Forecasting Retained Earnings of Privately Held Companies with PCA and L^1 Regression

Harish S. Bhat*† Dan Zaelit‡

October 15, 2012

Abstract

We use proprietary data collected by SVB Analytics, an affiliate of Silicon Valley Bank, to forecast the retained earnings of privately held companies. Combining methods of principal component analysis (PCA) and L^1/quantile regression, we build multivariate linear models that feature excellent in-sample fit and strong out-of-sample predictive accuracy. The combined PCA and L^1 technique effectively deals with multicollinearity and non-normality of the data, and also performs favorably when compared against a variety of other models. Additionally, we propose a variable ranking procedure that explains which variables from the current quarter are most predictive of the next quarter's retained earnings. We fit models to the top five variables identified by the ranking procedure and thereby discover interpretable models with excellent out-of-sample performance.

Keywords: L^1 regression; principal component analysis; private companies; quantile regression; forecasting

1 Introduction

In the United States, privately held companies are not typically required to file financial statements with the Securities and Exchange Commission (SEC). This is in sharp contrast to publicly owned companies, which are required to submit quarterly 10-Q and annual 10-K statements to the SEC. This contrast extends itself to statistical studies. The bulk of the literature on quantitative forecasting of financial variables deals with publicly owned companies, because financial data for such companies is relatively easy to acquire.

In this paper, we analyze data on privately held companies maintained in SVB Analytics' (SVBA) proprietary database. We use this data to develop models that use financial variables from the current quarter to predict retained earnings for the next quarter, and we also identify which variables are the most predictive of retained earnings. The data is described in more detail below in Section 2.

The findings of this paper are significant for two reasons. First, we expect that the statistical methodology developed here can be fruitfully applied to other data sets, both internal and external to SVBA. Principal component analysis (PCA) and quantile/L^1 regression have been applied separately in many studies, but, to our knowledge, this is the first study in which the combination of these techniques is explored in the context of forecasting. Second, the models we describe are the first statistical models built using SVBA's database of financial statements. The conclusions of this study are designed to confirm and extend domain-specific

*School of Natural Sciences, University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343 (email: hbhat@ucmerced.edu).

†The first author gratefully acknowledges Dave Krimm and Tony Yeh of SVB Analytics for supporting this collaboration, both by providing access to data and through discussions of this and future work.

‡SVB Analytics, 555 Mission St., San Francisco, CA, 94105.

knowledge of the dynamics of privately held companies. Given their predictive accuracy, the models built here can potentially be used to improve models for credit scoring [Fernandes, 2005] and warrant pricing [Lauterbach and Schultz, 1990] for privately held firms.

1.1 Summary of Results

Using principal component analysis (PCA) and 10-fold cross-validation, we reduce the dimension of the covariate space from 87 to 35. This dimensionality reduction solves the problem of significant multicollinearity in the original 87-variable data set.

We then apply L^1 regression—also known as Least Absolute Deviation (LAD) regression—to the PCA-transformed, lower-dimensional data set. Through a number of subsequent tests, we find that the models built using this combination of PCA and L^1 regression possess excellent in-sample fit: analyses of regression residuals from both L^1 and L^2 (ordinary least squares) models reveal that the residuals fit the Laplace distribution far better than they fit the normal distribution. This is the first of several indications of the appropriateness of L^1 regression. The development of these statistical methods and resulting in-sample tests are described in Section 3.

Once we have a model where quarter $q+1$ retained earnings have been fit to quarter q covariates, we perform out-of-sample tests. We apply the model to covariates from quarter $q+1$ and see how well we do at predicting quarter $q+2$ retained earnings. In Section 4, we report in detail the results of these tests, which show that PCA plus L^1 regression outperforms four competing methods: PCA plus L^2 regression and three nonlinear, nonparametric regression approaches. Moreover, when we apply quantile regression to the PCA-transformed data set, we are able to generate accurate interval forecasts.

While the models built using 35 PCA-transformed covariates are predictive, they do not by themselves help answer the question of which of the original variables in the data set are most predictive of retained earnings. To address this, we describe in Section 5 a method for using the PCA plus L^1 model to rank quarter q variables in order of importance to the regression model for quarter $q+1$ retained earnings. Using the top five such variables, we develop pruned and simplified models with improved out-of-sample performance. One of our main findings here is that once the most predictive variables have been identified using the PCA plus L^1 approach, different robust regression approaches may be applied to yield interpretable models with excellent out-of-sample predictive power.

In Section 5.3, we apply the PCA plus L^1 methodology to quarterly financial statement data for publicly traded companies in the S&P 500 index. The data includes 38 covariates and covers the same period of time as the SVBA data. The 38-dimensional data set again displays a high level of multicollinearity, so we use PCA to reduce the dimensionality to 28, after which we find that L^1 regression models again outperform L^2 regression models as well as competing nonlinear, nonparametric approaches. An interesting difference is that while the PCA plus L^1 results for the SVBA data set show strong consistency across 11 quarters of testing, for the S&P 500 data set, we observe one large jump in quarterly error that coincides with the onset of an economic recession. We note two other differences. Net income is a highly predictive variable for the SVBA models, and this variable is effectively replaced by net worth for the S&P 500 models. The relative error for the S&P 500 models is roughly one percentage point higher than for the SVBA models. Overall, though the S&P 500 results are preliminary, they serve as a concrete demonstration that the techniques developed in this paper can be effectively applied to other data sets.

1.2 Prior Work

The literature on statistical modeling of privately held firms is not nearly as large as that on publicly held firms. One way this asymmetry of information manifests is in the estimation of CAPM (Capital Asset Pricing Model) betas of privately held companies; a popular method is to use comparable public companies

for which data is readily available [Bowman and Bush, 2006]. Despite this asymmetry, there do exist various financial databases for privately held companies. Our review of research that studies these data sets is by no means exhaustive, as our aim is to put our work in the context of relevant studies.

There is a relatively large amount of work on the problems of assessing the probability that a privately held company will default on a loan or go bankrupt. Here we find, for example, a discrete-time hazards model applied to data on 7711 individual firms collected by Intesa SanPaolo [De Leonardis and Rocci, 2008], generalized additive modeling applied to data on Norweigian limited liability firms [Berg, 2007] and probit modeling applied to data from the Bureau van Dijk FAME database of U.K. firms [Bunn and Redwood, 2003]. As with the SVBA data analyzed here, these data sets comprise financial statements such as balance sheets and income statements. [Mramor and Valentincic, 2003] used a database of nearly 20,000 Slovenian companies to develop a liquidity forecasting model—note that in Slovenia, the government collects financial statements from all companies, including young startup companies, again enabling the authors to use balance sheets, income statements, and other data points for each company in their study. In the U.S., such data is not ordinarily collected from privately held companies, making the SVBA database a rare source.

A noteworthy study involving U.S. firms is that of [Hand, 2005]. Privately held companies that file for an initial public offering (IPO) must provide five years of audited historical financial statements; [Hand, 2005] uses this data source in conjunction with firm valuations data (obtained from Recombinant Capital) to establish a close relationship between financial statement data and equity values for privately held firms. More recently, [Minnis, 2011] has analyzed private firm financial statements collected by Sageworks to show that firms that provide their lenders with audited financial statements enjoy a significantly lower cost of debt.

Many studies on privately held U.S. firms have utilized commercial databases from Thomson VentureXpert [Bhat and Zaelit, 2011] or Venture Economics [Tolkamp, 2007]. These databases do not contain financial statements, either audited or otherwise, and are typically used for their qualitative data (such as which investors have invested in each company) or financial variables that have been aggregated across either time or industry sector.

2 Description of the Data

SVB Analytics (SVBA) compiles regularly submitted financial statements provided by clients. These financial statements are audited prior to delivery and are comprised of classical balance sheet and income statement metrics that are reliable and rich in detail.

The data utilized in this study is a subset of this data set, spanning 13 quarters from Q1 2008 to Q1 2011, and only consisting of those companies whose last twelve months of revenue is less than $75 million. Note that the names or other equivalent identifying information of the clients were *not* included in the data set analyzed here. The analysis focused on the performance of statistical models across the entire data set—no client's data was analyzed individually. However, it is known that, in the aggregate, these clients predominantly consist of privately held companies.

The primary focus of this paper is on modeling past, present, and future SVBA clients, not necessarily all possible privately held companies. The data used in this study reflects this in two underlying biases: companies represented in the data have debt in their capital structure and have passed SVBA's initial viability and risk assessments. Despite these biases, revenue and affiliated metrics are well dispersed, and the collection of companies represented in the data consist of a variety of technology and life science companies that are in different stages of their lives.

The subset of SVBA data used here has not been used for prior studies in either the statistical or financial literature. This paper represents the first attempt to utilize the data for any type of forecasting.

As the database is proprietary, we describe the variables in terms of broad categories rather than spe-

cific names. Balance sheet assets are measured by $\{X_j\}_{j=1}^{j=27}$, liabilities by $\{X_j\}_{j=28}^{j=47}$, and net worth by $\{X_j\}_{j=48}^{j=56}$. Income statement revenue is measured by $\{X_j\}_{j=57}^{j=59}$, expenses by $\{X_j\}_{j=60}^{j=69}$, and other items by $\{X_j\}_{j=70}^{j=85}$. We include two unitless ratios X_{86} and X_{87}; all other variables are measured in units of thousands of dollars.

In Table 1, we present summary statistics for each of the 87 covariates. More specifically, we present the mean (μ), standard deviation (σ), percent of samples that lie within one standard deviation of the mean (% Conc), excess kurtosis (γ), and range (Rng) for all 87 covariates. Here the covariates are aggregated across all quarters from Q1 2008 to Q1 2011.

By excess kurtosis, we mean the sample excess kurtosis computed using the default kurtosis function from the R utility package `e1071`. This function corresponds to the b_2 formula described in the literature [Joanes and Gill, 1998]. Since we have aggregated 13 quarters worth of data, we have a large sample size of $n = 15411$ and the differences between the sample excess kurtosis functions are negligible [Joanes and Gill, 1998].

For a normally distributed random variable, the excess kurtosis vanishes ($\gamma = 0$), and the probability of obtaining values within one standard deviation of the mean is ≈ 0.683. In Table 1, the large values of γ and % Conc indicate significant departure from normality for the marginal distributions of each X_j. We omit quantile-quantile plots comparing the empirical quantiles of each X_j to those of the normal distribution, but simply note that all such plots are clearly nonlinear, confirming non-normality of the X_j's.

By range, we mean the difference between the maximum and minimum sample values of the covariate. The large values of Rng in Table 1 together with the large values of % Conc indicate the following. For each $j \in [1, 87]$, despite the fact that the values of X_j are very likely to be within standard deviation of the mean, we can always find ≥ 1 companies that display extreme behavior in X_j.

2.1 Preliminary Considerations

We number the quarters from Q1 2008 to Q1 2011 using integers $q \in \{1, 2, \ldots, 13\}$. Let N_q denote the number of companies for which we have data from quarter q. Examining the raw data, it is clear that (a) N_q fluctuates as a function of q, and (b) each N_q is smaller than 1844, the total number of unique companies.

We plot in the left panel of Figure 1 the number of companies for which we have exactly z quarters worth of data, as z goes from 1 to 13. Less than 20% of companies are represented for all 13 quarters. For any given company, the actual list of quarters for which we have data may not be consecutive. This list will also vary from one company to the next. These facts motivate us to look at a sequence of one-quarter-ahead models rather than a single model fit to multiple quarters' worth of data.

2.2 Consecutive Quarter Intersections

Let C_q denote the set of companies for which we have data in two consecutive quarters q and $q+1$, as q varies from 1 to 12. Let $|C_q|$ denote the number of companies in the set C_q. In the right panel of Figure 1, we plot $|C_q|$ versus q. Note that $|C_q| > 1000$ for all but the last quarter $q = 12$. Moreover, one checks that $|\bigcup_{q=1}^{12} C_q| = 1746$. By looking at all intersections of consecutive quarters, our analysis covers $1746/1844 = 94.7\%$ of the companies in the data set.

Hence we form 12 pairs of matrices $\{(X_q^0, X_q^1)\}_{q=1}^{12}$. Here X_q^0 contains the quarter q data for all companies in C_q, and X_q^1 contains the quarter $q+1$ data for all companies in C_q. Both matrices X_q^j are of size $N \times p$ where $N = |C_q|$ and $p = 87$.

Let r_q^1 denote the vector of all quarter $q+1$ retained earnings for all companies in C_q. The elements of r_q^1 have units of thousands of dollars. Let m_q denote the median of r_q^1, and then define the median absolute deviation (MAD):

$$\mathrm{MAD}(r_q^1) = \mathrm{median}\left|r_q^1 - m_q\right|.$$

For $q = 1, \ldots, 12$, we give in Table 2 the minimum, median, maximum, and median absolute deviation of r_q^1. Note that the median values are all on the order of 10^4. Note also that we use the term retained earnings though the actual value may be negative and therefore represent accumulated loss.

2.3 Statistical Goal

With these definitions, we can state the first statistical goal of this paper: estimation of regression functions f_q that use quarter q information contained in X_q^0 to forecast quarter $q+1$ retained earnings r_q^1, i.e.,

$$r_q^1 = f_q(X_q^0) + \varepsilon_q. \tag{1}$$

We are interested in both the in-sample fit and out-of-sample performance of these models.

3 Statistical Methods and In-Sample Tests

3.1 PCA

Before proceeding, we review the basic theory behind PCA [Jolliffe, 2002]. Let $\mathbf{1}$ be an $N \times 1$ vector of ones. Let $\overline{X}$ be the $N \times p$ matrix such that the k-th column of $\overline{X}$ is the vector $\mu_k \mathbf{1}$, where μ_k is the mean of the k-th column of X. Then let

$$\widetilde{X} = X - \overline{X}, \tag{2}$$

a centered version of X where each column has zero mean. We now compute the singular value decomposition (SVD):

$$\widetilde{X} = V \Sigma W^T. \tag{3}$$

Here V is an orthogonal $N \times N$ matrix, W is an orthogonal $p \times p$ matrix, and Σ is an $N \times p$ matrix with p singular values along its diagonal and zero's elsewhere. The singular values are nonnegative and sorted in decreasing order.

Note that in the above discussion, we omitted superscripts and subscripts for readability. For our specific data matrices, we will have the decomposition $\widetilde{X}_q^j = V_q^j \Sigma_q^j (W_q^j)^T$ where all the matrices in the equation depend on j and q. In what follows, we will similarly omit superscripts/subscripts on the matrices Y and S.

The *principal components* are the columns of W, and the matrix

$$Y = \widetilde{X} W = V \Sigma \tag{4}$$

is the PCA-transformed data matrix. Note that the columns of W are the eigenvectors of the variance-covariance matrix $S = \frac{1}{N-1} \widetilde{X}^T \widetilde{X}$; the eigenvalues of S are given by $\frac{1}{N-1} \Sigma^T \Sigma$.

Since V is orthogonal, $Y^T Y = \Sigma^T \Sigma$, i.e., the variance-covariance matrix of Y is purely diagonal. By (4), multiplying $\widetilde{X}$ by W has the effect of decorrelating the covariates in the original data matrix. The columns of W can be interpreted as new covariates—each one a linear combination of the original covariates—that are perfectly decorrelated.

Let Σ' be the matrix obtained by starting with Σ and setting all but the p' largest diagonal entries to zero. Then define

$$\widetilde{X}' = V \Sigma' W^T.$$

This is a rank-p' approximation of $\widetilde{X}$. By the Eckart-Young theorem, the rank-p' approximation Z that minimizes the Frobenius norm $\|\widetilde{X} - Z\|_F$ is $Z = \widetilde{X}'$. This motivates the SVD/PCA as a tool for finding an optimal low-dimensional representation of the original data set, which we carry out below. This in turn gives us another interpretation of the columns of W_q as an optimal basis in which to represent the original data set.

We define the $N \times p'$ transformed data matrix by

$$Y' = \widetilde{X}'W' \equiv \widetilde{X}W', \tag{5}$$

where W' is the $p \times p'$ matrix obtained by retaining only the first p' columns of W.

Scaling. Note that the PCA described above is unscaled. Scaling refers to normalizing the columns of $\widetilde{X}$ so that they have variance one. We have found that with SVBA's data, the regression models using scaled PCA are worse (in both in-sample and out-of-sample tests) than those using unscaled PCA. In what follows, we omit further discussion of the scaled PCA.

3.2 Motivation and Results

To understand why the PCA is well-indicated for this data, we compute the condition numbers of the matrices Y_q^0—see the $p' = 87$ column of Table 3. The condition number is the ratio of the largest to the smallest singular value of Y_q^0. The enormity of these numbers indicates three issues: (i) the $p \times p$ matrices $(Y_q^0)^T Y_q^0$ are close to singular, (ii) the original data set possesses significant multicollinearity, and (iii) naïvely fitting an ordinary least squares (OLS) model of the form $r_q^1 = \alpha + Y_q^0\beta_q + \varepsilon_q$ is unsound. The multicollinearity of the data set is to expected—the columns of our original data set correspond to balance sheet and income statement variables, and many of these can be expected to be correlated, e.g., "accounts receivable" and "gross sales."

A standard idea to combat these problems is to use PCA-transformed, lower-dimensional representations of the data matrices [Jolliffe, 2002, Chap. 8]. In the $p' = 35$ and $p' = 20$ columns on the left half of Table 3, we record the condition numbers of the $N \times p'$ matrices $(Y_q^0)'$ computed using (5). The condition numbers for these matrices are much smaller than for the original data set.

Another way to view the effect of dimensionality reduction is to examine correlation matrices. Let Z_q (respectively, Z_q') denote the correlation matrix obtained from the data matrix Y_q^0 (respectively, $(Y_q^0)'$). The maximum absolute value of the non-diagonal entries of Z_q is given in the "Max" column of Table 3—for some quarters, there are covariates with significant correlation. Continuing into the right half of the table, the $p' = 87$ (respectively, $p' = 35$ and $p' = 20$) column gives the number of above-diagonal entries of Z_q (respectively, Z_q') that are at least 0.1 in absolute value. Note that the $p' = 35$ and $p' = 20$ columns are identically zero, again indicating that $(Y_q^0)'$ does not suffer from the multicollinearity of Y_q^0.

Putting together the results of Table 3, it is clear that using PCA to reduce the dimensionality of the data set remedies the three issues (i-iii) described above.

3.3 Selecting p'

The next PCA-related question to answer is: how do we choose p', the number of columns of $(Y_q^0)'$? In the left panel of Figure 2, we plot the j-th singular value Σ_{jj} of the centered data matrix $\widetilde{X}_1^0$ from Q1 2008. The plots for other quarters $q \geq 2$ look qualitatively the same. The plot shows that if our goal were merely to devise matrices $(Y_q^0)'$ that closely approximate Y_q^0, then we would expect $p' = 20$ to be an excellent choice.

As our goal is instead to use $(Y_q^0)'$ to predict r_q^1, we remind the reader that the PCA was performed only on the data matrices X_q^0. PCA variables that correspond to small singular values may in fact be good predictors of r_q^1. The right panel of Figure 2 shows $\log(\Sigma_{jj})$ versus j. This shows that choosing any $1 \leq p' \leq 80$ will yield a matrix $(Y_q^0)'$ with a far better (i.e., smaller) condition number than the full matrix Y_q^0.

Our strategy is to select p' based on 10-fold cross-validation. For each $q \in \{1, 2, \ldots, 12\}$, we randomly partition C_q into 10 disjoint subsets $C_{q;i}$ of approximately equal size. Let $X_{q;i}^j$ denote the data matrix X_q^j restricted to only those companies in $C_{q;i}$.

We loop over each fold i, each time taking $(X^0_{q;i}, X^1_{q;i})$ to be the test set. For each i, the training data consists of all rows of X^0_q and X^1_q that are *not* present in the test set. We apply the PCA to the training set data, reducing the number of columns in the $j = 0$ data matrix to $p' \in [5, 80]$. We then fit L^1 and L^2 regression models to the PCA-transformed training set data. Finally, we test the performance of these models on the test set data that has been held out. Our metric for test set performance is the median absolute deviation between true and predicted retained earnings. Note that the details of fitting and testing the models are described in detail below.

Averaging over the the 10 folds and over the 12 quarters, we obtain the results plotted in Figure 3. Note that both L^1 and L^2 test error curves decrease monotonically from $p' = 5$ until $p' = 35$. For $p' > 35$, the test set errors are either greater or only marginally less than the error at $p' = 35$. Therefore, in the remainder of the study, we use $p' = 35$ as our baseline value of p'.

3.4 Linear Models

In this section, we use Y'_q to denote $(Y^0_q)'$. We discuss two competing methods for fitting linear models.

L^2 (Ordinary Least Squares) Regression. We propose a model of the form

$$r^1_q = \begin{bmatrix}\mathbf{1} & Y'_q\end{bmatrix} \beta_q + \varepsilon_q \tag{6}$$

where β_q is a vector of $p' + 1$ unknown regression coefficients, and ε_q stands for the residual error. The column of 1's included in the matrix takes care of the intercept. We find β_q by minimizing the sum of squared residuals,

$$\|\varepsilon_q\|_2^2 = \varepsilon_q^T \varepsilon_q = \sum_{j=1}^{|C_q|} (\varepsilon_{q\,j})^2.$$

The solution β_q satisfies the normal equation

$$\begin{bmatrix}\mathbf{1}^T \\ (Y'_q)^T\end{bmatrix} \begin{bmatrix}\mathbf{1} & Y'_q\end{bmatrix} \beta_q = \begin{bmatrix}\mathbf{1}^T \\ (Y'_q)^T\end{bmatrix} r^1_q.$$

Note that $\mathbf{1}^T Y'_q = \mathbf{1}^T \widetilde{X}_q W'_q = \mathbf{0}^T$, since each column of $\widetilde{X}_q$ has mean zero. We let Σ''_q denote the matrix Σ'_q truncated to size $N \times p'$. Then

$$\begin{bmatrix}\mathbf{1}^T \\ (Y'_q)^T\end{bmatrix} \begin{bmatrix}\mathbf{1} & Y'_q\end{bmatrix} = \begin{bmatrix}N_q & \mathbf{0}^T \\ \mathbf{0} & (\Sigma''_q)^T \Sigma''_q\end{bmatrix},$$

which is a purely diagonal matrix, so the solution for β_q is trivial:

$$\beta_q = \begin{bmatrix}N_q^{-1} & \mathbf{0}^T \\ \mathbf{0} & ((\Sigma''_q)^T \Sigma''_q)^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{1}^T \\ (Y'_q)^T\end{bmatrix} r^1_q.$$

L^1 and Quantile Regression. We propose a model of precisely the same form as above:

$$r^1_q = \begin{bmatrix}\mathbf{1} & Y'_q\end{bmatrix} \beta_q + \varepsilon_q.$$

The only difference is in how we solve for β_q. In L^1 or Least Absolute Deviation (LAD) regression, we find β_q by minimizing the sum of absolute residuals,

$$\|\varepsilon_q\|_1 = \sum_{j=1}^{|C_q|} |\varepsilon_{q\,j}|. \tag{7}$$

Several numerical methods efficiently solve this minimization problem [Barrodale and Roberts, 1973; Bloomfield and Steiger 1980; Li and Arce, 2004].

Quantile regression [Koenker and Bassett, 1978; Koenker and Hallock, 2001; Koenker, 2005] is a generalization of the above procedure. For $\tau \in (0, 1)$, we define a tilted version of the absolute value function:

$$\rho_\tau(x) = \begin{cases} \tau x & x \geq 0 \\ (\tau - 1)x & x < 0. \end{cases}$$

Suppose we are given scalar data $\{\zeta_i\}_{i=1}^N$. Fix $\tau \in (0, 1)$ and consider

$$F_\tau(\xi) = \sum_{i=1}^{N} \rho_\tau(\zeta_i - \xi).$$

Let $\xi^* = \operatorname{argmin}_\xi F_\tau(\xi)$. One checks that ξ^* is equal to the τ-th quantile of the scalar data set $\{\zeta_i\}_{i=1}^N$. We thus define quantile regression as follows: for $\tau \in (0, 1)$, find β_q that minimizes

$$\|\varepsilon_q\|_{1;\tau} = \sum_{j=1}^{|C_q|} \rho_\tau(\varepsilon_{q\,j}). \tag{8}$$

Note that when $\tau = 0.5$, we have $\rho_\tau(x) = |x|/2$, and so minimizing $\|\varepsilon_q\|_{1;0.5}$ is the same as minimizing $\|\varepsilon_q\|_1$ as defined in (7).

Applied to our data set, we see that quantile regression builds a model for the τ-th quantile of the next quarter's retained earnings r_q^1. In the special case of $\tau = 0.5$, the procedure reduces to L^1 regression and produces a model for the median of the next quarter's retained earnings r_q^1.

3.5 Nonlinear Models

To compare against the linear models, we also fit three nonlinear, nonparametric regression models. Unlike the linear models, these models can be tuned using user-specified parameters. To determine optimal values of these parameters, we follow a 10-fold cross-validation procedure, similar to the one described in Section 3.3. As before, in the i-th fold of quarter q, the test set is $(X_{q;i}^0, X_{q;i}^1)$, and the training set consists of all rows of X_q^0 and X_q^1 that are *not* present in the test set. We fix $p' = 35$ and vary the parameters of the nonlinear model. For each parameter choice, we obtain the error between the test set retained earnings and the model retained earnings using the training set predictors. We measure this error using both root mean squared error (RMSE) and median absolute error (MAE).

Averaging over quarters and cross-validation folds, we determine the choice of parameters that minimizes each error metric in turn. We have taken care to choose intervals of parameters such that the minimum does not appear at the boundary of the interval—if it does, we rerun the cross-validation study using appropriately enlarged intervals.

Most importantly, we do not tune the parameters of the nonlinear models using out-of-sample data. We use 10-fold cross-validation on in-sample data to mimic the procedure that one would apply if one were creating true forecasts of the future using data that is only present now. We follow this philosophy of parameter selection for each of the nonlinear models.

Regression Trees. We use the R package `rpart` to test the performance of regression trees [Breiman, 1984]. The package allows the user to set several parameters; we use the default values of all parameters except for minsplit and cp. For a particular node, to determine whether the node should be split, rpart checks whether the number of instances associated to the node is at least equal to minsplit; if not, rpart will not split the node. Similarly, for a candidate split, the parameter cp is the minimum factor by which the loss function must be reduces.

Searching across 29 values of cp between 10^{-7} and 10^{-1} and 20 values of minsplit from 2 to 40, we find that the parameters that minimize test set RMSE are minsplit $= 2, \text{cp} = 4.0 \times 10^{-6}$, while the parameters that minimize test set MAE are minsplit $= 2, \text{cp} = 3.0 \times 10^{-6}$.

Random Forests. We use the R package `randomForest` to test the performance of random forests [Breiman, 2001]. Again, while there are many parameters that the user can set, we use the default values of all parameters except for ntree and mtry. Here ntree is the number of trees in the forest, and mtry is the number of covariates for which candidate splits are assessed. Searching across 20 values of ntree from 50 to 1000 and 17 values of mtry from 3 to 35, we find that the parameters that minimize test set RMSE are ntree $= 500, \text{mtry} = 35$, while the parameters that minimize test set MAE are ntree $= 950, \text{mtry} = 27$.

Boosted Trees. We use the `mboost` function in the R package `blackboost` to test the performance of boosted regression trees [Bühlmann and Hothorn, 2007]. The `mboost` function uses conditional inference trees as base learners for the boosting algorithm. We use default values of the parameters for the boosting algorithm, but vary two tree parameters, minsplit and cp, that have the same interpretation as the corresponding rpart parameters described above, except that "number of instances" is replaced by "sum of weights."

We search across 27 values of cp from 10^{-8} and 10^{-1} and 20 values of minsplit from 2 to 40, and find that the parameters that minimize test set RMSE are minsplit $= 22, \text{cp} = 0.01$, while the parameters that minimize test set MAE are minsplit $= 14, \text{cp} = 0.1$.

Remarks. In the remainder of this paper, when we report either $\|\cdot\|_2$ or RMS errors, the nonlinear regressions use the optimal RMSE parameters given above. Similarly, when we report either $\|\cdot\|_1$ or median absolute errors, the nonlinear regressions use the optimal MAE parameters given above. Also, when we describe the results of the above procedures, we refer to the models by the one word abbreviations "tree," "forest," and "boost."

3.6 In-Sample Results

For $1 \leq q \leq 11$, we fit each of the five models (L^1, L^2, tree, forest, boost) described above. Let us first examine the in-sample errors. For each model and each quarter q, we tabulate in Table 4 both the sum of squared errors $\|\varepsilon_q\|_2^2$ and the sum of absolute errors $\|\varepsilon_q\|_1$.

It is mathematically guaranteed that L^1 regression minimizes $\|\varepsilon_q\|_1$ and L^2 regression minimizes $\|\varepsilon_q\|_2^2$ across all *linear* models; these results are borne out in Table 4. There are two facts, consistent across all quarters, that were not predetermined. First, two of the three nonlinear methods, forest and boost, produced absolute and squared errors that exceed those produced by the linear models. Second, the model with the best in-sample fit is the single regression tree (tree). This indicates that for the data set considered here, regression trees do belong in the list of models against which to compare L^1 regression. However, as we find later, the tree method's out-of-sample results do not match its in-sample fit.

Next we examine the residuals ε_q for both the L^1 and L^2 models. For each quarter q, we use maximum likelihood estimation (MLE) to fit normal

$$f^N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

and Laplace

$$f^L(x;\lambda,b) = \frac{1}{2b} \exp\left(-\frac{|x-\lambda|}{b}\right) \tag{10}$$

densities to ε_q. These particular distributions arise from the following consideration: in the linear model (6), if the residuals ε_q are i.i.d. samples from a Laplace (respectively, normal) distribution, then L^1 regression (respectively, L^2 regression) yields the maximum likelihood estimate (MLE) for β_q.

We denote the MLE of a parameter (e.g., σ) using a hat (e.g., $\hat{\sigma}$). We form two hypotheses:

- H_0 (null): ε_q follows normal$(\hat{\mu}, \hat{\sigma})$
- H_1 (alternative): ε_q follows Laplace$(\hat{\lambda}, \hat{b})$

Let L_q^N and L_q^L denote the maximized values of the likelihood functions for the normal and Laplace densities, evaluated on the same set of residuals ε_q. We form the test statistic

$$T_q = \log \frac{L_q^N}{L_q^L}.$$

For both L^1 and L^2 residuals, we obtain one value of T_q for each $q \in \{1, 2, \ldots, 11\}$. The asymptotic distribution of T_q has been calculated [Kundu, 2005]—applying this calculation, we find that for all quarters, and for both L^1 and L^2 regression residuals, we can reject H_0 with a p-value less than 2.2×10^{-16}.

We can go further in our analysis of the residuals. Both the normal and Laplace densities are special cases of the exponential power distribution (EPD)

$$f(x) = \frac{1}{2p^{1/p}\Gamma(1+p^{-1})\sigma_p} \exp\left(-\frac{|x-\mu|^p}{p\sigma_p^p}\right). \tag{11}$$

For $p = 2$ and $p = 1$, respectively, $f(x)$ reduces to the normal and Laplace densities. Note that in the EPD, the decay rate of the distribution's tail is controlled by the parameter p. For both L^1 and L^2 residuals, we use numerical maximization of the likelihood function to find $\hat{\mu}$, $\hat{\sigma}_p$, and $\hat{p}$. The results are given in Table 5. In all cases, p is found to equal 1 to machine precision, again indicating that the Laplace distribution fits the residuals better than the normal.

In Figures 4 and 5, we plot three cumulative distribution functions (CDFs) for the residuals ε_q from L^1 regression. In black, we plot the empirical CDF, while in blue and in red, we plot the fitted Laplace and normal CDFs. These plots suggest that not only is the Laplace distribution a better fit to the residuals than the normal, but also that the Laplace distribution fits the residuals closely in absolute terms.

Overall, the analysis of residuals helps to understand why L^1 regression works well for this data set. For an established, publicly traded company, large fluctuations in retained earnings may be viewed as rare events. For young startup companies, on the other hand, such events are much more likely. This fact shows itself in our analysis; the residuals' distribution, like the Laplace distribution, has heavier tails than the Gaussian distribution. Quantile/L^1 regression is one of several regression procedures that are robust with respect to heavy-tailed errors [Fox, 2008, Chap. 19], making it well-suited for the problem considered here.

4 Out-of-Sample Tests

In what follows, hats signify statistical estimates. For example, r_q^1 and $\widehat{r_q^1}$ are, respectively, the true and estimated retained earnings from quarter $q+1$ for all companies in C_q. Suppose we have already used either L^2 or L^1 regression to determine the vector of unknown coefficients β_q. Then our model is

$$\widehat{r_q^1} = \begin{bmatrix} \mathbf{1} & (Y_q^0)' \end{bmatrix} \beta_q, \tag{12}$$

with residual error

$$\varepsilon_q = r_q^1 - \widehat{r_q^1}. \tag{13}$$

We now describe how to apply this model to forecast retained earnings for future quarters. Let us use (5) and (2) to write

$$\widehat{r_q^1} = f_q(X_q^0) = \begin{bmatrix} \mathbf{1} & (X_q^0 - \overline{X_q^0})(W_q^0)' \end{bmatrix} \beta_q. \tag{14}$$

The function f_q takes as input the quarter q data matrix X_q^0 and produces as output the vector of quarter $q+1$ retained earnings forecasts $\widehat{r_q^1}$.

Suppose we have the data matrix X_θ^0 from quarter $\theta > q$. We can use this data to forecast retained earnings r_θ^1 for quarter $\theta + 1$. Our forecast will be

$$\widehat{r_\theta^1} = f_q(X_\theta^0) = \begin{bmatrix} \mathbf{1} & (X_\theta^0 - \overline{X_q^0})(W_q^0)' \end{bmatrix} \beta_q. \tag{15}$$

4.1 Results

For each $q = 1, 2, \ldots, 11$, we use $(Y_q^0)'$ and r_q^1 to build L^1, L^2, and nonlinear regression models. To test the out-of-sample performance of these models, we apply them on the next quarter's data, i.e., $\theta = q + 1$. We therefore obtain retained earnings forecasts for 11 quarters $\theta + 1 = 3, 4, \ldots, 13$.

In Table 6, we report the out-of-sample performance for the five models (L^1, L^2, tree, forest, and boost) across 11 quarters of testing. To the left of the double vertical bar, we report the RMSE

$$E_\theta^{\text{RMSE}} = \sqrt{\frac{1}{|C_\theta|} \|r_\theta^1 - \widehat{r_\theta^1}\|_2^2}.$$

To the right of the double vertical bar, we report the MAE

$$E_\theta^{\text{MAE}} = \text{median} \left| r_\theta^1 - \widehat{r_\theta^1} \right|.$$

Across all quarters of testing, and in both RMSE and MAE metrics, the L^1 model has the lowest out-of-sample error. The performance of the L^1 model also shows strong consistency across different quarters. The difference between the performance of the linear and nonlinear models is striking. Note that the tree model, which had the smallest in-sample errors in Table 4, does not perform well out-of-sample. Based on our testing, there is no reason to prefer any of the nonlinear models over the L^1 model.

Recalling Table 2, we see that the median absolute out-of-sample error for the PCA plus L^1 model is, roughly, two orders of magnitude less than the median value of the quantity being forecast, retained earnings.

In Figure 6, we plot the median absolute errors E_θ^{MAE} as a function of θ. We obtain one curve for each of four models included in our test—the errors for the tree model are excluded because their average is near 1400. The L^1 model consistently displays out-of-sample errors less than any of the competing models. Moreover, the L^1 model's error is consistent from one quarter to the next.

We have reported results for models trained on the PCA-transformed data set $(Y_q^0)'$ with $p' = 35$ columns. Note that when we change p' to either 20 or 80 and keep all other parameters the same, the errors of all the models increase. Though the L^1 model still has the least error, the out-of-sample results again indicate that $p' = 35$ is a good choice for the number of columns to retain in the PCA-transformed data matrices.

Also in Figure 6, we plot scatterplots of the predicted retained earnings $\widehat{r_\theta^1}$ versus true retained earnings r_θ^1 aggregated across all 11 quarters of out-of-sample testing. Note that the correlation coefficient is 0.9952. We include a line of slope one that goes through the origin—if the models were perfect, all points would sit on the line. Note that we have omitted quarter-by-quarter scatterplots because they all look nearly the same—the large majority of points lie close to the line, and occasionally, we find large differences between true and predicted retained earnings. Overall, the L^1 models feature excellent out-of-sample performance.

4.2 Interval Forecasting

Thus far we have been focused on point forecasting. For a given company, a point forecast is a single number that is our best estimate of the next quarter's retained earnings. Here we explore a different approach, that of interval forecasting. For each company, we seek an interval $[a, b]$ that has a certain probability δ of containing the true retained earnings.

This approach is easily implemented using quantile regression. For a given value of δ, we take $\tau^\pm = 0.5 \pm \delta/2$, so that $\tau^+ - \tau^- = \delta$. We perform quantile regression with these two values of τ, obtaining regression coefficients β_q^+ and β_q^-. Using these coefficients in (15), we obtain the upper $\widehat{r_\theta^+}$ and lower $\widehat{r_\theta^-}$ ends of the forecast intervals for all quarter $\theta + 1$ retained earnings:

$$\widehat{r_\theta^\pm} = f_q(X_\theta^0) = \left[\mathbf{1} \quad (X_\theta^0 - \overline{X_q^0})(W_q^0)'\right] \beta_q^\pm$$

To evaluate accuracy, we examine

$$E_\theta^\delta = \frac{\#\left\{r_\theta^1 \in \left[\widehat{r_\theta^-}, \widehat{r_\theta^+}\right]\right\}}{|C_\theta|}, \tag{16}$$

the fraction of all true quarter $\theta + 1$ retained earnings that actually lie within the forecast intervals.

In Table 7, we give the results of this procedure for values of δ in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. For each value of δ, and for each of 11 quarters of out-of-sample testing, we see that E_θ^δ is close to δ for each quarter θ. In short, the interval forecast is accurate: the empirical probabilities E_θ^δ are close to the desired probabilities δ.

For reference, we also give the mean width of the forecast interval,

$$W_\theta^\delta = \text{mean}\left(\widehat{r_\theta^+} - \widehat{r_\theta^-}\right), \tag{17}$$

for each value of δ and each quarter θ. The mean is taken over all companies from quarter θ. As expected, these intervals grow in size as δ increases.

Given the consistency of the results from one quarter to the next, we could easily adapt this procedure in two different ways. First, if we desire empirical probabilities E_θ^δ that exceed some fixed value δ_0, we could determine how much larger than δ_0 we should take δ. Second, if we desire forecast interval widths W_θ^δ less than some fixed value, we could determine how large we can take δ. Table 7 indicates that both of these goals are achievable, and hence that this procedure could be adapted for various applications.

5 Variable Ranking and Model Pruning

Thus far our model (14) is expressed in terms of 35 PCA-transformed variables. This relatively large number of covariates, each of which is a linear combination of the 87 original covariates, yields a model that is not

easily interpretable. Moreover, these models do not help us to answer the question of which of the 87 original covariates is most predictive of retained earnings. In order to address these issues, we seek to use the information already present in the PCA plus L^1 model to determine which of the original variables is most predictive.

Before proceeding, let us give two additional motivating reasons for performing the study in this way. Suppose we were to use forward stepwise modeling on the original space of 87 covariates to determine the top J most predictive variables. In this approach, we start with no predictors and add one at a time, each time looping over all predictors that have not yet been chosen and selecting the one with the greatest predictive power, as defined either by hypothesis tests or by cross-validation test set error. This requires searching through $87!/(87-J)!$ possible models, a task that becomes computationally difficult as J increases.

Suppose we were to use backward stepwise modeling on the original space of 87 covariates to determine the top five most predictive variables. In this approach, we start with a model fit to all predictors and remove one at a time, each time looping over all predictors that have not yet been eliminated and removing the one with the least predictive power, as defined either by hypothesis tests or by cross-validation test set error. Here our progress would be blocked by multicollinearity, as manifested in the large ($\sim 10^{16}$) condition number of the data matrix with all 87 columns.

The approach we now describe is far more efficient than forward stepwise modeling and also avoids the problem of multicollinearity that prevents us from using backward stepwise modeling. Let us express the PCA plus L^1 model in terms of the original covariate space. We partition the $(p'+1)$-dimensional vector of coefficients β_q as follows:

$$\beta_q = \begin{bmatrix} \beta_q^0 & \beta_q^{1:p'} \end{bmatrix},$$

so that β_q^0 is the scalar intercept and $\beta_q^{1:p'}$ is a p'-dimensional vector of the remaining coefficients. Then (14) can be written

$$\widehat{r_q^1} = \beta_q^0 \mathbf{1} + \widetilde{X_q^0}\gamma_q, \tag{18}$$

where $\gamma_q = (W_q^0)'\beta_q^{1:p'}$ is a p-dimensional vector of regression coefficients. These regression coefficients can be viewed as multiplying the original balance sheet and income statement variables present in the mean-centered data matrix $\widetilde{X_q^0} = X_q^0 - \overline{X_q^0}$.

5.1 PCA plus L^1 Variable Ranking

We do not expect that all $p = 87$ of the coefficients in γ_q are equally important, and we seek a ranking of variables in order of importance to the forecasting model. Let $\gamma_q = \begin{bmatrix} \gamma_q^1 & \gamma_q^2 & \cdots & \gamma_q^p \end{bmatrix}$ denote the p components of γ_q. For $1 \leq j \leq p$, let $\mathbf{x}_q^j$ denote the j-th column of $\widetilde{X_q^0}$. Then (18) is

$$\widehat{r_q^1} = \beta_q^0 \mathbf{1} + \sum_{j=1}^{p} \gamma_q^j \mathbf{x}_q^j. \tag{19}$$

This statistical model is still written in terms of the columns of the mean-centered data matrix. We assume that each row of $\mathbf{x}_q^j$ is in fact a sample of a random variable x_q^j. This implies the probabilistic model

$$\hat{r} = \beta_q^0 + \sum_{j=1}^{p} \gamma_q^j x^j, \tag{20}$$

where $\hat{r}$ is itself a random variable. Suppose that each x^j is bounded, i.e., there exist finite (m_j, M_j) such that $m_j \leq x^j \leq M_j$. Then it is clear that $\hat{r}$ is also bounded. We may estimate the maximum value by

$$\hat{r}_{\max} \geq \beta_q^0 + \sum_{\gamma_q^j \geq 0} \gamma_q^j M_j + \sum_{\gamma_q^j < 0} \gamma_q^j m_j,$$

and the minimum value by

$$\hat{r}_{\text{min}} \leq \beta_q^0 + \sum_{\gamma_q^j \geq 0} \gamma_q^j m_j + \sum_{\gamma_q^j < 0} \gamma_q^j M_j.$$

These inequalities become equalities if, for example, the random variables $\{x^j\}$ are pairwise independent. This implies that the range of $\hat{r}$ can be estimated via

$$\begin{aligned}
\text{range}(\hat{r}) &= \hat{r}_{\text{max}} - \hat{r}_{\text{min}} \\
&\geq \sum_{\gamma_q^j \geq 0} \gamma_q^j (M_j - m_j) + \sum_{\gamma_q^j < 0} \gamma_q^j (m_j - M_j) \\
&\geq \sum_{j=1}^{p} \left|\gamma_q^j\right| (M_j - m_j) = \sum_{j=1}^{p} \left|\gamma_q^j\right| \text{range}\left(x^j\right).
\end{aligned}$$

Our variable ranking procedure relies on the principle that the term on the right-hand side that explains the largest fraction of the range of $\hat{r}$ corresponds to the most important variable in the model. We therefore use the sample range of $\mathbf{x}_q^j$ as an estimate of the range of x^j, and compute

$$\Gamma_j = \left|\gamma_q^j\right| \text{range}\left(\mathbf{x}_q^j\right)$$

for $j = 1, 2, \ldots, p$. We define the k-th most important variable to be the one that has the k-th largest value in the set $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_p\}$.

Suppose we were to take the variance of both sides of (20). Assuming that the variance-covariance matrix of the random variables $\{x^j\}_{j=1}^p$ is purely diagonal, we obtain

$$\text{Var}(\hat{r}) = \sum_{j=1}^{p} (\gamma_q^j)^2 \text{Var}\left(x^j\right). \tag{21}$$

The right-hand side is the sum of the squares of the standardized regression coefficients for the model (20). Using these standardized coefficients to rank the importance of variables in an L^2 multivariate model is a classical technique.

We therefore view each Γ_j as a regression coefficient that has been standardized using the range rather than the standard deviation. Other authors have referred to these coefficients as maximum impact coefficients [Alderson and Nielsen, 2002]: Γ_j measures the maximum possible impact that the j-th input x^j has on the output $\hat{r}$.

Since the SVBA data set consists of audited, cleaned data, we view extreme values in any of the variables as important indicators, rather than as outliers. The tight clustering of values around the mean in each of the covariates in Table 1 suggests that only by looking at extreme values would one be able to extract information that helps predict retained earnings. This motivates the use of the range, which is maximally sensitive to extreme values in each of the covariates.

We apply our ranking procedure to each of the PCA plus L^1 forecasting models developed above, one for each of 11 quarters. The rankings of the top five most important variables can be found in Table 8. Note that if one is interested in forecasting quarter $q + 1$ retained earnings, the most important variable is always the retained earnings from quarter q. Besides retained earnings, net profit is another variable that appears in all 11 top five lists. Both results are highly intuitive from the point of view of standard accounting principles. Note that the variable "other equity" also appears in all 11 top five lists.

5.2 Pruned/Simple Models

Let us define a *pruned model* to be one that predicts quarter $q+1$ retained earnings based on the top five quarter q variables. Starting from the mean-centered data matrix $\widetilde{X_q^0}$, we delete all but five columns corresponding to the top five variables for quarter q. The resulting $N \times 5$ matrix will be denoted P_q; let P_q^j denote the j-th column of this matrix. We assume that the first three columns of P_q correspond to retained earnings, net profit, and other equity, respectively—these are the three variables common to all 11 top five lists in Table 8. When we fit a model to these three variables, we refer to it as a *simple model*.

There are two reasons to build pruned and simple models: (1) to improve the interpretability of the models, and (2) to quantify how well the procedure from Section 5.1 identifies variables with predictive power.

With these definitions and our rankings in mind, we fit five linear models and one nonlinear model to the data. In all cases, residual error is as in (13). We first describe the linear models:

1. PCA plus RLM. The form of the model is the same as in (6):

$$r_q^1 = \begin{bmatrix} \mathbf{1} & Y_q' \end{bmatrix} \beta_q + \varepsilon_q$$

   As before, Y_q' is the PCA-transformed data matrix with $p' = 35$ columns. The only difference between this and the PCA plus L^1 model is that the coefficients β_q are found by minimizing

$$\|\varepsilon_q\|_\rho = \sum_{j=1}^{|C_q|} \rho\big(\varepsilon_{q_j}\big), \tag{22}$$

   where ρ is Tukey's bisquare function [Maronna et al., 2006]. We include this model because our earlier analysis indicated that the regression residuals have tails that are heavier than those of the normal distribution. Such behavior suggests the use of robust regression; L^1 and RLM are competing robust regression techniques.

2. L^1 pruned. This is an L^1 model fitted to each quarter's top 5 variables:

$$r_q^1 = \beta_q^0 + P_q \beta_q^{1:5} + \varepsilon_q. \tag{23}$$

   The regression coefficients β_q are found by minimizing $\|\varepsilon_q\|_1$.

3. RLM pruned. This is the same model as (23),

$$r_q^1 = \beta_q^0 + P_q \beta_q^{1:5} + \varepsilon_q, \tag{24}$$

   except that the regression coefficients β_q are found by minimizing $\|\varepsilon_q\|_\rho$ from (22).

4. L^1 simple. This is an L^1 model fitted to the three variables that we have found are common to all 11 top 5 lists in Table 8:

$$r_q^1 = \beta_q^0 + P_q^1 \beta_q^1 + P_q^2 \beta_q^2 + P_q^3 \beta_q^3 + \varepsilon_q, \tag{25}$$

   where the regression coefficients $\{\beta_q^0, \beta_q^1, \beta_q^2, \beta_q^3\}$ are found by minimizing $\|\varepsilon_q\|_1$.

5. RLM simple. This is the same model as (25),

$$r_q^1 = \beta_q^0 + P_q^1 \beta_q^1 + P_q^2 \beta_q^2 + P_q^3 \beta_q^3 + \varepsilon_q, \tag{26}$$

   except that the regression coefficients β_q are found by minimizing $\|\varepsilon_q\|_\rho$ from (22).

The only new nonlinear model we fit is an additive model of the form

$$r_q^1 = \beta_q^0 + \sum_{j=1}^{5} f_q^j(P_q^j) + \varepsilon_q. \tag{27}$$

The f_q^j functions are nonparametric smoothing functions that are calculated using a backfitting algorithm [Hastie et al., 2009]. We include this model to check whether a nonlinear, partially nonparametric model fitted to a small number of influential variables performs better than a linear model. As it turns out, for all quarters of testing, its error exceeds the linear models' errors by a large margin; we do not include these results in tables given below. We mention the results of the additive model to again confirm the suitability of linear models for this problem.

For the five new linear models described above, out-of-sample median absolute errors are given in Table 9. We also include errors for the PCA plus L^1 model, and we italicize the lowest error value for each quarter.

The PCA plus L^1 model is competitive. At no point does it achieve the lowest out-of-sample error. However, for a purely machine-generated model, into which we have invested no intuition or domain-specific knowledge, its performance is more than adequate. For the PCA plus L^1 model, the standard deviation of the errors σ_ε, a metric that measures the consistency of the model, is the smallest across all six models.

The PCA plus RLM model gives slightly smaller average error than the PCA plus L^1 model, but its σ_ε is the highest across all six models. This large value stems from the relatively large error incurred in the last quarter of testing.

Pruning both the L^1 and RLM models decreases their out-of-sample errors. For the L^1 model, this comes at the expense of a larger σ_ε. However, the pruned RLM model achieves what we believe is an excellent balance between low average error μ_ε and consistency σ_ε—both values are second best across all six models.

Finally, the simple L^1 and simple RLM models have the smallest mean out-of-sample errors, but both σ_ε values have increased slightly relative to the pruned versions of these models.

The differences between the pruned and simple models are slight. The first conclusion we draw is that the variable ranking procedure from Section 5.1, which uses the regression coefficients from the PCA plus L^1 model, succeeds in finding models with superior out-of-sample performance. The results from the most predictive RLM models are consistent with the hypothesis that the variables included in the pruned/simple models are the most predictive variables.

Our second conclusion is that once the top variables have been identified, we can obtain strong out-of-sample performance via either robust regression technique.

5.3 Comparison with Results for Publicly Traded Companies

In order to test whether PCA plus L^1 regression works well on other data sets, and also to understand some of the differences between models of privately held and publicly traded companies, we have applied the techniques described in this paper to financial data for companies in the Standard & Poor's (S&P) 500 index. The source of this data set is S&P Capital IQ, a division of Standard & Poor's (`http://www.capitaliq.com`). This data set consists of 38 financial variables extracted from quarterly financial statements over the same time span as the SVBA data. These 38 variables are standard balance sheet and income statement variables, all measured in units of millions of dollars. One of the variables is retained earnings, which will again serve as the dependent variable in our models.

Our goal here is to highlight similarities and differences between the PCA plus L^1 methodology applied to the SVBA data versus the S&P 500 data. We therefore present our results in summary form, mostly omitting detailed tables and figures. In future work, we plan to carry out a detailed analysis of forecasting models for the S&P 500 data.

Examining each of the 38 S&P 500 covariates in turn, in the same way as in Table 1, we find similar trends. For each covariate, the percent of samples that lie within one standard deviation of the mean is between 92.4% and 99.9%. For each covariate, the sample excess kurtosis is between 33 and 3603. Other metrics such as the sample range and sample standard deviation are not as consistent from one covariate to the next, unlike what we found in Table 1. We attribute this inconsistency to the different constituencies represented in the respective data sets. The companies in the S&P 500 index cover a broader range of industry sectors than the companies in the SVBA database. The companies in the SVBA database have been or currently are lending clients of SVB; in contrast, the companies in the S&P 500 index are not necessarily selected for their credit worthiness.

Importantly, the S&P 500 data displays a similar level of multicollinearity as the SVBA data. Plots of the singular values reveal the same trends as in Fig. 2, with several singular values lying close to zero. Applying the same 10-fold cross-validation study described in Section 3.3, we produce a plot that is similar to Fig. 3—the only difference is that the optimal number of PCA-transformed variables is now $p' = 28$.

Comparing the results of linear and nonlinear models as in Table 6, we find that the PCA plus L^1 model features the best out-of-sample predictive accuracy for the retained earnings of S&P 500 companies.

Next, we apply the variable ranking and model pruning techniques from Section 5.2 to the PCA plus L^1 model for the S&P 500 data. We present the top five variables for each quarter in Table 10, and the out-of-sample results for pruned and simple models in Table 11.

A key difference is that the PCA plus L^1 model's errors are now much less consistent from one quarter to the next. For the SVBA data, if we examine the L^1 errors to the right of the vertical bar in Fig. 6, all the errors lie in the tight interval $[333, 497]$—the coefficient of variation (standard deviation divided by mean) for these errors is 0.11.

However, for the PCA plus L^1 model applied to the S&P 500 data, the errors lie in the interval $[69, 145]$, with coefficient of variation equal to 0.33. Moreover, the error for Q3 2008 is 74.8, while the errors for Q4 2008 and Q1 2009 are 144.4 and 145.3. The increase in error from Q3 2008 to Q4 2008 is 96%, far higher than any increase (or decrease) in consecutive quarter errors observed for the PCA plus L^1 model on the SVBA data. Noting that Q4 2008 coincided with the beginning of a serious economic recession, we hypothesize that the consistency of the forecasting model for privately held companies may be due to their relative insulation from macroeconomic forces, as compared to publicly traded companies.

The simple models in Table 11 are trained on the three variables common to all 11 top five lists in Table 10: retained earnings, total assets, and total liabilities. We can see that model pruning/simplification decreases both μ_ε and σ_ε as compared to the original PCA plus L^1 and PCA plus RLM models. We conclude that while the identities of the most predictive variables turn out to be rather different for privately held versus publicly traded companies, the pruning method works well for both data sets.

Examining the regression coefficients for the pruned and simple models of S&P 500 retained earnings, we find that the coefficients of total assets and total liabilities are nearly equal in magnitude but have opposite sign. This implies the following simplification:

$$\begin{aligned}\widehat{r_q^1} &= \beta_q^0 + (1.017)\text{ret. earnings} + (0.0117)[\text{total assets} - \text{total liabilities}] \\ &= \beta_q^0 + (1.017)\text{ret. earnings} + (0.0117)\text{net worth}\end{aligned}$$

The coefficients we have reported are for the final quarter's pruned PCA plus L^1 model. The true coefficients for total assets/liabilities are very slightly larger/smaller than ± 0.0117—we have averaged the absolute values of the true coefficients to produce this number. The comparable model for the SVBA data reads as follows:

$$\widehat{r_q^1} = \beta_q^0 + (1.003)\text{ret. earnings} + (0.768)\text{net profit} + (0.00282)\text{other equity}.$$

For a publicly traded company, the optimal model appears to be one where quarterly retained earnings increase by a small percentage (in the above case, 1.17%) of the net worth (or book value) of the company.

Net income/profit does not enter directly into this model at all. For a privately held company, the optimal model appears to be one where retained earnings increase through a combination of the past quarter's net profit plus a small percentage of the company's other equity. We see that the identities of the top variables has changed, and that this change creates different types of models.

Because the data sets differ in units (thousands versus millions of dollars), we present one final set of results showing relative errors in the L^1 norm. Using the same notation as in Section 4.1, we define for each quarter θ the quantity

$$E_\theta^{\text{REL}} = \frac{\left\| r_\theta^1 - \widehat{r_\theta^1} \right\|_1}{\left\| r_\theta^1 \right\|_1}. \tag{28}$$

Here r_θ^1 (respectively, $\widehat{r_\theta^1}$) is the vector of true (respectively, predicted) retained earnings. Computing the component-wise relative errors between these vectors is not possible due to occasional zero entries in r_θ^1. Note that the E_θ^{REL} metric does not suffer from this problem.

In Table 12, we compare the relative out-of-sample errors for retained earnings forecasts using both SVBA and S&P 500 data sets. The overall performance of all models is roughly one percentage point better for SVBA data. While pruning/simplification of the models does improve the SVBA relative errors, the improvement is more pronounced for the S&P 500 errors.

6 Conclusion

The combined PCA plus L^1 model forecasts retained earnings with greater out-of-sample accuracy than a variety of other regression techniques: OLS, trees, forests, and boosting. Using the PCA plus L^1 model to select variables, we are able to develop reduced-order models where the out-of-sample accuracy has been improved still further. As we have explained, a key driver for the success of L^1 and other robust regression models is our finding that the retained earnings residuals are distributed with heavier-than-normal tails.

Based on the success of this method, we see three areas for future work. First, we seek to further explore the differences between privately held and publicly traded companies. We seek to investigate more deeply the financial statement data for companies in the S&P 500 index; the results of Section 5.3 give us a number of hypotheses to test. Second, the current study has focused on using a novel data set to develop a sound framework for predictive modeling. Incorporating domain-specific knowledge into this framework, e.g., developing separate models for different industry sectors, or using variable amounts of past data based on historical conditions, may lead to improvements. Finally, we seek to generalize the codes and algorithms developed here to forecast financial variables besides retained earnings.

References

A. S. Alderson and F. Nielsen. Globalization and the great U-turn: Income inequality trends in 16 OECD countries. *American Journal of Sociology*, 107(5):1244–1299, 2002. URL `http://www.jstor.org/stable/10.1086/341329`.

I. Barrodale and F. D. K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848, 1973. URL `http://dx.doi.org/10.2307/2156318`.

D. Berg. Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, 23:129–143, 2007. URL `http://dx.doi.org/10.1002/asmb.658`.

H. S. Bhat and D. Zaelit. Predicting private company exits using qualitative data. In J. Huang, L. Cao, and J. Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6634 of *Lecture Notes in Computer Science*, pages 399–410. Springer, Berlin, 2011. URL 10.1007/978-3-642-20841-6_33.

P. Bloomfield and W. Steiger. Least absolute deviations curve-fitting. *SIAM Journal on Scientific and Statistical Computing*, 1(2):290–301, 1980. URL http://dx.doi.org/10.1137/0901019.

R. Bowman and S. Bush. Using comparable companies to estimate the betas of private companies. *Journal of Applied Finance*, pages 71–81, Fall/Winter 2006. URL http://ssrn.com/abstract=956443.

L. Breiman. *Classification and Regression Trees*. Chapman & Hall, 1984.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. URL http://dx.doi.org/10.1023/A:1010933404324.

P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007. URL http://dx.doi.org/10.1214/07-STS242.

P. Bunn and V. Redwood. Company accounts-based modelling of business failures and the implications for financial stability. Technical Report 210, Bank of England, 2003. URL http://dx.doi.org/10.2139/ssrn.598276.

D. De Leonardis and R. Rocci. Assessing the default risk by means of a discrete-time survival analysis approach. *Applied Stochastic Models in Business and Industry*, 24:291–306, 2008. URL http://dx.doi.org/10.1002/asmb.705.

J. Fernandes. Corporate credit risk modeling: Quantitative rating system and probability of default estimation. *SSRN eLibrary*, 2005. URL http://dx.doi.org/10.2139/ssrn.722941.

J. Fox. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Thousand Oaks, CA, second edition, 2008.

J. R. M. Hand. The value relevance of financial statements in the venture capital market. *The Accounting Review*, 80(2):613–648, 2005. URL http://dx.doi.org/10.2307/4093071.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, second edition, 2009.

D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998. URL http://dx.doi.org/10.1111/1467-9884.00122.

I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, second edition, 2002.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. URL http://dx.doi.org/10.2307/2328720.

R. Koenker and K. F. Hallock. Quantile regression. *The Journal of Economic Perspectives*, 15(4):143–156, 2001. URL http://dx.doi.org/10.2307/2696522.

D. Kundu. Discriminating between normal and Laplace distributions. In N. Balakrishnan, N. Balakrishnan, H. N. Nagaraja, and N. Kannan, editors, *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, Statistics for Industry and Technology, pages 65–79. Birkhäuser Boston, 2005. URL http://dx.doi.org/10.1007/0-8176-4422-9_4.

B. Lauterbach and P. Schultz. Pricing warrants: An empirical study of the Black-Scholes model and its alternatives. *Journal of Finance*, 45(4):1181–1209, 1990. URL http://dx.doi.org/10.2307/2328720.

Y. Li and G. R. Arce. A maximum likelihood approach to least absolute deviation regression. *EURASIP Journal on Applied Signal Processing*, 2004(12):1762–1769, 2004. URL http://dx.doi.org/10.1155/S1110865704401139.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Chichester, UK, 2006.

M. Minnis. The value of financial statement verification in debt financing: evidence from private U.S. firms. *Journal of Accounting Research*, 49(2):457–506, 2011. URL http://dx.doi.org/10.1111/j.1475-679X.2011.00411.x.

D. Mramor and A. Valentincic. Forecasting the liquidity of very small private companies. *Journal of Business Venturing*, 18:745–771, 2003. URL http://dx.doi.org/10.1016/S0883-9026(03)00002-8.

C. T. Tolkamp. Predicting private equity performance. Master's thesis, University of Twente, 2007. URL http://essay.utwente.nl/771/. Department of Industrial Engineering & Management.

	μ	σ	% Conc	γ	Rng		μ	σ	% Conc	γ	Rng
X_1	6.5e+3	1.7e+4	95.2	3.8e+2	6.5e+5	X_{45}	1.9e+2	1.6e+3	97.9	5.3e+2	6.2e+4
X_2	2.5e+2	5.8e+3	99.7	1.4e+3	2.8e+5	X_{46}	4.4e+2	4.4e+3	98.2	5.6e+2	1.8e+5
X_3	4.3e+2	4.9e+3	98.6	3.4e+2	1.4e+5	X_{47}	3.9e+1	1.4e+3	99.9	1.8e+3	6.6e+4
X_4	3.1e+3	5.7e+3	91.2	2.9e+1	8.7e+4	X_{48}	2.8e+3	2.6e+4	98.0	7.1e+2	8.9e+5
X_5	5.1e+1	1.3e+3	98.9	9.5e+2	1.0e+5	X_{49}	1.7e+4	3.0e+4	89.2	2.5e+1	3.7e+5
X_6	2.3e+2	2.3e+3	98.5	4.4e+2	7.7e+4	X_{50}	-3.5e+4	6.3e+4	92.1	8.2e+1	1.4e+6
X_7	8.5e+1	8.7e+2	98.4	3.0e+2	2.6e+4	X_{51}	1.3e+1	4.1e+2	99.7	6.2e+3	4.1e+4
X_8	3.4e+1	3.7e+2	98.6	3.0e+2	1.2e+4	X_{52}	4.7e+2	5.0e+3	98.1	4.8e+2	1.6e+5
X_9	9.0e+2	4.8e+3	96.1	2.8e+2	1.4e+5	X_{53}	-7.6e+0	9.7e+2	98.4	5.5e+2	7.1e+4
X_{10}	2.8e+2	1.9e+3	96.7	1.9e+2	7.4e+4	X_{54}	7.2e+0	6.7e+2	98.6	6.2e+2	4.8e+4
X_{11}	2.8e+1	3.1e+2	98.3	1.3e+3	1.7e+4	X_{55}	2.2e+2	2.2e+3	98.0	5.3e+2	6.9e+4
X_{12}	6.5e+2	1.6e+3	94.2	1.2e+2	5.6e+4	X_{56}	2.3e+4	6.8e+4	93.5	7.3e+1	1.5e+6
X_{13}	4.3e+1	7.0e+2	99.2	7.0e+2	2.7e+4	X_{57}	5.0e+3	9.5e+3	92.4	4.5e+1	1.8e+5
X_{14}	7.8e+1	6.3e+2	97.5	2.8e+2	2.2e+4	X_{58}	2.2e+1	7.5e+2	99.7	5.9e+3	7.1e+4
X_{15}	6.7e+1	5.8e+2	97.8	5.9e+2	2.5e+4	X_{59}	2.5e+0	1.1e+2	99.7	5.7e+3	1.1e+4
X_{16}	1.8e+2	3.3e+3	99.2	1.2e+3	1.6e+5	X_{60}	1.7e+1	5.2e+2	99.6	5.5e+3	5.1e+4
X_{17}	1.8e+1	3.8e+2	99.6	8.1e+2	1.4e+4	X_{61}	2.3e+3	5.7e+3	93.4	7.5e+1	1.0e+5
X_{18}	2.9e+3	6.7e+3	93.7	6.2e+1	1.0e+5	X_{62}	8.9e+2	2.2e+3	92.5	1.8e+2	8.3e+4
X_{19}	1.1e+3	3.3e+3	94.2	1.3e+2	7.5e+4	X_{63}	8.4e+2	2.2e+3	94.3	3.5e+2	7.3e+4
X_{20}	5.7e+1	9.0e+2	99.4	5.2e+2	3.0e+4	X_{64}	1.2e+3	2.0e+3	91.0	9.3e+1	6.4e+4
X_{21}	2.2e+2	2.6e+3	98.7	1.0e+3	1.6e+5	X_{65}	9.0e+0	1.4e+2	98.7	2.4e+3	8.7e+3
X_{22}	5.0e+1	4.2e+2	98.1	2.3e+2	1.2e+4	X_{66}	1.8e+1	3.3e+2	98.5	5.2e+2	2.3e+4
X_{23}	2.1e+2	2.1e+3	98.2	3.0e+2	5.7e+4	X_{67}	5.6e+2	2.0e+3	94.2	4.4e+2	7.6e+4
X_{24}	1.3e+2	8.9e+2	97.6	5.0e+2	6.0e+4	X_{68}	1.6e+2	4.3e+2	93.6	9.1e+1	1.5e+4
X_{25}	2.0e+2	9.1e+2	96.2	8.7e+1	2.1e+4	X_{69}	4.2e+1	3.6e+2	97.7	8.9e+2	2.1e+4
X_{26}	1.0e+2	9.7e+2	98.1	4.2e+2	3.2e+4	X_{70}	1.8e+1	2.1e+2	98.7	7.8e+3	2.3e+4
X_{27}	3.6e+3	1.7e+4	95.8	1.6e+2	3.6e+5	X_{71}	8.8e+1	3.3e+2	95.8	1.2e+3	2.0e+4
X_{28}	1.6e+3	3.6e+3	93.4	1.1e+2	1.0e+5	X_{72}	4.2e-1	2.2e+1	99.9	3.7e+3	1.5e+3
X_{29}	3.5e+2	6.0e+3	99.3	1.1e+3	2.9e+5	X_{73}	1.5e+0	6.2e+1	99.8	2.3e+3	4.6e+3
X_{30}	7.2e+2	2.3e+3	93.6	8.6e+1	4.3e+4	X_{74}	2.7e+0	1.6e+2	99.6	7.4e+3	2.2e+4
X_{31}	1.7e+1	6.1e+2	99.5	1.2e+4	7.2e+4	X_{75}	2.3e+1	3.2e+2	98.3	2.4e+3	3.2e+4
X_{32}	5.5e+1	6.5e+2	98.4	7.5e+2	3.3e+4	X_{76}	4.8e+1	8.4e+2	99.1	2.9e+3	6.5e+4
X_{33}	2.7e+1	4.0e+2	99.2	4.8e+2	1.2e+4	X_{77}	7.0e+1	1.5e+3	99.4	3.7e+3	2.1e+5
X_{34}	7.4e+2	2.1e+3	94.0	2.7e+2	8.4e+4	X_{78}	1.8e+1	1.4e+3	99.8	1.4e+4	1.8e+5
X_{35}	2.1e+3	6.1e+3	93.4	8.1e+1	1.2e+5	X_{79}	1.8e+0	5.5e+1	99.6	1.8e+3	4.1e+3
X_{36}	1.7e+3	3.8e+3	93.6	8.7e+1	7.5e+4	X_{80}	4.9e+1	5.8e+2	98.4	2.3e+3	6.0e+4
X_{37}	1.4e+2	8.6e+2	96.6	2.3e+2	2.8e+4	X_{81}	1.5e+1	3.7e+2	99.3	5.0e+3	3.8e+4
X_{38}	2.0e+0	3.4e+1	99.3	9.5e+2	1.7e+3	X_{82}	1.8e+0	3.8e+2	99.3	2.0e+3	3.6e+4
X_{39}	9.2e+1	1.3e+3	98.6	6.0e+2	5.8e+4	X_{83}	7.7e-1	3.0e+1	99.9	3.0e+3	2.4e+3
X_{40}	1.9e+2	3.0e+3	98.9	2.3e+3	1.8e+5	X_{84}	-2.4e+0	7.3e+2	99.9	6.7e+3	1.1e+5
X_{41}	1.8e+3	5.6e+3	94.5	9.5e+1	1.0e+5	X_{85}	-1.2e+3	3.7e+3	92.7	4.8e+2	2.7e+5
X_{42}	5.2e+1	5.3e+2	98.1	5.2e+2	2.1e+4	X_{86}	-1.8e+2	8.1e+3	100.0	6.9e+3	7.2e+5
X_{43}	9.5e+1	1.2e+3	98.7	4.9e+2	3.3e+4	X_{87}	1.0e+0	1.8e+0	94.8	5.6e+2	8.7e+1
X_{44}	9.0e+2	6.8e+3	97.7	6.5e+2	2.8e+5						

Table 1: Mean (μ), standard deviation (σ), percent of samples that lie within one standard deviation of the mean (% Conc), excess kurtosis (γ), and range (Rng) for all 87 covariates, aggregated across all quarters from Q1 2008 to Q1 2011. All X_j's have units of thousands of dollars, except for the unitless ratios X_{86} and X_{87}. For a normal random variable, $\gamma = 0$ and % Conc $= 68.3$. Large values of γ and % Conc indicate significant departure from normality for the marginal distributions of each X_j. Also note the large values of Rng, implying that for each j, we can find ≥ 1 companies displaying extreme behavior in X_j.

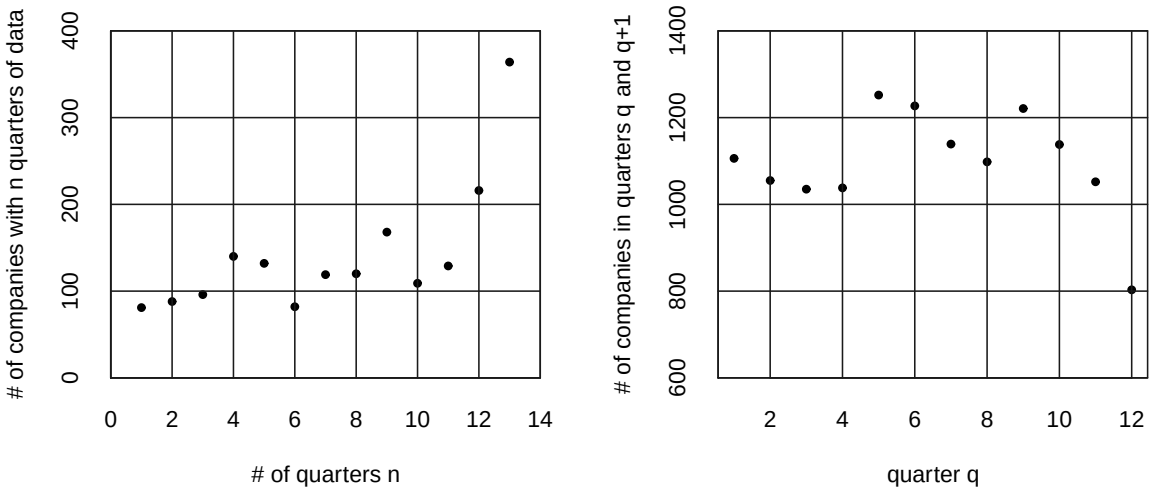

Figure 1: The total number of companies in the study is 1844. The left panel shows that if we try to study only those companies for which we have many quarters worth of data, we will leave out most companies. For example, less than 20% of the companies are represented for all 13 quarters. Let C_q denote those companies represented in consecutive quarters q and $q+1$. In the right panel, we see that $|C_q|$ is at least 800 for all q and exceeds 1000 except at $q = 12$.

	min	median	max	MAD
Q2 2008	-8.78e+05	-1.58e+04	6.56e+04	1.38e+04
Q3 2008	-9.17e+05	-1.76e+04	7.50e+04	1.47e+04
Q4 2008	-9.55e+05	-1.72e+04	8.33e+04	1.42e+04
Q1 2009	-9.91e+05	-1.80e+04	3.86e+04	1.46e+04
Q2 2009	-1.04e+06	-1.87e+04	6.45e+04	1.56e+04
Q3 2009	-1.06e+06	-1.99e+04	1.04e+05	1.61e+04
Q4 2009	-7.71e+05	-2.03e+04	1.08e+05	1.64e+04
Q1 2010	-7.74e+05	-2.14e+04	1.18e+05	1.71e+04
Q2 2010	-1.16e+06	-2.06e+04	1.99e+05	1.72e+04
Q3 2010	-1.16e+06	-2.10e+04	1.34e+05	1.75e+04
Q4 2010	-1.04e+06	-2.18e+04	1.43e+05	1.79e+04
Q1 2011	-1.04e+06	-2.27e+04	7.74e+04	1.83e+04

Table 2: For each of 12 quarters q, we give the minimum, median, maximum, and median absolute deviation (MAD) of the true retained earnings vector r_q^1.

	$p' = 87$	$p' = 35$	$p' = 20$	Max	$p' = 87$	$p' = 35$	$p' = 20$
Q1 2008	1.164e16	7.899e1	4.045e1	0.75	9	0	0
Q2 2008	1.161e16	9.032e1	4.370e1	0.35	6	0	0
Q3 2008	1.122e16	7.239e1	3.603e1	0.56	12	0	0
Q4 2008	9.710e4	7.524e1	3.651e1	0.00	0	0	0
Q1 2009	1.179e16	7.786e1	3.860e1	0.46	8	0	0
Q2 2009	2.925e6	7.360e1	3.877e1	0.00	0	0	0
Q3 2009	1.096e16	6.826e1	3.069e1	0.68	21	0	0
Q4 2009	1.090e16	6.544e1	3.023e1	0.60	9	0	0
Q1 2010	1.107e16	7.179e1	3.829e1	0.73	5	0	0
Q2 2010	1.103e16	8.035e1	3.918e1	0.24	20	0	0
Q3 2010	1.067e16	7.062e1	3.491e1	0.47	11	0	0
Q4 2010	1.060e16	7.355e1	3.253e1	0.56	11	0	0

Table 3: Condition numbers (left of double vertical bar) and correlation counts (right of double vertical bar) for full PCA-transformed data matrix Y_q^0 (with $p' = 87$) and reduced-dimension PCA-transformed data matrices $(Y_q^0)'$ with $p' = 35$ and $p' = 20$ columns. The condition numbers are calculated by taking the ratio of the largest to the smallest singular value. The correlation counts are the number of above-diagonal elements of the correlation matrix with absolute value greater than or equal to 0.1. The original data matrix possesses such entries, but the reduced-dimension data matrices do not. For reference, in the "Max" column, we also report the maximum absolute correlation in the original $p' = 87$ data matrices. The results show that the raw data sets possess significant multicollinearity, which can be remedied by using PCA with $p' = 35$.

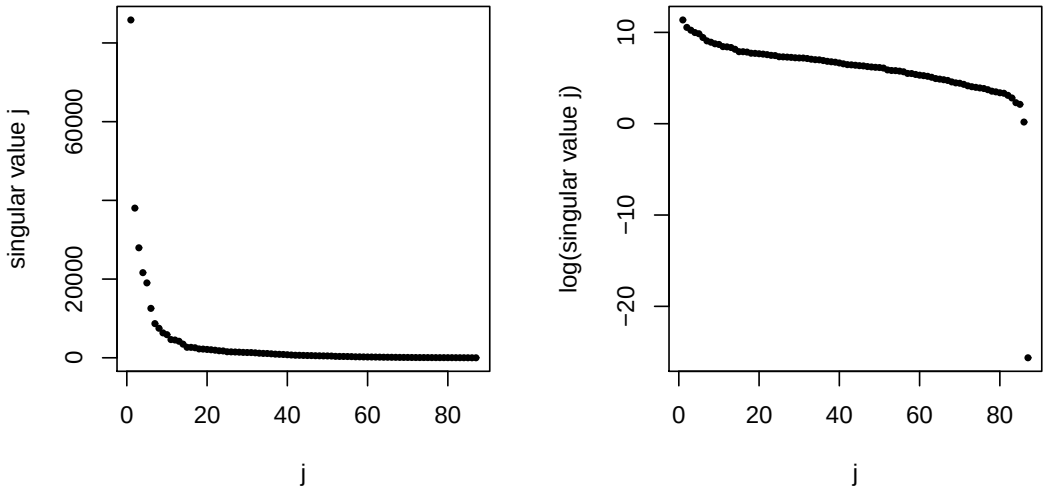

Figure 2: Let Σ_{jj} denote the j-th singular value of the centered Q1 2008 data matrix $\tilde{X}_1^0$. Then the left and right panels show, respectively, Σ_{jj} and $\log(\Sigma_{jj})$ versus j.

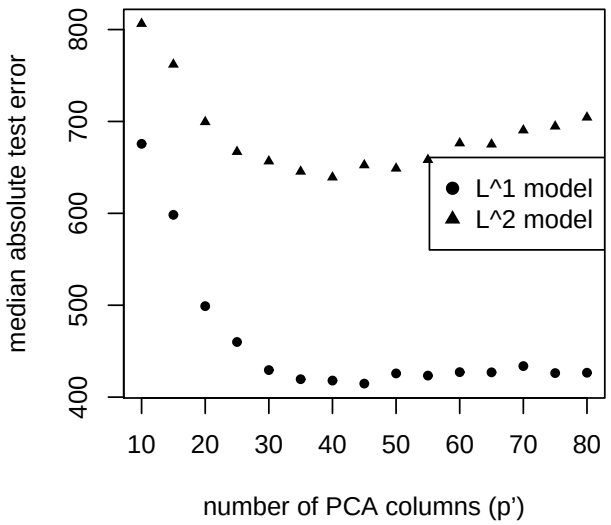

Figure 3: Results of 10-fold cross-validation to determine the optimal number of PCA columns. Note that both L^1 and L^2 test error curves are monotonically decreasing until $p' = 35$. For $p' > 35$, the test set errors are either greater or only marginally less than the error at $p' = 35$.

	L^1	L^2	tree	forest	boost	L^1	L^2	tree	forest	boost
Q1 2008	1.74e10	1.64e10	1.38e9	5.31e11	2.68e11	1.11e6	1.29e6	8.45e5	5.14e6	2.54e6
Q2 2008	1.26e10	1.18e10	1.42e9	5.99e11	4.37e11	1.11e6	1.26e6	7.97e5	5.65e6	3.05e6
Q3 2008	7.99e9	6.80e9	1.16e9	5.86e11	2.90e11	1.09e6	1.24e6	7.32e5	4.92e6	2.38e6
Q4 2008	2.59e10	2.35e10	1.32e9	5.98e11	2.07e11	1.41e6	1.70e6	7.61e5	5.06e6	2.75e6
Q1 2009	4.02e10	3.12e10	1.74e9	7.72e11	3.06e11	1.36e6	2.14e6	9.56e5	5.74e6	2.88e6
Q2 2009	2.13e10	1.88e10	1.69e9	7.39e11	3.16e11	1.29e6	1.61e6	9.77e5	5.89e6	2.77e6
Q3 2009	1.41e10	9.74e9	1.23e9	4.32e11	1.43e11	1.22e6	1.44e6	7.79e5	5.10e6	1.95e6
Q4 2009	2.31e10	1.91e10	1.25e9	4.41e11	1.44e11	1.48e6	1.75e6	7.97e5	5.51e6	2.37e6
Q1 2010	1.17e10	9.63e9	2.17e9	1.09e12	5.64e11	1.29e6	1.40e6	1.08e6	7.30e6	3.85e6
Q2 2010	2.81e10	2.52e10	2.12e9	1.03e12	6.44e11	1.28e6	1.59e6	1.05e6	6.96e6	3.61e6
Q3 2010	1.01e11	4.59e10	1.45e9	8.60e11	3.01e11	1.62e6	2.82e6	8.20e5	6.28e6	2.55e6
mean	2.75e10	1.98e10	1.54e9	6.99e11	3.29e11	1.30e6	1.66e6	8.72e5	5.78e6	2.79e6

Table 4: In-sample errors for models fitted using $p' = 35$ PCA-transformed covariates. In the left half, we report sum of squared errors ($\|\varepsilon_q\|_2^2$); in the right half, we report sum of absolute errors ($\|\varepsilon_q\|_1$). The results for the linear L^1 and L^2 models are expected—over all linear models, the L^2 model minimizes $\|\varepsilon_q\|_2^2$ and the L^1 model minimizes $\|\varepsilon_q\|_1$. Note that the linear models fare better than two of the nonlinear models: random forests (forest) and boosted trees (boost). However, the model with the best in-sample fit is the single regression tree model (tree), a result that does not carry over to the out-of-sample tests in Table 6.

	$\hat{\mu}$	$\hat{\sigma}_p$	$\hat{p}$	$\hat{\mu}$	$\hat{\sigma}_p$	$\hat{p}$
Q1 2008	5.46e-11	1.01e+03	1.00e+00	1.24e+00	1.17e+03	1.00e+00
Q2 2008	1.46e-11	1.05e+03	1.00e+00	1.59e+02	1.19e+03	1.00e+00
Q3 2008	5.82e-11	1.05e+03	1.00e+00	2.82e+01	1.20e+03	1.00e+00
Q4 2008	-3.64e-11	1.36e+03	1.00e+00	3.22e+02	1.59e+03	1.00e+00
Q1 2009	1.46e-11	1.09e+03	1.00e+00	2.12e+02	1.69e+03	1.00e+00
Q2 2009	2.91e-11	1.05e+03	1.00e+00	1.79e+02	1.30e+03	1.00e+00
Q3 2009	-4.09e-11	1.08e+03	1.00e+00	1.29e+02	1.26e+03	1.00e+00
Q4 2009	-4.37e-11	1.35e+03	1.00e+00	2.48e+02	1.56e+03	1.00e+00
Q1 2010	7.28e-12	1.06e+03	1.00e+00	7.93e+01	1.15e+03	1.00e+00
Q2 2010	-1.48e-11	1.13e+03	1.00e+00	1.06e+02	1.39e+03	1.00e+00
Q3 2010	8.75e-11	1.54e+03	1.00e+00	2.82e+02	2.66e+03	1.00e+00

Table 5: Maximum likelihood estimates of exponential power distribution (EPD) parameters, fitted to regression residuals from L^1 (left) and L^2 (right) models. Fitting was carried out using numerical maximization of the EPD likelihood. Note that in all cases, the estimated shape parameter $\hat{p}$ equals 1 up to machine precision. Since the EPD reduces to the Laplace and normal distributions when, respectively, $p = 1$ and $p = 2$, this is further indication that the Laplace distribution fits the regression residuals better than the normal distribution.

	L^1	L^2	tree	forest	boost	L^1	L^2	tree	forest	boost
Q3 2008	3558	3635	9434	11730	18257	380	495	1466	816	939
Q4 2008	5633	5727	8127	12153	21565	405	532	1400	793	939
Q1 2009	5203	5380	7582	13478	18462	408	559	1333	764	984
Q2 2009	6376	6110	10798	14122	18688	386	681	1368	901	1107
Q3 2009	4393	5046	8970	13363	19603	333	775	1296	721	917
Q4 2009	3736	3639	20475	13947	17952	378	578	1252	802	897
Q1 2010	5650	5679	7555	10239	13895	427	545	1308	902	960
Q2 2010	4666	4724	29758	29574	32214	430	633	1452	930	1115
Q3 2010	5541	5730	10129	16145	27374	393	509	1438	862	959
Q4 2010	12188	11967	14803	14690	21956	478	688	1505	1014	1108
Q1 2011	7192	10355	9954	15471	20834	497	1235	1519	1004	1126
mean	5831	6181	12508	14992	20982	410	657	1394	865	1005

Table 6: Out-of-sample errors for models fitted using $p' = 35$ PCA-transformed covariates. Two different metrics are separated by the double vertical bar: to the left, we report root mean squared errors (RMSE); to the right, we report median absolute errors (MAE). The L^1 model has the smallest MAE for all 11 quarters, and the smallest RMSE for 8 out of 11 quarters. Averaged over all quarters, the RMSE and MAE of the L^1 model are smaller than all other models.

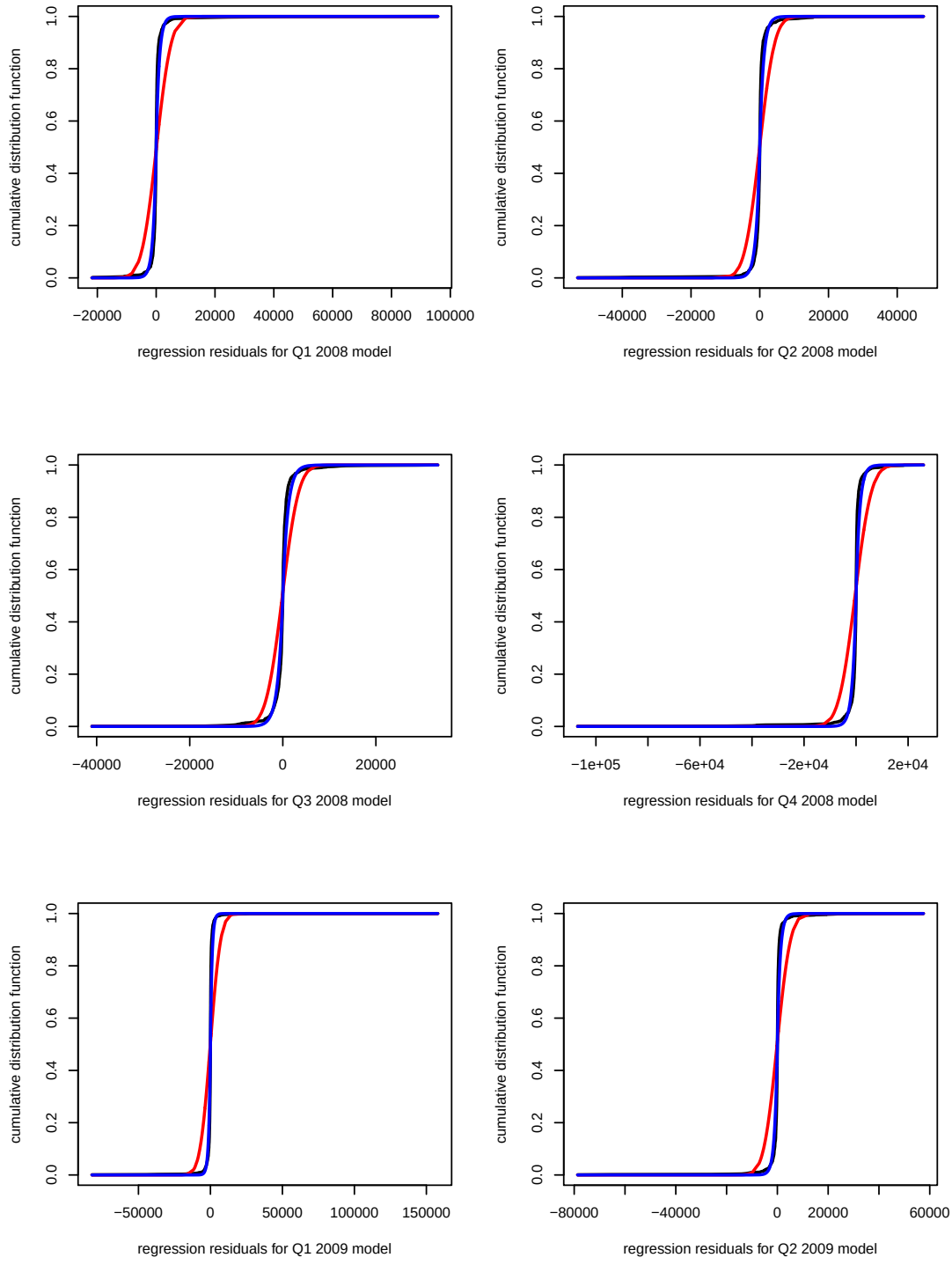

Figure 4: For each quarter, after fitting the PCA plus L^1 regression model, we plot three cumulative distribution functions (CDFs) for the residuals: empirical (black), fitted normal distribution (red), fitted Laplace distribution (blue). The results clearly show that the Laplace is a better fit to the residuals than the normal.

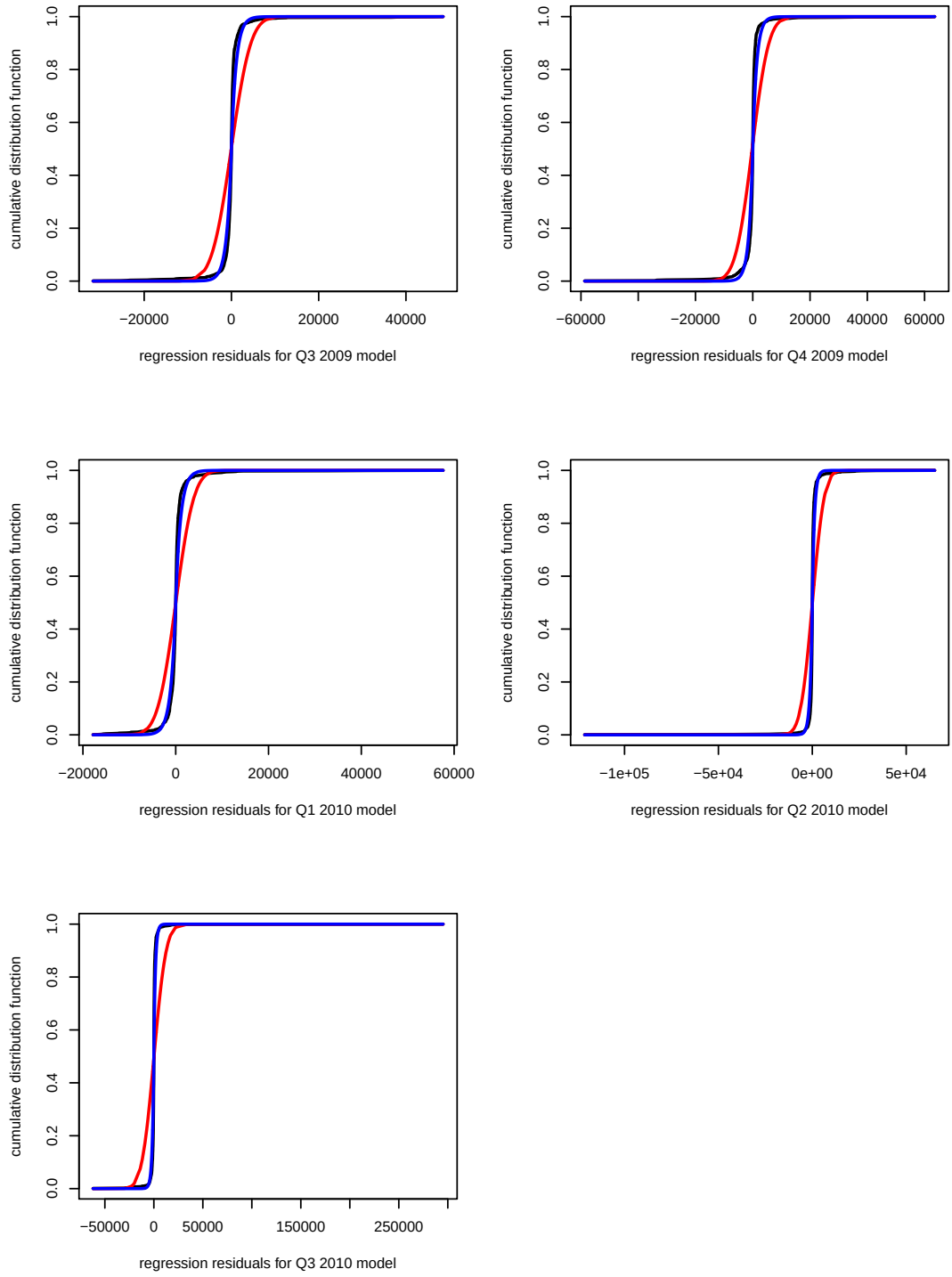

Figure 5: For each quarter, after fitting the PCA plus L^1 regression model, we plot three cumulative distribution functions (CDFs) for the residuals: empirical (black), fitted normal distribution (red), fitted Laplace distribution (blue). The results clearly show that the Laplace is a better fit to the residuals than the normal.

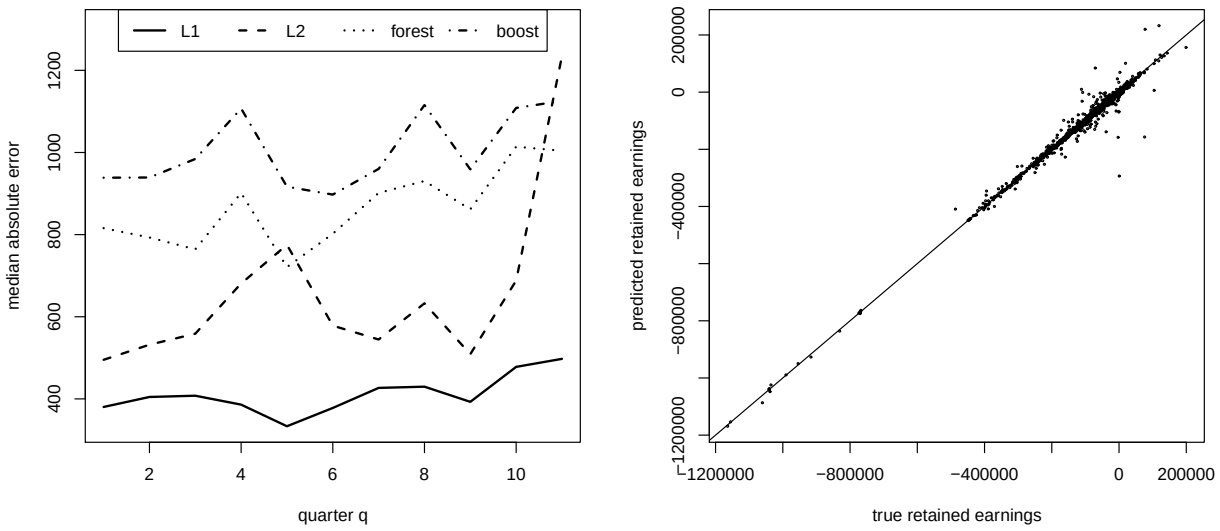

Figure 6: On the left, we plot the median absolute error as a function of quarter q for four different models—all errors plotted here are out-of-sample errors corresponding to the right block of numbers in Table 6. The PCA plus L^1 model errors are consistently smaller than the errors made by the other models. The errors are in units of thousands of dollars. On the right, using the PCA plus L^1 model, we provide a scatterplot of the out-of-sample predicted retained earnings vs. true retained earnings, aggregated across 11 quarters of testing. Note that the correlation coefficient is 0.9952. The black line is a line of slope 1 through the origin—if the model were perfect, all points would lie on this line.

	$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.3$		$\delta = 0.4$		$\delta = 0.5$	
	E_θ^δ	width	E_θ^δ	width	E_θ^δ	width	E_θ^δ	width	E_θ^δ	width
Q3 2008	0.106	198	0.188	424	0.283	596	0.365	830	0.481	1179
Q4 2008	0.073	151	0.197	347	0.290	528	0.358	756	0.428	1004
Q1 2009	0.118	218	0.214	432	0.295	681	0.385	974	0.480	1343
Q2 2009	0.098	213	0.178	495	0.284	742	0.399	1100	0.491	1493
Q3 2009	0.091	179	0.205	411	0.294	620	0.391	858	0.497	1193
Q4 2009	0.092	171	0.189	346	0.272	547	0.353	787	0.440	1062
Q1 2010	0.077	183	0.163	365	0.261	607	0.371	875	0.472	1281
Q2 2010	0.108	241	0.206	513	0.305	752	0.396	992	0.487	1324
Q3 2010	0.100	276	0.214	515	0.322	762	0.419	977	0.525	1288
Q4 2010	0.062	202	0.145	541	0.230	793	0.313	1039	0.393	1300
Q1 2011	0.093	284	0.214	622	0.325	995	0.435	1396	0.511	2105
mean	0.093	210	0.192	456	0.287	693	0.380	962	0.473	1325

Table 7: Out-of-sample interval forecasting results. For five different values of δ, and for 11 quarters, we provide interval forecasts using quantile regression with $\tau^\pm = 0.5 \pm \delta/2$. That is, we forecast the τ^+-th and τ^--th quantiles of the retained earnings for each private company for each quarter, thus obtaining interval forecasts $[\widehat{r_\theta^-}, \widehat{r_\theta^+}]$. The metric E_θ^δ is the fraction (16) of true retained earnings that lie within the respective forecast interval. Note that these numbers are close to δ, indicating accuracy of the interval forecast. We also show the widths of the forecast intervals (17), averaged across all companies, for all values of δ and θ.

	1	2	3	4	5
Q1 2008	ret. earnings	other equity	net profit	common stock	return on assets
Q2 2008	ret. earnings	other equity	net profit	common stock	preferred stock
Q3 2008	ret. earnings	restricted cash	other equity	common stock	net profit
Q4 2008	ret. earnings	net profit	nocl	other equity	other expenses
Q1 2009	ret. earnings	other equity	net profit	common stock	restricted cash
Q2 2009	ret. earnings	other equity	net profit	common stock	nocl
Q3 2009	ret. earnings	net profit	other equity	common stock	restricted cash
Q4 2009	ret. earnings	restricted cash	other equity	net profit	common stock
Q1 2010	ret. earnings	net profit	other equity	apo	return on assets
Q2 2010	ret. earnings	apo	other equity	net profit	restricted cash
Q3 2010	ret. earnings	net profit	restricted cash	apo	other equity

Table 8: We present the ranking of variables obtained using the PCA plus L^1 ranking described in Section 5.1. Here *apo* stands for accounts payable (other), and *nocl* stands for non-operating current liabilities. For each quarter q, we list the top five variables in order of how important they are to the PCA plus L^1 forecasting model of quarter $q+1$ retained earnings. Note that quarter q retained earnings, net profit, and other equity are present in all 11 top five lists.

	PCA+L^1	PCA+RLM	L^1 pruned	RLM pruned	L^1 simple	RLM simple
Q3 2008	380.3	390.8	370.8	373.1	*363.7*	367.5
Q4 2008	404.7	*403.4*	410.7	415.8	414.9	415.7
Q1 2009	407.7	417.2	*383.8*	392.9	396.2	394.3
Q2 2009	386.0	392.7	405.8	*353.8*	414.8	356.3
Q3 2009	333.4	323.2	296.9	293.3	288.5	*285.6*
Q4 2009	377.9	351.6	360.1	341.0	352.8	*340.3*
Q1 2010	426.8	411.8	410.3	397.7	410.0	*373.4*
Q2 2010	429.7	407.2	426.5	389.2	415.0	*385.9*
Q3 2010	392.8	352.6	360.1	348.2	351.3	*336.2*
Q4 2010	478.0	476.5	487.4	*466.3*	472.3	469.7
Q1 2011	497.5	530.9	477.3	471.9	487.1	*470.0*
mean (μ_ε)	410.4	405.3	399.1	385.7	397.0	381.4
sd (σ_ε)	46.5	58.2	54.3	53.1	56.4	55.5

Table 9: We present out-of-sample results for the PCA plus L^1 model trained on $p' = 35$ variables together with five models described in Section 5.2. We show the median absolute error (in units of thousands of dollars) made by forecasting quarter $\theta+1$ retained earnings using quarter θ covariates; the models themselves were fitted by regressing quarter θ retained earnings onto quarter $\theta - 1$ covariates. The smallest number in each row has been italicized. The pruned models use only the top five variables for quarter $\theta - 1$ indicated in Table 8; the simple models use only the three variables common to all 11 top five lists.

	1	2	3	4	5
Q1 2008	ret. earnings	total liabilities	total assets	ap	other liabilities
Q2 2008	ret. earnings	total assets	total liabilities	common stock	treasury stock
Q3 2008	ret. earnings	total liabilities	total assets	short-term debt	common stock
Q4 2008	ret. earnings	total assets	total liabilities	treasury stock	common stock
Q1 2009	total assets	ret. earnings	total liabilities	treasury stock	common stock
Q2 2009	ret. earnings	total assets	total liabilities	treasury stock	common stock
Q3 2009	ret. earnings	total liabilities	total assets	ap	other assets
Q4 2009	total assets	ret. earnings	total liabilities	ap	treasury stock
Q1 2010	ret. earnings	total assets	total liabilities	treasury stock	ap
Q2 2010	ret. earnings	total assets	total liabilities	short-term debt	ap
Q3 2010	total assets	ret. earnings	total liabilities	treasury stock	common stock

Table 10: We present the ranking of variables obtained using the PCA plus L^1 ranking described in Section 5.1, applied to financial statement data for companies in the S&P 500 index. Here *ap* stands for accounts payable, and *common stock* includes additional paid in capital (APIC). Note that quarter q retained earnings, total assets, and total liabilities are present in all 11 top five lists.

	PCA+L^1	PCA+RLM	L^1 pruned	RLM pruned	L^1 simple	RLM simple
Q3 2008	*74.8*	76.4	80.5	81.2	83.2	83.2
Q4 2008	144.4	135.5	*117.2*	123.8	129.6	127.0
Q1 2009	145.3	124.1	91.0	90.6	*86.4*	89.3
Q2 2009	*68.5*	72.7	72.7	73.1	75.5	74.6
Q3 2009	70.2	*62.7*	66.8	69.2	72.2	71.2
Q4 2009	*72.5*	75.1	88.1	82.6	86.1	89.7
Q1 2010	76.5	73.0	71.9	*70.9*	72.0	71.9
Q2 2010	70.4	73.2	69.6	67.5	*65.6*	71.1
Q3 2010	69.1	71.8	75.4	71.1	70.0	*68.3*
Q4 2010	87.4	*82.1*	83.4	82.5	83.2	82.4
Q1 2011	97.5	91.0	82.9	80.6	*78.7*	79.4
mean (μ_ε)	88.8	85.3	81.8	81.2	82.0	82.5
sd (σ_ε)	29.1	23.2	14.1	15.9	17.2	16.5

Table 11: For the S&P 500 data described in Section 5.3, we present out-of-sample results for the PCA plus L^1 model trained on $p' = 28$ variables together with five models described in Section 5.2. We show the median absolute error (in units of millions of dollars) made by forecasting quarter $\theta + 1$ retained earnings using quarter θ covariates; the models themselves were fitted by regressing quarter θ retained earnings onto quarter $\theta - 1$ covariates. The smallest number in each row has been italicized. The pruned models use only the top five variables for quarter $\theta - 1$ indicated in Table 10; the simple models use only the three variables common to all 11 top five lists.

	PCA+L^1	PCA+RLM	L^1 pruned	RLM pruned	L^1 simple	RLM simple
SVBA	3.70%	3.70%	3.68%	3.60%	3.58%	3.57%
S&P 500	5.40%	5.44%	4.72%	4.69%	4.71%	4.73%

Table 12: We compare the relative out-of-sample errors for retained earnings forecasts using both SVBA and S&P 500 data sets. The overall performance of all models is roughly one percentage point better for SVBA data. While pruning/simplification of the models does improve the SVBA relative errors, the improvement is more pronounced for the S&P 500 errors. The relative errors have been computed using the L^1 norm as in (28).