# Temporally-Reweighted Chinese Restaurant Process Mixtures for Clustering, Imputing, and Forecasting Multivariate Time Series

**Feras A. Saad**
Probabilistic Computing Project
Massachusetts Institute of Technology

**Vikash K. Mansinghka**
Probabilistic Computing Project
Massachusetts Institute of Technology

## Abstract

This article proposes a Bayesian nonparametric method for forecasting, imputation, and clustering in sparsely observed, multivariate time series data. The method is appropriate for jointly modeling hundreds of time series with widely varying, non-stationary dynamics. Given a collection of $N$ time series, the Bayesian model first partitions them into independent clusters using a Chinese restaurant process prior. Within a cluster, all time series are modeled jointly using a novel "temporally-reweighted" extension of the Chinese restaurant process mixture. Markov chain Monte Carlo techniques are used to obtain samples from the posterior distribution, which are then used to form predictive inferences. We apply the technique to challenging forecasting and imputation tasks using seasonal flu data from the US Center for Disease Control and Prevention, demonstrating superior forecasting accuracy and competitive imputation accuracy as compared to multiple widely used baselines. We further show that the model discovers interpretable clusters in datasets with hundreds of time series, using macroeconomic data from the Gapminder Foundation.

## 1 Introduction

Multivariate time series data is ubiquitous, arising in domains such as macroeconomics, neuroscience, and public health. Unfortunately, forecasting, imputation, and clustering problems can be difficult to solve when there are tens or hundreds of time series. One challenge in these settings is that the data may reflect underlying processes with widely varying, non-stationary dynamics [13]. Another challenge is that standard parametric approaches such as state-space models and vector autoregression often become statistically and numerically unstable in high dimensions [20]. Models from these families further require users to perform significant custom modeling on a per-dataset basis, or to search over a large set of possible parameter settings and model configurations. In econometrics and finance, there is an increasing need for multivariate methods that exploit sparsity, are computationally efficient, and can accurately model hundreds of time series (see introduction of [15], and references therein).

This paper presents a nonparametric Bayesian method for multivariate time series that aims to address some of the above challenges. The model is based on two extensions to Dirichlet process mixtures. First, we introduce a recurrent version of the Chinese restaurant process mixture to capture temporal dependences. Second, we add a hierarchical prior to discover groups of time series whose underlying dynamics are modeled jointly. Unlike autoregressive models, our approach is designed to interpolate in regimes where it has seen similar history before, and reverts to a broad prior in previously unseen regimes. This approach does not sacrifice predictive accuracy, when there is sufficient signal to make a forecast or impute missing data.

We apply the method to forecasting flu rates in 10 US regions using flu, weather, and Twitter data from the US Center for Disease Control and Prevention. Quantitative results show that the method outperforms several Bayesian and non-Bayesian baselines, including Facebook Prophet, multi-output Gaussian processes, seasonal ARIMA, and the HDP-HMM. We also show competitive imputation accuracy with widely used statistical techniques. Finally, we apply the method to clustering hundreds of macroeconomic time series from Gapminder, detecting meaningful clusters of countries whose data exhibit coherent temporal patterns.

Proceedings of the $21^{st}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. 

## 2 Related Work

The temporally-reweighted Chinese restaurant process (TRCRP) mixture we introduce in Section 3 can be directly seen as a time series extension to a family of nonparametric Bayesian regression models for cross-sectional data [18, 35, 28, 22, 23]. These methods operate on an exchangeable data sequence $\{x_i\}$ with exogenous covariates $\{\mathbf{y}_i\}$; the prior CRP cluster probability $p(z_i = k)$ for each observation $x_i$ is reweighted based on $\mathbf{y}_i$. Our method extends this idea to a time series $\{x_t\}$; the prior CRP cluster probability $p(z_t = k)$ for $x_t$ is now reweighted based on the $p$ previous values $\mathbf{x}_{t-1:t-p}$. Moreover, the hierarchical extension in Section 3.4 coincides with CrossCat [21], when all temporal dependencies are removed (by setting $p = 0$).

Temporal extensions to the Dirichlet process have been previously used in the context of dynamic clustering [38, 1]. The latter work derives a recurrent CRP as the limit of a finite dynamic mixture model. Unlike the method in this paper, those models are used for clustering batched data and dynamic topic modeling [7], rather than data analysis tasks such as forecasting or imputation in real-valued, multivariate time series.

For multivariate time series, recent nonparametric Bayesian methods include using the dependent Dirichlet process for dynamic density estimation [31]; hierarchical DP priors over the state in hidden Markov models [HDP-HMM; 12, 19]; Pitman-Yor mixtures of non-linear state-space models for clustering [26]; and DP mixtures [8] and Polya trees [27] for modeling noise distributions. As nonparametric Bayesian extensions of state-space models, all of these approaches specify priors that fall under distinct model classes to the one developed in this paper. They typically encode parametric assumptions (such as linear autoregression and hidden-state transition matrices), or integrate explicit specifications of underlying temporal dynamics such as seasonality, trends, and time-varying functionals. Our method instead builds purely empirical models and uses simple infinite mixtures to detect patterns in the data, without relying on dataset-specific customizations. As a multivariate interpolator, the TRCRP mixture is best applied to time series where there is no structural theory of the temporal dynamics, and where there is sufficient statistical signal in the history of the time series to inform probable future values.

To the best of our knowledge, this paper presents the first multivariate, nonparametric Bayesian model that provides strong baseline results without specifying custom dynamics on a problem-specific basis; and that has been benchmarked against multiple Bayesian and non-Bayesian techniques to cluster, impute, and forecast sparsely observed real-world time series data.

## 3 Temporally-Reweighted Chinese Restaurant Process Mixture Model

We first outline the notations and basic setup assumed throughout this paper. Let $\{\mathbf{x}^n : n = 1, \dots, N\}$ denote a collection of $N$ discrete-time series, where the first $T$ variables of the $n^{\text{th}}$ time series is $\mathbf{x}^n_{1:T} = (x^n_1, x^n_2, \dots, x^n_T)$. Slice notation is used to index subsequences of variables, so that $\mathbf{x}^n_{t_1:t_2} = (x^n_{t_1}, \dots, x^n_{t_2})$ for $t_1 < t_2$. Superscript $n$ will be often be omitted when discussing a single time series. The remainder of this section develops a generative process for the joint distribution of all random variables $\{x^n_t : t = 1, \dots, N, n = 1, \dots, N\}$ in the $N$ time series, which we proceed to describe in stages.

### 3.1 Background: CRP representation of Dirichlet process mixture models

Our approach is based on a temporal extension of the standard Dirichlet process mixture (DPM), which we review briefly. First consider the standard DPM in the non-temporal setting [11], with concentration $\alpha$ and base measure $\pi_\Theta$. The joint distribution of a sequence of $m$ exchangeable random variables $(x_1, \dots, x_m)$ is:

$$P \sim DP(\alpha, \pi_\Theta), \quad \theta^*_j \mid P \sim P, \quad x_j \mid \theta^*_j \sim F(\cdot \mid \theta^*_j).$$

The DPM can be represented in terms of the Chinese restaurant process [2]. As $P$ is almost-surely discrete, the $m$ draws $\{\theta^*_j\} \sim P$ contain repeated values, thereby inducing a clustering among data $x_j$. Let $\lambda_F$ be the hyperparameters of $\pi_\Theta$, $\{\theta_k\}$ be the unique values among the $\{\theta^*_j\}$, and $z_j$ denote the cluster assignment of $x_j$ which satisfies $\theta^*_j = \theta_{z_j}$. Define $n_{jk}$ to be the number of observations $x_i$ with $z_i = k$ for $i < j$. Using the conditional distribution of $z_j$ given previous cluster assignments $\mathbf{z}_{1:j-1}$, the joint distribution of exchangeable data sequence $(x_1, x_2, \dots)$ in the CRP mixture model can be described sequentially:

$$\begin{aligned}
&\{\theta_k\} \overset{\text{iid}}{\sim} \pi_\Theta(\cdot \mid \lambda_F) \\
&\Pr\left[z_j = k \mid \mathbf{z}_{1:j-1}; \alpha\right] \qquad (j = 1, 2, \dots) \\
&\quad \propto \begin{cases} n_{jk} & \text{if } 1 \le k \le \max(\mathbf{z}_{1:j-1}) \\ \alpha & \text{if } k = \max(\mathbf{z}_{1:j-1}) + 1 \end{cases} \\
&x_j \mid z_j, \{\theta_k\} \sim F(\cdot \mid \theta_{z_j})
\end{aligned} \tag{1}$$

The CRP mixture model (1), and algorithms for posterior inference, have been studied extensively for nonparametric modeling in a variety of statistical applications (for a survey see [37], and references therein).

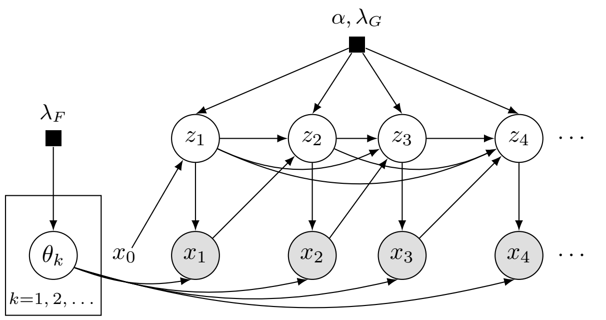



**Figure 1:** Graphical model for the TRCRP mixture in a single time series $\mathbf{x} = (x_1, x_2, \ldots)$ with lagged window size $p = 1$.

### 3.2 The temporally-reweighted CRP mixture for modeling a single time series

Our objective is to define a CRP-like process for a non-exchangeable discrete-time series $(x_1, x_2, \ldots)$, where there is now a temporal ordering and a temporal dependence among the variables. Instead of having $(x_t, z_t)$ be conditionally independent of all other data given $\mathbf{z}_{1:t-1}$ as in the CRP mixture (1), we instead consider using previous observations $\mathbf{x}_{1:t-1}$ when simulating $z_t$. The main idea in our approach is to modify the CRP prior by having the cluster probability $\Pr[z_t = k \mid \mathbf{z}_{1:t-1}]$ at step $t$ additionally account for (i) the $p$ most recent observations $\mathbf{x}_{t-p:t-1}$, and (ii) collection of lagged values $D_{tk} := \{\mathbf{x}_{t'-p:t'-1} \mid z_{t'} = k, 1 \le t' < t\}$ of earlier data points $x_{t'}$ assigned to cluster $k$. The distribution of time series $(x_1, x_2, \ldots)$ in the temporally-reweighted CRP (TRCRP) mixture is therefore:

$$\{\theta_k\} \overset{\text{iid}}{\sim} \pi_\Theta(\cdot \mid \lambda_F)$$

$$\Pr\left[z_t = k \mid \mathbf{z}_{1:t-1}, \mathbf{x}_{t-p:t-1}; \alpha, \lambda_G\right] \qquad (t = 1, 2, \ldots)$$

$$\propto \begin{cases} n_{tk}\, G(\mathbf{x}_{t-p:t-1}; D_{tk}, \lambda_G) & \text{if } 1 \le k \le \max(\mathbf{z}_{1:t-1}) \\ \alpha\, G(\mathbf{x}_{t-p:t-1}; \lambda_G) & \text{if } k = \max(\mathbf{z}_{1:t-1}) + 1 \end{cases}$$

$$x_t \mid z_t, \{\theta_k\} \sim F(\cdot|\theta_{z_t}) \tag{2}$$

The main difference between the TRCRP mixture (2) and the standard CRP mixture (1) is the term $G(\mathbf{x}_{t-p:t-1}; \lambda_G, D_{tk})$ which acts as a non-negative "cohesion" function $\mathbb{R}^p \to \mathbb{R}^+$, parametrized by $D_{tk}$ and a bundle of real values $\lambda_G$. This term measures how well the current lagged values $\mathbf{x}_{t-p:t-1}$ match the collection of lagged values of earlier data $D_{tk}$ in each cluster $k$, thereby introducing temporal dependence to the model. The smoothness of the process depends on the choice of the window size $p$: if $t_1$ and $t_2$ are close in time (relative to $p$) then they have overlapping lagged values $\mathbf{x}_{t_1-p:t_1-1}$ and $\mathbf{x}_{t_2-p:t_2-1}$, so $G$ increases the prior probability that $\{z_{t_1} = z_{t_2}\}$. More generally, any pair of time points $t_1$ and $t_2$ that share similar lagged values are a-priori more likely to have similar distributions for generating $x_{t_1}$ and $x_{t_2}$, because $G$ increases the probability that $\{z_{t_1} = z_{t_2} = k\}$, so that $x_{t_1}$ and $x_{t_2}$ are both drawn from $F(\cdot|\theta_k)$.

Figure 1 shows a graphical model for the TRCRP mixture (2) with window size $p = 1$. The model proceeds as follows: first assume the initial $p$ observations $(x_{-p+1}, \ldots, x_0)$ are fixed or have a known joint distribution. At step $t$, the generative process samples a cluster assignment $z_t$, whose probability of joining cluster $k$ is a product of (i) the CRP probability for $\{z_t = k\}$ given all previous cluster assignments $\mathbf{z}_{1:t-1}$, and (ii) the "cohesion" term $G(\mathbf{x}_{t-p:t-1}; \lambda_G, D_{tk})$. In Figure 1, edges between the $z_t$'s denote the CRP probabilities, while edges from $x_{t-1}$ up to $z_t$ represent reweighting the CRP by $G$. Cluster assignment $z_t$ identifies the temporal regime that dictates the distribution of $x_t \sim F(\cdot|\theta_{z_t})$. Observe that if $p = 0$ or $G \propto 1$, then the model reduces to a standard CRP mixture (1) with no temporal dependence, since $(z_t, x_t)$ are conditionally independent of the entire time series history $\mathbf{x}_{1:t-1}$ given $\mathbf{z}_{1:t-1}$. Also note that the model is not Markovian, due to the infinite coupling among the latent $z_t$ (compare to the recurrent switching linear dynamical system of [4]).

The data distribution $F$ in (2) is a Normal distribution with Normal-InverseGamma prior $\pi_\Theta$:

$$\pi_\Theta(\mu_k, \sigma_k^2 \mid m, V, a, b) = \mathrm{N}(\mu_k | m, \sigma_k^2 V)\mathrm{IG}(\sigma_k^2 | a, b)$$
$$F(x_t|\mu_k, \sigma_k) = \mathrm{N}(x_t \mid \mu_k, \sigma_k^2), \tag{3}$$

where $\theta_k = (\mu_k, \sigma_k^2)$ are the per-cluster parameters of $F$, and $\lambda_F = (m, V, a, b)$ the hyperparameters of $\pi_\Theta$. Conjugacy of $F$ and $\pi_\Theta$ [5] implies that $\theta_k$ can be marginalized out of the generative model (2) (see Appendix B). As for $G$, it may in general be any non-negative weighting function which assigns a high value to lagged data vectors that are "similar" to one another. Previous approaches Bayesian nonparametric regression constructed covariate-dependent probability measures using kernel-based reweighting [10]. Our method defines $G$ as a product of $p$ Student-T distributions whose location, scale, and degrees of freedom depend on lagged data $D_{tk}$ in cluster $k$:

$$\begin{aligned} G(\mathbf{x}_{t-p:t-1}; D_{tk}, \lambda_G) &= \prod_{i=1}^{p} G_i(x_{t-i}; D_{tki}, \lambda_{Gi}) \\ &= \prod_{i=1}^{p} \mathrm{T}_{2a_{tki}}\left(x_{t-i}; m_{tki}, b_{tki}\frac{1+V_{tki}}{a_{tki}}\right) \end{aligned} \tag{4}$$

where hyperparameter $\lambda_{Gi} = (m_{i0}, V_{i0}, a_{i0}, V_{i0})$ and data $D_{tki} = \{x_{t'-i} : z_{t'} = k, 1 \le t' < t\}$. Equations for the data-dependent terms $(m_{tki}, V_{tki}, a_{tki}, b_{tki})$ are given in Appendix A. We emphasize that $G$ itself is

**1**. Sample concentration parameter of CRP

$$\alpha \sim \text{Gamma}(1,1)$$

**2**. Sample model hyperparameters $\quad (n = 1, 2, \ldots, N)$

$$\lambda_G^n \sim H_G^n$$
$$\lambda_F^n \sim H_F^n$$

**3**. Sample distribution parameters of $F \quad (n = 1, 2, \ldots, N)$

$$\theta_1^n, \theta_2^n, \ldots \stackrel{\text{iid}}{\sim} \pi_\Theta(\cdot|\lambda_F^n)$$

**4**. Assume first $p$ values are known $\quad (n = 1, 2, \ldots, N)$

$$\mathbf{x}_{-p+1:0}^n := (x_{-p+1}^n, \ldots, x_0^n)$$

**5**. Sample time series observations $\quad (t = 1, 2, \ldots)$

**5.1** Sample temporal cluster assignment $z_t$

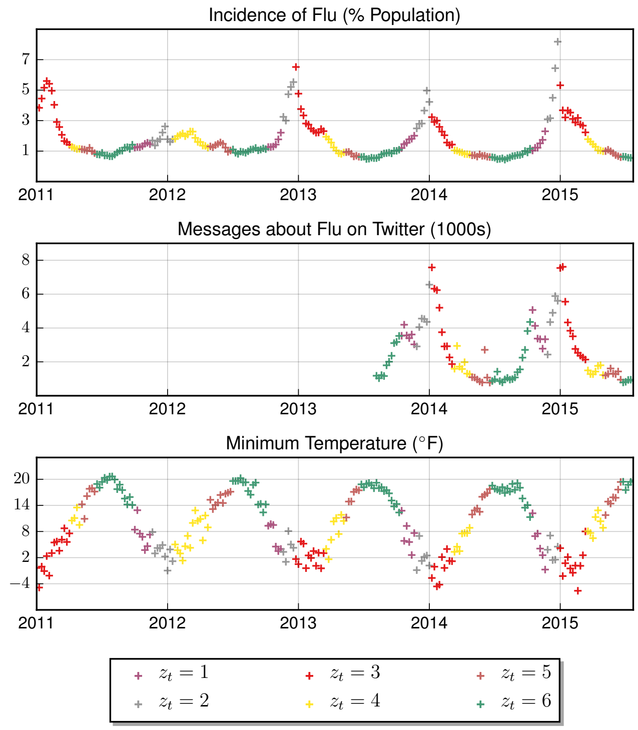
$$\Pr\left[z_t = k \mid \mathbf{z}_{1:t-1}, \mathbf{x}_{t-p:t-1}^{1:N}, \alpha, \lambda_G^{1:N}\right]$$
$$\propto \text{CRP}(k|\alpha, \mathbf{z}_{1:t-1}) \prod_{n=1}^N G(\mathbf{x}_{t-p:t-1}^n; D_{tk}^n, \lambda_G^n)$$

where $D_{tk}^n := \left\{\mathbf{x}_{t'-p:t'-1}^n \mid z_{t'} = k, 1 \le t' < t\right\}$
and $k = 1, \ldots, \max(\mathbf{z}_{1:t-1}) + 1$

**5.2** Sample data $x_t^n \quad (n = 1, 2, \ldots, N)$

$$x_t^n \mid z_t, \{\theta_k^n\} \sim F(\cdot|\theta_{z_t}^n)$$

**(a)** Generative process for the multivariate TRCRP mixture




**(b)** Discovering flu season dynamics with the method

**Figure 2:** (a) Generative model describing the joint distribution of $N$ dependent time series $\{\mathbf{x}^n\}$ in the multivariate temporally-reweighted CRP mixture. Lagged values for all time series are used for reweighting the CRP by $G$ in step 5.1. Dependencies between time series are mediated by the shared temporal regime assignment $z_t$, which ensures that all the time series have the same segmentation of the time course into the different temporal regimes. (b) Applying the TRCRP mixture with $p = 10$ weeks to model $\mathbf{x}^{\text{flu}}$, $\mathbf{x}^{\text{tweet}}$, and $\mathbf{x}^{\text{temp}}$ in US Region 4. Six regimes describing the seasonal behavior shared among the three time series are detected in this posterior sample. Purple, gray, and red are the pre-peak rise, peak, and post-peak decline during the flu season; and yellow, brown, and green represent the rebound in between successive seasons. In 2012, the model reports no red post-peak regime, reflecting the season's mild flu peak. See Section 5 for quantitative experiments.

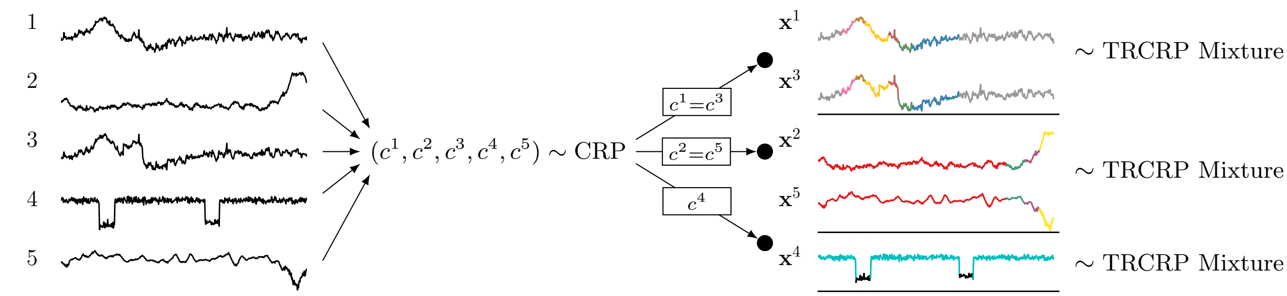
used for reweighting only; it does not define a probability distribution over lagged data. Mathematically, $G$ attracts $x_t$ towards a cluster $k$ that assigns $\mathbf{x}_{t-p:t-1}$ a high density, under the posterior predictive of an axis-aligned Gaussian having observed $D_{tk}$ [24].

### 3.3 Extending the TRCRP mixture to multiple dependent time series

This section generalizes the univariate TRCRP mixture (2) to handle a collection of $N$ time series $\{\mathbf{x}^n : n = 1, \ldots, N\}$, assumed for now to all be dependent. At time $t$, we let the temporal regime assignment $z_t$ be shared among all the time series, and use lagged values of all $N$ time series when reweighting the CRP probabilities by the cohesion term $G$. Figure 2a contains a step-by-step description of the multivariate TRCRP mixture, with an illustrative application in Figure 2b. It is informative to consider how $z_t$ mediates dependences between $\mathbf{x}^{1:N}$. First, the model requires all time series to be in the same regime $z_t$ at time $t$. However, each time series has its own set of per-cluster parameters $\{\theta_k^n\}$. Therefore, all the time series share the same segmentation $\mathbf{z}_{1:T}$ of the time course into various temporal regimes, even though the parametric distributions $F(\cdot|\theta_k^n), n = 1, \ldots, N$ within each temporal regime $k \in \mathbf{z}_{1:T}$ differ. Second, the model makes the "naive Bayes" assumption that data $\{x_t^n\}_{n=1}^N$ at time $t$ are independent given $z_t$, and that the reweighting term $G$ in step 5.1 factors as a product. This characteristic is essential for numerical stability of the method in high dimensional and sparse regimes, while still maintaining the ability to recover complex distributions due to the infinite CRP mixture.

### 3.4 Learning the dependence structure between multiple time series

The TRCRP mixture in Figure 2a makes the restrictive assumption that all time series $\mathbf{x}^{1:N}$ are dependent




(a) Original $N = 5$ time series (b) CRP over time series clusters (c) TRCRP mixture within each cluster

**Figure 3:** Hierarchical prior for learning the dependence structure between multiple time series. Given $N$ EEG time series, we first nonparametrically partition them by sampling an assignment vector $\mathbf{c}^{1:N}$ from an "outer" CRP. Time series assigned to the same cluster are jointly generated using the TRCRP mixture. Colored segments of each curve indicate the hidden states at each time step (the shared latent variables within the cluster).

with one another. However, with dozens or hundreds of time series whose temporal regimes are not well-aligned, forcing a single segmentation sequence $\mathbf{z}_{1:T}$ to apply to all $N$ time series will result in a poor fit to the data. We relax this assumption by introducing a hierarchical prior that allows the model to determine which subsets of the $N$ time series are probably well-described by a joint TRCRP model. The prior induces sparsity in the dependencies between the $N$ time series by first nonparametrically partitioning them using an "outer" CRP. Within a cluster, all time series are modeled jointly using the multivariate TRCRP mixture described in Figure 2a:

$$\begin{aligned}(c^1, c^2, \dots, c^N) &\sim \mathrm{CRP}(\cdot|\alpha_0) \\ \{\mathbf{x}^n : c^n = k\} &\sim \text{TRCRP Mixture} \\ &\qquad \left(k = 1, \dots, \max \mathbf{c}^{1:N}\right),\end{aligned} \tag{5}$$

where $c^n$ is the cluster assignment of $\mathbf{x}^n$. Figure 3 shows an example of this structure learning prior applied to five EEG time series. In the second cluster of panel (c), the final yellow segment illustrates two time series sharing the latent regime at each time step, but having different distributions within each regime.

## 4 Posterior Inferences via Markov Chain Monte Carlo

In this section, we give the full model likelihood and briefly describe MCMC algorithms for inference in the hierarchical TRCRP mixture (5). Since the model learns $M = \max(\mathbf{c}^{1:N})$ separate TRCRP mixtures (one for each time series cluster) we superscript latent variables of Figure 2a by $m = 1, \dots, M$. Namely, $\alpha^m$ is the CRP concentration, and $\mathbf{z}^m_{1:T}$ the latent regime vector, shared by all time series in cluster $m$. Further, let $K_m = \max(\mathbf{z}^m_{1:T})$ denote the number of unique regimes in $\mathbf{z}^m_{1:T}$. Given window size $p$ and initial observations $\left\{\mathbf{x}^n_{-p+1:0} : n = 1, \dots, N\right\}$, we have:

$$\begin{aligned}&P\Big(\alpha_0, \mathbf{c}^{1:N},\ \alpha^{1:M}, \lambda_G^{1:N}, \lambda_F^{1:N}, \{\theta^n_j : 1 \le j \le K_{c^n}\}_{n=1}^N, \\ &\qquad \mathbf{z}^{1:M}_{1:T}, \mathbf{x}^{1:N}_{1:T}\ ;\ \mathbf{x}^{1:N}_{-p+1:0}, p\Big) \\ &= \Gamma(\alpha_0; 1, 1)\mathrm{CRP}(\mathbf{c}^{1:N} \mid \alpha_0) \\ &\left(\prod_{n=1}^{N} H^n_G(\lambda^n_G)\right)\left(\prod_{n=1}^{N} H^n_F(\lambda^n_F)\right)\left(\prod_{n=1}^{N}\prod_{j=1}^{K_{c^n}} \pi^n_\Theta(\theta^n_j)\right) \\ &\prod_{m=1}^{M}\Bigg(\Gamma(\alpha^m; 1, 1)\prod_{t=1}^{T}\Big[b^m_t \mathrm{CRP}(z^m_t \mid \mathbf{z}^m_{1:t-1}, \alpha^m) \\ &\qquad \prod_{n|c_n=m} G(\mathbf{x}^n_{t-p:t-1}; D^n_{tz^m_t}, \lambda^n_G) F(x^n_t \mid \theta^n_{z^m_t})\Big]\Bigg),\end{aligned} \tag{6}$$

where $b^m_t$ normalizes the term between the square brackets, summed over $z'^m_t = 1, \dots, \max(\mathbf{z}^m_{1:t-1}) + 1$. Eq (6) defines the unnormalized posterior distribution of all latent variables given the data. Appendix B contains detailed algorithms for posterior inference. Briefly, temporal regime assignments $(z^m_t | \mathbf{z}^m_{1:T \setminus t} \cdots)$ are sampled using a variant of Algorithm 3 from [25], taking care to handle the temporal-coupling term $b^m_t$ which is not found in traditional DPM samplers. We also outline an alternative particle-learning scheme [9] to sample $(\mathbf{z}^m_{1:T} | \dots)$ jointly as a block. Time series cluster assignments $(c^n | \mathbf{c}^{1:N \setminus n}, \dots)$ are transitioned by proposing to move $\mathbf{x}^n$ to either an existing or a new cluster, and computing the appropriate MH acceptance ratio for each case. Model hyperparameters are sampled using an empirical Bayes approach [30] and the "griddy Gibbs" [29] sampler.

### 4.1 Making predictive inferences

Given a collection of approximate posterior samples $\left\{\hat{\xi}^1, \dots, \hat{\xi}^S\right\}$ of all latent variables produced by $S$ in-

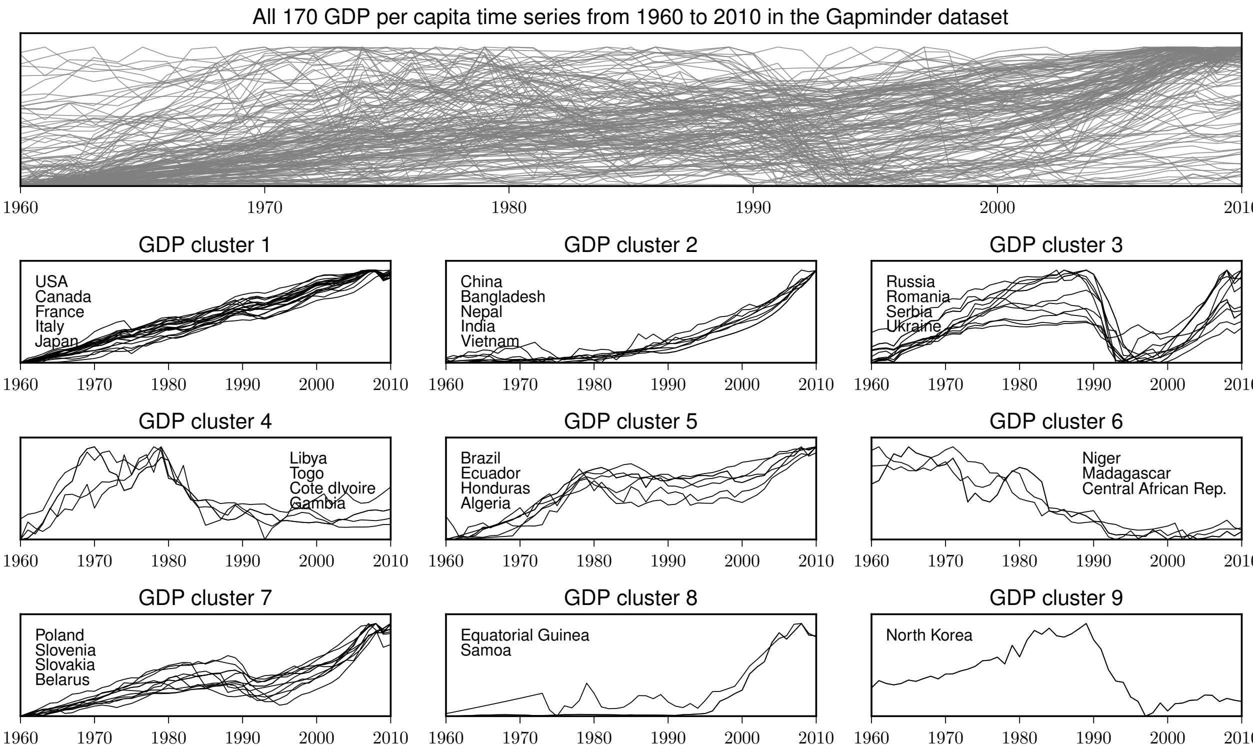



**Figure 4:** Given GDP per capita data for 170 countries from 1960-2010, the hierarchical TRCRP mixture (5) detects qualitatively distinct temporal patterns. The top panel shows an overlay of all the time series; nine representative clusters averaged over 60 posterior samples are shown below. Countries within each cluster, of which a subset are labeled, share similar political, economic, and/or geographic characteristics. For instance, cluster 1 contains Western democracies with stable economic growth over 50 years (slight dip in 2008 is the financial crash). Cluster 2 includes China and India, whose GDP growth rates have outpaced those of industrialized nations since the 1990s. Cluster 3 contains former communist nations, whose economies tanked after fall of the Soviet Union. Outliers such as Samoa, Equatorial Guinea, and North Korea can be seen in clusters 8 and 9.

dependent runs of MCMC, we can draw a variety of predictive inferences about the time series $\mathbf{x}^{1:N}$ which form the basis of the applications in Section 5.

**Forecasting** For out-of-sample time points, a forecast over an $h$ step horizon $T < t < T+h$ is generated by ancestral sampling: first draw a chain $\tilde{s} \sim$ Uniform$[1 \dots S]$, then simulate step 5 of Figure 2a using the latent variables in chain $\xi^{\tilde{s}}$ for $t = T, \dots, T+h$.

**Clustering** For a pair of time series $(\mathbf{x}^i, \mathbf{x}^k)$, the posterior probability that they are dependent is the fraction of samples in which they are in the same cluster:

$$\mathbb{P}\left[c^i = c^k \middle| \mathbf{x}^{1:N}\right] \approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left[\hat{c}^{i,s} = \hat{c}^{k,s}\right]. \qquad (7)$$

**Imputation** Posterior inference yields samples of each temporal regime $\hat{z}_t^{\cdot,s}$ for all in-sample time points $1 \le t \le T$; the posterior distribution of a missing value is:

$$\mathbb{P}\left[x_t^n \in B \middle| \mathbf{x}^{1:N} \setminus \{x_t^n\}\right] \approx \frac{1}{S}\sum_{s=1}^{S} F(B \mid \hat{\theta}^{n,s}_{\hat{z}_t^{\hat{c}^{n,s}}}). \qquad (8)$$

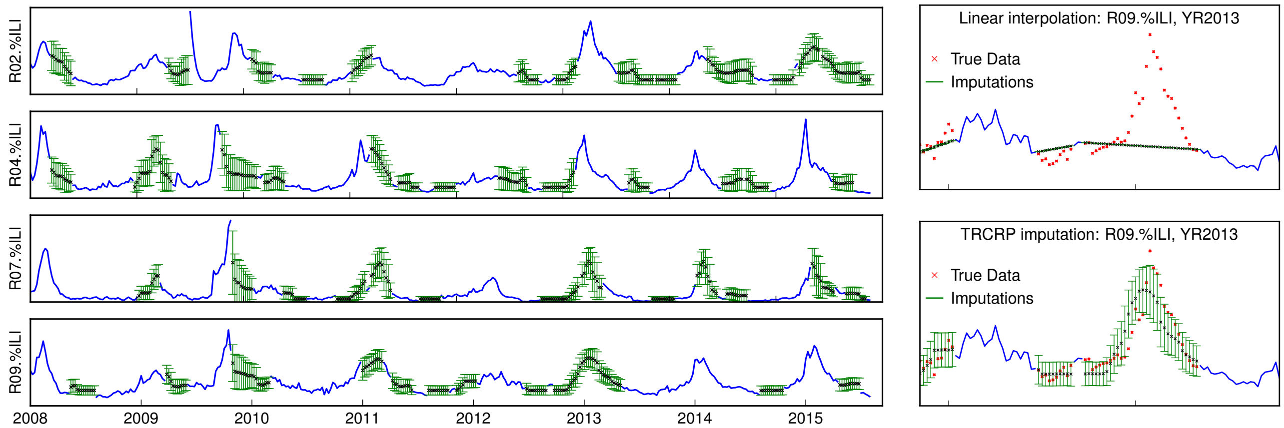
## 5 Applications

In this section, we apply the TRCRP mixture to clustering hundreds of time series using macroeconomic data from the Gapminder Foundation, as well as imputation and forecasting tasks on seasonal flu data from the US Center for Disease Control and Prevention (CDC). We describe the setup in the text below, with further commentary given in Figures 4, 5, and 6. Experimental methods are detailed in Appendix C[1].

We first applied the TRCRP mixture with hierarchical prior to cluster countries in the Gapminder dataset, which contains dozens of macroeconomic time series for 170 countries spanning 50 years. Because fluctuations due to events such as natural disasters, financial crises, or healthcare epidemics are poorly described by parametric or hand-designed causal models, a key objective is to automatically discover the number and kinds of patterns underlying the tempo-

[1]An implementation of the hierarchical TRCRP mixture is available at https://github.com/probcomp/trcrpm.

**(a)** Four representative flu time series imputed jointly **(b)** Example imputations in R09




**(c)** Mean absolute imputation errors in ten United States flu regions

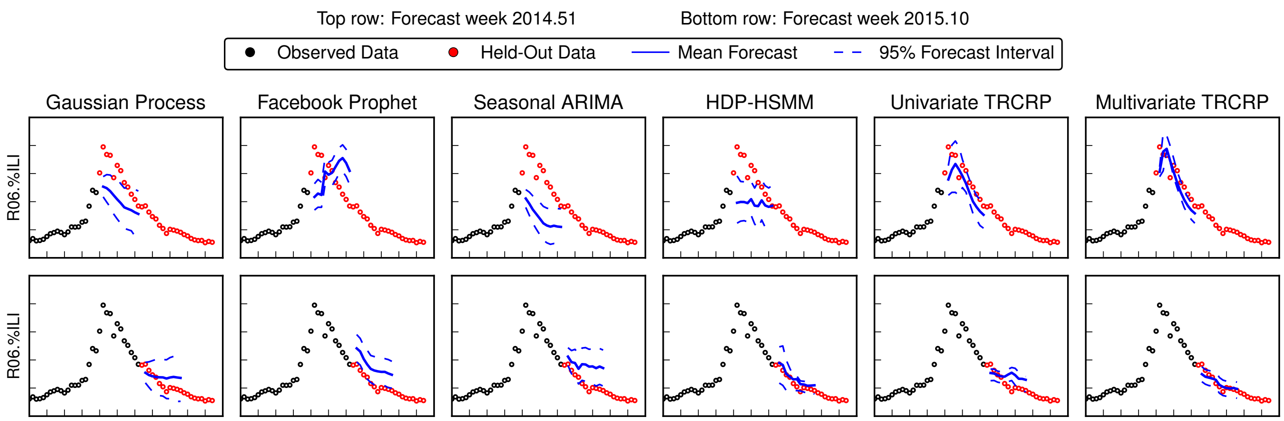
| | R01 | R02 | R03 | R04 | R05 | R06 | R07 | R08 | R09 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean Imputation | $0.65_{(0.04)}$ | $0.85_{(0.11)}$ | $0.91_{(0.04)}$ | $1.07_{(0.06)}$ | $0.66_{(0.04)}$ | $1.20_{(0.08)}$ | $1.17_{(0.10)}$ | $0.75_{(0.04)}$ | $0.80_{(0.05)}$ | $1.10_{(0.10)}$ |
| Linear Interpolation | $0.43_{(0.07)}$ | $0.63_{(0.08)}$ | $0.57_{(0.06)}$ | $\mathbf{0.42}_{(0.04)}$ | $0.44_{(0.04)}$ | $0.71_{(0.08)}$ | $0.71_{(0.09)}$ | $0.35_{(0.03)}$ | $0.43_{(0.05)}$ | $0.72_{(0.06)}$ |
| Cubic Splines | $1.01_{(0.15)}$ | $0.72_{(0.09)}$ | $0.61_{(0.06)}$ | $0.89_{(0.10)}$ | $0.69_{(0.06)}$ | $1.68_{(0.21)}$ | $1.42_{(0.22)}$ | $0.63_{(0.05)}$ | $0.99_{(0.22)}$ | $1.47_{(0.13)}$ |
| Multi-output GP | $0.36_{(0.04)}$ | $0.57_{(0.11)}$ | $0.32_{(0.02)}$ | $0.58_{(0.07)}$ | $0.30_{(0.03)}$ | $0.57_{(0.04)}$ | $0.62_{(0.04)}$ | $\mathbf{0.34}_{(0.03)}$ | $0.43_{(0.04)}$ | $0.56_{(0.04)}$ |
| Amelia II | $0.29_{(0.03)}$ | $0.52_{(0.11)}$ | $0.25_{(0.02)}$ | $0.45_{(0.03)}$ | $\mathbf{0.29}_{(0.03)}$ | $\mathbf{0.53}_{(0.04)}$ | $\mathbf{0.53}_{(0.05)}$ | $0.37_{(0.03)}$ | $0.39_{(0.04)}$ | $\mathbf{0.51}_{(0.03)}$ |
| TRCRP Mixture | $\mathbf{0.23}_{(0.03)}$ | $\mathbf{0.47}_{(0.09)}$ | $\mathbf{0.23}_{(0.02)}$ | $0.49_{(0.04)}$ | $0.31_{(0.03)}$ | $0.55_{(0.05)}$ | $0.75_{(0.07)}$ | $\mathbf{0.34}_{(0.03)}$ | $\mathbf{0.37}_{(0.03)}$ | $0.67_{(0.07)}$ |

**Figure 5:** Jointly imputing missing data in ten flu populations over eight seasons. (a) Imputations and standard errors in four of the time series. The TRCRP mixture accurately captures both seasonal behavior as well as non-recurrent characteristics, such as the very mild flu season in 2012. (c) Comparing imputation quality with several baseline methods. The TRCRP mixture ($p = 10$ weeks) achieves comparable performance to Amelia II. Cubic splines are completely ineffective due to long sequences without any observations. (b) While linear interpolation may seem to be a good performer given its simplicity and mean errors, unlike the TRCRP it cannot predict non-linear behavior when an entire flu season is unobserved and entirely misses seasonality.

ral structure. Figure 4 shows the outcome of structure discovery in GDP time series using the model with $p = 5$ years. Several common-sense, qualitatively distinct clusters are detected. Note that countries within each cluster share similar political, economic, and/or geographic characteristics; see caption for additional details. Appendix C.5 gives an expanded set of clusterings showing changepoint detection in cell phone subscription time series, and compares to a baseline using k-medoids clustering.

Predicting flu rates is a fundamental objective in public health policy. The CDC has an extensive dataset of flu rates and associated time series such as weather and vaccinations. Measurements are taken weekly from January 1998 to June 2015. Figure 2b shows the influenza-like-illness rate (ILI, or flu), tweets, and minimum temperature time series in US Region 4, as well as six temporal regimes detected by one posterior sample of the TRCRP mixture model ($p = 10$ weeks). We first investigated the performance of the proposed model on a multivariate imputation task. Windows of length 10 were dropped at a rate of 5% from flu series in US Regions 1-10. The top panel of Figure 5a shows flu time series for US Regions 2, 4, 7, and 9, as well joint imputations (and two standard deviations) obtained from the TRCRP mixture using (8). Quantitative comparisons of imputation accuracy to baselines are given in Table 5c. In this application, the TRCRP mixture achieves comparable accuracy to the widely used Amelia II [16] baseline, although neither method is uniformly more accurate. A sensitivity analysis showing imputation performance with varying $p$ is given in Appendix C.3.

To quantitatively investigate the forecasting abilities of the model, we next held out the 2015 season for 10 US regions and generated forecasts on a rolling basis. Namely, for each week $t = 2014.40, \dots, 2015.20$ we forecast $\mathbf{x}^{\text{flu}}_{t:t+h}$ given $\mathbf{x}^{\text{flu}}_{1:t-2}$ and all available covariate data up to time $t$, with horizon $h = 10$. A key challenge is that when forecasting $\mathbf{x}^{\text{flu}}_{t:t+h}$, the most recent flu measurement is two weeks old $x^{\text{flu}}_{t-2}$. Moreover, covariate time series are themselves sparsely observed in the training data (for instance, all Twitter data is missing before June 2013, top panel of Figure 2b).




Mean absolute flu prediction error for 10 forecast horizons (in weeks) averaged over 10 United States flu regions

| | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=7$ | $h=8$ | $h=9$ | $h=10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $^\dagger$Linear Extrapolation | $0.65_{(0.06)}$ | $0.79_{(0.05)}$ | $0.93_{(0.05)}$ | $1.08_{(0.05)}$ | $1.24_{(0.05)}$ | $1.39_{(0.05)}$ | $1.55_{(0.05)}$ | $1.70_{(0.05)}$ | $1.86_{(0.05)}$ | $2.01_{(0.05)}$ |
| $^\dagger$GP(SE+PER+WN) | $0.53_{(0.04)}$ | $0.60_{(0.03)}$ | $0.66_{(0.03)}$ | $0.71_{(0.03)}$ | $0.75_{(0.02)}$ | $0.79_{(0.02)}$ | $0.82_{(0.02)}$ | $0.85_{(0.02)}$ | $0.87_{(0.02)}$ | $0.89_{(0.02)}$ |
| $^\dagger$GP(SE$\times$PER+WN) | $0.50_{(0.04)}$ | $0.57_{(0.03)}$ | $0.62_{(0.03)}$ | $0.67_{(0.02)}$ | $0.71_{(0.02)}$ | $0.74_{(0.02)}$ | $0.78_{(0.02)}$ | $0.81_{(0.02)}$ | $0.84_{(0.02)}$ | $0.86_{(0.02)}$ |
| $^\dagger$Facebook Prophet | $0.83_{(0.04)}$ | $0.84_{(0.03)}$ | $0.85_{(0.02)}$ | $0.85_{(0.02)}$ | $0.85_{(0.02)}$ | $0.86_{(0.02)}$ | $0.86_{(0.02)}$ | $0.87_{(0.02)}$ | $0.87_{(0.01)}$ | $0.87_{(0.01)}$ |
| $^\dagger$Seasonal ARIMA | $0.64_{(0.04)}$ | $0.76_{(0.03)}$ | $0.84_{(0.03)}$ | $0.92_{(0.03)}$ | $0.98_{(0.03)}$ | $1.04_{(0.02)}$ | $1.08_{(0.02)}$ | $1.13_{(0.02)}$ | $1.16_{(0.02)}$ | $1.19_{(0.02)}$ |
| $^\dagger$TRCRP Mixture | $0.54_{(0.04)}$ | $0.58_{(0.03)}$ | $0.62_{(0.02)}$ | $0.67_{(0.02)}$ | $0.71_{(0.02)}$ | $0.76_{(0.02)}$ | $0.80_{(0.02)}$ | $0.83_{(0.02)}$ | $0.86_{(0.02)}$ | $0.89_{(0.02)}$ |
| $^\ddagger$HDP-HSMM | $0.69_{(0.05)}$ | $0.72_{(0.04)}$ | $0.76_{(0.03)}$ | $0.79_{(0.03)}$ | $0.82_{(0.02)}$ | $0.84_{(0.02)}$ | $0.86_{(0.02)}$ | $0.88_{(0.02)}$ | $0.89_{(0.02)}$ | $0.90_{(0.02)}$ |
| $^\star$Multi-output GP | $0.70_{(0.04)}$ | $0.77_{(0.03)}$ | $0.84_{(0.03)}$ | $0.88_{(0.03)}$ | $0.91_{(0.02)}$ | $0.93_{(0.02)}$ | $0.95_{(0.02)}$ | $0.97_{(0.02)}$ | $0.99_{(0.02)}$ | $1.01_{(0.02)}$ |
| $^\star$TRCRP Mixture | $\mathbf{0.46}_{(0.03)}$ | $\mathbf{0.49}_{(0.02)}$ | $\mathbf{0.51}_{(0.02)}$ | $\mathbf{0.53}_{(0.02)}$ | $\mathbf{0.56}_{(0.02)}$ | $\mathbf{0.58}_{(0.02)}$ | $\mathbf{0.59}_{(0.01)}$ | $\mathbf{0.61}_{(0.01)}$ | $\mathbf{0.62}_{(0.01)}$ | $\mathbf{0.64}_{(0.01)}$ |

Modeled time series: $^\dagger$flu $^\ddagger$flu+weather $^\star$flu+weather+tweets

**Figure 6:** Quantitative evaluation of forecasting performance on the 2015 flu season. The table shows mean prediction errors and (one standard error) of the flu rate, for various forecast horizons averaged over US Regions 1–10. Available covariate time series include minimum temperature and Twitter messages about the flu (not shown, see Figure 2b). Predictive improvement of the multivariate TRCRP mixture over baselines is especially apparent at longer horizons. The top two panels show sample forecasts in US Region 6 for week 2014.51 (pre-peak) and week 2015.10 (post-peak). The TRCRP mixture accurately forecasts seasonal dynamics in both cases, whereas baseline methods produce inaccurate forecasts and/or miscalibrated uncertainties.

Figure 6 shows the forecasting accuracy from several widely-used, domain-general baselines that do not require detailed custom modeling for obtaining forecasts, and that have varying ability to make use of covariate data (weather and tweet signals). The TRCRP mixture consistently produces the most accurate forecasts for all horizons (last row). Methods such as seasonal ARIMA [17] can handle covariate data in principle, but cannot handle missing covariates in the training set or over the course of the forecast horizon. Both Facebook Prophet [36] and ARIMA incorrectly forecast the peak behavior (Figure 6, top row), and are biased in the post-peak regime (bottom row). The HDP-HSMM [19] also accounts for weather data, but fails to detect flu peaks. The univariate TRCRP (only modeling the flu) performs similarly to periodic Gaussian processes, although the latter gives wider posterior error bars, even in the relatively noiseless post-peak regime. The multi-output GP [3] uses both weather and tweet covariates, but they do not result in an improvement in predictive accuracy over univariate methods.

## 6 Discussion

This paper has presented the temporally-reweighted CRP mixture, a domain-general nonparametric Bayesian method for multivariate time series. Experiments show strong quantitative and qualitative results on multiple real-world multivariate data analysis tasks, using little to no custom modeling. For certain application domains, however, predictive performance may improve by extending the model to include custom knowledge such as time-varying functionals. Further avenues for research include guidelines for selecting the window size; greater empirical validation; a stick breaking representation; improving inference scalability; and establishing theoretical conditions for posterior consistency. Also, it could be fruitful to integrate this method into a probabilistic programming platform [33], such as BayesDB. This integration would make it easy to query mutual information between time series [32], identify data that is unlikely under the model, and make the method accessible to a broader audience.

## Acknowledgments

This research was supported by DARPA PPAML program, contract number FA8750-14-2-0004. The authors wish to thank Max Orhai from Galois, Inc. for assembling the CDC flu dataset.

## References

[1] A. Ahmed and E. Xing. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 219–230. SIAM, 2008.

[2] D. J. Aldous. Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer, 1985.

[3] M. Alvarez and N. D. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems 21*, pages 57–64. Curran Associates, Inc., 2009.

[4] D. Barber. Expectation correction for smoothed inference in switching linear dynamical systems. *Journal of Machine Learning Research*, 7(Nov): 2515–2540, 2006.

[5] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley Series in Probability & Statistics. Wiley, 1994. ISBN 9780471924166.

[6] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Workshop on Knowledge Discovery in Databases*, AAAIWS-94, pages 359–370. AAAI Press, 1994.

[7] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning*, pages 113–120. ACM, 2006.

[8] F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 56(1): 71–84, 2008.

[9] C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson. Particle learning and smoothing. *Statistical Science*, 25(1):88–106, 02 2010.

[10] D. B. Dunson, N. Pillai, and J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 163–183, 2007.

[11] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430): 577–588, 1995.

[12] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems 21*, pages 457–464. Curran Associates, Inc., 2009.

[13] B. D. Fulcher and N. S. Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, 2014.

[14] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[15] L. F. Gruber and M. West. Bayesian online variable selection and scalable multivariate volatility forecasting in simultaneous graphical dynamic linear models. *Econometrics and Statistics*, 3(C): 3–22, 2017.

[16] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[17] R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008. ISSN 1548-7660.

[18] H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4): 1211–1235, 2003.

[19] M. J. Johnson and A. S. Willsky. Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, 14(Feb):673–701, 2013.

[20] G. M. Koop. Forecasting with medium and large Bayesian VARS. *Journal of Applied Econometrics*, 28(2):177–203, 2013.

[21] V. Mansinghka, P. Shafto, E. Jonas, C. Petschulat, M. Gasner, and J. B. Tenenbaum. CrossCat: A fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *Journal of Machine Learning Research*, 17(138): 1–49, 2016.

[22] P. Mueller and F. Quintana. Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808, 2010.

[23] P. Mueller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

[24] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia, 2007.

[25] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2): 249–265, 2000.

[26] L. E. Nieto-Barajas and A. Contreras-Cristan. A Bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9(1):147–170, 03 2014.

[27] L. E. Nieto-Barajas and F. A. Quintana. A Bayesian non-parametric dynamic AR model for multiple time series analysis. *Journal of Time Series Analysis*, 37(5):675–689, 2016.

[28] J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20(3):1203–1226, 2010.

[29] C. Ritter and M. Tanner. The griddy Gibbs sampler. Technical Report 878, University of Wisconsin-Madison, 1991.

[30] H. Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1–20, 1964.

[31] A. Rodriguez and E. ter Horst. Bayesian dynamic density estimation. *Bayesian Analysis*, 3(2):339–365, 6 2008.

[32] F. Saad and V. Mansinghka. Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric Bayes. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 632–641. PMLR, 2017.

[33] F. Saad and V. K. Mansinghka. A probabilistic programming approach to probabilistic data analysis. In *Advances in Neural Information Processing Systems 29*, pages 2011–2019. Curran Associates, Inc., 2016.

[34] U. Schaechtle, B. Zinberg, A. Radul, K. Stathis, and V. Mansinghka. Probabilistic programming with Gaussian process memoization. *arXiv preprint*, arXiv:1512.05665, 2015.

[35] B. Shahbaba and R. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850, 2009.

[36] S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, To appear, 2017.

[37] Y. W. Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2011.

[38] Z. G. Zhu, Xiaojin and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical Report CMU-CALD-05-104, Carnegie-Mellon University, School of Computer Science, 2005.