\boldsymbol{\alpha}_u)$, whith $\mathbb{E}_q[\rho_t]$, acting as a combination factor.

Rather than seeking to maximize $\mathcal{L}$, we will instead maximize $\hat{\mathcal{L}}$; see Equation (31). Thus, maximizing $\hat{\mathcal{L}}$ w.r.t. the variational parameters $\boldsymbol{\lambda}_t$ and $\boldsymbol{\phi}$ also maximizes $\mathcal{L}$. By the same observation, we also have that the (natural) gradients are consistent relative to the two bounds, as stated by the next corollary.

**Corollary 1.**

$$\begin{aligned}\nabla^{nat}_{\boldsymbol{\lambda}_t} \hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) &= \nabla^{nat}_{\boldsymbol{\lambda}_t} \mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) \\ &= \nabla^{nat}_{\boldsymbol{\lambda}_t} \mathcal{L}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t | \boldsymbol{x}_t, \mathbb{E}_q[\rho_t]\boldsymbol{\lambda}_{t-1} + (1 - \mathbb{E}[\rho_t])\boldsymbol{\alpha}_u)\end{aligned}$$

*The same result holds for $\boldsymbol{\phi}_t$.*

**Proof.** The proof can be found in Appendix A. □

Thus, updating the variational parameters $\boldsymbol{\lambda}_t$ and $\boldsymbol{\phi}_t$ in HPP models can be done in the same way as for regular conjugate exponential models of the form in Figure 1. A pseudo-code description of the updating process can be found in Algorithm 1 when $\rho_t$ is assumed to follow a truncated exponential distribution.

**Algorithm 1** Streaming variational Bayes (SVB) with Hierarchical Power Priors (SVB-HPP).

**Input:** A data batch $\boldsymbol{x}_t$, the variational posterior in previous time step $\boldsymbol{\lambda}_{t-1}$.
**Output:** $(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t)$, a new update of the variational posterior.

```
λ_t ← λ_{t−1}.
E_q[ρ_t] ← 0.5.
Randomly initialize φ_t.
repeat
    (λ_t, φ_t) = arg min_{λ_t, φ_t} L(λ_t, φ_t | x_t, E_q[ρ_t] λ_{t−1} + (1 − E[ρ_t]) α_u)
    ω_t = KL(q(β_t|λ_t) || p_u(β_t)) − KL(q(β_t|λ_t) || p_δ(β_t|λ_{t−1})) + γ
    Update E_q[ρ_t] according to Equation (28) or Equation (29).
until convergence
return (λ_t, φ_t, ω_t)
```

In order to update $\boldsymbol{\omega}_t$, we rely on $\hat{\mathcal{L}}$, which we can maximize using the natural gradient w.r.t. $\boldsymbol{\omega}_t$ [44] and which can be calculated in closed form for a restricted distribution family for $\rho_t$, as stated in the following result.

**Lemma 2.** *Assuming that the first component of the sufficient statistics function for $\rho_t$ is the identity function, i.e., $\boldsymbol{t}_1(\rho_t) = \rho_t$, it holds that*

$$\begin{aligned} \frac{\partial^{nat}\hat{\mathcal{L}}}{\partial \boldsymbol{\omega}_{t,1}} =& KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right) - KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})\right) + \gamma_1 - \boldsymbol{\omega}_{t,1}, \\ \frac{\partial^{nat}\hat{\mathcal{L}}}{\partial \boldsymbol{\omega}_{t,k}} =& \gamma_k - \boldsymbol{\omega}_{t,k} \quad (k \neq 1). \end{aligned} \tag{32}$$

**Proof.** The proof can be found in Appendix A. □

From the above lemma, we can easily deduce that the truncated exponential distributions, for which the sufficient statistics are $t(\rho_t) = \rho_t$, and the truncated normal distribution, for which the sufficient statistics are $t(\rho_t) = (\rho_t, \rho_t^2)^T$, satisfy the criteria to be considered as hierarchical priors for $\rho_t$.

The problem of the above result is that, for $k \neq 1$, the optimal $\boldsymbol{\omega}_{t,k}$ is just equal to the the prior value, i.e., $\boldsymbol{\omega}_{t,k} = \gamma_k$. In the case of the truncated normal, which has a two-dimensional natural parameter vector, it would imply that the variance of the posterior $q(\rho_t|\boldsymbol{\omega}_t)$, denoted by $\sigma_q^2$, will be equal to the variance of prior, denoted by $\sigma_p^2$, which has to be set manually (We dropped the t-index in $\sigma_q^2$ for simplicity.). To address the issue of having to manually fix the variance of the truncated normal prior, $\sigma_p^2$, we employ an empirical Bayes approach and consider $\sigma_p^2$ as another free parameter of the double lower bound that we want to optimize. Therefore, we need to compute the gradient of the double lower bound w.r.t. this parameter:

$$\begin{aligned} \frac{\partial \hat{\mathcal{L}}}{\partial \sigma_p^2} &= \frac{\partial \gamma_1}{\partial \sigma_p^2}\frac{\partial \hat{\mathcal{L}}}{\partial \gamma_1} + \frac{\partial \gamma_2}{\partial \sigma_p^2}\frac{\partial \hat{\mathcal{L}}}{\partial \gamma_2} \\ &= -\frac{\mu_p}{\sigma_p^4}\left(\mathbb{E}[\rho_t|\mu_q, \sigma_q^2] - \mathbb{E}[\rho_t|\mu_p, \sigma_p^2]\right) + \frac{1}{2\sigma_p^4}\left(\mathbb{E}[\rho_t^2|\mu_q, \sigma_q^2] - \mathbb{E}[\rho_t^2|\mu_p, \sigma_p^2]\right), \end{aligned}$$

where $\gamma = (\mu_p/\sigma_p^2, -1/(2\sigma_p^2))$ is the natural parameter vector of the truncated normal prior, and $\mu_p$ and $\mu_q$ denote the mean of the truncated normal prior and posterior over $\rho_t$, respectively. We set $\mu_p$ to 0.5, trying to define a non-informative and symmetric prior.

Note that, in this case, we have a plain gradient instead of a natural gradient w.r.t. $\sigma_p^2$. Also, note that there is no closed-form solution for the stationary point of $\sigma_p^2$. To optimize along this direction, we use a simple gradient ascent approach with backtracking line-search to set the learning rate.

### *5.3. Towards a Measure of Concept Drift*

Observe that the form of the natural gradient of $\boldsymbol{\omega}_t$ given in Lemma 2 has an intuitive semantic interpretation in terms of measure of concept drift. If we follow a coordinate ascent algorithm, at every iteration, we should set

$$\boldsymbol{\omega}_t = KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right) - KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})\right) + \gamma. \tag{33}$$

Specifically, using the constant $\gamma$ as a threshold, we see that, if the uniform prior $p_u(\boldsymbol{\beta}_t)$ is closer to the variational posterior at time $t$, in terms of KL divergence than the variational posterior at the previous time step (i.e., $KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right) + \gamma < KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})\right)$), then we will get a negative value for $\boldsymbol{\omega}_t$.

This in turn implies that $\mathbb{E}_q[\rho] < 0.5$, according to Equations (28) and (29) (plotted in Figure 3), which means that we have a higher degree of forgetting for past data. If $\boldsymbol{\omega}_t > 0$, then $\mathbb{E}_q[\rho] > 0.5$

and less past data is forgotten. Figure 3 (left) graphically illustrates this trade-off, while Figure 4 shows a particular example of this situation using Gaussian posterior distributions.

The difference between the use of a truncated normal over a truncated exponential is that, with the former, the relation between $\omega_t$ and $\mathbb{E}_q[\rho_t]$ can be tuned by a change in the precision of the truncated normal prior, as it is graphically illustrated in Figure 3 (right). By using a prior with higher precision, we impose a stronger belief about the fact that $\rho_t$ values are close to neither 1 nor 0. In this way, the approach has the possibility to enforce smooth drift regimes.

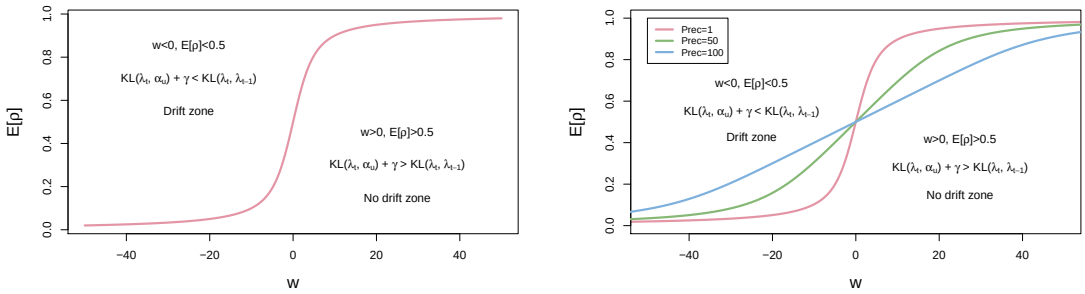



**Figure 3.** The relationship between $\omega_t$ and $\mathbb{E}_q[\rho_t]$ according to Equation (28) (**left**) and Equation (29) (**right**): see Section 5.3 for details.

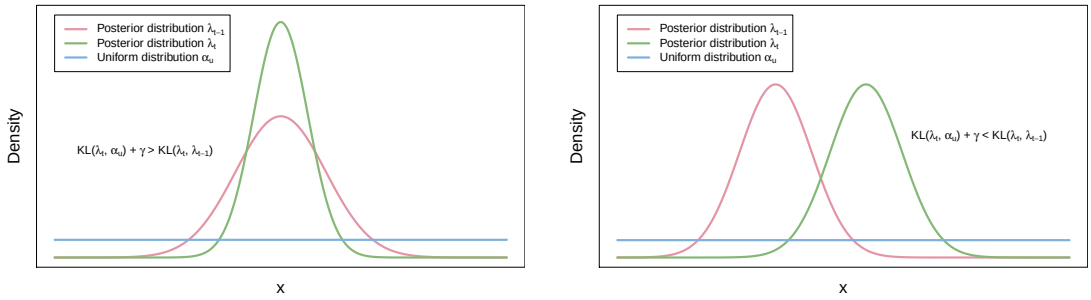



**Figure 4.** Example with Gaussian posterior distribution are shown. Two possible situations: no concept drift (**left**) when $\lambda_t$ is closer to $\lambda_{t-1}$ than to $\alpha_u$ (in terms of KL divergence); otherwise (**right**), there is concept drift. See Section 5.3 for details.

### *5.4. The Multiple Hierarchical Power Prior Model*

In this section, we propose a modification of our HPP model to deal with complex concept drift patterns which involve only a part of the model. For example, let us consider the application of an LDA model [4] for tracking over time the evolution of topics in a text corpus. Under these settings, a drift could eventually affect only a subset of the topics. Using our current approach, we might detect this drift and forget part of the data to adapt to the new situation and to learn the new topics. However, if some topics have not changed, we are losing information that could provide better estimations for these topics.

We propose an immediate extension of HPP, which include multiple power priors $\rho_t^{(i)}$, one for each parameter $\beta_i$. In this model, the $\rho_t^{(i)}$s are pair-wise independent. The latter ensures that optimizing $\hat{\mathcal{L}}$ can be performed as above, since the variational distribution for each $\rho_t^{(i)}$ can be updated independently of the other variational distributions over $\rho_t^{(j)}$ for $j \neq i$. This extended model allows local model substructures to have different forgetting mechanisms, thereby extending the expressiveness of the model. We shall refer to this extended model as a *multiple hierarchical power prior* (MHPP) model.

## 6. Results

In this section we will evaluate the following methods:

1. Streaming variational Bayes (**SVB**) as described in Section 3.3.
2. Four versions of Population Variational Bayes (**PVB**) (We do not compare with SVI because SVI is a special case of PVB when $M$ is equal the total size of the stream.) resulting from combining the values of the population size parameter $M = 1000$ (Section 6.1) and $M = 10,000$ (Section 6.2) with the learning rate values $\nu = 0.1$ and $\nu = 0.01$. In the four cases, the mini-batch size was set to 1000. Note that, however, for the LDA case, we set $M = 1000$ rather than $M = 10,000$ in Section 6.2 and use a mini-batch size of 1000 instead of 10,000.
3. Two versions of the SVB method with power priors (**SVB-PP**) or fixed exponential forgetting (as described in Section 4.1) with $\rho = 0.9$ or $\rho = 0.99$.
4. Three versions of our method based on the SVB method with adaptive exponential forgetting using hierarchical power priors (as described in Section 5):

    - **SVB-HPP-Exp** using a single shared $\rho$ with a truncated exponential distribution as a prior over $\rho$ with $\gamma = 0.1$ (i.e., close to uniform).
    - **SVB-MHPP-Exp** using separate $\rho^{(i)}$ for each parameter (as described in Section 5.4) with truncated exponential distributions as priors over each $\rho^{(i)}$ with $\gamma = 0.1$.
    - **SVB-MHPP-Norm** using also separate $\rho^{(i)}$ for each parameters but with truncated normal distributions as priors over each $\rho^{(i)}$. In this case, we use $\mu_p = 0.5$ and learn the variance $\sigma_p^2$ using the empirical Bayes approach described at the end of Section 5.2.

The underlying variational inference method used in the experiments is the variational message passing (VMP) algorithm [41] for all models; VMP was terminated after 100 iterations or if the relative increase in the lower bound fell below 0.01% (0.0001% for LDA). We considered *non-informative* priors, i.e., flat Gaussians, flat Gamma, or uniform Dirichlet (full details in Appendix B). For the LDA model [4], we use standard priors for this model which include a Dirichlet prior over topics with $\alpha = \frac{1}{|V|}$, where $|V|$ denotes the size of the vocabulary and another Dirichlet prior over topics assignments with $\alpha = 0.1$. Variational parameters were randomly initialized using the same seed for all methods.

### *6.1. Evaluation Using an Artificial Data Set*

First, we illustrate the behavior of the different approaches in a controlled experimental setting: We produced an artificial data stream by generating 100 samples (i.e., $|\boldsymbol{x}_t| = 100$) from a Binomial distribution at each time step. We artificially introduced concept drift by changing the parameter $p$ of the Binomial distribution: $p = 0.2$ for the first 30 time steps, then $p = 0.5$ for the following 30 time steps, and finally $p = 0.8$ for the last 40 time steps. The data stream was modeled using a beta-binomial model.

**Parameter estimation:** Figure 5 shows the evolution of $\mathbb{E}_q[\beta_t]$ for the different methods. We recognize that SVB simply generates a running average of the data, as it is not able to adapt to the concept drift. The results from PVB strongly depend on the learning rate $\nu$, where the highest learning rate, which results in the more aggressive forgetting, works better in this example. Recall, however, that $\nu$ has to be hand-tuned to achieve optimal performance. As expected, the choice of the size of the population $M$ for SVB does not have an impact because the present model has no local hidden variables (cf. Section 4.1). SVB-PP yields results almost identical to PVB when $\rho$ matches the learning rate of PVB (i.e., $\rho = 1 - \nu$). Finally, SVB-HPP provides the best results, almost mirroring the true model.

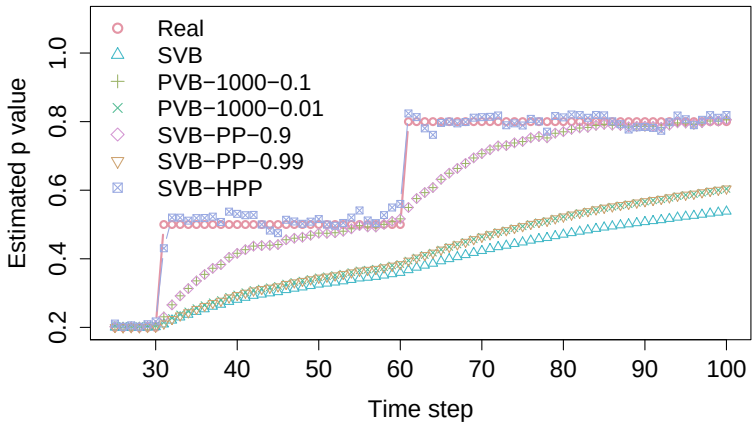



**Figure 5.** $E[\beta_t]$ in the beta-binomial model artificial data set: note that PVB-1000-0.1 overlaps with SVB-PP-0.9 and PVB-1000-0.01 overlaps with SVB-PP-0.99. The exact value is denoted with red circles.

**Equivalent sample size of the posterior ($ESS_{post}$):** Figure 6 (left) gives the evolution of the equivalent sample size of the posterior, $ESS_{post,t}$, for the different methods (For this model, $ESS_{post,t}$ is simply computed by summing the components of $\lambda_t$ defining the beta posterior.). $ESS_{post,t}$ of PVB is always given by the constant $M$. For SVB, $ESS_{post,t}$ monotonically increases as more data is used, while SVB-PP exhibits convergence to the limiting value computed in Equation (13). A different behaviour is observed for SVB-HPP: It is automatically adjusted. Notice that the values for this model are to be read off the alternative $y$-axis. We can detect the concept drift by identifying where ESS rapidly declines.

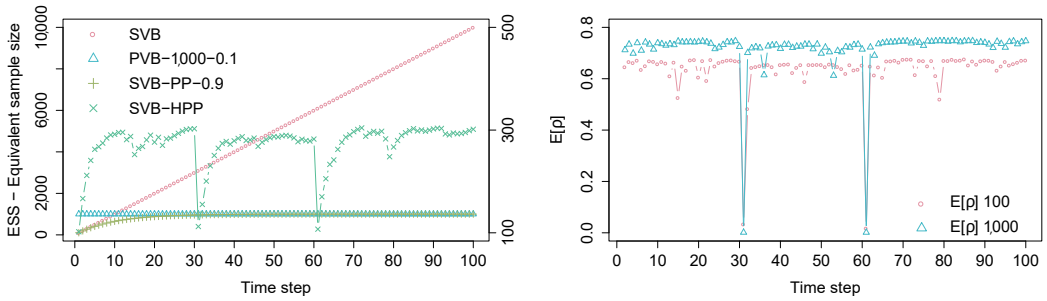



**Figure 6.** The results for the beta-binomial model artificial data set. Left panel: The equivalent sample size of the posterior, $ESS_{post,t}$, for the different methods; the values for SVB-HPP are shown on the right $y$-axis. Right panel: The expected values of $\rho_t$, $\mathbb{E}_q[\rho_t]$, for batches of size 100 and 1000, respectively.

**Evolution of expected forgetting factor:** In Figure 6 (right), the series denoted "$E[\rho] - 100$" shows the evolution of $\mathbb{E}_q[\rho_t]$ for the artificial data set. Notice how the model clearly identifies abrupt concept drift at time steps $t = 30$ and $t = 60$. The series denoted "$E[\rho] - 1000$" illustrates the evolution of the parameter when we increase the batch size to 1000 samples. We recognize a more confident assessment about the absence of concept drift as more data is made available.

### *6.2. Evaluation Using Real Data Sets*

For this evaluation, we consider four real data sets from four different domains:

**Electricity market** [45]: The data set describes the electricity market of two Australian states. It contains 45,312 instances of 6 attributes, including a class label comparing change of the electricity price related to a moving average of the last 24 h. Each instance in the data set represents 30 min of

trading; during our analysis, we created batches such that $\boldsymbol{x}_t$ contains all information associated with month $t$.

The data is analyzed using a Bayesian linear regression model. The binary class label is assumed to follow a Gaussian distribution in order to fit within the conjugate model class. Similarly, the marginal densities of the predictive attributes are also assumed to be Gaussian. The regression coefficients are given Gaussian prior distributions, and the variance is given a Gamma prior. Note that the overall distribution does not fall inside the conditional conjugate exponential family [26]; hence, we do not apply SVI (and PVB) in this setting.

**GPS** [46–48]: This data set contains 17,621 GPS trajectories (time-stamped $x$ and $y$ coordinates), totalling more than 4.5 million observations. To reduce the data size, we kept only one out of every ten measurements. We grouped the data so that $\boldsymbol{x}_t$ contains all data collected during hour $t$ of the day, giving a total of 24 batches of this stream.

Here, we employ a model with one independent Gaussian mixture model per day of the week, each mixture with 5 components. This enables us to track changes in the users' profiles across hours of the day and to monitor how the changes are affected by the day of the week.

**Finance** [5]: The data contains monthly aggregated information about the financial profile of around 50,000 customers over 62 (nonconsecutive) months. Three attributes were extracted per customer in addition to a class-label indicating whether the customer will default within the next 24 months.

We fit a naïve Bayes model to this data set, where the distribution at the leaf nodes is a 5-component mixture of Gaussians. The distribution over the mixture node is shared by all attributes but not between the two classes of customers.

**NIPS** [36]: This data set consists of the abstracts of published papers in the NIPS (Neural Information Processing Systems) conference between 1987 and 2015 (5804 documents in total). The data was preprocessed by choosing the most relevant individual terms across the whole dataset. This was done by ordering the words (11,463 in total) by their importance in the dataset, using the TF-IDF (term frequency-inverse document frequency) metric. The top 10 words after this filtering were *"policy"*, *"image"*, *"kernel"*, *"network"*, *"neurons"*, *"training"*, *"graph"*, *"images"*, *"matrix"*, and *"tree"*, while the last 5 words in the ranking were *"ralf"*, *"ciated"*, *"havior"*, *"references"*, and *"abstract"*. Only the top 100 words were kept, according to this criterion. In this way, we removed words that were not significant to tracking the concept drift in this data set. The documents were grouped by year, yielding a total of 29 batches of documents of different sizes. An LDA model with ten topics was employed to analyze the vocabulary and to detect changes in the evolution of the major topics of the papers of this conference every year. Note that the temporal extension of this model involves dealing with dynamics at the Dirichlet distributions over the topics. As commented in Section 2, there have been many previous approaches trying to deal with this problem [34–36], but none of them were applicable to general conjugate exponential family models and, in general, rely on much more complex inference schemes.

To evaluate the different methods discussed, we use the test marginal log-likelihood (TMLL). Specifically, each data batch is randomly split into a train data set, $\boldsymbol{x}_t$, and a test data set, $\tilde{\boldsymbol{x}}_t$, containing two thirds and one third of the data batch, respectively. Then, $\text{TMLL}_t$ is computed as $\text{TMLL}_t = \frac{1}{|\tilde{\boldsymbol{x}}_t|} \int p(\tilde{\boldsymbol{x}}_t, \boldsymbol{z}_t|\boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t|\boldsymbol{x}_t) d\boldsymbol{z}_t d\boldsymbol{\beta}_t$ (For LDA, $|\tilde{\boldsymbol{x}}_t|$ refers to the number of words in the test set; we then compute the so-called *per-word perplexity*.).

A detailed description of all models, including their structure and their variational families, is given in Appendix B.

### *6.3. Discussion of Results*

In this first part, we will highlight how the basic versions of SVB-HPP and SVB-MHPP outperform the rest of the approaches in most of the cases. Figure 7 shows for each method the difference between its $\text{TMLL}_t$ and that obtained by SVB (which is considered the baseline method). To improve readability, we only plot the results of the best performing method inside each group of methods. Figure 8

shows the development of $\mathbb{E}_q[\rho_t]$ over time for SVB-HPP-Exp, SVB-MHPP-Exp, and SVB-MHPP-Norm. For SVB-HPP-TExp, we only have one $\rho_t$-parameter, and its value is given by the solid line. SVB-MHPP utilizes one $\rho^{(i)}$ for each variational parameter. (The numbers of variational parameters are 14, 78, 33, and 10 for the electricity, GPS, financial, and NIPS models, respectively.) In this case, we plot $\mathbb{E}_q[\rho_t^{(i)}]$ at each point in time to indicate the variability between the different estimates throughout the series. We also report the average of the $\mathbb{E}_q[\rho_t^{(i)}]$ values at every time step. Finally, we compute the aggregated test marginal log-likelihood measure $\sum_{t=1}^{T} \text{TMLL}_t$ for each method and report these values in Table 1.

**Table 1.** Aggregated test marginal log-likelihood: maximum values for each data set are boldfaced.

| Data Set | SVB | PVB | | | | SVB-PP | | SVB-HPP | SVB-MHPP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | $\rho = 0.9$ | $\rho = 0.99$ | Exp | Exp | Norm |
| Electricity | −44.91 | −51.01 | −52.19 | −51.11 | −61.70 | −43.92 | −44.80 | −40.05 | −40.02 | **−39.91** |
| GPS | −1.98 | −2.10 | −2.77 | −1.97 | −4.49 | −1.94 | −1.97 | −1.97 | **−1.86** | **−1.86** |
| Finance | −19.84 | −22.29 | −22.57 | −20.40 | −20.73 | **−19.05** | −19.78 | −19.83 | −19.83 | −19.82 |
| NIPS | −4.07 | −4.04 * | −4.21 * | −4.01 | −4.12 | −4.02 | −4.06 | −4.01 | **−4.00** | **−4.00** |

Population variational Bayes (PVB) parameters: (1) $M = 10k, \nu = 0.1$; (2) $M = 10k, \nu = 0.01$; (3) $M = |\boldsymbol{x}_t|, \nu = 0.1$; and (4) $M = |\boldsymbol{x}_t|, \nu = 0.01$. (*) For NIPS, $M = 1k$ was used in (1) and (2).

For the electricity data set, we can see that the two proposed methods (SVB-HPP and SVB-MHPP) perform the best, as shown in Table 1. According to Figure 7 (electricity plot), all models are comparable during the first nine months, which is a period where our models detected no or very limited concept drift. However, after this period, both SVB-HPP and SVB-MHPP detected substantial drift and were able to adapt better than the other methods, which appeared unable to adjust to the complex concept drift structure in the latter part of the data. SVB-HPP and SVB-MHPP continued to behave at a similar level, mainly because when drift happened, it typically involved a high proportion of the parameters of the model (see the electricity plot in Figure 8).

For the GPS data set, we can observe how the SVB-MHPP perfoms quite well (see Table 1), particularly towards the end of the series (see the GPS plot in Figure 7). We can see in Figure 8 (GPS plot), that a significant proportion of the model parameters drifted (i.e., $\mathbb{E}_q[\rho_t^{(i)}] \leq 0.05$) at all times, while another proportion of the parameters showed a quite stable behavior ($\rho$-values above 0.9). This complex pattern is not properly captured by SVB-HPP, which ends up assuming no concept drift. PVB with $M = |\boldsymbol{x}_t|$ and $\nu = 0.1$ does well here, but it strongly depends on the hyperparameters, as in fact any other hyperparameter combination yields poorer results (see Table 1).

The financial data set shows a different behavior. As can be seen in Figure 7 (Finance plot), during the first months, no major differences among the methods were found. However, after month 30, SVB-PP with $\rho = 0.9$ is superior. Looking at the $E[\rho_t^{(i)}]$-values of SVB-MHPP in Figure 8 (Finance plot), we observe that there is remarkable concept drift in some of the parameters over the first few months. However, only a few parameters exhibit noteworthy drift after the first third of the sequence. Apparently, the simple SVB-PP approach has the upper hand when drift is constant and fairly limited, at least when the optimal forgetting factor $\rho$ has been identified.

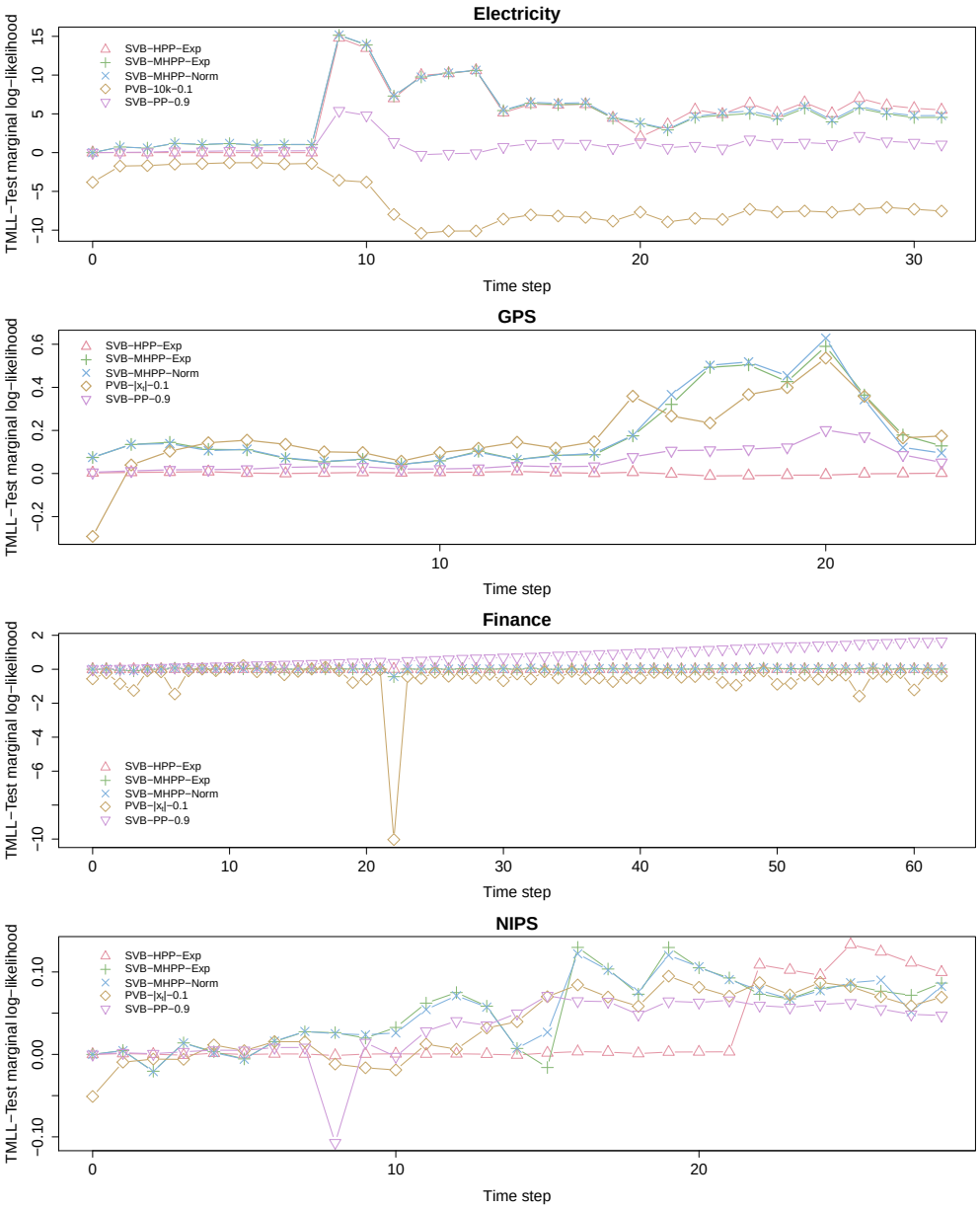



**Figure 7.** Results of the $\text{TMLL}_t$ improvement over SVB for the competing methods for the four real data sets.

In the case of the NIPS data set, we see again that HPP approaches capture drift in the data. As can be seen in Figure 7 (NIPS plot), SVB-HPP hardly detects any drift in the first 20 years and performs quite similarly to SVB (i.e., relative performance close to zero). However, in the last 10 years, SVB-HPP clearly outperforms SVB because it detects two strong drifts at years 23 and 29. Therefore, at these time steps, the whole LDA model is almost reestimated from scratch. In this case, SVB-MHPP is able to capture more fine-grained drifts in the data, as can been deduced from the $E[\rho_t^{(i)}]$-values of SVB-MHPP in Figure 8 (NIPS plot). Mainly, it detects changes in some topics while other topics remain constant over time. This allows SVB-MHPP to outperform SVB-HPP during some periods.

We have also observed that there are no major differences between SVB-MHPP-Exp and SVB-MHPP-Norm. Therefore, it seems that the inclusion of alternative priors does not have a big impact on the performance, at least if the variance parameter of the truncated normal is fixed automatically using an empirical Bayes approach. However, this is something that may require further investigations.

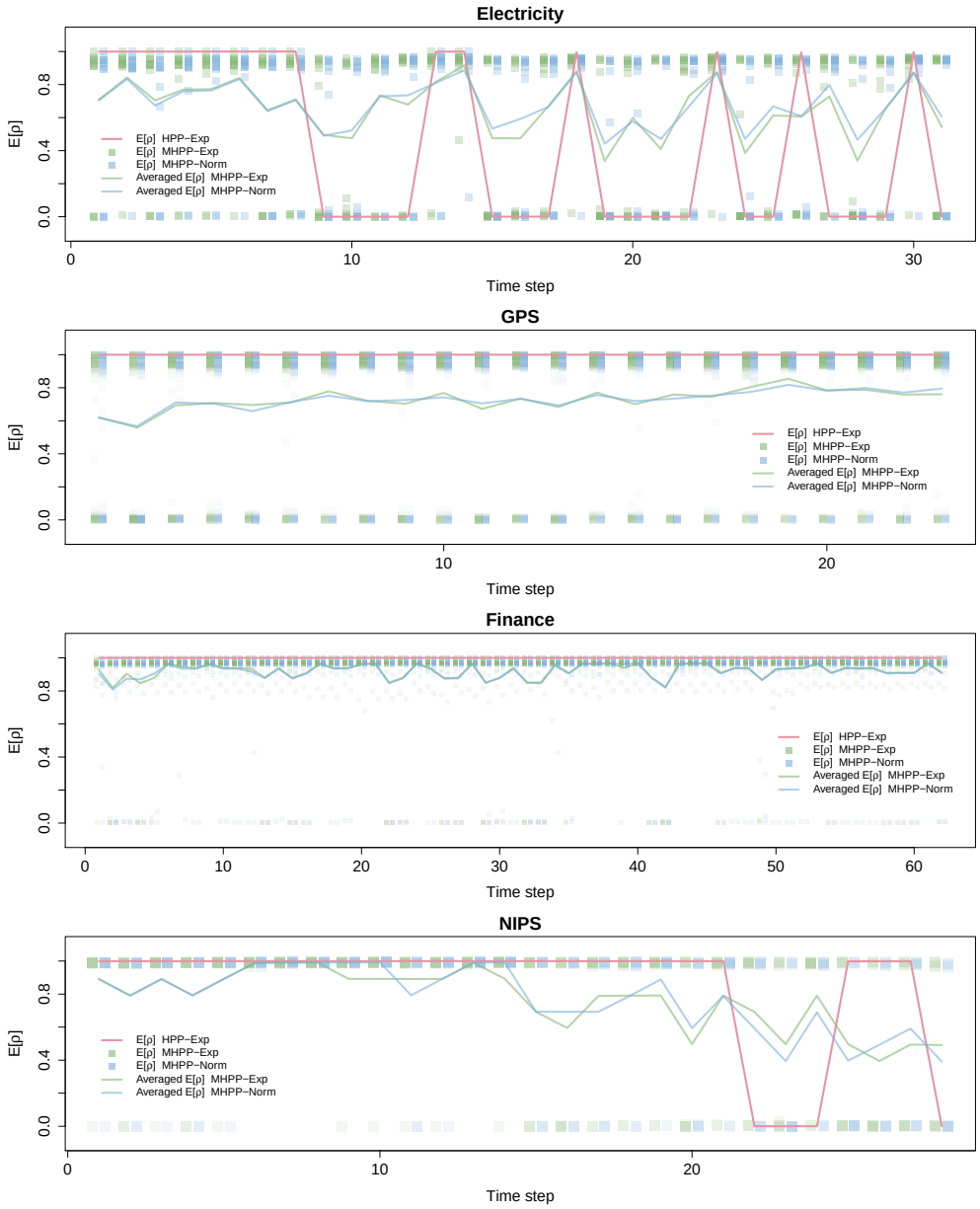



**Figure 8.** Evolution of $\mathbb{E}_q[\rho_t]$ for SVB-HPP and SVB-*multiple hierarchical power prior* (MHPP).

We conclude this section by highlighting that the performances of SVB-PP and PVB strongly depends on the hyperparameters of the model, cf. Table 1. As an example, consider SVB-PP and how its performance varies by changing the $\rho$ parameter. Similarly, PVB's performance is sensitive both to $\nu$ (see in particular the results for the GPS data) and $M$ (financial data). These hyperparameters are hard to fix, as their optimal values depend on data characteristics (see McInerney et al. [25] and Broderick et al. [27] for similar conclusions). We, therefore, believe that the fully Bayesian formulation is an important strong point of our approach.

## 7. Conclusions

In this paper, we have introduced a novel Bayesian approach for learning general latent variable models from nonstationary data streams. For this purpose, we introduce implicit transition models as a general method for transitioning the parameters of a latent variable model. We also show that previous approaches like exponential forgetting and power priors can be seen as specific cases of this general transition model. However, these approaches are only able to model slowly changing data

streams. Our approach is able to handle both abrupt and gradual drifts in the data stream by explicitly modeling the rate of change of the data stream. For this purpose, we introduce a novel hierarchical prior which allows the model to adapt to the different drifts that one can encounter in a data stream. We then develop an efficient variational inference scheme that optimizes a novel lower bound of the likelihood function.

As future work, we aim to provide a sound approach to semantically characterize concept drift by inspecting the $\mathbb{E}[\rho_t^{(i)}]$ values provided by SVB-MHPP and to investigate the effects in variational approximation introduced by the use of the double lower bound approximation.

**Author Contributions:** Conceptualization, A.R.M., A.S., H.L., and T.D.N.; methodology, A.R.M., D.R.-L., A.S., H.L., and T.D.N.; software, A.R.M. and D.R.-L.; validation, A.R.M. and D.R.-L.; formal analysis, A.R.M., D.R.-L., A.S., H.L., and T.D.N.; investigation, A.R.M., D.R.-L., A.S., H.L., and T.D.N.; resources, A.S., H.L., and T.D.N.; data curation, A.R.M. and D.R.-L.; writing—original draft preparation, A.R.M.; writing—review and editing, A.R.M., D.R.-L., A.S., H.L., and T.D.N.; visualization, D.R.-L.; supervision, A.S., H.L., and T.D.N.; project administration, A.S., H.L., and T.D.N.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript. .

**Funding:** This research was funded by the Spanish Ministry of Economy and Competitiveness through projects TIN2015-74368-JIN and TIN2016-77902-C3-3-P and by ERDF funds and by the Ministry of Science and Innovation through the project PID2019-106758GB-C32. A.M., D.R.L., and A.S. thank the support from the Center for Development and Transfer of Mathematical Research to Industry CDTIME (University of Almería). D.R.L. thanks also the research group FQM-229 (Junta de Andalucía) and the "Campus de Excelencia Internacional del Mar" CEIMAR (University of Almería).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proofs

**Proof of Theorem 1.** (The proof follows [30].) To solve the constrain optimization problem, we define the following Lagrangian:

$$J(q) = KL\left(q(\boldsymbol{\beta}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right) + \lambda\left(\int q(\boldsymbol{\beta}_t) d\boldsymbol{\beta}_t - 1\right) + \gamma\left(KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) - \kappa\right)$$

where the second component corresponds to the constrain imposing that $q$ must be a valid density and that the last term is the constraint given in Equation (18).

Taking a derivative w.r.t. to $q(\boldsymbol{\beta}_t)$ and regrouping, we have

$$\frac{\partial J(q)}{\partial q(\boldsymbol{\beta}_t)} = (1+\gamma)\ln q(\boldsymbol{\beta}_t) - \ln p_u(\boldsymbol{\beta}_t) + \gamma \ln p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}) + 1 + \lambda + \gamma$$

By setting the partial derivative to zero and regrouping, we have

$$\ln q(\boldsymbol{\beta}_t) = \frac{1}{1+\gamma}\ln p_u(\boldsymbol{\beta}_t) + \frac{\gamma}{1+\gamma}\ln p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1}) - \frac{1+\lambda+\gamma}{1+\gamma}$$

If we exponentiate both arguments, we get the first Karush–Kuhn–Tucker condition:

$$q(\boldsymbol{\beta}_t) = p_u(\boldsymbol{\beta}_t)^{\frac{1}{1+\gamma}} p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})^{\frac{\gamma}{1+\gamma}} e^{\frac{1+\gamma}{1+\lambda+\gamma}} \tag{A1}$$

The rest of Karush–Kuhn–Tucker conditions are

$$KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) \leq \kappa \tag{A2}$$

$$\gamma\left(KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) - \kappa\right) = 0 \tag{A3}$$

$$\int q(\boldsymbol{\beta}_t) d\boldsymbol{\beta}_t = 1 \tag{A4}$$

$$\gamma \geq 0. \tag{A5}$$

If $KL\left(p_u(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) \leq \kappa$, then we have that $q(\boldsymbol{\beta}_t) = p_u(\boldsymbol{\beta}_t)$ and $\gamma = 0$ satisfy Equations (A1)–(A4). If not, there must exist a $\gamma^\star > 0$ satisfying that $KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) = \kappa$ because we have that $q(\boldsymbol{\beta}_t)$ is smooth w.r.t. to $\gamma$; when $\gamma \to \infty$, we have that $KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) \to 0$; and when $\gamma \to 0$, we have that $KL\left(q(\boldsymbol{\beta}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{x}_{1:t-1})\right) > \kappa$. Then, Equations (A1)–(A4) are also satisfied by this $\gamma^\star$ value. Finally, we can deduce Equation (19) by setting $\rho_t = \frac{\gamma}{1+\gamma}$. □

**Proof of Lemma 1.** Translate the recursive Bayesian updating approach of power priors into an equivalent two time slice model, where $\boldsymbol{\beta}_0$ is given a prior distribution $p(\boldsymbol{\beta}_0)$ and $p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0)$ is a Dirac delta function. The distribution $p(\boldsymbol{\beta}_1|\boldsymbol{x}_0, \boldsymbol{x}_1, \rho)$ in this model is equivalent to $p(\boldsymbol{\beta}|\boldsymbol{x}_1, \boldsymbol{x}_0, \rho)$, which, in turn, is equivalent (up to proportionality) to $p(\boldsymbol{x}_1|\boldsymbol{\beta}_1)\hat{p}(\boldsymbol{\beta}_1|\boldsymbol{x}_0, \rho_t)$. Note that the last term can alternatively be expressed as $\hat{p}(\boldsymbol{\beta}_1|\boldsymbol{x}_0, \rho_t) \propto p_\delta(\beta_1|\boldsymbol{x}_0)^\rho p(\beta_1)^{1-\rho} \propto p_\delta(\boldsymbol{x}_0|\beta_1)^\rho p(\beta_1)$. □

**Proof of Theorem 2.** It follows from Equations (30) and (31) that

$$\hat{\mathcal{L}}_{HPP} - \mathcal{L}_{HPP} = \mathbb{E}_q[\ln \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)].$$

According to Equation (21), if we ignore the base measure, we can write

$$\mathbb{E}_q[\ln \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)] = \mathbb{E}_q[(\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u)\boldsymbol{t}(\boldsymbol{\beta}_t) - a_g(\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u)].$$

Since $a_g$ is convex [49], we have

$$a_g(\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u) \leq \rho_t a_g(\boldsymbol{\lambda}_{t-1}) + (1-\rho_t)a_g(\boldsymbol{\alpha}_u),$$

which combined with Equation (21) gives

$$\begin{aligned}
&\mathbb{E}_q[\ln p(\boldsymbol{x}_t, \boldsymbol{Z}_t|\boldsymbol{\beta}_t)] + \mathbb{E}_q[(\rho_t\boldsymbol{\lambda}_{t-1} + (1-\rho_t)\boldsymbol{\alpha}_u)\boldsymbol{t}(\boldsymbol{\beta}_t)\\
&- \rho_t a_g(\boldsymbol{\lambda}_{t-1}) - (1-\rho_t)a_g(\boldsymbol{\alpha}_u)] + \mathbb{E}_q[p(\rho_t|\boldsymbol{\gamma})]\\
&- \mathbb{E}_q[\ln q(\boldsymbol{Z}_t|\boldsymbol{\phi}_t)] - \mathbb{E}_q[q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)] - \mathbb{E}_q[q(\rho_t|\boldsymbol{\omega}_t)] \leq \mathcal{L}.
\end{aligned}$$

Finally, by exploiting the mean field factorization of $q$ and by using the exponential family form of $p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})$ and $p_u(\boldsymbol{\beta}_t)$, we get the desired result. □

**Proof of Corollary 1.** The first equality follows immediately from Equation (30) because the difference does not depend on $\boldsymbol{\lambda}_t$ and $\boldsymbol{\phi}_t$. The second equality holds because the difference between the natural gradients of $\mathcal{L}$ and $\mathcal{L}_{HPP}$ is equal to

$$\begin{aligned}
&\nabla^{nat}_{\boldsymbol{\lambda}_t}\mathcal{L}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t|\boldsymbol{x}_t, \boldsymbol{\alpha}_u) - \nabla^{nat}_{\boldsymbol{\lambda}_t}\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \boldsymbol{\omega}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}_{t-1}) =\\
&\qquad \nabla^{nat}_{\boldsymbol{\lambda}_t}\mathbb{E}_q[\ln p(\boldsymbol{\beta}_t|\boldsymbol{\alpha}_u)] - \nabla^{nat}_{\boldsymbol{\lambda}_t}\mathbb{E}_q[\ln \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)]
\end{aligned}$$

According to Equation (19), the above difference is null if we make the $\boldsymbol{\alpha}_u$ parameter of the $\mathcal{L}$ term equal to $\mathbb{E}_q[\rho_t]\boldsymbol{\lambda}_{t-1} + (1 - \mathbb{E}_q[\rho_t])\boldsymbol{\alpha}_u$. □

**Proof of Lemma 2.** Firstly, by ignoring the terms in $\hat{\mathcal{L}}$ (Equation (31)) that do not involve $\boldsymbol{\omega}_t$, we get

$$\begin{aligned}
\hat{\mathcal{L}}(\boldsymbol{\omega}_t) &= \mathbb{E}_q[\rho_t](\mathbb{E}_q[\ln(p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})) - \mathbb{E}_q[\ln p_u(\boldsymbol{\beta}_t)]) + \mathbb{E}_q[p(\rho_t|\boldsymbol{\gamma})] - \mathbb{E}_q[q(\rho_t|\boldsymbol{\omega}_t)]\\
&= \mathbb{E}_q[\rho_t](\mathbb{E}_q[\ln(p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})) - \mathbb{E}_q[\ln p_u(\boldsymbol{\beta}_t)]) + \boldsymbol{\gamma}^T\mathbb{E}_q[t[\rho_t]] - (\boldsymbol{\omega}_t^T\mathbb{E}_q[t[\rho_t]] - a_g(\boldsymbol{\omega}_t)) + cte
\end{aligned}$$

As we have assumed that the sufficient statistics function $\boldsymbol{t}(\rho_t)$ for $p(\rho_t|\boldsymbol{\gamma})$ and $q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)$ contains the identity function ($\boldsymbol{t}_1(\rho_t) = \rho_t$), we have

$$\hat{\mathcal{L}}(\boldsymbol{\omega}_t) = \begin{pmatrix} \mathbb{E}_q[\rho_t] \\ \mathbb{E}_q[t_{\neq 1}(\rho_t)] \end{pmatrix}^T \begin{pmatrix} (\mathbb{E}_q[\ln(p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})) - \mathbb{E}_q[\ln p_u(\boldsymbol{\beta}_t)]) + \boldsymbol{\gamma}_1 - \boldsymbol{\omega}_1 \\ \boldsymbol{\gamma}_{\neq 1} - \boldsymbol{\omega}_{\neq 1} \end{pmatrix} - a_g(\boldsymbol{\omega}_t) + cte$$

where the subindex $\neq 1$ refers to those subindexes different from 1.

Using the standard equality of exponential family distributions, $\mathbb{E}_q[\boldsymbol{t}(\rho_t)] = \nabla_{\boldsymbol{\omega}_t} a_g(\boldsymbol{\omega}_t)$, we have

$$\nabla_{\omega_t} \hat{\mathcal{L}} = \nabla^2_{\omega_t} a_g(\boldsymbol{\omega}_t) \begin{pmatrix} \mathbb{E}_q[\ln(p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})) - \ln p_u(\boldsymbol{\beta}_t)] + \boldsymbol{\gamma}_1 - \boldsymbol{\omega}_{t,1} \\ \boldsymbol{\gamma}_{\neq 1} - \boldsymbol{\omega}_{t,\neq 1} \end{pmatrix}.$$

We can now find the natural gradient by premultiplying $\nabla_{\omega_t} \hat{\mathcal{L}}$ by the inverse of the Fisher information matrix, which for the exponential family corresponds to the inverse of the Hessian of the log-normalizer:

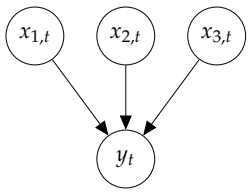
$$\begin{aligned} \hat{\nabla}_{\omega_t} \hat{\mathcal{L}} &= (\nabla^2_{\omega_t} a_g(\boldsymbol{\omega}_t))^{-1} \nabla_{\omega_t} \hat{\mathcal{L}} \\ &= \begin{pmatrix} \mathbb{E}_q[\ln(p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})) - \ln p_u(\boldsymbol{\beta}_t)] + \boldsymbol{\gamma}_1 - \boldsymbol{\omega}_{t,1} \\ \boldsymbol{\gamma}_{\neq 1} - \boldsymbol{\omega}_{t,\neq 1} \end{pmatrix}. \end{aligned}$$

Then, by introducing $q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) - q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)$ inside the expectation, we get the difference in Kullbach–Leibler divergence $KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_u(\boldsymbol{\beta}_t)\right) - KL\left(q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t) \,||\, p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})\right)$. □

## Appendix B. Probabilistic Models Used in the Experimental Evaluation

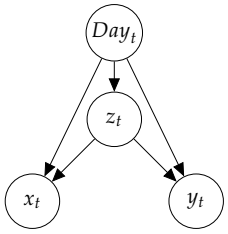
We provide a (simplified) graphical description of the probabilistic models used in the experiments. We also detail the distributional assumptions of the parameters, which are then used to define the variational approximation family.

### *Appendix B.1. Electricity Model*



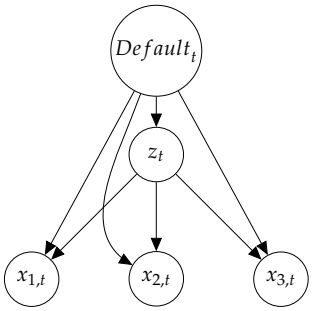

$$\begin{aligned} (\mu_i, \gamma_i) &\sim NormalGamma(1, 1, 0, 1e-10) \\ \gamma &\sim Gamma(1, 1) \\ b_i &\sim \mathcal{N}(0, +\infty) \\ x_{i,t} &\sim \mathcal{N}(\mu_i, \gamma_i) \\ y_t &\sim \mathcal{N}\left(b_0 + \sum_i b_i x_{i,t}, \gamma\right) \end{aligned}$$

*Appendix B.2. GPS Model*




$$
\begin{aligned}
p &\sim Dirichlet(1,\ldots,1)\\
p_k &\sim Dirichlet(1,\ldots,1)\\
(\mu_{j,k}^{(x)},\gamma_{j,k}^{(x)}) &\sim NormalGamma(1,1,0,1e-10)\\
(\mu_{j,k}^{(y)},\gamma_{j,k}^{(y)}) &\sim NormalGamma(1,1,0,1e-10)\\
Day_t &\sim Multinomial(p)\\
(z_t|Day_t=k) &\sim Multinomial(p_k)\\
(x_t|z_t=j,Day_t=k) &\sim \mathcal{N}(\mu_{j,k}^{(x)},\gamma_{j,k}^{(x)})\\
(y_t|z_t=j,Day_t=k) &\sim \mathcal{N}(\mu_{j,k}^{(y)},\gamma_{j,k}^{(y)})
\end{aligned}
$$

*Appendix B.3. Financial Model*




$$
\begin{aligned}
p &\sim Dirichlet(1,\ldots,1)\\
p_k &\sim Dirichlet(1,\ldots,1)\\
(\mu_{i;j,k},\gamma_{i;j,k}) &\sim NormalGamma(1,1,0,1e-10)\\
Default_t &\sim Binomial(p)\\
(z_t|Default_t=k) &\sim Multinomial(p_k)\\
(x_{i,t}|z_t=j,Day_t=k) &\sim \mathcal{N}(\mu_{i;j,k},\gamma_{i;j,k})
\end{aligned}
$$

## References


1. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109.
2. Gelfand, A.E.; Smith, A.F. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409.
3. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 1–37.
4. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
5. Borchani, H.; Martínez, A.M.; Masegosa, A.R.; Langseth, H.; Nielsen, T.D.; Salmerón, A.; Fernández, A.; Madsen, A.L.; Sáez, R. Modeling concept drift: A probabilistic graphical model based approach. In *International Symposium on Intelligent Data Analysis*; Springer: Cham, Switzerland , 2015; pp. 72–83.
6. Rabiner, L.R.; Juang, B.H. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16.
7. Bishop, C.M. Latent variable models. In *Learning in Graphical Models*; Springer: Dordrecht, The Netherlands, 1998; pp. 371–403.
8. Blei, D.M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 203–232.
9. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
10. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622.
11. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
12. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
13. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877.
14. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994; Volume 2.

15. Triantafyllopoulos, K. Inference of dynamic generalized linear models: On-line computation and appraisal. *Int. Stat. Rev.* **2009**, *77*, 430–450.
16. Chen, X.; Irie, K.; Banks, D.; Haslinger, R.; Thomas, J.; West, M. Scalable Bayesian Modeling, Monitoring, and Analysis of Dynamic Network Flow Data. *J. Am. Stat. Assoc.* **2018**, *113*, 519–533.
17. Aminikhanghahi, S.; Cook, D.J. A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **2017**, *51*, 339–367.
18. Adams, R.P.; MacKay, D.J. Bayesian online changepoint detection. *arXiv* **2007**, arXiv:0710.3742 .
19. Gaber, M.M.; Zaslavsky, A.; Krishnaswamy, S. Mining data streams: A review. *ACM Sigmod Rec.* **2005**, *34*, 18–26.
20. Aggarwal, C.C. *Data Streams: Models and Algorithms*; Springer: New York, NY, USA, 2007; Volume 31.
21. Gama, J.; Rodrigues, P.P. An overview on mining data streams. In *Foundations of Computational, Intelligence*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 6, pp. 29–45.
22. Aggarwal, C.C. *Managing and Mining Sensor Data*; Springer: New York, NY, USA, 2013.
23. Papadimitriou, S.; Sun, J.; Faloutsos, C. Streaming pattern discovery in multiple time-series. In Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, Trondheim, Norway, 30 August–2 September 2005; pp. 697–708.
24. Honkela, A.; Valpola, H. On-line variational Bayesian learning. In Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 1–4 April 2003; pp. 803–808.
25. McInerney, J.; Ranganath, R.; Blei, D. The population posterior and Bayesian modeling on streams. In *Advances in Neural Information Processing Systems 28*; Curran Associates, Inc.: Montreal, QC, Canada, 2015; pp. 1153–1161.
26. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
27. Broderick, T.; Boy, N.; Wibisono, A.; Wilson, A.C.; Jordan, M.I. Streaming variational Bayes. In *Advances in Neural Information Processing Systems 26*; Curran Associates, Inc.: Lake Tahoe, NV, USA, 2013; pp. 1727–1735.
28. Ibrahim, J.G.; Chen, M.H. Power prior distributions for regression models. *Stat. Sci.* **2000**, *15*,46–60.
29. Ibrahim, J.G.; Chen, M.H.; Sinha, D. On optimality properties of the power prior. *J. Am. Stat. Assoc.* **2003**, *98*, 204–213.
30. Ozkan, E.; Smidl, V.; Saha, S.; Lundquist, C.; Gustafsson, F. Marginalized adaptive particle filtering for nonlinear models with unknown time-varying noise parameters. *Automatica* **2013**, *49*, 1566–1575.
31. Kárnỳ, M. Approximate Bayesian recursive estimation. *Inf. Sci.* **2014**, *285*, 100–111.
32. Shi, T.; Zhu, J. Online Bayesian passive-aggressive learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–39 .
33. Williamson, S.; Orbanz, P.; Ghahramani, Z. Dependent Indian buffet processes. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 924–931.
34. Blei, D.M.; Lafferty, J.D. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 113–120.
35. Williamson, S.; Wang, C.; Heller, K.; Blei, D. The IBP compound Dirichlet process and its application to focused topic modeling. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
36. Perrone, V.; Jenkins, P.A.; Spano, D.; Teh, Y.W. Poisson random fields for dynamic feature models. *J. Mach. Learn. Res.* **2017**, *18*, 1–45.
37. Brown, L.D. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*; Institute of Mathematical Statistics: Hayward, CA, USA, 1986.
38. Bernardo, J.M.; Smith, A.F. *Bayesian Theory*; John Wiley & Sons: New York, NY, USA, 2009; Volume 405.
39. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.* **1995**, *20*, 197–243.
40. Masegosa, A.R.; Martinez, A.M.; Langseth, H.; Nielsen, T.D.; Salmerón, A.; Ramos-López, D.; Madsen, A.L. Scaling up Bayesian variational inference using distributed computing clusters. *Int. J. Approx. Reason.* **2017**, *88*, 435–451.
41. Winn, J.M.; Bishop, C.M. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.

42. Olesen, K.G.; Lauritzen, S.L.; Jensen, F.V. aHUGIN: A system creating adaptive causal probabilistic networks. In *Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence, Stanford, CA, USA, 17–19 July 1992*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1992; pp. 223–229.
43. Lauritzen, S.L. Propagation of probabilities, means, and variances in mixed graphical association models. *J. Am. Stat. Assoc.* **1992**, *87*, 1098–1108.
44. Sato, M.A. Online model selection based on the variational Bayes. *Neural Comput.* **2001**, *13*, 1649–1681.
45. Harries, M. *Splice-2 Comparative Evaluation: Electricity Pricing*; NSW-CSE-TR-9905; School of Computer Siene and Engineering, The University of New South Wales: Sydney, Australia, 1999.
46. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.Y. Understanding mobility based on GPS data. In *Proceedings of the 10th International Conference on Ubiquitous Computing, UbiComp '08, Seoul, Korea, 21–24 September 2008*; ACM: New York, NY, USA, 2008; pp. 312–321, doi:10.1145/1409635.1409677.
47. Zheng, Y.; Zhang, L.; Xie, X.; Ma, W.Y. Mining interesting locations and travel sequences from GPS trajectories. In P*roceedings of the 18th International Conference on World Wide Web, WWW '09, Madrid, Spain, 20–24 April 2009*; ACM: New York, NY, USA, 2009; pp. 791–800, doi:10.1145/1526709.1526816.
48. Zheng, Y.; Xie, X.; Ma, W.Y. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* **2010**, *33*, 32–39.
49. Wainwright, M.J.; Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends® Mach. Learn.* **2008**, *1*, 1–305.