A Deep Structural Model for Empirical Asset Pricing

Kristoffer Halskov[*]

November 14, 2023

Abstract

This paper proposes a new type of modelling framework that use machine learning techniques to estimate the parameters of structural models: Deep Structural Models (DSMs). I implement a DSM with a simple Merton (1974) model as a foundation, and show that the DSM *jointly* estimates expected equity returns and (co)variances with higher predictive power than leading benchmark models. I form long-short and mean variance efficient portfolios with significantly higher average excess returns, alphas, and Sharpe ratios, compared to those formed on the basis of a state-of-the-art machine learning model. Economically, the DSM suggests that systematic risk compensation is the largest contributor to the average expected equity return of firms, while mispricing is the primary driver of the dispersion of expected returns. Finally, the DSM provides evidence that firm leverage is the main reason for an increased equity premium during economic recessions.

Keywords: Empirical Asset Pricing, Machine Learning, Factor Models, Return Prediction, Structural Models of Credit Risk

JEL: C10, C45, G10, G11

[*] Copenhagen Business School; www.khalskov.com. I am grateful for the helpful comments and feedback from Alejandro Lopez-Lira (discussant), Amit Goyal, David Lando, Grace Xing Hu, Lasse Heje Pedersen, Peter Feldhütter, Stephen Schaefer, Tarun Ramadorai, Giovanni Venturato, members of the BIGFI center at Copenhagen Business School, and seminar and conference participants at the European FMA Doctoral Student Consortium 2023, the HEC Workshop on Incentives in Finance 2023, and the FMA Annual Meeting 2023.

1 Introduction

Over the past decade, the integration of machine learning (ML) models into financial research has led to significant advances. Despite these advances, ML models suffer from a lack of interpretability and theoretical foundation, limiting financial insight.[1] Structural models, in contrast to ML models, are inherently theory-driven. Despite the broad range of methodologies that "structural models" encompasses, they share a core characteristic: they are born from theory and are designed to offer explicit predictions and insights into the phenomena they represent. This does not come without a cost though: while structural models provide clear economic insights, it is not clear how to best estimate their parameters, let alone how to incorporate the vast amount of conditioning information available to econometricians. This apparent dichotomy between ML models and structural models naturally begs the question: can we combine the two and keep the flexibility and predictive power of ML, and the economic intuition and interpretability of structural models? As Giglio, Kelly, and Xiu (2022) write: *"...our view is that the most promising direction for future empirical asset pricing research is developing a genuine fusion of economic theory and machine learning. It is a natural marriage..."*. This paper is an attempt to officiate such a wedding. I propose a new model framework, Deep Structural Models (DSMs), that combine ML techniques with structural models.

Figure 1 illustrates how the DSM framework works by showcasing three different approaches for modelling an object of interest, $\mathbf{y}$: a deep learning model, a structural model, and a DSM.[2] When fitting the deep learning model to the data, we start from a point of observing a set of variables, denoted by $\mathbf{X}_{obs}$ in the figure, that has some functional relationship with $\mathbf{y}$. We then rely on the deep learning model, represented by the black-box, to find the functional relationship. In contrast, the structural approach relies on a set of estimated parameters, denoted by $\hat{\boldsymbol{\theta}}$, that feeds through an economic model to arrive at an estimate of $\mathbf{y}$. The estimation of $\hat{\boldsymbol{\theta}}$ depends on the specific choice of structural model, but usually relies

[1] Machine learning has been successful in various predictive tasks within finance, yet it faces unique challenges in this domain. Israel, Kelly, and Moskowitz (2020) provide an insightful discussion on these issues.

[2] Deep learning models are a specific type of ML models that are ideally suited for the purposes of this paper due to their flexibility in terms of customization and optimization. ML and deep learning will be used interchangeably even though one is a subset of the other.

on some sort of GMM, maximum likelihood, or simulation estimation. The DSM framework combines the two approaches and models the parameters of a structural model as functions of the observable data, i.e. $\mathbf{y}$ is now only indirectly a function of $\mathbf{X}_{obs}$ through $\hat{\boldsymbol{\theta}}$. This allows us to keep the ability of ML models to flexibly incorporate all observable information in the estimation of $\mathbf{y}$, while keeping the transparency of structural models. The DSM framework can therefore be viewed as a flexible methodology for estimating the parameters of structural models or, alternatively, as an economically motivated regularization of an ML model.

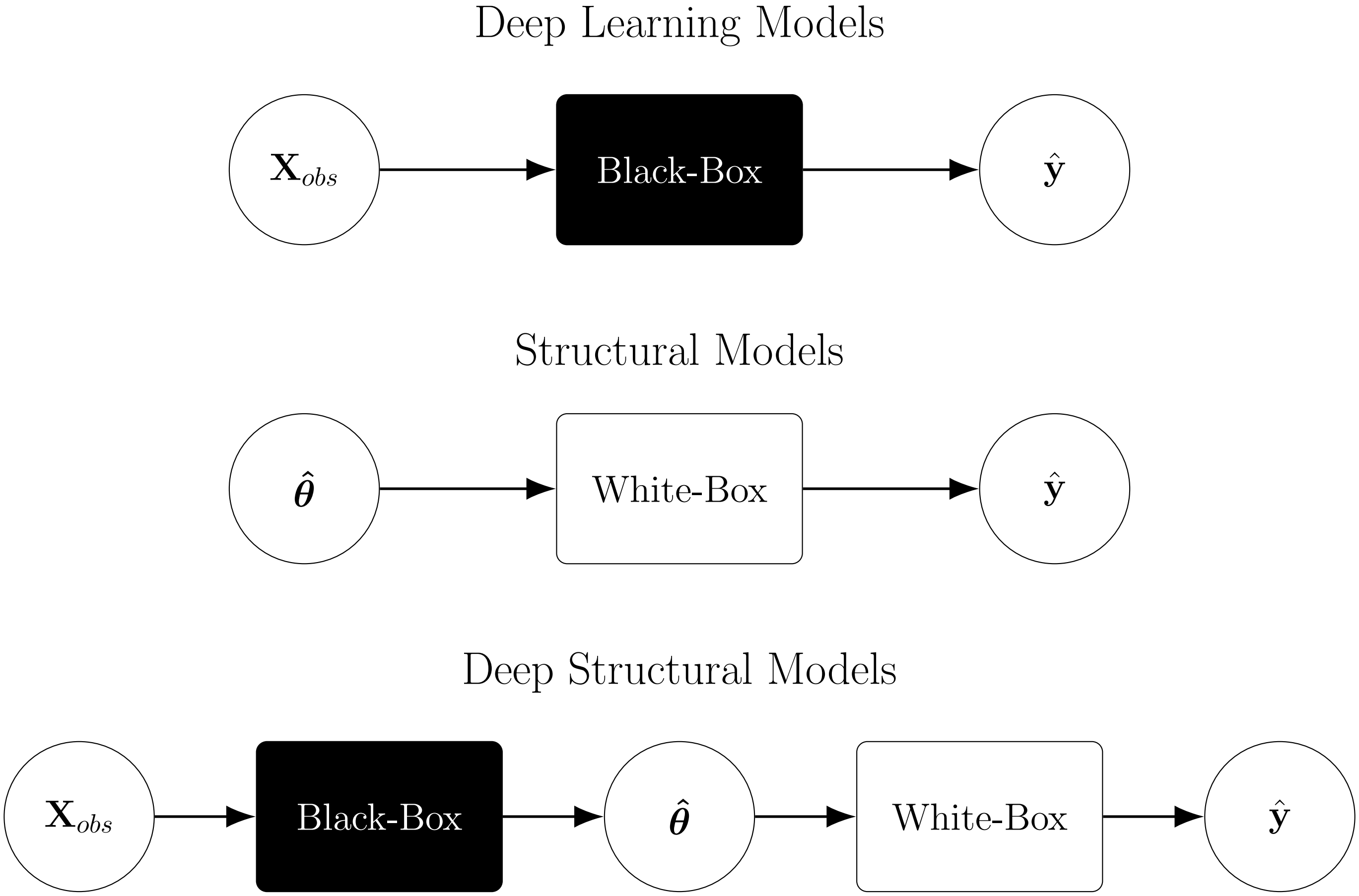


Figure 1: *Three Different Modelling Approaches.* This figures illustrates the modelling approach for three different types of models: deep learning models, structural models, and deep structural models (DSMs). For all three types of models, the object or phenomenon of interest is represented by $\mathbf{y}$. $\mathbf{X}_{obs}$ refers to a set of observable variables, while $\hat{\boldsymbol{\theta}}$ refers to a set of estimated parameters for some structural model. The "Black-Box" represents a deep learning model that transforms $\mathbf{X}_{obs}$ into an output, while the "White-Box" refers to the structurally determined transformation of $\hat{\boldsymbol{\theta}}$ to an estimate of $\mathbf{y}$.

The specific structural model examined in this paper is a modified version of the classic

Merton (1974) model wherein the assets of the firm follow a geometric brownian motion (GBM). The asset drift is the sum of the risk-free rate, a term representing mispricing, and systematic risk compensation, while asset volatility contains a systematic and idiosyncratic component. This model jointly estimates the conditional expected equity returns and (co)variances and enables the analysis of the importance of mispricing relative to systematic risk compensation, as well as the effect of firm leverage on expected equity returns. The GBM parameters are modelled as functions of 238 firm-specific characteristics and 45 macroeconomic variables, and the model is estimated on a comprehensive dataset of equity returns spanning 1950-2021 with around 3.2 million firm-month observations containing 23,422 unique firms. After fitting the model, I use analytically derived expressions for the expected equity returns and (co)variances and find the following key results:

1. **The Role of Mispricing.** Systematic risk compensation is the largest contributor to the average expected excess asset return, while mispricing is responsible for most of the dispersion. Systematic risk compensation contributes 63.73% to the average excess asset return, while mispricing only contributes 36.27%. The standard deviations of the systematic risk compensation and mispricing parameters are 6.85% and 12.51%, respectively.

2. **The Role of Leverage.** Firm leverage, rather than the underlying asset dynamics, is responsible for an increased equity premium during recessions. The estimated asset parameters are stable through recessions, yet the time series dynamics of leverage cause the equity premium to increase. I find that the equity premium peaked at 15% during the financial crisis of 2008-09.

3. **Equity Return Prediction.** The DSM provides more accurate firm-level estimates of the expected equity return than existing state-of-the-art ML models. Different specifications of the DSM provide out-of-sample R^2-values in the range of 0.74-0.80, compared to 0.56 for a neural network (NN) benchmark and -10.60 from a simple OLS model.

4. **Long-Short Portfolio Performance.** The more accurate equity return predictions from the DSM lead to better performing long-short portfolios. Portfolios based on the

DSM predictions outperform the NN benchmark in terms of both excess returns and annualized Sharpe ratios: the DSM portfolios have average monthly excess returns (Sharpe ratios) in the range of 2.63-3.10% (1.43-1.67) compared to the NN benchmark portfolio of 1.91% (1.08).

5. **Variance Forecasting.** The DSM estimates future firm-level equity return variances better than a GARCH(1,1) model. The DSM variance forecasts has an out-of-sample mean squared error that is 20.59%-21.27% lower than the GARCH benchmark. Regression results confirm that the DSM predictions explain a higher proportion of the variance with an R^2-value of 0.53 compared to 0.50 for the GARCH benchmark.

6. **Mean Variance Efficient Portfolios.** Mean variance efficient (MVE) portfolios, formed on the basis of the expected equity returns and covariance matrix, perform even better than the long-short portfolios. I form both a leverage constrained and a long-only MVE portfolio that are re-balanced on a monthly basis. The best performing leverage constrained MVE portfolio achieves a monthly average excess return of 4.89% with an annualized Sharpe ratio of 3.96, while the best performing long-only MVE portfolio has a monthly average excess return and Sharpe ratio of 4.58% and 1.93, respectively. For comparison, the S&P500 index delivered an average monthly excess return of 0.60% and a Sharpe ratio of 0.48 over the same time period.[3]

This paper touches upon several strands of the literature. While I implement a modified version of the Merton (1974) model, that was not the only possible choice. The literature on structural credit risk models has since the publication of Merton (1974) added additional economic mechanisms: Black and Cox (1976) introduce a default boundary and Leland (1994) accounts for bankruptcy costs and the tax benefits of debt. More recent advancements include Du, Elkamhi, and Ericsson (2019), who models the firm's asset volatility as stochastic, and Feldhütter and Schaefer (2023), who incorporate stochastic debt dynamics. Vassalou and Xing (2004) also implements the Merton (1974) model on a firm-level basis using an iterative estimation technique and uncover the so-called distress risk puzzle. Bharath and Shumway

[3] These numbers are gross of transaction fees and should not be viewed as achievable by an investor. The MVE portfolios serve as a testament to the DSM's ability to accurately model not only expected equity returns but also covariances.

(2008) use a much simpler firm-level estimation technique to show that it is the functional form of the Merton (1974) model, rather than the specific implementation of it, that matters when using it for default prediction. ML has, in part, gained popularity within finance for its ability to overcome the curse of dimensionality: a large number of factors and characteristics has been put forth in the literature for explaining the cross-section of equity returns, leading to the so-called "factor zoo" as Cochrane (2011) put it (see Harvey, Liu, and Zhu (2016), Hou, Xue, and Zhang (2020), and Jensen, Kelly, and Pedersen (2022) for an overview of the many factors proposed in the literature). Freyberger, Neuhierl, and Weber (2020), Feng, Giglio, and Xiu (2020), and Kozak, Nagel, and Santosh (2020) use various shrinkage methods on a large set of factors and characteristics to construct stochastic discount factors, while Bryzgalova, Pelger, and Zhu (2020) and Chen, Pelger, and Zhu (2023) extend this idea to non-linear ML techniques. The idea of modelling structural parameters as functions of contemporary observable variables is heavily inspired by Kelly, Pruitt, and Su (2019) and Gu, Kelly, and Xiu (2021) who use a multitude of characteristics to determine conditional betas for equity returns in a latent factor model. Bali, Goyal, Huang, Jiang, and Wen (2020) use the structural model of Du, Elkamhi, and Ericsson (2019) to motivate the use of hedge ratios for predicting bond returns. They use different statistical methods, including ML models, to estimate expected equity returns and hedge ratios, which they then use to predict bond returns. Their paper is a great example of how to use financial theory and ML techniques in conjunction with each other. While their methodology achieves significantly better bond return predictions than traditional models, they do not estimate the underlying parameters of their structural model, and so their predictions are still of a "black-box" nature. This paper differs since I directly estimate the underlying firm-level asset dynamics, which not only allows us to analyze these estimated parameters, but also enables us to use analytically derived predictions for firm-level expected equity returns and (co)variances.

The paper proceeds as follows: Section 2 introduces the structural model, its implications, and the empirical implementation using ML. Section 3 gives an overview of the data and analyzes the out-of-sample performance of the DSM, and finally, Section 4 concludes.

2 Model

2.1 A General Model

It is assumed that all systematic risk in the economy can be characterized by K independent Brownian motions B_{kt}, for $k = 1, ..., K$. The price of risk associated with each Brownian motion is time-varying and is denoted by λ_{kt}. In addition to the systematic risks, each firm, denoted by i, is also exposed towards an idiosyncratic risk represented by another independent Brownian motion, $\mathcal{E}_{it}$. The asset value for firm i at time t, V_{it}, is then assumed to follow a $(K+1)$-dimensional geometric Brownian motion:

$$dV_{it} = \left(r_{ft} - \delta_{it} + \alpha_{it} + \sum_{k=1}^{K} \beta_{kit} \lambda_{kt} \right) V_{it} dt + \sum_{k=1}^{K} \beta_{kit} V_{it} dB_{kt} + \epsilon_{it} V_{it} d\mathcal{E}_{it} \tag{1}$$

Where r_{ft} denotes the risk-free rate, δ_{it} is the firm-wide payout, α_{it} is compensation unassociated with any risk exposure and can be viewed as a mispricing or arbitrage term, β_{kit} is the risk exposure of firm i at time t towards the systematic Brownian motion B_{kt}, while ϵ_{it} is the risk exposure towards the idiosyncratic Brownian motion $\mathcal{E}_{it}$. For convenience, (1) can also be written in matrix form:

$$dV_{it} = \left(r_{ft} - \delta_{it} + \alpha_{it} + \boldsymbol{\beta}_{it}^T \boldsymbol{\lambda}_t \right) V_{it} dt + \boldsymbol{\beta}_{it}^T V_{it} d\boldsymbol{B}_t + \epsilon_{it} V_{it} d\mathcal{E}_{it} \tag{2}$$

Where $\boldsymbol{\beta}_{it}$, $\boldsymbol{\lambda}_t$, and $\boldsymbol{B}_t$ are now all $K \times 1$ vectors containing the systematic risk exposures, prices of risk, and systematic shocks, respectively. The stochastic process presented in (2) is standard in the asset pricing literature, although it is commonly expressed more succinctly as:

$$dV_{it} = \mu_{it} V_{it} dt + \sigma_{it} V_{it} dW_{it} \tag{3}$$

Which, given the representation in (2), means that:

$$\mu_{it} = r_{ft} - \delta_{it} + \alpha_{it} + \boldsymbol{\beta}_{it}^T \boldsymbol{\lambda}_t \tag{4}$$

$$\sigma_{it} = \sqrt{\boldsymbol{\beta}_{it}^T \boldsymbol{\beta}_{it} + \epsilon_{it}^2} \tag{5}$$

$$dW_{it} = \frac{1}{\sqrt{\boldsymbol{\beta}_{it}^T \boldsymbol{\beta}_{it} + \epsilon_{it}^2}} \left(\boldsymbol{\beta}_{it}^T d\boldsymbol{B}_t + \epsilon_{it} d\mathcal{E}_{it} \right) \tag{6}$$

Despite the subscript on the single Brownian motion, W_{it}, it is important to note that firm i is still exposed towards the K systematic risk factors.

2.2 Simplifying Assumptions

Now, assume that the contingent claims to the firm's assets are defined as in Merton (1974). That is, at time t, each firm has two contingent claims to its assets: A single class of debt with a market value of D_{it} that promises a single cash flow at time $t+1$ equal to F_{it+1}, and equity, E_{it}, which is the residual claim to the firm's assets. If, at time $t+1$, we have that $V_{it+1} \geq F_{it+1}$, then bondholders collectively receive F_{it+1} and equity holders receive $V_{it+1} - F_{it+1}$. On the other hand, if $V_{it+1} < F_{it+1}$, then bondholders receive V_{it+1} and equity holders receive nothing. Thus, the terminal values of debt and equity can be written as $D_{it+1} = \min[F_{it+1}, V_{it+1}]$ and $E_{it+1} = \max[V_{it+1} - F_{it+1}, 0]$. Additionally, the firm is restricted from issuing new debt, paying dividends, or buying back shares before time $t+1$. This means that in the context of the general model in Section 2.1 we have that $\delta_{it} = 0, \quad \forall t < t+1$.

Let $\boldsymbol{X}_{it}$ be a $N \times 1$ vector containing a set of observable firm characteristics for firm i at time t. Then, let $\boldsymbol{Y}_t$ be a $M \times 1$ vector of observable macroeconomic variables shared among all firms at time t. Both $\boldsymbol{X}_{it}$ and $\boldsymbol{Y}_t$ are assumed constant between t and $t+1$. For each systematic shock, k, it is assumed that there exists two sets of functions that map the observable variables into firm risk exposures and market prices of risks, respectively. Each risk exposure function, denoted β_k, transforms the firm characteristics into a single value, $\beta_k : \mathbb{R}^N \to \mathbb{R}$, while each market price of risk function, denoted λ_k, depend on the macroeconomic variables, $\lambda_k : \mathbb{R}^M \to \mathbb{R}$. Like the β_k-functions, both the mispricing term, α_{it}, and the idiosyncratic risk exposure, ϵ_{it}, are functions that solely depend on the firm characteristics, $\alpha : \mathbb{R}^N \to \mathbb{R}$ and $\epsilon : \mathbb{R}^N \to \mathbb{R}$. Finally, the risk-free rate r_{ft} is assumed to be constant between between t and $t+1$.

With these simplifying assumptions, we can rewrite the asset process of firm i from (3):

$$dV_{it} = \mu(\boldsymbol{X}_{it}, \boldsymbol{Y}_t) V_{it} dt + \sigma(\boldsymbol{X}_{it}) V_{it} dW_{it}, \quad \forall t < t+1 \tag{7}$$

Where the “transformation” equations (4)-(6) become:

$$\mu(\boldsymbol{X}_{it}, \boldsymbol{Y}_t) = r_{ft} + \alpha(\boldsymbol{X}_{it}) + \boldsymbol{\beta}(\boldsymbol{X}_{it})^T \boldsymbol{\lambda}(\boldsymbol{Y}_t) \tag{8}$$

$$\sigma(\boldsymbol{X}_{it}) = \sqrt{\boldsymbol{\beta}(\boldsymbol{X}_{it})^T \boldsymbol{\beta}(\boldsymbol{X}_{it}) + \epsilon(\boldsymbol{X}_{it})^2} \tag{9}$$

$$dW_{it} = \frac{1}{\sqrt{\boldsymbol{\beta}(\boldsymbol{X}_{it})^T \boldsymbol{\beta}(\boldsymbol{X}_{it}) + \epsilon(\boldsymbol{X}_{it})^2}} \left(\boldsymbol{\beta}(\boldsymbol{X}_{it})^T d\boldsymbol{B}_t + \epsilon(\boldsymbol{X}_{it}) d\mathcal{E}_{it}\right) \tag{10}$$

Since it is clear from (7)-(10) which functions depend on $\boldsymbol{X}_{it}$, $\boldsymbol{Y}_t$, or both, the observable variables are omitted in the notation for the sake of simplicity, and a subscript of i is added to functions that depend on $\boldsymbol{X}_{it}$. Similarly, a subscript of t is added to the parameter functions, but it is important to note that this indicates the parameters are time-varying because of time-varying function inputs and not because the parameter functions themselves are time-varying, i.e. $\boldsymbol{\beta}(\cdot)$ does not change but its input variables, $\boldsymbol{X}_{it}$, varies across firms and time.

2.3 Model Implications

This section describes the analytical properties of the model in Section 2.2. Some of these properties, such as the implied default probability, are well-known in the literature, while others, such as the expected equity return and (co)variance, are not.

2.3.1 Default Implications

Let $\mathbb{1}_{V_{it+1}<F_{it+1}}$ be an indicator variable equal to one if firm i defaults at time $t+1$. The probability of this event is the implied default probability of Merton (1974):

$$\begin{aligned} \pi_{it} &= \mathrm{E}[\mathbb{1}_{V_{it+1}<F_{it+1}}] \\ &= \Phi(-DD_{it}) \end{aligned} \tag{11}$$

Where $\Phi(\cdot)$ is the cumulative standard normal distribution and DD_{it} is the distance to default:

$$DD_{it} = \frac{\ln\left(\frac{V_{it}}{F_{it+1}}\right) + \mu_{it} - \frac{\sigma_{it}^2}{2}}{\sigma_{it}} \tag{12}$$

2.3.2 Equity Implications

In a setting such as this, we know that the equity of the firm can be viewed as a European call option on the underlying firm assets, i.e. the current equity value can be expressed as:

$$E_{it} = V_{it}\Phi(d_{1it}) - F_{it+1}\exp\{-r_{ft}\}\Phi(d_{2it}) \tag{13}$$

Where:

$$d_{1it} = \frac{\ln\left(\frac{V_{it}}{F_{it+1}}\right) + r_{ft} + \frac{\sigma_{it}^2}{2}}{\sigma_{it}} \tag{14}$$

$$d_{2it} = d_{1it} - \sigma_{it} \tag{15}$$

At time $t+1$, the equity value of a firm is equal to the asset value of the firm minus the face value of debt bounded below at 0. When the asset value of a firm follows a Geometric Brownian motion we know that the terminal (or in this case, the time $t+1$) asset value is log-normally distributed, which means that we can view the time $t+1$ equity value as a mixture distribution of a constant 0 and a shifted log-normal distribution truncated at 0:

$$\mathcal{L}^E_{it+1} = \pi_{it}\delta(E_{it+1}) + \frac{1}{(E_{it+1}+F_{it+1})\sqrt{2\pi\sigma^2_{it}}}\exp\left\{-\frac{\left(\ln\left(\frac{E_{it+1}+F_{it+1}}{V_{it}}\right)-\left(\mu_{it}-\frac{\sigma^2_{it}}{2}\right)\right)^2}{2\sigma^2_{it}}\right\}U(E_{it+1}) \tag{16}$$

Where $\delta(\cdot)$ and $U(\cdot)$ are the Dirac delta and the Heaviside step functions, respectively.[4] From an empirical standpoint, it is more convenient to work with the density function for the equity return:

$$\mathcal{L}^r_{it+1} = \pi_{it}\delta(1+r_{it+1}) + \frac{1}{\left(1+r_{it+1}+\frac{F_{it+1}}{E_{it}}\right)\sqrt{2\pi\sigma^2_{it}}}\exp\left\{-\frac{\left(\ln\left(\frac{(1+r_{it+1})E_{it}+F_{it+1}}{V_{it}}\right)-\left(\mu_{it}-\frac{\sigma^2_{it}}{2}\right)\right)^2}{2\sigma^2_{it}}\right\}U(1+r_{it+1}) \tag{17}$$

Equation (17) is employed to fit the model to the data, however, to conduct an out-of-sample analysis of equity return predictions, variance predictions, and portfolio optimization, we need expressions for the expected equity returns and (co)variances. These are shown in Appendix A.1 and A.2, respectively, to be:

$$\mathrm{E}[r_{it+1}] = \frac{V_{it}}{E_{it}}\exp\{\mu_{it}\}\Phi\left(DD_{it}+\sigma_{it}\right) - (1-\pi_{it})\frac{F_{it+1}}{E_{it}} - 1 \tag{18}$$

[4] Be aware of the slightly confusing notation in (16) and (17): π_{it} is the default probability of firm i, whereas π without a subscript refers to the mathematical constant.

And:

$$
\begin{aligned}
\text{Cov}[r_{it+1}, r_{jt+1}] = {} & (1 - \pi_{it} - \pi_{jt} + \text{Cov}[\mathbb{1}_{V_{it+1}<F_{it+1}}, \mathbb{1}_{V_{jt+1}<F_{jt+1}}] + \pi_{it}\pi_{jt}) \\
& \times \frac{1}{E_{it}E_{jt}} \text{E}\big[E_{it+1}E_{jt+1} | \min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{jt+1}>F_{jt+1}}] = 1\big] \\
& - (1 + \text{E}[r_{it+1}])(1 + \text{E}[r_{jt+1}])
\end{aligned} \tag{19}
$$

In the case where $i = j$, i.e. the variance, equation (19) has the closed-form solution:

$$
\begin{aligned}
\text{Var}[r_{it+1}] = \frac{1}{E_{it}^2} \bigg(& \text{E}[V_{it+1}^2]\Phi\left(DD_{it} + 2\sigma_{it}\right) - V_{it}^2 \exp\{2\mu_{it}\}\Phi\left(DD_{it} + \sigma_{it}\right)^2 \\
& - 2V_{it}\exp\{\mu_{it}\}\Phi\left(DD_{it} + \sigma_{it}\right)\pi_{it}F_{it+1} + (1 - \pi_{it})\pi_{it}F_{it+1}^2 \bigg)
\end{aligned} \tag{20}
$$

Where:

$$
\text{E}[V_{it+1}^2] = V_{it}^2 \exp\left\{2\mu_{it} + \sigma_{it}^2\right\} \tag{21}
$$

To avoid using the computationally expensive procedure of numerically estimating the covariance matrix through (19), I use asset return correlations as a proxy for equity return correlations. Specifically, the asset return correlation between firm i and j, at time t, can be analytically calculated as:

$$
\rho_{ijt}^V = \frac{\boldsymbol{\beta}_{it}^T\boldsymbol{\beta}_{jt} + \mathbb{1}_{i=j}\epsilon_{it}\epsilon_{jt}}{\sigma_{it}\sigma_{jt}} \tag{22}
$$

Then, using ρ_{ijt}^V as a proxy for the equity return correlation, the equity return covariance between firm i and j is estimated as:

$$
\text{Cov}[r_{it+1}, r_{jt+1}] = \rho_{ijt}^V \sqrt{\text{Var}[r_{it+1}]\text{Var}[r_{jt+1}]} \tag{23}
$$

2.4 Empirical Implementation

While the theoretical framework assumes knowledge of the model parameters, this is not the case in practice. In fact, the only value we can reasonably assume to be observable is E_{it}, which is calculated as the total number of shares outstanding, S_{it}, times the price of each share, P_{it}^S:

$$
E_{it} = S_{it}P_{it}^S \tag{24}
$$

Since each firm has a lot of different debt instruments in practice, it is not clear how we should measure the debt value of the theoretical framework, D_{it}, or the face value F_{it+1}. I follow the convention of previous literature, such as Vassalou and Xing (2004) and Bharath and Shumway (2008), and estimate the face value of the debt as short-term debt, F_{it+1}^{SD}, plus half of long-term debt, F_{it+1}^{LD}:[5]

$$F_{it+1} = F_{it+1}^{SD} + 0.5F_{it+1}^{LD} \tag{25}$$

The market value of debt is then estimated by discounting the face value with the risk-free rate:[6]

$$D_{it} = F_{it+1}\exp\{-r_{ft}\} \tag{26}$$

To find the total firm value add (24) and (26):

$$V_{it} = E_{it} + D_{it} \tag{27}$$

There are no theoretical constraints on the mispricing function, α, the risk-exposure functions, β_k, the prices of risk functions, λ_k, or the idiosyncratic asset volatility function, ϵ. This is where the "Deep" part of the "Deep Structural Model" comes into play: in order to keep the overall model as flexible as possible, the α, β_k, λ_k, and ϵ functions are all modelled as neural networks. Specifically, they will be structured as standard feed-forward neural networks with 1 hidden layer.[7] An illustrative example of the parameter functions, can be seen in Figure 2. From the figure, we see that each function takes its input variables (either $\boldsymbol{X}_{it}$ or $\boldsymbol{Y}_t$) and passes it to a set of hidden nodes. At each hidden node, some linear combination of the input variables is passed through an activation function. The machine learning literature proposes a wide range of activation functions, but a particular common one is the rectified linear unit (ReLU) function, which will be utilized by all hidden nodes across the parameter functions:

$$ReLU(x) = \max[x, 0] \tag{28}$$

[5] Short-term debt and long-term debt is defined as "Debt in Current Liabilities - Total" and "Long-Term Debt - Total", respectively, from the Compustat database.

[6] This is obviously a major simplification as there should be some yield spread added to the discounting. However, due to the lack of good firm-level yield spread proxies that covers the entire data set described in Section 3.1, I follow a similar simplified debt value estimation as Bharath and Shumway (2008).

[7] It is possible to add more hidden layers to the functions, however, the results shown in Section 3 generally deteriorate when more hidden layers are added.

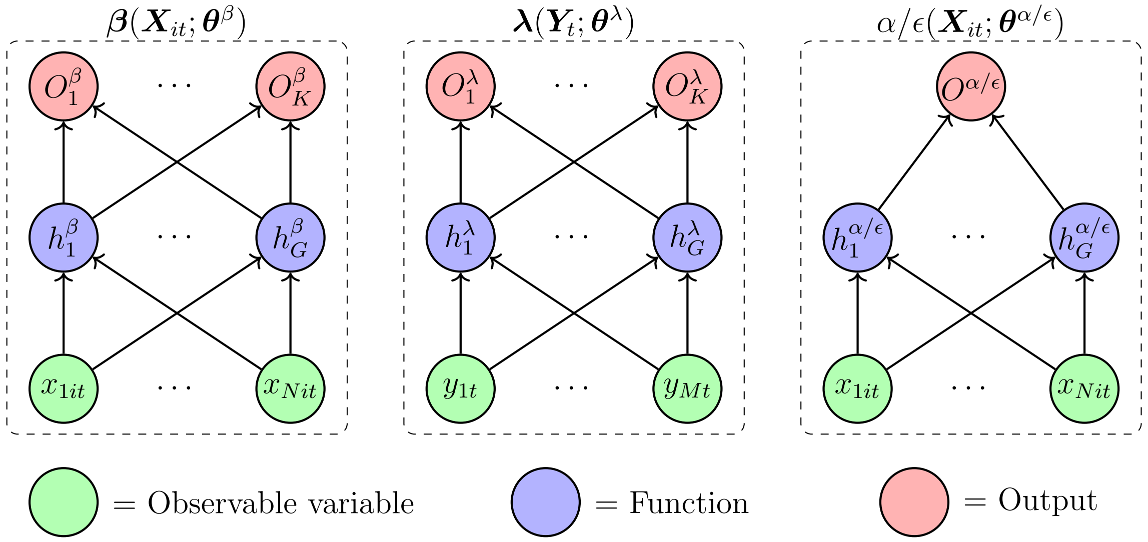


Figure 2: *The Parameter Functions.* This figure shows the general structure of the parameter functions of the asset value process, where α/ϵ indicates that α and ϵ has the same function structure. Each function takes a vector of inputs ($\boldsymbol{X}_{it}$ for α, $\boldsymbol{\beta}$ and ϵ, and $\boldsymbol{Y}_t$ for $\boldsymbol{\lambda}$) and passes them along to G hidden nodes. Each of the hidden nodes transform a linear combination of its inputs to a single positive real number through the ReLU function ($ReLU(x) = \max[x, 0]$). Finally, the outputs from each hidden node are passed to a number of output nodes (K output nodes for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and 1 for α and ϵ) each of which outputs some linear combination of its inputs, i.e. the activation function of all the output nodes is the identity function, $I(x) = x$.

The hidden nodes in Figure 2 will therefore have the following functional forms:

$$h_{git}^{\alpha/\beta/\epsilon} = \max\left[a_g^{\alpha/\beta/\epsilon} + \sum_{n=1}^{N} b_{gn}^{\alpha/\beta/\epsilon} x_{nit}, 0\right] \tag{29}$$

$$h_{gt}^{\lambda} = \max\left[a_g^{\lambda} + \sum_{m=1}^{M} b_{gm}^{\lambda} y_{mt}, 0\right] \tag{30}$$

Where $\alpha/\beta/\epsilon$ indicates that the function structure is the same for α, β, and ϵ, while $g \in 1, ..., G$ with G being the number of hidden nodes.[8] All outputs from each of these hidden nodes then feed into a set of output nodes, that, similarly to the hidden nodes, transforms the linear combination of its inputs into a single real number. All output nodes utilize the identity function, $I(x) = x$, which means their functional forms are:

$$O_{it}^{\alpha/\epsilon} = a^{\alpha/\epsilon} + \sum_{g=1}^{G} b_g^{\alpha/\epsilon} h_{git}^{\alpha/\epsilon} \tag{31}$$

[8] The choice of G is arbitrary and could be treated as a hyperparameter, however, for simplicity this paper use $G = 32$.

$$O^{\beta}_{kit} = a^{\beta}_{k} + \sum_{g=1}^{G} b^{\beta}_{gk} h^{\beta}_{git} \tag{32}$$

$$O^{\lambda}_{kt} = a^{\lambda}_{k} + \sum_{g=1}^{G} b^{\lambda}_{gk} h^{\lambda}_{gt} \tag{33}$$

All parameters associated with the α, β_k, λ_k, and ϵ functions in (29)-(33) are denoted $\boldsymbol{\theta}^{\alpha}$, $\boldsymbol{\theta}^{\beta}$, $\boldsymbol{\theta}^{\lambda}$, and $\boldsymbol{\theta}^{\epsilon}$, respectively.

Putting it all together, the complete Deep Structural Model of this paper can be viewed as a neural network, with an architecture imposed by the structural model of Section 2. A full general model illustration can be seen in Figure 3. The specific loss function used to train the model, along with the training procedure itself, can be found in Appendix B.

3 Empirical Results

3.1 Data

The primary data source for the empirical analysis is the combined US Compustat and CRSP data from Jensen, Kelly, and Pedersen (2022), which has been provided by the authors. The data set includes 4,135,225 firm-month observations from 1925-2021. I exclude firm-month observations where one or more of the following variables is missing: (company wide) market equity value, short-term debt, long-term debt, or one-month ahead equity return. This removes all firm-months in the early part of the sample, leaving a total of 3,197,609 firm-month observations from 1950-2021 for the actual empirical analysis. For the firm specific characteristics, I use the 153 variables used as the basis for the 153 factors explored in Jensen, Kelly, and Pedersen (2022). Additionally, a set of industry dummies based on the first two digits of a firm's SIC code (including a missing SIC dummy) are added as firm characteristics. This means that $\boldsymbol{X}_{it} \in \mathbb{R}^{238}$. For the macroeconomic variables I use the 14 variables in Welch and Goyal (2008) that covers the full time-period 1950-2021, alongside with the monthly S&P500 return.[9] I augment these 15 macroeconomic variables by taking the quarterly and yearly changes,[10] so that $\boldsymbol{Y}_t \in \mathbb{R}^{45}$. The macroeconomic variables are

[9] Specifically, b/m, d/e, d/p, d/y, dfr, dfy, e/p, infl, ltr, lty, ntis, sp500ret, svar, tbl, tms.

[10] For sp500ret and dfr, the quarterly and yearly returns are used instead.

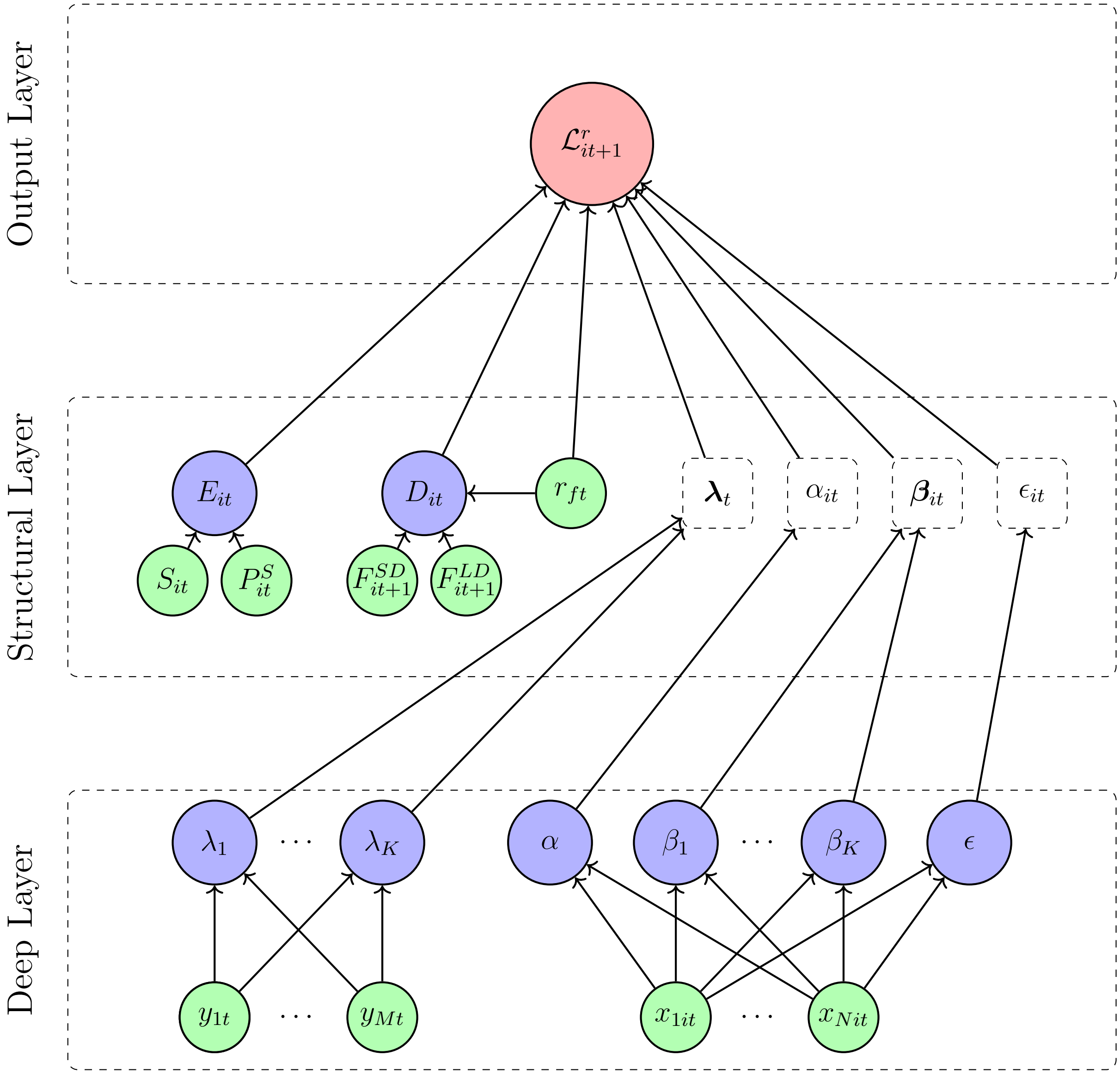


Figure 3: *Deep Structural Model Architecture.* This figure shows the full architecture of the DSM of this paper. At the bottom we have the "Deep Layer" where all observable inputs, $\boldsymbol{X}_{it}$ and $\boldsymbol{Y}_t$, feed into the α function, the risk exposure functions, $\beta_1, ..., \beta_K$, market price of risk functions, $\lambda_1, ..., \lambda_K$, and the idiosyncratic asset volatility function, ϵ. In the middle we have the "Structural Layer" containing all the structural parameters. Finally, at the top, we have the "Output Layer". This layer is essentially all the implications associated with the structural model, however, the output shown in this figure is limited to the equity return likelihood function as that is the only object used for training the model.

all extracted from Amit Goyal's personal website, from which an estimate of the monthly risk-free interest rate, r_{ft}, is also extracted. The data is then split into a training set, $\mathcal{T}_1$ (1950-1974), a validation set, $\mathcal{T}_2$ (1975-1984), and a test set, $\mathcal{T}_3$ (1985-2021). Additional information regarding data preprocessing can be found in Appendix C.

3.2 The Estimated Model Parameters

Panel A of Table 1 reports the out-of-sample distributional properties of the parameters, for the DSM with $K = 5$, on an annualized basis.[11] The μ_{it} parameter has an average of 9.39% and a standard deviation of 12.92%, and the empirical distribution has heavy tails as indicated by a p1 and p99 value of -26.07% and 50.85%, respectively. Breaking down μ_{it} into its constituent parts: r_{ft}, α_{it}, and $\beta_{it}\lambda_t$, we see that they contribute 36.00% ($\frac{3.38}{9.39}$), 23.21% ($\frac{2.18}{9.39}$), and 40.79% ($\frac{3.83}{9.39}$), respectively, to the average value of μ_{it}. Thus, the DSM suggests that, on average, the biggest contributor to the average expected asset return is systematic risk compensation, rather than mispricing. The contribution of α_{it} ranges from 19.95% to 29.11% across the 6 different DSM specifications with $K \in 1, ..., 6$. In terms of the average excess asset return, $\mu_{it} - r_{ft}$, systematic risk compensation constitutes 63.73% ($\frac{3.83}{6.01}$), while mispricing only constitutes 36.27% ($\frac{2.18}{6.01}$). It is worth noting that while α_{it} is the smallest contributor to the average asset drift, it is highly dispersed as indicated by a standard deviation of 12.51%, which is significantly higher than the standard deviations of 2.50% and 6.85% for r_{ft} and $\beta_{it}\lambda_t$, respectively. This means that while the location of the asset drift distribution is primarily determined by systematic risk compensation and the risk-free rate, the scale and tails are driven by the mispricing term. This effect seem to be largest at the left tail of the distribution, meaning that a negative expected asset return is more likely to be caused by overpricing (negative α_{it}), rather than because the firm's assets act as a hedge against systematic risk exposure (negative $\beta_{it}\lambda_t$).

Looking at σ_{it}, the DSM estimates an annualized average asset volatility of 34.69%. This is somewhat higher than previous literature such as Schaefer and Strebulaev (2008) and Feldhütter and Schaefer (2018) who estimate an average annualized asset volatility of 22% and 25%, respectively. These estimates, however, are based on samples of corporate bonds which are likely skewed towards larger firms. Limiting the sample of this paper to the 1,000 largest firms of each cross-section, as measured by market equity, reduces the average annualized asset volatility to 25.96%. Looking at the constituents of σ_{it}: $\sqrt{\beta_{it}^T\beta_{it}}$ and $\sqrt{\epsilon_{it}^2}$, it is clear that the vast majority of asset volatility is coming from systematic,

[11] As will become clear in the coming sections, the DSM specification with $K = 5$ is the best performing model, however, similar distributional results are obtained for Table 1 when using $K \in 1, 2, 3, 4, 6$.

	Mean	Std.	p1	p5	p10	p25	p50	p75	p90	p95	p99
Panel A: Annualized Parameters											
μ_{it}	9.39	12.92	-26.07	-10.78	-3.90	3.99	8.77	14.38	23.08	30.73	50.85
- r_{ft}	3.38	2.50	0.00	0.00	0.00	0.96	3.72	5.28	6.60	7.44	8.40
- α_{it}	2.18	12.51	-35.45	-19.23	-12.01	-3.25	2.69	8.24	15.29	21.30	36.48
- $\beta_{it}^T\lambda_t$	3.83	6.85	-6.13	-4.35	-3.22	-0.50	2.80	6.19	12.30	16.10	28.57
σ_{it}	34.69	16.40	10.36	13.79	17.26	23.08	30.78	43.40	57.61	66.86	85.44
- $\sqrt{\beta_{it}^T\beta_{it}}$	33.78	15.88	10.21	13.39	16.54	22.01	29.44	41.70	55.43	64.44	82.85
- $\sqrt{\epsilon_{it}^2}$	7.90	6.97	0.05	0.27	0.59	2.84	6.04	11.19	17.44	21.85	31.13
$L_{it} = \frac{D_{it}}{V_{it}}$	18.94	21.16	0.00	0.00	0.00	1.49	11.09	29.68	52.65	65.15	80.72
Panel B: Annualized Expected Equity Returns											
E$[r_{it+1}]$	11.62	19.33	-33.04	-14.01	-5.61	4.44	10.53	17.42	28.19	38.79	71.84

Table 1: *DSM Parameter and Expected Equity Return Distributions.* This table shows the out-of-sample distributions of the annualized parameters for the DSM with $K = 5$ (Panel A), along with the distribution of the annualized expected equity return (Panel B). The two first columns indicate the mean and standard deviation, while the rest denote specific percentiles of the distributions. All values are reported in percentages.

rather than idiosyncratic, volatility. Both the mean and standard deviation of $\sqrt{\beta_{it}^T\beta_{it}}$ is almost identical to that of σ_{it} itself. The actual proportion of the average asset volatility coming form systematic risk exposure is 94.82% ($\frac{33.78^2}{34.69^2}$) and this proportion ranges from 88.05%-98.19% across the six different DSM specifications with $K \in 1, ..., 6$.

Leverage, as defined by the ratio of the estimated market value of debt, D_{it}, to the overall market value of the assets, V_{it}, has an average value of 18.94%, but is heavily right-skewed with a substantial minority of firms having little to no debt. Leverage has a profound effect when moving from the distribution of μ_{it} in Panel A to the expected equity return distribution of Panel B: the average expected equity return is 11.62% which represents an increase of 23.75% over the average value of μ_{it}. Even more striking is the increase in the dispersion when moving from expected asset returns to expected equity returns: the standard deviation jumps to 19.33%, which represents a 49.61% increase. This is consistent with the empirical findings of Doshi, Jacobs, Kumar, and Rabinovitch (2019) that leverage has a large impact on the dispersion of equity returns.

Figure 4 plots the time series of the cross-sectional average for each of the three primary parameters: $\bar{\mu}_t$, $\bar{\sigma}_t$, and $\bar{L}_t$. The time series of $\bar{\mu}_t$ shows a high degree of short-term time series variation, but has no clear time trend or interaction with recessions (as indicated by the shaded red areas). The average asset volatility, $\bar{\sigma}_t$, is quite stable over short time periods[12] but has generally trended up during the out-of-sample time period, from a low of around 30% in the late 1980's to around 40% at the end of 2021. Finally, the cross-sectional average leverage, $\bar{L}_t$, spikes heavily during recessions, but remains relatively stable in a range of 15%-20% during normal times.

Figure 5 plots the time series of the estimated equity premium, EP_t^{DSM}, which is calculated as the cross-sectional value-weighted average of the expected equity return minus the risk-free rate. For comparison, the SVIX time series of Martin (2017) is also shown in the figure.[13] EP_t^{DSM} is generally above the SVIX, which is theoretically justified as it represents a lower bound on the equity premium, and their correlation is 0.41. Both time series peak during the financial crisis of 2008-09, although EP_t^{DSM} peaks at just over 15%, whereas the SVIX peaks above 25%. The fact that EP_t^{DSM} increase during recessions is interesting given that the time series of $\bar{\mu}_t$ in Figure 5 shows no such tendency. This suggests that the underlying asset dynamics of firms might be more stable than suggested by the literature; the time series dynamics of leverage is enough to create time series dynamics of the equity premium that are consistent with empirical observations.

3.3 Equity Return Prediction

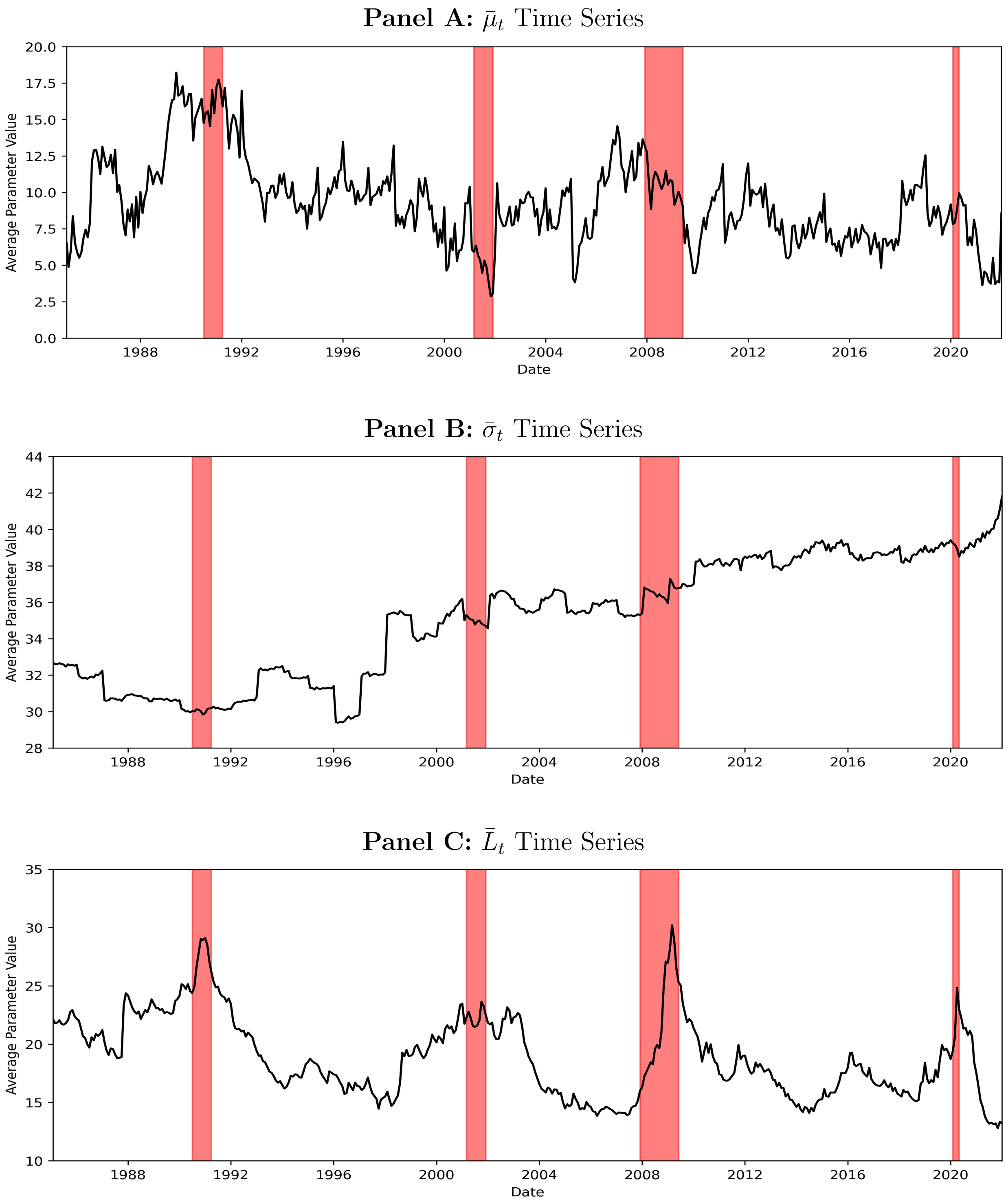
The DSM predicts firm-level equity returns by inserting the estimated model parameters into equation (18). The performance measure used to evaluate these predictions, is the zero-mean out-of-sample R^2_{oos}, also used in Gu, Kelly, and Xiu (2020):

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3} \left(r_{it+1} - \hat{r}_{it+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{it+1}} \tag{34}$$

Where $\hat{r}_{it+1}$ is the predicted equity return. The performance of the DSM is compared to two benchmark models: A machine learning benchmark (NN Benchmark), which is chosen

[12] The sudden spikes and drops that occur during the first half of the out-of-sample period is a consequence of re-training the model each year. Incorporating another year of training data has a comparatively larger impact on the estimated model parameter functions of Figure 2 in the first half of the sample compared to the second half, since another year represents a larger fraction of additional data during this time.

[13] The SVIX time-series is downloaded directly from Ian Martin's personal website and covers time time-period from 1996-2011.


Panel A: $\bar{\mu}_t$ Time Series

Panel B: $\bar{\sigma}_t$ Time Series

Panel C: $\bar{L}_t$ Time Series

Figure 4: *DSM Parameter Time Series.* This figure plots the out-of-sample time series of the cross-sectional average annualized parameter value for the three primary model parameters: $\bar{\mu}_t$ (Panel A), $\bar{\sigma}_t$ (Panel B), and $\bar{L}_t$ (Panel C), from the DSM specification with $K = 5$. The time periods shaded in red indicate (NBER) recessions.

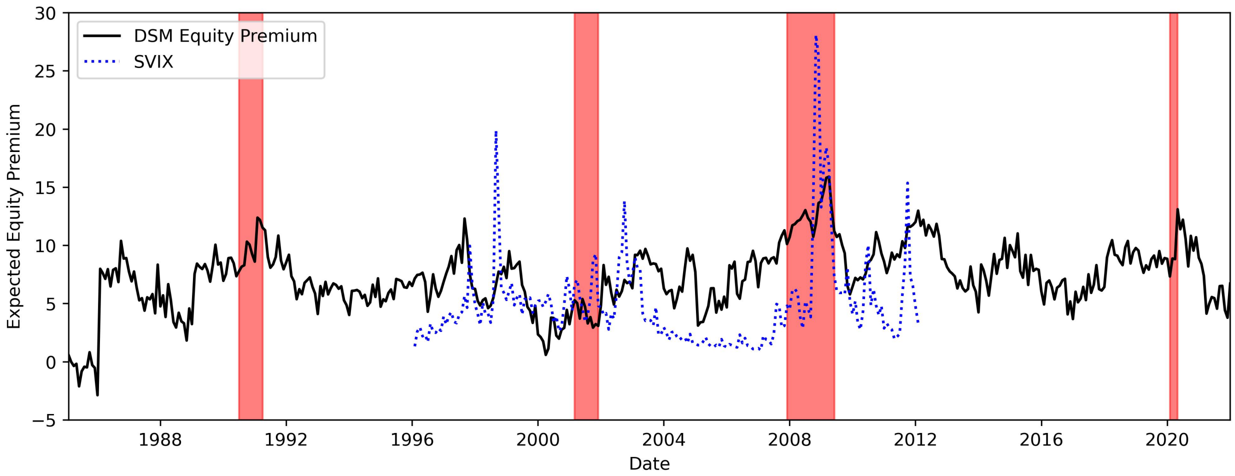


Figure 5: *Equity Premium Time Series.* This figure plots the out-of-sample time series of the cross-sectional value-weighted average expected excess equity return, from the DSM specification with $K = 5$, calculated across all firms (full black line). Additionally, the figure also plots the SVIX time series of Martin (2017) (dotted blue line).

to be a standard feed-forward neural network, as it is the best performing model in Gu, Kelly, and Xiu (2020), and a linear benchmark (OLS Benchmark), which is simply an OLS regression. Both of these are trained on the same data as the DSM, with the neural network also having a similar training procedure as the DSM (see Appendix B.2 for details). Both of the benchmark models use the combined set of firm characteristics and macroeconomic variables as input. Table 2 reports the performance of the DSM, along with the NN and OLS benchmarks. The first six columns of the table report the performance of the DSM with an increasing number of systematic risk factors, indicated by $K \in 1, ..., 6$, while the two last columns report the performance of the benchmark models.[14] Each row of Table 2 indicate which subset of the data is used for calculating R^2_{oos}, with "All" referring to the entire test dataset, while "Top 1,000" ("Bottom 1,000") refers to the subset containing only the 1,000 largest (smallest) firms of each cross-section, as measured by market equity. From the table it is clear that the DSMs outperform both of the benchmark models in terms of R^2_{oos}: all DSM specifications have values between 0.74 and 0.80, when estimated across all out-of-sample observations, compared to 0.56 for the NN benchmark and -10.60 for the OLS benchmark.[15]

[14] Five different versions of the NN benchmark model have been trained, with a varying number of hidden layers, $H \in 1, ..., 5$, but only the best performing one in terms of R^2_{oos} ($H = 2$) is used as a benchmark model throughout this section.

[15] The R^2_{oos} value for the NN benchmark is similar to the one reported in Gu, Kelly, and Xiu (2020).

This outperformance is not due to the presence of small-caps as the majority of the DSMs actually have higher R^2_{oos}-values for the data subset containing the largest firms.[16]

	DSM						NN	OLS
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	Benchmark	Benchmark
All	0.75	0.74	0.76	0.77	0.80	0.75	0.56	-10.60
Top $1,000$	1.03	1.02	1.03	1.04	1.11	1.04	0.85	-35.33
Bottom $1,000$	0.89	0.85	0.90	0.91	0.93	0.88	0.67	-4.47

Table 2: *Equity Return Prediction Performance.* This table reports the out-of-sample equity return predictive performance for the DSM and two benchmark models: a neural network model and an OLS regression. The performance measure is the zero-mean R^2_{oos} measure, $R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{it+1}-\hat{r}_{it+1})^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{it+1}}$, reported in percentages. The table reports the performance for six different specifications of the DSM, with a varying number of systematic shocks $K \in 1, ..., 6$, indicated by the first six columns. Each row of the table denotes the specific subset of the data used for calculating R^2_{oos}. "All" refers to the entire test dataset, $\mathcal{T}_3$, while "Top 1,000" ("Bottom 1,000") refers to the sub-sample consisting of only the 1,000 largest (smallest) firms of each cross-section.

Next, to examine if the higher R^2_{oos}-values of the DSMs translate into portfolios with higher returns and Sharpe ratios, I form decile portfolios each month based on the expected equity returns from each model. In addition to the decile portfolios, I also create a long-short portfolio that is long the 10th decile portfolio and short the 1st. Table 3 reports the out-of-sample performance of these portfolios in terms of the average monthly excess return (Panel A), the standard deviation of the monthly excess returns (Panel B), and the annualized Sharpe ratios (Panel C). Looking at Panel A, we see that all DSMs have a strictly monotone increase in the realized average excess returns, as we move down the panel, which is not the case for the benchmark models (although it is close in the case of the NN benchmark). Furthermore, the realized excess returns of the long-short portfolios are much higher for the DSM portfolios compared to the benchmark models: the lowest average excess return of the DSM based long-short portfolios is 2.63% ($K = 1$), which is still 0.72 percentage points higher than the NN benchmark, while the best performing DSM portfolio has an average monthly excess return of 3.10% ($K = 5$).

Looking at Panel B, a curious pattern emerges: the DSMs indicate a pronounced convex relationship between the realized excess return and the standard deviation of the decile

[16] Interestingly, the DSMs and the NN benchmark generally have higher R^2_{oos} for both of the data subsets examined here, indicating that these models perform worse, in terms of R^2_{oos}, for mid-sized firms compared to the firms at the ends of the size spectrum.

	DSM						NN	OLS
Decile	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	Benchmark	Benchmark
Panel A: Average Monthly Excess Return								
1	-0.78	-0.87	-0.84	-0.90	-1.01	-0.90	-0.16	0.33
2	-0.15	-0.15	-0.12	-0.12	-0.03	-0.11	0.30	0.56
3	0.33	0.29	0.32	0.36	0.24	0.26	0.63	0.73
4	0.53	0.58	0.57	0.58	0.51	0.48	0.78	0.73
5	0.73	0.72	0.72	0.78	0.73	0.74	0.73	0.97
6	0.85	0.87	0.89	0.95	0.99	0.94	0.98	0.95
7	1.09	0.99	1.11	1.08	1.03	1.03	1.11	1.02
8	1.20	1.26	1.20	1.23	1.25	1.23	1.20	1.00
9	1.50	1.44	1.46	1.43	1.33	1.36	1.39	1.37
10	1.85	1.88	1.99	1.86	2.09	1.89	1.75	1.17
10-1	2.63	2.75	2.83	2.76	3.10	2.79	1.91	0.85
Panel B: Std. of Monthly Excess Returns								
1	8.40	8.50	8.65	8.33	8.50	8.78	6.40	4.99
2	6.57	6.54	6.49	6.56	6.77	6.91	5.44	4.31
3	5.04	5.16	4.95	5.03	5.12	5.48	5.22	4.31
4	4.59	4.45	4.43	4.38	4.48	4.35	4.97	4.45
5	4.30	4.31	4.29	4.37	4.35	4.33	5.05	4.64
6	4.31	4.38	4.44	4.45	4.47	4.40	5.06	4.73
7	4.61	4.63	4.67	4.67	4.73	4.70	5.13	5.38
8	5.12	5.26	5.21	5.29	5.12	5.24	5.50	5.86
9	5.93	6.15	6.16	6.64	6.28	6.09	5.97	6.41
10	7.34	7.63	7.56	7.58	7.70	7.69	6.90	7.66
10-1	6.38	6.53	6.44	6.26	6.44	6.40	6.10	5.64
Panel C: Annualized Sharpe Ratio								
1	-0.32	-0.36	-0.34	-0.38	-0.41	-0.36	-0.16	0.23
2	-0.08	-0.08	-0.07	-0.06	-0.02	-0.05	0.19	0.45
3	0.23	0.20	0.22	0.25	0.16	0.16	0.42	0.59
4	0.40	0.45	0.45	0.46	0.40	0.39	0.54	0.56
5	0.59	0.58	0.58	0.61	0.58	0.59	0.50	0.73
6	0.69	0.69	0.70	0.74	0.80	0.74	0.67	0.69
7	0.82	0.74	0.82	0.80	0.73	0.76	0.75	0.65
8	0.81	0.83	0.80	0.81	0.85	0.81	0.76	0.59
9	0.88	0.81	0.82	0.75	0.74	0.75	0.81	0.74
10	0.87	0.85	0.91	0.85	0.94	0.85	0.88	0.53
10-1	1.43	1.46	1.52	1.53	1.67	1.51	1.08	0.52

Table 3: *Decile Portfolio Performance.* This table reports the out-of-sample performance of decile portfolios based on monthly sorts on the expected equity returns, $\hat{r}_{it+1}$, from one of seven different models as indicated by the columns. In addition to the decile portfolios, the performance of a long-short portfolio, which is long the 10th decile portfolio and short the 1st, is also reported. Panel A of the table reports the average monthly excess return (in percentages) for each portfolio, Panel B reports the standard deviation of monthly excess returns, while Panel C reports the annualized Sharpe ratios. All portfolios are value-weighted.

portfolios, i.e. the most volatile portfolios are those with the lowest and highest average excess returns. One might suspect that this convex relationship between excess returns and volatility would cause the Sharpe ratios of the highest decile portfolios to be lower than the middle ones, but this is generally not the case, as is evident from Panel C: the Sharpe ratios of the DSM portfolios are generally increasing. Looking at the DSM based long-short portfolios, we see that they handily outperform the benchmark portfolios with Sharpe ratios between 1.43 ($K = 1$) and 1.67 ($K = 5$), compared to 1.08 and 0.52 for the NN benchmark and OLS benchmark, respectively. Investors can therefore achieve significant benefits in terms of both absolute returns and Sharpe ratios by adopting the DSM framework of this paper, when forming long-short portfolios, compared to an off-the-shelf machine learning approach.

3.4 Equity Return Variance Prediction

One of the advantages of the DSM presented in this paper is its versatility: it is not confined to being an equity return model but offers insights into all the model implications presented in Section 2.3. To explore this further, I use the analytical expression in (20) to produce out-of-sample equity return variance predictions. Unlike equity returns, variances are not observed over a single period and so the "realized" variance is based on the daily variance of equity returns between t and $t + 1$:

$$\text{Var}[r_{it+1}] = \sum_{d=1}^{D}(r_{id} - \bar{r}_{it+1})^2 \tag{35}$$

Where D is the number of days between time t and $t + 1$, r_{id} is the daily return on day d, and $\bar{r}_{it+1}$ is the average daily return across all D days. As a benchmark, a GARCH(1,1) model is recursively fit to each individual firm.[17]

[17] Because of the estimation procedure of the GARCH(1,1) model, the predictions are restricted to firm-month observations without any data gaps and with at least 12 prior observations, e.g. if a firm enters the dataset in July, 1995, and exits after July, 1997, then it is required that there are a total of 24 firm-month observations for this particular firm and the first 12 observations will not be used for the variance prediction analysis. With this restriction, the test dataset shrinks from 2,334,603 to 1,605,489 observations for this particular section of the paper. The fact that we need this restriction also highlights a strength of the DSM framework: once trained, all that is needed for predicting the next-period variance is a single contemporary firm observation, i.e. no historical information is needed.

	DSM						GARCH
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	Forecast
Panel A: Regression Results							
Constant	0.02***	0.25***	0.37***	0.42***	0.50***	0.53***	-1.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\log(\widehat{\text{Var}}[r_{it+1}])$	1.01***	1.04***	1.07***	1.08***	1.10***	1.11***	0.76***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.53	0.53	0.53	0.53	0.53	0.53	0.50
N	1,605,489	1,605,489	1,605,489	1,605,489	1,605,489	1,605,489	1,605,489
Panel B: MSE							
All	3.257	3.276	3.280	3.275	3.280	3.285	4.137
Top 1,000	0.096	0.096	0.096	0.095	0.095	0.096	0.168
Bottom 1,000	18.812	18.905	18.919	18.894	18.926	18.950	21.420

Table 4: *Equity Return Variance Prediction.* This table reports the out-of-sample performance of seven different models for predicting the equity return variance over the coming month. The first six columns are DSMs with an increasing number of systematic shocks, as indicated by K, while the "GARCH Benchmark" refers to the performance of a GARCH(1,1) model recursively fit to each individual firm. Panel A reports the results of the regression $\log(\text{Var}[r_{it+1}]) = a + b\log(\widehat{\text{Var}}[r_{it+1}])$, where $\text{Var}[r_{it+1}]$ is the daily equity return variance between time t and $t+1$, multiplied by the number of days in that time period, and $\widehat{\text{Var}}[r_{it+1}]$ is the predicted variance. The parentheses report the estimated standard errors, while ***, **, *, indicate statistical significance at the 0.05, 0.01, and 0.001 level, respectively. Panel B reports the mean squared error, $MSE_{oos} = (\text{Var}[r_{it+1}] - \widehat{\text{Var}}[r_{it+1}])^2$, scaled by 100, across all out-of-sample observations, as well as the data subsets consisting of the 1,000 largest and smallest firms of each cross-section.

Panel A of Table 4 reports the out-of-sample results of regressing the log of the realized return variance onto the log of the variance prediction, for each of the six DSM specifications and the GARCH benchmark:

$$\log(\text{Var}[r_{it+1}]) = a + b\log(\widehat{\text{Var}}[r_{it+1}]) \tag{36}$$

Where $\widehat{\text{Var}}[r_{it+1}]$ is the predicted equity return variance for the next period. The panel shows that the DSM predictions explain a higher proportion of the return variance compared to the GARCH benchmark, as measured by R^2. Interestingly, it seems like the DSMs and the GARCH model produce variance estimates that are biased in opposite directions: the DSMs (GARCH model) tend to underestimate (overestimate) the realized variance as indicated by a positive (negative) constant and a coefficient above (below) one. The six different DSM specifications result in different regression coefficients, especially in terms of the intercept which ranges from 0.02 ($K = 1$) to 0.53 ($K = 6$), however, this does not translate into differences in terms of R^2. Panel B of Table 4 report the out-of-sample mean squared error, MSE_{oos}, for each of the seven models:

$$MSE_{oos} = \frac{1}{N_{\mathcal{T}_3}} \sum_{(i,t)\in\mathcal{T}_3} (\text{Var}[r_{it+1}] - \widehat{\text{Var}}[r_{it+1}])^2 \tag{37}$$

From Panel B, we see that the DSMs have MSEs that are between 20.59%-21.27% lower than that of GARCH benchmark. This outperformance is consistent across the size spectrum of firms, as indicated by the second and third row of the panel.

These results suggest that, even though the DSMs have only been trained on equity return data, the conditional parameter estimates, used in conjunction with the model implications of Section 2.3, are able to accurately describe and predict, not only the first, but the second moment of equity returns, on a firm-level basis.

3.5 Mean-Variance Efficient Portfolios

To examine how well the DSM estimate the conditional covariance matrix of equity returns, I construct a classic mean-variance efficient (MVE) portfolio in the spirit of Markowitz (1952), at each point in time, by solving the following portfolio choice problem:

$$\max_{\boldsymbol{w}_t} \quad \frac{\boldsymbol{w}_t^T(\hat{\boldsymbol{r}}_{t+1} - \mathbf{1}r_{ft})}{\boldsymbol{w}_t^T\hat{\Sigma}_{t+1}\boldsymbol{w}_t} \tag{38}$$

$$\text{s.t.} \quad \boldsymbol{w}_t^T\mathbf{1} = 1 \tag{39}$$

$$||\boldsymbol{w}_t||_1 \leq 3 \tag{40}$$

Where $\boldsymbol{w}$ is the vector containing the MVE portfolio weights, $\hat{\boldsymbol{r}}_{t+1}$ is the vector of expected returns, and $\hat{\Sigma}_{t+1}$ is the estimated conditional covariance matrix with elements calculated according to (23). The first constraint in (39) ensures the weights sum to one. The second in (40) states that the ℓ_1-norm of the weights cannot exceed 3, which has the practical effect of limiting the leverage of the MVE portfolio such that the sum of the negative weights do not exceed -1. In addition to the MVE portfolios created by solving (38)-(40), a long-only version is also constructed by replacing the constraint in (40) with $w_{it} \geq 0 \quad \forall i \in 1, ..., I_t$, where I_t is the number of firms in cross-section t. These are referred to as "long-only MVE portfolios", while the others are referred to as "unconstrained MVE portfolios".[18] At each point in time, both the unconstrained and long-only MVE portfolios are constructed using three different investment universes: all firms, the largest 3,000 firms, and the largest 1,000 firms.

Table 5 and 6 reports the monthly average excess returns (Panel A), the monthly standard deviation of excess returns (Panel B), and the annualized Sharpe ratios (Panel C), for the unconstrained and long-only MVE portfolios, respectively, for each of the six DSM specifications. The rows of each panel indicate which investment universe has been used for constructing the MVE portfolios. Even when only considering the 1,000 largest firms of each cross-section, the unconstrained MVE portfolios have average monthly excess returns of over 2% and Sharpe ratios in the range of 1.07-1.54, which is comparable to the DSM based long-short portfolios of Section 3.3. Increasing the investment universe to all firms has the effect of increasing the average excess return and lowering the portfolio volatility, resulting in extremely high Sharpe ratios: the unconstrained MVE portfolios, based on the entire investment universe, have average monthly excess returns in the range of 4.74-5.19 and Sharpe ratios between 2.92 and 3.96. The long-only MVE portfolios constructed from the entire investment universe generally have slightly lower average excess returns than their unconstrained counterparts. They exhibit around the twice the volatility, resulting in Sharpe ratios in the range of 1.34-1.93. Even though these Sharpe ratios seem small compared to the Sharpe ratios of the unconstrained MVE portfolios, it is worth noting that the S&P 500 index had a Sharpe ratio of 0.48 over the same time period. The best performing long-only MVE portfolio, constructed from only the 1,000 largest firms of each cross-section, had a Sharpe ratio of almost double that (0.90 for the DSM with $K = 6$).

Figure 6 plots the cumulative log returns, over the out-of-sample period, of all MVE portfolios for the DSM specification with $K = 5$. Additionally, the figure also plots the cumulative log returns of the long-short portfolio based on the DSM with $K = 5$, the long-

[18] While the "unconstrained MVE portfolios" are subject to a leverage constraint, they produce portfolios that a similar to a scaled version of the completely unrestricted MVE portfolios, which is why they are labelled as "unconstrained". The leverage constraint has the effect of stabilizing the portfolio performance across the the different DSM specifications, but is not strictly needed to produce portfolios with high Sharpe ratios.

	DSM					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
Panel A: Average Monthly Excess Return						
All	4.79	5.19	4.74	4.78	4.89	4.96
Top 3,000	3.52	3.91	3.52	3.58	3.61	3.61
Top 1,000	2.21	2.34	2.17	2.17	2.25	2.32
Panel B: Std. of Monthly Excess Return						
All	5.69	4.72	4.22	4.22	4.28	4.72
Top 3,000	5.41	4.59	4.28	4.35	4.21	4.69
Top 1,000	7.14	5.79	4.95	4.89	5.18	5.65
Panel C: Annualized Sharpe Ratio						
All	2.92	3.81	3.89	3.92	3.96	3.64
Top 3,000	2.25	2.95	2.85	2.85	2.97	2.66
Top 1,000	1.07	1.40	1.52	1.54	1.51	1.42

Table 5: *Unconstrained MVE Portfolio Performance.* This table reports the out-of-sample performance of the MVE portfolios formed on the basis of the six DSM specifications with $K \in 1, ..., 6$. Panel A reports the average monthly excess return (in percentages), Panel B reports the standard deviation of excess returns, while Panel C reports the annualized Sharpe ratio. The rows of each panel indicate which subset of the data has been used to construct the MVE portfolios.

	DSM					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
Panel A: Average Monthly Excess Return						
All	4.08	4.56	4.56	4.85	4.55	4.58
Top 3,000	2.34	2.49	2.40	2.53	2.44	2.31
Top 1,000	1.23	1.44	1.49	1.60	1.46	1.57
Panel B: Std. of Monthly Excess Return						
All	10.53	10.20	8.84	8.85	8.52	8.24
Top 3,000	8.23	8.51	7.15	7.29	7.02	6.57
Top 1,000	7.84	7.10	6.35	6.24	6.00	6.01
Panel C: Annualized Sharpe Ratio						
All	1.34	1.55	1.79	1.90	1.85	1.93
Top 3,000	0.99	1.01	1.16	1.20	1.20	1.22
Top 1,000	0.55	0.70	0.81	0.89	0.84	0.90

Table 6: *Long-Only MVE Portfolio Performance.* This table reports the same statistics as Table 5, but for MVE portfolios that prohibits short-selling.

short portfolio based on the NN benchmark model, the S&P 500 index, and the CRSP value-weighted index. All portfolios have been scaled to have an overall annualized volatility of 10%. All of the model based portfolios have outperformed the market portfolios during this period.

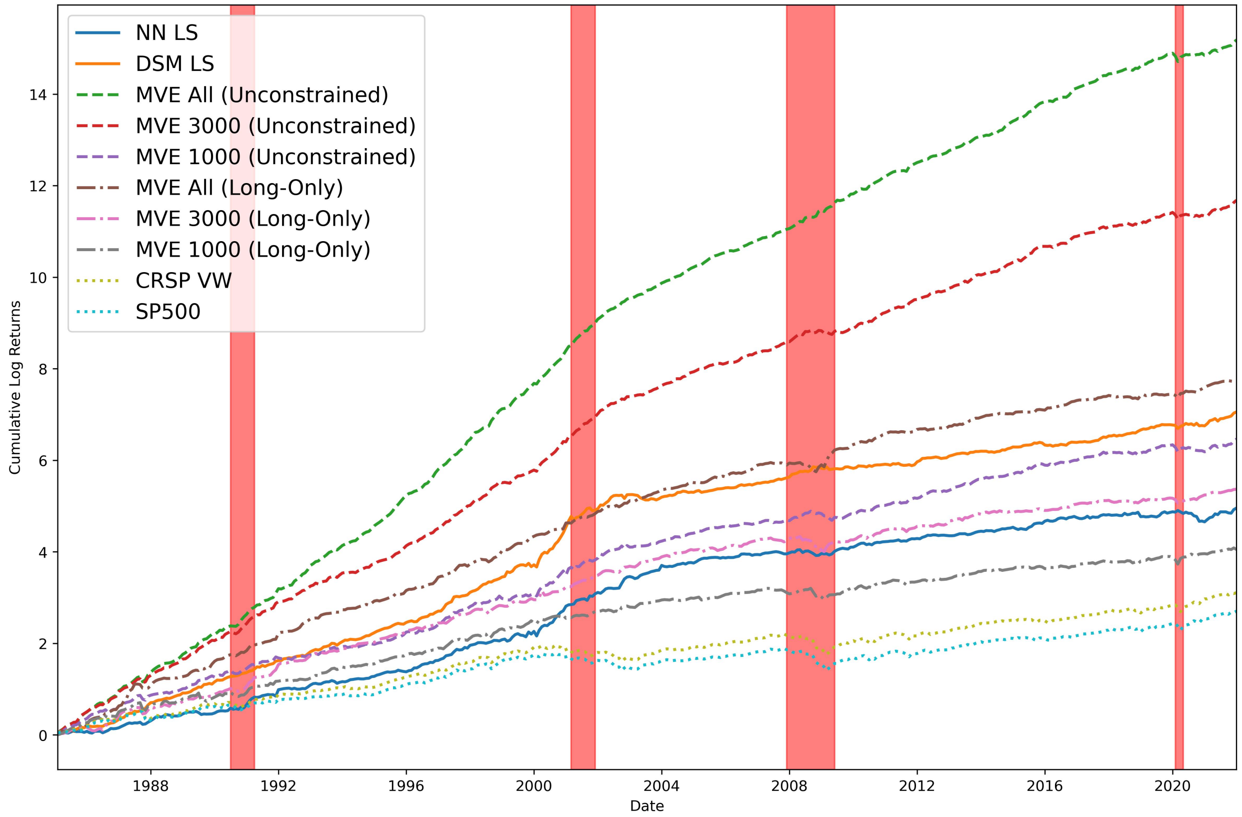


Figure 6: *MVE Portfolio Performances.* This figure shows the out-of-sample cumulative log returns for the MVE portfolios based on the DSM with $K = 5$. The figure includes 3 versions of an unrestricted and long-only MVE portfolio: one that includes all firms in the investment universe, one that only considers the largest 3,000 firms when forming the portfolio, and one which only considers the largest 1,000 firms when forming the portfolio. Additionally, the "NN LS" and "DSM LS" are the long-short portfolios formed on the basis of the expected returns of the NN benchmark model and the DSM model with $K = 5$, respectively, from Section 3.3. Finally, the figure also includes the cumulative log return of the S&P500 index and the CRSP value-weighted index. All portfolios are re-balanced monthly and scaled to have an annualized volatility of 10%.

It is also clear from the figure that the unconstrained MVE portfolio, based on the entire investment universe, has performed several orders of magnitude better than the other portfolios and its performance seem completely uncorrelated with recessions. Table 7 explores this further by reporting various other performance measures for the scaled portfolios of Figure 6. The first two rows report the average excess monthly returns and the annualized Sharpe ratios. The third row reports the estimated intercept (or α) of regressing the portfolio return onto the five factors of Fama and French (2015) and the momentum factor of Carhart (1997) (the "FF6 model"). The fourth and fifth row report the maximum portfolio drawdown and the highest 1 month loss, respectively. They key takeaway from the table is the fact that the returns of all the different MVE portfolios are considered to be α in the FF6 model, and they do not experience drawdowns of the same magnitude as those of the market portfolios.

	Long-Short		Unconstrained MVE's			Long-Only MVE's			Market Portfolios	
	NN	DSM	All	Top 3,000	Top 1,000	All	Top 3,000	Top 1,000	CRSP VW	SP500
ER	0.91	1.39	3.26	2.45	1.25	1.54	1.00	0.70	0.49	0.40
SR	1.08	1.67	3.96	2.97	1.51	1.85	1.20	0.84	0.59	0.48
FF6 α	0.83	1.31	3.29	2.52	1.49	0.93	1.02	0.83	-0.01	-0.14
Max DD.	-21.95	-11.10	-16.51	-13.95	-18.25	-11.02	-27.84	-21.43	-36.32	-38.02
Max 1M Loss	-10.04	-6.98	-11.18	-9.05	-7.66	-11.02	-12.94	-15.03	-14.44	-14.28

Table 7: *MVE Portfolio Performance Measures.* This table reports various out-of-sample portfolio measures for the portfolios depicted in Figure 6. The first and second row report the average excess returns and annualized Sharpe ratios, respectively. The third row reports the intercept from regressing the monthly portfolio returns onto the five factors of Fama and French (2015) and the momentum factor of Carhart (1997). The fourth row shows the maximum drawdown of the portfolio, while the fifth row reports the largest 1 month loss experienced during the out-of-sample period.

The performance of the MVE portfolios is impressive, especially considering no form of covariance shrinkage is needed (see Ledoit and Wolf (2022) for a literature review on shrinking the covariance matrix). Still, it is important to note that this paper does *not* claim that these returns and Sharpe ratios can be achieved by an investor. While the leverage and shorting restrictions, along with the investment universe limits, seek to make the portfolios more realistic than a completely unrestricted MVE portfolio, there are still no considerations given to trading costs or limits to short-selling. With that being said, this section shows that not only does the DSM provide accurate return and variance predictions, it also provides useful information about the covariance of equity returns.

3.6 Enforcing No-Arbitrage

The distribution of the estimated model parameters shown in Table 1 seem to suggest that α_{it} plays an important role for the dispersion of the asset drift, μ_{it}, and by extension, the dispersion of expected equity returns. Removing the α_{it} term from the model would force the DSM to find parameter solutions in which all excess asset returns are compensation for systematic risk exposure. It might be the case that this solution produce equity return predictions that are just as accurate as the full model. To test whether this is the case, the six different DSM specifications are re-trained with the α_{it} parameter set to zero. The equity return prediction exercise of Section 3.3 is then repeated.

Table 8 report R^2_{oos} for the no-arbitrage DSMs. The first row contains the absolute values while the second contains the ratio of the no-arbitrage version to the full model counterpart. The table shows that, for low values of K, the no-arbitrage versions of the DSM are much worse than the full model for predicting equity returns. For $K \geq 4$, the performance is better than the NN benchmark, but still not as good as their full model counterparts.

	DSM					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
R^2_{oos} Value	0.19	0.26	0.47	0.63	0.67	0.61
% of Full Model	25.33	35.14	61.84	81.82	83.75	81.33

Table 8: *No-Arbitrage Equity Return Prediction Performance.* This table reports the same R^2_{oos} measure as Table 2, but for the equity return predictions from the six DSM versions with α_{it} set to zero. The first row is the absolute R^2_{oos}-value across all out-of-sample observations, while the second is the relative value compared to the full model performance (as found in the first row of Table 2), reported in percentages.

.

As in Section 3.3, I form decile portfolios based on the expected returns from the no-arbitrage DSMs and look at their out-of-sample performances. The results can be seen in Table 9. Here we see that the lower R^2_{oos}-values for the no-arbitrage DSMs translate into lower average realized excess returns and Sharpe ratios. The best performing long-short portfolio is (again) the portfolio based on the predictions from the DSM with $K = 5$, however, this portfolio "only" has a Sharpe ratio of 1.06, which is not only lower than the full model DSM, but also lower than the NN benchmark long-short portfolio.[19]

The results of this section confirm the results of Section 3.2, namely, that α_{it} is an important piece of the puzzle. It not only helps the DSM provide more accurate equity return predictions, but is also crucial when constructing portfolios based on the estimated model parameters.

[19] Constructing MVE portfolios based on the no-arbitrage DSMs have higher Sharpe ratios than the NN benchmark. They do not, however, outperform the MVE portfolios constructed from the full model DSMs and their relative performance mirror that of the long-short portfolios.

	DSM					
Decile	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
Panel A: Average Monthly Excess Return						
1	0.69	0.12	-0.58	-0.64	-0.75	-0.84
2	0.75	0.66	0.19	-0.02	0.07	0.09
3	0.80	0.78	0.39	0.33	0.22	0.32
4	0.84	0.88	0.74	0.66	0.68	0.63
5	0.91	1.03	0.94	0.86	0.81	0.85
6	0.92	1.19	1.09	1.07	1.00	0.88
7	0.77	1.09	1.19	1.22	1.20	1.12
8	0.53	0.83	1.16	1.25	1.10	1.05
9	0.38	0.79	1.04	1.12	1.16	1.20
10	-0.08	0.82	1.04	1.28	1.52	1.32
10-1	-0.77	0.70	1.62	1.92	2.27	2.16
Panel B: Std. of Monthly Excess Returns						
1	3.62	5.16	7.93	8.78	9.18	9.29
2	4.24	4.55	6.06	6.53	7.19	6.97
3	4.92	4.48	4.80	4.90	4.98	4.95
4	5.65	4.78	4.33	4.36	4.33	4.28
5	6.41	5.35	4.57	4.14	4.28	4.29
6	7.28	5.71	5.15	4.61	4.61	4.60
7	7.99	6.57	5.06	5.19	5.04	5.13
8	8.49	6.95	6.01	5.63	5.69	5.39
9	9.62	7.53	6.37	6.23	6.08	6.49
10	10.72	8.53	7.78	7.85	7.51	7.53
10-1	9.58	7.00	6.73	6.76	7.39	7.10
Panel C: Annualized Sharpe Ratio						
1	0.66	0.08	-0.25	-0.25	-0.28	-0.31
2	0.62	0.51	0.11	-0.01	0.03	0.05
3	0.56	0.60	0.28	0.23	0.15	0.23
4	0.51	0.64	0.59	0.53	0.55	0.51
5	0.49	0.66	0.72	0.68	0.66	0.68
6	0.44	0.72	0.73	0.81	0.76	0.66
7	0.34	0.58	0.81	0.81	0.82	0.76
8	0.21	0.41	0.67	0.77	0.67	0.68
9	0.14	0.36	0.57	0.62	0.66	0.64
10	-0.03	0.33	0.46	0.57	0.70	0.61
10-1	-0.28	0.34	0.83	0.98	1.06	1.05

Table 9: *No-Arbitrage Decile Portfolio Performance.* This table reports the out-of-sample performance of decile portfolios based on monthly sorts on the expected equity returns, from the six DSM versions with α_{it} set to zero. The table is structured the same way as Table 3.

4 Conclusion

Using a novel modelling framework, coined Deep Structural Models (DSMs), I estimate conditional monthly firm-level parameters of a Merton type model that allows for mispricing and decomposes the total risk of the firm's assets into a systematic and idiosyncratic component. The framework suggests that systematic risk compensation is the largest contributor to the average asset return, while the mispricing component is the primary driver of the dispersion of asset returns. Furthermore, the effect of leverage on the asset returns of firms is the main mechanism responsible for an increased equity premium during recessions. In fact, the DSM indicate that both the asset drift and volatility remain unaffected by the onset of a recession. The estimated parameters of the DSM jointly model the expected equity returns and (co)variances through analytically derived expressions. The estimated expected equity returns have higher predictive power than an "off-the-shelf" neural network model and the DSM predictions enables investors to form long-short portfolios with higher out-of-sample absolute returns and Sharpe ratios. Additionally, the variance predictions of the DSM have higher predictive power than a GARCH(1,1) model. Finally, I use the estimated expected returns and covariance matrix to construct leverage constrained and long-only mean variance efficient portfolios, re-balanced on a monthly basis. These portfolios have much higher returns and Sharpe ratios than the long-short portfolios, indicating that the estimated conditional covariance matrix does indeed carry relevant information about the covariances of equity returns.

References

Bali, T. G., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2020). Predicting corporate bond returns: Merton meets machine learning. *Georgetown McDonough School of Business Research Paper* (3686164), 20–110.

Bharath, S. T. and T. Shumway (2008). Forecasting default with the merton distance to default model. *The Review of Financial Studies 21*(3), 1339–1369.

Black, F. and J. C. Cox (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance 31*(2), 351–367.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

Chen, L., M. Pelger, and J. Zhu (2023). Deep learning in asset pricing. *Management Science*.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance 66*(4), 1047–1108.

Doshi, H., K. Jacobs, P. Kumar, and R. Rabinovitch (2019). Leverage and the cross-section of equity returns. *The Journal of Finance 74*(3), 1431–1471.

Du, D., R. Elkamhi, and J. Ericsson (2019). Time-varying asset volatility and the credit spread puzzle. *The Journal of Finance 74*(4), 1841–1885.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Feldhütter, P. and S. Schaefer (2023). Debt dynamics and credit risk. *Journal of Financial Economics 149*(3), 497–535.

Feldhütter, P. and S. M. Schaefer (2018). The myth of the credit spread puzzle. *The Review of Financial Studies 31*(8), 2897–2942.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance 75*(3), 1327–1370.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Giglio, S., B. Kelly, and D. Xiu (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics 14*, 337–368.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*(1), 429–450.

Harvey, C. R., Y. Liu, and H. Zhu (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *The Review of financial studies 33*(5), 2019–2133.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr.

Israel, R., B. T. Kelly, and T. J. Moskowitz (2020). Can machines' learn'finance? *Journal of Investment Management*.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2022). Is there a replication crisis in finance? *The Journal of Finance*.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Ledoit, O. and M. Wolf (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics 20*(1), 187–218.

Leland, H. E. (1994). Corporate debt value, bond covenants, and optimal capital structure. *The journal of finance 49*(4), 1213–1252.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics 132*(1), 367–433.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance 29*(2), 449–470.

Schaefer, S. M. and I. A. Strebulaev (2008). Structural models of credit risk are useful: Evidence from hedge ratios on corporate bonds. *Journal of Financial Economics 90*(1), 1–19.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*(1), 1929–1958.

Vassalou, M. and Y. Xing (2004). Default risk in equity returns. *The journal of finance 59*(2), 831–868.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

A Proofs

A.1 Expected Equity Return

The expected equity value at time $t+1$ can be written as:

$$\mathrm{E}[E_{it+1}] = \mathrm{E}[(V_{it+1} - F_{it+1})\mathbb{1}_{V_{it+1} > F_{it+1}}]$$

By the law of iterated expectation, this can be rewritten to:

$$\begin{aligned}
\mathrm{E}[E_{it+1}] &= \mathrm{E}\big[\mathrm{E}[(V_{it+1} - F_{it+1})\mathbb{1}_{V_{it+1} > F_{it+1}} | \mathbb{1}_{V_{it+1} > F_{it+1}} = 1] \\
&\quad + \mathrm{E}[(V_{it+1} - F_{it+1})\mathbb{1}_{V_{it+1} > F_{it+1}} | \mathbb{1}_{V_{it+1} > F_{it+1}} = 0]\big] \\
&= (1 - \pi_{it})\mathrm{E}[(V_{it+1} - F_{it+1}) | \mathbb{1}_{V_{it+1} > F_{it+1}} = 1] + \pi_{it}0 \\
&= (1 - \pi_{it})(\mathrm{E}[V_{it+1} | \mathbb{1}_{V_{it+1} > F_{it+1}} = 1] - F_{it+1})
\end{aligned}$$

Where the second equality comes from the fact that if $\mathbb{1}_{V_{it+1} > F_{it+1}} = 0$ then obviously $(V_{it+1} - F_{it+1})\mathbb{1}_{V_{it+1} > F_{it+1}} = 0$. The third equality is from realizing that F_{it+1} is a constant. Now use the general result that if a random variable X is log-normally distributed, such that $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$, then the expectation of X conditional on $X \geq K$ is:

$$\mathrm{E}[X | X \geq K] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \frac{\Phi\left(\frac{\mu - \ln(K) + \sigma^2}{\sigma}\right)}{1 - \Phi\left(\frac{\ln(K) - \mu}{\sigma}\right)}$$

Using this expression with $\mu = \ln(V_{it}) + \mu_{it}$, $\sigma = \sigma_{it}$, and $K = F_{it+1}$ we can write the expectation of V_{it+1} conditional on $V_{it+1} \geq F_{it+1}$ as:

$$\begin{aligned}
\mathrm{E}[V_{it+1} | \mathbb{1}_{V_{it+1} > F_{it+1}} = 1] &= V_{it} \exp\{\mu_{it}\} \frac{\Phi\left(\frac{\ln\left(\frac{V_{it}}{F_{it+1}}\right) + \left(\mu_{it} - \frac{\sigma_{it}^2}{2}\right) + \sigma_{it}^2}{\sigma_{it}}\right)}{1 - \Phi\left(\frac{\ln\left(\frac{F_{it+1}}{V_{it}}\right) - \left(\mu_{it} - \frac{\sigma_{it}^2}{2}\right)}{\sigma_{it}}\right)} \\
&= V_{it} \exp\{\mu_{it}\} \frac{\Phi\left(DD_{it} + \sigma_{it}\right)}{1 - \pi_{it}}
\end{aligned}$$

Inserting this in the equation for the expected terminal equity value, we get:

$$\begin{aligned}\mathrm{E}[E_{it+1}] &= (1-\pi_{it})\left(V_{it}\exp\{\mu_{it}\}\frac{\Phi\left(DD_{it}+\sigma_{it}\right)}{1-\pi_{it}} - F_{it+1}\right)\\ &= V_{it}\exp\{\mu_{it}\}\Phi\left(DD_{it}+\sigma_{it}\right) - (1-\pi_{it})F_{it+1}\end{aligned}$$

Finally, divide by E_{it} and subtract 1 to get the expected equity return:

$$\mathrm{E}[r_{it+1}] = \frac{V_{it}}{E_{it}}\exp\{\mu_{it}\}\Phi\left(DD_{it}+\sigma_{it}\right) - (1-\pi_{it})\frac{F_{it+1}}{E_{it+1}} - 1$$

A.2 Equity Return Covariance

The covariance between the terminal equity of firm i and j is:

$$\mathrm{Cov}[E_{it+1}, E_{jt+1}] = \mathrm{E}[E_{it+1}E_{jt+1}] - \mathrm{E}[E_{it+1}]\mathrm{E}[E_{jt+1}]$$

Since we know the value of $\mathrm{E}[E_{it+1}]$ and $\mathrm{E}[E_{jt+1}]$ from Appendix A.1, we only need to focus on the first term, $\mathrm{E}[E_{it+1}E_{jt+1}]$. From the law of iterated expectation, and the fact that $E_{it+1} = (V_{it+1} - F_{it+1})\mathbb{1}_{V_{it+1}>F_{it+1}}$ and $E_{jt+1} = (V_{jt+1} - F_{jt+1})\mathbb{1}_{V_{jt+1}>F_{jt+1}}$, we can write it as:

$$\begin{aligned}\mathrm{E}[E_{it+1}E_{jt+1}] &= \mathrm{E}\big[\mathrm{E}[(V_{it+1}-F_{it+1})\mathbb{1}_{V_{it+1}>F_{it+1}}(V_{jt+1}-F_{jt+1})\mathbb{1}_{V_{jt+1}>F_{jt+1}}|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 1]\\ &\quad + \mathrm{E}[(V_{it+1}-F_{it+1})\mathbb{1}_{V_{it+1}>F_{it+1}}(V_{jt+1}-F_{jt+1})\mathbb{1}_{V_{jt+1}>F_{jt+1}}|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 0]\big]\\ &= \Pr[\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 1]\mathrm{E}[(V_{it+1}-F_{it+1})(V_{jt+1}-F_{jt+1})|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 1]\\ &\quad + \Pr[\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 0]0\\ &= (1-\pi_{it}-\pi_{jt}+\mathrm{Cov}[\mathbb{1}_{V_{it+1}<F_{it+1}}, \mathbb{1}_{V_{jt+1}<F_{jt+1}}]+\pi_{it}\pi_{jt})\\ &\quad \times \mathrm{E}[E_{it+1}E_{jt+1}|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{it+1}>F_{it+1}}] = 1]\end{aligned}$$

Where the second equality is due to the fact that $(V_{iT}-F_i)\mathbb{1}_{V_{it+1}>F_{it+1}}(V_{jt+1}-F_{jt+1})\mathbb{1}_{V_{jt+1}>F_{jt+1}} = 0$ if either $\mathbb{1}_{V_{it+1}>F_{it+1}} = 0$ or $\mathbb{1}_{V_{jt+1}>F_{jt+1}} = 0$. For the third equality we utilize the fact that the probability of at least one firm defaulting is $\mathrm{E}[\mathbb{1}_{V_{it+1}<F_{it+1}\cup V_{jt+1}<F_{jt+1}}] = \pi_{it}+\pi_{jt} - \mathrm{Cov}[\mathbb{1}_{V_{it+1}<F_{it+1}}, \mathbb{1}_{V_{jt+1}<F_{jt+1}}] - \pi_{it}\pi_{jt}$. Inserting this into the equation for the covariance, we get:

$$\begin{aligned}\mathrm{Cov}[E_{it+1}, E_{jt+1}] &= (1-\pi_{it}-\pi_{jt}+\mathrm{Cov}[\mathbb{1}_{V_{it+1}<F_{it+1}}, \mathbb{1}_{V_{jt+1}<F_{jt+1}}]+\pi_{it}\pi_{jt})\\ &\quad \times \mathrm{E}\big[E_{it+1}E_{jt+1}|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{jt+1}>F_{jt+1}}] = 1\big] - \mathrm{E}[E_{it+1}]\mathrm{E}[E_{jt+1}]\end{aligned}$$

Finally, we divide by $E_{it}E_{jt}$ to obtain the return covariance:

$$\begin{aligned}\mathrm{Cov}[r_{it+1}, r_{jt+1}] = (1 - \pi_{it} - \pi_{jt} + \mathrm{Cov}[\mathbb{1}_{V_{it+1}<F_{it+1}}, \mathbb{1}_{V_{jt+1}<F_{jt+1}}] + \pi_{it}\pi_{jt}) \\ \times \frac{1}{E_{it}E_{jt}}\mathrm{E}\big[E_{it+1}E_{jt+1}|\min[\mathbb{1}_{V_{it+1}>F_{it+1}}, \mathbb{1}_{V_{jt+1}>F_{jt+1}}] = 1\big] \\ - (1 + \mathrm{E}[r_{it+1}])(1 + \mathrm{E}[r_{jt+1}])\end{aligned}$$

B DSM Implementation Details

B.1 The DSM Loss Function

Let the total set of parameters for the DSM be denoted $\boldsymbol{\theta}^{DSM} = [\boldsymbol{\theta}^{\alpha}, \boldsymbol{\theta}^{\beta}, \boldsymbol{\theta}^{\lambda}, \boldsymbol{\theta}^{\epsilon}]$. Then, the optimal set of parameters, $\hat{\boldsymbol{\theta}}^{DSM}$, are found by minimizing some loss function, $L(\cdot, \boldsymbol{\theta}^{DSM})$, over some sample, $\mathcal{S}$:

$$\hat{\boldsymbol{\theta}}^{DSM} = \underset{\boldsymbol{\theta}^{DSM}}{\operatorname{argmin}}\, L(\mathcal{S}; \boldsymbol{\theta}^{DSM})$$

The basis for our loss function will be the negative log likelihood of observing the realized equity returns of the $\mathcal{S}$ sample:

$$L^{LL}(\mathcal{S}; \boldsymbol{\theta}^{DSM}) = -\sum_{(i,t)\in\mathcal{S}} \ln\left(\mathcal{L}^{r}_{it+1}(r_{it+1}; \boldsymbol{\theta}^{DSM})\right)$$

However, as is common in the machine learning literature, a parameter penalty term, $L^{P}(\boldsymbol{\theta}^{DSM})$, is added to avoid overfitting. Here, the parameter penalty is chosen to be the ℓ_1-norm of the parameter vector:

$$L^{P}(\boldsymbol{\theta}^{DSM}) = ||\boldsymbol{\theta}^{DSM}||_1$$

Additionally, since we are estimating the parameters of a structural model, it is possible to implement what can best be described as economically motivated parameter penalties. This could either be outright restrictions or a set of penalties on the size or sign of the parameters. Here, the only economically motivated parameter penalty is a penalty on the variance of the

ℓ_2-norm of $\boldsymbol{\lambda}_t$:[20]

$$L^E(\boldsymbol{\lambda}_t|\mathcal{S}) = \text{Var}_{\mathcal{S}}\big[||\boldsymbol{\lambda}_t||_2\big]$$

Where the $\mathcal{S}$ subscript indicates that the variance is calculated over the sample, $\mathcal{S}$. The complete DSM loss function can thereby be expressed as:

$$L(\mathcal{S};\boldsymbol{\theta}^{DSM}) = \sum_{t=1}^{T_{\mathcal{S}}}\sum_{i=1}^{I_t} L^{LL}(\mathcal{S};\boldsymbol{\theta}^{DSM}) + w^P L^P(\boldsymbol{\theta}^{DSM}) + w^E L^E(\boldsymbol{\lambda}_t|\mathcal{S})$$

Where $t \in 1,...,T_{\mathcal{S}}$, with $T_{\mathcal{S}}$ being the number of cross-sections in $\mathcal{S}$, I_t is the number of firms in cross-section t, and w^P and w^E is the weight given to the parameter penalty term and the economically motivated penalty, respectively.

B.2 The DSM Training Procedure

The algorithm used to train the DSM, i.e. minimize the loss function, is the Adam algorithm of Kingma and Ba (2014), with batch sizes of 10,000. The learning rate of the algorithm follows a decaying schedule, in which it is reduced by 5% after each epoch. The initial learning rate is set to 10^{-2}, and the learning rate decay stops when/if it reaches 10^{-4}. In addition to the penalties described in Appendix B.1, the training procedure also involves other regularization techniques. Specifically, the parameter functions, shown in Figure 2, also employ batch normalization (Ioffe and Szegedy (2015)) and dropout (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014)).[21] When training the DSM, the total dataset is split into a training, validation and test dataset. The actual training procedure, outlined above, is done on the training data, while the validation data is used for early stopping and choosing the optimal set of hyperparameters. The early stopping procedure terminates training when the loss function has not improved across the validation data for 10 epochs. The choice of the optimal hyperparameters are determined through a grid search. That is, for the two hyperparameters of the model, w^P and w^E, the DSM is trained for all

[20] This particular penalty seem to be important to avoid overfitting on the training data. Other economically motivated parameter penalties have been tested, such as a penalizing high values of the mispricing term, α_{it}, encouraging a positive risk compensation, $\boldsymbol{\beta}_{it}^T\boldsymbol{\lambda}_t$, etc. Yet, they all lead to a worse performance score on the validation and test data.

[21] The dropout rate could be treated as a hyperparameter but is fixed in this paper to 0.5 for the sake of simplicity and computational costs.

combinations of $w^P \in 10^{-3}, 10^{-4}, 10^{-5}$ and $w^E \in 10^{-5}, 10^{-6}, 0$, and the set of hyperparameters with the lowest loss function value, across the validation data, is chosen. Finally, the optimal solution is sensitive to the choice of starting parameters and so 10 DSMs are trained and the final model output is the average across the 10 models. For the neural network benchmark model in Section 3.3, the training procedure is identical to that of the DSM, with two changes: The loss function is the mean squared error, $MSE = (r_{it+1} - \hat{r}_{it+1})^2$, and the hyperparameter grid search only involves searching across $w^P \in 10^{-3}, 10^{-4}, 10^{-5}$.

All models are trained on a rolling one year basis. After training the models on the initial training and validation set, predictions are made for all months in the first year of the test set. Then, the validation set is rolled forward one year, such that the training data grows by one year, the validation data has the same length in years, and the test data shrinks by one year. The models are then re-estimated using this new training and validation split and predictions are made on the first 12 months of the new test data. This process continues until all observations of the *initial* test data has a set of model predictions.

C Data Preprocessing

As is standard in the financial machine learning literature, the 153 firm characteristics are ranked at each cross-section and then transformed to lie in the $[-1, 1]$ range. The 45 macroeconomic variables are discretized based on the combined training and validation data and then, similarly to the firm characteristics, transformed to lie in the $[-1, 1]$ range. That is, for each macroeconomic variable, all observations are assigned a value between one and ten, based on a decile sort of the combined training and validation data, meaning that no information in the test data is used for the discretization process. Then, each macroeconomic variables is squeezed into the $[-1, 1]$ range. Both the equity returns and the monthly "realized" variances in Section 3.4 have been winsorized at the 99.99th percentile, using information from the entire data set, to exclude the most extreme outliers. Finally, the Compustat values for short-term debt, F^{SD}_{it+1}, and long-term debt, F^{LD}_{it+1}, have been modified to minimize the effect of extremely levered firms:

1. All negative values of F^{SD}_{it+1} and F^{LD}_{it+1} are set to zero.

2. A simple short-term and long-term leverage value is calculated as $L_{it+1}^{SD} = \frac{F_{it+1}^{SD}}{E_{it}}$ and $L_{it+1}^{LD} = \frac{F_{it+1}^{LD}}{E_{it}}$, respectively.

3. Both L_{it+1}^{SD} and L_{it+1}^{LD} are cross-sectionally winsorized at the 98th percentile.

4. The final values for F_{it+1}^{SD} and F_{it+1}^{LD} are then calculated based on the winsorized leverages, i.e. $F_{it+1}^{SD} = E_{it}L_{it+1}^{SD}$ and $F_{it+1}^{LD} = E_{it}L_{it+1}^{LD}$.