A New Class of Time Dependent Latent Factor Models with Applications

Sinead A. Williamson
Departments of Statistics and Data Science/Information, Risk and Operations Management, University of Texas at Austin, Austin, TX 78757, sinead.williamson@mccombs.utexas.edu

Michael Zhang
Department of Statistics and Data Science, University of Texas at Austin, Austin, TX 78757, michael_zhang@utexas.edu

Paul Damien
Department of Information, Risk and Operations Management, University of Texas at Austin, Austin, TX 78757, paul.damien@mccombs.utexas.edu

In many applications, observed data are influenced by some combination of latent causes. For example, suppose sensors are placed inside a building to record responses such as temperature, humidity, power consumption and noise levels. These random, observed responses are typically affected by many unobserved, latent factors (or features) within the building such as the number of individuals, the turning on and off of electrical devices, power surges, etc. These latent factors are usually present for a contiguous period of time before disappearing; further, multiple factors could be present at a time.

This paper develops new probabilistic methodology and stochastic simulation methods for random object generation influenced by latent features. Every datum is associated with subsets of a potentially infinite number of hidden, persistent features that account for temporal dynamics in an observation. The ensuing class of dynamic models constructed by adapting the Indian Buffet Process — a probability measure on the space of unbounded binary matrices — finds use in a variety of applications arising in operations, signal processing, biomedicine, marketing, image analysis, etc. Illustrations using synthetic and real data are provided.

Key words: Bayesian nonparametrics, latent feature model, simulation

---

1. Introduction

Random object generation is a broad topic, since the word "object" has many connotations in mathematics and applied probability. For example, "object" could refer to a matrix or a polynomial.

Indeed, observed data are random objects; for instance, a vector of observables in a regression context satisfies transparently the idea of a probabilistic object (Leemis 2006).

Of late, a class of random object models is growing in popularity, namely Latent Factor (or Feature) Models, abbreviated LFM. The theory and use of these models lie at the intersection of probability theory, Bayesian inference, and simulation methods, specifically Markov chain Monte Carlo (MCMC). Saving the formal description of LFMs for future sections, consider the following heuristics of certain key ideas central to the paper.

Latent variables are unobserved, or are not directly measurable. Parenting skill, speech impediments, socio-economic status, and quality of life are some examples of these. Latent variables could also correspond to a "true" variable observed with error. Examples would include iron intake measured by a food frequency, self-reported weight, and lung capacity measured by forced expiratory volume in one second. In Bayesian hierarchical modeling, latent variables are often used to represent unobserved properties or hidden causes of data that are being modeled (Bishop 1998). Typically, these variables have a natural interpretation in terms of certain underlying but unobserved features of the data; as examples, thematic topics in a document or motifs in an image. The simplest such models, which we will refer to simply as Latent Variable Models (LVMs), use a finite number of latent variables, with each datum related to a single latent variable (Bishop 1998, McLauchlan 2000). This class of models includes finite mixture models, where a datum is associated with a single latent mixture component, and Hidden Markov Models (HMMs), where each point in a time series is associated with a single latent state (Baum and Petrie 1996). All data associated with a given latent parameter are assumed to be independently and identically simulated according to a distribution parametrized by that latent parameter.

Greater flexibility could be obtained by allowing multiple latent features for each datum. This allows different aspects of a datum to be shared with different subsets of the dataset. For example, two articles may share the theme "science", but the second article may also exhibit the theme "finance". Similarly, a picture of a dog in front of a tree has aspects in common with both tree

pictures and dog pictures. Models that allow multiple features are typically referred to as LFMs. Examples of LFMs include Bayesian Principle Component Analysis where data are represented using a weighted superposition of latent factors, and Latent Dirichlet Allocation where data are represented using a mixture of latent factors; see Roweis and Ghahramani (1999) for a review of both LVMs and LFMs.

In the majority of LVMs and LFMs, the number of latent variables is finite and pre-specified. The appropriate cardinality is often hard to determine *a priori* and, in many cases, we do not expect our training set to contain exemplars of all possible latent variables. These difficulties have led to the increasing popularity of LVMs and LFMs where the number of latent variables associated with each datum or object is potentially unbounded. In the LVM setting, this yields mixture models and latent state models where the number of parameters are unbounded, such as Dirichlet Process Mixture Models and infinite HMMs (Antoniak 1974, Teh et al. 2006). In the LFM setting, this leads to models where both the total number of latent features used to represent a dataset, and the number of latent features underlying a given datum, are unbounded; examples include models based on the Indian Buffet Process, the infinite gamma-Poisson process, and the beta-negative binomial process (Griffiths and Ghahramani 2005, Titsias 2007, Broderick et al. 2015).

These latter probabilistic models with an infinite number of parameters are referred to as nonparametric latent variable models (npLVMs) and nonparametric latent factor models (npLFMs)—an unfortunate misnomer to be sure. These models generally tend to provide richer inferences than their finite-dimensional counterparts, since deeper relationships between the unobserved variables and the observed data could be obtained by relaxing finite distributional assumptions about the probability generating mechanism.

In many applications, data are assumed *exchangeable*, in that no information is conveyed by the order in which data are observed. Even though exchangeability is a weaker (hence preferable) assumption than independent and identically distributed data, often times, observed data are time-stamped emissions from some evolving process; that is, the ordering (or dependency) is crucial to

understanding the entire random data-generating mechanism. There are two types of dependent data that, typically, arise in practice. It is convenient to use terminology from the medical literature to distinguish the two. *Longitudinal dependency* refers to situations where one records multiple entries from the same random process over a period of time. In AIDS research, a biomarker such as a CD4 lymphocyte cell count is observed intermittently for a patient and its relation to time of death is of interest. In a different context, the ordering of frames in a video sequence or the ordering of windowed audio spectra in a piece of music within a time interval are crucial to our understanding of the entire video or musical piece. *Epidemiological dependency* corresponds to situations where, at a fixed covariate value, more than one random process is involved in the data generating mechanism; that is, single records from multiple entities constitute the observed data. For instance, in an annual survey on diabetic indicators one might interview different people each year; the observations correspond to different random processes (i.e. different individuals), but still capture global trends. Or consider the following: at any fixed moment in time, the location of each weather sensor is likely to capture many of the relationships in the different outputs emanating from a network of weather sensors. Alternatively, this could be a network of sensors recording air quality or seismic activity.

Both LFMs and LVMs have been developed for non-dependent as well as the two types of dependent data. Here we provide a brief summary of some of these application areas.

Choice Models: These are ubiquitous in decision theory, marketing, management science, psychology, machine learning, etc. As one example, purchase decisions are dictated by attributes or product features. Thus, if one were interested in buying a laptop computer from an array of computer product offerings, attributes such as price, speed, battery life, portability, etc. would play a significant role in the final decision. As another example, using an HMM, Shi, Wedel, and Pieters (2013) shed critical insight on the impact of eye-tracking movements on information acquisition patterns with the aim of assessing how consumers gather product and attribute information from moment to moment. Among the noteworthy findings, they demonstrate that horizontal and contiguous eye movements are significant in information acquisition, leading to useful operational and managerial implications for Web design, online retailing and online choice.

Causal Graphical Models: The aim of probabilistic graphical models is to learn about an unknown number of hidden or latent factors (e.g. diseases) that are causes for a set of observed variables (e.g. symptoms). Wood, Griffiths, and Ghahramani (2006) develop an infinite-dimensional, stochastic simulation, Bayesian LFM that allows the practitioner to gain insights on the causal structure underlying the onset of strokes. Ruiz et al. (2014) use a related infinite-dimensional LFM for inferring hidden causes implicit in multiple-choice survey questions that are used in discovering important co-morbidity patterns in psychiatric disorders.

Dyadic Objects: Several applications involve two sets of objects for which data are observed in pairs. As examples, the sets could be movies and viewers where the response of interest is the ratings given by the latter group regarding the former. In Biology, the sets might be genes and biological tissues with expression levels for specific genes in different tissues acting as the observed data. Predictive inference is of interest in these types of settings where, for instance, in the first example the ratings of a new viewer for a movie, conditioned on other viewers ratings, could be of value. This general notion of borrowing strength to improve predictions using npLFMs is tackled by Meeds et al. (2007), who use a hybrid of simulation methods including Gibbs sampling, Metropolis, and split-merge techniques. For a film dataset, they represent movies and viewers via latent vectors with an unbounded number of elements that correspond to viewer appreciation of movie features such as "like comedies", "dislike lengthy movies", and so forth. Using their Bayesian npLFM, the authors obtain predictions of viewer ratings.

Image Analysis: Here, the primary aim is to extract features in both still and moving images. The following citations cover a wide spectrum of applications including medicine, engineering, operations, business, machine learning, etc. (Stark 1987, Hanson 1993, Winkler 2003, Acton and Ray 2006, Kopperapu and Desai 2006, Wang et al. 2013, Tavares and Jorge 2011).

A number of papers have developed npLFMs for epidemiological dependence (Foti et al. 2013, Ren et al. 2011, Zhou et al. 2011, Rao and Teh 2009); in these settings we are often able to make use of conjugacy to develop reasonably efficient stochastic simulation schemes. In addition, several

nonparametric priors for LFMs have been proposed for longitudinally dependent data (Williamson, Orbanz, and Ghahramani 2010, Gershman, Frazier, and Blei 2015) — but unfortunately these latter papers, by virtue of their modeling approaches, lead to computationally complex inference protocols. Furthermore, these methods are often invariant under time reversal, rendering them unsuitable for temporal dynamics.

In this paper, we introduce a new class of npLFMs that is suitable for time (or longitudinally) dependent data. From a modeling perspective, the focus is on npLFMs rather than npLVMs since the separability assumptions underlying LVMs is overly restrictive for most real data. Specifically, we follow the tradition of generative or simulation-based npLFMs. A Bayesian approach is natural in this framework since the form of npLFMs needed to better model temporal dependency involves the use of probability distributions on function spaces; the latter idea is commonly referred to as Bayesian nonparametric inference (Walker et al. 1999).

The primary aims of this research are the following. First, to develop a class of npLFMs with practically useful attributes to generate random objects in a variety of applications. These attributes include an unbounded number of latent factors; capturing temporal dynamics in the data; and the tracking of *persistent* factors over time. The significance of this class of models is best described with a simple, yet meaningful, example. Consider a flautist playing a musical piece. At very short time intervals if the flautist is playing a B♭ at time t, it is likely that note would still be playing at time $t+1$. Arguably, this is a continuation of a single note instance that begins at time t and persists to time $t+1$ (or beyond). Unlike current approaches, our proposed time-dynamic model captures this (persistent latent factor) dependency in the musical notes from time t to $t+1$. The second goal of this research is to develop a general Markov chain Monte Carlo algorithm to enable full Bayesian implementation of the new npLFM family. Finally, applications of time-dependent npLFMs are shown via simulated and real data analysis.

In Section 2, finite and nonparametric LFMs are described. Section 3 discusses the Indian Buffet Processes that form the kernel for the new class of npLFMs introduced in Section 4. Section 5

details the stochastic simulation method used to implement the models in Section 4, followed by synthetic and real data illustrations in Section 6. A brief discussion in Section 7 concludes the paper.

2. Latent Variable Models (LVMs) and Latent Factor Models (LFMs)

An LVM posits that the variation within a dataset of size N could be described using some smaller set of $K < N$ latent variables. As an example, consider a mixture of K Gaussian distributions where each datum belongs to one of the K mixture components parametrized by different means and variances. These parameters, along with the cluster allocations, comprise the latent variables.

Mixture models can be represented through a hierarchical representation (Lindley and Smith 1972). This layered structure of the observed data and the unobserved latent quantities reflects the true underlying generative process; i.e., the assumed probabilistic mechanism by which the latent variables and the observations arise. For the above mixture of K Gaussians, a well-known generative process is given by:

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(\gamma) \\
\mu_k &\sim \text{Normal}(\mu_0, \sigma_0^2), \\
\sigma_k^2 &\sim \text{Inverse-Gamma}(\alpha, \beta), \quad k = 1, \ldots, K \\
z_n &\sim \text{Multinomial}(\pi), \\
x_n &\sim \text{Normal}(\mu_{z_n}, \sigma_{z_n}^2), \qquad n = 1, \ldots, N
\end{aligned}$$

The lowest level of the hierarchy comprises the observed data, conditioned on the mean and variance parameters that define a Normal distribution for each datum, n. Each upward step in the hierarchy correspond to probability models for the latent parameters. Noting that Bayes theorem is implicit in the above representation (Lindley and Smith 1972), inference about the latent parameters, via their posterior distributions, can be obtained using stochastic simulation methods such as Markov chain Monte Carlo. The above mixture of K Gaussians, where the observations are conditionally independent, is wildly popular; hundreds of papers on the topic may be found.

The key point to note in this model is that the appearance of all the data points associated with the kth cluster are controlled entirely by μ_k and σ_k^2. This is viable when observations are entirely separable.

However, there are many practical applications where the observed data exhibit multiple underlying features. For example, in image modeling we may have two pictures, one of a dog beside a tree, and one of a dog beside a bicycle. If we assign both images to a single cluster, we ignore the difference between tree and bicycle. If we assign them to different clusters, we ignore the commonality of the dogs. In these situations, LVMs should allow each datum to be associated with multiple latent variables. This extension of LVMs is typically referred to as Latent Feature Models or Latent Factor Models (LFMs). For clarity, throughout this paper, LVMs refer exclusively to models where each datum is associated with a single latent parameter, and LFMs refer to models where each datum is associated with multiple latent parameters.

A classic example of an LFM is Factor Analysis (Cattell 1952), wherein one assumes K D-dimensional latent features (or factors f_k) which are typically represented as a $K \times D$ matrix F. Each datum, x_n, is associated with a vector of weights, λ_n, known as the factor loading, which determines the degree to which the datum exhibits each factor. Letting X be the $N \times D$ data matrix and Λ be the $N \times K$ factor loading matrix, we can write $X = \Lambda F + \mathbf{e}$, where $\mathbf{e}$ is a matrix of random noise terms. Factor Analysis can be placed in a Bayesian framework by placing appropriate priors on the factors, loadings and noise terms (Press and Shigemasu 1989). Such analysis is used in many contexts; as examples: micro array data (Hochreiter, Clevert, and Obermayer 2006), dietary patterns (Venkaiah, Brahmam, and Vijayaraghavan 2011), and psychological test responses (Tait 1986). Independent Component Analysis (ICA; see Hyvärinen, Karhunen, and Oja (2001)) is a related model with independent non-Gaussian factors; ICA is commonly used in blind source separation of audio data.

A serious disadvantage of LFMs such as Factor Analysis is that they assume a fixed, finite number of latent factors. In many settings, such an assumption is hard to justify. Even with a fixed, finite

dataset, picking an appropriate number of factors, *a priori*, requires expensive cross-validation. In an online setting, where the dataset is constantly growing, it may be unreasonable to consider any finite upper bound. As illustrations, the number of topics that may appear in a newspaper, or the number of image features that may appear in an online image database, could grow unboundedly over time. One way of obviating this difficulty is to allow an infinite number of latent features *a priori*, and by ensuring every datum exhibits only a finite number of features whereby popular features tend to get reused. Such a construction would allow the number of exhibited features to grow in an unbounded manner as sample size grows, while still borrowing (statistical) strength from repeated features. The new family of LFMs developed in Section 4 of this paper will allow for the possibility of an infinite number of latent features.

3. The Indian Buffet Process (IBP)

From the last section, if an infinite LFM representation is needed to model the observed data via a generative process similar to the finite-dimensional Gaussian mixture model detailed earlier, then it is necessary to use an appropriate family of probability distributions for such a task. The transition from finite to infinite dimensional latent factors implies that the probability distributions on these factors in the generative process would now be elements in some function space; i.e., we enter the realm of Bayesian nonparametric inference. There is a vast literature on Bayesian nonparametric models; the classic references are Ferguson (1973) and Lo (1984).

Recently, a new class of nonparametric distributions of particular relevance to LFMs was developed by Griffiths and Ghahramani (2005) who labeled their stochastic process prior as the Indian Buffet Process (IBP). This prior is the Machine Learning attempt to adopt Bayesian nonparametric inference into the generative process of an LFM where the goal of unsupervised learning is to discover the latent variables responsible for generating the observed properties of a set of objects.

The IBP provides a mechanism for selecting sets of features. This mechanism can be broken down into two components: a global random sequence of feature probabilities that assigns probabilities to infinitely many features, and a local random process that selects a finite subset of these features for each datum.

The global sequence of feature probabilities is distributed according to a stochastic process known as the beta process (Hjort 1990, Thibaux and Jordan 2007). Loosely speaking, the beta process is a random measure, $B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k}$, that assigns finite mass to a countably infinite number of locations; these atomic masses μ_k are independent, and are distributed according to the infinitesimal limit of a beta distribution. The locations, θ_k, of the atoms parametrize an infinite sequence of latent features.

The subset selection mechanism is a stochastic process known as the Bernoulli process (Thibaux and Jordan 2007). This process samples a random measure $\zeta_n = \sum_{k=1}^{\infty} z_{n,k} \delta_{\theta_k}$, where each $z_{n,k} \in \{0, 1\}$ indicates the presence or absence of the kth feature θ_k, and are sampled independently as $z_{n,k} \sim \text{Bernoulli}(\mu_k)$. We can use these random measures ζ_n to construct a binary feature allocation matrix Z by ordering the features according to their popularity and aligning the corresponding ordered vector of indicators. This matrix will have a finite but unbounded number of columns with at least one non-zero entry; the re-ordering allows us to store the non-zero portion of the matrix in memory. It is often convenient to work directly with this *random*, binary matrix, and doing so offers certain insights into the properties of the IBP. This representation depicts the IBP as a (stochastic process) prior probability distribution over equivalence classes of binary matrices with a specified number of rows and potentially infinite columns. Viewed from this perspective, under this prior, binary matrices could grow or shrink with more data, effectively letting models adapt to the complexity of the observations.

Consider a mathematical representation of the above discussion. Let Z denote a random, binary matrix. Then, following Griffiths and Ghahramani (2005), the IBP prior distribution for Z is given by

$$p(Z) = \frac{\alpha^K}{\prod_{h=1}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!} \tag{1}$$

where N is the number of objects; K is the number of multisensory features expressed by at least one object; K_h is the number of features with history h where h is the matrix column for that feature interpreted as a binary number; H_N is the Nth harmonic number; m_k is the number of

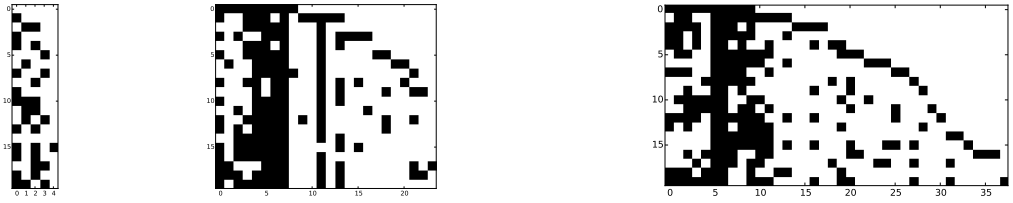
objects with feature k; and α is a variable influencing the number of features. Succinctly, Equation 1 is stated as: $Z \sim \text{IBP}(\alpha)$, where α is the parameter of the process; that is, Z has an IBP distribution with parameter α.

What is the meaning of α? Perhaps the most intuitive way to understand the answer to this question is to recast $p(Z)$ in Equation 1 through the spectrum of an Indian restaurant serving an infinite number of dishes at a buffet.

Customers (observations) sequentially arrive at an Indian buffet that offers an infinite number of dishes (features). The first customer arrives and samples the first Poisson(α) dishes, $\alpha > 0$. The ith customer samples one of the already sampled dishes with probability proportional to the popularity of that dish prior to her arrival; that is, the probability is proportional to m_k/i where m_k is the number of previous customers who tried dish k. When she is done sampling dishes previously sampled, customer i further samples a Poisson(α/i) number of new dishes. This process continues until all N customers visit the buffet. Now, represent the outcome of this buffet process in a binary matrix Z where the rows of the matrix are customers and the columns are the dishes. The element $z_{i,k}$ is 1 if observation i possesses feature k. Then, after some algebra, it follows that the probability distribution over the random, binary matrix Z induced by this buffet process is the expression given in Equation 1.

The meaning of α is now clear. The smaller the α, the smaller the number of features with $\sum_i z_{i,k} > 0$. That is, for equal number of observations, the parameter α influences the likelihood of multiple observations sharing the same features. Hence, α is called the concentration parameter of the IBP process.

Figure 1 shows the effect of α in the IBP for simulated data where we let α equal 1, 5, or 10. The number of customers (rows) was fixed at 20. The binary matrices shown in the graph are the IBP samples where the black and white pixels correspond to 1 and 0, respectively. Navigating from the left most panel in the figure, $\alpha = 1$ resulted in a total of 4 sampled dishes, $\alpha = 5$ in 23 sampled dishes, and $\alpha = 10$ in 37 sampled dishes.



(a) $\alpha = 1$ (b) $\alpha = 5$ (c) $\alpha = 10$

Figure 1 Samples from the IBP with varying concentration parameter α

Thus when the IBP is used in the generative process of an LFM, the total number of features exhibited by N data points will be finite, but random, and this number would grow in expectation with the number of data points. Further, it is also evident from Figure 1 that the subset selection procedure behaves in a "rich-get-richer" manner—if a dish had been selected by previous customers, it would likely be selected by new arrivals to the buffet. Stating generically, therefore, if a feature appears frequently in previously observed data points it would likely continue to appear again in subsequent observations as well.

We could use the IBP as the basis for an LFM by specifying a prior on the latent factors (henceforward denoted A), as well as a likelihood model for generating observations, as shown in the following examples.

If our data are real-valued vectors, an appropriate choice for the likelihood model data could be a weighted superposition of Gaussian features:

$$\begin{aligned} Z = (Z_n)_{n=1}^N &\sim \text{IBP}(\alpha) \\ y_{nk} &\sim f \\ A_k &\sim \text{Normal}(0, \sigma_A^2 I), \qquad k = 1, 2, \dots \\ X_n &\sim \text{Normal}((Z \circ Y)A, \sigma_X^2 I), \quad n = 1, \dots, N. \end{aligned} \tag{2}$$

Here, Y is the $N \times \infty$ matrix with elements y_{nk}; A is the $\infty \times D$ matrix with rows A_k; $\circ$ is the Hadamard product; and f is a distribution over the weights for a given feature instance. Note that, while we are working with infinite-dimensional matrices, the number of non-zero columns of Z is finite almost surely, so we only need to represent finitely many columns of Y and rows of A. If we

let $f = \delta_1$, we have a model where features are either included or not in a data point, and where a feature is the same each time it appears; this straightforward model was proposed by Griffiths and Ghahramani (2005), but is generally inflexible for real-life modeling scenarios.

Letting $f = \text{Normal}(\mu_f, \sigma_f^2)$ gives Gaussian weights, yielding a nonparametric variant of Factor Analysis (Knowles and Ghahramani 2007, Teh, Görür, and Ghahramani 2007). Such a model is appropriate where Factor Analysis is commonly used, as in modeling psychometric test data, or analyzing marketing survey data.

Letting $f = \text{Laplace}(\mu_f, b_f)$ results in a heavier-tailed distribution over feature weights, yielding a nonparametric version of Independent Components Analysis (Knowles and Ghahramani 2007). This allows one to perform blind source separation where the number of sources is unknown, making it a potentially useful tool in signal processing applications.

Often, one encounters binary-valued data: for example, an indicator vector corresponding to disease symptoms (where a 1 indicates the patient exhibits that symptom), or purchasing patterns (where a 1 indicates that a consumer has purchased that product). In these cases, a weighted superposition model is not directly applicable, but it may be reasonable to believe there are multiple latent causes influencing whether an element is turned on or not. One option in such cases is to use the IBP with a likelihood model (Wood, Griffiths, and Ghahramani 2006) where observations are generated according to:

$$Z = (Z_n)_{n=1}^N \sim \text{IBP}(\alpha)$$

$$y_{nk} \sim \text{Bernoulli}(p)$$

$$P(x_{nd} = 1 | Z, Y) = 1 - (1 - \lambda)^{Z_i Y_i^T} (1 - \epsilon),$$

where Y is the $N \times \infty$ matrix with elements y_{nk}; Z_i and Y_i are the ith rows of Z and Y respectively.

The above illustrations exemplify the value of IBP priors in LFMs. While these illustrations cover a vast range of applied problems, there are limitations. Notable among them is that the above LFMs do not encapsulate time dynamics. The aim of this paper is to develop a new family of IBP-based LFMs that obviates this crucial shortcoming. Additionally, unlike the afore-described

models, the new class also allows one to capture the repeat occurrence of a feature through time; i.e., persistence of latent factors. (Recall from the Introduction the example of a flautist's musical note persisting in successive time intervals.)

4. Temporal Dynamics in npLFMs

The IBP, like its finite-dimensional analogues, assumes that the data are exchangeable; that is, no information is conveyed by the order in which data are observed. But, as described in the Introduction, this could be a restrictive assumption in many applications where we discussed the differences between *longitudinal (time) dependence* and *epidemiological dependence* in observed data. While the latter has been more commonly studied in the nonparametric literature (see Foti and Williamson (2015) for a review), it is not directly applicable to the longitudinal case which is the focus of this paper. And so, from a modeling perspective, to include temporal dynamics, it is first necessary to mathematically recast the IBP to better exploit its useful properties while modeling longitudinal dependence.

4.1. A New Family of npLFMs for Time-Dependent Data

Some nonparametric priors have been proposed to model longitudinally (or time) dependent data, including temporally dependent versions of the IBP (Williamson, Orbanz, and Ghahramani 2010, Gershman, Frazier, and Blei 2015). However, these methods rely on explicitly or implicitly varying the underlying latent feature probabilities—a difficult task— and inference tends to be computationally complex. Furthermore, with the exception of Gershman, Frazier, and Blei (2015), these methods are invariant under time reversal, rendering them unsuitable for temporal dynamics.

Our proposed method obviates these limitations. In a nutshell, unlike existing dependent npLFMs, we build our model on top of a single IBP, as described in Section 3. Temporal dependence is encapsulated via a *likelihood model*.

The value of our approach could be best understood via some simple examples. Consider audio data. A common approach to modeling audio data is to view them as a superposition of multiple

sources; for example, individual speakers or different instruments. The IBP has previously been used for blind source separation (Knowles and Ghahramani 2007) and for modeling music data (Doshi-Velez 2009), but these approaches ignore the *temporal dynamics* present in most audio data. Once again recall the example from the Introduction: at very short time intervals, if a flautist is playing a B♭ at time t, it is likely that note would still be playing at time $t+1$. Our proposed model captures this dependency in the musical notes from time t to $t+1$. In Section 6, using real data, we show the benefit of incorporating this dynamic, temporal *feature persistence* when compared to a static IBP model.

As noted in the Abstract, another illustration is the modeling of sensor outputs over time. Sensors record responses to a variety of external events: for example, in a building we may have sensors recording temperature, humidity, power consumption and noise levels. These are all altered by events happening in the building—the presence of individuals; the turning on and off of electrical devices; and so on. Latent factors influencing the sensor output are typically present for a contiguous period of time before disappearing; besides, multiple factors could be present at a time. Thus, for instance, our model should capture the effect on power consumption due to an air conditioning unit being turned on from 9am to 5pm, and which could be subject to latent disturbances during that time interval such as voltage fluctuations.

Consider a third illustration involving medical signals such as EEG or ECG data. Here, we could capture latent features (or factors) causing patterns in the data, as well as infer the duration of their influence. As in previous examples, we expect factors to contribute for a contiguous period of time: for instance, a release of stress hormones would affect all time periods until the levels decrease below a threshold. Note that this cannot be accurately captured with standard factor models where the probability of a factor varies smoothly, but the actual presence or absence of that feature is independently sampled given appropriate probabilities. This leads to noisy data where a feature randomly oscillates between on and off.

Under the linear Gaussian likelihood model described in Equation 2, conditioned on the latent factors A_k, the nth datum is characterized entirely by the nth row of the IBP-distributed matrix

Z thereby ensuring that the data, like the rows of Z, are exchangeable. In the following, the key point of departure from the npLFMs described earlier is this: we now let the nth datum depend not only on the nth row of Z, but also on the $n-1$ preceding rows, thus breaking the exchangeability of the X_n data sequence. This is the mathematical equivalent of dependency in the data that we now formalize.

Associate each non-zero element z_{nk} of Z with a geometrically-distributed "lifetime", namely $\ell_{nk} \sim \text{Geometric}(\rho_k)$. An instance of the kth latent factor is then incorporated from the nth to the $(n+\ell_{nk}-1)$th datum. The nth datum is therefore associated with a set $\mathcal{Y}_n$ of feature indices $\{(i,j) : z_{ij} = 1, i+\ell > n\}$. We use the term "feature" to refer to a factor, and the term "feature instance" to refer to a specific realization of that factor. For example, if each factor corresponds to a single note in an audio recording, the global representation of the note C would be a factor, and the specific instance of note C that starts at time n and lasts for a geometrically distributed time would be a factor instance. If we assume a shared lifetime parameter, $\rho_k = \rho$ for all features, then the number of features at any time point is given, in expectation, by a geometric series $E[|\mathcal{Y}_n|] = \sum_{i=0}^{n-1} \alpha\rho^i \to \frac{\alpha}{1-\rho}$ as $n \to \infty$, i.e. as we forget the start of the process. More generally, we allow ρ_k to differ between features, and place a $\text{Beta}(a_\rho, b_\rho)$ prior on each ρ_k. By a judicious choice of the hyper-parameters, this prior could be easily tailored to encapsulate vague prior knowledge or contextual knowledge. (As an added bonus, it leads to simpler stochastic simulation methods which will be discussed later on.)

This geometric lifetime is the source of dependency in our new class of IBP-based LFMs. It captures the idea of feature *persistence*: a feature instance "turned on" at time t appears in a geometrically distributed number of future time steps. Since any feature instance that contributes to x_n also contributes to x_{n+1} with probability ρ_k, we expect x_n to share $\frac{\alpha+\rho-1}{1-\rho}$ feature instances with x_{n-1}, and to introduce α new feature instances. Of these new feature instances, we expect α/n to be versions of previously unseen features.

Note that this construction allows a specific datum to exhibit multiple instances of a given latent factor. For example, if $\mathcal{Y}_n = \{(n,1),(n,3),(n-1,1)\}$, then the nth datum will exhibit two copies of

the first feature and one copy of the third feature. In many settings, this is a reasonable assumption: two trees appearing in a movie frame, or two instruments playing the same note at the same time.

The construction of dependency detailed above could now be combined with a variety of likelihood functions (or models) appropriate for different data sources or applications. Armed with this kernel of geometric dependency and likelihood functions, we now illustrate the broad scope of the proposed family of time-dependent npLFMs via two generalizations. Later, we demonstrate these using real or synthetic data.

Adapting the Gaussian IBP LFM of Equation 2 to our dynamic time-dependent model, where each datum is given by a linear superposition of Gaussian features, results in:

$$\begin{aligned} Z \sim & \text{IBP}(\alpha) \\ A_k \sim & \text{Normal}(0, \sigma_A^2) \\ \ell_{nk} \sim & \text{Geometric}(\rho_k), \qquad k = 1, 2, \dots \\ \mu_n = & \textstyle\sum_{i=1}^{n} \sum_{k=1}^{\infty} z_{ik} I(i + \ell_{ik} > n) A_k \\ X_n \sim & \text{Normal}(\mu_k, \sigma_X^2), \qquad n = 1, \dots, N, \end{aligned} \tag{3}$$

where $I()$ is the indicator function.

Consider a second generalization where one wishes to model variations in the appearance of a feature. For example, in modeling audio data, a note or chord might be played at different volumes throughout a piece. In this case, it is appropriate to incorporate a per-factor-instance gamma-distributed weight as follows:

$$\begin{aligned} Z \sim & \text{IBP}(\alpha) \\ A_k \sim & \text{Normal}(0, \sigma_A^2) \\ \ell_{nk} \sim & \text{Geometric}(\rho_k), \qquad k = 1, 2, \dots \\ b_{nk} \sim & \text{Gamma}(\alpha_B, \beta_B) \\ \mu_n = & \textstyle\sum_{i=1}^{n} \sum_{k=1}^{\infty} z_{ik} b_{ik} I(i + \ell_{ik} > n) A_k \\ X_n \sim & \text{Normal}(\mu_k, \sigma_X^2), \qquad n = 1, \dots, N. \end{aligned} \tag{4}$$

Having introduced the new family of time-dependent npLFMs, we now turn to the implementational aspects of these models.

5. Inference Methods for npLFMs

A number of inference methods have been proposed for the IBP, including Gibbs samplers (Griffiths and Ghahramani 2005, Teh, Görür, and Ghahramani 2007), variational inference algorithms (Doshi et al. 2009), and sequential Monte Carlo samplers (Wood and Griffiths 2006). In this work, we focus on Markov chain Monte Carlo (MCMC) approaches (like the Gibbs sampler) since, under certain conditions, they are guaranteed to asymptotically converge to the true posterior distributions of the random parameters. Additionally, having tested various simulation methods for the dynamic models introduced in this paper, we found that the MCMC approach is easier to implement with good mixing properties.

5.1. An MCMC Algorithm for the Indian Buffet Process

MCMC is the most commonly used inference technique for the IBP. Under this broad technique, one could either use Gibbs sampling, Metropolis-Hastings sampling, or a combination of the two, to iteratively sample the latent feature allocations, feature parameters, and various hyperparameters.

We first construct MCMC algorithms for the weighted model in Equation 2, where the feature instance weights b_{nk} are distributed according to some arbitrary distribution $f(b)$. We choose this model to start with, because it is relatively easy to modify the resulting algorithm to implement temporal dynamics detailed in the previous section.

When working with nonparametric models, we are faced with a choice. One, perform inference on the full nonparametric model by assuming infinitely many features *a priori* and inferring the appropriate number of features required to model the data. Two, work with a large, K-dimensional model that converges (in a weak-limit sense) to the true posterior distributions as K tends to infinity. The former approach will asymptotically sample from the true posterior distributions, but the latter approximation approach is often preferred in practice due to lower computational costs. We describe algorithms for both approaches.

Consider Equation 2. Conditioned on Z, the distribution over A and B are independent whether we use a weak limit approximation or not. In the model under consideration, the latent features A_k are normally distributed; one could either sample these features or integrate them out. Suppose the features are integrated out, then one only needs to sample the binary matrix Z and the feature instance weights y_{nk} (Griffiths and Ghahramani 2005). This leads to a faster mixing rate in terms of number of iterations in the overall MCMC, but the per-iteration speed is slower (Doshi-Velez and Ghahramani 2009). Here, we instantiate the A_k, making it easier for the reader to substitute other distributions for A_k that may not be easily marginalized over; details of the appropriate conditional distributions for the marginalized sampler can be found in Griffiths and Ghahramani (2005), Doshi-Velez and Ghahramani (2009) and Knowles and Ghahramani (2007). In this uncollapsed setting, the data likelihood is given by

$$P(X|Z,A,B,\sigma_X) = \prod_n \text{Normal}(X_n; \textstyle\sum_k z_{nk} b_{nk} A_k, \sigma_X^2 \mathbf{I}). \tag{5}$$

From an MCMC perspective, once the likelihood function and priors have been specified, it only remains to be able to sample from the full conditional distributions of all the parameters in the posterior joint distribution. The following provides details on sampling these conditional distributions for the above model.

Sampling Z in the Full Nonparametric Model: When sampling from the conditional distribution over the nth row of Z in the full nonparametric model, we first consider the elements z_{nk} where $m_k^{-n} = \sum_{i \neq n} z_{ik} > 0$. Here, the conditional probability $P(z_{nk} = 1 | \{z_{ij} : (i,j) \neq (n,k)\}) = \frac{m_k^{-n}}{N}$; we can use this conditional probability, combined with the data likelihood stated above, to accept or reject a Metropolis-Hastings proposal for z_{nk}.

Concretely, if $z_{nk} = 1$, we propose Z^* and B^*, where $z_{nk}^* = 0$ and $b_{nk}^* = 0$, and accept with probability

$$\min\left(1, \frac{N - m_k^{-n}}{m_k^{-n}} \frac{P(X|Z^*, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)}\right).$$

If $z_{nk} = 0$, we propose Z^* and B^*, where $z^*_{nk} = 1$ and b^*_{nk} is sampled from the prior distribution $f(b)$ on elements of B, and accept with probability

$$\min\left(1, \frac{m_k^{-n}}{N - m_k^{-n}} \frac{P(X|Z^*, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)}\right)$$

where the data likelihood is given in Equation 5. An alternative approach is to use a Gibbs sampler to choose between 0 and 1; however a Metropolis proposal will tend to mix faster.

For the nth row of Z, we can sample the number of singleton features — i.e., features where $z_{nk} = 1$ but $\sum_{i \neq n} z_{ik} = 0$ — using a Metropolis Hastings step. We sample the number K^* of singletons in our proposal from a Poisson(α/N) distribution, and sample corresponding values of $b^*_{nk} \sim f(b)$. We then accept the new Z and B with probability

$$\min\left(1, \frac{P(X|Z^*, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)}\right).$$

The difficulty in sampling Z using the full nonparametric model is that there is no mathematically explicit form to calculate the probability of each element in Z. This is one of the main reasons the weak-limit approximation to sampling Z (given below) is often preferred in practice.

Sampling Z using a Weak-Limit Approximation: In this set up, we assume

$$\begin{aligned} \pi_k \sim &\text{Beta}\left(\frac{\alpha}{K}, 1\right), \quad k = 1, \ldots, K \\ z_{nk} \sim &\text{Bernoulli}(\pi_k), \quad n = 1, \ldots, N \end{aligned} \tag{6}$$

and note that this converges (in a weak limit sense, as $K \to \infty$) to the beta-Bernoulli process representation of the IBP. Inference is more straightforward since we always have an explicit form for the probability of a given element z_{nk}, even if $m_k^{-n} = 0$. Concretely, we know that $P(z_{nk} = 1 | \{z_{ij} : (i,j) \neq (n,k)\}, \alpha) = \frac{m_k^{-n} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$. We can use this in a Metropolis-Hastings sampler. If $z_{nk} = 1$, we propose Z^* and B^*, where $z^*_{nk} = 0$ and $b^*_{nk} = 0$, and accept with probability

$$\min\left(1, \frac{N - m_k^{-n}}{m_k^{-n} + \frac{\alpha}{K}} \frac{P(X|Z^*, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)}\right).$$

If $z_{nk} = 0$, we propose Z^* and B^*, where $z^*_{nk} = 1$ and b^*_{nk} is sampled from the prior distribution $f(b)$ on elements of B, and accept with probability

$$\min\left(1, \frac{m_k^{-n} + \frac{\alpha}{K}}{N - m_k^{-n}} \frac{P(X|Z^*, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)}\right)$$

where the data likelihood is given in Equation 5.

Sampling A: Conditioned on Z and B, the feature matrix A is normally distributed with mean

$$\mu_A = \left((Z \circ B)^T (Z \circ B) + \frac{\sigma_X^2}{\sigma_A^2} \mathbf{I} \right)^{-1} (Z \circ B)^T X$$

and block-diagonal covariance, with each column of A having the same covariance

$$\Sigma_A = \sigma_x^2 \left((Z \circ B)^T (Z \circ B) + \frac{\sigma_X^2}{\sigma_A^2} \mathbf{I} \right)^{-1}.$$

Sampling B: In general, the conditional distribution for b_{nk} will not be available in analytic form; however, we could still use a Metropolis-Hastings proposal to sample from the conditional distribution $P(b_{nk}|X, Z, A, \sigma_X^2)$. For example, we could sample $b_{nk}^* \sim f(b)$ and accept with probability

$$\min \left(1, \frac{P(X|Z, A, B^*, \sigma_X)}{P(X|Z, A, B, \sigma_X)} \right).$$

Sampling Hyperparameters: Without loss of generality, place inverse gamma priors on σ_X^2 and σ_A^2; then, we can easily sample from their conditional distributions due to conjugacy. Similarly, if we place a Gamma(a_α, b_α) prior on α, we can sample from its conditional distribution

$$\alpha|Z \sim \text{Gamma} \left(K + a_\alpha, \frac{b_\alpha}{1 + b_\alpha H_n} \right)$$

where H_N is the Nth harmonic number.

The reason these prior distributions are general stems from the fact they can be made non-informative or informative by a judicious choice of the corresponding prior parameters that define them; see Chick (2006) for a review of Bayesian methodology.

5.2. An MCMC Algorithm for the New Class of Time-Dependent npLFMs

To extend the MCMC algorithms in Section 5.1 to the dynamic npLFM described in Section 4.1, we must sample not only whether feature k is instantiated in observation n, but also for the number of observations for which the particular feature remains active. We can jointly infer the IBP-distributed matrix Z and the lifetimes ℓ_{nk} by extending the Metropolis-Hastings algorithms

in the previous section; as before, we can either operate in a fully nonparametric setting, or in a weak-limit approximation.

The IBP-distributed matrix Z and the set of lifetimes $\{\ell_{nk}\}$ fully define the sets $\{\mathcal{Y}_n\}$ of feature indices contributing to each datum. Conditioned on $\mathcal{Y}_n$, inference for A and B is similar to the static setting, reducing to the latter case when all the lifetimes $\ell_{nk} = 1$.

Sampling Z and the ℓ_{nk} in the Full Nonparametric Model: We jointly sample the feature instance matrix Z and the corresponding lifetimes ℓ using a combination of Metropolis-Hastings moves. Let Λ be the matrix whose elements are given by $\lambda_{nk} := z_{nk}\ell_{nk}$. To sample a new value for λ_{nk} where $\sum_{i\neq n} \lambda_{ik} > 0$, we sample $\lambda^*_{nk} \sim q(\lambda^*_{nk}; \lambda_{nk})$. If both λ^*_{nk} and λ_{nk} are greater than zero, we accept with probability

$$\min\left(1, \frac{q(\lambda_{nk}; \lambda^*_{nk})}{q(\lambda^*_{nk}; \lambda_{nk})}(1-\rho_k)^{\lambda^*_{nk}-\lambda_{nk}} \frac{P(X|\Lambda^*, A, B, \sigma_X^2)}{P(X|\Lambda, A, B, \sigma_X^2)}\right).$$

If $\lambda_{nk} = 0$ and $\lambda^*_{nk} \neq 0$, the underlying z_{nk} and z^*_{nk} are different, so the acceptance probability becomes

$$\min\left(1, \frac{q(\lambda_{nk}; \lambda^*_{nk})}{q(\lambda^*_{nk}; \lambda_{nk})}\rho_k(1-\rho_k)^{\lambda^*_{nk}} \frac{m_k^{-n}}{N - m_k^{-n}} \frac{P(X|\Lambda^*, A, B, \sigma_X^2)}{P(X|\Lambda, A, B, \sigma_X^2)}\right).$$

Similarly, if $\lambda^*_{nk} = 0$ and $\lambda_{nk} \neq 0$, the acceptance probability is

$$\min\left(1, \frac{q(\lambda_{nk}; \lambda^*_{nk})}{q(\lambda^*_{nk}; \lambda_{nk})} \frac{1}{\rho_k(1-\rho_k)^{\lambda_{nk}}} \frac{N - m_k^{-n}}{m_k^{-n}} \frac{P(X|\Lambda^*, A, B, \sigma_X^2)}{P(X|\Lambda, A, B, \sigma_X^2)}\right).$$

We alternate between two proposal distributions. We propose small changes by letting

$$q(\lambda^*_{nk}; \lambda_{nk}) = \begin{cases} 0.5\delta_{\lambda_{nk}-1} + 0.5\delta_{\lambda_{nk}+1} & \text{if } \lambda_{nk} > 0 \\ \delta_1 & \text{otherwise} \end{cases}$$

If $\lambda_{nk} > 0$, we can also propose larger changes by letting

$$q(\lambda_{nk}*; \lambda_{nk}) = \text{Geometric}(\rho_k).$$

For the nth row of Z, we can sample the number of singleton features — i.e. features where $z_{nk} = 1$ but $\sum_{i\neq n} z_{ik} = 0$ — using a Metropolis Hastings step that is very similar to the non-dynamic

case. As before, we sample the number K^* of singletons in our proposal from a Poisson(α/N) distribution, and sample corresponding values of $b^*_{nk} \sim f(b)$. We also sample corresponding lifetime probabilities $\rho^*_k \sim \text{Beta}(a_\rho, b_\rho)$ and lifetimes $\ell^*_{nk} \sim \text{Geometric}(\rho^*_k)$ for the proposed singleton features. We then accept the new Λ and B with probability

$$\min\left(1, \frac{P(X|\Lambda^*, A, B^*, \sigma_X)}{P(\Lambda, A, B, \sigma_X)}\right).$$

Sampling Z and the ℓ_{nk} using a Weak-Limit Approximation: As before, inference in the weak-limit setting is more straightforward since we do not have to worry about adding and deleting new features. We modify the Metropolis-Hastings methods for the full nonparametric models, proposing a new value λ^*_{nk} and accepting with probability

$$\begin{cases} \min\left(1, \frac{q(\lambda_{nk};\lambda^*_{nk})}{q(\lambda^*_{nk};\lambda_{nk})}(1-\rho_k)^{\lambda^*_{nk}-\lambda_{nk}} \frac{P(X|\Lambda^*,A,B,\sigma_X^2)}{P(X|\Lambda,A,B,\sigma_X^2)}\right) & \text{if } \lambda_{nk}>0 \text{ and } \lambda^*_{nk}>0 \\ \min\left(1, \frac{q(\lambda_{nk};\lambda^*_{nk})}{q(\lambda^*_{nk};\lambda_{nk})}\rho_k(1-\rho_k)^{\lambda^*_{nk}} \frac{m_k^{-n}+\frac{\alpha}{K}}{N-m_k^{-n}} \frac{P(X|\Lambda^*,A,B,\sigma_X^2)}{P(X|\Lambda,A,B,\sigma_X^2)}\right) & \text{if } \lambda_{nk}=0 \text{ and } \lambda^*_{nk}>0 \\ \min\left(1, \frac{q(\lambda_{nk};\lambda^*_{nk})}{q(\lambda^*_{nk};\lambda_{nk})} \frac{1}{\rho_k(1-\rho_k)^{\lambda_{nk}}} \frac{N-m_k^{-n}}{m_k^{-n}+\frac{\alpha}{K}} \frac{P(X|\Lambda^*,A,B,\sigma_X^2)}{P(X|\Lambda,A,B,\sigma_X^2)}\right) & \text{if } \lambda_{nk}>0 \text{ and } \lambda^*_{nk}=0 \end{cases} \tag{7}$$

We use the same proposal distributions as in the full nonparametric setting.

Sampling A and B: Conditioned on Z and the ℓ_{nk}, inferring A and B is similar to the static setting, and does not depend on whether we used a weak-limit approximation for sampling Z. Recall that $\mathcal{Y}_n$ is the vector of feature indices $\{(i,j) : z_{ij}=1, i+\ell>n\}$. Let Y be the matrix with elements $y_{nk} = \sum_{i:(i,k)\in\mathcal{Y}_n} b_{nk}$ – i.e. the total weight given to the kth feature in the nth observation. Then conditioned on Y and B, the feature matrix A is normally distributed with mean

$$\mu_A = \left(Y^TY + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{I}\right)^{-1} Y^TX$$

and block-diagonal covariance, with each column of A having the same covariance

$$\Sigma_A = \sigma_x^2\left(Y^TY + \frac{\sigma_X^2}{\sigma_A^2}\mathbf{I}\right)^{-1}.$$

As before, we can use a Metropolis-Hastings proposal to sample from the conditional distribution $P(b_{nk}|X, Z, \{\ell_{nk}\}, A, \sigma_X^2)$ — for example, sampling $b^*_{nk} \sim f(b)$ and accepting with probability

$$\min\left(1, \frac{P(X|Z, \{\ell_{nk}\}, A, B^*, \sigma_X)}{P(X|Z, \{\ell_{nk}\}, A, B, \sigma_X)}\right).$$

Sampling Hyperparameters: We can use the same approach to sample σ_X, σ_A and α as in the static case. We can sample the parameters $\rho_k \sim \text{Beta}(a_\rho + m_k, b_\rho + \sum_{n:z_{nk}=1} \ell_{nk})$, due to conjugacy.

6. Experimental Evaluation

Here the proposed models and stochastic simulation methods are exemplified via synthetic and real data illustrations. In the synthetic illustration, we used the full nonparametric simulation method; in the real data examples, we used the weak-limit approximation version of the MCMC algorithm to sample the nonparametric component by setting the number of unobserved latent features K to 20; see earlier sections for details on this and related ideas.

The "gold standard" in assessing npLFMs is to first set aside a hold-out sample. Then, using the estimated parameters one predicts these held-out data; i.e., comparing actual versus predicted values. Since the aim is to compare the traditional, static npLFM models with the dynamic (or time-dependent) npLFM models developed in this paper, the $L2$ norm is used to contrast the performance of these approaches on the held-out samples. The $L2$ norm minimizes the sum of the squared differences between the actual and predicted values. For a variety of reasons (such as less sensitivity to outliers), $L2$ is generally preferred to the $L1$ norm, where the latter minimizes the sum of absolute differences between the actual and predicted values.

Also, in the interests of space we have not shown the convergence charts of the various parameters from the MCMC algorithms; they are available on request.

6.1. Synthetic Data

To show the benefits of explicitly addressing temporal dependence, we carried out the following.

- Generate a synthetic dataset with eight binary 8×8 features, shown in Figure 2; these features were used to generate a longitudinally varying dataset.
- Simulate a sequence of $N = 500$ data points, corresponding to N time steps.
- For each time step, add a new instance of each feature with probability 0.2, then sample an active lifetime for that feature instance according to a geometric distribution with parameter 0.5.

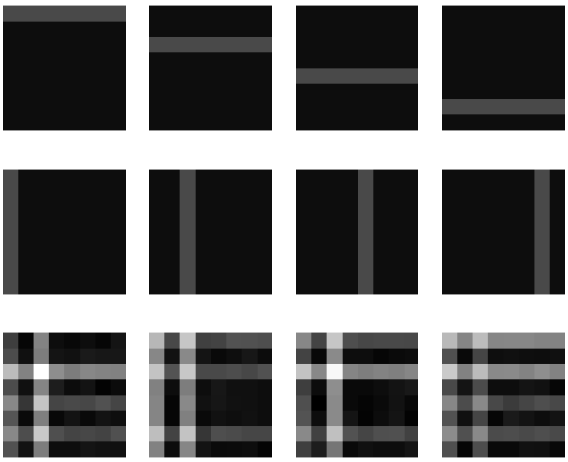


Figure 2 Top two rows: Eight synthetic features used to generate data. Bottom row: Four consecutive observations.

- Each datum was generated by superimposing all the active feature instances (i.e. those whose lifetimes had not expired) and adding Gaussian noise to give an 8×8 real-valued image.
- We designated 10% of the observations as our test set. For each test set observation, we held out 56 of the 64 pixels. The remaining 8 pixels allowed us to infer the features associated with the test set observations.

Table 1 shows the $L2$ error obtained on the held-out data; the number of features; and the average feature persistence; all values are averaged over 500 samples from the appropriate posterior distributions following convergence of the MCMC chain. We see that the average $L2$ error is significantly lower (0.4962 ± 0.0052) for the dynamic model in contrast to the static model (0.5960 ± 0.0061). Next, consider Figure 3 that shows the total number of times each feature contributes to a data point (i.e., the sum of that feature's lifetimes), based on a single iteration from each model. It is clear that the dynamic model reuses common features a larger number of times than the static model.

There are two reasons for this superior performance. First, consider a datum with two instances of a given feature: one that has just been introduced, and one that has persisted from a previous time-point. The dynamic model is able to use the same latent feature to model both feature instances, while the static model must use two separate features (or model this double-instance as a separate

	MSE	Number of features	Average persistence
Dynamic npLFM	0.4962 ± 0.0052	15.80 ± 0.897	2.856 ± 0.010
Static npLFM	0.5960 ± 0.0061	23.13 ± 0.336	1

Table 1 Average MSE; number of features; and feature persistence on synthetic data under static and dynamic npLFMs.

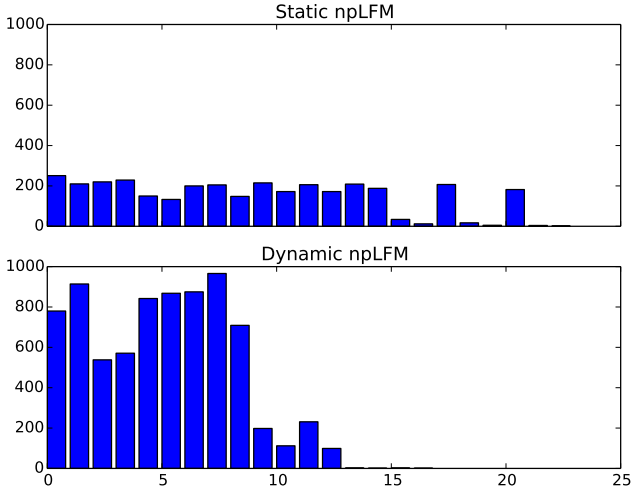


Figure 3 Number of times each feature contributes to a data point under static and dynamic npLFMs. Note that under the dynamic model, a feature can contribute multiple times to the same data point.

feature from a single-instance). This is seen in the lower average number of features required by the dynamic model (Table 1, and in the much greater number of times common features are reused (Figure 3).

Second, the dynamic npLFM makes use of the ordering information, and anticipates that feature instances will persist for multiple time periods; this means that the latent structure for a given test-set observation is informed not only by the eight observed pixels, but also by the latent structures of the adjacent observations. We see that the average feature persistence is 2.856, which confirms that the dynamic model makes use of the temporal dynamics inherent in the data.

6.2. Audio Real Data Illustration

It is natural to think of a musical piece in terms of latent features, for it is made up of one or more instruments playing one or more notes simultaneously. There is clearly persistence of features,

making the longitudinal model described in Section 4.1 a perfect fit. For this illustration, we carried out the following data steps before implementing the model. (A similar dataset, gathered by Poliner and Ellis (2007), was modeled using a static npLFM by Doshi-Velez and Ghahramani (2009).)

- We evaluated the model on Bach's "Sonata No. 1 in B minor, MWV 1030 for Flute and Harpsichord".
- A midi-synthesized digital piano recording of the piece, downloaded from `http://www.jsbach.net/midi/midi_chambermusic.html`, was converted to a mono wave recording with an 8kHz sampling rate.
- We generated a sequence of $D = 64$-dimensional observations by applying a short-time Fourier transform using a 64-point discrete Fourier transform, a 64-point Hanning window, and an 80 point advance between frames.
- Hence, each datum corresponds to an 8ms segment with a 10ms advance between segments.
- The data were pre-processed by whitening each frequency component to have zero mean and unit variance.
- To evaluate the model, a hold-out sample of 10% of the data, evenly spaced throughout the piece, was set aside. We held out all but eight randomly selected dimensions.

Since notes can be played at different volumes, we used our dynamic npLFM with gamma-distributed amplitudes, described in Equation 4, to model the data. For comparison, we also used a static npLFM to model these data. In both cases, we performed inference with a maximum of $K = 20$ features. (Recall K is the bound used in the weak-limit approximation.)

Average $L2$ errors, along with confidence bands, were obtained by averaging over 500 samples from the Gibbs sampler following convergence. Static model ($L2$ average error and bounds): 1.310 ± 0.7072; Dynamic model ($L2$ average error and bounds): 0.7945 ± 0.0189. The time-dependent npLFM model clearly outperforms the static model for these data. Note that in addition to the average error being less for our model, the Monte Carlo standard error for these held-out samples is strikingly lower under our approach. One key reason for this is that our model better captures the transition of notes during successive time periods, since persistence of latent features in the musical composition is formally accounted for in the mathematical construction.

6.3. Household Power Consumption Real Data Illustration

A number of different appliances contribute to a household's overall power consumption, and each appliance will have different energy consumption and operating patterns. We analyzed a subset of the "Individual household electric power consumption data set" available from the UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`. This dataset records overall minute-averaged active power, overall minute-averaged reactive power, minute-averaged voltage, overall minute-averaged current intensity, and watt-hours of active energy on three sub-circuits within one house.

We examined 200 recordings taken at five-minute spacings. Based on the assumption that a given appliance's energy demands are approximately constant, we applied our dynamic npLFM with constant weights described in Equation 3. As with the audio data, we used a weak limit sampler with a maximum of 20 features. For validation, 10% of the data were set aside, with a randomly selected six out of seven dimensions being held out.

The held-out set average $L2$ errors with bounds are: Static model: 45.18 ± 6.494; Dynamic model: 17.78 ± 4.874. Like the previous example, our dynamic model outperforms its static counterpart. This makes sense when considering the underlying data generating process: electricity demand is dictated by which appliances and systems are currently drawing power. Most appliances are used for contiguous stretches of time — for example, we turn a light on when we enter a room, and turn it off when we leave some time later. Further, many appliances have characteristic periods of use: a microwave is typically on for a few minutes, while a washing machine is on for around an hour. The static, exchangeable model ignores these patterns, whereas the dynamic model captures these persistent latent features.

7. Conclusion

This paper introduces a new family of latent factor (or feature) models for time-dependent data. Unobserved latent features are often subject to temporal dynamics for data arising in a multitude of applications in industry. Static models for time-dependence exist but, as shown in this work,

such approaches disregard key insights that could be gained if time dependency was modeled dynamically. Synthetic and real data illustrations exemplify the improved predictive accuracy while using time-dependent, nonparametric latent feature models. General algorithms to sample from the new family developed here could be easily adapted to model data arising in different applications where the likelihood function changes. In other words, the algorithms are fairly generic.

This paper focused on temporal dynamics for random, *fixed*, time-dependent data using nonparametric LFMs. But if data are *changing* in real time, as in moving images in a film, then the notion of temporal dependency needs a different treatment than the one developed here. In addition to the mathematical challenges this proposed extension presents, the computational challenges are daunting as well. Preliminary theoretical and empirical work show promise. Details will be reported elsewhere.

Acknowledgments

Sinead Williamson and Michael Zhang are supported by NSF grant 1447721.

References

Acton RS, Ray N (2006) *Biomedical Image Analysis: Tracking* (Morgan and Claypool Publishers).

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2(6):1152–1174.

Baum LE, Petrie T (1996) Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics* 37(6):1554–1563.

Bishop CM (1998) Latent variable models. *Learning in graphical models* 371–403.

Broderick T, Mackay L, Paisley J, Jordan MI (2015) Combinatorial clustering and the beta negative binomial process. *Pattern Analysis and Machine Intelligence* 37(2):290–306.

Cattell RB (1952) *Factor analysis* (Harper).

Chick SE (2006) Subjective probability and Bayesian methodology. *Handbooks in Operations Research and Management Science* 13:225–257.

Doshi F, Miller K, Van Gael J, Teh YW (2009) Variational inference for the Indian buffet process. *Artificial Intelligence and Statistics*.

Doshi-Velez F (2009) *The Indian buffet process: Scalable inference and extensions*. Master's thesis, University of Cambridge.

Doshi-Velez F, Ghahramani Z (2009) Accelerated sampling for the Indian buffet process. *International Conference on Machine Learning*.

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1(2):209–230.

Foti NJ, Futoma JD, Rockmore DN, Williamson S (2013) A unifying representation for a class of dependent random measures. *Artificial Intelligence and Statistics*.

Foti NJ, Williamson SA (2015) A survey of non-exchangeable priors for Bayesian nonparametric models. *Pattern Analysis and Machine Intelligence* 37(2):359–371.

Gershman SJ, Frazier P, Blei DM (2015) Distance dependent infinite latent feature models. *Pattern Analysis and Machine Intelligence* 37(2):334–345.

Griffiths TL, Ghahramani Z (2005) Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*.

Hanson KM (1993) Introduction to Bayesian image analysis. *Medical Imaging 1993*, 716–731 (International Society for Optics and Photonics).

Hjort NL (1990) Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* 18(3):1259–1294.

Hochreiter S, Clevert DA, Obermayer K (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics* 22(8):943–949.

Hyvärinen A, Karhunen J, Oja E (2001) *Independent Component Analysis* (Wiley).

Knowles D, Ghahramani Z (2007) Infinite sparse factor analysis and infinite independent component analysis. *Independent Component Analysis*.

Kopperapu KS, Desai BU (2006) *Bayesian Approach to Image Interpretation* (Springer Science and Business Media).

Leemis ML (2006) Arrival processes, random lifetimes, and random objects. *Handbook in Operations Research and Management Science* 13:155–180.

Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* 35(1):1–41.

Lo AY (1984) On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* 12(1):351–357.

McLauchlan GJ (2000) *Finite mixture models* (Wiley).

Meeds E, Ghahramani Z, Neal R, Roweis ST (2007) Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*.

Poliner GE, Ellis DPW (2007) A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing* 2007(1):154–154.

Press S, Shigemasu K (1989) Bayesian inference in factor analysis. *Contributions to probability and statistics*, 271–287 (Springer).

Rao V, Teh YW (2009) Spatial normalized gamma processes. *Advances in Neural Information Processing Systems*.

Ren L, Wang Y, Carin L, Dunson DB (2011) The kernel beta process. *Advances in Neural Information Processing Systems*, 963–971.

Roweis ST, Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Computation* 11(2):305–345.

Ruiz FJR, Valera I, Blanco C, Perez-Cruz F (2014) Bayesian nonparametric comorbidity analysis of psychiatric disorders. *Journal of Machine Learning Research* 15:1215–1247.

Shi W, Wedel M, Pieters R (2013) Information acquisition during online decision making: A model-based exploration using eye-tracking data. *Management Science* 59(5):1009–1026.

Stark H (1987) *Image Recovery: Theory and Application* (Academic Press).

Tait M (1986) The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel psychology* 986:39.

Tavares J, Jorge N (2011) *Computational Vision and Medical Image Processing* (Springer).

Teh TW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476):1566–1581.

Teh YW, Görür D, Ghahramani Z (2007) Stick-breaking construction for the Indian buffet process. *Artificial Intelligence and Statistics*.

Thibaux R, Jordan MI (2007) Hierarchical beta processes and the Indian buffet process. *Artificial Intelligence and Statistics*, 564–571.

Titsias M (2007) The infinite gamma-Poisson feature model. *Advances in Neural Information Processing Systems*.

Venkaiah K, Brahmam GNV, Vijayaraghavan K (2011) Application of factor analysis to identify dietary patterns and use of factor scores to study their relationship with nutritional status of adult rural populations. *Journal of Health, Population, and Nutrition* 29(4):327–338.

Walker SG, Damien P, Laud PW, Smith AFM (1999) Bayesian inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society-B* 61:485–527.

Wang L, Zhao G, Cheng L, Pietikainen M (2013) *Machine Learning for Vision-Based Motion Analysis: Theory and Techniques* (Springer).

Williamson S, Orbanz P, Ghahramani Z (2010) Dependent Indian buffet processes. *Artificial Intelligence and Statistics*.

Winkler G (2003) *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods* (Springer), 2 edition.

Wood F, Griffiths TL (2006) Particle filtering for nonparametric Bayesian matrix factorization. *Advances in Neural Information Processing Systems*.

Wood F, Griffiths TL, Ghahramani Z (2006) A non-parametric Bayesian method for inferring hidden causes. *Uncertainty in Artificial Intelligence*.

Zhou M, Yang H, Sapiro G, Dunson D, Carin L (2011) Dependent hierarchical beta processes for image interpolation and denoising. *Artificial Intelligence and Statistics*.