A Dimensionality Reduction Model for Complex Data Grouping

Xiang Yang¹,*, Chen Xiaojun¹ and Luo cheng²

¹Department of Computer and Science, College of Electronic and Information Engineering, Tongji University, 4800 Cao'an Highway, Shanghai 201804, P.R. China
²Ford AP Shanghai office, Information Building, Pudong New Area, Shanghai, 200120, P.R. China
*txjdxiangyang@gmail.com

Abstract. Given the existing packet dimensionality reduction model, the simple distance hypothesis is only used as a simple assumption that there is a certain relationship between packet data. This document proposes to share information between packet data with relevant random measures as a priori. We explicitly calculated the Lévy measure of the mixed random measure and offered the inference steps of detailed parameter a posteriori. Compared with the traditional method, the grouping dimensionality reduction model can achieve faster convergence and can well maintain the original information of data. The experimental results on the public dataset show that the grouping dimensionality reduction model is an effective dimensionality reduction algorithm and can be employed to extract characteristics on big data.

1. Introduction

The era of big data comes with a sharply increasing amount of data and dramatically improving data dimensionality. In the research and application of machine learning and data mining, the harm caused by the curse of dimensions is also becoming increasingly serious. An inherent but more serious problem of high-dimensional data in data analysis rests in that it affects the effect of further data analysis and processing. Studies have shown that the classification effect of data after dimensionality reduction is superior to the direct data classification effect [1-4].

In big data analysis, a common situation is that samples are presented by group, data of different groups share some information and each group has their respective unique information at the same time. For example, in medical research, experimenters usually divide subjects into an experimental group and a control group, give normal drugs to the experimental group and placebo to the control group and then compare the performance of the subjects in these two groups. Apparently, the experimental and control groups will have shared unique information. Similarly, in document classification, different documents have both similarities and differences. Each document can be regarded as a set of data and the entire corpus as the whole sample. At present, important correlated dimensionality reduction studies include: 1) a dimensionality reduction method based on transfer learning[5], which assumes that data of different groups have a closer distance in low-dimensional space, so they try to find low-dimensional representations of data of different groups. 2) a dimensionality reduction method based on the scattering matrix[6], which sets up a scatter matrix that links different groups and maintains the distinguishability of labels during dimensionality reduction by making use of the data
label information and the Fisher Discriminant Analysis (FDA) framework.

However, these models only use the traditional method and need to set the dimensionality of dimensionality reduction in advance. To address these defects, we intend to introduce the widely studied Bayesian Non-parametric Model. An important concept in the Bayesian Non-parametric Model is the random distribution. Ferguson proposed the dependent Dirichlet process [7], which uses the stick-breaking paradigm of DP to share weights of atoms and keep atoms independent of each other. Teh et al. advanced the Hierarchical Dirichlet Process (HDP) based on the Dirichlet process [8]. The model first assumes a base distribution $H$ and then samples a discrete distribution $G_0$ to be used as a pool of public resources according to DP $(\alpha, H)$. Wherein, DP represents the Dirichlet process, $\alpha$ is a constant and $H$ is a probability measure. The corresponding distributions $G_1, ..., G_n$ of each group select atoms from $G_0$ as their unique information. Leisen et al. proposed the vectors of Dirichlet process [9] to construct vectors consisting of random processes, each edge of which is a Dirichlet process. Epifani et al. constructed a correlated random process class by Lévy copula [10]. With the given random process of each edge, the method can determine different correlated random processes with different copulas. MRMI (mixed of random measures with independent increment) [11, 12] is composed of different linear combinations of independent random processes. This method is characterized in great flexibility and convenience and also good operability.

In the traditional Bayesian non-parametric statistical methods, only the hierarchical Dirichlet process can analyse and process data containing grouping information. We propose a dimensionality reduction model for complex data grouping, which can ensure the sharing of grouping information in the dimensionality reduction process. Subsequently, the paper studies in detail the Lévy measure of linearly mixed random probability and the corresponding inference process and apply the random probability to analyse and model the dimensionality reduction of documents.

2. Related Work

In this section, we review some existing dimensionality reduction methods.

2.1. Hierarchical Dirichlet Process

The Dirichlet process is a measure of the probability measure, which comprises the base probability measure and the concentration parameter. The Dirichlet process works well in clustering data and does not need to specify the number of clusters. The hierarchical Dirichlet Process (HDP) is a Dirichlet process of hierarchy and introduces another hierarchy of DP over DP. A text can be described by DP, but the topics are shared among texts, a hierarchy of DP needs to be introduced. First, the entire text set is divided into topics by the DP process; then, the DP of each text divides the text into topics. Here, each topic comes from the superior DP, that is, a global measure $G_0$ is generated from a DP of which the base probability measure and the concentration parameter is $H$ and $\gamma$ respectively. Then, the random measure $G_j$ of the specific set is generated from a DP whose base measure is $G_j$:

$$
\begin{cases}
    G_0 \sim DP(\gamma, H) \\
    (G_j|G_0) \sim DP(\alpha_0, G_0)
\end{cases}
$$

Wherein, $\alpha_0$ is the concentration parameter of the measure $G_j$.

2.2. Linear discriminant analysis (LDA)

LDA get the highly discriminative low-dimensional representation by maximizing within-class scatter while minimizing intra-class scatter. Wherein, the within-class scatter matrix $S_b$ and the intra-class scatter matrix $S_w$ are defined as follows.

$$
S_b = \frac{1}{n} \sum_{i=1}^{C} n_i (\mu_i - \mu)(\mu_i - \mu)^T
$$

$$
S_w = \frac{1}{n} \sum_{i=1}^{C} \sum_{x_j \in C_j} (x_j - \mu)(x_j - \mu)^T
$$
Where $c_j$ is the set of Class $i$ samples, $\mu_i$ is the sample mean of Class $i$ samples and $\mu$ is the mean of all samples. Thus, the LDA algorithm can come down to the optimization problems below:

$$
\max_{v} \frac{\sum_{i=1}^{n} n_i ||v^T \mu_i - v^T \mu||^2}{\sum_{i=1}^{m} \sum_{j \in c_i} ||v^T x_i - v^T x_j||^2}
$$

(4)

Equivalently, the above equation can be rewritten as a trace ratio optimization form:

$$
\max_{V} \frac{\text{tr}(V^T S_b V)}{\text{tr}(V^T S_w V)}
$$

(5)

Since the objective function (5) is non-convex, no optimal global solution of the analytical form is available, and it is usually approximately transformed into the following trace ratio optimization problem:

$$
\max tr((V^T S_w V)^{-1} V^T S_b V)
$$

(6)

When the within-class scatter matrix $S_w$ is non-singular, the optimal projection matrix $V$ obtained by Equation (6) is composed of the eigenvectors corresponding to $r$ largest eigenvalues of $S_w^{-1}S_b$.

2.3. Maximum margin criteria (MMC)

Similar to LDA, the optimization goal of MMC is also to maximize the intra-class scatter and minimize the within-class of samples in the low-dimensional space after projection. The difference lies in that MMC uses a trace difference optimization model:

$$
J(V) = \text{tr}(V^T (S_b - S_w) V^T)
$$

(7)

Each column defined $V$ is a unit vector and thus $V$ can be acquired by the following eigendecomposition:

$$
(S_b - S_w) v_i = \lambda_i v_i
$$

(8)

3. Mixed random measure model in data dimensionality reduction

This section will introduce the main model and its correlated properties. The model presented in this section is a framework applicable to any kind of regular random measure.

3.1. Dimensionality reduction model based on regular correlated random measure

Now, we will build a dimensionality reduction model based on regular correlated random measure. Note that in general we do not directly make samples themselves, but samples’ parameters, obey a random measure. Here, we are building a text dimensionality reduction model, of which is the basic idea is to regard a corpus as a grouped sample set, each document corresponds to a set of data, each word is a sample in the document and then all the documents constitute a corpus. $D = \{D_1, ..., D_m \}$ is used to represent a corpus, where $D_i$ is the $i$th document in the corpus, with $i = 1, ..., m$. In the model, each document corresponds to a regular random measure in a regular correlated random measure, i.e. a component of vector $p$. Each random probability in vector $p$ is a mixed probability concerning the characteristics of the document, each atom of which corresponds to a characteristic and the weight on the atom corresponds to the probability that the corresponding characteristic appears in the document.

The sampling method of each word in each document is as below:

1. Determine the random probability corresponding to each document and the distribution of each topic;
2. Extract a characteristic $T_k$ for the document $r$ through the regular correlated random measure $p_r$;
3. Sample a word from the topic $T_k$.

Formally, assuming that $T_{r,j}$ denotes a document characteristic, i.e., a distribution of the dictionary $V$ and that $p_r$ a random probability in the regular correlated random measure, then its form is:

$$
p_r = \sum_{i=1}^{m} p_{r,i} \delta_{T_i}
$$

(9)

And each $T_{r,j}$ must be from $(T_i)_{i \in N}$. Let $w_{r,j}$ be a word in the document $r$ and use RP to represent the regular correlated random measure and then the following model can be got:

$$
\max_{V} \frac{\text{tr}(V^T S_b V)}{\text{tr}(V^T S_w V)}
$$

(5)

Since the objective function (5) is non-convex, no optimal global solution of the analytical form is available, and it is usually approximately transformed into the following trace ratio optimization problem:

$$
\max tr((V^T S_w V)^{-1} V^T S_b V)
$$

(6)

When the within-class scatter matrix $S_w$ is non-singular, the optimal projection matrix $V$ obtained by Equation (6) is composed of the eigenvectors corresponding to $r$ largest eigenvalues of $S_w^{-1}S_b$. 

2.3. Maximum margin criteria (MMC)

Similar to LDA, the optimization goal of MMC is also to maximize the intra-class scatter and minimize the within-class of samples in the low-dimensional space after projection. The difference lies in that MMC uses a trace difference optimization model:

$$
J(V) = \text{tr}(V^T (S_b - S_w) V^T)
$$

(7)

Each column defined $V$ is a unit vector and thus $V$ can be acquired by the following eigendecomposition:

$$
(S_b - S_w) v_i = \lambda_i v_i
$$

(8)

3. Mixed random measure model in data dimensionality reduction

This section will introduce the main model and its correlated properties. The model presented in this section is a framework applicable to any kind of regular random measure.

3.1. Dimensionality reduction model based on regular correlated random measure

Now, we will build a dimensionality reduction model based on regular correlated random measure. Note that in general we do not directly make samples themselves, but samples’ parameters, obey a random measure. Here, we are building a text dimensionality reduction model, of which is the basic idea is to regard a corpus as a grouped sample set, each document corresponds to a set of data, each word is a sample in the document and then all the documents constitute a corpus. $D = \{D_1, ..., D_m \}$ is used to represent a corpus, where $D_i$ is the $i$th document in the corpus, with $i = 1, ..., m$. In the model, each document corresponds to a regular random measure in a regular correlated random measure, i.e. a component of vector $p$. Each random probability in vector $p$ is a mixed probability concerning the characteristics of the document, each atom of which corresponds to a characteristic and the weight on the atom corresponds to the probability that the corresponding characteristic appears in the document.

The sampling method of each word in each document is as below:

1. Determine the random probability corresponding to each document and the distribution of each topic;
2. Extract a characteristic $T_k$ for the document $r$ through the regular correlated random measure $p_r$;
3. Sample a word from the topic $T_k$.

Formally, assuming that $T_{r,j}$ denotes a document characteristic, i.e., a distribution of the dictionary $V$ and that $p_r$ a random probability in the regular correlated random measure, then its form is:

$$
p_r = \sum_{i=1}^{m} p_{r,i} \delta_{T_i}
$$

(9)

And each $T_{r,j}$ must be from $(T_i)_{i \in N}$. Let $w_{r,j}$ be a word in the document $r$ and use RP to represent the regular correlated random measure and then the following model can be got:
\[(p_1, \ldots, p_R) \sim RP\]
\[T_{r,j} \sim \pi_r\]
\[w_{r,j} \sim \tau_{r,j}\]

The weight on the atom activated in \(p_r\) is used to represent the document of a paper after dimensionality reduction. Although \(p_r\) has a myriad of atoms, the above model is indeed a dimensionality reduction model given there is at most a finite number of atoms activated. If \(K_r\) atoms are activated in the mixing probability \(p_r\) corresponding to the \(r\)th document, it can be expressed as

\[p_r = \sum_{i=1}^{K_r} \pi_{r,i} \delta_{r,i}\]  

Certainly, the specific number is associated with the parameters in the Lévy measure, but still reflects an advantage of the non-parametric model that there is no need to pre-set the dimension of dimensionality reduction, which can be reflected by data in the calculation of the model.

3.2. Inference method of mixed random measure

The Gibbs sampling is the most frequently used among Bayesian non-parametric inference methods. Specifically, in Bayesian non-parametric methods, we mainly calculate two probabilities, that is, the probability of the existing atom acquired by a new sample \(X_{new}\) after a constant sample set \(X\) and the probability of a new atom acquired by it.

3.2.1. Posterior inference method. By exchangeability, we can use any sample as the last one and all other samples as the observed sample set to make a posterior inference on the last sample. In actual calculation, we remove a sample each time to infer its situation and this time, only one round of Gibbs sampling is done among all sample sets. Then, the Gibbs sampling is iterated for several times until the algorithm converges.

Let \(Y_1, \ldots, Y_K\) be values different from each other after a sample \(X\) is removed from sample \(X\). We add the auxiliary variable \(u = (\mu_1, \ldots, \mu_R)\). Here, \(q_k\) is defined as a vector with its dimension as \(R\) to denote that the \(k\)th atom is directly proportional to

\[\mathbb{P}(X_r = Y_k | X \setminus \{\cdot \}) \propto \frac{\pi_{q_k} + \delta_{r}(u)}{\tau_{r}(u)}\]  

a new atom is proportional to

\[\mathbb{P}(X_r = Y_k | X \setminus \{\cdot \}) \propto \tau_{r}(u)\]  

In actual sampling, the above two equations need to be normalized and their respective probabilities can be used for sampling after calculation. When \(X_r\) gets a new atom, Equation (13) simply gives the probability that it acquires a new atom, but not the position of the new atom. According to the assumptions in this paper, the weight and position of each atom is independent of each other, so the position of the atom can be directly sampled from \(H(dx)\).

Finally, each component of the auxiliary variable \(u\) needs to be sampled. After variables unrelated to \(\mu_r\) are removed, the density function of \(\mu_r\) is in direct proportion to

\[\mu_r^{n_r-1} \psi(u) \prod_{k=1}^{K} \tau_{q_k}(u)\]  

3.2.2. Linear mixed random measure. In this section, a special correlated random measure is defined with its Lévy measure calculated and then the inference process of this regular correlated random measure is calculated in detail by making use of the a posteriori calculation equation given in the preceding section.

The correlated random measure in this section is known as linearly mixed random measure and comprises different linear combinations of several completely random measures. As a result, the component of this correlated random measure is composed of different linear combination methods. Let \(\tilde{\mu}_1, \ldots, \tilde{\mu}_q\) be \(d\) completely random measure(s) with exactly the same Lévy measure \(\tilde{\psi}(ds)\) and then the \(r\)th linearly mixed random measure is
\[
\mu_r = \sum_{i=1}^{d} w_{r,j} \tilde{\mu}_i
\]

(15)

Where, \( w_{r,j} \geq 0 \) is a constant.

For any function \( f_1, ..., f_R \) measurable almost everywhere, the Laplace transform of \( \mu(f) \) is calculated:

\[
\mathbb{E}[e^{-\mu(f)}] = \mathbb{E} \left[ e^{-\sum_{i=1}^{d} w_{r,j} \tilde{\mu}_i(f_r)} \right]
\]

(16)

According to the linear additivity of the integral, the above equation can be rewritten into:

\[
\mathbb{E}[e^{-\mu(f)}] = \mathbb{E} \left[ e^{-\sum_{i=1}^{d} \tilde{\mu}_i(\sum_{j=1}^{R} w_{r,j} f_r)} \right]
\]

(17)

Due to \( \tilde{\mu}_i \) mutual independence, the right end of the equation above can be written as:

\[
\mathbb{E}[e^{-\mu(f)}] = \prod_{i=1}^{d} \mathbb{E} \left[ e^{-\tilde{\mu}_i(\sum_{j=1}^{R} w_{r,j} f_r)} \right]
\]

(18)

Then, the Lévy-Khintchine equation of the completely random measure can be used to expand the right end of the above equation.

\[
\mathbb{E}[e^{-\mu(f)}] = \prod_{i=1}^{d} \exp \left\{ - \int (1 - e^{-s \sum_{j=1}^{R} w_{r,j} f_r(x)}) \tilde{v}(ds) H(dx) \right\}
\]

(19)

Now, a measurable function \( g(w_1, ..., w_R) \) is constructed as

\[
g(w_1, ..., w_R) = \left\{ \begin{array}{ll}
1, & \text{if } (w_1, ..., w_R) = (w_{1,i}, ..., w_{R,i}), \text{among them } i = 1, ..., d \\
0, & \text{other cases}
\end{array} \right.
\]

(20)

Then the Laplace transform of \( \mu(f) \) can be written as:

\[
\mathbb{E}[e^{-\mu(f)}] = \exp \left\{ - \int (1 - e^{-s \sum_{j=1}^{R} w_{r,j} f_r g(w_1, ..., w_R)(ds) H(dx)}) \right\}
\]

(21)

Further, after substitution with \( s_r = w_r s \), the final result of the Laplace transform can be got.

\[
\mathbb{E}[e^{-\mu(f)}] = \exp \left\{ - \int (1 - e^{-s \sum_{j=1}^{R} w_{r,j} f_r g(w_1, ..., w_R)(ds) H(dx)}) \right\}
\]

(22)

Thus, the Lévy measure of the linear random mixed random measure is:

\[
v(ds_1, ..., ds_R) = \sum_{i=1}^{d} \int 1_{(w_1, ..., w_R)}(ds_1, ..., ds_R) \tilde{v}(ds)
\]

(23)

In the end, for actual calculation and experimental demonstration, the Lévy measure of the linearly mixed random measure is calculated when \( \tilde{v} \) is the Lévy measure of the gamma process, i.e., presenting a real example by Equation (23). Let \( t > 0 \) be a constant and make:

\[
\tilde{v}(ds) = \alpha s^{-1} e^{-s} ds
\]

By directly substituting the above equation into Equation (23), the Lévy measure of the linearly mixed random measure can be acquired, known as Linearly Mixed Gamma Measure (LMGM). After the Lévy measure of LMGM is got, the values of the two essential variables \( \psi(u) \) and \( \tau_{q_k}(u) \) can be calculated. After bringing \( v(ds_1, ..., ds_R) \) given by Equation (23), the following can be got:

\[
\psi(u) = \sum_{r=1}^{R} \log(h_l + 1)
\]

(24)

\[
\tau_{q_k}(u) = \Gamma(t_k) \sum_{r=1}^{R} \prod_{l=1}^{R} q_{r,k}^{w_{r,k}}
\]

(25)

Where,

\[
t_k = \sum_{r=1}^{R} q_{r,k}
\]

is the number of samples owned by the \( k \)th atom, and

\[
h_l = \sum_{r=1}^{R} u_r w_{r,j}
\]
A linear combination of weights on the $i$th completely random measure and the weight of the linear combination is the auxiliary variable $u$. After putting the results of Equations (24) and (25) into Equations (12) and (13), a probability that a new sample $X_{new}$ in LMGM takes the existing atom $k$ is proportional to:

$$
t_k \left( \sum_{i=1}^{d} w_{r,i}^{q_r k+1} \prod_{l=1}^{m} w_{l,i}^{q_l k} \right) \left( \sum_{i=1}^{d} \prod_{l=1}^{m} \frac{w_{l,i}^{q_l k}}{(h_{i+l})^{q_l k}} \right)^{-1}
$$

(26)

Moreover, the probability that $X_{new}$ gets a new atom is in direct proportion to:

$$
\alpha \sum_{i=1}^{d} \frac{w_{r,i}}{w_{r,i}+1}
$$

(27)

Comparing the above update equation with that in the Dirichlet process, it can be made clear that the most significant difference is that there is an auxiliary variable $u$ in the LMGM. Then, let $d$ in LMGM be 1. In other words, with the availability of only one completely random measure, LMGM can be degenerated to the Dirichlet process. In fact, LMGM can be regarded as a generalization of the normalized random measure. When $d = 1$, the update equation above has a same update equation with the normalized random measure. Please see Reference [13] for information concerning how to get the Dirichlet process through the normalized gamma process.

The auxiliary variable $u$ and weight $w_{r,i}$ are updated by the stochastic gradient MCMC proposed in Reference [14]. Here, the gradients of the two parameters are given for reference. They are:

$$
\frac{\partial p}{\partial u_r} = \frac{n_r - u_r}{u_r} - \alpha \sum_{i=1}^{d} \frac{w_{r,i}}{h_{i+1}} - \sum_{k=1}^{K} \frac{\tau_{k} q_k(u_1, \ldots, u_R)}{\tau_{k}(u_1, \ldots, u_R)}
$$

(28)

$$
\frac{\partial p}{\partial w_{r,i}} = - \alpha \frac{u_r}{h_{i+1}} + \sum_{k=1}^{K} \frac{\tau_{r,k} q_{r,k}}{w_{r,i}} - \frac{r_{k} u_r}{w_{r,i}} \left( \sum_{l=1}^{d} \prod_{l=1}^{m} \frac{w_{l,i}^{q_l k}}{(h_{i+l})^{q_l k}} \right)
$$

(29)

After some algebraic operations, we get:

$$
\frac{\tau_{q_k u_1, \ldots, u_R}}{\tau_{q_k u_1, \ldots, u_R}} = t_k \left( \sum_{i=1}^{d} \frac{w_{r,i}}{h_{i+1}} \sum_{l=1}^{d} \frac{w_{l,i}^{q_l k}}{(h_{i+l})^{q_l k}} \right) \prod_{l=1}^{m} \frac{w_{l,i}^{q_l k}}{w_{r,i}^{q_r k}}
$$

(30)

4. Experiments

4.1. Artificial dataset

To simplify the experiment and verification, the following model is used to generate data. Let $\mu_0$, $\mu_1$ and $\mu_2$ be three fixed distributions with their respective density functions being $f_0$, $f_1$ and $f_2$. To simulate the linearly mixed random measure, we construct

$$
\mu_1 = \tilde{\mu}_0 + \mu_1
$$

(31)

$$
\mu_2 = \tilde{\mu}_0 + \mu_2
$$

(32)

That is, the two random measures share one distribution, while the other distribution is different. We hope that the completely random measure in the posterior inference will have the corresponding result. In the specific calculation, we adopt

$$
\tilde{\mu}_1 = \frac{1}{2N(-10,1)} + 1/2N(-5,1)
$$

(33)

$$
\tilde{\mu}_2 = \frac{1}{2N(0,1)} + 1/2N(5,1)
$$

(34)

$$
\tilde{\mu}_2 = \frac{1}{2N(10,1)} + 1/2N(15,1)
$$

(35)

Where $N(\alpha, b)$ is a normal distribution with a mean of $\alpha$ and a standard deviation of $b$. The structure of $\mu_0$, $\mu_1$ and $\mu_2$ is:

$$
\mu_r = w_{r,1} \mu_1 + w_{r,2} \mu_2 + w_{r,3} \mu_3, \ r = 1, 2
$$

(36)
In order to verify the validity of the model inference, we set
\[ w_{1,1} = 0.3, \quad w_{1,2} = 0.0, \quad w_{1,3} = 0.69, \]
\[ w_{2,1} = 0.3, \quad w_{2,2} = 0.69, \quad w_{2,3} = 0.01. \]
With a view to demonstrating the correctness of the inference,
we use some extreme values, that is, each mixed random measure has a completely random measure with a weight approaching to zero. Thus, one of the two mixed measures must have \( w_{r,i} \) approaching to 0. Another reason for setting weights this way rests in that we are attempting to discuss the effect of the number \( R \) of completely random measures. Given the completely random measures are fully integrated a posteriori, the impact of \( R \) cannot be investigated directly, but indirectly determined by observing the weight. Apparently, when \( d \) is set to be greater than or equal to the true value, there must be a weight \( w_{r,i} \) in each \( \mu_r \) approaching to 0. On the contrary, if \( d \) is smaller than the true value, there is no completely random measure with its weight approaching to 0 in any \( \mu_r \). The histogram of the data generated is as illustrated in figure 1.

![Figure 1. Histogram of the Two Groups of Data](image)

According to the figure, the first two classes are shared by the two groups of data, while the third and fourth classes are only used by the second group and the fifth and sixth classes only by the first group.

Table 1. Mean of Weights and Number of Activated Classes.

|   | \( w_{1,1} \) | \( w_{1,2} \) | \( w_{1,3} \) | \( w_{1,4} \) | \( w_{2,1} \) | \( w_{2,2} \) | \( w_{2,3} \) | \( w_{2,4} \) | \( K \) |
|---|---|---|---|---|---|---|---|---|---|
| \( i = 1 \) | 2.1563 | 0.0743 | 0.0504 | 2.0894 | 8.0785 |
| \( i = 2 \) | 2.0379 | 0.0635 | 0.0346 | 1.6804 | 0.8997 | 6.0480 |
| \( i = 3 \) | 2.3880 | 1.0882 | 0.0351 | 1.3252 | 2.3666 | 11.3251 | 6.1414 |

In the experiment, Gibbs sampling is done for 10,000 times and with the first 2,000 results ignored as burn-in. The mean of \( w_{r,i} \) is as listed in Table 1. As shown in the table, when \( d = 3 \) (the second line), 6 classes inferred by the posterior inference are inferred and \( r = 1 \) or \( r = 2 \). (\( w_{1,2} \) and \( w_{2,1} \)) has its weight approaching to 0. Further, it can be seen that the ratio of weights (0:1:0 and 0.9:1.7) in each group is close to the ratio of true values (0.01:0.3:0.69). We also need to note that the largest and smallest number in the first group is \( w_{1,1} \) and \( w_{1,2} \) separately and that in the second group is \( w_{2,2} \) and \( w_{2,1} \) respectively. That is to say, when the first group has the largest weight, the second group has the smallest one and when the first group has the smallest weight, the second group has the largest one, which is the same as we assumed.

When \( d = 4 \) (Table 1), the mean number of activated atoms also approximates the true value of 6. Similarly, each group has one weight approximating 0 and the maximum and minimum weights appear in pairs, that is, when the maximum value in the first group is \( w_{1,1} \), the minimum value in the
second group is \( w_{2,1} \) and when the minimum value in the first group is \( w_{1,3} \), the maximum value in the second group is \( w_{2,3} \). This indicates that the posterior inference roughly gives the structure of the model.

### Table 2. The cluster and the percentages of each group.

| Cluster | Group 1 | Group 2 |
|---------|---------|---------|
|         | Mean    | Count   | Percent | Count   | Percent |
| 1       | 5.0820  | 1       | 0.33%   | 104     | 34.67%  |
| 2       | 15.0416 | 96      | 32.00%  | 1       | 0.33%   |
| 3       | -9.8725 | 52      | 17.33%  | 3       | 1.00%   |
| 4       | -0.0499 | 5       | 1.67%   | 91      | 30.33%  |
| 5       | -5.0027 | 4       | 1.33%   | 55      | 18.33%  |
| 6       | 9.9423  | 104     | 34.67%  | 3       | 1.00%   |
| 7       | -5.0607 | 38      | 12.67%  | 2       | 0.67%   |
| 8       | -10.0084| 9       | 0.00%   | 41      | 13.67%  |

When \( d = 2 \), our inference algorithm fails to infer the correct model structure, since there are too few completely random measures \( m \), not enough for allocation to each linear mixed measure. However, the result under such this condition is more interesting. According to the inference results, the algorithm splits one of the classes into two and assigns them to each linearly mixed random measure. First, Table 1 shows that there are 8 activated atoms, rather than 6, the correct value. Second, as for the position of each atom and the proportion of samples concentrated on them (Table 2), though atom 3 and atom 8 have almost the same position, and atom 5 and atom 7 also have almost the same position, the algorithm divides them into four atoms and then assigns samples to each atom. According to Table 2, there are 17\% and 12\% samples on atom 3 and atom 7 respectively in the first group but almost 0 in the second group. However, the number of sample on atom 8 and atom 5 is just the opposite.

After putting these experimental results together, it becomes clear that when the number \( d \) set of the completely random measures is larger than or equal to the true value, the model can well infer the structure of the sample, while when \( d \) is smaller than the true value, it is hard to correctly reflect the structure of the sample, in which case the algorithm assigns some shared classes to each group.

#### 4.2. Real dataset

In the experiment of real dataset, the Open American National Corpus is employed for experimental analysis. It is a large electronic collection of American English and includes all types of text and spoken transcripts since 1990. All data and comments included in it can be used for all purposes. This corpus comprises two datasets, i.e., OANC and MASC.

In the parameter setting of the model, a linear mixed gamma random measure is employed, where the parameter of the gamma Lévy measure is set as \( \alpha = 3 \). The prior of the subject is set as a symmetric Dirichlet distribution, where the parameter is set as \( \lambda = 0.05 \). The initial values of all weights and auxiliary variables are set as 1 and the number \( d \) of the completely random measure as 40. The perplexity, a common standard in document mining and representing the energy of corpus, is used to judge the algorithm convergence. The smaller the value of the perplexity is, the effect the structure will be.

The convergence of the dimensionality reduction model is as illustrated in figure 2(a). The value of the perplexity of each round is used to determine the algorithm convergence. According to the figure, the algorithm has a faster rate of convergence. When Gibbs sampling proceeds to the 100th rounds, the algorithm has basically converged when the value of the perplexity stabilizes at around 70. Figure 2(b) reflects the number of atoms activated in each document. In the theoretical assumption of LMGM, the weights of all atoms are positive with some having smaller values. Atoms with their weight lower than 0.01 are regarded as not to be activated and then, the number of atoms activated in each document is
counted. The weights of these activated atoms are the representation of each document after dimensionality reduction. The figure shows that the dimensions of all documents after dimensionality reduction are different when the dimension of each document is not preset when the model is built, but reflected by data. Such a setting is significantly superior to the dimensions of the designated dimensionality reduction.

![Figure 2. Posterior inference result of MASC dataset](image)

(a) Value of the MASC Dataset Perplexity in Each Round of the Loop  
(b) Histogram of the Number of Atoms Activated in the MASC Dataset

5. Conclusion
This document proposes a grouped data model with the correlated random measure being a priori for information sharing of complex grouped data containing group information. This model can share information among data of groups. For each group of data, random measures are given being a priori and these random measures are correlated. By the co-joint Lévy measure of the correlated random measure, the parameter posterior inference framework of the grouped data is offered. Notably, we explicitly calculated the Lévy measure of the mixed random measure and offered the inference steps of detailed parameter a posteriori. This inference method is similar to the Chinese restaurant process, in which all the random measures are integrated, and only the values of parameters are finally activated are retained, thus with a faster rate of convergence.

Acknowledgments
This work is supported by the Project of Science and Technology Commission of Shanghai Municipality under Grant No. 16JC1403000 and the National Natural Science Foundation of China under Grant No. 71571136. We thank the reviewers for their helpful comments.

References
[1] Cunningham JP, Ghahramani Z 2015 Linear dimensionality reduction: Survey, insights, and generalizations. The Journal of Machine Learning Research. 16(1) 2859-900
[2] Weinberger KQ, Sha F, Saul LK 2010 Convex optimizations for distance metric learning and pattern classification [applications corner]. IEEE Signal Processing Magazine. 27(3) 146-58
[3] Liaw A, Wiener M 2002 Classification and regression by randomForest. R news. 2(3) 18-22
[4] Kim H, Howland P, Park H 2005 Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research. 6(Jan) 37-53
[5] Pan SJ, Kwok JT, Yang Q Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence Transfer Learning via Dimensionality Reduction. Palo Alto, AAAI Press 8 677-82
[6] Tu W, Sun S Proceedings of 23rd IEEE International Conference on Tools with Artificial Intelligence Transferable discriminative dimensionality reduction. Piscataway, IEEE 865-8
[7] Ferguson TS 1973 A Bayesian analysis of some nonparametric problems. The annals of statistics. 209-30
[8] Teh YW, Jordan MI, Beal MJ, Blei DM 2006 Hierarchical Dirichlet Processes. Publications of the American Statistical Association. 101(476) 1566-81
[9] Leisen F, Lijoi A, Spanò D 2013 A vector of Dirichlet processes. Electronic Journal of Statistics. 7(62-90
[10] Kallsen J, Tankov P 2006 Characterization of dependence of multidimensional Lévy processes using Lévy copulas. Journal of Multivariate Analysis. 97(7) 1551-72
[11] Griffin JE, Kolossiatis M, Steel MF 2013 Comparing distributions by using dependent normalized random - measure mixtures. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 75(3) 499-529
[12] Lijoi A, Nipoti B, Prünster I 2014 Dependent mixture models: clustering and borrowing information. Computational Statistics & Data Analysis. 71(417-33
[13] Favaro S, Teh YW 2013 MCMC for normalized random measure mixture models. Statistical Science. 28(3) 335-59
[14] Ma Y-A, Chen T, Fox E Proceedings of Neural Information Processing Systems 2015 A complete recipe for stochastic gradient MCMC. Palais des Congrès de Montréal, 2917-25