A Semiparametric Gaussian Copula Regression Model for Predicting Financial Risks from Earnings Calls

William Yang Wang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
yww@cs.cmu.edu

Zhenhao Hua
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
zhua@cs.cmu.edu

Abstract

Earnings call summarizes the financial performance of a company, and it is an important indicator of the future financial risks of the company. We quantitatively study how earnings calls are correlated with the financial risks, with a special focus on the financial crisis of 2009. In particular, we perform a text regression task: given the transcript of an earnings call, we predict the volatility of stock prices from the week after the call is made. We propose the use of *copula*: a powerful statistical framework that separately models the uniform marginals and their complex multivariate stochastic dependencies, while not requiring any prior assumptions on the distributions of the covariate and the dependent variable. By performing *probability integral transform*, our approach moves beyond the standard count-based bag-of-words models in NLP, and improves previous work on text regression by incorporating the correlation among local features in the form of semiparametric Gaussian copula. In experiments, we show that our model significantly outperforms strong linear and non-linear discriminative baselines on three datasets under various settings.

1 Introduction

Predicting the risks of publicly listed companies is of great interests not only to the traders and analysts on the Wall Street, but also virtually anyone who has investments in the market (Kogan et al., 2009). Traditionally, analysts focus on quantitative modeling of historical trading data. Today, even though earnings calls transcripts are abundantly available, their distinctive communicative practices (Camiciottoli, 2010), and correlations with the financial risks, in particular, future stock

performances (Price et al., 2012), are not well studied in the past.

Earnings calls are conference calls where a listed company discusses the financial performance. Typically, a earnings call contains two parts: the senior executives first report the operational outcomes, as well as the current financial performance, and then discuss their perspectives on the future of the company. The second part of the teleconference includes a question answering session where the floor will be open to investors, analysts, and other parties for inquiries. The question we ask is that, even though each earnings call has distinct styles, as well as different speakers and mixed formats, can we use earnings calls to predict the financial risks of the company in the limited future?

Given a piece of earnings call transcript, we investigate a semiparametric approach for automatic prediction of future financial risk[1]. To do this, we formulate the problem as a text regression task, and use a Gaussian copula with probability integral transform to model the uniform marginals and their dependencies. Copula models (Schweizer and Sklar, 1983; Nelsen, 1999) are often used by statisticians (Genest and Favre, 2007; Liu et al., 2012; Masarotto and Varin, 2012) and economists (Chen and Fan, 2006) to study the bivariate and multivariate stochastic dependency among random variables, but they are very new to the machine learning (Ghahramani et al., 2012; Han et al., 2012; Xiang and Neville, 2013; Lopez-paz et al., 2013) and related communities (Eickhoff et al., 2013). To the best of our knowledge, even though the term "copula" is named for the resemblance to grammatical copulas in linguistics, copula models have not been explored in the NLP community. To evaluate the performance of our approach, we compare with a standard squared loss linear regression baseline, as well as strong baselines such as linear and non-linear support

---

[1] In this work, the risk is defined as the measured volatility of stock prices from the week following the earnings call teleconference. See details in Section 5.

vector machines (SVMs) that are widely used in text regression tasks. By varying different experimental settings on three datasets concerning different periods of the Great Recession from 2006-2013, we empirically show that our approach significantly outperforms the baselines by a wide margin. Our main contributions are:

- We are among the first to formally study transcripts of earnings calls to predict financial risks.

- We propose a novel semiparametric Gaussian copula model for text regression.

- Our results significantly outperform standard linear regression and strong SVM baselines.

- By varying the number of dimensions of the covariates and the size of the training data, we show that the improvements over the baselines are robust across different parameter settings on three datasets.

In the next section, we outline related work in modeling financial reports and text regression. In Section 3, the details of the semiparametric copula model are introduced. We then describe the dataset and dependent variable in this study, and the experiments are shown in Section 6. We discuss the results and findings in Section 7 and then conclude in Section 8.

2 Related Work

Fung et al. (2003) are among the first to study SVM and text mining methods in the market prediction domain, where they align financial news articles with multiple time series to simulate the 33 stocks in the Hong Kong Hang Seng Index. However, text regression in the financial domain have not been explored until recently. Kogan et al. (2009) model the SEC-mandated annual reports, and performs linear SVM regression with ϵ-insensitive loss function to predict the measured volatility. Another recent study (Wang et al., 2013) uses exactly the same max-margin regression technique, but with a different focus on the financial sentiment. Using the same dataset, Tsai and Wang (2013) reformulate the regression problem as a text ranking problem. Note that all these regression studies above investigate the SEC-mandated annual reports, which are very different from the earnings calls in many aspects such as length, format, vocabulary, and genre. Most recently, Xie et al. (2013) have proposed the use of frame-level semantic features to understand financial news, but they treat the stock movement

prediction problem as a binary classification task. Broadly speaking, our work is also aligned to recent studies that make use of social media data to predict the stock market (Bollen et al., 2011; Zhang et al., 2011).

Despite our financial domain, our approach is more relevant to text regression. Traditional discriminative models, such as linear regression and linear SVM, have been very popular in various text regression tasks, such as predicting movie revenues from reviews (Joshi et al., 2010), understanding the geographic lexical variation (Eisenstein et al., 2010), and predicting food prices from menus (Chahuneau et al., 2012). The advantage of these models is that the estimation of the parameters is often simple, the results are easy to interpret, and the approach often yields strong performances. While these approaches have merits, they suffer from the problem of not explicitly modeling the correlations and interactions among random variables, which in some sense, corresponding to the impractical assumption of independent and identically distributed (i.i.d) of the data. For example, when bag-of-word-unigrams are present in the feature space, it is easier if one does not explicitly model the stochastic dependencies among the words, even though doing so might hurt the predictive power, while the variance from the correlations among the random variables is not explained.

3 Copula Models for Text Regression

In NLP, many statistical machine learning methods that capture the dependencies among random variables, including topic models (Blei et al., 2003; Lafferty and Blei, 2005; Wang et al., 2012), always have to make assumptions with the underlying distributions of the random variables, and make use of informative priors. This might be rather restricting the expressiveness of the model in some sense (Reisinger et al., 2010). On the other hand, once such assumptions are removed, another problem arises — they might be prone to errors, and suffer from the overfitting issue. Therefore, coping with the tradeoff between expressiveness and overfitting, seems to be rather important in statistical approaches that capture stochastic dependency.

Our proposed semiparametric copula regression model takes a different perspective. On one hand, copula models (Nelsen, 1999) seek to explicitly model the dependency of random variables by separating the marginals and their correlations. On the other hand, it does not make use of any as-

sumptions on the distributions of the random variables, yet, the copula model is still expressive. This nice property essentially allows us to fuse distinctive lexical, syntactic, and semantic feature sets naturally into a single compact model.

From an information-theoretic point of view (Shannon, 1948), various problems in text analytics can be formulated as estimating the probability mass/density functions of tokens in text. In NLP, many of the probabilistic text models work in the discrete space (Church and Gale, 1995; Blei et al., 2003), but our model is different: since the text features are sparse, we first perform kernel density estimates to smooth out the zeroing items, and then calculate the empirical cumulative distribution function (CDF) of the random variables. By doing this, we are essentially performing *probability integral transform*— an important statistical technique that moves beyond the count-based bag-of-words feature space to marginal cumulative density functions space. Last but not least, by using a parametric copula, in our case, the Gaussian copula, we reduce the computational cost from fully nonparametric methods, and explicitly model the correlations among the covariate and the dependent variable.

In this section, we first briefly look at the theoretical foundations of copulas, including the Sklar's theorem. Then we describe the proposed semiparametric Gaussian copula text regression model. The algorithmic implementation of our approach is introduced at the end of this section.

3.1 The Theory of Copula

In the statistics literature, copula is widely known as a family of distribution function. The idea behind copula theory is that the cumulative distribution function (CDF) of a random vector can be represented in the form of uniform marginal cumulative distribution functions, and a copula that connects these marginal CDFs, which describes the correlations among the input random variables. However, in order to have a valid multivariate distribution function regardless of n-dimensional covariates, not every function can be used as a copula function. The central idea behind copula, therefore, can be summarize by the Sklar's theorem and the corollary.

Theorem 1 (Sklar's Theorem (1959)) *Let* F *be the joint cumulative distribution function of* n *random variables* $X_1, X_2, ..., X_n$. *Let the corresponding marginal cumulative distribution functions of the random variable be* $F_1(x_1), F_2(x_2), ..., F_n(x_n)$. *Then, if the marginal*

functions are continuous, there exists a unique copula C, *such that*

$$F(x_1, ..., x_n) = C[F_1(x_1), ..., F_n(x_n)]. \quad (1)$$

Furthermore, if the distributions are continuous, the multivariate dependency structure and the marginals might be separated, and the copula can be considered independent of the marginals (Joe, 1997; Parsa and Klugman, 2011). Therefore, the copula does not have requirements on the marginal distributions, and any arbitrary marginals can be combined and their dependency structure can be modeled using the copula. The inverse of Sklar's Theorem is also true in the following:

Corollary 1 *If there exists a copula* $C : (0, 1)^n$ *and marginal cumulative distribution functions* $F_1(x_1), F_2(x_2), ..., F_n(x_n)$, *then* $C[F_1(x_1), ..., F_n(x_n)]$ *defines a multivariate cumulative distribution function.*

3.2 Semiparametric Gaussian Copula Models

The Non-Parametric Estimation

We formulate the copula regression model as follows. Assume we have n random variables of text features $X_1, X_2, ..., X_n$. The problem is that text features are sparse, so we need to perform non-parametric kernel density estimation to smooth out the distribution of each variable. Let $f_1, f_2, ..., f_n$ be the unknown density, we are interested in deriving the shape of these functions. Assume we have m samples, the kernel density estimator can be defined as:

$$\hat{f}_h(x) = \frac{1}{m} \sum_{i=1}^{m} K_h(x - x_i) \quad (2)$$

$$= \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right) \quad (3)$$

Here, $K(\cdot)$ is the kernel function, where in our case, we use the Box kernel[2] $K(z)$:

$$K(z) = \frac{1}{2}, |z| \leq 1, \quad (4)$$

$$= 0, |z| > 1. \quad (5)$$

Comparing to the Gaussian kernel and other kernels, the Box kernel is simple, and computationally inexpensive. The parameter h is the bandwidth for smoothing[3].

---

[2]It is also known as the original Parzen windows (Parzen, 1962).

[3]In our implementation, we use the default h of the Box kernel in the *ksdensity* function in Matlab.

Now, we can derive the empirical cumulative distribution functions $\hat{F}_{X_1}(\hat{f}_1(X_1)), \hat{F}_{X_2}(\hat{f}_2(X_2)), ..., \hat{F}_{X_n}(\hat{f}_n(X_n))$ of the smoothed covariates, as well as the dependent variable y and its CDF $\hat{F}_y(\hat{f}(y))$. The empirical cumulative distribution functions are defined as:

$$\hat{F}(\nu) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{I}\{x_i \leq \nu\} \qquad (6)$$

where $\mathbf{I}\{\cdot\}$ is the indicator function, and ν indicates the current value that we are evaluating. Note that the above step is also known as *probability integral transform* (Diebold et al., 1997), which allows us to convert any given continuous distribution to random variables having a uniform distribution. This is of crucial importance to modeling text data: instead of using the classic bag-of-words representation that uses raw counts, we are now working with uniform marginal CDFs, which helps coping with the overfitting issue due to noise and data sparsity.

The Parametric Copula Estimation

Now that we have obtained the marginals, and then the joint distribution can be constructed by applying the copula function that models the stochastic dependencies among marginal CDFs:

$$\hat{F}(\hat{f}_1(X_1), ..., \hat{f}_1(X_n), \hat{f}(y)) \qquad (7)$$
$$= C[\hat{F}_{X_1}(\hat{f}_1(X_1)), ..., \hat{F}_{X_n}(\hat{f}_n(X_n)), \hat{F}_y(\hat{f}_y(y))] \quad (8)$$

In this work, we apply the parametric Gaussian copula to model the correlations among the text features and the label. Assume x_i is the smoothed version of random variable X_i, and y is the smoothed label, we have:

$$F(x_1, ..., x_n, y) \qquad (9)$$
$$= \Phi_\Sigma\left(\Phi^{-1}[F_{x_1}(x_1)], ..., , \Phi^{-1}[F_{x_n}(x_n)], \Phi^{-1}[F_y(y)]\right) \qquad (10)$$

where Φ_Σ is the joint cumulative distribution function of a multivariate Gaussian with zero mean and Σ variance. Φ^{-1} is the inverse CDF of a standard Gaussian. In this parametric part of the model, the parameter estimation boils down to the problem of learning the covariance matrix Σ of this Gaussian copula. In this work, we perform standard maximum likelihood estimation for the Σ matrix.

To calibrate the Σ matrix, we make use of the power of randomness: using the initial Σ from MLE, we generate random samples from the Gaussian copula, and then concatenate previously generated joint of Gaussian inverse marginal CDFs with the newly generated random copula

numbers, and re-estimate using MLE to derive the final adjusted Σ. Note that the final Σ matrix has to be symmetric and positive definite.

Computational Complexity

One important question regarding the proposed semiparametric Gaussian copula model is the corresponding computational complexity. This boils down to the estimation of the $\hat{\Sigma}$ matrix (Liu et al., 2012): one only needs to calculate the correlation coefficients of $n(n-1)/2$ pairs of random variables. Christensen (2005) shows that sorting and balanced binary trees can be used to calculate the correlation coefficients with complexity of $O(n \log n)$. Therefore, the computational complexity of MLE for the proposed model is $O(n \log n)$.

Efficient Approximate Inference

In this regression task, in order to perform exact inference of the conditional probability distribution $p(F_y(y)|F_{x_1}(x_1), ..., F_{x_n}(x_n))$, one needs to solve the mean response $\hat{\mathbf{E}}(F_y(y)|F_{x_1}(x_1), ..., F_{x_1}(x_1))$ from a joint distribution of high-dimensional Gaussian copula.

Assume in the simple bivariate case of Gaussian copula regression, the covariance matrix Σ is:

$$\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ & \Sigma_{22} \end{array} \right]$$

We can easily derive the conditional density that can be used to calculate the expected value of the CDF of the label:

$$C(F_y(y)|F_{x_1}(x_1); \Sigma) = \frac{1}{|\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}|^{\frac{1}{2}}}$$
$$\exp\left(-\frac{1}{2}\delta^T \left([\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}]^{-1} - \mathbf{I} \right)\delta \right) \qquad (11)$$

where $\delta = \Phi^{-1}[F_y(y)] - \Sigma_{12}^T \Sigma_{11}^{-1} \Phi^{-1}[F_{x_1}(x_1)]$.

Unfortunately, the exact inference can be intractable in the multivariate case, and approximate inference, such as Markov Chain Monte Carlo sampling (Gelfand and Smith, 1990; Pitt et al., 2006) is often used for posterior inference. In this work, we propose an efficient sampling method to derive y given the text features — we sample $F_y(y)$ s.t. it maximizes the joint high-dimensional Gaussian copula density:

$$\hat{F_y}(y) \approx \underset{F_y(y) \in (0,1)}{\arg\max} \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}\Delta^T \cdot (\Sigma^{-1} - \mathbf{I}) \cdot \Delta \right) \qquad (12)$$

where

$$\Delta = \begin{pmatrix} \Phi^{-1}(F_{x_1}(x_1)) \\ \vdots \\ \Phi^{-1}(F_{x_n}(x_n)) \\ \Phi^{-1}(F_y(y)) \end{pmatrix}$$

Again, the reason why we perform approximated inference is that: exact inference in the high-dimensional Gaussian copula density is nontrivial, and might not have analytical solutions, but approximate inference using maximum density sampling from the Gaussian copula significantly relaxes the complexity of inference. Finally, to derive $\hat{y}$, the last step is to compute the inverse CDF of $F_y(y)$.

3.3 Algorithmic Implementation

The algorithmic implementation of our semiparametric Gaussian copula text regression model is shown in Algorithm 1. Basically, the algorithm can be decomposed into four parts:

- Perform nonparametric Box kernel density estimates of the covariates and the dependent variable for smoothing.

- Calculate the empirical cumulative distribution functions of the smoothed random variables.

- Estimate the parameters (covariance Σ) of the Gaussian copula.

- Infer the predicted value of the dependent variable by sampling the Gaussian copula probability density function.

4 Datasets

We use three datasets[4] of transcribed quarterly earnings calls from the U.S. stock market, focusing on the period of the Great Recession.

The *pre-2009* dataset consists of earnings calls from the period of 2006-2008, which includes calls from the beginning of economic downturn, the outbreak of the subprime mortgage crisis, and the epidemic of collapses of large financial institutions. The *2009* dataset contains earnings calls from the year of 2009, which is a period where the credit crisis spreads globally, and the Dow Jones Industrial Average hit the lowest since the beginning of the millennium. The *post-2009* dataset includes earnings calls from the period of 2010 to 2013, which concerns the recovery of global economy. The detailed statistics is shown in Table 1.

---

[4]http://www.cs.cmu.edu/~yww/data/earningscalls.zip

Algorithm 1 A Semi-parametric Gaussian Copula Model Based Text Regression Algorithm

---

Given:
(1) training data $(X^{(tr)}, \vec{y}^{(tr)})$;
(2) testing data $(X^{(te)}, \vec{y}^{(te)})$;

Learning:
for $i = 1 \rightarrow n$ dimensions **do**
   $X_i^{(tr)'} \leftarrow BoxKDE(X_i^{(tr)}, X_i^{(tr)})$;
   $U_i^{(tr)} \leftarrow EmpiricalCDF(X_i^{(tr)'})$;
   $X_i^{(te)'} \leftarrow BoxKDE(X_i^{(tr)}, X_i^{(te)})$;
   $U_i^{(te)} \leftarrow EmpiricalCDF(X_i^{(te)'})$;
end for
$y^{(tr)'} \leftarrow BoxKDE(y^{(tr)}, y^{(tr)})$;
$v^{(tr)} \leftarrow EmpiricalCDF(y^{(tr)'})$;
$Z^{(tr)} \leftarrow GaussianInverseCDF([U^{(tr)}\ v^{(tr)}])$;
$\hat{\Sigma} \leftarrow CorrelationCoefficients(Z^{(tr)})$;
$r \leftarrow MultiVariateGaussianRandNum(0, \hat{\Sigma}, n)$;
$Z^{(tr)'} = GaussianCDF(r)$;
$\hat{\Sigma} \leftarrow CorrelationCoefficients([Z^{(tr)}\ Z^{(tr)'}])$;

Inference:
for $j = 1 \rightarrow m$ instances **do**
   $\max_j \leftarrow 0$;
   $\hat{Y}' = 0$;
   for $k = 0.01 \rightarrow 1$ **do**
      $Z^{(te)} \leftarrow GaussianInverseCDF([U^{(te)}\ k])$;
      $p_j = \frac{MultiVariateGaussianPDF(Z^{(te)}, \hat{\Sigma})}{\prod_n GaussianPDF(Z^{(te)})}$;
      if $p_j \geq \max_j$ **then**
         $\max_j = p_j$;
         $\hat{Y}' = k$;
      end if
   end for
end for
$\hat{y} \leftarrow InverseCDF(\vec{y}^{(tr)}, \hat{Y}')$;

Dataset	#Calls	#Companies	#Types	#Tokens
Pre-2009	3694	2746	371.5K	28.7M
2009	3474	2178	346.2K	26.4M
Post-2009	3726	2107	377.4K	28.6M

Table 1: Statistics of three datasets. *Types: unique words. Tokens: word tokens.*

Note that unlike the standard news corpora in NLP or the SEC-mandated financial report, Transcripts of earnings call is a very special genre of text. For example, the length of WSJ documents is typically one to three hundreds (Harman, 1995), but the averaged document length of our three earnings calls datasets is 7677. Depending on the amount of interactions in the question answering session, the complexities of the calls vary. This mixed form of formal statement and informal speech brought difficulties to machine learning algorithms.

5 Measuring Financial Risks

Volatility is an important measure of the financial risk, and in this work, we focus on predicting the future volatility following the earnings teleconfer-

ence call. For each earning call, we have a week of stock prices of the company after the day on which the earnings call is made. The *Return* of Day t is:

$$r_t = \frac{x_t}{x_{t-1}} - 1 \qquad (13)$$

where x_t represents the share price of Day t, and the *Measured Stock Volatility* from Day t to $t + \tau$:

$$y_{(t,t+\tau)} = \sqrt{\frac{\sum_{i=0}^{\tau}(r_{t+i} - \bar{r})^2}{\tau}} \qquad (14)$$

Using the stock prices, we can use the equations above to calculate the measured stock volatility after the earnings call, which is the standard measure of risks in finance, and the dependent variable y of our predictive task.

6 Experiments

6.1 Experimental Setup

In all experiments throughout this section, we use 80-20 train/test splits on all three datasets.

Feature sets:

We have extracted lexical, named entity, syntactic, and frame-semantics features, most of which have been shown to perform well in previous work (Xie et al., 2013). We use the unigrams and bigrams to represent lexical features, and the Stanford part-of-speech tagger (Toutanova et al., 2003) to extract the lexicalized named entity and part-of-speech features. A probabilistic frame-semantics parser, SEMAFOR (Das et al., 2010), is used to provide the FrameNet-style frame-level semantic annotations. For each of the five sets, we collect the top-100 most frequent features, and end up with a total of 500 features.

Baselines:

The baselines are standard squared-loss linear regression, linear kernel SVM, and non-linear (Gaussian) kernel SVM. They are all standard algorithms in regression problems, and have been shown to have outstanding performances in many recent text regression (Kogan et al., 2009; Chahuneau et al., 2012; Xie et al., 2013; Wang et al., 2013; Tsai and Wang, 2013). We use the Statistical Toolbox's linear regression implementation in Matlab, and LibSVM (Chang and Lin, 2011) for training and testing the SVM models. The hyperparameter C in linear SVM, and the γ and C hyperparameters in Gaussian SVM are tuned on the training set using 10-fold cross-validation. Note that since the kernel density estimation in the proposed copula model is nonparametric, and we only need to learn the Σ in the

Gaussian copula, there is no hyperparameters that need to be tuned.

Evaluation Metrics:

Spearman's correlation (Hogg and Craig, 1994) and Kendall's tau (Kendall, 1938) have been widely used in many regression problems in NLP (Albrecht and Hwa, 2007; Yogatama et al., 2011; Wang et al., 2013; Tsai and Wang, 2013), and here we use them to measure the quality of predicted values $\hat{y}$ by comparing to the vector of ground truth y. In contrast to Pearson's correlation, Spearman's correlation has no assumptions on the relationship of the two measured variables. Kendall's tau is a nonparametric statistical metric that have shown to be inexpensive, robust, and representation independent (Lapata, 2006). We also use paired two-tailed t-test to measure the statistical significance between the best and the second best approaches.

6.2 Comparing to Various Baselines

In the first experiment, we compare the proposed semiparametric Gaussian copula regression model to three baselines on three datasets with all features. The detailed results are shown in Table 2. On the *pre-2009* dataset, we see that the linear regression and linear SVM perform reasonably well, but the Gaussian kernel SVM performs less well, probably due to overfitting. The copula model outperformed all three baselines by a wide margin on this dataset with both metrics. Similar performances are also obtained in the *2009* dataset, where the result of linear SVM baseline falls behind. On the *post-2009* dataset, none of results from the linear and non-linear SVM models can match up with the linear regression model, but our proposed copula model still improves over all baselines by a large margin. Comparing to second-best approaches, all improvements obtained by the copula model are statistically significant.

6.3 Varying the Amount of Training Data

To understand the learning curve of our proposed copula regression model, we use the 25%, 50%, 75% subsets from the training data, and evaluate all four models. Figure 1 shows the evaluation results. From the experiments on the *pre-2009* dataset, we see that when the amount of training data is small (25%), both SVM models have obtained very impressive results. This is not surprising at all, because as max-margin models, soft-margin SVM only needs a handful of examples that come with nonvanishing coefficients (support vectors) to find a reasonable margin. When in-

Method	Pre-2009		2009		Post-2009	
	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall
linear regression:	0.377	0.259	0.367	0.252	0.314	0.216
linear SVM:	0.364	0.249	0.242	0.167	0.132	0.091
Gaussian SVM:	0.305	0.207	0.280	0.192	0.152	0.104
Gaussian copula:	**0.425***	**0.315***	**0.422***	**0.310***	**0.375***	**0.282***

Table 2: Comparing the learning algorithms on three datasets with all features. The best result is highlighted in **bold**. * indicates $p < .001$ comparing to the second best result.
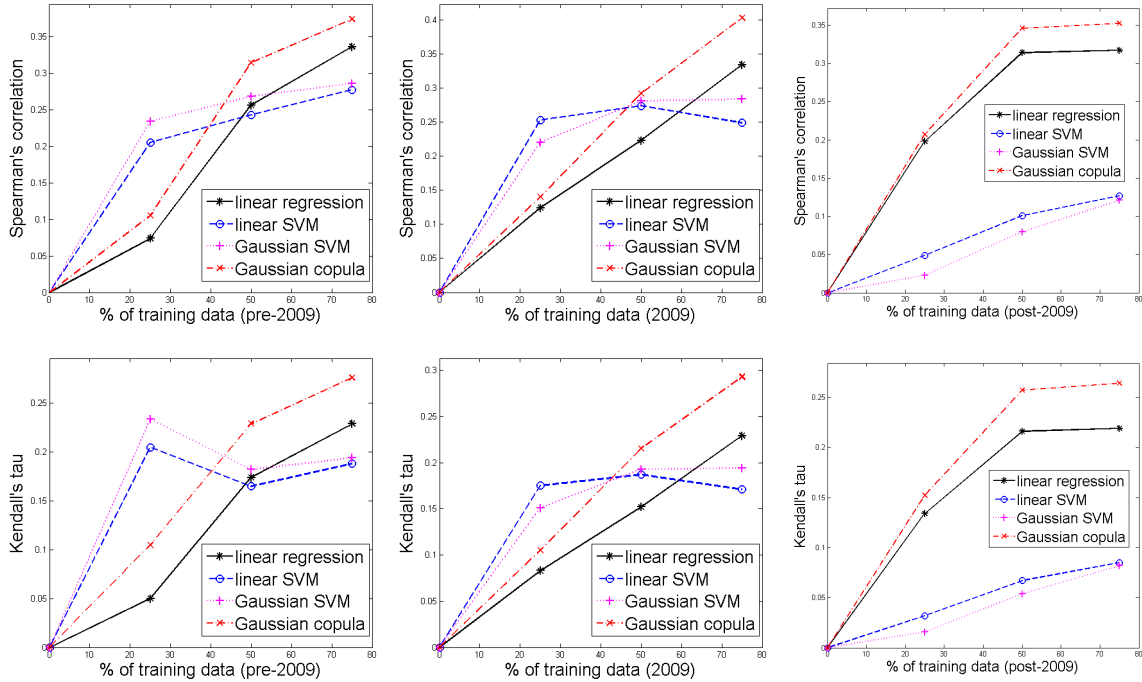

Figure 1: Varying the amount of training data. Left column: pre-2009 dataset. Middle column: 2009 dataset. Right column: post-2009 dataset. Top row: Spearman's correlation. Bottom row: Kendall's tau.

creasing the amount of training data to 50%, we do see the proposed copula model catches up quickly, and lead all baseline methods undoubtably at 75% training data. On the *2009* dataset, we observe very similar patterns. Interestingly, the proposed copula regression model has dominated all methods for both metrics throughout all proportions of the "post-2009" earnings calls dataset, where instead of financial crisis, the economic recovery is the main theme. In contrast to the previous two datasets, both linear and non-linear SVMs fail to reach reasonable performances on this dataset.

6.4 Varying the Amount of Features

Finally, we investigate the robustness of the proposed semiparametric Gaussian copula regression model by varying the amount of features in the covariate space. To do this, we sample equal amount of features from each feature set, and concatenate them into a feature vector. When increasing the amount of total features from 100 to 400, the results are shown in Figure 2. On the *pre-2009* dataset, we see that the gaps between the best-perform copula model and the second-best linear regression model are consistent throughout all feature sizes. On the *2009* dataset, we see that the performance of Gaussian copula is aligned with the linear regression model in terms of Spearman's correlation, where the former seems to perform better in terms of Kendall's tau. Both linear and non-linear SVM models do not have any advantages over the proposed approach. On the *post-2009* dataset that concerns economic growth and recovery, the boundaries among all methods are very clear. The Spearman's correlation for both SVM baselines is less than 0.15 throughout all settings, but copula model is able to achieve 0.4 when using 400 features. The improvements of copula
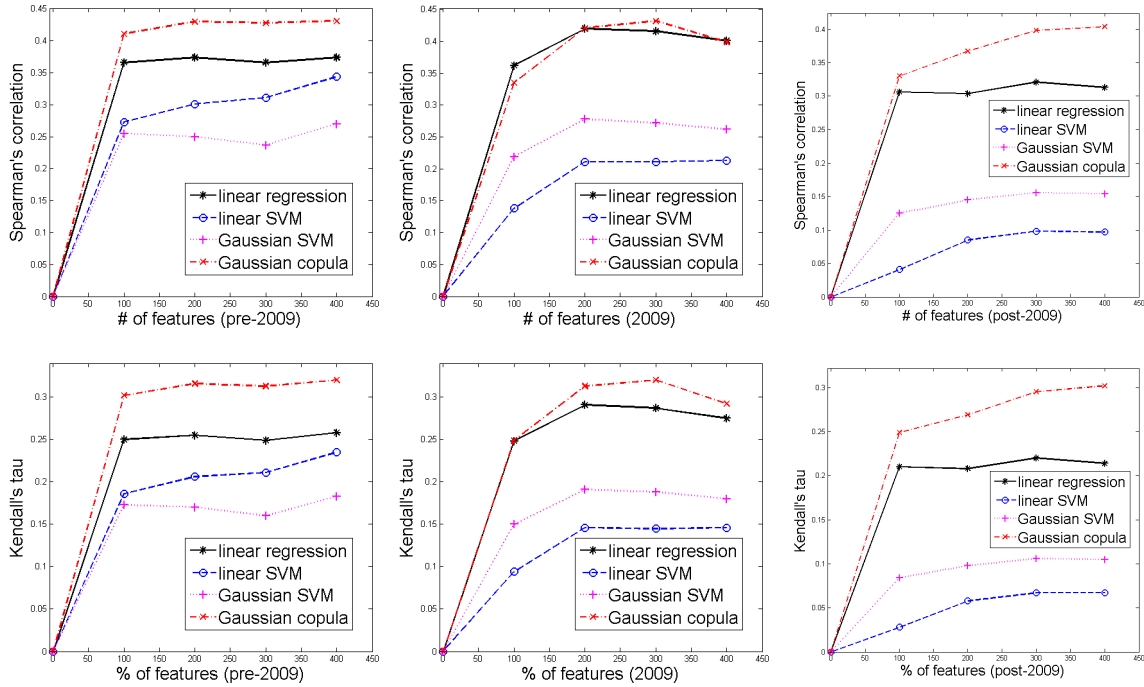
Figure 2: Varying the amount of features. Left column: pre-2009 dataset. Middle column: 2009 dataset. Right column: post-2009 dataset. Top row: Spearman's correlation. Bottom row: Kendall's tau.

Pre-2009	2009	Post-2009
2008/CD	2008	first_quarter
2008	million/CD	revenue/NN
third_quarter	2008/CD	revenue
third	million	quarter_of
third/JJ	million_in	compared_to
the_third	the_fourth	million_in
million/CD	fourth_quarter	Peter/PERSON
capital	fourth	call
million	fourth/JJ	first/JJ
FE_Trajector_entity	$/$	million/CD

Table 3: Top-10 features that have positive correlations with stock volatility in three datasets.

model over squared loss linear regression model are increasing, when working with larger feature spaces.

6.5 Qualitative Analysis

Like linear classifiers, by "opening the hood" to the Gaussian copula regression model, one can examine features that exhibit high correlations with the dependent variable. Table 3 shows the top features that are positively correlated with the future stock volatility in the three datasets. On the top features from the "pre-2009" dataset, which primarily (82%) includes calls from 2008, we can clearly observe that the word "2008" has strong correlation with the financial risks. Interestingly, the phrase "third quarter" and its variations, not only play an important role in the model, but also highly correlated to the timeline of the financial crisis: the Q3 of 2008 is a critical period in the

recession, where Lehman Brothers falls on the Sept. 15 of 2008, filing $613 billion of debt — the biggest bankruptcy in U.S. history (Mamudi, 2008). This huge panic soon broke out in various financial institutions in the Wall Street. On the top features from "2009" dataset, again, we see the word "2008" is still prominent in predicting financial risks, indicating the hardship and extended impacts from the center of the economic crisis. After examining the transcripts, we found sentences like: "*...our specialty lighting business that we discontinued in the fourth quarter of 2008...*", "*...the exception of fourth quarter revenue which was $100,000 below our guidance target...*", and "*...to address changing economic conditions and their impact on our operations, in the fourth quarter we took the painful but prudent step of decreasing our headcount by about 5%...*", showing the crucial role that Q4 of 2008 plays in 2009 earnings calls. Interestingly, after the 2008-2009 crisis, in the recovery period, we have observed new words like "revenue", indicating the "back-to-normal" trend of financial environment, and new features that predict financial volatility.

7 Discussions

In the experimental section, we notice that the proposed semiparametric Gaussian copula model has obtained promising results in various setups on three datasets in this text regression task. The

main questions we ask are: how is the proposed model different from standard text regression/classification models? What are the advantages of copula-based models, and what makes it perform so well?

One advantage we see from the copula model is that it does not require any assumptions on the marginal distributions. For example, in latent Dirichlet allocation (Blei et al., 2003), the topic proportion of a document is always drawn from a $Dirichlet(\alpha)$ distribution. This is rather restricted, because the possible shapes from a $K-1$ simplex of Dirichlet is always limited in some sense. In our copula model, instead of using some priors, we just calculate the empirical cumulative distribution function of the random variables, and model the correlation among them. This is extremely practical, because in many natural language processing tasks, we often have to deal with features that are extracted from many different domains and signals. By applying the *Probability Integral Transform* to raw features in the copula model, we essentially avoid comparing apples and oranges in the feature space, which is a common problem in bag-of-features models in NLP.

The second hypothesis is about the semiparametirc parameterization, which contains the nonparametric kernel density estimation and the parametric Gaussian copula regression components. The benefit of a semiparametric model is that here we are not interested in performing completely nonparametric estimations, where the infinite dimensional parameters might bring intractability. In contrast, by considering the semiparametric case, we not only obtain some expressiveness from the nonparametric models, but also reduce the complexity of the task: we are only interested in the finite-dimensional components Σ in the Gaussian copula with $O(n \log n)$ complexity, which is not as computationally difficult as the completely nonparametric cases. Also, by modeling the marginals and their correlations seperately, our approach is cleaner, easy-to-understand, and allows us to have more flexibility to model the uncertainty of data. Our pilot experiment also aligns with our hypothesis: when not performing the kernel density estimation part for smoothing out the marginal distributions, the performances dropped significantly when sparser features are included.

The third advantage we observe is the power of modeling the covariance of the random variables. Traditionally, in statistics, independent and identically distributed (i.i.d) assumptions among the instances and the random variables are often used in various models, such that the correlations among the instances or the variables are often ignored. However, this might not be practical at all: in image processing, the "cloud" pixel of a pixel showing the blue sky of a picture are more likelihood to co-occur in the same picture; in natural language processing, the word "mythical" is more likely to co-occur with the word "unicorn", rather than the word "popcorn". Therefore, by modeling the correlations among marginal CDFs, the copula model has gained the insights on the dependency structures of the random variables, and thus, the performance of the regression task is boosted.

In the future, we plan to apply the proposed approach to large datasets where millions of features and millions of instances are involved. Currently we have not experienced the difficulty when estimating the Gaussian copula model, but parallel methods might be needed to speedup learning when significantly more marginal CDFs are involved. The second issue is about overfitting. We see that when features are rather noisy, we might need to investigate regularized copula models to avoid this. Finally, we plan to extend the proposed approach to text classification and structured prediction problems in NLP.

8 Conclusion

In this work, we have demonstrated that the more complex quarterly earnings calls can also be used to predict the measured volatility of the stocks in the limited future. We propose a novel semiparametric Gausian copula regression approach that models the dependency structure of the language in the earnings calls. Unlike traditional bag-of-features models that work discrete features from various signals, we perform kernel density estimation to smooth out the distribution, and use probability integral transform to work with CDFs that are uniform. The copula model deals with marginal CDFs and the correlation among them separately, in a cleaner manner that is also flexible to parameterize. Focusing on the three financial crisis related datasets, the proposed model significantly outperform the standard linear regression method in statistics and strong discriminative support vector regression baselines. By varying the size of the training data and the dimensionality of the covariates, we have demonstrated that our proposed model is relatively robust across different parameter settings.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level mt evaluation with pseudo references. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*.

Belinda Camiciottoli. 2010. Earnings calls: Exploring an emerging financial reporting genre. *Discourse & Communication*.

Victor Chahuneau, Kevin Gimpel, Bryan R Routledge, Lily Scherlis, and Noah A Smith. 2012. Word salad: Relating food prices and descriptions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Xiaohong Chen and Yanqin Fan. 2006. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*.

David Christensen. 2005. Fast algorithms for the calculation of kendalls τ. *Computational Statistics*.

Kenneth Church and William Gale. 1995. Poisson mixtures. *Natural Language Engineering*.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith. 2010. Probabilistic frame-semantic parsing. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*.

Francis X Diebold, Todd A Gunther, and Anthony S Tay. 1997. Evaluating density forecasts.

Carsten Eickhoff, Arjen P. de Vries, and Kevyn Collins-Thompson. 2013. Copulas for information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Pui Cheong Fung, Xu Yu, and Wai Lam. 2003. Stock prediction: Integrating text mining approach using real-time news. In *Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering*.

Alan Gelfand and Adrian Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*.

Christian Genest and Anne-Catherine Favre. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*.

Zoubin Ghahramani, Barnabás Póczos, and Jeff Schneider. 2012. Copula-based kernel dependency measures. In *Proceedings of the 29th International Conference on Machine Learning*.

Fang Han, Tuo Zhao, and Han Liu. 2012. Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*.

Donna Harman. 1995. Overview of the second text retrieval conference (trec-2). *Information Processing & Management*.

Robert V Hogg and Allen Craig. 1994. Introduction to mathematical statistics.

Harry Joe. 1997. *Multivariate models and dependence concepts*.

Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*.

Shimon Kogan, Dimitry Levin, Bryan Routledge, Jacob Sagi, and Noah Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

John Lafferty and David Blei. 2005. Correlated topic models. In *Advances in neural information processing systems*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. 2012. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*.

David Lopez-paz, Jose M Hernández-lobato, and Ghahramani Zoubin. 2013. Gaussian process vine copulas for multivariate dependence. In *Proceedings of the 30th International Conference on Machine Learning*.

Sam Mamudi. 2008. *Lehman folds with record $613 billion debt*. MarketWatch.com.

1164

Guido Masarotto and Cristiano Varin. 2012. Gaussian copula marginal regression. *Electronic Journal of Statistics*.

Roger B Nelsen. 1999. *An introduction to copulas*. Springer Verlag.

Rahul A Parsa and Stuart A Klugman. 2011. Copula regression. *Variance Advancing and Science of Risk*.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*.

Michael Pitt, David Chan, and Robert Kohn. 2006. Efficient bayesian inference for gaussian copula regression models. *Biometrika*.

McKay Price, James Doran, David Peterson, and Barbara Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. 2010. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning*.

Berthold Schweizer and Abe Sklar. 1983. *Probabilistic metric spaces*.

Claude Shannon. 1948. A mathematical theory of communication. In *The Bell System Technical Journal*.

Abe Sklar. 1959. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Ming-Feng Tsai and Chuan-Ju Wang. 2013. Risk ranking from financial reports. In *Advances in Information Retrieval*.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

Rongjing Xiang and Jennifer Neville. 2013. Collective inference for network data with copula latent markov networks. In *Proceedings of the sixth ACM international conference on Web search and data mining*.

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. 2013. Semantic frames to predict stock price movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Xue Zhang, Hauke Fuehres, and Peter A Gloor. 2011. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*.