# Text Regression with Unequal Penalty Factors

$ZhendongZhao^{1,2}$ $NataliyaSokolovska^{1}$ $MarkJohnson^{1}$
(1)Department of Computer Science, Macquarie University
(2) The Capital Markets Cooperative Research Centre (CMCRC), Sydney, Australia
zhaozhendong@gmail.com, mark.johnson@mq.edu.au
Technical Report

**Abstract**

Regularization hyper-parameters play a critical role in problems where features are organized into groups or structures with significant noise. Usually, all features are penalized equally. In this paper, we propose a novel regression approach with unequal penalty factors on different groups of features and apply the approach to predict the daily returns of a stock. Two distinctive groups of features, i.e., financial texts and quantitative information, are given in this prediction task to investigate the effectiveness of unequal penalty factors. We conduct experiments on all companies listed in Australia Stock eXchanges (ASX), the results show that different penalty factors for different groups of features significantly reduces the mean square error.

**Keywords:** Sparse linear regression, group penalty factors, financial risk prediction

## 1 Introduction

Predictability of the daily return, the ratio of money gained or lost on a stock, is of critical interest to investors. Recent literatures, focus on either pure text data (i.e. announcements) or financial quantitative data (i.e. historical daily returns) for predicting the stock daily returns. The text data includes official self-report (e.g., company announcements), formal media news (e.g., Reuters news, Wall Street Journal, Dow Jones News Service), social media (e.g., twitter); the financial quantitative data includes market-wide investor sentiment [Baker and Wurgler2006], global, local and contagious investor sentiment [Baker et al.2012], past volatility [Kogan et al.2009], and anomalies [Stambaugh et al.2012].

The combination of these two heterogeneous sources has also been investigated by Kogan et al. [Kogan et al.2009], in which the authors predict the volatility of coming year using annual reports and historical volatilities. Although the work is the pioneer of text regression, the number of quantitative variables seems to be insufficient and the method of combining two heterogeneous data is improvable, since two heterogeneous sources, the text and financial quantitative features, possess rather different properties.

For example, the size of text features (unigram and bigram model) is huge but sparse; the size of quantitative features is small but dense.

In this paper, therefore, we address a regression model with unequal penalty factors for heterogeneous groups of features to predict the daily returns of stocks and apply this model to Australia Security Exchange (ASX). The official announcements are used as text data, and the past daily returns, capital size and the stock price are used as financial quantitative data.

Our contributions are threefold:

1. We introduce a unique text regression data set, which has half-year announcements and the quantitative data of all companies listed on Australia Security Exchange (ASX). The data set will be made publicly available;

2. We propose a novel regression scheme that combines both text and financial quantitative features with unequal penalty factors to predict the daily returns of stocks and apply this model to Australia Security Exchange (ASX). To the best of our knowledge, there is no published paper addressing the regression problem to predict the daily returns using heterogeneous features and unequal penalty factors. We show that the combination of quantitative and text features outperforms models trained on data issued from one source;

3. We demonstrate that unequal penalty factors are important for different heterogenous sources.

The structure of the paper is as follows. In Section 2 we consider the text regression model. In Section 3 we describe the model and formalize the training procedure. Section 4 introduces the dataset used in our experiments. The data set can be obtained on demand, and will be made publicly available. In Section 5 we demonstrate our results on a newly introduced data set. Concluding remarks and perspectives close the paper.

## 2 Text Regression

It is widely known that a stock's daily return is difficult to predict using either public information or quantitative data. Yet prediction is a key role of efficient markets and the "efficient market hypothesis" [Fama1970]. A new approach, called text regression [Kogan et al.2009], however, has shown the improved prediction power by using both of the data. In this approach, the authors tried to predict the yearly volatility, in which both text and financial quantitative features are used to boost the prediction performance. With this inspiration, we extend their methodology to predict the daily returns. Unlike their work, our regression scheme predicts the daily returns, including more quantitative features, applying unequal penalty factors to different groups of features.

Following their methodology, we use the similar methodology to predict the daily returns, in turn, to solve this regression problem:

$$R_t = \underbrace{\sum_{j=1}^{k}(\theta_j Q_j)}_{\text{Quantitative Features}} + \underbrace{\sum_{j=k+1}^{m}(\theta_j T_j)}_{\text{Text Features}} \quad (1)$$

Where $R_t$ is the daily returns when the announcement is exposed. Assume we have $k$ quantitative features and $m$ unigram and bigram, $Q_j$ is $j-th$ quantitative feature. $T_j$ is $j-th$ unigram or bigram in the vocabulary. Similar to Kogan et al.'s method [Kogan et al.2009], we work in logarithmic domain for the output variable.

The daily returns we are using in this paper are log daily returns:

$$R_t = log(\frac{price_t}{price_{t-1}}) \quad (2)$$

where $price_t$ is the close price of today ($t$) while $price_{t-1}$ is the close price of yesterday ($t-1$).

# 3 Learning Framework

## 3.1 Standard Regression Model

For a typical regression problem, we apply a standard machine learning method of the least squares. Let $\{\mathbf{x}^i, y_i\}_{i=1}^N$ be our training corpus, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{im})^{\mathrm{T}}$ is a vector of observations, and $y_i$ is a predictor. We solve the following function.

$$\ell(\boldsymbol{\theta}) = \arg\min \{ \sum_{i=1}^{N}(y_i - \sum_{j=1}^{m}\theta_j x_{ij})^2 \}. \quad (3)$$

The function $\ell(\boldsymbol{\theta})$ can be penalized by the $L_2$ penalty term $\lambda_2 \sum_{j=1}^{m}\theta_j^2$ to avoid over-fitting the model, and to avoid numerical problems.

Tibshirani et al. [Tibshirani1996] proposed to penalize the objective function by the $L_1$ penalty term to induce sparsity on features:

$$f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda_1 \|\theta\|_1, \quad (4)$$

where $\|\theta\|_1 = \sum_{j=1}^{m}|\theta_j|$.

The criterion presented in Eq. (4) is not differentiable at zero. Therefore, standard gradient-based approaches cannot be applied directly. A number of efficient optimization procedures has been already developed to cope with this problem. In our experiments, we use the Orthant-Wise-Limited Quasi-Newton [Andrew and Gao2007].

A compromise between sparsity and numerical stability of the solution is the elastic net penalty term, introduced by Friedman [Friedman et al.2007]:

$$f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{m}|\theta|_j + \lambda_2 \sum_{j=1}^{m}\theta_j^2, \quad (5)$$

where $\lambda_1$ and $\lambda_2$ are the hyper-parameters associated with the $L_1$ and $L_2$ norms. In our experiments, we exploit this regularization.

## 3.2 Regression Model with Unequal Penalty Factors

We address a regression problem using heterogeneous sources for predicting the daily returns of stocks with unequal penalty factors. We modify Eq. ( 5) to handle the unequal penalty factors:

$$f(\boldsymbol{\theta}) \quad = \quad \ell(\boldsymbol{\theta}) + \lambda_1^q \sum_{j=1}^{k} |\theta_j| + \lambda_2^q \sum_{j=1}^{k} \theta_j^2 + \lambda_1^t \sum_{j=k+1}^{m} |\theta_j| + \lambda_2^t \sum_{j=k+1}^{m} \theta_j^2 \qquad (6)$$

where $\lambda_1^q$, $\lambda_1^t$, and $\lambda_2$ are the hyper-parameters associated with the $L_1$ and $L_2$ norms. $\lambda_1^q$, $\lambda_1^t$ are the hyper-parameters associated with the $L_1$ norm for quantitative features and text features, respectively. $\lambda_2^q$ and $\lambda_2^t$ are hyper-parameters associated with $L_2$ norm for quantitative and text features, respectively.

## 3.3 Orthant-Wise-Limited Quasi-Newton with Unequal Penalty Factors (OWL-QN-UPF)

For solving the Eq.( 6), we modify the Orthant-Wise-Limited Quasi-Newton [Andrew and Gao2007] for the purpose of handling unequal penalty factors.

**Algorithm 1** OWL-QN-UPF

choose initial point $x^0$
$S \Leftarrow \{\}, Y \Leftarrow \{\}$
**for** $k = 0$ to MaxIters **do**
  Compute $\boxed{v^k = - \diamond_i f(\boldsymbol{\theta}^k)}$ (1)
  Compute $d^k \Leftarrow \mathbf{H_k} v^k$ using $S$ and $K$
  $p^k \Leftarrow \pi(d^k; v^k)$
  Find $\boldsymbol{\theta}^{k+1}$ with constrained line search
  **if** termination condition satisfied **then**
    Stop and return $\boldsymbol{\theta}^{k+1}$
  **end if**
  Update $S$ with $s^k = \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k$
  Update $Y$ with $y^k = \nabla \ell(\boldsymbol{\theta}^{k+1}) - \nabla \ell(\boldsymbol{\theta}^k)$
**end for**

A pseudo-code description of OWL-QN-UPF is given in algorithm 1, which is the same as original OWL-QN except one step (1) in calculating $\diamond f(\boldsymbol{\theta}^k)$, according to

$$\diamond_i f(\boldsymbol{\theta}) = \begin{cases} \partial_i^- f(\boldsymbol{\theta}) & \text{if} \quad \partial_i^- f(\boldsymbol{\theta}) > 0 \\ \partial_i^+ f(\boldsymbol{\theta}) & \text{if} \quad \partial_i^+ f(\boldsymbol{\theta}) < 0 \\ 0 & \text{otherwise,} \end{cases} \qquad (7)$$

where the left and right partial derivatives of $f$ are

$$\partial_i^{\pm} f(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_i} \ell(\boldsymbol{\theta}) + \begin{cases} \lambda_i \sigma(\theta_i) & \text{if} \quad \theta_i \neq 0 \\ \pm \lambda_i & \text{if} \quad \theta_i = 0 \end{cases} \tag{8}$$

The only difference between OWL-QN and OWL-QN-UPF is in Eq.( 8). The OWL-QN uses unique penalty (one constant number) on all weights, while OWL-QN-UPF applies particular penalty on each weight. In our case, we use two distinctive penalties on quantitative and text features, thus, we set $\lambda_i$ with same value if corresponding feature belongs to the same group, say, $\lambda_1^q$ and $\lambda_1^t$.

# 4 Australia Security eXchange (ASX) Data Set

Australia Security eXchange (ASX) provides announcements of all listing companies on its website. The announcement is an official report made by the company itself, which covers all important affairs, e.g. takeover, periodic reports, issued capital, asset acquisition & disposal, dividend announcements.

The announcements are collected during the first half of 2010 from ASX. In total, we collect 19, 282 documents, all announcements published in the same day of a company are aggregated. The announcements have been downloaded from the ASX website[1]. All announcements were initially in PDF format then converted to text documents. Since the conversion between PDF and text is noisy, we filter out low frequent terms (we remove all terms from vocabulary if its term frequency in the corpus is lower than 50). Finally, we get 62,115 unigram and 802,788 bigram as the vocabulary.

The quantitative features (mixture of time-series and cross-sectional features) used in our experiments include the daily returns that one to six days before announcements ($R_{t-1}$ *to* $R_{t-6}$) and five cross-sectional features: stock close price($price$), ASX price sensitive label ($asx\_ps$), whether published in trading time ($isInTradingTime$), previous trading value ($pre\_value$), and decile by capital ($cap\_decile$). The quantitative features and announcements are aligned according to their publishing time stamp.

# 5 Experiments on Australia Security eXchange (ASX) Corpus

In this section, we apply the model to ASX corpus. The data set we collected from ASX is split into two parts (80% and 20%) for training and testing algorithms, respectively. The training corpus is divided in its turn, into two parts (80% and 20%), into a main training and a holdout set. We use the holdout data set to search the optimal hyper-parameter values.

## 5.1 Significance Test

In many systems, the levels of error reduction that we report would be unlikely to have much effect on the utility of the system. For instance, for searching and translation, the

[1] `http://www.asx.com.au/`

output is presented to a user who is somewhat able to ignore system errors. However, in the finance domain, any statistically significant improvement may be of commercial benefit, for instance, in improving a high-frequency trading strategy. We therefore rely on statistical significance tests to assess our models, rather than attempting to comment on whether the magnitude of error reduction we achieve is substantial or not.

During each comparison of two methods, we train two models using the same training data set. For each model and for each instance in the test data set, we calculate the squared error. Then calculate the number of test examples for which each model is better. Then do a t-test with the null hypothesis that each model is equally likely to be best.

## 5.2 Quantitative, Text versus Quantitative + Text Features

Table 1 illustrates our results. As aforementioned, there are two groups of features in our corpus: quantitative and text features. Their impacts are quite different, in order to obtain a baseline, we train (and test) the regression on qualitative and text features separately.

We employ an average predictor as the baseline for comparison purpose. The predictor averages the daily returns in training data set, and use this constant as the daily returns for all examples in testing data set.

| Feature Set | Optimal hyper-parameter | MSE | MSE Reduction(%) |
|---|---|---|---|
| Baseline | | 0.0027756 | 0% |
| Qualitative | $\lambda_1 = 2.5, \lambda_2 = 0.1$ | 0.0027661 | 0.34% |
| Text | $\lambda_1 = 2.5, \lambda_2 = 0.1$ | 0.0027556 | 0.721% |
| Quantitative + Text | $\lambda_1 = 2.5, \lambda_2 = 0.1$ | **0.00271809** | 2.072%$^*$ |

Table 1: Performance comparison between Baseline, Quantitative, Text, and Quantitative+Text Features.$^*$ indicates that $p$-value is less than 0.05.

We calculate the MSE reduction for each method against baseline. The two lines of Table 1 under the baseline demonstrate that both methods outperform the baseline (although the reduction is not significant), and text features alone lead to slightly better performance than the qualitative features only.

As Kogan et al. [Kogan et al.2009] shows in their paper, the model using combination of qualitative and text features outperforms the models that use each of them and achieves significantly lower MSE compared against baseline. The fourth line indicates that their conclusion has been demonstrated on our data set as well.

## 5.3 Unequal Penalty factors on Heterogeneous Features

The most accurate result we obtain was to penalize the two groups of features with different penalty factors. As shown in table 2, the model significantly decreased the MSE reduction and obtained almost 3% improvement. Although 3% is not a big improvement in the computer science, it has a significant impact in finance. Imagine that we invest 1 million dollars in stock market, 3% improvement equals to 30,000 dollars.

As a fair comparison, we use the same $\lambda_1^t$ on text features and different $\lambda_1^q$ penalty on quantitative features. The optimal penalty value for the text group is higher than for the quantitative group. We expect the difference, since the number of text features (bigrams and unigrams) is quite large, and we expect the solution to be sparse.

Note that the results reported in the table are the values obtained with the optimal hyper-parameter values.

| Feature Set | Optimal hyper-parameter | MSE | MSE Reduction (%) |
|---|---|---|---|
| Baseline | | 0.0027756 | 0% |
| Qualitative + Text | $\lambda_1^q = 5, \lambda_1^t = 5, \lambda_2^q = 1, \lambda_2^t = 1$ | 0.0027176 | 2.09% *** |
| Qualitative + Text with different penalty | $\lambda_1^q = 1.3, \lambda_1^t = 5, \lambda_2^q = 1, \lambda_2^t = 1$ | **0.0026945** | 2.92%*** |

Table 2: Performance comparison between equal and unequal penalties. *** indicates that $p$-value is less than 0.001.

### 5.4 Qualitative Analysis of Results

In our experiments, we consider a high number of text features. Table 3 shows the bigram and unigram features which have the maximal weights, in other words, these features are considered by the regression model to be the most important. For instance, *high*, *pleased*, *high-grade* are obvious positive terms.

Table 4 shows the features with the most negative weights. For example, *issues*,*may*, *energy-limited*, *corporation-limited*, *group-limited* are negative terms.

| Unigram Feautre | Weight | Bigram Feature | Weight |
|---|---|---|---|
| stage | 0.00207106 | securities-exchange | 0.000816 |
| very | 0.00169215 | has-been | 0.00018 |
| high | 0.00148873 | pty-ltd | 6.29E-05 |
| employee | 0.00137991 | company-has | 3.55E-05 |
| info | 0.00135883 | australian-securities | 3.54E-05 |
| australian | 0.00126658 | high-grade | 3.33E-05 |
| units | 0.00116373 | fair-value | 3.12E-05 |
| billion | 0.00091421 | income-tax | 3.12E-05 |
| pleased | 0.00079896 | financial-assets | 2.91E-05 |
| april | 0.00075961 | which-has | 2.47E-05 |

Table 3: Unigram and bigram features with the maximal weights

## 6 Conclusion

In this paper, we have introduced a newly created text regression corpus used to predict the stock's daily returns and proposed a novel text regression model with different penalty factors. Furthermore, the model has been applied to predict the daily returns of

| Unigram Feautre | Weight | Bigram Feature | Weight |
|---|---|---|---|
| indices | -0.0063438 | during-quarter | -0.00138 |
| issues | -0.0034011 | and-has | -0.00095 |
| quarterly | -0.0031819 | energy-limited | -0.0007 |
| quarter | -0.0020532 | corporation-limited | -0.00066 |
| june | -0.001882 | group-limited | -0.00049 |
| raising | -0.0017045 | share-purchase | -0.00044 |
| minerals | -0.0013457 | code-name | -0.00025 |
| spp | -0.0012596 | purchase-plan | -0.00025 |
| continue | -0.0008822 | cash-end | -0.00021 |
| may | -0.000839 | consolidated-statement | -0.00021 |

Table 4: Unigram and bigram features with the least weigths

Australia Security Exchange. A number of experiments with different feature sets were conducted with the finding that the best performance can be achieved by combining different features of groups but with different penalty factors.

Text features in this paper, however, were considered as a "bag-of-words" model, ignoring all possible structures. We are interested applying conditional random fields model, which takes the structure of the data into consideration.

# References


[Andrew and Gao2007] Andrew, G. and Gao, J. (2007). Scalable training of $l_1$-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA. ACM.

[Baker and Wurgler2006] Baker, M. and Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, 61(4):1645–1680.

[Baker et al.2012] Baker, M., Wurgler, J., and Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2):272–287.

[Fama1970] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

[Friedman et al.2007] Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 2(1):302–332.

[Kogan et al.2009] Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risks from financial reports with regression. In *NAACL*.

[Stambaugh et al.2012] Stambaugh, R. F., Yu, J., and Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2):288–302.

[Tibshirani1996] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.