Coupling Implicit and Explicit Knowledge for Customer Volume Prediction

Jingyuan Wang,[†] Yating Lin,[†] Junjie Wu,[‡]* Zhong Wang,[†] Zhang Xiong[†§]
† School of Computer Science and Engineering, Beihang University, Beijing, China
‡ School of Economics and Management, Beihang University, Beijing, China
§ Research Institute of Beihang University in Shenzhen, Shenzhen, China
Email: {jywang, linyating, wujj, wangzhong, xiongz}@buaa.edu.cn, *Corresponding author

Abstract

Customer volume prediction, which predicts the volume from a customer source to a service place, is a very important technique for location selection, market investigation, and other related applications. Most of traditional methods only make use of partial information for either supervised or unsupervised modeling, which cannot well integrate overall available knowledge. In this paper, we propose a method titled GR-NMF for jointly modeling both implicit correlations hidden inside customer volumes and explicit geographical knowledge via an integrated probabilistic framework. The effectiveness of GR-NMF in coupling all-round knowledge is verified over a real-life outpatient dataset under different scenarios. GR-NMF shows particularly evident advantages to all baselines in location selection with the cold-start challenge.

Introduction

Customer volume prediction refers to the problem of predicting customer volumes (footfalls) from customer sources, *e.g.*, residential zones, to service places, *e.g.*, theaters, shopping malls, schools, and hospitals. It plays a key role in many business and public affair applications. For example, in location selection applications, customer volume prediction is utilized to predict potential footfalls to a location candidate (Xu et al. 2016; Hernandez and Bennison 2000). In business and public investigation applications, it is also utilized to estimate the competitiveness or investment value of service places (Fu et al. 2014b).

Traditional methods for customer volume prediction usually adopt intuitive information, such as spatial interactions (Athiyaman 2011), road structure (Bafna 2003), and accessibility (Medda 2012), to predict unknown footfall from a customer source point to a service place. In recent years, with the rapid development of data sciences, more and more researchers turn to data-driven methods (Li et al. 2015; Karamshuk et al. 2013) and empirical models (Simini et al. 2012; Jensen 2006) for higher-quality prediction. The rich studies along this line, however, are mostly concerned with modeling partial information as either a supervised or unsupervised learning problem. Further study is still in great need to integrate all-round knowledge available for robust customer volume prediction.

Driven by this motivation, in this paper, we propose a *Geographical Regression and Non-negative Matrix Factorization* (GR-NMF) model for high-quality customer volume prediction. The main contributions of our study are summarized as follows.

First, GR-NMF is able to jointly model implicit knowledge hidden inside customer volumes and explicit knowledge expressed as geographical relations. Specifically, the regression model for the explicit knowledge is formulated as an order-unsymmetrical matrix factorization problem, which is naturally compatible with the matrix factorization framework for implicit knowledge mining.

Second, GR-NMF can be regarded as a semi-supervised learning framework that models unobserved customer volumes as the calibration of implicit knowledge modeling with explicit knowledge modeling. This turns out to be the coupling item and fits the general matrix factorization framework nicely.

Third, GR-NMF has a unified probabilistic interpretation, which makes the model theoretically solid. All the parameters have definite statistical meanings, which can help users to understand the relations and contribution weights of implicit and explicit knowledge. Clear guidance to the fine setting of these parameters is carefully given for practical use.

Extensive experiments are conducted on a real-life outpatient dataset obtained from the Shenzhen city of China. The results show that GR-NMF outperforms competitive baselines consistently in various application scenarios with different sampling rates. In particular, equipped with explicit geographical knowledge, GR-NMF shows prominent advantages to all baselines in location selection with the cold-start challenge. The rationale of the approximate method for parameter setting is also testified empirically.

Model and Inference

Throughout the paper, we use lowercase symbols such as a, b to denote scalars, bold lowercase symbols such as $\mathbf{a}$, $\mathbf{b}$ to denote vectors, bold uppercase symbols such as $\mathbf{A}$, $\mathbf{B}$ to denote matrices, and calligraphy symbols such as $\mathcal{A}$, $\mathcal{B}$, to denote tensors.

Assume $S = \{s_1, s_2, \ldots, s_M\}$ contains M *service points*, and $C = \{c_1, c_2, \ldots, c_N\}$ contains N *customer-source points* (or *customer points* for short). Let $\tilde{x}_{ij}$ denote the absolute customer volume from c_j to s_i, then the *customer volumes matrix* $\mathbf{X} \in \mathbb{R}^{M\times N}$ is defined as

$$x_{ij} = \log(\tilde{x}_{ij} + 1), \forall\, i, j. \tag{1}$$

Note that we adopt the logarithmic customer volume to avoid modeling biases due to the severe imbalance of absolute customer volumes for different service-customer point pairs (Wang et al. 2014). We also define a *binary sampling matrix* $\mathbf{Y} \in \mathbb{R}^{M\times N}$ for $\mathbf{X}$, where the element y_{ij} equals to 1 when customer volume from c_j to s_i is sampled and 0 otherwise. In other words, we know the customer volumes with sampling indicator 1, and the target of our model is to predict the customer volumes with sampling indicator 0.

Probabilistic Matrix Factorization for Implicit Correlations

We first propose a probabilistic matrix factorization to explore latent correlations in the customer volumes matrix $\mathbf{X}$. Suppose there exist H latent patterns in $\mathbf{X}$. The correlations between the i-th service point and the H patterns is measured by a projection vector $\mathbf{s}_i \in \mathbb{R}^H$, where the h-th element s_{ih} is the coefficient projecting s_i to pattern h. Similarly, the projection vector from the customer point c_j to the H latent patterns is denoted as $\mathbf{c}_j \in \mathbb{R}^H$. As a result, we have two projection matrices $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M] \in \mathbb{R}^{H\times M}$ and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N] \in \mathbb{R}^{H\times N}$ for service and customer points, respectively. We then model $\mathbf{X}$ using $\mathbf{S}$ and $\mathbf{C}$ as

$$\mathbf{X} = \mathbf{S}^\top\mathbf{C} + \mathbf{E}_1, \tag{2}$$

where $\mathbf{E}_1$ is an error matrix.

We adopt a Gaussian observation noise to model $\mathbf{E}_1$. That is, for each element e_{ij} in $\mathbf{E}_1$, we have $e_{ij} \sim \mathcal{N}(0, \sigma_{X1}^2), \forall\, i, j$. Hence, the conditional distribution over the sampled elements in $\mathbf{X}$ is defined as

$$P(\mathbf{X}|\mathbf{S}, \mathbf{C}, \sigma_{X1}^2) = \prod_{i=1}^{M}\prod_{j=1}^{N}\left(\mathcal{N}(x_{ij}|\mathbf{s}_i^\top\mathbf{c}_j, \sigma_{X1}^2)\right)^{y_{ij}}. \tag{3}$$

Since a service point usually absorbs customers who live in the neighborhood, it is intuitive that most of the elements of $\mathbf{X}$ are zeros, *i.e.*, $\mathbf{X}$ is a sparse matrix. As a result, it is reasonable to assume zero-mean Laplace priors on the projection vectors $\mathbf{s}_i$ and $\mathbf{c}_j$, which gives

$$\begin{aligned} P(\mathbf{S}|\sigma_S^2) &= \prod_{i=1}^{M}\mathcal{L}(\mathbf{s}_i|\mathbf{0}, \sigma_S^2\mathbf{I}), \\ P(\mathbf{C}|\sigma_C^2) &= \prod_{j=1}^{N}\mathcal{L}(\mathbf{c}_j|\mathbf{0}, \sigma_C^2\mathbf{I}). \end{aligned} \tag{4}$$

According to the Bayes' theorem as well as Eq. 3 and Eq. 4, the log posterior distribution of the projection vectors can be formulated as

$$\begin{aligned} &\log P(\mathbf{S}, \mathbf{C}|\mathbf{X}, \sigma_{X1}^2, \sigma_S^2, \sigma_R^2) \\ &\propto -\frac{1}{\sigma_{X1}^2}\sum_{i,j} y_{ij}(x_{ij} - \mathbf{s}_i^\top\mathbf{c}_j)^2 - \frac{1}{\sigma_S^2}\sum_i \|\mathbf{s}_i\|_1 \\ &\quad -\frac{1}{\sigma_C^2}\sum_j \|\mathbf{c}_j\|_1. \end{aligned} \tag{5}$$

Therefore, the *Maximum A Posteriori* (MAP) estimation of $\mathbf{S}$ and $\mathbf{C}$ is to minimize the objective function $\mathcal{J}_1$ as

$$\mathcal{J}_1 = \frac{1}{\sigma_{X1}^2}\left\|\mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top\mathbf{C})\right\|_F^2 + \frac{1}{\sigma_S^2}\|\mathbf{S}\|_1 \frac{1}{\sigma_R^2}\|\mathbf{C}\|_1, \tag{6}$$

where $\|.\|_F^2$ is the Frobenius Norm, $\|.\|_1$ is the L1 Norm, and $\odot$ is the Hardamard Product.

Geographical Regression for Explicit Correlations

We here propose a regression-based matrix factorization method for the modeling of explicit geographical factors. Suppose we altogether have $K-1$ explicit geographical factors that could make impact on the footfall from a customer point to a service point. That is, for any $x_{ij} \in \mathbf{X}$, we have a geographical feature vector $\mathbf{a}_{ij} = \left[a_{ij1}, \ldots, a_{ijk}, \ldots, a_{ij(K-1)}\right]^\top$, where a_{ijk} is the k-th feature of footfall x_{ij}. As a result, we can use a linear regression to model the relations between x_{ij} and $\mathbf{a}_{ij}$ as

$$x_{ij} = \mathbf{w}^\top\mathbf{a}_{ij} + b, \tag{7}$$

where $\mathbf{w} = [w_1, \ldots, w_k, \ldots, w_{K-1}]^\top$ is a weight vector to learn from Eq. 7. For the sake of concision, we define $a_{ijK} = 1$ and $w_K = b$, which gives: $x_{ij} = \mathbf{w}^\top\mathbf{a}_{ij}, \forall\, i, j$.

Furthermore, if we regard $\mathbf{a}_{ij}$ as the (i, j) fiber of a tensor $\mathcal{A}$, the linear regression in Eq. 7 can be rewritten as an order-unsymmetrical matrix factorization form:

$$\mathbf{X} = \mathcal{A} \times_k \mathbf{w} + \mathbf{E}_2, \tag{8}$$

where $\times_k$ is the k-mode product (Kolda and Bader 2009) between tensor $\mathcal{A}$ and vector $\mathbf{w}$, *i.e.*, a $K \times 1$ matrix $\mathbf{W}$. Again we adopt a Gaussian observation noise with variance σ_{X2}^2 to model the error $\mathbf{E}_2$. The conditional distribution over the sampled entries in $\mathbf{X}$ is given by

$$P(\mathbf{X}|\mathbf{w}, \sigma_{X2}^2) = \prod_{i=1}^{M}\prod_{j=1}^{N}\left(\mathcal{N}(x_{ij}|\mathbf{w}^\top\mathbf{a}_{ij}, \sigma_{X2}^2)\right)^{y_{ij}}. \tag{9}$$

We then introduce a zero-mean Gaussian prior on the regression weight vector, which gives

$$P(\mathbf{w}|\sigma_{W1}^2) = \prod_{k=1}^{K}\mathcal{N}(w_k|0, \sigma_{W1}^2). \tag{10}$$

According to the Bayes' theorem, the log posterior distribution over the regression weight vector is calculated by

$$\begin{aligned} &\log P(\mathbf{w}|\mathbf{X}, \sigma_{X2}^2, \sigma_{W1}^2) = \log\frac{P(\mathbf{X}|\mathbf{w}, \sigma_{X2}^2)P(\mathbf{w}|\sigma_{W1}^2)}{P(\mathbf{X})} \\ &\propto -\frac{1}{\sigma_{X2}^2}\sum_{i,j} y_{ij}(x_{ij} - \mathbf{w}^\top\mathbf{a}_{ij})^2 - \frac{1}{\sigma_{W1}^2}\sum_k w_k^2. \end{aligned} \tag{11}$$

Therefore, the MAP estimation of $\mathbf{w}$ is equivalent to minimizing the objective function $\mathcal{J}_2$ as

$$\mathcal{J}_2 = \frac{1}{\sigma_{X2}^2}\|\mathbf{Y} \odot (\mathbf{X} - \mathcal{A} \times_k \mathbf{w})\|_F^2 + \frac{1}{\sigma_{W1}^2}\|\mathbf{w}\|_2^2. \tag{12}$$

Now we remain to identify some useful geographical factors to feed Eq. 12. Specifically, three types of geographical features are adopted in our study as follows.

Geographical Relation. Geographical relation means the geographic relationship between service points and customer points, which in our paper includes:

1. Reciprocal of square distance from the customer point c_j to the service point s_i, *i.e.*, $\frac{1}{d_{ij}^2}$, where d_{ij} is the distance from c_j to s_i, which follows the assumption of the gravity model (Zipf 1946).
2. Geographical competition of the service point s_i with other service points to the customer point c_j, which is measured by $\frac{1}{u_{ij}+1}$, where u_{ij} is the number of service points located nearer than s_i to c_j.
3. Geographical competition of the customer point c_j with other customer points to the service point s_i, which is measured by $\frac{1}{v_{ij}+1}$, where v_{ij} is the number of customer points located nearer than c_j to s_i.

Geographical Similarity. Geographical similarity means the footfall similarity among customer-service point pairs that are geographically close. The features about geographical similarity we introduced include:

1. The average footfall to the service point s_i from the five customer points nearest to the customer point c_j;
2. The average footfall from the customer point c_j to the five service points nearest to the service point s_i.

If any of the above two features is unavailable, we use average footfall from the five customer points nearest to c_j to the five service points nearest to s_i as an alternate.

Social Geography Features. Social geography means social connections between customer and service points. The features about social geography we introduced include:

1. Whether the customer point c_j and the service point s_i are in the same administrative region;
2. The population flow intensity from the block of the customer point c_j to the block of the service point s_i (which is approximated by the taxi flow between the two blocks in our study).

It is very easy to extend the above features using other useful geographical information, which we will not elaborate any more.

Modeling Unobserved Volumes

In both the probabilistic matrix factorization model and the geographical regression model, we only considered the observed footfall samples in $\mathbf{X}$ with $y_{ij} = 1$. However, $\mathbf{X}$ could be a huge matrix with majority of unobserved elements (unknown footfalls). In order to model these elements, we introduce a coupling item to calibrate the probabilistic matrix factorization model with that of the geographical regression results on the footfalls with $y_{ij} = 0, \forall\, i, j$.

Specifically, for the footfall elements with $y_{ij} = 0$, we require

$$\mathbf{S}^\top \mathbf{C} = \mathcal{A} \times_k \mathbf{w} + \mathbf{E}_3, \tag{13}$$

where the error $\mathbf{E}_3$ is also a Gaussian noise. The conditional distribution over the elements in the reconstructed matrix $\mathbf{S}^\top \mathbf{C}$ with $y_{ij} = 0$ is thus defined as

$$P(\mathbf{S}, \mathbf{C} | \mathbf{w}, \sigma_{W2}^2) = \prod_{i=1}^{M} \prod_{j=1}^{N} \left(\mathcal{N}(\mathbf{s}_i^\top \mathbf{c}_j | \mathbf{w}^\top \mathbf{a}_{ij}, \sigma_{W2}^2) \right)^{\bar{y}_{ij}}, \tag{14}$$

where $\bar{y}_{ij}$ is negation of y_{ij}, *i.e.*, $\bar{y}_{ij} = 1 - y_{ij}, \forall\, i, j$. The log posterior distribution over $\mathbf{S}$ and $\mathbf{C}$ is thus given as

$$\log P(\mathbf{S}, \mathbf{C} | \mathbf{w}, \sigma_{W2}^2) \propto -\frac{1}{\sigma_{W2}^2} \sum_{i,j} \bar{y}_{i,j} (\mathbf{s}_i^\top \mathbf{c}_j - \mathbf{w}^\top \mathbf{a}_{ij})^2. \tag{15}$$

Therefore, the MAP estimation is to minimize a sum-of-squared errors objective function $\mathcal{J}_3$ as

$$\mathcal{J}_3 = \frac{1}{\sigma_{W2}^2} \left\| \overline{\mathbf{Y}} \odot (\mathcal{A} \times_k \mathbf{w} - \mathbf{S}^\top \mathbf{C}) \right\|_F^2. \tag{16}$$

GR-NMF: Integrated Model for Footfall Prediction

We here integrate all the above objective functions $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ to get a joint model titled *Geographical Regression and Non-negative Matrix Factorization* (GR-NMF) for footfall prediction. The objective function of GR-NMF is

$$\begin{aligned} \min\ \mathcal{J} = & \left\| \mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top \mathbf{C}) \right\|_F^2 \\ & + \alpha \left\| \mathbf{Y} \odot (\mathbf{X} - \mathcal{A} \times_k \mathbf{w}) \right\|_F^2 \\ & + \beta \left\| \overline{\mathbf{Y}} \odot (\mathcal{A} \times_k \mathbf{w} - \mathbf{S}^\top \mathbf{C}) \right\|_F^2 \\ & + \gamma \left\| \mathbf{w} \right\|_2^2 + \delta \left\| \mathbf{S} \right\|_1 + \zeta \left\| \mathbf{C} \right\|_1 \\ s.t.\ & \mathbf{S} \geq 0, \mathbf{C} \geq 0, \mathbf{w} \geq 0, \end{aligned} \tag{17}$$

where $\alpha = \frac{\sigma_{X1}^2}{\sigma_{X2}^2}, \beta = \frac{\sigma_{X1}^2}{\sigma_{W2}^2}, \gamma = \frac{\sigma_{X1}^2}{\sigma_{W1}^2}, \delta = \frac{\sigma_{X1}^2}{\sigma_{S}^2}, \zeta = \frac{\sigma_{X1}^2}{\sigma_{R}^2}$, which can be well estimated in advance by minimizing $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ separately. In other words, these parameters are to be set before the optimization of $\mathcal{J}$. Also note that since all elements in the footfall matrix are nonnegative and the geographical features are positively related to footfalls, we introduce non-negativity constraints to the projection matrices and the geographical features weight vector.

To predict unknown customer volumes, *i.e.*, x_{ij} with $y_{ij} = 0$, is straightforward using $\mathbf{S}$ and $\mathbf{C}$ estimated from Eq. 17. That is, $\hat{x}_{ij} = \mathbf{s}_i^\top \mathbf{c}_j, \forall\, i, j, y_{ij} = 0$.

Inference of GR-NMF

In this section, we introduce an Alternating Proximal Gradient Descent (APGD) method to solve Eq. 17. While $\mathcal{J}$ is not jointly convex *w.r.t.* $\mathbf{S}$, $\mathbf{C}$ and $\mathbf{w}$, it is convex *w.r.t.* each of these variables with the other two fixed. Therefore, we can update $\mathbf{S}$, $\mathbf{C}$ and $\mathbf{w}$ alternatively in an iterative algorithm. Moreover, since $\mathcal{J}$ contains non-differentiable parts, *i.e.*, L1-norms of $\mathbf{S}$ and $\mathbf{C}$, we introduce a Proximal Gradient Descent method to update each variable (Xu and Yin 2013). For easier discussion, we here express the objective function as $\mathcal{J} = \mathcal{F} + \mathcal{H}$, where

$$\begin{aligned} \mathcal{F} = & \left\| \mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top \mathbf{C}) \right\|_F^2 + \alpha \left\| \mathbf{Y} \odot (\mathbf{X} - \mathcal{A} \times_k \mathbf{w}) \right\|_F^2 \\ & + \beta \left\| \overline{\mathbf{Y}} \odot (\mathcal{A} \times_k \mathbf{w} - \mathbf{S}^\top \mathbf{C}) \right\|_F^2 + \gamma \left\| \mathbf{w} \right\|_2^2 \end{aligned} \tag{18}$$

is the differentiable part and

$$\mathcal{H} = \delta \left\| \mathbf{S} \right\|_1 + \zeta \left\| \mathbf{C} \right\|_1 \tag{19}$$

is the non-differentiable part.

In the Proximal Gradient Descent method, the variable $\mathbf{Z}$ of the objective function at the t-th iteration is updated as

$$\begin{aligned} \mathbf{Z}_t &= \arg\min_{\mathbf{Z}} \hat{\mathcal{J}}(\mathbf{Z}, \mathbf{Z}_{t-1}) \\ &= \arg\min_{\mathbf{Z}} \frac{L}{2} \left\| \mathbf{Z} - \left(\mathbf{Z}_{t-1} - \frac{1}{L} \frac{\partial \mathcal{F}}{\partial \mathbf{Z}_{t-1}} \right) \right\|_F^2 + \mathcal{H}(\mathbf{Z}), \end{aligned} \tag{20}$$

where $\hat{\mathcal{J}}$ is a quadratic approximation of the objective function, and L is a Lipschitz constant of $\frac{\partial \mathcal{F}}{\partial \mathbf{Z}_t}$, namely

$$\left\| \frac{\partial \mathcal{F}}{\partial \mathbf{Z}_{t1}} - \frac{\partial \mathcal{F}}{\partial \mathbf{Z}_{t2}} \right\|_F^2 \leq L \left\| \mathbf{Z}_{t1} - \mathbf{Z}_{t2} \right\|_F^2, \forall\, \mathbf{Z}_{t1}, \mathbf{Z}_{t2}. \tag{21}$$

Let $\sigma\left(\mathbf{Z}_{t-1}\right) = \mathbf{Z}_{t-1} - \frac{1}{L} \frac{\partial \mathcal{F}}{\partial \mathbf{Z}_{t-1}}$. The updates of $\mathbf{S}$ and $\mathbf{C}$ at the t-th iteration are then given as

$$\begin{aligned} \mathbf{S}_t &= \max\left(0, \sigma\left(\mathbf{S}_{t-1}\right) - \delta\right), \\ \mathbf{C}_t &= \max\left(0, \sigma\left(\mathbf{C}_{t-1}\right) - \zeta\right), \end{aligned} \tag{22}$$

and w is given as

$$\mathbf{w}_t = \max\left(0, \sigma\left(\mathbf{w}_{t-1}\right)\right), \tag{23}$$

where a max operator is introduced to satisfy the nonnegative constrains. Specifically, the gradients $\frac{\partial \mathcal{F}}{\partial \mathbf{Z}_{t-1}}$ in Eq. 20 for $\mathbf{S}$ and $\mathbf{C}$ are given as

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \mathbf{S}} &= -2\left[\mathbf{C}(\mathbf{Y}^\top \odot (\mathbf{X}^\top - \mathbf{C}^\top \mathbf{S}))\right] \\ &\quad - 2\beta \left[\mathbf{C}(\overline{\mathbf{Y}}^\top \odot ((\mathcal{A} \times_k \mathbf{w})^\top - \mathbf{C}^\top \mathbf{S}))\right], \\ \frac{\partial \mathcal{F}}{\partial \mathbf{C}} &= -2\left[\mathbf{S}(\mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top \mathbf{C}))\right] \\ &\quad - 2\beta \left[\mathbf{S}(\overline{\mathbf{Y}} \odot (\mathcal{A} \times_k \mathbf{w} - \mathbf{S}^\top \mathbf{C}))\right], \end{aligned} \tag{24}$$

and for $\mathbf{w}$ the k-th element is given as

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial w_k} &= -2\alpha \sum_{ij} \left[y_{ij}(x_{ij} - \mathbf{a}_{ij}^\top \mathbf{w}) a_{ijk}\right] \\ &\quad - 2\beta \sum_{ij} \left[\bar{y}_{ij}(\mathbf{a}_{ij}^\top \mathbf{w} - \mathbf{s}_i^\top \mathbf{c}_j) a_{ijk}\right] + 2\gamma w_k. \end{aligned} \tag{25}$$

Experiments

Experimental Setup

Dataset: We perform our experiments on an outpatient service data set collected from the public hospital system of Shenzhen, a major city in southern China[1]. In the data set, service points are all public hospitals of Shenzhen, and customer points are all residential zones of Shenzhen. The data set contains all outpatient records of 321 public hospitals from January to December, 2014, whose patients came from 1343 residential zones. As a result, we finally obtain a 321 × 1343 patients volume matrix $\mathbf{X}$, with a high sparsity (the ratio of zero elements) equal to 94.87%. Note that we only select common diseases without hospitalization, such as cold, cough, and upper respiratory infection. For these diseases, the service qualities of all hospitals are very close so that their service ability differences can be neglected.

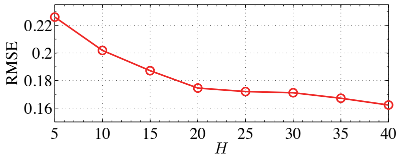


Figure 1: Influence of dimensionality of latent space.

Baselines: In the experiments, we adopt the following methods as baselines to the GR-NMF model.

- **Linear Regression (LR)**: This baseline uses a linear combination of the geographical features to predict customer volumes for every customer-service point pair. The objective function is: $\min\ \sum_{y_{ij}=1}(x_{ij} - \mathbf{w}^\top \mathbf{a}_{ij})^2$.
- **Singular Value Decomposition (SVD)**: This baseline decomposes the customer volumes matrix as an inner product of two matrices $(\mathbf{U}, \mathbf{V})$ that respectively project customer and service points into orthogonal latent spaces, and a diagonal component weight matrix $(\mathbf{\Sigma})$. The objective function is: $\min\ \|\mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)\|_F^2$.
- **Basic Non-Negative Matrix Factorization (bNMF)**: This baseline models the customer volumes matrix as an inner product of two nonnegative projection matrices that project customer and service points into nonnegative latent spaces. The objective function is

$$\begin{aligned} &\min\ \|\mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top \mathbf{C})\|_F^2, \\ &s.t.\ \mathbf{S} \geq 0, \mathbf{C} \geq 0. \end{aligned}$$

- **Sparse Non-Negative Matrix Factorization (sNMF)**: This baseline introduces a sparse prior for the projection matrices $\mathbf{S}$ and $\mathbf{C}$ compared with bNMF. The objective function is

$$\begin{aligned} &\min\ \|\mathbf{Y} \odot (\mathbf{X} - \mathbf{S}^\top \mathbf{C})\|_F^2 + \alpha\|\mathbf{S}\|_1 + \beta\|\mathbf{C}\|_1, \\ &s.t.\ \mathbf{S} \geq 0, \mathbf{C} \geq 0. \end{aligned}$$

Note that LR only uses explicit geographical correlations for modeling, while the rest three baselines use only the implicit correlations hidden inside $\mathbf{X}$. By comparing GR-NMF with these baselines, we can evaluate the benefit from coupling both implicit and explicit information.

Dimensionality of Latent Space: The selection of the dimensionality H in the latent space is essentially a tradeoff between model precision and computational complexity. A very low dimensionality will lead to a big lost of detailed information about the customer volume matrix and result in poor predictions. A very high dimensionality, however, will incur unaffordable computational costs. To set a proper H, therefore, we take a "warmup" experiment on a sampled

[1] https://en.wikipedia.org/wiki/Shenzhen

dataset with 10% of all hospitals, and watch the predictive precision of GR-NMF with H varying from 5 to 40. As can be seen from Fig. 1 , when $H > 20$ the increasing trend of the predictive precision of GR-NMF tends to be flattened. As a result, we set $H = 20$ as the default setting in the following experiments.

Evaluation Measure: We evaluate the predictive power of competing models by comparing the predicted customer volumes with the ground truths. Specifically, we adopt Root Mean Square Error (RMSE) as the evaluation measure, which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{\sum_{i,j} \bar{y}_{ij}} \sum_{y_{ij}=0} \left(x_{ij} - \hat{x}_{ij}\right)^2},$$

where $\hat{x}_{ij}$ is the predictive patients volume from residential zone c_j to the hospital s_i and $x_{ij} \in \mathbf{X}$ is the ground truth with $y_{ij} = 0$. This means that unsampled elements in $\mathbf{X}$ are to be served as ground truths for performance evaluation.

Performance in General Scenario

We here test the prediction power of GR-NMF on the patients volumes in $\mathbf{X}$. To this end, we first set the binary sampling matrix $\mathbf{Y}$ as an all-one matrix, and then do random sampling on it and turn the sampled elements to zeros. In this way, each (i, j) element of $\mathbf{X}$ with $y_{ij} = 0$ is treated as "unknown" patient volume (test data), and its true value x_{ij} is treated as the ground truth. We increase the sampling rate gradually from 10% to 50%, and watch the performance variation of the competing methods, as shown in Fig. 2(a). Note that for each sampling rate, we repeat the experiment 10 times and return the average result.

As can be seen from Fig. 2(a), while all the methods show declining performance with increasing unknown rates, GR-NMF consistently outperforms the competitors in all cases. This well demonstrates the advantage of GR-NMF gained by coupling both the implicit and explicit correlations between residential zones and hospitals. It is also interesting to see that LR outperforms all other baselines, which implies higher value of explicit information than implicit information for the customer volume prediction problem. This is somewhat unexpected when we recall the dominant position of matrix factorization methods in various matrix completion applications in recent years. Nevertheless, the implicit correlations yet take effects — that is why GR-NMF can beat LR, although the margins are small in all cases. Finally, the much worse performances of SVD testify the necessity of setting non-negative constraint to matrix factorization.

Performance in Location Selection Scenario

In this section, we further evaluate GR-NMF in a more difficult scenario of location selection. In this scenario, we assume the Shenzhen government is planning to build a new public hospital that can serve as many people as possible. As a result, we need to predict the patients volumes from different residential zones given each candidate hospital location and then select the one with a highest sum. To simulate this scenario, we do random sampling on the hospitals in the hospital set S, and then set the row vector $\mathbf{y}_i$ in all-one $\mathbf{Y}$ as a zero vector if hospital s_i is sampled. In this way, the location of each sampled hospital s_i is treated as a candidate location for the new hospital, and all the true values of $x_{ij}, j = 1, \dots, 1343$, are treated as the ground truths for the predicted patient volumes to s_i. We also vary the sampling rate from 10% to 50%, and for each setup we again repeat the experiment 10 times so as to return the average result.

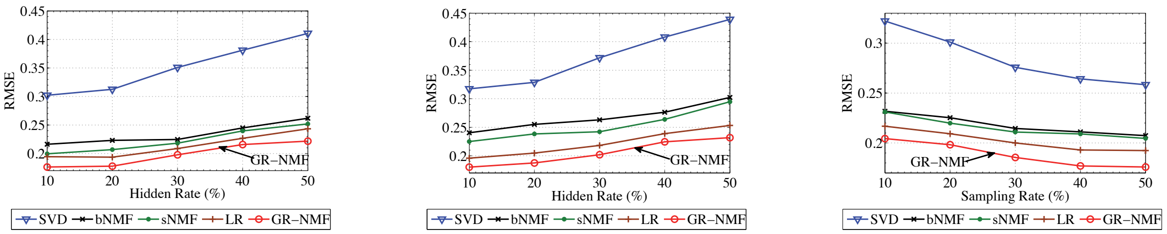
Fig. 2(b) shows the predictive performances of the competing methods. As can be seen from the figure, in general the prediction accuracies of all models exhibit very similar trends as the ones in the general scenario. Compared with the baseline methods, the GR-NMF model again achieves the best prediction performance, and the second best is still LR, which is followed by the three matrix factorization methods. This observation nicely verifies the finding in the general scenario; that is, while explicit information seems more valuable than implicit one in customs volume prediction, coupling both of them into GR-NMF can result in the best performance.

Furthermore, it is also interesting to compare the performances of the models in different scenarios. By putting Fig. 2(a) and Fig. 2(b) together for comparison, we can see that for each same sampling rate (which means the numbers of sampled y_{ij} elements are equal), all the methods seem perform worse in the location selection scenario, especially for the three matrix factorization baselines. This implies that to predict all the patient volumes to a hospital (*i.e.*, an empty row in $\mathbf{X}$) in the location selection scenario is more challenging. This phenomenon is in fact not surprising since matrix factorization models usually have problems in processing matrices with empty rows or columns, which is also described as the intractable *cold-start* problem from a recommender systems perspective. This, in turn, illustrates the importance of coupling implicit correlations with explicit geographical correlations; that is, the generation of explicit geographical features are immune to the cold-start problem. This well explains why GR-NMF and LR show more evident superiority to the other three baselines in the location selection scenario.

Performance in Market Investigation Scenario

We here continue to evaluate GR-NMF from a market investigation perspective. In this scenario, we assume an agency is investigating specific patient volumes of every residential zones to all hospitals in Shenzhen. However, given the limited budget, they can only obtain the patient volumes of some sampled hospitals and residential zones. We therefore need to predict the patient volumes from unsampled residential zones to unsampled hospitals. To simulate this scenario, we first set $\mathbf{Y}$ as an all-zero matrix, and then do random sampling on hospitals and residential points, respectively. If the hospital s_i is sampled, we turn all the elements in the i-th row of $\mathbf{Y}$ to one. Analogously, if the residential zone c_j is sampled, all the elements in the j-th column of $\mathbf{Y}$ will be set to one. We vary the sampling rate for both hospitals and residential zones from 10% to 50%, and return the average results of ten repetitions.

The comparison results of GR-NMF and baseline meth-

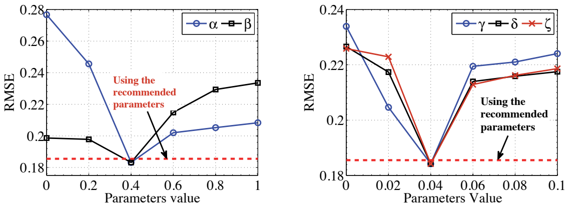


(a) General Scenario

(b) Location Selection

(c) Market Investigation

Figure 2: Comparison of prediction performances in different scenarios.



(a) α and β

(b) γ, δ and ζ

Figure 3: The setting of paraments.

ods are shown in Fig. 2(c). As shown in the figure, GR-NMF achieves the highest prediction power compared with the baseline methods. Moreover, the performance improvement of GR-NMF to LR is obviously higher than that in the location selection scenario. This is due to the fact that the sampling method for the market investigation scenario keeps more implicit correlation information between different hospitals. As a result, GR-NMF can obtain more implicit knowledge to improve the prediction performance.

Validation of Parameter Setting Method

When introducing the GR-NMF model above, we proposed an approximate method to set the parameters α, β, γ, δ, and ζ. We here verify its effectiveness by comparing it with a traversal method. In this traversal method, we alternately traverse the optimal value of each parameter while keeping other parameters fixed. In this way, we get a set of quasi-optimal parameters. The experiment is over the market investigation scenario and the sampling rate is 30%.

Fig. 3(a) shows a group of GR-NMF prediction results, with α and β varying from 0 to 1, and γ, δ and ζ set to the quasi-optimal values. Fig. 3(b) shows the results with quasi-optimal α and β but varying γ, δ and ζ. The red dash lines indicate our approximate method. As shown in the figures, the performance of our approximate method is just slightly worse than the optimal performance of the traversal method. Further considering the huge advantage in computational complexity of the approximate method, we adopt it as a default way to set parameters in practice.

Related Work

Customer volume prediction plays a key role in various application domains such as consumer market analysis (Fu et al. 2014b) and business location selection (Xu et al. 2016). In recent years, data-driven approaches become very popular in this area (Qu and Zhang 2013; Wang et al. 2015). For instance, Geo-spotting (Karamshuk et al. 2013) adopts nine geographic features including density, neighbors entropy, competitiveness, *etc.*, for mining online LBS data for optimal retail store placement. The study in (Li et al. 2015) adopts average travel time calculated from real traffic GPS trajectory data to optimize location selection of ambulance stations. The works in (Fu et al. 2014a; 2014b) exploit geographic contexts for real estate appraisal. The work in (Xu et al. 2016) mines service requirement information from LBS search engine query data for store location selection. Most of these approaches, however, are based on explicit geographic contexts, and few methods have the ability to exploit implicit correlations in data.

Matrix factorization is widely admitted as a very powerful model to mine hidden correlations, with explicit correlations often formulated as regularization to introduce external knowledge (Baltrunas, Ludwig, and Ricci 2011). For instance, PHF-MF (Cui et al. 2011b; 2011a) is a hybrid-factor matrix factorization model to mine latent correlations of a user-post matrix, and two regularization factors are introduced to fuse explicit user-user and post-post correlations into the model. The Geo-MF model (Lian et al. 2014) integrates hidden correlations in user visiting data with spatial clustering phenomena to recommend points of interests. The study in (Cheng et al. 2012) fuses hidden geographical influences with a social influence regulation to achieve personalized POI recommendation. The CLAR model (Zheng et al. 2010a) exploits location-activity information, location-feature, and activity-activity correlation for mobile recommendation, which is extended as UCLAF (Zheng et al. 2010b; 2012) with a tensor framework. This kind of models often treat implicit correlations as the most important knowledge and use explicit information as their regulations. In our GR-NMF model, however, both implicit and explicit correlations are treated equally and inter-calibrated on unobserved customer volumes, which enables the cold-start location selection nicely.

Conclusions

In this paper, a novel model called GR-NMF is proposed for integrating implicit footfall knowledge and explicit geographical knowledge into a unified matrix factorization framework for customer volume prediction. Experiments on a real-world outpatient dataset from Shenzhen City demonstrate the advantages of GR-NMF to the competitive baselines, which is even more evident in the location selection scenario with cold-start problem.

Acknowledgments

Dr. Jingyuan Wang's work was supported in part by NSFC (61572059, 61332018), the State's Key Project of Research and Development Plan (2016YFC1000307), and the Basic Science Research Program of Shenzhen (JCYJ20140828113329138). Dr. Junjie Wu's work was partially supported by NSFC (71322104, 71531001, 71471009, 71490723), National High Technology Research and Development Program of China (SS2014AA012303), National Center for International Joint Research on E-Business Information Processing (2013B01035), and Fundamental Research Funds for the Central Universities.

References

Athiyaman, A. 2011. Location decision making: the case of retail service development in a closed population. *Academy of Marketing Studies Journal* 15(1):87.

Bafna, S. 2003. Space syntax a brief introduction to its logic and analytical techniques. *Environment and Behavior* 35(1):17–29.

Baltrunas, L.; Ludwig, B.; and Ricci, F. 2011. Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, 301–304. ACM.

Cheng, C.; Yang, H.; King, I.; and Lyu, M. R. 2012. Fused matrix factorization with geographical and social influence in location-based social networks. In *Aaai*, volume 12, 17–23.

Cui, P.; Wang, F.; Liu, S.; Ou, M.; Yang, S.; and Sun, L. 2011a. Who should share what?: item-level social influence prediction for users and posts ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 185–194. ACM.

Cui, P.; Wang, F.; Yang, S.; and Sun, L. 2011b. Item-level social influence prediction with probabilistic hybrid factor matrix factorization. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 331–336. AAAI Press.

Fu, Y.; Ge, Y.; Zheng, Y.; Yao, Z.; Liu, Y.; Xiong, H.; and Yuan, J. 2014a. Sparse real estate ranking with online user reviews and offline moving behaviors. In *2014 IEEE International Conference on Data Mining*, 120–129. IEEE.

Fu, Y.; Xiong, H.; Ge, Y.; Yao, Z.; Zheng, Y.; and Zhou, Z.-H. 2014b. Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1047–1056. ACM.

Hernandez, T., and Bennison, D. 2000. The art and science of retail location decisions. *International Journal of Retail & Distribution Management* 28(8):357–367.

Jensen, P. 2006. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E* 74(3):035101.

Karamshuk, D.; Noulas, A.; Scellato, S.; Nicosia, V.; and Mascolo, C. 2013. Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 793–801. ACM.

Kolda, T. G., and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review* 51(3):455–500.

Li, Y.; Zheng, Y.; Ji, S.; Wang, W.; Gong, Z.; et al. 2015. Location selection for ambulance stations: a data-driven approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 85. ACM.

Lian, D.; Zhao, C.; Xie, X.; Sun, G.; Chen, E.; and Rui, Y. 2014. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 831–840. ACM.

Medda, F. 2012. Land value capture finance for transport accessibility: a review. *Journal of Transport Geography* 25:154–161.

Qu, Y., and Zhang, J. 2013. Trade area analysis using user generated mobile location data. In *Proceedings of the 22nd international conference on World Wide Web*, 1053–1064.

Simini, F.; González, M. C.; Maritan, A.; and Barabási, A.-L. 2012. A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.

Wang, J.; Gao, F.; Cui, P.; Li, C.; and Xiong, Z. 2014. Discovering urban spatio-temporal structure from time-evolving traffic networks. In *Asia-Pacific Web Conference*, 93–104. Springer.

Wang, Y.; Yuan, N. J.; Lian, D.; Xu, L.; Xie, X.; Chen, E.; and Rui, Y. 2015. Regularity and conformity: location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1275–1284. ACM.

Xu, Y., and Yin, W. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences* 6(3):1758–1789.

Xu, M.; Wang, T.; Wu, Z.; Zhou, J.; Li, J.; and Wu, H. 2016. Store location selection via mining search query logs of baidu maps. *arXiv preprint arXiv:1606.03662*.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2010a. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, 1029–1038. ACM.

Zheng, V. W.; Cao, B.; Zheng, Y.; Xie, X.; and Yang, Q. 2010b. Collaborative filtering meets mobile recommendation: A user-centered approach. In *AAAI*, volume 10, 236–241.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2012. Towards mobile intelligence: Learning from gps history data for collaborative recommendation. *Artificial Intelligence* 184:17–37.

Zipf, G. K. 1946. The P_1P_2/D hypothesis: on the intercity movement of persons. *American sociological review* 11(6):677–686.