SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2017, Vol. 60 092103:1–092103:14
doi: 10.1007/s11432-015-0905-6

A Gaussian copula regression model for movie box-office revenues prediction

Junwen DUAN, Xiao DING & Ting LIU*

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001*, China*

Received April 11, 2016; accepted August 15, 2016; published online April 25, 2017

Abstract In this article, we revisit the task of movie box-office revenues prediction using multi-type features. The movie box-office revenues are affected by numerous factors. Previous work with discriminative models assumes these factors are identically and independently distributed. The correlations between these factors are rarely considered, which limited the performances of discriminative models in this task. To address these problems, we investigate a novel Gaussian copula regression model. Based on this model, we do not need to make any prior assumptions about the marginal distributions of the features. In particular, we perform a cumulative probability estimation on each of the smoothed features. The estimation learns the marginal distributions and maps all features into a uniform vector space. Sequentially, we bridge the marginal distributions with a copula function to create their joint distribution, and learn the dependency structure between them. Moreover, we propose a computational-efficient approximate algorithm for responsible variable inference. Experimental results on two movie datasets from Chinese and U.S. market show that our approach outperforms strong discriminative regression baselines.

Keywords Gaussian copula, movie box-office revenue, multi-variate regression, text regression, social media

Citation Duan J W, Ding X, Liu T. A Gaussian copula regression model for movie box-office revenues prediction. Sci China Inf Sci, 2017, 60(9): 092103, doi: 10.1007/s11432-015-0905-6

1 Introduction

Statistical analysis of historical Chinese movie market data reveals that, of all the movies released in the first half of 2013, only a fraction make profits [1]. Thus, designing effective models to predict the future market performances of upcoming movies, will obviously benefit the producers, sponsors, and theaters.

Previous work has investigated numerous predictive indicators for the task, but the results are still not satisfying. The uncertainties of movie box-office revenues lie throughout the productions and distributions of movies, ranging from the actors, directors and budgets before the releases and word-of-mouth marketings, screen arrangements after the releases.

Traditionally, analysts make predictions based on easily accessible movie meta-data, such as genres, budgets, or by referring to market performances of similar movies in history. Although critic reviews and blog contents are abundantly available, they are not exploited until recent advances in natural language

* Corresponding author (email: tliu@ir.hit.edu.cn)

processing. The sentiment analysis techniques are then used in the task. Recently, the advent of social media has brought massive user generated content and user social network [2]. The social media "big data" have immediately shown distinctive power in user intention prediction [3, 4], presidential election prediction [5], stock market prediction [6, 7] and also movie box-office revenue predictions [1,8]. The scope of predictions is significantly broadened. However, challenges such as data of heterogeneity for prediction models are acompanied with the social media "big data".

To model heterogeneous data, linear regression, support vector regression and neural network model are commonly used. They are easy to train and highly descriptive. However, they also suffer some limitations. On the one hand, discriminative models consider features independently of each other, which rarely use the underlying dependency structure between the features. Consequently, their applications to cases where features are highly correlated with each other are thus limited. For the task of box-office revenues prediction, the wordings in movie reviews are highly correlated with genres of movies. For instance, the word "romantic" is more frequent in reviews of movies with genre "Romance" than genre "War". Thus, jointly modeling meta-data and user generated content is very important in box-office revenues prediction. On the other hand, historical movie market data suggest that features in movie meta-data subject to different statistical distributions [9]. For example, movie box-office revenues subject to a Pareto law distribution, while the number of theaters in which a movie is shown follows a bimodal distribution.

To address the above challenges, we investigate a novel Gaussian copula regression model for this task. The theory of copula [10] was first proposed in 1959. Owing to its capability in multi-variate modeling, it was later widely used in quantitative finance [11]. Recently, researchers in machine learning domain begin to exploit it for information retrieval [12], multi-modal regression [13] and many other tasks [14, 15]. The key idea underlying copula theory is to decompose multi-variate joint distribution to univariate marginal distributions and encode the dependency structure between the margins into a correlation matrix. In this article, we first perform a cumulative probability estimation to each of the normalized features, to obtain their arbitrary marginal distributions. We then connect the marginal distributions with the Gaussian copula function and learn the dependency structure between the margins. This process helps create the joint distributions of the input features. We further propose an approximate inference algorithm for response variable inference in high dimensions. The algorithm is computational efficient and can easily scale to high dimensions.

To evaluate the effectiveness of our approach, we collect two datasets with rich features from U.S. and Chinese market. We compare our approach with two state-of-the-art systems in movie box-office revenues prediction. In the baseline systems, linear regression with elastic net regularization and support vector regression with both linear and non-linear kernels, are used. The experimental results on the two datasets highlight the effectiveness of our approach under various settings. Feature combination results indicate that critic reviews on news media and user posts on social media are important indicators for predicting movies' market performances. The main contributions of this article are as follows.

- We propose using a novel Gaussian copula regression model for movie box-office revenues prediction task, which proves to be effective and efficient.
- We propose an approximate inference algorithm for multivariate copula regression which is scalable.
- We investigate the critics' movie reviews and user activities before movies' official releases and study their impacts on movie box-office revenues.

The rest of this paper is organized as follows. We summarize related work in Section 2. In Section 3, we first give a brief introduction to the theory of copula and then describe our proposed model in detail. Experimental details and analysis are shown in Section 4. We conclude this article in Section 5.

2 Related work

Owing to its entertaining and profitable characteristics, movie box-office revenues prediction has attracted intensive interests from areas of statistics, machine learning and natural language processing. According to the data sources used for prediction, previous work can be classified into three classes, i.e., meta-data based, text-based and behavior-based.

Metadata-based. Traditionally, researchers focus on modeling structured meta-data of movies, such as their genres, budgets. The meta-data of movies are easily accessible before their official releases. Sharda et al. [16] were among the first to model movie meta-data with a neural network in the revenue prediction task. Instead of predicting the exact revenues, they formulated it as a multi-classification problem. Movies were classified into nine classes by their gross box-office revenues, ranging from "flop" to "blockbusters". Zhang et al. [17] applied a back-propagation neural network model with similar settings. This line of work are early trials on this task. Although their performances are not very satisfying, they have discovered numerous strong predictive indicators.

Text-based. Today, movie relevant user generated data are abundantly available on public media. However, their strong correlations with the box-office revenues were not explored until recently, especially after the advances in natural language processing. Mishne et al. [18] analyzed bloggers' sentiments towards the movies and found that positive sentiments are good indicators of the movies' future market performances. Zhang et al. [19] incorporated both movie meta-data and sentiments in movie-associated news in their model, which achieved a remarkable improvement over previous approaches. Joshi et al. [20] formulated the task as a text regression problem. They extracted both lexical and syntactic features from the critics' movie reviews and employed a linear regression model with elastic net as regularization. Text features begin to play a role alternative to movie meta-data. However, the textual features in these models are used independently of each other, the correlations among these covariates are less considered.

Behavior-based. The advent of social media has brought rich user activity data, which are successfully applied to numerous prediction tasks, such as presidential election prediction [5] and stock market prediction [6]. Liu et al. [1] were among the first to investigate movie-specific user activities on social media. By analyzing users' intention towards the movies, tracing the number of movie mentions in a particular time window and detecting user purchase intentions towards the movies, they achieved state-of-the-art performance on this task. Mestyán et al. [21] addressed the problem from a different prospective, they traced the users' edit and page view activities on the movie-specific Wikipedia pages and recoded the frequencies and time intervals. They found strong correlations between these activities and gross box-office revenues. In particular, by tracing such activities, they were able to make predictions one month ahead of the movies' official releases.

3 Problem statement

In this article, we revisit the task of predicting the box-office revenues of upcoming movies with multi-modal data. A more formal definition is given as follows. For each movie i in the n training examples, we have data in the form of $\{m_i, d_i, y_i\}$, in which m_i are the relevant meta-data, and d_i is a collection of user generated content from public media corresponding to the movies, and y_i is the ground-truth weekend or gross box-office revenue. We can extract syntactic, semantic and even user behavior information $t_i = \{z_1, z_2, \ldots, z_n\}$ towards the movie i from d_i. As summarized in Eq. (1), our ultimate goal is to automatically predict the box-office revenue $\hat{y_i}$ given the metadata m_i and feature set t_i.

$$\hat{y_i} = f(m_i, t_i; d_i). \tag{1}$$

We summarize the framework of our approach in Figure 1, which consists of four sequential steps.

Representation. Feature set is extracted from the structured meta-data and unstructured user generate content. Textual features are temporarily represented in the classical count-based form.

Transformation. We first normalize the features so as to accelerate the computation and partly relieve the increasing data sparsity in high dimensions. Afterwards, we derive the marginal distributions of the stochastic variables and transform the count-based representations into dense probability vectors. The details are presented in Subsection 4.2.

Dependency learning. We connect the marginal distributions with copula function and learn the underlying dependency structure between them.

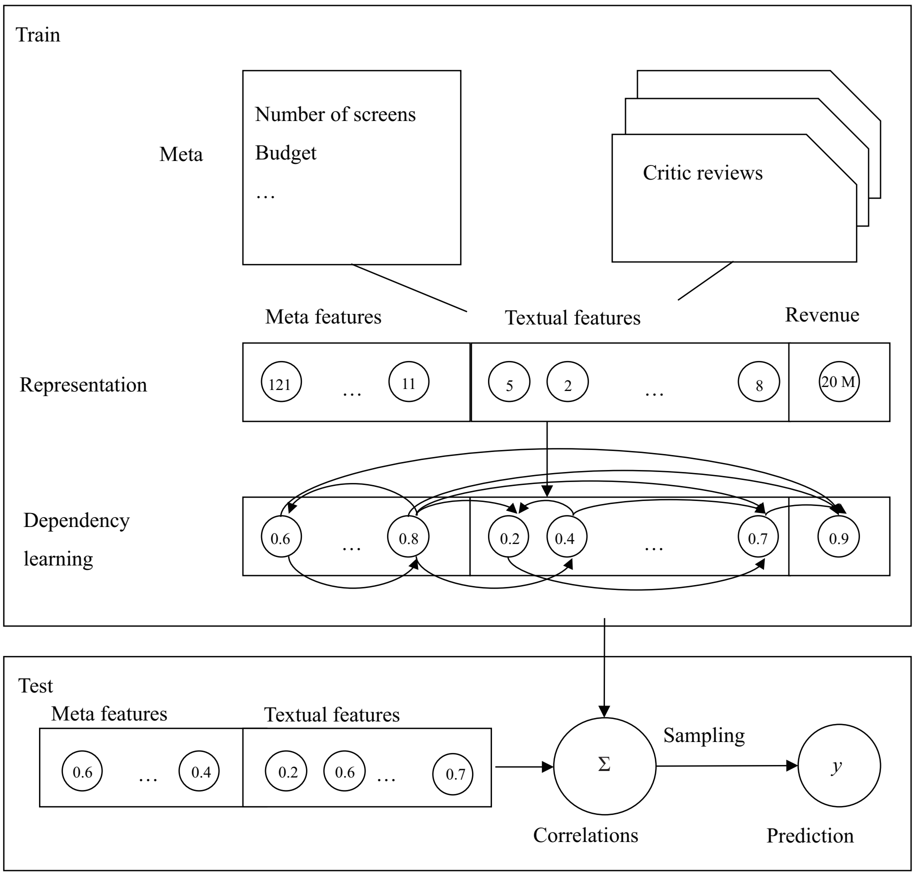

Figure 1 The framework of our proposed method.

Prediction. With the learned dependency structure and the approximate inference algorithm, we make predictions of the box-office revenues on the test set.

4 Copula regression for prediction

Discriminative regression models, such as linear regression and support vector regression (SVR), have been successfully applied to numerous regression tasks, in some they have achieved state-of-the-art performance. However, in the two models, the correlations among the covariates have rarely been considered. On the contrary, variables are modeled independently of each other. Obviously, these limitations will restrict the generalizations of the models to cases where variables are strongly correlated with each other. Another problem with traditional regression model is to cope with large numbers of features. Linear regression often resorts to $L1$ or $L2$ norm to eliminate the duplicated features while SVR is not immune to either large numbers of training data or large numbers of features.

The Gaussian copula regression model addresses the problems from a quite different prospective. It can model the correlations between stochastic variables without depending on their priors. It scales to high dimensions at inexpensive computational costs. Even with these settings, the model is still expressive. Since the theory of copula is rather new to natural language processing researchers, we give a brief introduction to it.

4.1 A brief introduction to copula theory

The theory of copula was first introduced by Sklar in 1959 [10]. It was then known as a family of distribution functions in probability theory. The successful applications of copula theory in quantitative finance [11], civil engineering [22] have evoked interests from many other areas. Until recently, researchers in machine learning begin to explore it for information retrieval [12], text regression [23] and multi-modal data modeling [13].

Sklar's Theorem states that any multi-variate joint distribution could be further decomposed into univariate marginal distributions and a copula function that describes the dependency structure between the stochastic variables. A more formal definition is summarized in Sklar's Theorem [10].

Theorem 1 (Sklar's theorem)**.** Suppose that there are n random variables $X_1, X_2, \ldots, X_n$. Let $F(x_1, x_2, \ldots, x_n)$ be their n-dimensional cumulative distribution function and $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ be their corresponding marginal cumulative distribution functions. Then, if the marginal distributions are continuous, there exists a unique Coupla function C, such that

$$F(x_1, x_2, \ldots, x_n) = C[F_1(x_1), F_2(x_2), \ldots, F_n(x_n)], \tag{2}$$

in which $F_1(x_1), \ldots F_n(x_n)$ are uniformly distributed in $[0, 1]$. Specially, C is independent of the margins when the marginal distributions are continuous [25]. Based on the definition, C can be considered as a function that maps $[0, 1]^n$ to $[0, 1]$.

The inverse of Sklar's Theorem, which states marginal distributions and copula could define a multi-variate joint distribution, is also true. For example, we can define a bivariate joint distribution by

$$C_\sigma[F_1(x_1), F_2(x_2)], \tag{3}$$

in which the correlations among variables are encoded in the σ. The inverse copula theory enables us to construct the multi-variate joint distribution with the marginal distributions and the copula function [26]. It is particularly useful in cases where one can hardly derive the multi-variate joint distribution directly while the marginal distributions are easily accessible.

The copula function is central to the theory. There are various copula function families available currently, including the Clayton, Gumbel, Frank and Elliptical. They differ in the manner of bridging the dependency structure between marginal distributions.

The copula function we use throughout this paper is called Gaussian copula:

$$C_{\text{Gauss}}(x_1, x_2, \ldots, x_n) = \Phi_{\mathbf{\Sigma}}(\Phi^{-1}(x_1), \Phi^{-1}(x_2), \ldots, \Phi^{-1}(x_n)). \tag{4}$$

As a member of the Elliptical copula family, it is also among the most widely used copula functions.

From above introductions, we can find that the copula theory is particularly good for multi-variate and multi-modal data modeling. First, copula functions can construct complex multi-variate joint distribution with simple univariate marginal distributions. Second, the inputs of copula functions are from the uniform space $[0, 1]^n$, variables with arbitrary distributions can perform space transformations through techniques like probability integral transformation.

4.2 Cumulative distribution probability estimation

A key challenge in jointly modeling meta-data features and textual features is that meta-data features are usually dense and continuous while textual features are often sparse and discrete. Our solution to the heterogeneous data involves a two-step process. We first normalize the data to $[0, 1]$, so as to accelerate the computation as well as partly relieve the data sparsity in high dimensions. Afterwards, we perform a non-parametric kernel-based cumulative distribution estimation on the normalized data and thus obtained their estimated marginal distributions. The cumulative distribution $\hat{F}(x)$ is the integral of its continuous probability density $\hat{f}(x)$:

$$\hat{F}(x) = \int_0^x \hat{f}(x)\mathrm{d}x. \tag{5}$$

Thus, the key to the distribution estimation is to estimate the probability density. In this article, we adopt the commonly used non-parametric kernel-based approach for probability density estimation.

For a sequence with m observed cases, the kernel-based probability density estimation can be summarized as

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} K_h(x, x_i), \tag{6}$$

in which $K_h(\cdot)$ is the kernel function and h is the bandwidth parameter.

Due to the data heterogeneity, we apply Gaussian kernel (Eq. (7)) to the meta-data features and Box kernel (Eq. (8)) to the textual features:

$$\text{Gaussian kernel} \qquad K_h(x, x_i) = \exp\left(-\frac{||x - x_i||_2^2}{2\sigma^2}\right), \tag{7}$$

$$\text{Box kernel} \qquad K_h(x, x_i) = \begin{cases} \frac{1}{2}, & |x - x_i| \leqslant 1, \\ 0, & |x - x_i| > 1. \end{cases} \tag{8}$$

The estimations are implemented using the ksdensity package in Matlab with the default bandwidth estimation algorithm.

By performing a non-parametric distribution estimation, we derive the marginal distributions. All the features are mapped into a uniform vector space $[0, 1]$ regardless of their arbitrary distributions. The process, therefore, enables us to cope with high dimensions and arbitrary distributions. The margin are then connected by the copula functions in the following step.

4.3 Copula parameter estimation

Let z_i be the estimated marginal distribution $\hat{F}_i(x_i)$. Based on the inverse Sklar's Theorem, we can construct the joint distribution $F(z_1, z_2, \ldots, z_n)$ through (10),

$$\begin{aligned} F(z_1, z_2, \ldots, z_n) &= C_{\mathbf{\Sigma}}(z_1, z_2, \ldots, z_n, z_y) \qquad (9) \\ &= \Phi_{\mathbf{\Sigma}}(\Phi^{-1}(z_1), \Phi^{-1}(z_2), \ldots, \Phi^{-1}(z_n), \Phi^{-1}(z_y)). \qquad (10) \end{aligned}$$

In (10), $z_1, z_2, \ldots, z_n$ and z_y are the marginal distributions of the covariates and the response variable respectively. The response variable y is the box-office revenue in our case. $\Phi_{\mathbf{\Sigma}}$ is the standard multivariate Gaussian distribution with zero means and $\mathbf{\Sigma}$ variance. Thus, the correlations between the marginal distributions are encoded in $\mathbf{\Sigma}$. Φ^{-1} is the inverse cumulative distribution function of standard Gaussian. The problem of learning the dependency structure among the marginal distributions is thus reduced to estimate the covariance matrix $\mathbf{\Sigma}$. We solve the problem by performing a widely used maximum likelihood estimation. The likelihood is in the form of (11).

$$p(\Delta|\mathbf{0}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \cdot \det \mathbf{\Sigma}}} \exp\left\{ -\frac{1}{2}\Delta^\top \cdot (\mathbf{\Sigma}^{-1} - \mathbf{I}) \cdot \Delta \right\}, \tag{11}$$

where

$$\Delta = \left\{ \begin{matrix} \Phi^{-1}(z_1) \\ \cdots \\ \Phi^{-1}(z_n) \\ \Phi^{-1}(z_y) \end{matrix} \right\}. \tag{12}$$

There are two tricks [23] in learning a well-generalized $\mathbf{\Sigma}$. The first is to reestimate the $\mathbf{\Sigma}$ with random generated cases. Another is the $\mathbf{\Sigma}$ matrix has to be positive definite, while variables that follow similar distributions may easily break such property. Thus, we add tiny Gaussian noise ϵ to the matrix to maintain the property.

4.4 Inference

The ultimate objective is to infer the response variable through the learned dependency structure between covariates and the response variable. However, previous approaches concentrate on modeling the correlations between bi-variates. In bivariate copula regression, it is easy to infer $F_y(y)$ from $F(x, y)$ by maximizing the conditional expectation $E(F_y(y)|F_x(x); \sigma)$, where σ is the bivariate covariance matrix. However, it is complex and intractable in multivariate distributions. Wang et al. [23] propose to sample y to maximize the joint probability density $f(x_1, x_2, \ldots, x_n, y)$, however, it is not working well in our case. Therefore, we propose to approximate the y by sampling in $[0, 1]$ to maximize the conditional density $f(y|x_1, x_2, \ldots, x_n)$.

In copula theory, the conditional density $f(y|x_1, x_2, \ldots, x_n)$ can be further decomposed to the form of (13), where $c(\cdot)$ and $f(\cdot)$ are the joint copula density and univariate probability density respectively. Because the covariates $(x_1, x_2, \ldots, x_n)$ are fixed across the same observation, the conditional density is therefore proportional to (14),

$$f(y|x_1, x_2, \ldots, x_n) = f(y) \cdot \frac{c(x_1, x_2, \ldots, x_n, y)}{c(x_1, x_2, \ldots, x_n)} \tag{13}$$

$$\propto f(y) \cdot c(x_1, x_2, \ldots, x_n, y). \tag{14}$$

Consequently, our objective in the approximate inference algorithm is to sample y that maximizes

$$\hat{y} = \operatorname*{argmax}_{y \in [0,1]} f(y) \cdot c(x_1, x_2, \ldots, x_n, y), \tag{15}$$

where

$$c(x_1, x_2, \ldots, x_n) = \Phi_{\mathbf{\Sigma}}(\Phi^{-1}(x_1), \Phi^{-1}(x_2), \ldots, \Phi^{-1}(x_n)) \cdot \prod_{i=1}^{n} \frac{1}{\Phi(\Phi^{-1}(x_i))} \tag{16}$$

$$= C_{\mathbf{\Sigma}}(x_1, x_2, \ldots, x_n) \cdot \prod_{i=1}^{n} \frac{1}{\Phi(\Phi^{-1}(x_i))}. \tag{17}$$

For each of the test cases, we only have to sample 100 times to obtain the optimal y. It therefore significantly reduces the computational complexity and enables the approximate algorithm easily generalize to high dimensions. Ultimately, we apply the inverse cumulative distribution function to $\hat{F_y}(y)$ to obtain the predicted value.

4.5 Algorithm implementation

We present the pseudo code of the Gaussian copula regression model in Algorithm 1. The implementation is composed of the following four main components.

(1) Normalize and apply the kernel-based cumulative distribution estimation to each dimension of the training and testing examples respectively to derive their uniform marginal distributions.

(2) Learn the dependency structure (the parameter $\mathbf{\Sigma}$) between the uniform marginal distributions by means of maximum likelihood estimation.

(3) Based on the dependency structure among the stochastic variables, sample the response variable to maximize the conditional density and obtain the optimal approximate inference of the response variable.

(4) Infer the exact revenue by inverse cumulative distribution function.

5 Experiments

5.1 Datasets

We evaluate our proposed approach on two different movie datasets, i.e., Datasets S1 and S2. Dataset S1 is released by Joshi et al. [20] with 1718 movies between 2005 and 2009 on U.S. market. While Dataset S2 is collected by ourself with 188 movies between 2012 and 2014 on Chinese market.

Algorithm 1 Gaussian copula regression algorithm

```
Require:
    % Training and testing data
    Training data: $X^{\rm tr}_{\rm meta}, X^{\rm tr}_{\rm text}, y^{\rm tr}$;
    Testing data: $X^{\rm te}_{\rm meta}, X^{\rm te}_{\rm text}, y^{\rm te}$;
    Output: the predicted value $y$;
    % Normalize the data
normalize($X^{\rm tr}_{\rm meta}, X^{\rm te}_{\rm meta}$);
normalize($X^{\rm tr}_{\rm text}, X^{\rm te}_{\rm text}$);
normalize($y^{\rm tr}, y^{\rm te}$);

    % Kernel-based CDF estimation
$U^{\rm tr}_{\rm meta}$ = GaussianKernel($X^{\rm tr}_{\rm meta}, X^{\rm tr}_{\rm meta}$);
$U^{\rm tr}_{\rm text}$ = BoxKernel($X^{\rm tr}_{\rm text}, X^{\rm tr}_{\rm text}$);
$U^{\rm te}_{\rm meta}$ = GaussianKernel($X^{\rm tr}_{\rm meta}, X^{\rm te}_{\rm meta}$);
$U^{\rm te}_{\rm text}$ = BoxKernel($X^{\rm tr}_{\rm text}, X^{\rm te}_{\rm text}$);
$U^{\rm tr}_{y}$ = GaussianKernel($y^{\rm tr}, y^{\rm tr}$);
$U^{\rm te}_{y}$ = GaussianKernel($y^{\rm tr}, y^{\rm te}$);
$Z^{\rm tr}$ = GaussianInverseCDF($U^{\rm tr}_{\rm meta}, U^{\rm tr}_{\rm text}, U^{\rm tr}_{y}$);
$\Sigma$ = MLE($Z^{\rm tr}$);

    % Approximate inference
for $i = 1 \to m$ testing examples do
    max_density = 0;
    probability = 0;
    for $k = 0.01 \to 1$ do
        dens = Density($k$) $*$ CopulaDensity($U^{\rm te}_{\rm meta}, U^{\rm te}_{\rm text}, k$);
        if dens $\geqslant$ max_density then
            max_density = dens;
            probability = $k$;
        end if
    end for
    $y$ = InverseCDF($y^{\rm tr}$, probability);
end for
```

Table 1 Train-, dev- and test-split and details of Dataset S1

	Train	Dev	Test	Total
Movies	1147	317	254	1718
Reviews	4818	1268	1042	7044

Dataset S1 contains structured movie metadata and critic reviews collected before movies' official releases from seven main-stream news media, including New York Time and Boston Globe. There is at least one review for each movie in the dataset. For comparison, we split of the train-, dev- and test-set identically with Joshi et al. [20]. The details of the split are shown in Table 1.

Dataset S2 is made up of meta-data and behavior data, namely, number of screens (NS), purchase intention rate (PIR) and post rate (PR). The details of each feature type are illustrated in Subsection 5.2. We follow previous work (Liu et al. [1]) and collect the PIR and PR in seven days before the movies' official releases from Sina Weibo[1]. The NS, weekend and gross box-office revenues are obtained from Wangpiao[2]. PIR and PR are calculated based on user discussions about the movies on social media, we exclude movies with PR less than threshold 20. Movies that have not arouse hot discussions on social media are thus eliminated. For Dataset S2, all the experimental results showed in this section are achieved under 5-fold cross validations.

[1] Sina Weibo. http://weibo.com. A Twitter-like social media platform in China.

[2] Wangpiao. http://www.wangpiao.com. Movie box-office data provider.

Table 2 Metadata features extracted from Datasets S1 and S2

Feature	Possible value	Description
Num_of_Screen	Positive integer	Screens scheduled for the movie
Log_Budget	Positive real number	The logarithm of movie budget
Running_time	Positive integer	Running time of the movie in minutes
Num_of_Oscar_Winning_Actor	Positive integer	Number of Oscar-winning actors
Num_of_High_Gross_Actor	Positive integer	Number of high-gross actors
Oscar_Winning_Director_Present	Boolean	Presented by a Oscar-winning director

5.2 Feature set

Our feature set is made up of three types of features, i.e., meta-data features, textual features and behavior features.

Metadata features. The meta-data features are extracted from structured data of movies, which are summarized in Table 2. Dataset S1 contains all the features listed in Table 2, while Dataset S2 only includes the Num_of_Screen feature.

Textual features. To better represent the information encoded in texts, we use a variety of shallow and deep textual features. Specially, we stem and down-case the words in all our features. Each of the textual features are listed as follows.

(1) ***n*-gram features.** We extract unigrams, bigrams, trigrams from the movie reviews. We apply a stop word list from NLTK [27] to eliminate all the meaningless words.

(2) **Part-of-speech (POS) features.** We consider a bidirectional POS-tagging from Stanford part-of-speech tagger [28] to capture the shallow syntactic representation.

(3) **Dependency relation features.** The dependency relations are extracted to characterize the syntactic dependency among the keywords. We use the Stanford parser [29] and represent dependency relations as triples (e.g., amod(movies, good)).

Behavior features. Behavior features are extracted from user activities on social media. In this article, Post-rate and Purchase intention rate are considered.

(4) **Post rate (PR).** This index measures the popularity of a movie on the social media. A formal definition is the number of tweets N_{total} that have mentioned a particular movie in a given time window T_{window}. We set the time window to seven days before the movie's official release. We retrieve the tweets via the Sina Weibo API[3] with a keyword filter.

$$\text{PR} = \frac{|N_{\text{total}}|}{|T_{\text{window}}|}. \tag{18}$$

(5) **Purchase intention rate (PIR).** This indicator measures the user expectation about a movie. A formal definition is given a collection of tweets associated with a movie, the number of tweets $N_{\text{intention}}$ in a given time window T_{window} that have shown intention to see the movie. The intention detection algorithm extracts a variety of features from tweets, such as bag-of-words, emoticon, url and user mentions. Refer to Liu et al. [1] for a detailed description about the inention detection algorithm.

$$\text{PIR} = \frac{|N_{\text{intention}}|}{|T_{\text{window}}|}. \tag{19}$$

We extract (1)–(3) from Dataset S1 and (4), (5) from Dataset S2. We do not use all features in (1)–(3). Instead, we pick the most frequent 100 features from each type which makes a total of 500 textual features.

5.3 Baselines

The baselines are two state-of-the-art discriminative regression models, the linear regression

$$\hat{y} = \beta_0 + x^\top \beta, \tag{20}$$

[3] http://api.weibo.com.

Table 3 Performance of our approach compared to baselines in weekend movie box-office revenue prediction on Dataset S1

Model	Weekend	
	MAE (million in U.S. dollars)	r
Linear regression (Joshi et al. [20])	5.738	0.812
SVR (linear kernel, Liu et al. [1])	6.104	0.696
SVR (RBF kernel, Liu et al. [1])	5.348	0.803
Our method	**5.245**	**0.879**

$$\hat{\theta} = \underset{\theta=(\beta_0,\beta)}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\beta_0 + x_i^{\top}\beta))^2 + \lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right). \tag{21}$$

and support vector regression (SVR). To enhance the prediction accuracy of linear regression, Joshi et al. [20] performed a feature selection and regularization with elastic net [30] (Eq. (21)), which is a mixture of both $L1$ norm (lasso regression) and $L2$ norm (ridge regression). The parameter α in Eq. (21) copes with the trade-off between $L1$ norm and $L2$ norm. Joshi et al. have released their model and data[4)], therefore the results reported on Dataset S1 are reproductions of their model.

Liu et al. [1] employed support vector regression [31] with both linear and non-linear kernels. The standard form of SVR is to solve the approximate problem of (22), in which $K(\cdot)$ is the kernel function and b is the bias. The α^* and α are the coefficients which can be obtained by solving the optimization of (23), where C is a regularization constant and ϵ controls the tolerance to training errors.

$$\hat{y} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(x, x_i) + b, \tag{22}$$

$$\langle \alpha^*, \alpha \rangle = \underset{\langle \alpha^*, \alpha \rangle}{\operatorname{argmin}} \left\{ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max(0,\ |\hat{y} - y| - \epsilon) \right\}. \tag{23}$$

The two models are good at capturing the linearity and non-linearity among the features and prove to work well on a variety of regression tasks.

5.4 Evaluation metrics

We adopt the same evaluation metrics with previous work (Joshi et al. [20] and Liu et al. [1]) for comparisons. They are mean absolute error (MAE) and Pearson's correlation (r).

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{24}$$

is a measurement of the bias between the predicated value $\hat{y}$ and the groud-truth value y_i, for which the smaller one is better. We do not consider other similar metrics like relative absolute error (RAE) and root mean squared error (RMSE) because MAE can interpret the prediction precision in a more intuitive manner in this task. The lower MAE value indicates the better experimental result.

Pearson's r measures the correlations between the predicted values $\hat{y}$ and the ground-truth values y. The value of Pearson's r falls in the range $[-1, 1]$, in which $r > 0$ is positively correlated, $r < 0$ is negatively correlated and $r = 0$ is non-correlated.

5.5 Comparison to the baselines

We compare our proposed Gaussian copula regression model to the baselines with all features on the two datasets. The experimental results on the two datasets are shown in Tables 3 and 4, respectively.

[4)] http://www.cs.cmu.edu/ ark/movie$-data/.

Table 4 Performance of our approach compared to baselines in weekend and gross movie box-office revenue prediction on Dataset S2

Model	Weekend		Gross	
	MAE (million in rmb)	r	MAE (million in rmb)	r
Linear regression	6.50	0.868	81.7	0.787
SVR (linear kernel)	6.06	0.861	69.9	0.776
SVR (RBF kernel)	7.72	0.858	79.5	0.773
Our Method	**5.26**	**0.907**	**54.7**	**0.824**

In Dataset S1, compared with linear regression model, our approach achieves a nearly 10% relative error reduction in MAE and is comparable with the support vector regression model. For the metric Pearson's r, our method outperforms all the baselines by a wide margin.

In Dataset S2, we outperform the baselines on weekend and gross box-office revenue prediction task and obtain a relative error reduction of 13% and 21% on weekend and gross box-office revenue prediction tasks respectively. The MAE in gross box-office revenue prediction task are almost ten-fold over the MAEs in weekend task among all methods. A movie usually screens on the theaters for more than one month. Many more other factors may continue to affect its final revenue in the meantime. The results further support that predicting the gross box-office revenue is even more challenging.

The performances of SVR with RBF kernel and linear kernel are inconsistent across the two datasets. RBF kernel outperforms linear kernel in Dataset S1, while linear kernel beats RBF kernel in Dataset S2. The inconsistency is probably caused by the heterogeneity of data. (1) The two datasets differ in feature types and feature size. Dataset S1 has more than 500 features, most of which are sparse while Dataset S2 has less than 5 features, all of which are dense. The linear kernel does not scale to large numbers of features. (2) The two datasets are different in data size. Dataset S1 has more than 1000 observations for training while Dataset S2 only has 188 samples in total, but the RBF kernel is much more sensitive to training data size than linear kernel. The capability of our approach in coping with heterogeneous data help achieve consistent performances on both datasets.

5.6 Feature combination

To compare the performance of different features, we experiment on different feature combinations. The results are shown in Table 5 and Figure 2.

For Dataset S1, our model can achieve comparable performance with the best baseline when we only use meta-data features. While the performance drops dramatically if only textual features are used. Unigrams contribute the most while the dependency relations are less useful. Adding textual features to meta-data features helps further boost the performances, although the improvements vary. Our approach achieves the best performance when both meta-data features and textual features are combined.

For Dataset S2, in the weekend revenue prediction task, the MAE is over 10 million rmb when only the meta-data feature — number of screens is used. However, by adding either post rate or purchase intention rate feature, the MAE drops dramatically. Post rate feature is much better than the purchase intention rate. The performance also reaches the best when we combine all the features.

The feature combination experiments indicate that integrating different forms of features, such as textual features and user behavior features may help greatly improve the performance of the proposed model.

5.7 Qualitative analysis

We qualitatively analyze the wordings in the critics' reviews and their impacts on the box-office revenues. We locate the features that have exhibited highly positive correlations with the box-office revenues by referring to the covariance matrix $\mathbf{\Sigma}$. The matrix stores the dependency structure between the stochastic variables. We list the top 30 features from Dataset S1 in Table 6. To be interpretive, we only show meta-data and n-gram features.

Table 5 Performance of our approach on Dataset S1 combining metadata and textual features

Feature sets	MAE (million in U.S. dollars)	Pearson's r
Metadata feartures	5.793	0.859
+ Unigrams	5.451	0.867
+ Bigrams	5.526	0.873
+ Trigrams	5.654	0.856
+ Part-of-speech	5.675	0.863
+ Dependency-relation	5.679	0.859
Textual features	7.950	0.598
Metadata + textual features	5.245	0.879

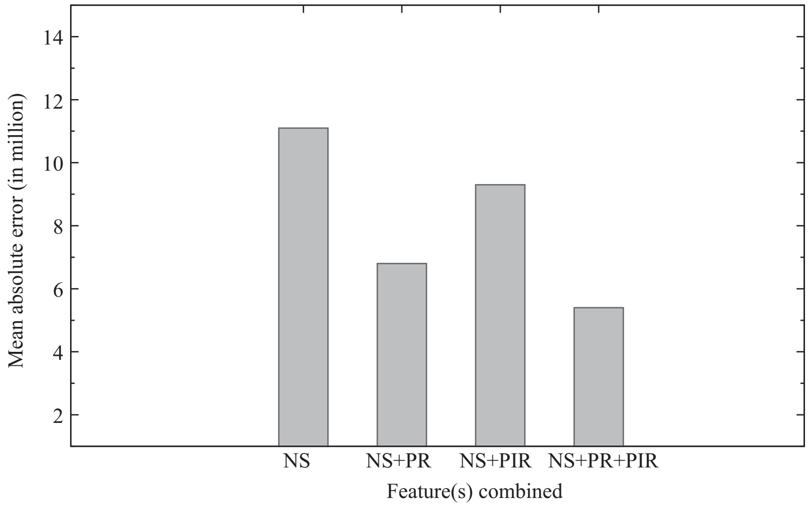

Figure 2 Performance of our approach on Dataset S2 combining different features. NS: number of screens; PR: post rate; PIR: purchase intention rate.

Not surprisingly, Num_of_Screen and Log_Budget are among the top two positively correlated features. Movies with more budgets before the productions and more scheduled screens during the distributions are more likely to have good market performances. It may also serve as indirectly evidences that our model have captured the most decisive indicators for the task. Compared with the feature Num_of_Oscar_Winning_Actor, Oscar_Winning_Director_Present gains a much higher correlation score with the box-office revenue. This may imply that as a decisive factor, actors are secondary to the directors, audiences show more expectations for the Oscar-winning directors rather than the actors.

As with other top-ranked items, they describe the movies from different aspects. 'beautiful' and 'pretty' probably express the critics' positive sentiments towards the heroines or the scenes while 'murder', 'in_blood' is likely to be disclosures of the plots. And 'be_loved', 'devotion' is probably talking about the themes.

Besides the positive correlated ones, we also list some that have exhibited negative correlations with the box-office revenues in Table 7. Interestingly, the items have revealed obvious negative sentiments towards the movies, such as 'a_brutal', 'clich_s' and 'and_the_sad'.

The critics' reviews are published before the movies' office releases. Thus the words inside the reviews may later indirectly influence the readers — the potential audiences of the movies. Analysis of impact of the wordings to the movie box-office revenues is left for future work.

6 Conclusion

In this article, we revisit the movie box-office revenue prediction task and propose a novel Gaussian copula regression model. The model jointly model meta-data and textual data. In particular, we propose an

Table 6 Top-ranked features that have exhibited positive correlations with the box-office revenues

Top 1–10	Top 11–20	Top 21–30
Num_of_Screen	less	murder
Log_Budget	know	phone
try	vincent	atmosphere
father_was	hell	pretty
co	joke	beloved
lawn	or_worse	klein
beautiful	moving_in	base
spurious	histrionics	devotion
comparison	could_almost_be	neo
Oscar_Winning_Director_Present	to_put	in_blood

Table 7 Top-ranked features that have exhibit negative correlations with the box-office revenues

in_a_black	the_detail	BOL_the_director
a_brutal	in_prison_for	clich_s
BOL_mostly_though	after_serving	and_the_sad

approximate inference algorithm which is computational inexpensive in high dimensions. By learning the dependency structure among the stochastic variables of arbitrary marginal distributions, we show that the correlations between variables can help boost the performance of the system. Experimental results show that our model outperforms two strong discriminative baseline models under various settings. Our results once again prove that apart from movie meta-data, user generated content and user behavior data are also powerful predictive indicators.

Acknowledgements This work was supported by National Basic Research Program of China (Grant No. 2014CB340503), and National Natural Science Foundation of China (Grant Nos. 71532004, 61133012, 61472107).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Liu T, Ding X, Chen Y, et al. Predicting movie box-office revenues by exploiting large-scale social media content. Multimedia Tools Appl, 2016, 75: 1509–1528

2 Zhou D H, Han W B, Wang Y J, et al. Information diffusion network inferring and pathway tracking. Sci China Inf Sci, 2015, 58: 092111

3 Duan J, Chen Y, Liu T, et al. Mining intention-related products on online q&a community. J Comput Sci Tech, 2015, 30: 1054–1062

4 Ding X, Liu T, Duan J, et al. Mining user consumption intention from social media using domain adaptive convolutional neural network. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, 2015. 2389–2395

5 Wang H, Can D, Kazemzadeh A, et al. A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics System Demonstrations, Jeju Island, 2012. 115–120

6 Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci, 2011, 2: 1–8

7 Ding X, Zhang Y, Liu T, et al. Using structured events to predict stock price movement: an empirical investigation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, 2014. 1415–1425

8 Asur S, Huberman B A. Predicting the future with social media. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Washington: IEEE Computer Society, 2010. 492–499

9 Pan R K, Sinha S. The statistical laws of popularity: universal properties of the box-office dynamics of motion pictures. New J Phys, 2010, 12: 5004

10 Sklar M. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris, 1959, 8: 229–231

11 Härdle W, Kleinow T, Stahl G. Applied Quantitative Finance: Theory and Computational Tools. Berlin: Springer, 2013

12 Eickhoff C, Vries A P, Collins-Thompson K. Copulas for information retrieval. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2013. 663–672

13 Wang W Y, Wen M. I can has cheezburger? A nonparanormal approach to combining textual and visual information

for predicting and generating popular meme descriptions. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, 2015. 355–365

14 Elidan G. Copula bayesian networks. Advances Neural Inf Process Syst, 2010, 23: 559–567

15 Fujimaki R, Sogawa Y, Morinaga S. Online heterogeneous mixture modeling with marginal and copula selection. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 2011. 645–653

16 Sharda R, Delen D. Predicting box-office success of motion pictures with neural networks. Expert Syst Appl, 2006, 30: 243–254

17 Zhang L, Luo J, Yang S. Forecasting box office revenue of movies with bp neural network. Expert Syst Appl, 2009, 36: 6580–6587

18 Mishne G, Glance N S. Predicting movie sales from blogger sentiment. In: Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Stanford, 2006. 155–158

19 Zhang W B, Skiena S. Improving movie gross prediction through news analysis. In: Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. Washington: IEEE Computer Society, 2009. 301–304

20 Joshi M, Das D, Gimpel K, et al. Movie reviews and revenues: an experiment in text regression. In: Proceedings of Human Language Technologies: the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, 2010. 293–296

21 Mestyán M, Yasseri T, Kertész J. Early prediction of movie box office success based on wikipedia activity big data. Plos One, 2013, 8: e71226

22 Zhang L, Singh V. Bivariate flood frequency analysis using the copula method. J Hydrol Eng, 2006, 11: 150–164

23 Wang W Y, Hua Z. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, 2014. 1155–1165

24 Nelsen R B. An Introduction to Copulas. New York: Springer, 2013

25 Joe H. Multivariate Models and Multivariate Dependence Concepts. Boca Raton: CRC Press, 1997

26 Yan J, Leeuw J D, Zeileis A. Enjoy the joy of copulas: with a package copula. J Stat Softw, 2007, 21: 1–21

27 Bird S. Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, Sydney, 2006. 69–72

28 Toutanova K, Manning C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction With the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13, Hong Kong, 2000. 63–70

29 Manning C D, Surdeanu M, Bauer J, et al. The stanford corenlp natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, 2014. 55–60

30 Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B, 2005, 67: 301–320

31 Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst, 1997, 9: 155–161