

---

# Bayesian Models of Data Streams with Hierarchical Power Priors

---

Andrés Masegosa<sup>1,2</sup> Thomas D. Nielsen<sup>3</sup> Helge Langseth<sup>2</sup> Darío Ramos-López<sup>1</sup> Antonio Salmerón<sup>1</sup>  
Anders L. Madsen<sup>3,4</sup>

## Abstract

Making inferences from data streams is a pervasive problem in many modern data analysis applications. But it requires to address the problem of continuous model updating, and adapt to changes or drifts in the underlying data generating distribution. In this paper, we approach these problems from a Bayesian perspective covering general conjugate exponential models. Our proposal makes use of non-conjugate hierarchical priors to explicitly model temporal changes of the model parameters. We also derive a novel variational inference scheme which overcomes the use of non-conjugate priors while maintaining the computational efficiency of variational methods over conjugate models. The approach is validated on three real data sets over three latent variable models.

## 1. Introduction

Flexible and computationally efficient models for streaming data are required in many machine learning applications, and in this paper we propose a new class of models for these situations. Specifically, we are interested in models suitable for domains that exhibit changes in the underlying generative process (Gama et al., 2014). We envision a situation, where one receives batches of data at discrete points in time. As each new batch arrives, we want to glean information from the new data, while also retaining relevant information from the historical observations.

Our modelling is inspired by previous works on *Bayesian recursive estimation* (Özkan et al., 2013; Kárný, 2014), *power priors* (Ibrahim & Chen, 2000) and exponential for-

getting approaches (Honkela & Valpola, 2003). However, all of these methods were developed for slowly changing processes, where the rate of change anticipated by the model is controlled by a quantity that must be set manually. Our solution, on the other hand, can accommodate both gradual and abrupt concept drift by continuously assessing the similarity between new and historic data using a fully Bayesian paradigm.

Building Bayesian models for data streams raises computational problems, as data may arrive with high velocity and is unbounded in size. We therefore develop an approximate variational inference technique based on a novel lower-bound of the data likelihood function. The appropriateness of the approach is investigated through experiments using both synthetic and real-life data, giving encouraging results. The proposed methods are released as part of an open-source toolbox for scalable probabilistic machine learning (<http://www.amidsttoolbox.com>) (Masegosa et al., 2017; 2016b; Cabañas et al., 2016).

## 2. Preliminaries

In this paper we focus on conjugate exponential Bayesian network models for performing Bayesian learning on streaming data. To simplify the presentation, we shall initially focus on the model structure shown in Figure 1 (a). This model includes the observed data  $\mathbf{x} = \mathbf{x}_{i=1:N}$ , global hidden variables (or parameters)  $\beta = \beta_{1:M}$ , a set of local hidden variables  $\mathbf{z} = \mathbf{z}_{1:N}$ , and a vector of fixed (hyper) parameters denoted by  $\alpha$ . Notice how the dynamics of the process is not included in the model of Figure 1 (a); the model will be set in the context of data streams in Section 4, where we extend it to incorporate explicit dynamics over the (global) parameters to capture concept drift.

With the conditional distributions in the model belonging to the exponential family, we have that all distributions are of the following form

$$\ln p(Y|\text{pa}(Y)) = \ln h_Y + \eta_Y(\text{pa}(Y))^T \mathbf{t}_Y(Y) - a_Y(\eta_Y(\text{pa}(Y))),$$

where  $\text{pa}(Y)$  denotes the parents of  $Y$  in the directed acyclic graph of the induced Bayesian network model. The scalar functions  $h_Y$  and  $a_Y(\cdot)$  are the base measure and

---

<sup>1</sup>Department of Mathematics, University of Almería, Almería, Spain <sup>2</sup>Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway <sup>3</sup>Department of Computer Science, Aalborg University, Aalborg, Denmark <sup>4</sup>Hugin Expert A/S, Aalborg, Denmark. Correspondence to: Andrés Masegosa <andresmasegosa@ual.es>.

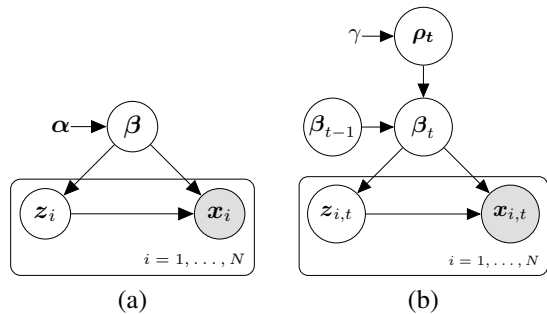


Figure 1. Left figure displays the core of the probabilistic model examined in this paper. Right figure includes a temporal evolution model for  $\beta_t$  as described in Section 4.

the log-normalizer, respectively; the vector functions  $\eta_Y(\cdot)$  and  $t_Y(\cdot)$  are the *natural parameters* and the *sufficient statistics* vectors, respectively. The subscript  $Y$  means that the associated functional forms may be different for the different factors of the model, but we may remove the subscript when clear from the context. By also requiring that the distributions are conjugate, we have that the posterior distribution for each variable in the model has the same functional form as its prior distribution. Consequently, learning (i.e. conditioning the model on observations) only changes the values of the parameters of the model, and not the functional form of the distributions.

Variational inference is a deterministic technique for finding tractable posterior distributions, denoted by  $q$ , which approximates the Bayesian posterior,  $p(\beta, z | \mathbf{x})$ , 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:

$$\min_{q(\beta, z) \in \mathcal{Q}} KL(q(\beta, z) | p(\beta, z | \mathbf{x})), \quad (1)$$

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. Extending the notation of Hoffman et al. (2013), we have that

$$q(\beta, z | \lambda, \phi) = \prod_{k=1}^M q(\beta_k | \lambda_k) \prod_{i=1}^N \prod_{j=1}^J q(z_{i,j} | \phi_{i,j}),$$

where  $J$  is the number of local hidden variables, which is assumed fixed for all  $i = 1, \dots, N$ . The parameterizations of the variational distributions are made explicit, in that  $\lambda$  parameterize the variational distribution of  $\beta$ , while  $\phi$  has the same role for the variational distribution of  $z$ .

To solve the minimization problem in Equation (1), the

variational approach exploits the transformation

$$\ln P(\mathbf{x}) = \mathcal{L}(\lambda, \phi | \mathbf{x}, \alpha_u) + KL(q(\beta, z | \lambda, \phi) | p(\beta, z | \mathbf{x})), \quad (2)$$

where  $\mathcal{L}(\cdot | \cdot)$  is a *lower bound* of  $\ln P(\mathbf{x})$  since  $KL$  is non-negative.  $\mathbf{x}$  and  $\alpha_u$  are introduced in  $\mathcal{L}$ 's notation to make explicit the function's dependency on  $\mathbf{x}$ , the data sample, and  $\alpha_u$ , the natural parameters of the prior over  $\beta$ . As  $\ln P(\mathbf{x})$  is constant, minimizing the  $KL$  term is equivalent to maximizing the lower bound. Variational methods maximize this lower bound by applying a coordinate ascent that iteratively updates the individual variational distributions while holding the others fixed (Winn & Bishop, 2005). The key advantage of having a conjugate exponential model is that the gradients of the  $\mathcal{L}$  function can be always computed in closed form (Winn & Bishop, 2005).

### 3. Related Work

Bayesian inference on streaming data has been widely studied (Ahmed et al., 2011; Doucet et al., 2000; Yao et al., 2009). In the context of variational inference, there are two main approaches. Ghahramani & Attias (2000); Broderick et al. (2013) propose recursive Bayesian updating of the variational approximation. The streaming variational Bayes (SVB) algorithm (Broderick et al., 2013) is the most known approach of this category. Alternatively, one could cast the inference problem as a stochastic optimization problem. Stochastic variational inference (SVI) (Hoffman et al., 2013) and the closely related population variational Bayes (PVB) (McInerney et al., 2015) are prominent examples from this group. 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 this 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 the time.  $M$  parameterizes the size of the population, and helps control the variance of the population posterior. Unfortunately,  $M$  must be specified by the user. No clear rule exists regarding how to set it, and McInerney et al. (2015) show that its optimal value may differ from one data stream to another.

The problem of Bayesian modeling of non-stationary data streams (i.e., with concept drift (Gama et al., 2014)) is not addressed by SVB, as it assumes data exchangeability. An online variational inference method, which exponentially forgets the variational parameters associated with old data, was proposed by Honkela & Valpola (2003). The so-called *power prior* approach (Ibrahim & Chen, 2000) is also based on an exponential forgetting mechanisms, and has nice theoretical properties (Ibrahim et al., 2003). Nevertheless, both approaches rely on a hyper-parameter determining forgetting, which has to be set manually. PVB can

also adapt to concept drift, because the variance of the variational posterior never decreases below a given threshold indirectly controlled by  $M$ , but again, the hyper-parameter has to be set manually.

A time series based modeling approach for concept drift using implicit transition models was pursued by Özkan et al. (2013); Kárný (2014). Unfortunately, the implicit transition model depends on a hyper-parameter determining the forgetting-factor, which has to be manually set. In this paper we build on this approach, adapt it to variational settings, and place a hierarchical prior on its forgetting parameter. This greatly improves the flexibility and accuracy of the resulting model when making inferences over drifting data streams.

## 4. Hierarchical Power Priors

In this section we extend the model in Figure 1 (a) to also account for the dynamics of the data stream being modeled. We shall here assume that only the parameters  $\beta$  in Figure 1 (a) are time-varying, which we will indicate with the subscript  $t$ , i.e.,  $\beta_t$ . First we briefly describe the approach on which the proposed model is based. Afterwards, we introduce the hierarchical power prior and detail a variational inference procedure for this model class.

### 4.1. Power Priors as Implicit Transition Models

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

$$p(\beta_t|\mathbf{x}_{1:t-1}) = \int p(\beta_t|\beta_{t-1})p(\beta_{t-1}|\mathbf{x}_{1:t-1})d\beta_{t-1}. \quad (3)$$

However, this approach introduces two problems. First of all, in non-stationary 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  $\beta_t$  is its own conjugate distribution, thereby severely limiting model expressivity (we can, e.g., not assign a Dirichlet distribution to  $\beta_t$ ).

Rather than explicitly modeling the evolution of the  $\beta_t$  parameters as in Equation (3), we instead follow the approach of Kárný (2014) and Özkan et al. (2013) who define the time evolution model implicitly by constraining the maximum  $KL$  divergence over consecutive parameter distributions. Specifically, by defining

$$p_\delta(\beta_t|\mathbf{x}_{1:t-1}) = \int \delta(\beta_t - \beta_{t-1})p(\beta_{t-1}|\mathbf{x}_{1:t-1})d\beta_{t-1} \quad (4)$$

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

$$KL(p(\beta_t|\mathbf{x}_{1:t-1}), p_\delta(\beta_t|\mathbf{x}_{1:t-1})) \leq \kappa. \quad (5)$$

Kárný (2014) and Özkan et al. (2013) seek to approximate  $p(\beta_t|\mathbf{x}_{1:t-1})$  by the distribution  $\hat{p}(\beta_t|\mathbf{x}_{1:t-1})$  having maximum entropy under the constraint in (5); for continuous distributions the maximum entropy can be formulated relative to an uninformative prior density  $p_u(\beta_t)$ , which corresponds to the Kullback-Leibler divergence between the two distributions. This approach ensures that we will not underestimate the uncertainty in the parameter distribution and the particular solution being sought takes the form

$$\hat{p}(\beta_t|\mathbf{x}_{1:t-1}, \rho_t) \propto p_\delta(\beta_t|\mathbf{x}_{1:t-1})^{\rho_t} p_u(\beta_t)^{(1-\rho_t)}, \quad (6)$$

where  $0 \leq \rho_t \leq 1$  is indirectly defined by (5) which in turn depends on the user defined parameter  $\kappa$ .

In our streaming data setting we follow *assumed density filtering* (Lauritzen, 1992) and the SVB approach (Broderick et al., 2013) and employ the approximation  $p(\beta_{t-1}|\mathbf{x}_{1:t-1}) \approx q(\beta_{t-1}|\lambda_{t-1})$ , where  $q(\beta_{t-1}|\lambda_{t-1})$  is the variational distribution calculated in the previous time step. Using this approximation in (3) and (4), we can express  $p_\delta$  in terms of  $\lambda_{t-1}$  in which case (6) becomes

$$\hat{p}(\beta_t|\lambda_{t-1}, \rho_t) \propto p_\delta(\beta_t|\lambda_{t-1})^{\rho_t} p_u(\beta_t)^{(1-\rho_t)}, \quad (7)$$

which we use as the prior density for time step  $t$ . Now, if  $p_u(\beta_t)$  belong to the same family as  $q(\beta_{t-1}|\lambda_{t-1})$ , then  $\hat{p}(\beta_t|\lambda_{t-1}, \rho_t)$  will stay within the same family and have natural parameters  $\rho_t\lambda_{t-1} + (1-\rho_t)\alpha_u$ , where  $\alpha_u$  are the natural parameters of  $p_u(\beta_t)$ . 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 (3).

So, 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,

$$\arg \max_{\lambda_t, \phi_t} \mathcal{L}(\lambda_t, \phi_t|\mathbf{x}_t, \rho_t\lambda_{t-1} + (1-\rho_t)\alpha_u).$$

As stated in the following lemma, this approach coincides with the so-called *power priors* approach (Ibrahim & Chen, 2000), a term that we will also adopt in the following.

**Lemma 1.** *The Bayesian updating scheme described by Figure 1 (b) and Equation 6, but with  $\rho_t$  fixed to a constant value, is equivalent to the recursive application of*

the Bayesian updating scheme of power priors (Ibrahim & Chen, 2000). This scheme is expressed as follows:

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

where  $\mathbf{x}_0$  and  $\mathbf{x}_1$  is the observation at time 0 (historical observation) and time 1 (current observation), respectively.

*Proof sketch.* 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  $p$  and  $p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0)$  is a Dirac delta function. The distribution  $p(\boldsymbol{\beta}_1|\mathbf{x}_0, \mathbf{x}_1, \rho)$  in this model is equivalent to  $p(\boldsymbol{\beta}|\mathbf{x}_1, \mathbf{x}_0, \rho)$ , which, in turn, is equivalent (up to proportionality) to  $p(\mathbf{x}_1|\boldsymbol{\beta}_1)\hat{p}(\boldsymbol{\beta}_1|\mathbf{x}_0, \rho_t)$ . Note that the last  $\hat{p}$  term can alternatively be expressed as  $\hat{p}(\boldsymbol{\beta}_1|\mathbf{x}_0, \rho_t) \propto p_\delta(\boldsymbol{\beta}_1|\mathbf{x}_0)^\rho p(\boldsymbol{\beta}_1)^{1-\rho} \propto p_\delta(\mathbf{x}_0|\boldsymbol{\beta}_1)^\rho p(\boldsymbol{\beta}_1)$ .  $\square$

The perspective provided by Lemma 1 introduces a well known result of power priors, which is also applicable in the current context (see the discussion after Theorem 1 in (Ibrahim et al., 2003)): “the power prior is an optimal prior to use and in fact minimizes the convex combination of KL 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.” As noted in (Olesen et al., 1992; Özkan et al., 2013), this schema works as a moving window with exponential forgetting of past data, where the effective number of samples or, more technically, the so-called *equivalent sample size* of the posterior (Heckerman et al., 1995), converges to,

$$\lim_{t \rightarrow \infty} ESS_t = \frac{|\mathbf{x}_t|}{1 - \rho} \quad (8)$$

if the size of the data batches is constant<sup>1</sup>.

For the experimental results reported in Section 5 we shall refer to the method outlined above as SVB with power priors (SVB-PP).

## 4.2. The Hierarchical Power Prior Model

In the approach taken by Özkan et al. (2013) (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.

As we shall see below, in order to support a variational updating scheme we need to restrict the prior distribution over  $\rho_t$ , effectively limiting the choice of prior distribution to

<sup>1</sup>For instance, the ESS of a Beta distribution is equal to the sum of the components of  $\boldsymbol{\lambda}_t$  and, in turn, equal to the number of data samples seen so far plus the prior’s pseudo-samples.

either an exponential distribution or a normal distribution with fixed variance, both of which should be truncated to the interval  $[0, 1]$ . Unless explicitly stated otherwise, we shall for now assume a truncated exponential distribution with natural parameter  $\gamma$  as prior distribution over  $\rho_t$ :

$$p(\rho_t|\gamma) = \frac{\gamma \exp(-\gamma \rho_t)}{1 - \exp(-\gamma)}. \quad (9)$$

The resulting model can be illustrated as in Figure 1 (b). We shall refer to models of this type as *hierarchical power prior* (HPP) models.

## 4.3. Variational Updating

For updating the model distributions we pursue a variational approach, where we seek to maximize the evidence lower bound  $\mathcal{L}$  in Equation (2) for time step  $t$ . However, since the model in Figure 1 (b) does not define a conjugate exponential distribution due to the introduction of  $p(\rho_t)$  we cannot maximize  $\mathcal{L}$  directly. Instead we will derive a (double) lower bound  $\hat{\mathcal{L}}$  ( $\hat{\mathcal{L}} \leq \mathcal{L}$ ) and use this lower bound as a proxy for the updating rules for the variational posteriors.

First of all, by instantiating the lower bound  $\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \omega_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1})$  in Equation (2) for the HPP model we obtain

$$\begin{aligned} \mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \omega_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1}) &= \mathbb{E}_q[\ln p(\mathbf{x}_t, \mathbf{Z}_t|\boldsymbol{\beta}_t)] \\ &+ \mathbb{E}_q[\ln \hat{p}(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1}, \rho_t)] \\ &+ \mathbb{E}_q[p(\rho_t|\gamma)] - \mathbb{E}_q[\ln q(\mathbf{Z}_t|\boldsymbol{\phi}_t)] \\ &- \mathbb{E}_q[q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)] - \mathbb{E}_q[q(\rho_t|\omega_t)], \end{aligned} \quad (10)$$

where  $\omega_t$  is the variational parameter for the variational distribution for  $\rho_t$ ; as we shall see later,  $\omega_t$  is a scalar and is therefore not shown in boldface. For ease of presentation we shall sometimes drop from  $\mathcal{L}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \omega_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1})$  the subscript as well as the explicit specification of the parameters when these is otherwise clear from the context.

We now define  $\hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \omega_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1})$  as

$$\begin{aligned} \hat{\mathcal{L}}_{HPP}(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t, \omega_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1}) &= \mathbb{E}_q[\ln p(\mathbf{x}_t, \mathbf{Z}_t|\boldsymbol{\beta}_t)] \\ &+ \mathbb{E}_q[\rho_t] \mathbb{E}_q[\ln p_\delta(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_{t-1})] + (1 - \mathbb{E}_q[\rho_t]) \mathbb{E}_q[\ln p_u(\boldsymbol{\beta}_t)] \\ &+ \mathbb{E}_q[p(\rho_t|\gamma)] - \mathbb{E}_q[\ln q(\mathbf{Z}_t|\boldsymbol{\phi}_t)] \\ &- \mathbb{E}_q[q(\boldsymbol{\beta}_t|\boldsymbol{\lambda}_t)] - \mathbb{E}_q[q(\rho_t|\omega_t)], \end{aligned} \quad (11)$$

which provide a lower bound for  $\mathcal{L}$ .

**Theorem 1.**  $\hat{\mathcal{L}}_{HPP}$  gives a lower bound for  $\mathcal{L}_{HPP}$ :

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

*Proof sketch.* The inequality derives by using Equation (12) and observing that  $a_g(\rho_t \boldsymbol{\lambda}_{t-1} + (1 - \rho_t) \boldsymbol{\alpha}_u) \leq \rho_t a_g(\boldsymbol{\lambda}_{t-1}) + (1 - \rho_t) a_g(\boldsymbol{\alpha}_u)$  because the log-normalizer



$a_g$  is always a convex function (Wainwright et al., 2008). Full details are given in the supplementary material.  $\square$

Rather than seeking to maximize  $\mathcal{L}$  we will instead maximize  $\hat{\mathcal{L}}$ . The gap between the two bounds is determined only by the log-normalizer of  $\hat{p}(\beta_t | \lambda_{t-1}, \rho_t)$ :

$$\hat{\mathcal{L}} - \mathcal{L} = \mathbb{E}_q[\rho_t a_g(\lambda_{t-1}) + (1 - \rho_t) a_g(\alpha_u)] + a_g(\rho_t \lambda_{t-1} + (1 - \rho_t) \alpha_u) \quad (12)$$

Thus, maximizing  $\hat{\mathcal{L}}$  wrt. the variational parameters  $\lambda_t$  and  $\phi$  also maximizes  $\mathcal{L}$ . By the same observation, we also have that the (natural) gradients are consistent relative to the two bounds:

**Corollary 1.**

$$\hat{\nabla}_{\lambda_t} \mathcal{L} = \hat{\nabla}_{\lambda_t} \hat{\mathcal{L}} \quad \hat{\nabla}_{\phi_t} \mathcal{L} = \hat{\nabla}_{\phi_t} \hat{\mathcal{L}}.$$

*Proof.* Follows immediately from Equation (12) because the difference does not depend of  $\lambda_t$  and  $\phi_t$ .  $\square$

Thus, updating the variational parameters  $\lambda_t$  and  $\phi_t$  in HPP models can be done as for regular conjugate exponential models of the form in Figure 1.

In order to update  $\omega_t$  we rely on  $\hat{\mathcal{L}}$ , which we can maximize using the natural gradient wrt.  $\omega_t$  (Sato, 2001) and which can be calculated in closed form for a restricted distribution family for  $\rho_t$ .

**Lemma 2.** Assuming that the sufficient statistics function for  $\rho_t$  is the identity function,  $\mathbf{t}(\rho_t) = \rho_t$ , then we have

$$\hat{\nabla}_{\omega_t} \hat{\mathcal{L}} = \text{KL}(q(\beta_t | \lambda_t), p_u(\beta_t)) - \text{KL}(q(\beta_t | \lambda_t), p_\delta(\beta_t | \lambda_{t-1})) + \gamma - \omega_t \quad (13)$$

*Proof sketch.* Based on a straightforward algebraic derivation of the gradient using standard properties of the exponential family. Full details are given in the supplementary material.  $\square$

Note that the truncated exponential distribution (see Equation (9)) satisfies the restriction expressed in Lemma 2, and also note that the variational posterior  $q(\rho_t | \omega_t)$  will be a truncated exponential density too.

On the other hand, observe that the form of the natural gradient of  $\omega_t$  have an intuitive semantic interpretation, which also extends to the coordinate ascent variational message passing framework (Winn & Bishop, 2005) as shown by Masegosa et al. (2016a). Specifically, using the constant  $\gamma$  as a threshold, we see that if the uninformed prior  $p_u(\beta_t)$  provides a better fit to the variational posterior at time  $t$  than the variational parameters  $\lambda_t$  from the previous time step ( $\text{KL}(q(\beta_t | \lambda_t), p_u(\beta_t)) +$

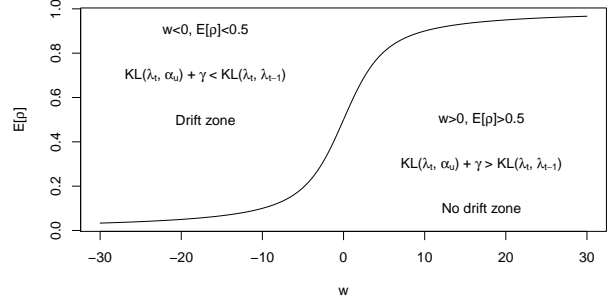


Figure 2. Relationship between  $\omega_t$  and  $E_q[\rho_t]$ .

$\gamma < \text{KL}(q(\beta_t | \lambda_t), p_\delta(\beta_t | \lambda_{t-1}))$ ), then we will get a negative value for  $\omega_t$  when performing coordinate ascent using Equation (13). This in turn implies that  $E_q[\rho] < 0.5$  because  $E_q[\rho] = 1/(1 - e^{-\omega_t}) - 1/\omega_t$  (plotted in Figure 2), which means that we have a higher degree of forgetting for past data. If  $\omega_t > 0$  then  $E_q[\rho] > 0.5$  and less past data is forgotten. Figure 2 graphically illustrates this trade-off.

#### 4.4. The Multiple Hierarchical Power Prior Model

The HPP model can immediately be extended to include multiple power priors  $\rho_t^{(i)}$ , one for each global parameter  $\beta_i$ . In this model the  $\rho_t^{(i)}$ 's are pair-wise independent. The latter ensures that optimizing the  $\hat{\mathcal{L}}$  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 expressivity of the model. We shall refer to this extended model as a *multiple hierarchical power prior* (MHPP) model.

## 5. Experiments

### 5.1. Experimental Set-up

In this section we will evaluate the following methods:

- Streaming variational Bayes (SVB).
- Four versions of Population Variational Bayes (PVB)<sup>2</sup>: Population-size  $M$  equal to the average size of each data-batch, or  $M$  equal to a fixed value ( $M = 1000$  in Section 5.2 and  $M = 10000$  in Section 5.3). Learning-rate  $\nu = 0.1$  or  $\nu = 0.01$ .
- Two versions of SVB-PP:  $\rho = 0.9$  or  $\rho = 0.99$ .
- Two versions of SVB-HPP: A single shared  $\rho$  (denoted SVB-HPP) or separate  $\rho^{(i)}$  parameters (SVB-MHPP).

The underlying variational engine is the VMP algorithm (Winn & Bishop, 2005) for all models; VMP was termi-

<sup>2</sup>We do not compare with SVI, because SVI is a special case of PVB when  $M$  is equal to the total size of the stream.

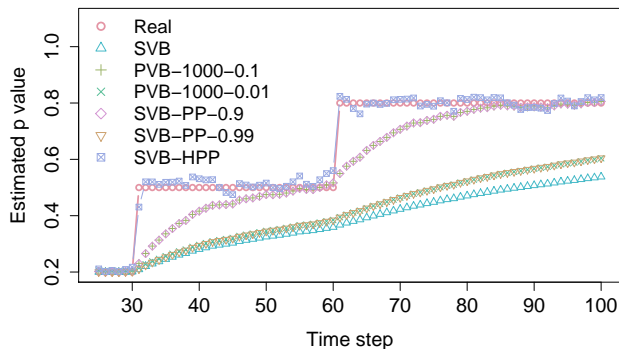


Figure 3.  $E[\beta_t]$  in the Beta-Binomial model artificial data set

nated after 100 iterations or if the relative increase in the lower bound fell below 0.01%. All priors were uninformative, using either flat Gaussians, flat Gamma priors or uniform Dirichlet priors. We set  $\gamma = 0.1$  for the HPP priors. Variational parameters were randomly initialized using the same seed for all methods.

## 5.2. 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.,  $|\mathbf{x}_t| = 100$ ) from a Binomial distribution at each time step. We artificially introduce 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 modelled using a Beta-Binomial model.

**Parameter Estimation:** Figure 3 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 depend heavily on the learning rate  $\nu$ , where the higher learning rate, which results in the more aggressive forgetting, works better in this example. Recall, though, that  $\nu$  needs to be hand-tuned to achieve an optimal performance. As expected, the choice of  $M$  does not have an impact, because the present model has no local hidden variables (cf. Section 3). SVB-PP produces 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.

**Equivalent Sample Size (ESS):** Figure 4 (left) gives the evolution of the equivalent sample size,  $ESS_t$ , for the different methods<sup>3</sup>. The ESS of PVB is always given by the constant  $M$ . For SVB, the ESS monotonically increases as more data is seen, while SVB-PP exhibits convergence to the limiting value computed in Equation (8). A different behaviour is observed for SVB-HPP: It is automatically ad-

<sup>3</sup>For this model, ESS is simply computed by summing up the components of the  $\lambda_t$  defining the Beta posterior.

justed. Notice that the values for this model is to be read off the alternative  $y$ -axis. We can detect the the concept drift, by identifying where the ESS rapidly declines.

**Evolution of Expected Forgetting factor:** In Figure 4 (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.

## 5.3. Evaluation using Real Data Sets

### 5.3.1. DATA AND MODELS

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

**Electricity Market (Harries, 1999):** The data set describes the electricity market of two Australian states. It contains 45312 instances of 6 attributes, including a class label comparing the change of the electricity price related to a moving average of the last 24 hours. Each instance in the data set represents 30 minutes of trading; during our analysis we created batches such that  $\mathbf{x}_t$  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 (Hoffman et al., 2013), hence PVB cannot be applied here, because lower-bound’s gradient cannot be computed in closed-form.

**GPS (Zheng et al., 2008; 2009; 2010):** 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  $\mathbf{x}_t$  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 also to monitor how the changes are affected by the day of the week.

**Finance (reference withheld):** The data contains monthly aggregated information about the financial profile of

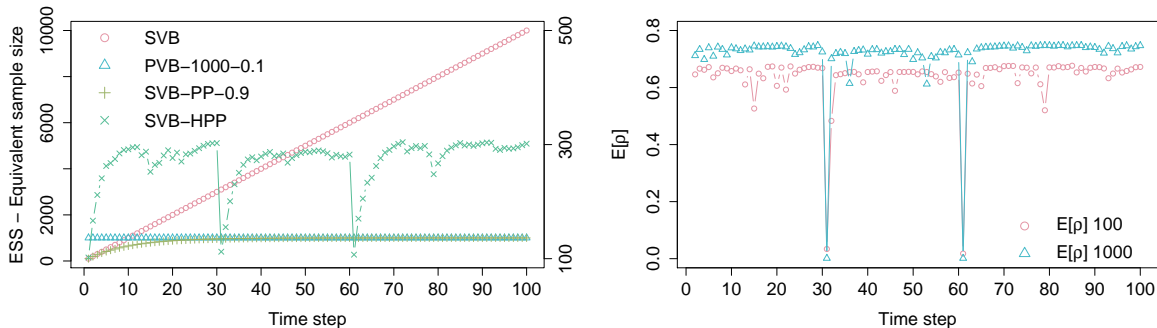


Figure 4.  $ESS_t$  (left) and  $\mathbb{E}_q[\rho_t]$  (right) in the Beta-Binomial model artificial data set

around 50 000 customers over 62 (non-consecutive) months. Three attributes were extracted per customer, in addition to a class-label telling whether or not 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 5-component mixture of Gaussians distribution. The distribution over the mixture node is shared by all the attributes, but not between the two classes of customers.

A detailed description of all the models, including their structure and their variational families, is given at the supplementary material.

### 5.3.2. EVALUATION AND DISCUSSION

To evaluate the different methods discussed, we look at the test marginal log-likelihood (TMLL). Specifically, each data batch is randomly split in a train data set,  $\mathbf{x}_t$ , and a test data set,  $\tilde{\mathbf{x}}_t$ , containing two thirds and one third of the data batch, respectively. Then,  $TMLL_t$  is computed as  $TMLL_t = \frac{1}{|\tilde{\mathbf{x}}_t|} \int p(\tilde{\mathbf{x}}_t, \mathbf{z}_t | \beta_t) p(\beta_t | \mathbf{x}_t) d\mathbf{z}_t d\beta_t$ . Figure 5 (left) shows for each method the difference between its  $TMLL_t$  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. The right-hand side of Figure 5 shows the development of  $\mathbb{E}_q[\rho_t]$  over time for SVB-HPP and SVB-MHPP. For SVB-HPP 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.<sup>4</sup> 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. Finally, we compute each method’s aggregated test marginal log-likelihood measure  $\sum_{t=1}^T TMLL_t$ , and report these values in Table 1.

For the electricity data set, we can see that the two proposed methods (SVB-HPP and SVB-MHPP) perform best. All models are comparable during the first nine months, which is a period where our models detect no or very limited con-

cept drift (cf. top right plot or Figure 5). However, after this period, both SVB-HPP and SVB-MHPP detects substantial drift, and is able to adapt better than the other methods, which appear unable to adjust to the complex concept drift structure in the latter part of the data. SVB-HPP and SVB-MHPP continue to behave at a similar level, mainly because when drift happens it typically includes a high proportion of the parameters of the model.

For the GPS data set, we can observe how the SVB-MHPP is superior to the rest of the methods, particularly towards the end of the series. When looking at Figure 5 (middle right panel), we can see that a significant proportion of the model parameters are drifting (i.e.,  $\mathbb{E}_q[\rho_t^{(i)}] \leq 0.05$ ) at all times, while another proportion of the parameters show a quite stable behavior ( $\rho$ -values above 0.9). This complex pattern is not captured well by SVB-HPP, which ends up assuming no concept drift after the initial time-step.

The financial data set shows a different behavior. During the first months, SVB-MHPP slightly outperforms the rest of the approaches, but after month 30, SVB-PP with  $\rho = 0.9$  is superior, with SVB-MHPP second. Looking at the  $E[\rho_t^{(i)}]$ -values of SVB-MHPP, we observe that there is significant 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 the drift is constant and fairly limited, at least when the optimal forgetting factor  $\rho$  has been identified.

We conclude this section by highlighting that the performance of SVB-PP and PVB depend heavily on the hyper-parameters of the model, cf. Table 1. As an example, consider SVB-PP for the financial data set. While it was the best overall with  $\rho = 0.9$ , it is inferior to SVB-MHPP if  $\rho = 0.99$ . Similarly, PVB’s performance is sensitive both to  $\nu$  (see in particular the results for the GPS data) and  $M$  (financial data). These hyper-parameters are hard to fix, as their optimal values depend on data characteristics (see Broderick et al. (2013); McNerney et al. (2015) for similar conclusions). We therefore believe that the fully Bayesian formulation is an important strong point of our approach.

<sup>4</sup>The numbers of variational parameters are 14, 78 and 33 for the Electricity, GPS and Financial model, respectively.

DATA SET	SVB	PVB $M = 10k$ $\nu = 0.1$	PVB $M = 10k$ $\nu = 0.01$	PVB $M =  \mathbf{x}_t $ $\nu = 0.1$	PVB $M =  \mathbf{x}_t $ $\nu = 0.01$	SVB-PP $\rho = 0.9$	SVB-PP $\rho = 0.99$	SVB-HPP	SVB-MHPP
ELECTRICITY	-44.91					-43.92	-44.80	-40.06	<b>-40.03</b>
GPS	-1.93	-2.03	-2.72	-1.88	-2.42	-1.89	-1.92	-1.86	<b>-1.74</b>
FINANCE	-19.84	-22.29	-22.57	-21.81	-22.07	<b>-19.05</b>	-19.78	-19.83	-19.40

Table 1. Aggregated Test Marginal Log-Likelihood.

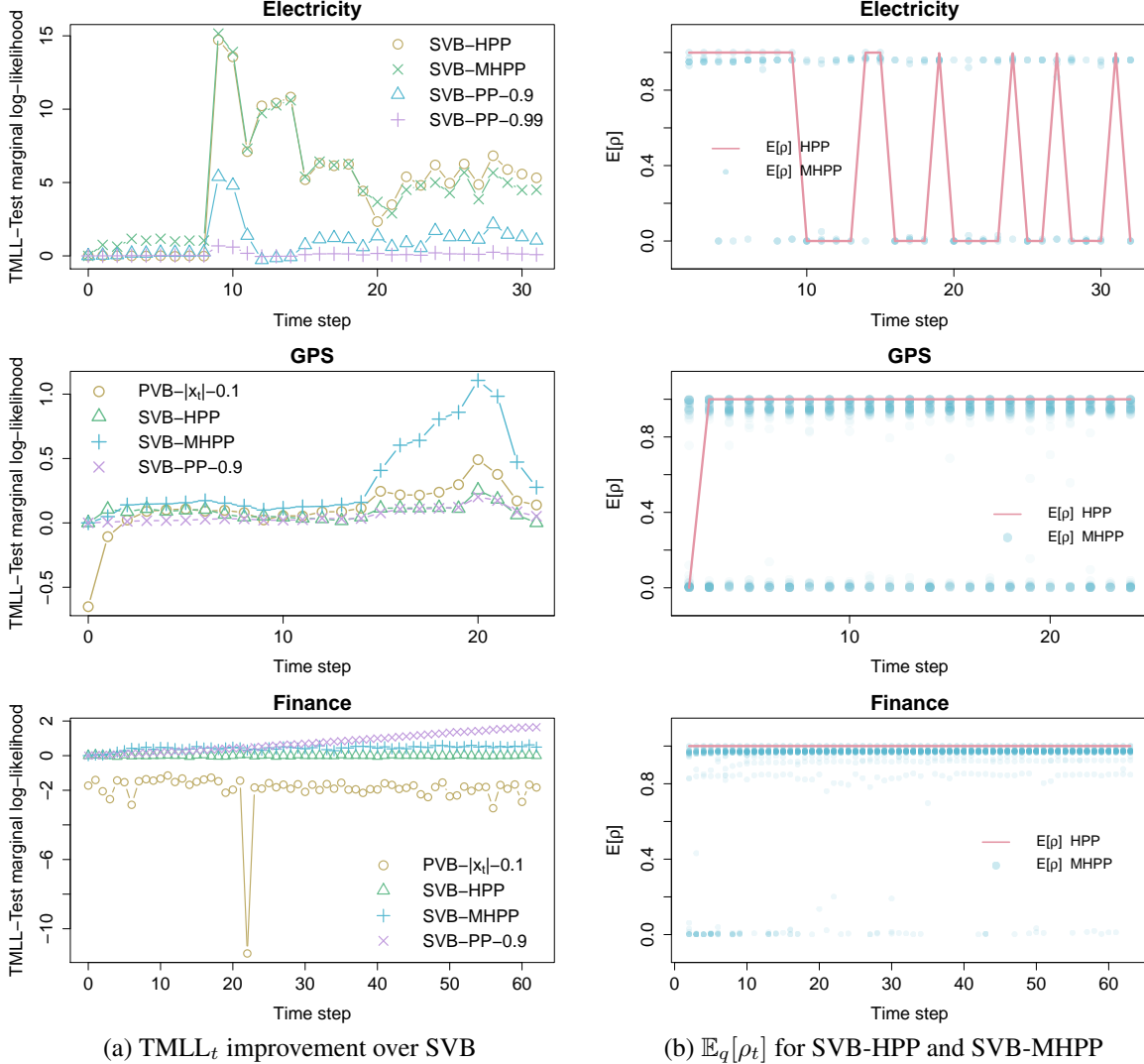


Figure 5. Results of the competing methods for the three real-life data sets. See text for discussion.

## 6. Conclusions and Future Work

We have introduced a new class of Bayesian models for streaming data, able to capture changes in the underlying generative process. Unlike existing solutions to this problem, aimed at modeling slowly changing processes, our proposal is able to handle both abrupt and gradual concept drift following a Bayesian approach. The new model accounts for the dynamics of the data stream by assuming that only the global parameters evolve over time. We intro-

duce the so-called hierarchical power priors, where a prior on the learning rate is given allowing it to adapt to the data stream. We have addressed the complexity of the underlying inference tasks by developing an approximate 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.



## Acknowledgements

This work was partly carried out as part of the AMIDST project. AMIDST has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209. Furthermore, this research has been partly funded by the Spanish Ministry of Economy and Competitiveness, through projects TIN2015-74368-JIN, TIN2013-46638-C3-1-P, TIN2016-77902-C3-3-P and by ERDF funds.

## References

- Ahmed, Amr, Ho, Qirong, Teo, Choon Hui, Eisenstein, Jacob, Smola, Alexander J, and Xing, Eric P. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *AISTATS*, pp. 101–109, 2011.
- Broderick, Tamara, Boy, Nicholas, Wibisono, Andre, Wilson, Ashia C., and Jordan, Michael I. Streaming variational Bayes. In *Advances in Neural Information Processing Systems 26*, pp. 1727–1735. Curran Associates, Inc., 2013.
- Cabañas, Rafael, Martínez, Ana M, Masegosa, Andrés R, Ramos-López, Darío, Samerón, Antonio, Nielsen, Thomas D, Langseth, Helge, and Madsen, Anders L. Financial data analysis with PGMs using AMIDST. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pp. 1284–1287. IEEE, 2016.
- Doucet, Arnaud, Godsill, Simon, and Andrieu, Christophe. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.
- Gama, João, Žliobaitė, Indrė, Bifet, Albert, Pechenizkiy, Mykola, and Bouchachia, Abdelhamid. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.
- Ghahramani, Zoubin and Attias, H. Online variational Bayesian learning. In *Slides from talk presented at NIPS workshop on Online Learning*, 2000.
- Harries, Michael. Splice-2 comparative evaluation: Electricity pricing. NSW-CSE-TR-9905, School of Computer Science and Engineering, The University of New South Wales, 1999.
- Heckerman, David, Geiger, Dan, and Chickering, David M. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3): 197–243, 1995.
- Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.
- Honkela, Antti and Valpola, Harri. On-line variational Bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 803–808, 2003.
- Ibrahim, Joseph G and Chen, Ming-Hui. Power prior distributions for regression models. *Statistical Science*, pp. 46–60, 2000.
- Ibrahim, Joseph G, Chen, Ming-Hui, and Sinha, Debajyoti. On optimality properties of the power prior. *Journal of the American Statistical Association*, 98(461):204–213, 2003.
- Kárný, Miroslav. Approximate Bayesian recursive estimation. *Information Sciences*, 285:100–111, 2014.
- Lauritzen, Steffen L. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.
- Masegosa, A. R., Martínez, A. M., Langseth, H., Nielsen, T. D., Salmerón, A., Ramos-López, D., and Madsen, A. L. d-VMP: Distributed variational message passing. In *PGM'2016. JMLR: Workshop and Conference Proceedings*, volume 52, pp. 321–332, 2016a.
- Masegosa, Andres R, Martinez, Ana M, and Borchani, Hanen. Probabilistic graphical models on multi-core cpus using Java 8. *IEEE Computational Intelligence Magazine*, 11(2):41–54, 2016b.
- Masegosa, Andrés R, Martínez, Ana M, Ramos-López, Darío, Cabañas, Rafael, Salmerón, Antonio, Nielsen, Thomas D, Langseth, Helge, and Madsen, Anders L. Amidst: a Java toolbox for scalable probabilistic machine learning. *arXiv preprint arXiv:1704.01427*, 2017.
- McInerney, James, Ranganath, Rajesh, and Blei, David. The population posterior and Bayesian modeling on streams. In *Advances in Neural Information Processing Systems 28*, pp. 1153–1161. Curran Associates, Inc., 2015.
- Olesen, Kristian G, Lauritzen, Steffen L, and Jensen, Finn V. ahugin: A system creating adaptive causal probabilistic networks. In *Proceedings of the Eighth international conference on Uncertainty in artificial intelligence*, pp. 223–229. Morgan Kaufmann Publishers Inc., 1992.

- Özkan, Emre, Šmđl, Václav, Saha, Saikat, Lundquist, Christian, and Gustafsson, Fredrik. Marginalized adaptive particle filtering for nonlinear models with unknown time-varying noise parameters. *Automatica*, 49(6):1566–1575, 2013. ISSN 0005-1098.
- Sato, Masa-Aki. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.
- Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305, 2008.
- Winn, John M. and Bishop, Christopher M. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.
- Yao, Limin, Mimno, David, and McCallum, Andrew. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 937–946. ACM, 2009.
- Zheng, Yu, Li, Quannan, Chen, Yukun, Xie, Xing, and Ma, Wei-Ying. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing, UbiComp '08*, pp. 312–321, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-136-1. doi: 10.1145/1409635.1409677.
- Zheng, Yu, Zhang, Lizhu, Xie, Xing, and Ma, Wei-Ying. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pp. 791–800, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526816.
- Zheng, Yu, Xie, Xing, and Ma, Wei-Ying. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.