Non-parametric Bayesian Modeling and Fusion of Spatio-temporal Information Sources

Priyadip Ray
Department of Electrical and Computer Engineering
Duke University
Durham, U.S.A.
Email: priyadip.ray@duke.edu

Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, U.S.A
Email: lcarin@duke.edu

Abstract—We propose a Gaussian process (GP) factor analysis approach for modeling multiple spatio-temporal datasets with non-stationary spatial covariance structure. A novel kernel stick-breaking process based mixture of GPs is proposed to address the problem of non-stationary covariance structure. We also propose a joint GP factor analysis approach for simultaneous modeling of multiple heterogeneous spatio-temporal datasets. The performance of the proposed models are demonstrated on the analysis of multi-year unemployment rates of various metropolitan cities in the United States and counties in Michigan.

Keywords: Gaussian process, infinite mixture model, Dirichlet Process, non-parametric Bayesian analysis, information fusion

I. INTRODUCTION

The modeling and analysis of multiple time series exhibiting spatio-temporal dependencies is a particularly important research problem. Typical application areas include the analysis of radar reflectivities for rainfall forecasting, modeling of surface sea temperatures for climate variability research, modeling of air-quality monitoring data and the analysis of economic/financial data. Typical spatio-temporal datasets are non-stationary in nature and have missing values. Popular multivariate time series analysis techniques such as spectral methods, wavelet analysis and Box-Jenkins models require stationary time series without missing data, and are not particularly suitable for inferring inter-relationships across data [1]. Hence, recent advances in computing technologies have encouraged the application of more computationally intensive techniques, such as factor analysis for characterizing multivariate datasets. Factor analysis is a linear technique capable of inferring latent inter-relationships across multiple non-stationary data streams with missing values.

For spatio-temporal datasets, the spatial locations as well as the time stamps of the observations are often accurately known. Taking into account this information during data modeling typically leads to more accurate models. In this context, Gaussian process (GP) priors [2] provide a particularly effective and theoretically elegant solution to incorporating knowledge of the spatial locations and the time stamps in spatio-temporal models. In [3], the authors propose a Gaussian process factor analysis (GPF A) technique for modeling spatio-temporal datasets, where the spatial dependence is incorporated in the factor loadings and the temporal dependence is incorporated in the factor scores. The form of the model may be represented as,

\[ X = DW + E \]  

where, \( X \) represents the spatio-temporal dataset, \( D \) represents the factor loading matrix, \( W \) represents the factor score matrix, and \( E \) represents noise and/or model residual. Similar approaches have also been proposed in [4] and [5]. However, these approaches are primarily parametric in nature, i.e., the number of factor loadings is assumed to be known a priori.

In this paper, for modeling spatio-temporal datasets, we adopt a GPFA approach similar to [3] but our model is completely non-parametric in nature, i.e., the number of factor loadings as well as the noise variance is inferred from the available data. This is the first intended contribution of the paper.

In the GPFA model considered in [3], the factor loadings capture the spatial correlation and each factor loading is drawn from a single Gaussian process. But the underlying assumption, that the correlation pattern is spatially invariant, may not be a valid for many complex real-life datasets. For example, for the unemployment rates in various cities across United States, it is likely that the spatial correlation pattern among the unemployment rates for cities in the northeast will be very different from the spatial correlation pattern for cities in the midwest. Also, for large number of spatial locations drawn from a single GP, the computational complexity will be very high, since for GPs the computational complexity of inference scales cubically with respect to the number of data points. To model datasets with non-stationary spatial covariance as well as to increase the computational efficiency for large datasets, we propose a new GP factor model, where the factor loadings are drawn from a mixture of GPs. Though not specifically in the context of factor models, mixture of GPs have been applied previously to model data with non-stationary covariance structure. Most mixture of GP approaches, such as [6], assume a known number of mixture components. Non-parametric approaches, where the number of mixture components is inferred from the data, have been proposed in [7] and [8]. In [8], the authors propose a tree-based partitioning approach to divide the input space and fit different base models to data independently in each region.
[7], the authors propose an input dependent Dirichlet process mixture of GPs to model non-stationary covariance functions. In this paper, we propose a novel GP mixture model based on the kernel stick breaking process (KSBP) [9]. Our approach is similar in spirit to [7], with the added advantage that the kernel parameters related to the stick-breaking process may be inferred efficiently via Gibbs sampling. We also demonstrate that our model performs better than [7] in inferring the number of GP mixture components on the classic motorcycle dataset [10]. Both the KSBP based GP mixture model as well as the integration of the KSBP based GP mixture model in the GPFA is the second intended contribution of the paper.

To demonstrate the performance of the proposed GP factor model, we analyze the time series of unemployment rates at various geographical locations in United States. We have two datasets: one dataset contains time series of unemployment rates at 83 counties in the state of Michigan and the other dataset contains time series of unemployment rates at 187 metro cities across the entire United States. We first apply our proposed GPFA technique to model these datasets separately. However, it is likely that the two datasets share some underlying common factors and by appropriately borrowing statistical strength between the datasets we may be able to obtain a better statistical representation of each dataset. Hence we propose a joint GP factor analysis technique to model heterogenous sources of information. The proposed model may also be viewed as an approach to fuse information from multiple information sources with the objective of inferring a better statistical model for each source. To the best of the authors’ knowledge, the proposed approach to combine multiple space-time information sources has not been attempted before and is the third intended contribution of the paper.

The remainder of the paper is organized as follows. In Section II we present the proposed GP factor analysis approach for multiple spatio-temporal datasets, and in Section III we outline an MCMC inference algorithm. In Section IV we provide experimental results on the unemployment dataset. Finally, we provide concluding remarks in Section V.

II. GAUSSIAN PROCESS FACTOR ANALYSIS

Let \( X \in \mathbb{R}^{N \times M} \) represent a spatio-temporal dataset, where the number of spatial locations is denoted by \( N \) and the number of time samples is denoted by \( M \). The Gaussian process (GP) factor model may be represented as,

\[
x_i = D w_i + e_i
\]

For the unemployment datasets analyzed in this paper, \( x_i \in \mathbb{R}^N \), represents the unemployment rates at \( N \) spatial locations at the \( i^{th} \) time sample. Note, that unemployment rates lie in the interval \((0, 1)\) and so for ease of modeling, we employ a probit link function to map the the components of \( x_i \) to a number between \((0, 1)\). Matrix \( D \in \mathbb{R}^{N \times K} \) represents factor-loadings, \( w_i \in \mathbb{R}^K \) represents the factor scores, and \( e_i \in \mathbb{R}^N \) represents the indiosynchratic noise/residue.

Though we analyze unemployment rates in this paper, the above model is completely general and may be applied to model any spatio-temporal dataset. However, to better integrate our prior knowledge about the spatial covariance pattern, different priors may be imposed on the factor loadings \( \{d_i\}_{i=1}^{K} \) for different datasets. We discuss choice of priors for \( D \) for the specific datasets considered in this paper in the next section.

We wish to impose the condition that any \( x_i \) is a sparse linear combination of the factor loadings \( \{d_i\}_{i=1}^{K} \). Hence, the factor scores \( w_i \) are represented as

\[
w_i = s_i \oplus b_i
\]

where \( s_i \in \mathbb{R}^K \), \( b_i \in \{0,1\}^K \) and \( \oplus \) represents the Hadamard vector product. The sparse binary vectors \( b_i \) are drawn from the following beta-Bernoulli process,

\[
b_i \sim \prod_{k=1}^{K} Bernoulli(\pi_k) \tag{4}
\]

\[
\pi \sim \prod_{k=1}^{K} Beta(a_1/K, b_1(K-1)/K) \tag{5}
\]

with \( \pi_k \) representing the \( k^{th} \) component of \( \pi \). The Bernoulli prior imposes \( b_{i,k} = 1 \) with probability \( \pi_k \) and \( b_{i,k} = 0 \) with probability \( 1 - \pi_k \). The beta distribution is a prior on a continuous real random number between \((0,1)\), and is represented as \( Beta(\pi; a, b) = c \pi^{a-1} (1 - \pi)^{b-1} \), where \( a > 0, b > 0 \) and \( c = \Gamma(a+b)/\Gamma(a)\Gamma(b) \). In the limit \( K \rightarrow \infty \) this construction reduces to a generalization of the Indian buffet process [11], [12]. In practice we truncate \( K \), and the number of non-zero components of each \( b_i \) is a random number drawn from \( Binomial\left( K, \frac{aK}{a+b(K-1)} \right) \), and in the limit \( K \rightarrow \infty \) this reduces to \( Poisson(\frac{a}{a+b}) \). If the truncation level \( K \) is set to a relatively large value, the beta process allows us to infer the number of factors based on the available data.

The rows of \( S \), denoted by \( s_{(k)} \), are drawn from a Gaussian process. Note that \( s_{(k)} \) denotes the \( k^{th} \) row of \( S \), whereas \( s_i \) denotes the \( i^{th} \) column of \( S \). Hence we have,

\[
s_{(k)} \sim \mathcal{N}(0, \mathbf{P}_k) \tag{6}
\]

\[
\mathbf{P}_k(i,j) = \gamma_k^{(t)} \exp \left( \frac{-\beta_k^{(t)} \| t_i - t_j \|^2}{2} + \sigma_k^{(t)} \delta_{i,j} \right) \tag{7}
\]

Here, \( t_i \) and \( t_j \) represent the time stamps for the \( i^{th} \) and \( j^{th} \) time samples. Equation (6) imposes the belief that the unemployment rates are likely to be correlated in time and the correlation decays with increasing time difference. The kernel given by (7) is widely employed for Gaussian process priors and is characterized by three parameters: \( \gamma_k^{(t)} \) controls the signal variance, \( \sigma_k^{(t)} \) controls the noise variance and \( \beta_k^{(t)} \) is the bandwidth or scale parameter and controls the amount of smoothing. Other popular kernel choices may be found in [2] and references therein.

We do not assume any knowledge of the noise/residual variance. Instead, a prior is placed on this variable and it is
inferred from the available data. Also non-informative gamma hyperprior is placed on the precision parameter $\gamma_c$:

$$
\epsilon_i \sim \mathcal{N}(0, \gamma_c^{-1}I_N)
$$

$$
\gamma_c \sim \Gamma(a_0, b_0)
$$

with $a_0 = b_0 = 10^{-3}$.

A. Priors for factor loadings

As discussed earlier, to best incorporate our a priori knowledge of the spatial covariance structure, priors for the factor loadings are chosen specific to the dataset analyzed. For modeling the unemployment rates across the 83 counties of Michigan, it is assumed that the spatial covariance structure remains the same across the entire state. Hence the factor loadings are drawn from a GP as shown below,

$$
d_k \sim \mathcal{N}(0, C_k) \quad k = 1, \ldots, K
$$

$$
C_k(i, j) = \tau_k(s) \exp \left\{ -\frac{\|\mathbf{r}_i - \mathbf{r}_j\|^2}{2} \right\} + \sigma_k(s) \delta_{i,j}
$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ represent the spatial locations of the counties $i$ and $j$ respectively. Equation (11) imposes the belief that the unemployment rates of neighboring counties are likely to be correlated and the correlation decays with increasing distance between the counties.

However, it is highly unlikely that the spatial covariance structure will remain the same across the entire United States. For example, the density of metro cities in the northeast is very high compared to that in the midwest and it is likely that the scale parameters of the GP required to model cities in these two regions will be very different. Ideally we wish to divide similar cities into clusters with the cities in each cluster sharing the same spatial covariance structure, priors for the factor loadings will have appreciable weight. The parameter $\gamma$ determines the number of clusters that are constituted and in practise one typically places a non-informative Gamma prior on $\gamma$.

Based on the stick-breaking process prior, a possible solution for a mixture of GP approach will be to draw the observations of the cities associated with a particular “stick” from a single GP with unique kernel parameters $(\tau_k(s), \sigma_k(s))$. However, a fundamental weakness of this approach is that though the model imposes our prior belief that the cities cluster, it does not impose our additional belief that cities that are geographically proximate are likely to belong to the same cluster. Hence, we propose a mixture of GP approach based on the kernel stick-breaking process (KSBP) [9]. We next present our model followed by an intuitive explanation of how the model imposes the condition that cities that are spatially proximate are likely to belong to the same cluster. The factor loadings are now drawn from a mixture of GPs as described below,

$$
d_{z(i),l,k} \sim \mathcal{N}(0, C_l) \quad k = 1, \ldots, K, \quad i = 1, \ldots, N
$$

Here, $d_{z(i),l,k}$ represents the $l^{th}$ element of $d_k$. We draw each factor loading $\{d_k\}_{k=1,K}$ from a separate KSBP mixture of GPs. Hereafter we drop the index $k$, and the covariance associated with sample $i$ is

$$
C_{z(i)}(n, m) = \theta_{z(i)} \exp \left\{ -\frac{\|\mathbf{r}_n - \mathbf{r}_m\|^2}{2} \right\} + \sigma_{z(i)} \delta_{n,m}
$$

$$
z(i) \sim \sum_{j=1}^{J} w_j \delta_j \quad i = 1, \ldots, N
$$

$$
w_j(r) = V_j K(r; r_i^*, \phi_j) \prod_{l=1}^{j-1} [1 - V_l K(r; r_i^*, \phi_l)]
$$

$$
K(r; r_i^*, \phi_j) = \exp \left( -\frac{\|r - r_i^*\|^2}{\phi_j} \right)
$$

$$
\gamma \sim \text{Gamma}(a_2, b_2)
$$

$$
\phi_j \sim H
$$

As seen from (13) and (18), the primary difference between a stick-breaking prior and a kernel stick-breaking prior is that the stick weights are modulated by an additional bounded kernel $K(r; r_i^*, \phi_l) \rightarrow [0, 1]$ which is a function of the spatial location $r$. Thus if two cities are spatially proximate, they will have similar stick weights $w_j(r)$ and hence are likely to share

1980
the same GP.

B. Joint Gaussian Process Factor Model

In the previous section we presented a GPFA technique for modeling a single spatio-temporal dataset. However, we often have access to multiple heterogenous spatio-temporal datasets that are statistically correlated. In such cases learning all the models jointly, thereby appropriately leveraging all available data as well as inferring important similarities between these data, we may be able to obtain better statistical models for each dataset. In this paper, we propose a joint Gaussian process factor approach for jointly modeling the unemployment rates across various counties in Michigan and the unemployment rates across various metropolitan cities in United States. The model may be represented as,

\[
X^{(M)} = D^{(M)}(S_c \oplus B_c + S^{(M)} \oplus B^{(M)}) + e^{(M)}
\]  

(24)

\[
X^{(U)} = D^{(U)}(S_c \oplus B_c + S^{(U)} \oplus B^{(U)}) + e^{(U)}
\]  

(25)

Here, \(X^{(M)}\) represents the unemployment rates across Michigan and \(X^{(U)}\) represents the unemployment rates across metro cities in US. To account for the heterogeneity of the two datasets we allow unique dictionaries as well as unique model noise/residue. However, the dimensionality of the latent space is shared across the two datasets and it is also assumed that each dataset will have some unique latent features, denoted by \((S^{(M)} \oplus B^{(M)})\) and \((S^{(U)} \oplus B^{(U)})\) and some shared latent features denoted by \((S_c \oplus B_c)\). As demonstrated in Section IV, jointly learning these shared latent features can provide a better statistical representation of each dataset.

III. MCMC INFERENCEx

Due to conjugacy, the full conditional posterior distribution of all the model parameters, except the parameters related to the Gaussian processes \((\tau, \beta, \sigma)\), may be derived analytically. Hence, we use a Gibbs sampler to draw samples from the posterior distribution of the model parameters. For the GP parameters, the full conditional posteriors are not available in closed form. We obtain point estimates for these parameters via maximum likelihood estimation.

IV. EXPERIMENTAL RESULTS

A. Motorcycle dataset

We first demonstrate the proposed KSBP-GP mixture model on the classic motorcycle data [10], which has been frequently used in recent literature to demonstrate the success of non-stationary models [7]. This dataset is clearly non-stationary and has input-dependent noise. The dataset is comprised of measurements of the acceleration of the head of a motorcycle rider in the first moments after impact. As shown in Figure 1, the dataset shows three distinct regions with different covariance structures. In Figure 2 we show a typical inferred clustering result obtained from the KSBP-GP mixture model. Note that for this dataset we do not perform a GPFA and instead of the factor loadings, the data is directly modeled as a KSBP-GP mixture. In Figure 3 we show the frequency of the number of inferred clusters computed over the Gibbs collection samples. We observe that the KSBP-GP mixture model infers two to six clusters with the highest probability mass on 3 clusters. This is significantly better than [7], which reported that the posterior distribution uses between 3 and 10 experts to fit the dataset, with even 10-15 experts still having considerable posterior mass.

B. Unemployment dataset across United States

In this section we provide results for the proposed KSBP-GP factor model on the unemployment dataset across United States. The dataset contains annual unemployment rates of 187 metro cities (see Figure 4) across United States over the period 1991 to 2009. As shown in Figure 5, the KSBP-GP factor model infers 6 dominant factor loadings. Since each factor loading is drawn from a separate KSBP-GP mixture, typical clusters associated with three such factor loadings are shown in Figure 6. Note that in our model the unemployment rates are a sparse linear combination of the factor loadings, which are drawn from separate KSBP mixtures of Gaussian processes. Hence, our model is capable of modeling datasets
for which the spatial clustering pattern may change over time. This aspect was validated on simulated toy data which is not included in this paper for brevity.

C. Joint modeling of unemployment rates across Michigan and United States

We next provide results for the joint GP factor model on the analysis of unemployment rates across counties of Michigan and cities of United States. As discussed in Section II-B, the key idea is to jointly learn the shared latent features so as obtain a better statistical representation of each data source than could be obtained by modeling the datasets separately. A possible application is to model undersampled datasets with the aid of any available corelated datasets. This may happen when data is missing or when sampling is very expensive.

We perform the following experiment to demonstrate our proposed joint GPFA model. We assume that we have very coarsely sampled unemployment data across Michigan. We simulate this by randomly selecting only 10% and 20% of the counties across Michigan (see Figure 7). We also assume we have access to coarsely sampled (15% of the total cities) unemployment rates across United States. We show interpolation results at the remaining counties in Michigan for a GPFA model (using only coarsely sampled Michigan data)
and a joint GPFA model (using both Michigan and US data) in Table I and Figure 8. We also show the inferred factor loadings for Michigan for the GPFA model as well as the joint GPFA model in Figures 9 and 10. The RMSE results are obtained by averaging across 100 Gibbs collection samples. The simulation results demonstrate that jointly learning the Michigan data model, specially when the dataset is severely undersampled, can provide significant improvement in terms of RMSE of interpolation.

Figure 7. 20% of the counties (shown on the left) are used to learn the model and the remaining 80% of the counties are interpolated based on the learnt model.

Table I

<table>
<thead>
<tr>
<th></th>
<th>Separate</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>90% data missing</td>
<td>0.2596</td>
<td>0.1735</td>
</tr>
<tr>
<td>80% data missing</td>
<td>0.2</td>
<td>0.1652</td>
</tr>
</tbody>
</table>

Figure 8. Typical interpolation result for GPFA and joint GPFA models for 20% missing data

V. CONCLUSIONS

We propose a non-parametric Gaussian process factor analysis technique for modeling spatio-temporal datasets. A new kernel stick-breaking process based GP mixture model is introduced to handle datasets with spatially varying covariance structure. The performance of the proposed framework is demonstrated on the analysis of unemployment rates sampled at multiple geographical locations. We also propose a joint Gaussian process factor approach for simultaneous modeling of multiple heterogenous datasets. The joint approach is shown to be particularly effective in improving the model learning of undersampled data with the aid of other available correlated datasets.

REFERENCES


