Variable Selection and Corporate Bankruptcy Forecasts

Shaonan Tian

College of Business, San Jose State University

Yan Yu

Carl H. Lindner College of Business, University of Cincinnati

Hui Guo

Carl H. Lindner College of Business, University of Cincinnati

Abstract

We investigate the relative importance of various bankruptcy predictors commonly used in the existing literature by applying a variable selection technique, the least absolute shrinkage and selection operator (LASSO), to a comprehensive bankruptcy database. Over the 1980 to 2009 period, LASSO admits the majority of Campbell, Hilscher, and Szilagyi's (2008) predictive variables into the bankruptcy forecast model. Interestingly, by contrast with recent studies, some financial ratios constructed from only accounting data also contain significant incremental information about future default risk, and their importance relative to that of market-based variables in bankruptcy forecasts increases with prediction horizons. Moreover, LASSO-selected variables have superior out-of-sample predictive power and outperform (1) those advocated by Campbell, Hilscher, and Szilagyi (2008) and (2) the distance to default from Merton's (1974) structural model.

Key Words: Discrete Hazard Model; Financial Ratios; LASSO; Market Information

JEL Classification: G33; G17; M41; C25

1. Introduction

Accounting and finance researchers have considered various predictive variables in the reduced-form corporate bankruptcy forecast model. Earlier studies, e.g., Beaver (1966), Altman (1968), Ohlson (1980), and Zmijewski (1984), have routinely used accounting variables, i.e., financial ratios constructed from only accounting data, as a gauge of default risk. In an attempt to improve the empirical performance of the reduced-form model, Shumway (2001) advocates for incorporating market-based variables in bankruptcy forecasts, in addition to two commonly used accounting-based variables. In a similar vein, Campbell, Hilscher, and Szilagyi (2008; CHS thereafter) introduce new market variables and accounting variables; and they also propose a modification of the accounting variables adopted in Shumway (2001) by using the market value of assets rather than the book value. While Shumway (2001) and CHS (2008) have shown their models exhibit noticeable improvement over the models proposed in previous studies, none of existing studies has provided a formal analysis on the relative importance of a comprehensive set of bankruptcy predictors using the advanced variable selection technique. In particular, there is no conclusive evidence on the role of accounting variables in bankruptcy forecasts. We try to fill the gap by introducing a state-of-art variable selection technique proposed by Tibshirani (1996)—the least absolute shrinkage and selection operator (LASSO).

Statisticians develop variable-selection methods to achieve two main objectives—(1) identifying relevant predictive variables and (2) improving prediction accuracy (see, e.g., Fan and Li (2001)). A formal variable-selection analysis thus allows us to shed new light on the corporate bankruptcy forecast literature in two important ways. First, it enables us to identify from an exhaustive set of bankruptcy predictors proposed in existing studies a parsimonious subset of the most relevant ones. Such identification has important implications for testing

bankruptcy theories, designing regulations in credit markets, and conducting credit risk analysis. Second, as we confirm in this paper, the selected reduced-form model shows improved in-sample and out-of-sample performance, when comparing with the prominent models in the existing literature.

LASSO penalizes regression coefficients through a shrinkage method and thus provides a sparse variable-set solution. It has been widely used in variable-selection studies (see, Tibshirani (1996) and Efron, Hastie, Johnstone, and Tibshirani (2004)) and is a state-of-the-art variable selection tool. LASSO enjoys the easy interpretability as the traditional subset variable selection does but has additional advantages of (1) the stability of model selection and (2) potential improvement in prediction accuracy. Compared with other commonly used variable selection methods such as the subset or stepwise selection, LASSO has several desirable statistical properties that suit particularly well for the main empirical issues that we try to address in this paper. First, given the rareness of default events, stability is a necessary requirement of variable selection techniques used for bankruptcy forecasts. LASSO is quite stable to small perturbations of data changes. Second, the shrinkage method may improve prediction accuracy. Third, LASSO produces an entire variable selection path that we can use to gauge the relative importance of the selected variables. Fourth, LASSO naturally overcomes the multicollinearity problem. Last, LASSO is computationally efficient, especially when there are a large number of candidate predictors.

We construct a comprehensive bankruptcy database by merging daily and monthly equity data from the Center for Research in Security Prices (CRSP) with annual financial information

from COMPUSTAT[1]. A company is in default if it files for either Chapter 7 (liquidation) or Chapter 11 (reorganization) bankruptcy protection. We include an exhaustive list of 39 accounting-based variables and market-based variables that have been used in the bankruptcy literature as candidate default-risk predictors. As in Shumway (2001), Chava and Jarrow (2004), CHS (2008), and others, we model the bankruptcy risk using the discrete hazard model. Shumway (2001) emphasizes that the discrete hazard model using time-varying panel data has important advantages compared with static models using cross-sectional data (e.g., Altman (1968), Ohlson (1980), and Zmijewski (1984)). This is because the latter ignore the fact that firms change over time and thus may produce biased and inconsistent bankruptcy probability estimates. In this paper, we adopt LASSO variable selection technique on time-varying covariates for the panel bankruptcy data.

We first focus on the forecast of one-year-ahead bankruptcy—the most commonly used forecast horizon in the existing literature, and then investigate how variable-selection results vary with forecast horizons. Over the full sample spanning the 1980 to 2009 period, LASSO selects seven predictive variables into the reduced-form bankruptcy forecast model. We find strong support for Shumway's (2001) argument of including market-based variables in bankruptcy forecasts. Two market variables advocated by Shumway (2001), i.e., stock return volatility and the excess stock return, and one market variable proposed by CHS (2008), i.e. stock price, enter into the LASSO-selected reduced-form model. Shumway (2001) shows that, consistent with the previous accounting studies, (1) the net income to total assets ratio and (2) the total liabilities to total assets ratio constructed from accounting data are significant predictors even when controlling for market-based variables in bankruptcy forecasts. CHS (2008), however, suggest

[1] Vassalou and Xing (2004) also use only CRSP and COMPUSTAT data to construct their bankruptcy database. While CHS (2008) incorporate additional proprietary data sources, our database is qualitatively similar to theirs. See Ding, Tian, Yu, and Guo (2012) for details.

that we should modify these two variables using the market value of assets instead of the book value. Our variable selection analysis allows us to shed light on this issue: LASSO selects CHS's modified financial ratios but not Shumway's (2001) original variables. Of CHS's eight predictive variables, five enter into our reduced-form bankruptcy forecast model, indicating that CHS have done a reasonably good job in selecting the bankruptcy predictors.[2] Nevertheless, the LASSO variable selection analysis differs from CHS's model in two ways. First, LASSO identifies two additional predictive variables—(1) the current liabilities to total assets ratio and (2) the total debts to total assets ratio constructed from only accounting data. This result reaffirms the important role of accounting variables in bankruptcy forecasts. Second, three of CHS's predictive variables, i.e., the market capitalization, the market to book ratio, and the ratio of cash and short-term assets to the market value of assets, do not enter into the LASSO-selected reduced-form model.[3] In our study, the LASSO variable selection results are strikingly consistent across subsample periods: The identical sets of predictive variables are selected over the 1980 to 2000, 1980 to 2002, 1980 to 2005, and 1990 to 2009 periods.

The distance to default (DD) constructed from Merton's (1974) structural model is a popular bankruptcy risk measure among practitioners. CHS (2008) and Bharath and Shumway (2008), however, find that DD provides relatively little information about future bankruptcy beyond the variables used in their reduced-form models. When we include DD as a candidate predictor along with the other 39 predictive variables, it does not enter into the LASSO-selected reduced-form model and the set of selected predictors is identical to that obtained without DD as a candidate predictor. In the out-of-sample forecast, our limited empirical study shows the

[2] Of five predictive variables proposed by Shumway (2001), four variables enter into our LASSO-selected reduced-form model either directly or in a modified form.

[3] Using our data, we confirm CHS's finding that these three variables have statistically significant in-sample predictive power for the default risk, although these variables are not selected by LASSO.

performance of the DD only model is similar to, or slightly better than, that of the CHS reduced-form model. By contrast, the LASSO-selected reduced-form model performs noticeably better than the DD only model over various out-of-sample testing periods.

CHS (2008) have advocated for constructing financial ratios using the market value of assets in default forecasts. By contrast, accounting researchers, e.g., Beaver, McNichols, and Rhie (2005), have reiterated the relevance of accounting-based variables by showing that their predictive power is strikingly consistent across time. In a similar vein, Das, Hanouna, and Sarin (2009) show that accounting data provide significant supplementary information in distress risk pricing, especially for firms with limited or no trading activity. We provide support for both arguments. LASSO selects the market value of assets for the net income to total assets ratio and the total liabilities to total assets ratio but chooses the book value for the current liabilities to total assets ratio and the total debts to total assets ratio.

As CHS (2008) emphasize, a firm's market value of equity is a more accurate gauge of its prospects than is the book value of equity. Hence, the predictive power of market leverage for bankruptcy risk likely reflects the conventional wisdom that firms with lower leverage are more able to pay off its debts. In addition, a firm with higher bankruptcy costs has more incentives to lower its bankruptcy risk by taking precautionary actions, for example, adopting low target leverage or taking low-risk project (e.g., George and Hwang (2010)). The book leverage is a more reliable measure of target leverage than is market leverage because many studies (e.g., Welch (2004) and Graham and Harvey (2001)) show that firms rarely counteract to changes in their capital structure caused by fluctuations in their stock prices. Therefore, book leverage correlates negatively with bankruptcy risk possibly because it is a proxy for precautionary actions taken by firms to reduce their bankruptcy risk.

Our conjectures have an interesting implication. Because individual firms are susceptible to large idiosyncratic shocks or their fortune can change quite drastically over time, market variables are more useful in short-run forecasts than in long-run forecasts. By contrast, because ceteris paribus a prudent firm is more able to withstand a financial market storm due to its precautionary actions intended for reducing bankruptcy risk than is a reckless firm, the relative importance of accounting-based variables may increase with forecast horizons when idiosyncratic risk averages out. We find strong support for this implication. When the prediction horizon is within two years, LASSO selects five market-based variables and two accounting-based variables. In contrast, for the three-year and five-year prediction horizons, of seven LASSO-selected covariates, there are five accounting-based variables but only two market-based variables. To the best of our knowledge, these results are novel. Moreover, our out-of-sample evaluations show the improvement of using LASSO-selected variables over CHS (2008)'s model at various prediction horizons.

The remainder of the paper proceeds as follows. We describe data in Section 2. We discuss the discrete hazard model and the LASSO variable selection results in Section 3. We report the model evaluation for one-year-ahead bankruptcy forecasts in Section 4. We analyze the DD measure from Merton's (1974) structural model in Section 5. We study different prediction horizons in Section 6. We offer some concluding remarks in Section 7.

2. Data

We construct the bankruptcy database by merging the daily and monthly CRSP equity data with the annually updated COMPUSTAT accounting data, spanning the 1980 to 2009 period. We have an entry for each firm-month with (1) bankruptcy information and (2) market and

financial variables. Companies report their accounting data with a delay. To ensure that the accounting information that we use is available at the time of bankruptcy forecasts, we lag all the annually updated accounting measures by four months. That is, all the forecasting variables are available to investors in real time. To alleviate the effect of outliers, we winsorize bankruptcy predictors at the 1st and 99th percentiles. Our bankruptcy database includes 17,570 firms and 1,571,115 firm-months with no missing observations.

To estimate the default risk, we need to construct a bankruptcy indicator. A company is in default if it files for bankruptcy under either Chapter 7 (liquidation) or Chapter 11 (reorganization) protection code. The bankruptcy indicator of a firm equals one only when the firm exited the database due to those bankruptcy filings. We assign a value of zero for the bankruptcy indicator for (1) firms that stayed or survived in the database through the end of the sampling period and (2) firms that exited from the database due to other reasons such as mergers and acquisitions. We identify 1,383 bankruptcy filings over the 1980 to 2009 period in our database. In Figure 1, we plot the number of bankruptcies for each year. Consistent with previous findings, Figure 1 shows that bankruptcy filings exhibit strong countercyclical patterns with peaks following the 1981-82, 1990-91, 2001, and 2007-09 business recessions.

We consider an exhaustive list of 39 financial and market variables as candidate bankruptcy predictors and briefly explain them in Table 1.[4] The Appendix also provides details of how we construct each variable using CRSP and/or COMPUSTAT data items. Those predictive variables are drawn from previous empirical bankruptcy studies, including Beaver (1966), Altman (1968), Ohlson (1980), Zmijewski (1984), Shumway (2001), Chava and Jarrow

[4] The variables used in this study are available only for publicly traded companies. Dwyer, Kocagil, and Stein (2004) propose some alternative predictive variables in the forecast of credit risk for privately held companies. As a robustness check, we also include these variables as candidate predictors and find that LASSO selects none of them. For brevity, we do not report these results but they are available upon request.

(2004), Dwyer, Kocagil, and Stein (2004), Beaver, McNichols, and Rhie (2005), Härdle, Lee, Schäfer, and Yeh (2009), Bharath and Shumway (2008), CHS (2008), Ding, Tian, Yu, and Guo (2012), and many others.

Earlier studies, e.g., Beaver (1966), Altman (1968), Ohlson (1980), and Zmijewski (1984), use various accounting variables in bankruptcy forecasts, and Altman's (1968) Z-score and Ohlson's (1980) O-score have been the standard distress risk measures for both practitioners and academic researchers. Some accounting researchers, e.g., Ohlson (1980), have also conjectured that including market-based variables may improve substantially bankruptcy forecasts. These authors, however, do not pursue this investigation because their main research interest is the informativeness of accounting data for bankruptcy rather than the search for a good bankruptcy forecast model. Shumway (2001) first provides empirical support for this conjecture by applying the discrete hazard model to panel data. He shows that three market variables—the relative market capitalization (RSIZE), the stock return in excess to the market return (EXCESS RETURN), and stock return volatility (SIGMA)—have significant predictive power for bankruptcy risk. Shumway (2001) finds that two accounting-based variables, the net income to total assets ratio (NIAT) and the total liabilities to total assets ratio (LTAT), are also significant bankruptcy predictors. Overall, Shumway's (2001) reduced-form model outperforms those proposed in earlier accounting studies, e.g., Altman (1968) and Zmijewski (1984).

Shumway's (2001) market-variable-augmented reduced-form model has become popular in the bankruptcy forecast literature, and CHS (2008) try to improve its empirical performance in three ways. First, CHS add a new market variable, the stock price, as a default predictor. Second, CHS advocate for constructing financial ratios using the market value of assets rather than the book value. That is, CHS replace NIAT and LTAT by the net income to the market

value of total assets ratio (NIMTA) and the total liabilities to the market value of total assets ratio (LTMTA), respectively. Last, CHS include two new financial ratios as bankruptcy predictors: the market-to-book equity ratio (MB) and the ratio of cash and short-term investment to the market value of total asset (CASHMTA).[5] CHS find that their model has a better in-sample fit than does Shumway's (2001) model. Nevertheless, neither Shumway (2001) nor CHS choose the variables in their reduced-form models via a statistical variable selection analysis, and we try to fill the gap in this paper.

3. Discrete Hazard Model and LASSO Variable Selection

3.1. Discrete Hazard Model

To evaluate a firm's bankruptcy risk over a given future period, we assume a logistic regression link between the bankruptcy indicator variable and time-varying covariates, following Shumway (2001), or equivalently the discrete logistic model (e.g., Cox (1972) and Ding, Tian, Yu, Guo (2012)). Specifically, the discrete hazard model for twelve-month or one-year-ahead default risk prediction is given by the following equation

$$\mathrm{P}\left(\mathrm{Y}_{i,t+12}=1 \middle| \mathrm{Y}_{i,t+12-1}=0, \mathrm{X}_{i,t}\right)=\frac{e^{\beta_0+\beta^{'}\mathrm{X}_{i,t}}}{1+e^{\beta_0+\beta^{'}\mathrm{X}_{i,t}}}, \tag{1}$$

where $\mathrm{X}_{i,t}$ is a covariate vector of time-varying firm-specific explanatory variables that are observable at time *t* for each of 17,570 public firms in our bankruptcy database or *i* = 1 to 17,570. β is a vector of covariate effect parameters and β_0 is a scalar parameter. The subscript *t* denotes the calendar time; in this paper, *t* represents the month end for monthly data. The dependent variable $\mathrm{Y}_{i,t+12}$ is a default indicator, which equals one if firm *i* files for bankruptcy

[5] MB correlates negatively with the cross-section of stock returns, and Fama and French (1996) suggest that this relation possibly reflects the fact that MB correlates negatively with distress risk.

protection in the month twelve given that it survives through the month eleven from time *t* and equals zero otherwise.

3.2. LASSO Variable Selection

Existing studies, e.g. Beaver (1966), Altman (1968), Beaver et al. (2005), Shumway (2001), Chava and Jarrow (2004), and CHS (2008), have introduced numerous accounting-based variables and market-based variables to improve the prediction accuracy of reduced-form bankruptcy forecast models. These authors often motivate their proposed predictive variables using arguably subjective expert or field judgment, and there is no consensus on which variables should be included in the reduced-form bankruptcy model of equation (1). To identify the most relevant variables from a comprehensive variable set considered in the literature, we introduce the state-of-the-art LASSO variable selection method for U.S. bankruptcy database described in the previous section[6]. Amendola, Restaino, and Sensini (2011) have applied LASSO to a binary classification problem using a small Italian sample with only accounting variables. In contrast, to the best of our knowledge, our study is the first one that offers the empirical insights on the debate between accounting-based and market-based variables in predicting default risk using an extensive U.S. bankruptcy database that is readily available to most academics.

Variable selection, which is essential for identifying relevant predictive variables and potentially improving prediction accuracy, has long been an important research topic in the statistics literature. Recent development in variable selection literature suggests a promising role for penalized shrinkage approaches (Tibshirani (1996, 2011), Zou (2006), Meier, Geer, and Bühlmann (2008)), which select predictive variables through shrunken coefficients under a pre-

[6] See Lane, Looney, and Wansley (1986) for a preliminary application of stepwise selection and Kumar and Ravi (2007) for a comprehensive survey study in bankruptcy literature.

specified roughness penalty. In this paper, we introduce the pioneer work proposed by Tibshirani (1996), namely the least absolute shrinkage and selection operator (LASSO), within the shrinkage context to select a parsimonious set of default predictor variables.

For the discrete hazard model in equation (1), we obtain LASSO parameter estimates by minimizing the negative log-likelihood function with a roughness penalty placed on the sum of the absolute value of the covariate parameters—the so-called "l_1 penalty" or "l_1 constraints"

$$\sum_{i=1}^{n}\left(-\mathrm{Y}_{i,t+12}\left(\beta_0+\beta'\mathrm{X}_{i,t}\right)+\log\left(1+\exp\left(\beta_0+\beta'\mathrm{X}_{i,t}\right)\right)\right) \qquad \text{subject to } \sum_{k=1}^{p}|\beta_k|\le s,$$

or equivalently,

$$\sum_{i=1}^{n}\left(-\mathrm{Y}_{i,t+12}\left(\beta_0+\beta'\mathrm{X}_{i,t}\right)+\log\left(1+\exp\left(\beta_0+\beta'\mathrm{X}_{i,t}\right)\right)\right)-\lambda\sum_{k=1}^{p}|\beta_k|,$$

where n is the number of firms and p is the number of predictive variables used in the hazard model. We control the amount of shrinkage through the roughness penalty tuning parameter s or λ. Note that a smaller value of s usually leads to a more parsimonious set of the selected predictive variables.[7]

The inclusion of the "l_1 norm penalty" or the constraint formulation in LASSO through simultaneous estimation and variable selection leads to some nice properties theoretically and computationally. LASSO selects variables by zeroing some coefficients and shrinking others. Hence, it retains the easy interpretability as subset selection and the stability as ridge regression

[7] In this paper, we implement empirical analysis using SAS software. The selection path is computed through an efficient least angle regression (LARS) algorithm developed by Efron, Hastie, Johnstone, and Tibshirani (2004). The best solution on the selected path, or equivalently, the number of the selected predictive variables, is identified by the optimal BIC criteria, which is the SAS default option. For the linear models, Knight and Fu (2000) show consistency of LASSO type estimators under some mild conditions. Wang and Leng (2007) establish consistency for adaptive LASSO and a BIC-type tuning parameter via least square approximation. As a robustness check, we find the same LASSO variable selection results using the LARS R routine provided by Efron et al. (2004) (downloadable from http://cran.r-project.org/web/packages/lars/index.html). Due to the big scope of the data we used, SAS is adopted in this study. Recent researchers also proposed some generalizations and variants of the LASSO (see Tibshirani (2011) for a review). However, most of the methods are not readily available in SAS.

(Hoerl and Kennard (1970)).[8] Stability is relevant in this setting because it helps us find the empirical variables that better predict bankruptcy, and these variables can inform us about the fundamental economic determinants of bankruptcy. Furthermore, the regularization shrinkage approach by LASSO naturally handles the multicollinearity problem. In bankruptcy prediction study, the predictors we use such as the accounting variables and market variables have strong correlations among themselves. For example, we include both total debts over total assets (FAT) and equity over total assets (SEQAT) in the candidate predictor set, whereas, the summation of FAT and SEQAT, in fact, is one by an accounting identity. With the presence of the multicollinearity, the superiority of shrinkage methods such as the ridge regression is always noted (Mahajan, Jain, and Bergier (1977), Vinod (1978), Mason and Perreault (1991)). LASSO is also computationally efficient so it can be easily implemented (e.g., Efron, Hastie, Johnstone, and Tibshirani (2004)).

Best-subset is a commonly used traditional variable selection approach that yields easily interpretable results. It, however, has some potential drawbacks. As a discrete process, subset selection is instable even with small changes of data that could in turn compromise the prediction accuracy (Breiman (1995, 1996), Tibshirani (1996), and Zou (2006)). The feasibility of applying subset selection in corporate bankruptcy prediction may also be a potential problem. With a comprehensive set of 39 candidate predictive variables, an exhaustive search of best-subset selection involves selection of the best model from 2^{39-1} or about 275 billion different combinations according to some criterion such as Akaike information criterion (AIC) (Akaike

[8] Breiman (1995, 1996) defines: "If the regression equations generated by a procedure do not change drastically with small changes in the data, the procedure is called stable". A comprehensive simulation study and pioneering theoretical analysis are conducted. Meinshausen and Bühlmann (2010) formally consider stability selection and stability paths for variable selection. Recently, Sun, Wang, and Fang (2013) propose a new tuning parameter selection via variable selection stability. In a given period of our bankruptcy study, stability refers to stable variable selection in the presence of small perturbations of data sampling, whereas, instability intuitively implies that small changes of the sample data could lead to large changes in selected variables.

(1974)). In practice, we usually adopt the stepwise-subset selection as a surrogate for the best-subset selection by sequentially deleting or adding one variable at a time according to some test statistics. The stepwise selection, however, may yield a local optimal solution rather than the global optimal solution due to its heuristic algorithm. It also ignores the stochastic errors in the variable selection stage (Fan and Li (2001)). Therefore, in this paper, we rely on LASSO in the selection of the most relevant bankruptcy predictors.

3.3. LASSO Variable Selection Results

We identify the set of the most important default predictors by applying the LASSO variable selection method to the full sample spanning the 1980 to 2009 period and report the results in Figure 2. The upper panel illustrates the evolution of estimated coefficients on all candidate predictive variables listed in Table 1 over the LASSO variable selection process. The horizontal axis indicates the constraint—the maximum value allowed for the sum of the absolute coefficient values. The vertical axis reports the coefficient estimates that correspond to the constraint on the horizontal axis. For restrictive constraints, LASSO parameter estimates are all close to zero. As the constraint gets relaxed, variables sequentially enter into the predictive regression as their LASSO parameter estimates increase in magnitude and become nonzero. Variables with stronger predictive power enter the process sooner, indicating their higher importance. The lower panel of Figure 2 illustrates the evolution of the in-sample AIC corrected with a factor of the finite sample size (AICC) as the constraint becomes less restrictive. AICC is a goodness-of-fit measurement with regard to information loss (see, e.g., Hurvich and Tsai (1989)): A smaller value of AICC indicates a better model fitting. Figure 2 shows that both market information and accounting data are important in forecasting default. This finding is

novel and interesting, and we elaborate on the importance of both sets of information below.

Shumway (2001) and CHS (2008) advocate for incorporating market information in bankruptcy forecast because compared with accounting information, market information has several advantages. First, the stock price is a forward looking variable that incorporates all available information. Second, the stock return volatility is a direct determinant of the default probability in Merton's (1974) structural model. Third, the market value of a firm's assets is a more accurate measure of the true value than is the book value. Consistent with Shumway (2001) and CHS's conjecture, Figure 2 shows that the stock price (PRICE) and the stock return volatility (SIGMA) are the first two predictive variables that LASSO selects. LASSO also selects the excess stock return (EXCESS RETURN) albeit at a relatively late stage. In addition, we document support for CHS's argument of using the market value of assets instead of the book value in the construction of financial ratios. Specifically, Shumway (2001) use the book value of equity in the construction of the net income to total assets (NIAT) and the total liabilities to total assets ratio (LTAT). CHS use the market value of equity for the net income to total assets (NIMTA) and the total liability to total assets (LTMTA). Figure 2 shows that both NIMTA and LTMTA enter into the bankruptcy forecast model but NIAT and LTAT do not.

Overall, our variable selection results are quite consistent with the variables advocated by Shumway (2001) and CHS (2008), the most widely recognized studies in default prediction literature. Specifically, except for the market capitalization, all the variables proposed by Shumway (2001) enter into the LASSO variable selection either directly or in a modified form. Similarly, LASSO chooses five out of eight variables used in CHS; two other CHS variables, corporate cash holdings (CASHMTA) and the market-to-book equity ratio (MB), do not enter the

LASSO-selected reduced-form model, however.[9]

Moreover, LASSO identifies important bankruptcy predictors, i.e., two accounting-based leverage measures constructed from only accounting data, that Shumway (2001) and CHS (2008) do not include in their models. In Figure 2, the current liabilities to total book assets ratio (LCTAT) and the total debts to total book assets ratio (FAT) enter into the LASSO-selected bankruptcy forecast model. In the next subsection, we further confirm that those two accounting-based variables are statistically significant in in-sample estimate and improve the out-of-sample performance of the reduced-form model. Leverage is a commonly used bankruptcy predictor because it is a gauge of a company's ability to pay off its debts. CHS point out that because the market value is a more accurate measure of the true value of a firm's assets than is the book value, financial ratios constructed using the market value of assets is a better measure of the firm's ability to pay off its debts. Below, we propose a novel explanation for why book leverage measures such as LCTAT and FAT provide information about future default risk that is complement to information conveyed by market leverage measures.

Specifically, we argue that book leverage correlates negatively with future bankruptcy risk because it is a measure of target leverage that serves as a proxy for actions taken by firms to reduce their bankruptcy costs. Ceteris paribus, a firm with higher bankruptcy costs has more incentives to reduce its bankruptcy risk by taking precautionary actions, e.g., adopting low target leverage and taking low-risk projects.[10] In particular, the tradeoff theory of capital structure stipulates that firms with higher bankruptcy costs should adopt lower target leverage (e.g.,

[9] Seven predictive variables are identified by LASSO using the optimal BIC criteria. We experimented with adding the next (eighth) selected variable, Cash/Total Assets (CHAT), to the empirical model. We also experimented with removing the last (seventh) selected variable, Current Liabilities/Total Asset (LCTAT), from the empirical model. We find that, with the addition of Cash/Total Assets ratio or the removal of Current Liabilities/Total Asset, both in-sample and out-of-sample performance is qualitatively similar to that of the empirical model with seven predictor variables, when comparing with the CHS model and the DD only model.

[10] Davydenko, Strebulaev, and Zhao (2012) find that firms with lower bankruptcy costs have lower credit ratings.

George and Hwang (2010), Johnson, Chebonenko, Cunha, D'Almeida, and Spencer (2011), and Glover (2013)).[11] Hence, we expect a negative relation between target leverage and bankruptcy risk. From a theoretical perspective, Shyam-Sunder and Myers (1999) argue that book leverage is a better measure of target leverage than is market leverage. Empirical studies, e.g., Cole, Daniel, and Naveen (2006), Welch (2004), Graham and Harvey (2001), also point out that market leverage is a poor measure of target leverage because firms rarely adjust their market leverage as a response to changes in their stock prices: Variation in market leverage is mainly due to stock price fluctuations. Overall, Parsons and Titman (2008; pp. 6) conclude that *many researchers prefer to scale debt by book assets instead*.

That is, we argue that leverage forecasts bankruptcy risk in two ways. First, market leverage is a measure of a firm's ability to pay off its debts. Second, book leverage is a proxy for precautionary actions taken by a firm to reduce its bankruptcy risk. These conjectures have an important implication: The predictive power of market leverage relative to that of book leverage decreases with forecast horizons. As mentioned above, market leverage comoves strongly with stock prices, which have a large idiosyncratic component because a firm's fortune can change quite drastically over time. Therefore, while a firm's fortune in the recent past is a more important determinant of the firm's bankruptcy risk in the near future, ceteris paribus, we expect that a prudent firm should have lower bankruptcy risk than should a reckless firm in the long run when idiosyncratic risk averages out.[12] We can also illustrate this difference from the statistical point of view. Recall that while the market leverage comoves strongly with stock prices in both

[11] Empirical evidence is broadly consistent with this implication (e.g., Parsons and Titman (2008)) and financial economists routinely use this implication to interpret their empirical findings (e.g., Andrade and Kaplan (1998)).

[12] This argument is analogous to the determinants of heart attack risk. While the blood pressure level and the cholesterol level are good predictors of heart attacks in the near future, life styles are a more important determinant of heart attacks over long horizons. That is, among people who currently have good blood pressure and cholesterol levels, those who have unhealthy life styles are more likely to have heart attacks eventually. Conversely, among people who currently have poor blood pressure and cholesterol levels, those who have healthy life styles are more likely to avoid heart attacks eventually.

short and long horizons, mangers rarely attempt to counteract the influence of stock prices on their target or optimal capital structure. Therefore, book leverage is more persistent than market leverage and thus has relatively stronger predictive power for bankruptcy risk over long horizons.[13] We document strong empirical support for this implication in Section 6. To the best of our knowledge, such an empirical link between accounting-based variables and the bankruptcy risk is novel.

Lastly, as a robustness check, we repeat the LASSO analysis using various subsample periods, including the 1980 to 2000, 1980 to 2002, 1980 to 2005, and 1990 to 2009 periods. Interestingly, we find the set of LASSO-selected variables are strikingly consistent across time.[14]

4. Empirical Results

4.1. Model Evaluation

To evaluate the performance of bankruptcy forecast models, we provide a comprehensive list of measurements from both in-sample and out-of-sample dimensions. First, we use formal model information criteria based on the negative log-likelihood and a complexity penalty to evaluate the overall in-sample performance of the discrete hazard model. For example, AIC is a popular goodness-of-fit measurement for likelihood-based model selection using two times the number of parameters as a penalty. A model with a smaller AIC is more desirable. In general, a good model attempts to balance its accuracy and complexity, which are often termed as the tradeoff between bias and variance by statisticians. For instance, a bankruptcy prediction model

[13] This point is analogous to that made by Campbell, Lo, and MacKinlay (1997) in the context of forecasting excess stock market returns. These authors show that stock market return predictability tends to increase with forecast horizons when conditional equity premium is persistent. It is important to note that target leverage is persistent because we assume stable bankruptcy costs over time. Extant empirical studies have not investigated this important assumption formally. Addressing this issue is beyond the scope of this paper and we leave it for future research.

[14] As a robustness check, we apply the stepwise variable selection technique to the same bankruptcy data over the aforementioned sampling periods and find that the resulting sets of the selected predictive variables vary drastically over different sample periods. For brevity, we do not report these results here but they are available upon request.

with a larger number of explanatory variables always yields a better in-sample likelihood but not necessarily a better AIC; and most importantly, it might have worse out-of-sample prediction due to overfitting or data-snooping.

AUC, the area under the receiver operating characteristic (ROC) curve, is a popular measure of a model's discriminatory power (Hosmer and Lemeshow (2000)). We commonly use AUC to evaluate a model's ability to discriminate between the binary events, e.g., bankruptcy versus non-bankruptcy, based on its predicted bankruptcy probabilities. The accuracy ratio, defined as the difference between AUC and 0.5 multiplied by two, is another commonly used gauge for corporate bankruptcy model evaluation (Duffie, Saita, and Wang (2007)). A value of zero for the accuracy ratio or a value of 0.5 for AUC indicates a random forecast, while a value of one for the accuracy ratio or AUC corresponds to a perfect forecast.

McFadden's pseudo-R^2 (McFadden (1974)) is a log-likelihood-based information measure and equals one minus the log-likelihood ratio of the fitted model over the intercept-only model. McFadden's pseudo-R^2 is commonly used to evaluate the goodness of fit for the estimated model (CHS (2008)), and the model with higher McFadden's pseudo-R^2 value is more desirable.

For practitioners, it is crucially important to develop a default forecast model with an accurate out-of-sample prediction. For example, Basel II standard for internal-rating-based approach uses one-year-ahead default probability prediction and out-of-sample backtesting for default model validation. In the out-of-sample analysis, we implement a strategy similar to that used in Shumway (2001). We first build our discrete hazard model using the bankruptcy data over the training period. With the variables selected by LASSO and the coefficient estimates from the discrete hazard model fitting, we then predict the probabilities of bankruptcy for the firms over the testing period and report the out-of-sample accuracy ratio and AUC.

In addition, we also report the out-of-sample Brier score, which is the average of the squared differences between the predicted values and the actual outcomes. The Brier score is a popular statistic for assessing a model's overall prediction accuracy. When evaluating the bankruptcy prediction accuracy, the Brier score measures how close the predicted default probability is to the company's true health status. A Brier score of zero indicates a perfect prediction, where all defaults were predicted with a default probability of one and all healthy firms were predicted with a default probability of zero.

Last, we evaluate the out-of-sample performance using decile rankings, as in Shumway (2001), Chava and Jarrow (2004), and others. For each year in the testing period, we rank companies in deciles by their predicted bankruptcy probabilities. Specifically, the first decile contains the companies with the highest default probabilities, and the tenth decile is for the companies with the lowest default probabilities. We tabulate the percentage of actual bankruptcy firms in each decile. A high percentage in the high bankruptcy probability deciles implies good out-of-sample performance.

4.2. In-Sample Estimation and Out-of-Sample Forecasts

Table 2 summarizes the estimation results of the discrete hazard model over the entire bankruptcy database in the first two columns. Column 1 reports the results for the reduced-form model with LASSO-selected variables. Panel A shows that all LASSO-selected predictive variables are statistically significant at the 1% level with expected signs. For comparison, in Column 2 of panel A, we report the results for the CHS model, which are similar to those reported in CHS, although we use an updated sample period. Panel B shows that the LASSO-selected model has a lower AIC value and a higher McFadden's pseudo-R^2 than does the CHS

model, indicating that the former has less information loss and thus provides a better fit for the bankruptcy data. Similarly, with a higher AUC value, the LASSO-selected model conveys a better discriminatory ability than does the CHS model. As a robustness check, we re-estimate the LASSO-selected model and the CHS model using the subsample spanning the 1980 to 2002 period, and results reported in columns 3 and 4 are qualitatively similar to their counterpart reported in columns 1 and 2 of Table 2[15]. To summarize, the LASSO-selected model provides a better in-sample explanation for the bankruptcy data than does the CHS model.

We then evaluate our model's out-of-sample predictive ability by splitting the bankruptcy data into a training sample ending in 2002 and a testing sample over the 2003 to 2009 period. Specifically, we employ the discrete hazard model on the bankruptcy records over the training period, and evaluate its out-of-sample predictive performance using the testing sample. Over the holdout sample spanning the 2003 to 2009 period, we sort stocks equally into ten portfolios by their predicted default probabilities, which decrease from the first decile to the tenth decile. Table 3 reports the out-of-sample evaluation measures, including the percentage of actual bankruptcy filings in each decile, the out-of-sample accuracy ratio, AUC, and the Brier score. Consistent with the in-sample estimation results reported in Table 2, we find that LASSO-selected model exhibits a better out-of-sample performance than does the CHS model. The LASSO-selected model delivers an almost 80 percent correct prediction rate in the top two deciles (column 1), comparing with 66 percent for the CHS model (column 2). The CHS model yields an out-of-sample accuracy ratio of 0.636 (equivalent to AUC of 0.818), which is lower

[15] Following the suggestion from one of the referees, we have also investigated an extended sample spanning the 1970 to 2009 period and compared LASSO and CHS by decade. We find that for most recent three decades (i.e., 1980 to 1989, 1990 to 1999, and 2000 to 2009), the LASSO-selected variables consistently outperform the CHS model, although the improvement is rather moderate. Overall, our main finding is that the LASSO-selected reduced-form model consistently outperforms the CHS model across subsamples.

than that of 0.682 (equivalent to AUC of 0.841) for the LASSO-selected model.

As a robustness check, we evaluate our model's out-of-sample performance over two different (2001 to 2009 and 2006 to 2009) periods. The training samples for those two testing data sets are from the 1980 to 2000 period and the 1980 to 2005 period, respectively. Table 4 shows that the LASSO-selected model again outperforms the CHS model with a higher accuracy ratio and a better performance in the decile ranking, while the two models have similar out-of-sample Brier scores. To summarize, our results show that the model based on the LASSO selection has overall better out-of-sample performance than the popular CHS model.

5. A Comparison with Distance to Default

In Merton's (1974) bond pricing model, DD depends on the difference between a firm's asset value and the face value of its debts, scaled by the volatility of the firm's asset value. DD is a leading alternative bankruptcy risk measure, and there is an ongoing debate about the relative performance of the structural versus the reduced-form bankruptcy forecast models. Hillegeist, Keating, Cram, and Lundstedt (2004) find that the default probability derived from the structural model performs substantially better than the Z-score or O-score in bankruptcy forecasts. CHS (2008) and Bharath and Shumway (2008), however, find that DD provides relatively little information about future default risk beyond the market variables and financial ratios employed in their reduced-form models. To address this issue, we follow Vassalou and Xing (2004) and construct the DD measure using CRSP and COMPUSTAT data.[16] We add DD to the candidate predictor set and then apply the LASSO variable selection method to determine the most

[16] Practitioners, e.g., Moody's KMV, adopt the empirical distribution of DD estimated from a large database to obtain an implied probability of default, called the expected default frequency or EDF, which may potentially yield better fitting and prediction results (see Hamilton, Sun, and Ding (2011)). However, the empirical distribution is proprietary data, which are unavailable to us for comparison.

important forecasting variables over the full sample. Interestingly, LASSO variable selection coefficient path is virtually identical to that reported in Figure 2, which we obtained by excluding DD from the candidate predictor set. That is, LASSO does not select DD, and including DD as a candidate variable does not affect our results in any qualitatively manner.

In Table 3, we report the predictive power of DD in the out-of-sample tests over the 2003 to 2009 period. For the one-year-ahead prediction, the DD only model (column 3) has an out-of-sample AUC of 0.824 (or equivalently, an accuracy ratio of 0.648). Comparing with CHS (2008) model (column 2), the DD only model shows some improved overall discriminatory ability; nevertheless, its predictive ability is noticeably weaker than that of the LASSO-selected reduced-form model (column 1). On the other hand, the DD only model has a higher Brier score than the CHS (2008) model and the LASSO-selected reduced-form model, suggesting that the DD only model has a poorer out-of-sample performance. Table 4 shows that validation tests over the 2001 to 2009 and 2006 to 2009 periods provide qualitatively similar results. The LASSO-selected model (column 1) consistently demonstrates improved discriminatory ability over both the CHS model (column 2) and the DD only model (column 3).

Furthermore, as in CHS (2008), we investigate DD's in-sample explanatory power by adding it to the LASSO-selected reduced-form model for 1-month, 6-month, 12-month, and 24-month forecast horizons (see Section 6.2 for details of models with different forecast horizons). Over the full sample from 1980 to 2009, our results are qualitatively similar to those reported in CHS (2008). Specifically, for 1-month-ahead prediction, DD enters the model with an unexpected positive sign. DD becomes insignificant for 6-month-ahead prediction. For longer prediction horizons, for example, 12-month and 24-month prediction, DD enters the model with the expected negative sign. Overall, from our limited empirical study, we concur with CHS

(2008) and Bharath and Shumway (2008) that including other relevant variables beyond DD such as the variables selected by LASSO in the reduced-form model would be useful. For brevity, we do not tabulate these results but they are available upon request.

6. Different Prediction Horizons

Foreseeing the default risk at different horizons is of great interests to practitioners and researchers. An important question is whether we should use the same set of predictive variables for different forecast horizons. For example, we conjecture in Section 3 that the relative importance of accounting variables should increase with bankruptcy forecast horizons if accounting variables forecast default risk because they are proxies of precautionary actions taken by firms to reduce their default risk. In this section, we attempt to address these issues using LASSO variable selection.

6.1. Model

To investigate a firm's default risk at different prediction horizons, we make some minor modifications to the discrete hazard model in equation (1). We assume the same logistic link between the default probability and the covariates,

$$\mathrm{P}\left(\mathrm{Y}_{i,t+j}=1\middle|\mathrm{Y}_{i,t+j-1}=0,\mathrm{X}_{i,t}\right)=\frac{e^{\beta_{0,j}+\beta_j'\mathrm{X}_{i,t}}}{1+e^{\beta_{0,j}+\beta_j'\mathrm{X}_{i,t}}}\quad, \tag{2}$$

where the subscript j denotes the prediction horizon. For example, $j=12$ corresponds to the 12-month or 1-year-ahead bankruptcy prediction model, as in Section 2.1. In this Section, we set j to different values of one, six, twelve, twenty four, thirty six, and sixty, to investigate one-month, six-month, one-year, two-year, three-year, and five-year-ahead bankruptcy predictions,

respectively. Again, $X_{i,t}$ is a vector of time-varying firm-specific explanatory variables and $Y_{i,t+j}$ is the default indicator conditional on survival in the past *j*-1 months. $\beta_{0,j}$ and β_j are the scalar parameter and the covariate effect vector, respectively. We use the additional subscript j on the covariate effect because the parameter estimates change with forecast horizons.

6.2. Variable Selection Results

Table 5 reports the LASSO variable selection results for different prediction horizons. The first six columns report the sets of selected variables for one-month, six-month, one-year, two-year, three-year, and five-year forecasting horizons. For comparison, the last column lists the variables used in CHS (2008), which we highlight using shaded areas. Note that the variables without highlight are financial variables constructed from only accounting information.

Variable selections are strikingly consistent for relatively short forecast horizons. LASSO selects an identical set of predictive variables for the horizons of two years or shorter, and market-based variables dominate accounting-based variables. For example, five market-based variables are in the set of selected covariates, while LASSO selects only two accounting-based variables. Interestingly, when the prediction horizon increases, more accounting-based variables enter into the default prediction model, while some market-based variables, including the profitability variable (NIMTA) and the excess return (EXCESS RETURN), drop out from the LASSO-selected reduced-form bankruptcy forecast model. For example, both three-year-ahead and five-year-ahead default prediction models include five accounting-based variables but only two market-based variables (PRICE and SIGMA for the three-year-ahead prediction and PRICE and LTMTA for the five-year-ahead prediction). Thus, as conjectured, the forecasting power of the market-based variables becomes weaker relative to that of the accounting-based variables

when the prediction horizon increases.

CHS (2008) also investigate the in-sample fit of their reduced-form bankruptcy forecast model over several different forecast horizons. They find that the predictive power of stock return volatility, market capitalization, and market-to-book equity ratio increases with forecast horizons. Based on these findings, CHS (pp. 2914) suggest that "*overall, market-based variables become more important relative to accounting variables as we increase the forecast horizon"*. Our proposed variable-selection analysis confirms the CHS finding that stock return volatility and the stock price (which correlates closely with market capitalization) are important long-run bankruptcy predictors. However, we also shed new light on bankruptcy forecasts over long horizons. Specifically, many accounting variables omitted from the CHS model are significant determinants of long-run corporate bankruptcy. That is, in contrast with the CHS conclusion, we find that the accounting-based variables become more important relative to the market-based variables when forecast horizon increases. More importantly, as we discuss in the next subsection, LASSO-selected reduced-form models have better both in-sample and out-of-sample predictive power than does the CHS model. To the best of our knowledge, these results, which highlight the importance of selecting bankruptcy predictors from a comprehensive set of candidate forecasting variables, are novel.

6.3. In-Sample and Out-of-Sample Results

For each prediction horizon, we estimate the discrete hazard model with the LASSO-selected covariates using the data from 1980 to 2002 period, and assess its out-of-sample performance over the 2003 to 2009 period. For comparison, we also conduct the same out-of-sample forecast for the CHS model. The results are in Table 6. Panels A through F report the in-

sample and out-of-sample forecast results for the one-month, six-month, one-year, two-year, three-year, and five-year-ahead predictions. We use AIC and AUC to measure in-sample performance and use the accuracy ratio and AUC for out-of-sample performance.

For all prediction horizons, the LASSO-selected models exhibit better in-sample fits in terms of AIC than does the CHS (2008) model. With regard to the in-sample AUC, the LASSO-selected models' performance is comparable to, if not better than, that of the CHS (2008) model. Similarly, for the out-of-sample forecast, except for the one-month-ahead prediction, the LASSO-selected models consistently show a noticeably better accuracy ratio and AUC than those of the CHS model.

7. Conclusion

Using a comprehensive U.S. bankruptcy database constructed from CRSP and COMPUSTAT, we apply a state-of-the-art variable selection method, LASSO, to the discrete hazard model of corporate bankruptcy and document several important results. First, we find that the accounting-based variables provide significant supplemental information about future default risk beyond that of (1) the market-based variables and (2) financial ratios constructed using the market value of assets. Second, the reduced model selected via the LASSO method performs better in out-of-sample prediction than do the models adopted in the previous studies, including the CHS (2008) model. Third, the distance to default has negligible predictive power when we control for the LASSO-selected predictive variables. Last, the importance of accounting-based variables relative to that of market-based variables increases with forecast horizons.

Acknowledgement

The authors thank Alex Borisov, Jens Hilscher, Xiaowen Jiang, Carol Alexander (the editor), and two anonymous referees for their helpful comments.

Appendix

In this appendix, we explain how we construct each candidate predictive variable using the CRSP and/or COMPUSTAT data items. We follow CHS (2008) in the construction of the variables EXCESS RETURN, SIGMA, PRICE, and MB. EXCESS RETURN is a firm's log excess return on its equity relative to that on the S&P 500 index. SIGMA is the standard deviation of the daily stock return observed over the previous three months. PRICE is the equity price per share truncated from the above at the value of $15 and then takes the logarithm. MB is the ratio of the market equity to the adjusted book equity to which we add a 10% difference between the market equity and book equity. All series are available to investors in real time. Below we provide details for the other 35 predictive variables.

ACTLCT=ACT/LCT; APSALE=AP/SALE; CASHAT=CHE/AT;

CASHMTA= CHE/(PRICE*SHROUT+LT+MIB);

CHAT=CH/AT; CHLCT=CH/LCT; (EBIT+DP)/AT=(EBIT+DP)/AT;

EBITAT=EBIT/AT; EBITSALE=EBIT/SALE; FAT=(DLC+0.5*DLTT)/AT;

FFOLT= FFO/LT; INVCHINVT=INVCH/INVT; INVTSALE=INVT/SALE;

(LCT-CH)/AT=(LCT-CH)/AT; LCTAT=LCT/AT; LCTLT=LCT/LT;

LCTSALE= LCT/SALE; LT/(LT+MKET)= LT/(LT+MKET); LTAT=LT/AT;

LTMTA= LT /(PRICE*SHROUT+LT+MIB); LOG(AT)= log(AT);

LOG(SALE)= log(abs(SALE)); MVEF=(abs(PRCC_F)*CSHO)/(DLC+0.5*DLTT);

NIAT=NI/AT; NIMTA=NI/(PRICE*SHROUT+LT+MIB); NISALE=NI/SALE;

OIADPAT=OIADP/AT; OIADPSALE=OIADP/SALE; QALCT= (ACT – INVT)/LCT;

REAT=RE/AT; RELCT=RE/LCT; RSIZE= log(PRICE*SHROUT/TOTVAL);

SALEAT=SALE / AT; SEQAT=SEQ/AT; WCAPAT=WCAP/AT.

References

Akaike, H., 1974. “A New Look at the Statistical Model Identification”, IEEE Transactions, Automatic Control, 19(6), 716-723.

Altman, E. I., 1968. “Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy”, Journal of Finance, 23, 589-610.

Amendola, A., Restaino M., and Sensini L., 2011. “Variable selection in default risk models”, Journal of Risk Model Validation, 5 (1), 3-19.

Andrade, G., and Kaplan, S., 1998. “How Costly is Financial (Not Economic) Distress? Evidence from Highly Leveraged Transactions that Became Distressed”, Journal of Finance, 53, 1443-1493.

Beaver, W. H., 1966. “Financial Ratios as Predictors of Failure”, Journal of Accounting Research, 4, 71-111.

Beaver, W. H., McNichols, M. F. and Rhie, J., 2005. “Have Financial Statements Become Less Informative? Evidence from the Ability of Financial Ratios to Predict Bankruptcy”, Review of Accounting Studies, 10, 93-122.

Bharath, S. and Shumway, T., 2008. “Forecasting Default with the Merton Distance to Default Model”, Review of Financial Studies, 21 (3), 1339-1369.

Breiman, L., 1995. “Better Subset Regression Using the Nonnegative Garotte”, Technometrics, 37, 373-384.

Breiman, L., 1996. “Heuristics of Instability and Stabilization in Model Selection”, Annals of Statistics, 24, 2297-2778.

Campbell, J., Hilscher, J., and Szilagyi, J., 2008. “In Search of Distress Risk”, Journal of Finance, 63, 2899-2939.

Campbell, J., Lo, A., and MacKinlay, A., 1997. “The Econometrics of Financial Markets”, Princeton, NJ: Princeton University Press.

Chava, S. and Jarrow, R. A., 2004. “Bankruptcy prediction with industry effects”, Review of Finance, 8, 537-569.

Coles, J., Daniel, N., and Naveen, L., 2006, “Managerial Incentives and Risk-Taking”, Journal of Financial Economics, 79, 431-468.

Cox, D. R., 1972. “Regression Models and Life-Tables”, Journal of the Royal Statistical Society, Series B, 34, 187–220.

Das, S., Hanouna, P., and Sarin, A., 2009. “Accounting-Based versus Market-Based Cross-Sectional Models of CDS Spreads”, Journal of Banking & Finance, 33, 719-730.

Davydenko, S., Strebulaev, I., and Zhao, X., 2012. “A Market-Based Study of the Cost of Default", Review of Financial Studies, 25, 2959-2999.

Ding, A. A., Tian, S., Yu, Y., and Guo, H., 2012. “A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction”, Journal of the American Statistical Association, 107, 990-1003.

Duffie, D., Saita, L., and Wang, K., 2007. “Multi-period Corporate Default Prediction with Stochastic Covariates”, Journal of Financial Economics, 83, 635-665.

Dwyer, D. W., Kocagil, A. E., and Stein, R. M., 2004. “MOODY’S KMV RISKCALC v3.1 MODEL”, Moody’s KMV.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., 2004. “Least Angle Regression”, Annals of Statistics, 32, 407-499.

Fama, E. F., and French, K. R., 1996. “Multifactor Explanations of Asset Pricing Anomalies”, Journal of Finance, 51, 55-84.
Fan, J., and Li, R., 2001. “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” Journal of the American Statistical Association, 96, 1348-1360.
George, T. J., and Hwang, C., 2010. “A resolution of the distress risk and leverage puzzles in the cross section of stock returns,” Journal of Financial Economics, 96, 56-79.
Glover, B., 2013. “The Expected Cost of Default”, Journal of Financial Economics, forthcoming.
Graham, J., and Harvey, C., 2001. “The Theory and Practice of Corporate Finance: Evidence from the Field”, Journal of Financial Economics, 60, 187–243.
Hamilton, D. T., Sun, Z., and Ding, M., 2011. “Through-the-Cycle EDF Credit Measures”, Moody’s KMV.
Härdle, W., Lee, Y. J., Schäfer, D., and Yeh, Y. R., 2009. “Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for predicting the default risk of Companies”, Journal of Forecasting, 28, 512-534.
Hillegeist, S., Keating, E., Cram, D., and Lundstedt, K., 2004. “Assessing the Probability of Bankruptcy”, Review of Accounting Studies, 9, 5-34.
Hoerl, A. E. and Kennard, R.W., 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, Technometrics, 12(1), 55-67.
Hosmer, D. W., and Lemeshow, S., 2000. “Applied Logistic Regression”, 2nd Ed., Wiley-Interscience.
Hurvich, C. M., and Tsai, C. L., 1989. “Regression and time series model selection in small samples”, Biometrika, 76, 297-307.
Johnson, T.C., Chebonenko, T., Cunha, I., D’Almeida, F., and Spencer, X., 2011. “Endogenous Leverage and Expected Stock Returns”, Finance Research Letters, 8(3), 132-145.
Knight, K., and Fu, W., 2000. “Asymptotics for Lasso-Type Estimators”, the Annuals of Statistics, 28(5), 1356-1378.
Kumar, P., and Ravi, V., 2007. “Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques: A Review”, European Journal of Operational Research, 180, 1-28.
Lane, W., Looney, S., and Wansley, J., 1986. “An Application of the Cox Proportional Hazards Model to Bank Failure”, Journal of Banking and Finance, 10, 511-531.
Mahajan, V., Jain, A. K., and Bergier, M., 1977. “Parameter Estimation in Marketing Models in the Presence of Multicollinearity: An Application of Ridge Regression”, Journal of Marketing Research, 14(4), 586-591.
Mason, C. H., and Perreault, W. D. Jr., 1991. “Collinearity, Power, and Interpretation of Multiple Regression Analysis”, Journal of Marketing Research, 28(3), 268-280.
McFadden D., 1974. “Conditional logit analysis of qualitative choice behavior”, In Frontiers in Econometrics, Zarembka, P. (ed.). Academic: New York: 105–142.
Meier, L., Geer, S., and Bühlmann, P., 2008. “The group lasso for logistic regression”, Journal of the Royal Statistical Society: Series B, 70(1), 53-71.
Meinshausen, N., and Bühlmann, P., 2010. “Stability Selection”, Journal of the Royal Statistical Society: Series B, 72, 414-473.
Merton, R. C., 1974. “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates”, Journal of Finance, 29, 449-470.
Ohlson, J. S., 1980. “Financial Ratios and the Probabilistic Prediction of Bankruptcy”, Journal of Accounting Research, 19, 109-131.

Parsons, C., and Titman, S., 2008, “Empirical Capital Structure: A Review”, Foundations and Trends in Finance, 3, 1-93.
Shyam-Sunder, L., and Myers, S., 1999, “Testing Static Tradeoff against Pecking Order Models of Capital Structure”, Journal of Financial Economics, 51, 219-244.
Shumway, T., 2001. “Forecasting Bankruptcy More Accurately: A Simple Hazard Model”, Journal of Business, 74, 101-124.
Sun., W., Wang, J., and Fang, Y., 2013. “Consistent Selection of Tuning Parameters via Variable Selection Stability”, Journal of Machine Learning Research, 14, 3419-3440.
Tibshirani, R., 1996. “Regression Shrinkage and Selection via the Lasso”, Journal of the Royal Statistical Society, Series B, 58, 267-288.
Tibshirani, R., 2011. “Regression shrinkage and selection via the lasso: a retrospective”, Journal of the Royal Statistical Society: Series B, 73(3), 273-282.
Vassalou, M., and Xing, Y., 2004. “Default Risk in Equity Returns”, Journal of Finance, 59(2), 831-868.
Vinod, H. D., 1978. “A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares”, Review of Economics and Statistics, 60(1), 121-131.
Wang, H., and Leng, C., 2007. “Unified LASSO Estimation by Least Squares Approximation”, Journal of the American Statistical Association, 102, 1039-1048.
Welch, I., 2004. “Capital Structure and Stock Returns”, Journal of Political Economy, 112, 106-132.
Zmijewski, M., 1984. “Methodological Issues Related to the Estimation of Financial Distress Prediction Models”, Journal of Accounting Research, 22, 59-82.
Zou, H., 2006. “The Adaptive Lasso and Its Oracle Properties”, Journal of the American Statistical Association, 101, 1418–1429.

Table 1: Variable Description

Note: The table provides the description of the 39 bankruptcy predictors used in the variable selection analysis.

Variable	Description	Variable	Description
ACTLCT	Current Assets/ Current Liabilities	LTMTA	Total Liabilities /(Market Equity + Total Liabilities)
APSALE	Accounts Payable / Sales	LOG(AT)	log(Total Assets)
CASHAT	Cash and Short-term Investment / Total Assets	LOG(SALE)	log(Sale)
CASHMTA	Cash and Short-term Investment /(Market Equity + Total Liabilities)	MB	Market-to-Book Ratio
CHAT	Cash / Total Assets	MVEF	Market Equity (Yearly) / Total Debit
CHLCT	Cash / Current Liabilities	NIAT	Net Income / Total Asset
(EBIT+DP)/AT	(Earnings before Interest and Tax + Amortization and Depreciation) / Total Asset	NIMTA	Net Income /(Market Equity + Total Liabilities)
EBITAT	Earnings before Interest and Tax / Total Asset	NISALE	Net Income / Sales
EBITSALE	Earnings before Interest and Tax / Sales	OIADPAT	Operating Income / Total Asset
EXCESS RETURN	Excess Return Over S&P 500 Index	OIADPSALE	Operating Income / Sales
FAT	Total Debts / Total Assets	PRICE	Log(Price)
FFOLT	Funds from Operations / Total Liabilities	QALCT	Quick Assets / Current Liabilities
INVCHINVT	Growth of Inventories / Inventories	REAT	Retained Earnings / Total Asset
INVTSALE	Inventories / Sales	RELCT	Retained Earnings / Current Liabilities
(LCT-CH)/AT	(Current Liabilities – Cash) / Total Asset	RSIZE	Log(Market Capitalization)
LCTAT	Current Liabilities / Total Asset	SALEAT	Sales / Total Assets
LCTLT	Current Liabilities / Total Liabilities	SEQAT	Equity / Total Asset
LCTSALE	Current Liabilities / Sales	SIGMA	Stock Volatility
LT/(LT+MKET)	Total Liabilities / (Total Liabilities + Market Equity)	WCAPAT	Working Capital / Total Assets
LTAT	Total Liabilities / Total Assets		

Table 2: Discrete Hazard Model Estimations

Note: Panel A reports the parameter estimation results of the discrete hazard model for the full sample spanning the 1980 to 2009 period unless otherwise indicated. Column "LASSO" is the LASSO-selected reduced-model. Column "CHS" is the CHS (2008) model. Column "LASSO (1980-2002)" is the LASSO-selected reduced-model for the subsample spanning the 1980 to 2002 period. Column "CHS (1980-2002)" is the CHS model (2008) for the subsample spanning the 1980 to 2002 period. The absolute z-statistics are reported in the parenthesis, and ** denotes significance at the 1% level. Panel B reports the in-sample AIC, AUC (the area under the ROC curve), and McFadden Pseudo-R^2 for the corresponding discrete hazard model estimations.

	LASSO	CHS	LASSO (1980-2002)	CHS (1980-2002)
Panel A: Parameter Estimations				
LCTAT	0. 5641 (3.30)**		0. 6557 (3.63)**	
FAT	0. 0013 (5.57)**		0. 0013 (5.41)**	
NIMTA	-1.0104 (5.74)**	-1.1949 (6.63)**	-1.1475 (6.20)**	-1.3940 (7.38)**
LTMTA	1.3582 (10.26)**	1.7785 (13.22)**	1.1910 (8.45)**	1.6707 (11.74)**
CASHMTA		-0.7096 (3.07) **		-0.9904 (3.81)**
RSIZE		-0.0939 (3.81)**		-0.1130 (4.24)**
PRICE	-0.5644 (17.05)**	-0.5330 (13.63)**	-0.5630 (16.53)**	-0.5142 (12.58)**
MB		0.0693 (3.92)**		0.0810 (4.48)**
SIGMA	0.5491 (7.92)**	0.5367 (7.76)**	0.4472 (6.22)**	0.4293 (5.99)**
EXCESS RETURN	-0.8803 (5.22)**	-0.8769 (5.18)**	-0.8320 (4.71)**	-0.8332 (4.69)**
INTERCEPT	-7.8232 (63.23)**	-8.8070 (26.74)**	-7.6472 (59.83)**	-8.8584 (25.03)**
Panel B: Goodness-of-Fit Statistics				
AIC	14683	14712	13035	13053
AUC	0.711	0.710	0.720	0.717
Pseudo-R^2	0.1060	0.1043	0.1026	0.1015

Table 3: Out-of-Sample Performance over the 2003 to 2009 Period

Note: The table reports the out-of-sample performance measures for the testing sample spanning the 2003 to 2009 period, including the accuracy ratio, AUC (area under the ROC curve), the Brier score, and the decile ranking. For the decile ranking, we sort firms in the testing sample equally into deciles based on their predicted default probabilities. The first decile (decile 1) contains firms with the highest predicted default probability, and the last five deciles (decile 6-10) include the firms with the lowest predicted default probability. We then tabulate the percentage of actual bankruptcy filings observed in each decile. Column "LASSO" refers to the LASSO-selected reduced-model. Column "CHS" refers to the CHS (2008) model. Column "DD" refers to the reduced-form model with the distance to default as the only predictive variable.

	LASSO	CHS	DD
Accuracy Ratio	0.682	0.636	0.648
AUC	0.841	0.818	0.824
Brier Score (10^{-3})	0.408	0.408	0.412
Percentage of Bankruptcy Filings			
1	59.62	58.65	55.77
2	19.23	7.69	20.19
3	5.77	12.5	7.69
4	5.77	7.69	3.84
5	0.96	5.77	1.92
6-10	8.65	7.69	10.57

Table 4: Out-of-Sample Performance for Different Testing Samples

Note: The table reports the out-of-sample performance measures for two different testing samples (2001 to 2009 in panel A and 2006 to 2009 in panel B), including the accuracy ratio, AUC (area under the ROC curve), the Brier score, and the decile ranking. For the decile ranking, we sort firms in the testing sample equally into deciles based on their predicted default probabilities. The first decile (decile 1) contains firms with the highest predicted default probability, and the last five deciles (decile 6-10) include the firms with the lowest predicted default probability. We then tabulate the percentage of actual bankruptcy filings observed in each decile. Column "LASSO" refers to the LASSO-selected reduced-model. Column "CHS" refers to the CHS (2008) model. Column "DD" refers to the reduced-form model with the distance to default as the only predictive variable.

	LASSO	CHS	DD
Panel A: Testing Period from 2001 to 2009			
Accuracy Ratio	0.686	0.642	0.668
AUC	0.843	0.821	0.834
Brier Score (10^{-3})	0.517	0.517	0.523
Percentage of Bankruptcy Filings			
1	57.53	59.68	56.99
2	19.89	10.75	15.05
3	8.60	4.30	10.22
4	2.15	10.21	6.99
5	2.69	5.38	2.15
6-10	9.14	9.68	8.60
Panel B: Testing Period from 2006 to 2009			
Accuracy Ratio	0.696	0.634	0.694
AUC	0.848	0.817	0.847
Brier Score (10^{-3})	0.585	0.584	0.586
Percentage of Bankruptcy Filings			
1	63.89	58.33	59.72
2	18.06	5.56	19.44
3	2.78	15.28	6.94
4	4.17	6.94	4.17
5	1.39	5.56	2.78
6-10	9.72	8.33	6.94

Table 5: Discrete Hazard Model Estimations

Note: The table reports the variable selection results for different prediction horizons over the full sample spanning the 1980 to 2009 period. Variables in grey shades are market variables; the others are financial ratios constructed using only accounting information. First six columns are the LASSO-selected reduced-form models for 1-month, 6-month, 12-month, 24-month, 36-month, and 60-month-ahead prediction, respectively. Column "CHS" is the CHS (2008) model. For each LASSO-selected forecast model, we highlight the selected variables by the letter "X".

	1 Month	6 Month	12 Month	24 Month	36 Month	60 Month	CHS
Panel A: LASSO Variable Selection Result							
FAT	X	X	X	X	X	X	
LCTAT	X	X	X	X	X	X	
LCTSALE						X	
LOG(AT)					X	X	
OIADPAT					X	X	
SEQAT					X		
LTMTA	X	X	X	X		X	X
NIMTA	X	X	X	X			X
CASHMTA							X
RSIZE							X
PRICE	X	X	X	X	X	X	X
MB							X
SIGMA	X	X	X	X	X		X
EXCESS RETURN	X	X	X	X			X

Table 6: Goodness-of-Fit Statistics

Note: The table reports the in-sample and out-of-sample performance for both the LASSO-selected reduced-form model and the CHS (2008) model over different prediction horizons. The in-sample performance measures, including AIC, AUC (area under the ROC curve), are calculated using the training data over the 1980 to 2002 period. The out-of-sample performance measures, including the accuracy ratio and AUC, are calculated using the testing data over the 2003 to 2009 period.

	In-Sample		Out-of-Sample	
	AIC	AUC	Accuracy Ratio	AUC
Panel A: 1 Month Ahead Prediction				
LASSO	15461	0.878	0.840	0.920
CHS	15535	0.876	0.840	0.920
Panel B: 6 Month Ahead Prediction				
LASSO	17485	0.796	0.766	0.883
CHS	17523	0.796	0.756	0.878
Panel C: 12 Month Ahead Prediction				
LASSO	13035	0.720	0.682	0.841
CHS	13053	0.717	0.636	0.818
Panel D: 24 Month Ahead Prediction				
LASSO	16487	0.601	0.454	0.727
CHS	16520	0.605	0.386	0.693
Panel E: 36 Month Ahead Prediction				
LASSO	14060	0.555	0.350	0.675
CHS	14128	0.550	0.290	0.645
Panel F: 60 Month Ahead Prediction				
LASSO	10362	0.530	0.184	0.592
CHS	10462	0.517	0.176	0.588

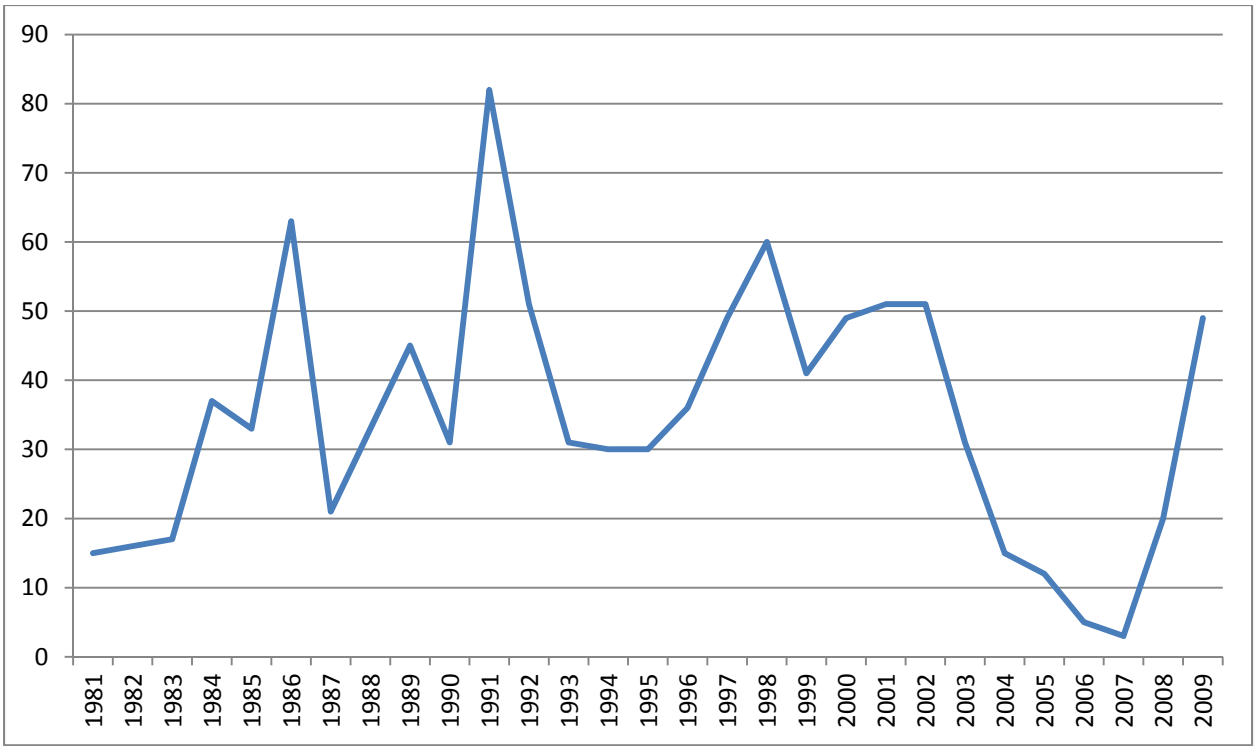

Figure 1: Number of Corporate Bankruptcy Filings in Each Year: 1980 to 2009

Note: The figure plots the number of the firms that filed for bankruptcy in each year over the 1980 to 2009 period. We define bankruptcy as filling under either Chapter 7 or Chapter 11 bankruptcy protection code.

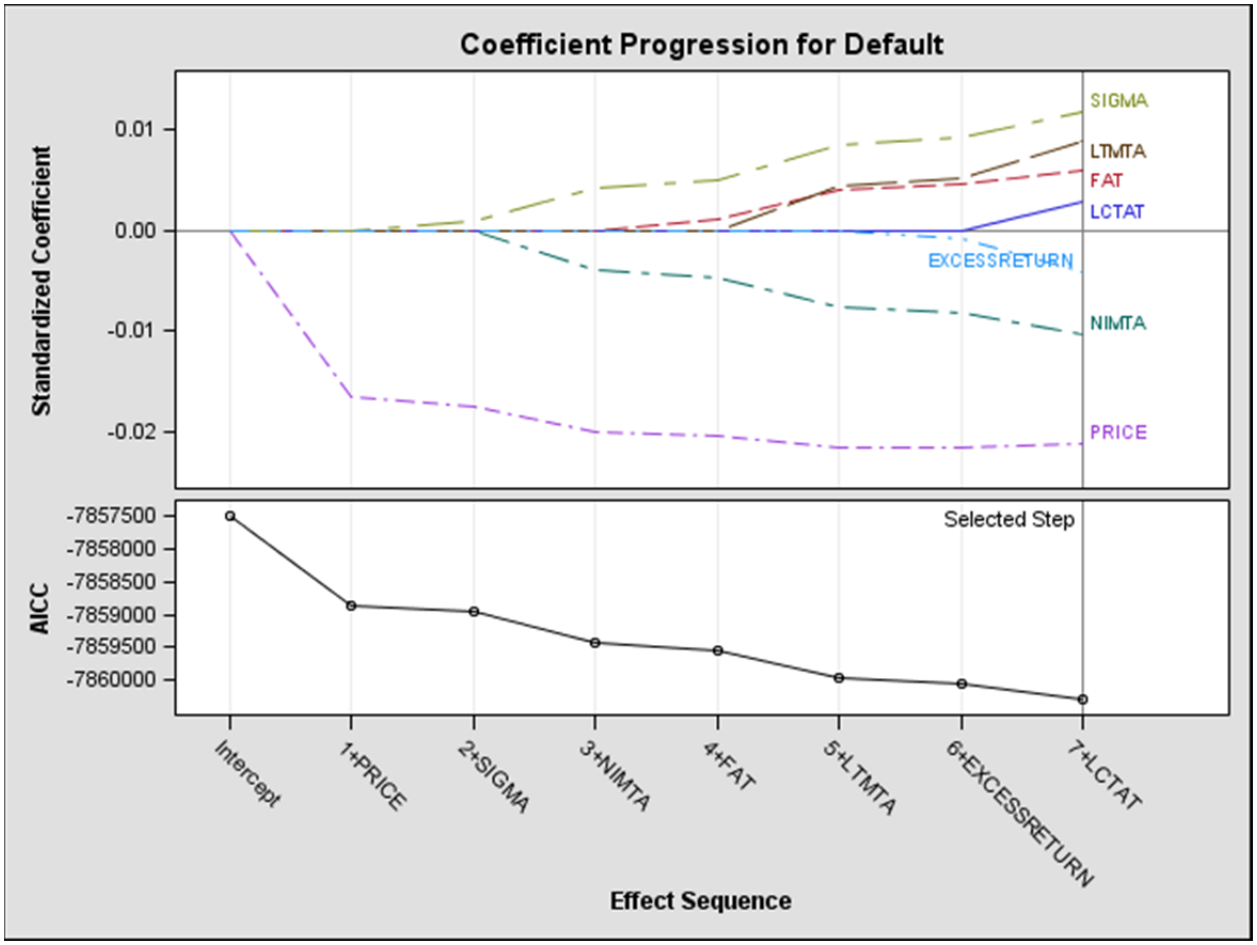

Figure 2: Coefficient Path using LASSO Variable Selection with 39 Explanatory Variables
Note: The figure plots the coefficient path of the LASSO selected predictive variables from a set of 39 candidate predictors over the 1980 to 2009 period. The upper panel illustrates the evolution of the estimated coefficients on all selected candidate predictive variables. As some LASSO parameter estimates increase in magnitude and become nonzero, explanatory variables sequentially enter into the bankruptcy forecast model. The figure shows that LASSO first selects PRICE, followed by SIGMA, NIMTA, FAT, LTMTA, EXCESSRETURN, and LCTAT. The lower panel illustrates the corresponding evolution of the model's AICC. AICC is a goodness-of-fit measurement with regard to information loss, similar to AIC.