Calibration of credit default probabilities in discrete default intensity and logit models

Anand Deo, Sandeep Juneja*

Tata Institute of Fundamental Research

Aakash Kalyani

June 17, 2017

Abstract

Discrete default intensity based or logit type models are commonly used as reduced form models for conditional default probabilities for corporate loans where this default probability depends upon macroeconomic as well as firm-specific covariates. Typically, maximum likelihood (ML) methods are used to estimate the parameters associated with these models. Since defaults are rare, a large amount of data is needed for this estimation resulting in a computationally time consuming optimization. In this paper, we observe that since the defaults are typically rare, say, on average $1-2\%$ per annum, under the Gaussian assumption on covariates (which may be achieved via transforming them), the first order equations from ML estimation suggest a simple, accurate and intuitively appealing *closed form estimator* of the underlying parameters. To gain further insights, we analyze the properties of the proposed estimator as well as the ML estimator in a statistical asymptotic regime where the conditional probabilities decrease to zero, the number of firms as well as the data availability time period increases to infinity. The covariates are assumed to evolve as a stationary Gaussian process. We characterize the dependence of the mean square error of the estimator on the number of firms as well as time period of available data. Our conclusion, validated by numerical analysis, is that when the underlying model is correctly specified, the proposed estimator is typically similar or only slightly worse than the ML estimator. Importantly however, since usually any model is misspecified due to hidden factor(s), then both the proposed and the ML estimator are equally good *or equally bad!* Further, in this setting, beyond a point, both are more-or-less insensitive to increase in data, in number of firms and in time periods of available data. This suggests that gathering excessive expensive data may add little value to model calibration. The proposed approximations should also have applications outside finance where logit type models are used and probabilities of interest are small.

1 Introduction

Calibration of credit risk models to predict firm default probability is of great practical importance in financial credit risk and has generated considerable academic literature over the past fifty years (see, e.g., Altman 1968, Merton 1974, Ohlson 1980, Zmijewski 1984, Shumway 2001, Chava and

*This work was initiated and substantially carried out when the second author was working as an adjunct with CAFRAL (Centre for Advanced Financial Research and Learning), a research wing of Reserve Bank of India, and the third author was a research associate there.

Jarrow 2004, Giesecke and Goldberg 2004, Giesecke, Longstaff, Schaefer, Strebulaev 2011). In addition, one looks to develop a modelling regime that accurately models the probabilities of joint defaults of many firms dynamically as a function of time - this becomes particularly useful in measuring dynamic evolution of portfolio credit risk exposure of financial institutions (see, e.g., Duffie, Saita and Wang 2007, Bharath and Shumway 2008 , Duan, Sun and Wang 2012, Eymen, Giesecke, and Goldberg 2010, Juneja 2017).

In this paper we revisit the well studied problem of estimating the conditional default probability of a firm as a function of common macroeconomy dependent and firm specific covariates. Traditionally, dynamic evolution of risk is modelled using doubly stochastic continuous time processes. Default intensity of each firm is modelled as a function of continuous time stochastic covariates. These covariates may be a solution to a stochastic differential equation. Popular covariates include prevalent treasury rates, trailing stock index return, distance to default of each firm and some firm-specific financial ratios. Dependence between obligors is captured by allowing the underlying covariates to evolve in a dependent manner. Conditioned on the realized default intensities, the firm default times are assumed to occur independently, each as the first event of a non-homogeneous Poisson process whose intensity corresponds to the realized default intensity of the firm.

In this paper, we work in a similar doubly stochastic framework, the only difference being that we model the covariates as well as the corresponding obligor default intensities as discrete time stochastic processes (as in Duan, Sun and Wang 2012, Duan and Fulop 2013). The benefits are that discrete time processes are usually simpler to analyze. Even continuous time models are typically first discretized both for calibration purposes as well as for simulating sample paths, so this too makes analysis of discrete time models important (see, e.g., Duffie, Saita and Wang 2007, Bharath and Shumway 2008).

In discrete time, with time restricted to integers, typically, one expresses conditional default probabilities at time t as some function $f(\beta^{\mathsf{T}} X_t)$, where the underlying covariates are denoted by a random vector X_t, and β denotes coefficients calibrated from data. An econometric model may be fit to explain the time dynamics of the covariates process X_t. Typically separately, the coefficients β are estimated from data often using the maximum likelihood (ML) estimation method. With this in place, one has a dynamic model of portfolio credit risk. The actual framework maybe more nuanced - the covariates may be firm dependent, and the coefficients may be a function of the grouping or a class to which the firm belongs.

Standard methods to estimate parameters, such as β above, involve solving a complex optimization problem using non linear optimization methods or even sequential Monte Carlo methods (see, Duffie et. al. 2007, Duan et. al. 2012, Duan and Fulop 2013). These can be extremely time consuming given the huge datasets that are often used for calibration purposes. These problems are also difficult since the defaults are rare events and sufficient amount of data is needed to contain enough defaults to allow for accurate calibration. Further, these computational procedures provide little insight on the underlying factors that determine these parameters.

It is this parameter estimation problem that we address in this paper for a few popular classes of conditional default probability models. For presentation ease, we restrict times in our discrete time models to integers. Consider a popular discretised default intensity based model, for a firm surviving till time $t \in Z^+$, the non-negative hazard rate intensity within times $[t, t+1)$ is assumed to have the form $\exp(\beta^T X_t - \alpha)$ where X_t denotes the value of the covariates at time t, while β and α are parameters that may be estimated from data. This implies that the conditional probability of default between times t and $t+1$ has the form

$$1 - \exp(-\exp(\beta^T X_t - \alpha)) \tag{1.1}$$

(see, e.g., Duffie et. al. 2007, Bharath and Shumway 2008, Duan et. al. 2012, Duan and Fulop

2013).

Logit type models are another class of popular models to which our analysis is applicable. Here, the conditional default probabilities have the form

$$\frac{\exp(\beta^\intercal X_t - \alpha)}{1 + \exp(\beta^\intercal X_t - \alpha)} \tag{1.2}$$

(see, e.g., Shumway 2001, Chava and Jarrow 2004).

We develop *closed form* approximate expressions for estimated parameters. In particular, we show that *each parameter maybe approximated by the weighted average of the corresponding covariate observed just before default occurrences.* In typical calibration settings, for instance when the default probabilities in a small period (say, a month or a quarter) are of order one in a thousand, when the number of firms observed is less than a few thousand, if the data generating model is correctly specified (that is, the structure of the conditional default probability that we assume is indeed the true structure that governs the default generation) we observe analytically as well as numerically that our estimator has similar root mean square error compared to the ML estimator. When the number of firms increases beyond a few thousands, our estimator is slightly worse compared to the ML estimator. Similarly, when the default probabilities increase to say five per thousand in a small time period, the accuracy of the proposed estimator somewhat worsens.

However, typically, the model that we assume for generating default probabilities may be misspecified - it may lack some hidden co-variates (see, e.g., Duffie, Eckner, Horel and Saita 2009), or it may be structurally misrepresentative, or very likely, both. In such settings, we observe that the proposed approximations are equally good (or equally bad!) compared to ML estimators. This is also borne by performance of the proposed estimator on real data where it is seen to be as accurate as ML methods in predicting defaults. The proposed closed form expressions also provide crisp, and intuitively appealing insights into factors that influence the estimated parameters.

To gain further insights into the performance of the proposed estimator, we embed the calibration problem in a sequence of correctly specified statistical models indexed by the rarity of the underlying defaults. The resulting asymptotic analysis sheds further light into the accuracy of proposed approximations and the amount of samples needed for accurate estimation as a function of the rarity of the underlying default probabilities.

In a simple set up, we also compare the proposed estimator with the ML estimator in our asymptotic regime. Our key observations are that when the model is misspecified even by a small amount, the misspecification error dominates other errors so that the proposed estimator and the ML estimator perform similarly. In particular, increasing data both in terms of the number of firms and the time periods considered, beyond a point, lead to virtually no improvement in the estimator quality. This may have ramifications both in search for more accurate models as well as on cost-benefit tradeoffs in gathering data for model calibration.

Our analysis relies on two straightforward observations:

1. In practice, when the covariates have substantially non-Gaussian marginal distributions, it is easy to find simple functional transformations so that the resulting random variables have distributions that are well modelled as multi-variate Gaussian.

2. Under the assumption that the covariates are multi-variate Gaussian, and $\exp(\beta^T X - \alpha)$ is small, expectations such as

$$EH(X)\frac{\exp(\beta^T X - \alpha)}{1 + \exp(\beta^T X - \alpha)}$$

are well approximated by

$$EH(X)\exp(\beta^T X - \alpha). \tag{1.3}$$

When $H(X)$ is a simple polynomial in its components, then (1.3) has a closed form solution.

This then provides a closed form expression for the dominant term in the solution of resulting first order equation in the maximum likelihood method.

It can be shown (this discussion will be added in an updated version) that when the underlying coefficients are small for some covariates, that is, the conditional default probabilities are less sensitive to some covariates, even if these covariates do not have a Gaussian distribution, a simple Taylor series based analysis supports the approximations that we propose.

The proposed approximations should have wider utility to applications where logit type conditional probability models are used, and associated conditional probabilities typically take small values. See for instance Milton, Shankar and Mannering (2008) and references therein for related literature in transportation safety, and Bagley, White and Golomb (2001) and references therein for related literature in medical community.

The structure of the remaining paper is as follows: In Section 2, we first discuss how the maximum likelihood estimators are arrived at in a set-up that includes the two popular regimes: When the conditional probabilities have the forms (1.1) and (1.2). We also identify the proposed estimator suggested by these equations under the assumption that covariates have a multivariate Gaussian distribution. In Section 3, we introduce the mathematical framework including a statistical asymptotic regime under which we conduct our analysis. In Section 4, we discuss the proposed estimators, and conduct their asymptotic performance analysis. We also discuss the performance of the ML estimators under correct and misspecified models, in this section. Numerical results based on simulation generated default data are presented in Section 5. In Section 6, we compare the proposed estimator with the maximum likelihood estimator on a small sample of US corporate default data. It is well known in corporate default literature that defaults tend to cluster displaying a contagion effect. We also observe here that appropriately including contagion effect as a factor improves the performance of the proposed estimator. In Section 7 we end with a brief conclusion. Proofs except for the simplest ones, are all kept in the Appendix.

2 Calibration via the maximum likelihood method

We first discuss the maximum likelihood (ML) estimation methodology for estimating parameters in the models suggested by (1.1) and (1.2).

Suppose that the data available involves m firms, observed over discrete set of time periods $\{0, 1, \ldots, T\}$. Further:

- For integer $d_1 \geq 1$, $(y_t \in \Re^{d_1} : t = 0, 1, \ldots, T)$ denotes the value of the common factors. These may include prevailing interest rates, stock market returns over the last some months, etc.

- We suppose that Firm i it came into existence at time $s_i \leq T - 1$ with $s_i = 0$ if it already exists at time 0. Let τ_i denote its default time if it defaults by time T. Specifically, set $\tau_i = t$ if the firm defaults between time periods t and $t + 1$. Else, if the firm does not default by time T, set $\tau_i = \infty$. Let $f_i = \min(\tau_i, T)$.

- For integer $d_2 \geq 1$, let $(x_{i,t} \in \Re^{d_2} : s_i \leq t \leq f_i - 1)$ denote firm specific information available. This may correspond to firm's estimated distance to default, its size, its net income to total asset ratio. Let $(d_{i,t} : s_i < t \le f_i)$ indicate when the firm defaults. Thus, $d_{i,t} = 1$ if the firm defaults between time $t-1$ and t. Else, it equals 0.

In this framework, let $p(y_t, x_{i,t})$ denote the conditional probability that a Firm i, surviving at time t, defaults between time t and $t+1$. This is assumed to be a function of $(y_t, x_{i,t})$ given $(y_s, x_{j,s} : s \le t, j \le m)$. This probability also depends upon underlying parameters that need to be estimated from data.

Then, the likelihood, call it $\mathcal{L}$ of seeing the default data $(d_{i,t} : s_i < t \le f_i)$ for each $i \le m$, is given by

$$\mathcal{L} = \prod_{i \le m} \prod_{t=s_i}^{f_i - 1} \left(p(y_t, x_{i,t})^{d_{i,t+1}} (1 - p(y_t, x_{i,t}))^{1-d_{i,t+1}} \right) \tag{2.1}$$

ML estimation of underlying parameters then corresponds to finding parameters that maximize $\mathcal{L}$, or equivalently, $\log \mathcal{L}$.

Consider the specific setting of the default intensity model where

$$p(y_t, x_{i,t}) = 1 - \exp(-e^{\beta^\intercal y_t + \eta^\intercal x_{i,t} - \alpha})$$

for parameters $\beta \in \Re^{d_1}$, $\eta \in \Re^{d_2}$ and $\alpha \in \Re$. For notational convenience, let $v_{i,t} = (y_t, x_{i,t}) \in \Re^{d_1+d_2}$, and $\tilde{\beta} = (\beta, \eta)$.

Setting the partial derivatives with each component of $\tilde{\beta}$ and α to zero, the following relations are well known and easily derived: Component wise,

$$\sum_{i \le m, t=s_i}^{f_i - 1} \frac{v_{i,t} e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}}}{1 - \exp(-e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}})} d_{i,t+1} = \sum_{i \le m, t=s_i}^{f_i - 1} v_{i,t} e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}} \tag{2.2}$$

and

$$\sum_{i \le m, t=s_i}^{f_i - 1} \frac{e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}}}{1 - \exp(-e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}})} d_{i,t+1} = \sum_{i \le m, t=s_i}^{f_i - 1} e^{\hat{\beta}^\intercal v_{i,t} - \hat{\alpha}}, \tag{2.3}$$

where $\hat{\beta}$ and $\hat{\alpha}$ are a solution to these equations.

Similarly, when the conditional default probabilities have the logit structure,

$$p(v_{i,t}) = \frac{\exp(\tilde{\beta}^\intercal v_{i,t} - \alpha)}{1 + \exp(\tilde{\beta}^\intercal v_{i,t} - \alpha)}.$$

As is well known, in this case the function $\log \mathcal{L}$ is a convex function of underlying parameters $(\tilde{\beta}, \alpha)$. Setting the partial derivatives with these parameters to zero, we get, component wise, the well known results

$$\sum_{i \le m, t=s_i}^{f_i - 1} v_{i,t} d_{i,t+1} = \sum_{i \le m, t=s_i}^{f_i - 1} v_{i,t} \frac{\exp(\hat{\beta}^\intercal v_{i,t} - \hat{\alpha})}{1 + \exp(\hat{\beta}^\intercal v_{i,t} - \hat{\alpha})}. \tag{2.4}$$

and

$$\sum_{i \le m, t=s_i}^{f_i - 1} d_{i,t+1} = \sum_{i \le m, t=s_i}^{f_i - 1} \frac{\exp(\hat{\beta}^\intercal v_{i,t} - \hat{\alpha})}{1 + \exp(\hat{\beta}^\intercal v_{i,t} - \hat{\alpha})}, \tag{2.5}$$

where the LHS is simply the number of defaults observed during the periods of observation, and again where $\hat{\beta}$ and $\hat{\alpha}$ are a solution to these equations.

2.1 Gaussian approximations

Let $\tau = \sum_{i\le m}(f_i - s_i)$ denote the firm periods of data available.

In our analysis we assume that $\{y_t\}$ for $0 \le t \le T$ and $x_{i,t} : i \le m$ for surviving firms, are realizations of a stationary process $\{(Y_t, (X_{i,t}, i \le m))\}$ (also denoted by $\{(V_{i,t}, i \le m)\}$) observed at integer times $0 \le t < T$, which is further assumed to be multivariate Gaussian. As discussed in the introduction, the covariates originally may not be Gaussian, but we assume that they are suitably transformed to have a Gaussian marginal distribution. More specifically, the transformed variables form a stationary multivariate Gaussian process where each marginal has been normalized to have stationary mean zero and variance one.

When the defaults occur with small probabilities, we may approximate RHS of (2.3) and (2.5) divided by τ by

$$E\left(\exp(\tilde{\beta}^\intercal V_{i,t} - \alpha)\right),$$

which, as is well known, equals

$$\exp\left(\frac{1}{2}\tilde{\beta}^\intercal \Sigma \tilde{\beta} - \alpha\right),$$

where Σ denotes the correlation matrix of $(V_{i,t})$ and is assumed to be independent of i.

Similarly, the RHS of (2.2) and (2.4) divided by τ may be approximated by

$$E(V_{i,t}\exp(\tilde{\beta}^\intercal V_{i,t} - \alpha)).$$

This equals

$$\Sigma\tilde{\beta}\exp\left(\frac{1}{2}\tilde{\beta}^\intercal \Sigma \tilde{\beta} - \alpha\right).$$

Assume that Σ is known and invertible. Then, the above discussion suggests that

$$\tilde{\beta} \approx \Sigma^{-1}\frac{\sum_{i\le m, t=s_i}^{f_i-1} v_{i,t}d_{i,t+1}}{\sum_{i\le m, t=s_i}^{f_i-1} d_{i,t+1}}. \tag{2.6}$$

Then, the RHS above, call it $\hat{\beta}$, is our proposed estimator for $\tilde{\beta}$. Motivated by (2.3) and (2.5), our estimator for α, $\hat{\alpha}$, is set as

$$\hat{\alpha} \triangleq \log\left(\frac{\sum_{i\le m, t=s_i}^{f_i-1}\exp(\hat{\beta}^\intercal v_{i,t})}{\sum_{i\le m, t=s_i}^{f_i-1} d_{i,t+1}}\right). \tag{2.7}$$

Remark 1. Let the weighted average of the covariates observed before defaults $\frac{\sum_{i\le m, t=s_i}^{f_i-1} v_{i,t}d_{i,t+1}}{\sum_{i\le m, t=s_i}^{f_i-1} d_{i,t+1}}$ be denoted by $\hat{w}$. Then, we have

$$\hat{\beta} = \Sigma^{-1}\hat{w}.$$

Now suppose that Σ^{-1} is not known but is estimated from data by $\hat{\Sigma}^{-1}$. Then, a reasonable estimator of $\tilde{\beta}$ is

$$\breve{\beta} = \hat{\Sigma}^{-1}\hat{w}. \tag{2.8}$$

This may be re-expressed as

$$\Sigma^{-1}\hat{w} + (\hat{\Sigma}^{-1} - \Sigma^{-1})\hat{w}.$$

In our asymptotic analysis later in Section 4, if $\hat{\Sigma}^{-1}$ is a sufficiently accurate estimator of Σ^{-1}, then the contribution to mean square error $E(\breve{\beta} - \tilde{\beta})^2$ due to replacing $\hat{\Sigma}^{-1}$ by Σ^{-1} in (2.8) can be negligible. This, and the need to avoid undue tediousness in analysis, motivates our assumption that the variance covariance matrix Σ is known.

Remark 2. An easy extension of above estimators to multiple K classes is to assume that the parameters $\tilde{\beta}$ are same across all classes and are estimated as above by $\hat{\beta}$ assuming that the data comes from a single class. The parameters $(\alpha_i : i \leq K)$ are allowed to be class dependent and they measure the riskiness of each class. These, for each k, can be estimated as

$$\hat{\alpha}_k \triangleq \log\left(\frac{\sum_{i\leq m_k, t=s_{i,k}}^{f_{i,k}-1} \exp(\hat{\beta}^\intercal u_{i,k,t})}{\sum_{i\leq m, t=s_{i,k}}^{f_{i,k}-1} d_{i,k,t+1}}\right)$$

where the subscript k attached to original notation $(m, s, f_i, d_{i,t})$ denotes that the associated data is class specific.

As mentioned earlier, in the next sections we characterize the performance of the proposed estimators.

3 Mathematical framework

We consider m firms observed over discrete set of time periods $t = 0, 1, 2, \ldots, T$. For simplicity, we assume that all firms belong to the same class, or that they are statistically homogeneous.

Recall that $Y_t \in \Re^{d_1}$ denotes the vector of common market information at time t, and for firm $i \leq m$, $X_{i,t} \in \Re^{d_2}$ denotes a vector of company specific information at time t, and $V_{i,t} = (Y_t, X_{i,t})$. Further, $(V_{i,t} : i \leq m)_{0\leq t\leq T}$ denote a stationary Gaussian process restricted to integer times $0 \leq t \leq T$. We let $(V_i : i \leq m)$ denote the random variables with the associated stationary distribution. These, as indicated earlier, are all assumed to be normalized to have marginal mean zero and variance one.

Let $\Sigma_{YY} \in \Re^{d_1\times d_1}$ denote the correlation matrix corresponding to Y, $\Sigma_{YX} \in \Re^{d_1\times d_2}$ denote the correlation matrix between Y and X_i which we assume to be same for all $i \leq m$ (also denoted by Σ_{XY}^T). Similarly, let $\Sigma_{XX} \in \Re^{d_2\times d_2}$ denote the correlation matrix between the components of X_i again assumed to be same for all i.

Further, let

$$\Sigma \triangleq \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}. \tag{3.1}$$

Thus, Σ denotes the correlation matrix of V_i.

3.1 Asymptotic formulation

Since the conditional default probabilities are typically very small, we analyze the calibration problem in a regime indexed by γ as $\gamma \downarrow 0$. We assume that the stochastic process $\{V_{i,t} : i \leq m\}_{0\leq t\leq T}$ is independent of γ while the parameters $\beta(\gamma) \in \Re^{d_1}$, $\eta(\gamma) \in \Re^{d_2}$ (let $\tilde{\beta}(\gamma)$ denote $(\beta(\gamma), \eta(\gamma))$ and $\alpha(\gamma) \in \Re$ are deterministic functions of γ. Later, when we move on to parameter estimation methodology, we will allow T and each m to increase with γ, although that is not needed at this stage. In addition to $V_{i,t}$, we define a sequence of iid uniform (between 0 and 1) random variables, $U_{i,t}$ independent of $V_{i,t}$'s. A firm i which has survived up to time t defaults between time t and $t+1$ if

$$U_{i,t+1} \leq p(\gamma, V_{i,t}). \tag{3.2}$$

Let $D_{i,t+1} = 1$ if firm $i \leq m(\gamma)$ that survives up to time t in our framework, defaults between times t and $t+1$. Let $\mathcal{F}_t$ denote the sigma algebra generated by $(U_{i,s}, V_{i,s} : s \leq t)$. Further, we define τ_i as

$$\tau_i = \min\{T(\gamma) - 1, \min\{t \geq 0 : D_{i,t+1} = 1\}\} \tag{3.3}$$

Thus, if Firm i defaults between times t and $t+1$, $\tau_i = t$. Then by (3.2)

$$E(D_{i,t+1}|\mathcal{F}_t) = p(\gamma, V_{i,t})\mathbb{I}(\tau_i \geq t). \tag{3.4}$$

We posit that for firm $i \leq m$, the conditional default probability $p(\gamma, V_{i,t})$ of defaulting in period $(t, t+1]$, conditioned on $\mathcal{F}_t$, and it surviving at time t, is small and is given by:

$$p(\gamma, V_{i,t}) = \exp(\tilde{\beta}^{\intercal}(\gamma)V_{i,t} - \alpha(\gamma))(1 + H(\gamma, V_{i,t})), \tag{3.5}$$

where we assume that $H(\gamma, V_{i,t}) \to 0$ as $\gamma \to 0$, almost surely.

Let

$$p(\gamma) \triangleq Ep(\gamma, V_i) = E\left(\exp(\tilde{\beta}^{\intercal}(\gamma)V_i - \alpha(\gamma))(1 + H(\gamma, V_i))\right) \tag{3.6}$$

denote the steady state conditional probability of default of a customer (these as per our assumptions are same for each customer).

We assume that $p(\gamma) \to 0$ as $\gamma \to 0$.

Note that

$$E\left(\exp(\tilde{\beta}^{\intercal}(\gamma)V_i - \alpha(\gamma))\right) = \exp\left(\frac{1}{2}\tilde{\beta}^{\intercal}(\gamma)\Sigma\tilde{\beta}(\gamma) - \alpha(\gamma)\right). \tag{3.7}$$

As suggested earlier, our aim is to develop and exploit the approximation

$$p(\gamma) \sim \exp\left(\frac{1}{2}\tilde{\beta}^{\intercal}(\gamma)\Sigma\tilde{\beta}(\gamma) - \alpha(\gamma)\right).$$

To this end we impose some simple conditions on $\tilde{\beta}(\gamma)$, $\alpha(\gamma)$ and $H(\gamma, V_i)$.

Assumption 1. *Suppose that*

1. $$\alpha(\gamma) = \log(1/\gamma) - \log c$$

   where $c > 0$ is a constant.

2. *The vector $\tilde{\beta}(\gamma)$ is independent of γ and is denoted by $\tilde{\beta} = (\beta, \eta)$.*

3. *Further,*

   $$|H(\gamma, V_i)| \leq C\gamma \exp(\tilde{\beta}^{\intercal}V_i) \tag{3.8}$$

   a.s. for a constant $C > 0$.

Remark 3. One instance of (3.8) is when the conditional default probabilities at any time t have the form

$$1 - \exp\left(-e^{\tilde{\beta}^{\intercal}V_{i,t} - \alpha(\gamma)}\right).$$

Here, since $e^x(1 - e^x/2) \leq 1 - \exp(-e^x) \leq e^x$, it is easily seen that

$$C = \frac{c}{2}$$

satisfies (3.8).

Another instance of (3.8) is when conditional default probabilities at any time t have a logit representation

$$\frac{\exp(\tilde{\beta}^{\intercal}V_{i,t} - \alpha(\gamma))}{1 + \exp(\tilde{\beta}^{\intercal}V_{i,t} - \alpha(\gamma))}.$$

Here, since for each x, $e^x(1-e^x) \leq \frac{e^x}{1+e^x} \leq e^x$, it is easily seen that

$$C = c$$

satisfies (3.8).

Remark 4. It is easy to extend our analysis if we assume more generally that

$$\alpha(\gamma) \sim \log(1/\gamma) - \log c$$

and

$$\tilde{\beta}(\gamma) \sim \log(1/\gamma)^{\psi}\tilde{\beta}$$

for some $\psi \in [0, 1/2)$. However, this leads to increase in tediousness of developed analysis without adding to meaningful insights. Further, the data does not suggest that the parameters associated with the covariates take large values.

As mentioned earlier, the proposed estimators remain effective even when some of the covariates do not have a Gaussian distribution, as long as the corresponding components of $\tilde{\beta}(\gamma) \to 0$ as $\gamma \to 0$.

Let $V_i^{(j)}$ denote the component j of V_i, and Σ_j denotes the row j of matrix Σ. The following observations are easily seen under Assumption 1:

$$p(\gamma) = c\gamma \exp(\frac{1}{2}\tilde{\beta}^{\intercal}\Sigma\tilde{\beta}) + O(\gamma^2). \tag{3.9}$$

Further,

$$E\left(V_i^{(j)} \exp(\tilde{\beta}^{\intercal} V_i - \alpha(\gamma))(1 + H(\gamma, V_i))\right) = \Sigma_j \tilde{\beta} c\gamma \exp(\frac{1}{2}\tilde{\beta}^{\intercal}\Sigma\tilde{\beta}) + O(\gamma^2). \tag{3.10}$$

4 Parameter Estimation

For notational simplicity we assume that all m firms are functional at time 0. We enhance our asymptotic regime by allowing the number of firms, as well as the total time for which the system is observed to increase as γ reduces to zero. Specifically, suppose that the number of firms, $m(\gamma) = \tilde{m}/\gamma^{\delta}$ for $\delta > 0$ and total number of periods $T(\gamma) = \gamma^{-\zeta}$ for $\zeta \in (0, 1)$.

The rationale for this form is data driven: Typical time periods that we have in mind are in months or quarters. The conditional default probabilities then are typically of order 10^{-3}, so one may heuristically view $\gamma \sim 10^{-3}$. We are typically looking at data involving tens of thousands of firms, so $\delta \in [1, 2)$, is reasonable although there may be cases where the data is limited and $\delta \in (0, 1)$ is a better representation. The time period can be in order of tens of years, so $\zeta \in (0, 1)$ appears reasonable.

Note that the expected total number of defaults observed,

$$\sum_{t=0}^{T(\gamma)-1} \sum_{i=1}^{m(\gamma)} ED_{i,t+1} = \Theta(\gamma^{-(\zeta+\delta-1)})$$

and increases to infinity as $\gamma \to 0$ when $\zeta + \delta > 1$. If $\zeta + \delta < 1$, the number of observed defaults converges to zero as $\gamma \to 0$.

Guided by our discussion in Section 2 leading to (2.6), we now develop our parameter estimation methodology.

4.1 Proposed estimators

Consider the following sequence of random variables indexed by γ

$$\hat{D}_\gamma = \frac{1}{\gamma T(\gamma) m(\gamma)} \sum_{t=0}^{T(\gamma)-1} \sum_{i=1}^{m(\gamma)} D_{i,t+1}. \tag{4.1}$$

Similarly, define the sequences of random vectors,

$$\hat{V}_\gamma = \frac{1}{\gamma T(\gamma) m(\gamma)} \sum_{t=0}^{T(\gamma)-1} \sum_{i=1}^{m(\gamma)} V_{i,t} D_{i,t+1}. \tag{4.2}$$

As suggested in (2.6), the proposed estimator for β is

$$\hat{\beta}(\gamma) \triangleq \Sigma^{-1} \times \frac{\hat{V}_\gamma}{\hat{D}_\gamma}. \tag{4.3}$$

Recall that

$$E(D_{i,t+1}|\mathcal{F}_t) = \exp(\tilde{\beta}^\intercal V_{i,t} - \alpha)(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t),$$

and thus,

$$\alpha = \log\left(\frac{\sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E(\exp(\tilde{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t))}{\sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E D_{i,t+1}} \right) \tag{4.4}$$

as $\gamma \to 0$. This motivates our empirical estimator for α,

$$\hat{\alpha}(\gamma) \triangleq \log\left(\frac{\sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} \exp(\hat{\beta}^\intercal(\gamma) V_{i,t})\mathbb{I}(\tau_i \geq t)}{\sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} D_{i,t+1}} \right), \tag{4.5}$$

where $\hat{\beta}(\gamma)$ is the estimator for $\tilde{\beta}$.

4.2 Covariates as a stationary Vector Autoregressive Process

To support further analysis we make a reasonable assumption that the covariates $(V_{i,t} : t \geq 0)$ for each $i \leq m$, follow an order 1 vector autoregressive process (VAR (1)). Specifically, we assume that

$$Y_t = AY_{t-1} + \epsilon_t \tag{4.6}$$

and for each i,

$$X_{i,t} = BY_{t-1} + CX_{i,t-1} + \psi_{i,t} \tag{4.7}$$

where matrix A takes values in $\Re^{d_1 \times d_1}$, matrix B takes values in $\Re^{d_2 \times d_1}$, matrix C takes values in $\Re^{d_2 \times d_2}$, ϵ_t is a vector of i.i.d. Gaussian mean zero variance one, random variables taking values in $\Re^{d_1}$, and $(\psi_{i,t} : i \leq m, t \geq 0)$ is a collection of independent Gaussian vectors taking values in $\Re^{d_2}$, whose components are independent standard (mean zero, variance 1) Gaussian random variables, that are also independent of $(\epsilon_t : t \geq 0)$.

Let

$$M \triangleq \begin{pmatrix} A & 0 \\ B & C \end{pmatrix}. \tag{4.8}$$

Then, (4.6) and (4.7) may be re-expressed as,

$$V_{i,t} = M \cdot V_{i,t-1} + \phi_{i,t}, \tag{4.9}$$

where, $\phi_{i,t} \in \Re^{d_1+d_2}$ is a column vector $(\epsilon_t, \psi_{i,t})^\intercal$. Let $\|A\|$ denote the operator norm for the matrix A, that is,

$$\|A\| = \inf_{r \in \Re^+} (r \colon \|Ax\|_2 \leq r\|x\|_2),$$

where $\|y\|_2$ denotes the Euclidean norm for any vector y. It follows that $\|Ax\|_2 \leq \|A\|\|x\|_2 \; \forall x$. As is well known, the process $(V_{i,t} : t \geq 0)$ has a stationary distribution if $\|M\| < 1$. In that case the stationary distribution is Gaussian in $\Re^{d_1+d_2}$, with componentwise mean zero and variance Σ (notationally denoted by $N(0, \Sigma)$),where Σ is the solution to the matrix equation

$$M\Sigma M^\intercal = \Sigma - \mathbf{I}.$$

Remark 5. It can be shown that $\Sigma = \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal$. The existence of a solution is guaranteed by the fact that $\|M\| < 1$.

In summary, we make the following assumption for $(V_{i,t} : t \geq 0)$ for each i.

Assumption 2. *The covariates $(V_{i,t} : t \geq 0)$ for each i follow the VAR (1) process (4.9). Further, $\|M\| < 1$, and $V_{i,0}$ is distributed as $N(0, \Sigma)$ so that each $(V_{i,t} : t \geq 0)$ is a stationary process.*

4.3 Main results

Recall that $\hat{\beta}(\gamma)$ and $\hat{\alpha}(\gamma)$, respectively, denote the proposed estimators for the true parameters $\tilde{\beta}$ and α. Theorem 4.1 specifies the order of the mean square error of the proposed estimators. Let

$$\|X\|_2 = E\left(\sum_{i=1}^{n} X_i^2\right)^{\frac{1}{2}}$$

denote the $\mathcal{L}_2$ norm for any $\Re^n$ valued random vector $X = (X_1, \ldots, X_n)$.

All the results that follow are under Assumptions 1 and 2.

Theorem 4.1. *The following relations hold:*

1.

$$\|\tilde{\beta} - \hat{\beta}(\gamma)\|_2^2 = O(\gamma^\zeta) + O(\gamma^{\zeta+\delta-1}), \tag{4.10}$$

and

2.

$$\|\alpha - \hat{\alpha}(\gamma)\|_2^2 = O(\gamma^\zeta) + O(\gamma^{\zeta+\delta-1}). \tag{4.11}$$

Remark 6. Theorem 4.1 makes an interesting observation related to the sensitivity of the proposed estimator to the systemic risk. Recall that $\zeta \in (0, 1)$. Observe the obvious fact that if $\delta + \zeta < 1$ then the mean square error increases as $\gamma \to 0$. This is because asymptotically, no defaults are observed in the data. Now consider two regimes

1. $\delta \geq 1$. In this case, further increase in δ does not reduce the mean square error of the parameters. Thus, having more than order γ^{-1} firms does not help in improving accuracy of calibration.

2. $\delta < 1$. In this case, it is clear that increasing δ does reduce the mean square error of the parameters.

Thus, having more firms data is useful up to order γ^{-1}, thereafter its utility to proposed estimator is marginal.

In Theorem 4.2, we find the order of the mean square of relative error of the firm default conditional probability when the covariates have a stationary distribution. This is perhaps a better measure of error vis-a-vis mean square error in estimating parameters, since ultimately our interest is in the error made in forecasting conditional probabilities. Also, given that in our asymptotic regime, default probabilities are decreasing to zero, relative error is a more appropriate measure of estimation error vis-a-vis absolute error.

Theorem 4.2. *Assuming that V_i below is independent of $\hat{\beta}$ and $\hat{\alpha}$,*

$$E\left(\frac{p(\gamma,V_i)-\exp(\hat{\beta}^\intercal(\gamma)V_i-\hat{\alpha}(\gamma))}{p(\gamma,V_i)}\right)^2 = O(\gamma^{\zeta})+O(\gamma^{\zeta+\delta-1}). \tag{4.12}$$

Remark 7. Since $\hat{\beta}$ and $\hat{\alpha}$ are estimated from past data, and these are used to provide conditional probabilities in future, there may be mild dependence between future covariates and these estimated parameters. The assumption above that these are independent is somewhat reasonable. It can be relaxed to allow for mild dependence without affecting the conclusions.

4.4 Maximum likelihood estimator in the proposed asymptotic regime

We provide a brief heuristic explanation for the performance of the maximum likelihood estimator (MLE). Recall that the proposed estimator requires that for $T(\gamma)=\gamma^{-\zeta}$ and number of obligors $m(\gamma)=\tilde{m}\gamma^{-\delta}$, $\delta+\zeta>1$ and $\zeta\in(0,1)$ for the mean square error (MSE) to decrease to zero as $\gamma\to 0$. Increasing the number of obligors so that $\delta>1$, no longer improves the rate of decay of MSE. We highlight here that for MLE even when $T(\gamma)$ is small, e.g., $\zeta=0$, but $\zeta+\delta>1$, the MSE of MLE asymptotically reduces to zero at the rate $\gamma^{\zeta+\delta-1}$. This difference is corroborated by the numerical experiments. Rigorous analysis will appear in a more elaborate version of this report.

Consider the first order conditions for MLE in our asymptotic framework when the underlying model is either logit or default intensity based. Further, for simplicity, assume that $\alpha(\gamma)$ is known. Recall that

$$\tau_i=\min\{T(\gamma)-1,\min\{t\geq 0: D_{i,t+1}=1\}\}.$$

Also, $\tau_i\sim T(\gamma)$ as $\gamma\to 0$. In both the cases, we have

$$\sum_{i\leq m(\gamma),t\leq\tau_i} V_{i,t}D_{i,t+1} = \left[\sum_{i\leq m(\gamma),t\leq\tau_i} V_{i,t}\exp(\hat{\beta}^\intercal V_{i,t}-\alpha(\gamma))\right](1+O(\gamma)), \tag{4.13}$$

Expanding the RHS in (4.13) using Taylor's expansion around $\tilde{\beta}$, dividing both sides by $\gamma\, m(\gamma)\, T(\gamma)$, and observing that along the set $\tau_i\geq t$,

$$E(D_{i,t+1}|\mathcal{F}_t)=\exp(\tilde{\beta}^\intercal V_{i,t}-\alpha(\gamma))(1+O(\gamma)),$$

we get

$$\frac{1}{\gamma m(\gamma)T(\gamma)}\sum_{i\leq m(\gamma),t\leq\tau_i} V_{i,t}\left(D_{i,t+1}-E(D_{i,t+1}|\mathcal{F}_t)\right) = \frac{1}{\gamma m(\gamma)T(\gamma)}(\hat{\beta}-\tilde{\beta})^\intercal \sum_{i\leq m(\gamma),t\leq\tau_i} V_{i,t}^2\exp(\tilde{\beta}^\intercal V_{i,t}-\alpha(\gamma)) \tag{4.14}$$

plus remainder terms, where the vector $V_{i,t}^2$ denotes $(V_{i,t}^2(j) : 1 \leq j \leq d_1 + d_2)$, that is, the vector obtained by squaring each component of $V_{i,t}$.

Since, the terms $(D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t))$ are zero mean random variables uncorrelated for each t and i, with variance of order γ (as $\gamma \to 0$), it is easy to see that the expectation of the square of LHS of (4.14) converges to zero if $\gamma m(\gamma)T(\gamma) \to 0$, or in our set up, when $\zeta + \delta > 1$. In particular, even when $\zeta = 0$, for $\delta > 1$, this term decreases to zero as $\gamma \to 0$.

To complete our argument, observe that

$$\frac{1}{m(\gamma)T(\gamma)} \sum_{i \leq m, t \leq \tau_i} V_{i,t}^2 \exp(\tilde{\beta}^\intercal V_{i,t})$$

converges as $\gamma \to 0$. (recall that $\tau_i \sim T(\gamma)$). If $T(\gamma) = O(1)$ $(\zeta = 0)$, and $m(\gamma) \to \infty$ $(\delta > 0)$, this converges to a random vector. If both $T(\gamma)$ and $m(\gamma)$ increase to infinity, this converges to a constant. In either case, heuristically, we see that mean square error of $\hat{\beta} - \tilde{\beta}$ converges to zero.

The above discussion also suggests that $\hat{\beta} - \tilde{\beta}$ suitably normalized has a non-trivial limiting distribution, that may be useful in constructing confidence intervals. We explore this as well in our ongoing research.

4.5 Performance of MLE under model misspecification

In this section, we heuristically examine the effect of model misspecification on the MLE as well as our proposed estimator in a simple illustrative setting. We capture misspecification by assuming that the model generating defaults has two Gaussian factors common to all obligors, while the modeller assumes that only one of the two factors exists; the other is hidden or latent. Further, the model in both cases is assumed to be logit for simplicity. Let $(Y_{1,t}, Y_{2,t} : 1 \leq t \leq T(\gamma))$ denote the time series corresponding to the two factors. Further assume that $(Y_{1,t}, Y_{2,t})$ have a stationary distribution under which random variables $Y_{1,t}$ and $Y_{2,t}$ are assumed to have zero mean, variance 1 and correlation ρ amongst them. Let $(\beta_1, \beta_2, \alpha(\gamma))$ denote the parameters of default generation, where as before $\alpha(\gamma) = \log(1/\gamma) - \log c$. Thus, the default probability of obligor $i \leq m(\gamma)$ defaulting at time $t + 1 \leq T(\gamma)$ is given by

$$\frac{\exp(\beta_1 Y_{1,t} + \beta_2 Y_{2,t} - \alpha(\gamma))}{1 + \exp(\beta_1 Y_{1,t} + \beta_2 Y_{2,t} - \alpha(\gamma))}.$$

Suppose it is thought (wrongly) that only the first factor with time series $(Y_{1,t} : 1 \leq t \leq T(\gamma))$ impact the conditional default probabilities of obligors. The parameters $(\hat{\beta}_1, \hat{\alpha})$ are then estimated from the first order MLE equations:

$$\frac{1}{\gamma m(\gamma)T(\gamma)} \sum_{i \leq m(\gamma), t \leq \tau_i} Y_{1,t} D_{i,t+1} = \frac{1}{\gamma m(\gamma)T(\gamma)} \sum_{i \leq m(\gamma), t \leq \tau_i} Y_{1,t} \frac{\exp(\hat{\beta}_1 Y_{i,t} - \hat{\alpha})}{1 + \exp(\hat{\beta}_1 Y_{i,t} - \hat{\alpha})}, \tag{4.15}$$

and

$$\frac{1}{\gamma m(\gamma)T(\gamma)} \sum_{i \leq m(\gamma), t \leq \tau_i} D_{i,t+1} = \frac{1}{\gamma m(\gamma)T(\gamma)} \sum_{i \leq m(\gamma), t \leq \tau_i} \frac{\exp(\hat{\beta}_1 Y_{i,t} - \hat{\alpha})}{1 + \exp(\hat{\beta}_1 Y_{i,t} - \hat{\alpha})}. \tag{4.16}$$

When $\gamma m(\gamma)T(\gamma) \to \infty$, the LHS in (4.15) converges to

$$cEY_{1,t} \exp(\beta_1 Y_{1,t} + \beta_2 Y_{2,t}) = c(\beta_1 + \rho\beta_2) \exp\left(\frac{1}{2}(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2)\right),$$

and the RHS in (4.15) is asymptotically similar (as $\gamma \to 0$) to

$$\frac{1}{\gamma} E Y_{1,t} \exp(\hat{\beta}_1 Y_{1,t} - \hat{\alpha}) = \frac{1}{\gamma} \hat{\beta}_1 \exp\left(\frac{1}{2}\hat{\beta}_1^2 - \hat{\alpha}\right).$$

Similarly, the LHS in (4.16) converges to

$$E \exp(\beta_1 Y_{1,t} + \beta_2 Y_{2,t}) = c \exp\left(\frac{1}{2}(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2)\right),$$

and the RHS in (4.16) is asymptotically similar (as $\gamma \to 0$) to

$$\frac{1}{\gamma} E \exp(\hat{\beta}_1 Y_{1,t} - \hat{\alpha}) = \frac{1}{\gamma} \exp\left(\frac{1}{2}\hat{\beta}_1^2 - \hat{\alpha}\right).$$

Equating for parameters, it is easily seen that

$$\hat{\beta}_1 = \beta_1 + \rho\beta_2$$

and

$$\hat{\alpha} = \alpha(\gamma) - \frac{\beta_2^2(1-\rho^2)}{2}.$$

It can be similarly seen that the proposed estimator under the same assumptions also converges to these values. The upshot is that while the MLE and the proposed estimator converge to the same value under this model misspecification, they both are equally wrong in the limit. Thus, practitioner may as well use the simpler proposed estimator. These observations are validated by our numerical experiments in Section 6.

5 Numerical Experiments

In this section we use simulation to generate default data and on this data compare the estimated root mean square error (RMSE) of the proposed estimator of the underlying parameters with the estimated RMSE of their MLE. The default generating model comprises three Gaussian distributed factors for each firm, two common and one firm specific. For the most part the generating model is logit, towards the end of this section, we also provide comparisons when the data is generated using the default intensity model. We compare the two estimators when we know the underlying factors that are used to generate the data as well as when one of the factors is not known. The former corresponds to correctly specified model while the latter is an example of a misspecified model.

We first consider the case where default probabilities are about 1% per annum (one twelfth of that per each discrete time period). Our broad conclusions are that when the model is correctly specified, our estimator is close in accuracy to the MLE when the number of firms is around 3,000 or less. Consistent with the theory, MLE performs relatively better when the number of firms increases as well as when the number of time periods increase, although even for large time periods (corresponding to sixty years of data), when the number of firms is not too large, our estimator has only slightly larger RMSE compared to MLE.

We also consider the case where default probabilities are of order 5% and observe MLE performs better than the proposed estimator, although in all cases, the RMSE of both the estimators is quite small.

When the underlying model is misspecified, as when one of he common factors is latent to the calibrator, we observe that both our proposed estimator as well as MLE have more or less identical RMSE even for large number of firms and long time periods of data availability, and for small and large values of annual default probabilities.

5.1 Setup

$(Y_{1,t}, Y_{2,t} : t \leq T)$ denote the underlying Gaussian factors while $(X_{i,t} : i \leq m, t \leq T)$ denote the idiosyncratic factors. These are generated as follows:

$$\begin{aligned} Y_{1,t} &= 0.3 \cdot Y_{1,t-1} + N_{1,t}(0,1) \\ Y_{2,t} &= 0.3 \cdot Y_{2,t-1} + N_{2,t}(0,1) \\ X_{i,t} &= N_{i+2,t}(0,1) \end{aligned}$$

where $N_{1,t}(0,1)$, $N_{2,t}(0,1)$ and $(N_{i,t}(0,1) : 3 \leq i \leq m+2)$ for each t are i.i.d. mean zero variance 1 standard Gaussian variables. $(Y_{1,0}, Y_{2,0})$ are generated from the stationary distribution, that is

$$Y_{1,0} \sim N\left(0, \frac{1}{1-0.3^2}\right)$$

and

$$Y_{2,0} \sim N\left(0, \frac{1}{1-0.3^2}\right).$$

The conditional default probability for firm i at time t for default at time $t+1$ is set to

$$P(D_{i,t+1} = 1 | X_{i,t}, Y_{1,t}, Y_{2,t}) = \frac{\exp(\beta_1 X_{i,t} + \beta_2 Y_{1,t} + \beta_3 Y_{2,t} - \alpha)}{1 + \exp(\beta_1 X_{i,t} + \beta_2 Y_{1,t} + \beta_3 Y_{2,t} - \alpha)}$$

and defaults are generated by comparing independently generated standard uniformly distributed random variables with conditional default probabilities. Initially, parameters $(\beta_1, \beta_2, \beta_3, \alpha)$ are selected so that average default probability is about 1%.

First consider case where the calibrator is aware of the three underlying factors and estimates $(\beta_1, \beta_2, \beta_3, \alpha)$ from the generated default data. In that case, in each experiment, we generate the default data for various values of m and T, arrive at estimators for the parameters $(\beta_1, \beta_2, \beta_3, \alpha)$ using the two approaches. These experiments are repeated 500 times and the RMSE is estimated under the two approaches for $(\beta_1, \beta_2, \beta_3)$ as the square root of the average of the square of the sum of the discrepancy in estimated values with the true values. These estimated values are referred to as RMSE(β_{prop}) and RMSE(β_{ML}) under the two methods. Similarly, the errors associated with estimators for α are referred to as RMSE(α_{prop}) and RMSE(α_{ML}).

The experiments are conducted in two sets: In the first set, we let the number of firms m vary from 1000 to 10,000. We let the number of time periods $T = 200$. In the second set, T varies from 100 to 700, and $m = 2,000$.

The results are reported in Table 1. These experiments are repeated when the true parameters are selected so that average default probability is about 5% in Table 2. As mentioned earlier, in all cases, the RMSE of both the estimators is quite small. For instance, when the number of firms is 3,000 and the data is generated for 200 months, when the annual default probability is about 1%, the RMSE of the proposed estimator is about 14% of the absolute value of the underlying β, while that of the MLE is about 13%. When the annual default probability is about 5%, all else being the same, the RMSE of the proposed estimator is about 11% of the absolute value of the underlying β, while that of the MLE is about 6%.

When the model is misspecified in the sense that the calibrator assumes that only factors $(Y_{1,t}, X_{i,t})$ determine the default likelihood for firm i at time $t+1$, and uses data to estimate parameters

$(\beta_1, \beta_2, \alpha)$. The RMSE is estimated under the two approaches for (β_1, β_2) in Table 3. These experiments are also repeated when the true parameters are selected so that average default probability is about 5% in Table 4.

Further, in Table 5, we change the data generating model to correspond to the default intensity model. The parameters are kept so that the default probabilities are again around 1% per annum. We compare the RMSE of the proposed estimator with the MLE where for the latter we assuming that the underlying model is logit. It is seen that since this model mis-specification is minor in nature, it leads to a negligible increase in mean square error for both the estimators. We do not use the default intensity based model to estimate MLE's in our experiments in this Section due to the enormous time needed to arrive at the correct values. This is discussed further in Section 6.

Table 1: Comparison of RMSE for default probability 1% per annum, model correctly specified

Time in months	No. of firms	RMSE(β_{prop})	RMSE(α_{prop})	RMSE(β_{ML})	RMSE(α_{ML})
200	1000	0.1280	0.1195	0.1248	0.1144
200	3000	0.0787	0.0663	0.0707	0.0608
200	5000	0.0685	0.0648	0.0574	0.0519
200	7000	0.0574	0.0616	0.0435	0.0424
200	10000	0.0547	0.0547	0.0374	0.0331
100	2000	0.1232	0.1260	0.1157	0.1081
300	2000	0.0774	0.0670	0.0714	0.0565
500	2000	0.0608	0.0529	0.0547	0.0479
700	2000	0.0565	0.0489	0.0509	0.0424

True Parameters: $(\alpha = 7.5, \beta_1 = -0.2, \beta_2 = 0.5, \beta_3 = 0.5)$. RMSE of the proposed estimator is only slightly larger than that of MLE except when the no. of companies is large.

Table 2: Comparison of RMSE for default probability 5% per annum, correctly specified model

Time in months	No. of firms	RMSE(β_{prop})	RMSE(α_{prop})	RMSE(β_{ML})	RMSE(α_{ML})
200	1000	0.0741	0.0663	0.0556	0.0489
200	3000	0.0608	0.0547	0.0331	0.0223
200	5000	0.0574	0.0538	0.0244	0.0173
200	7000	0.0547	0.0538	0.0223	0.0173
200	10000	0.0547	0.0529	0.0173	0.0173
100	2000	0.0761	0.0812	0.0479	0.0412
300	2000	0.0663	0.0583	0.0360	0.0300
500	2000	0.0670	0.0565	0.0331	0.0300
700	2000	0.0741	0.0747	0.0316	0.0282

True Parameters: $(\alpha = 5.5, \beta_1 = -0.2, \beta_2 = 0.5, \beta_3 = 0.5)$. RMSE of the proposed estimator worsens compared to MLE as the number of firms or the time period of data generated becomes large. In both the estimators, in all cases, the relative error is not very large.

Table 3: Comparison of RMSE for default probability 1% per annum, missing covariate with small and large coefficient

β_3	No. of firms	RMSE(β_{prop})	RMSE(α_{prop})	RMSE(β_{ML})	RMSE(α_{ML})
0.5	1000	0.1403	0.1493	0.1392	0.1489
0.5	3000	0.0871	0.1363	0.0842	0.1367
0.5	5000	0.0741	0.1303	0.0721	0.1307
0.5	7000	0.0754	0.1276	0.0707	0.1280
2	1000	0.3109	1.8770	0.3231	1.8785
2	3000	0.2958	1.8810	0.3041	1.8828
2	5000	0.3046	1.8852	0.3135	1.8874
2	7000	0.3014	1.8733	0.3072	1.8775

True Parameters: $(\alpha = 7.5, \beta_1 = -0.2, \beta_2 = 0.5)$, β_3 as specified above. Time period in above experiments is set to 200. Both the proposed estimator and MLE estimate parameters $(\alpha, \beta_1, \beta_2)$ only. The RMSE of the two methods is nearly identical. It worsens as model misspecification increases, that is as value of β_3 increases.

Table 4: Comparison of RMSE for default probability 5% per annum, missing covariate with small and large coefficient

β_3	No. of firms	RMSE(β_{prop})	RMSE(α_{prop})	RMSE(β_{ML})	MSE(α_{ML})
0.5	1000	0.0871	0.1288	0.0748	0.1345
0.5	3000	0.0721	0.1204	0.0608	0.1260
0.5	5000	0.0678	0.1204	0.0574	0.1252
0.5	7000	0.0678	0.1187	0.0547	0.1256
2	1000	0.3376	1.6765	0.3474	1.7172
2	3000	0.3399	1.6714	0.3489	1.7189
2	5000	0.3459	1.6652	0.3439	1.7090
2	7000	0.3911	1.6331	0.3919	1.6813

True Parameters: ($\alpha = 5.5, \beta_1 = -0.2, \beta_2 = 0.5$), β_3 as specified above. Time period in above experiments is set to 200. Both the proposed estimator and MLE estimate parameters (α, β_1, β_2) only. The RMSE of the two methods is nearly identical and higher than in Table 4. It worsens as model misspecification increases, that is as the β_3 value increases.

Table 5: Comparison of RMSE for default probability 1% per annum. Data generated using default intensity model, calibration conducted using logit model.

No. of firms	MSE(β_{prop})	MSE(α_{prop})	MSE(β_{ML})	MSE(α_{ML})
1000	0.1322	0.1144	0.1272	0.1063
3000	0.0824	0.0854	0.0714	0.0640
5000	0.0670	0.0574	0.0574	0.0479
7000	0.0583	0.0538	0.0469	0.0412

True Parameters: ($\alpha = 7.5, \beta_1 = -0.2, \beta_2 = 0.5, \beta_3 = 0.5$). Time periods are fixed at 200. This illustrates that when the default probabilities are small and the model is somewhat misspecified, both the proposed method and the MLE perform similarly and quite well.

6 US Corporate Default Data

6.1 Description of Data

We use a publicly available dataset provided by Research Management Institute, National University of Singapore, for calibration and testing of corporate defaults in the US. This data is a subset of a larger dataset used in [5]. The parent data contains a sample of 12,268 US public firms over the period from 1991 to 2011 assembled from the CRSP monthly and daily files and the Compustat quarterly file. However, the publicly available sample only contains 2,000 companies over the same time period.

The data consists of standard market, accounting and macroeconomic variables used for identification in other default studies. Specifically, variables available to us are distance to default (DTD) (see Merton 1974), ratio of cash to total assets (CASHTA), ratio of net income to total assets

(NITA), market size of the company (SIZE), ratio of market to book value (M/B), interest rate of the economy, and the treasury bill rate. The firm-specific variables have been de-constructed into 'level' and 'trend' components, where level is the moving 12 month average and trend is the difference of value at time (t) and the moving average of the past 12 months (t,t-12).

Summary statistics of the data are provided in below:

- Number of Companies: 2,000
- Time Duration: 251 months (1991-2011)
- Number of Effective Company-Month Observations: 180,329
- Total Number of Defaults: 168 (For year-wise split, refer to Table 6)
- There are no missing observations for active companies in our panel.

Table 6: Number of Defaults by Calendar Year

S No.	Year	Number of Defaults	S No.	Year	Number of Defaults
1	1991	5	11	2001	29
2	1992	8	12	2002	19
3	1993	2	13	2003	11
4	1994	0	14	2004	6
5	1995	4	15	2005	2
6	1996	7	16	2006	1
7	1997	10	17	2007	3
8	1998	8	18	2008	6
9	1999	15	19	2009	10
10	2000	15	20	2010	5

6.2 Pre-Calibration Processing

We short list six variables of the set of twelve which give us the best accuracy ratios (through trial and error). These variables are transformed using log, square and square transformations so that they resemble normal distribution. DTD_ level is log transformed to give log_ DTD_ level (log(DTD_ level)), DTDtrend and SIZEtrend are square transformed to sq_ DTDtrend and sq_ SIZEtrend, respectively. The variables are then standardized by subtracting their empirical mean and dividing the result by variables empirical standard deviation.

Fig 1 and Fig 2 provide histograms of variables pre and post transformations respectively. Table 7 shows that we do improve on our specification post the transformation.

6.3 Comparing different estimators

In order to compare the proposed estimator with those obtained by using the MLE associated with default intensity model and the logit model, comparison tests (same as ones used by Duffie et al. 2007 and Duan et al. 2012) were conducted. Transformed variables were used while implementing the MLE as well, although the comparative performance of the MLE estimator (discussed later in Section 6.4) was insensitive to this transformation.

Figure 1: Frequency Plots of Variables without Transformation

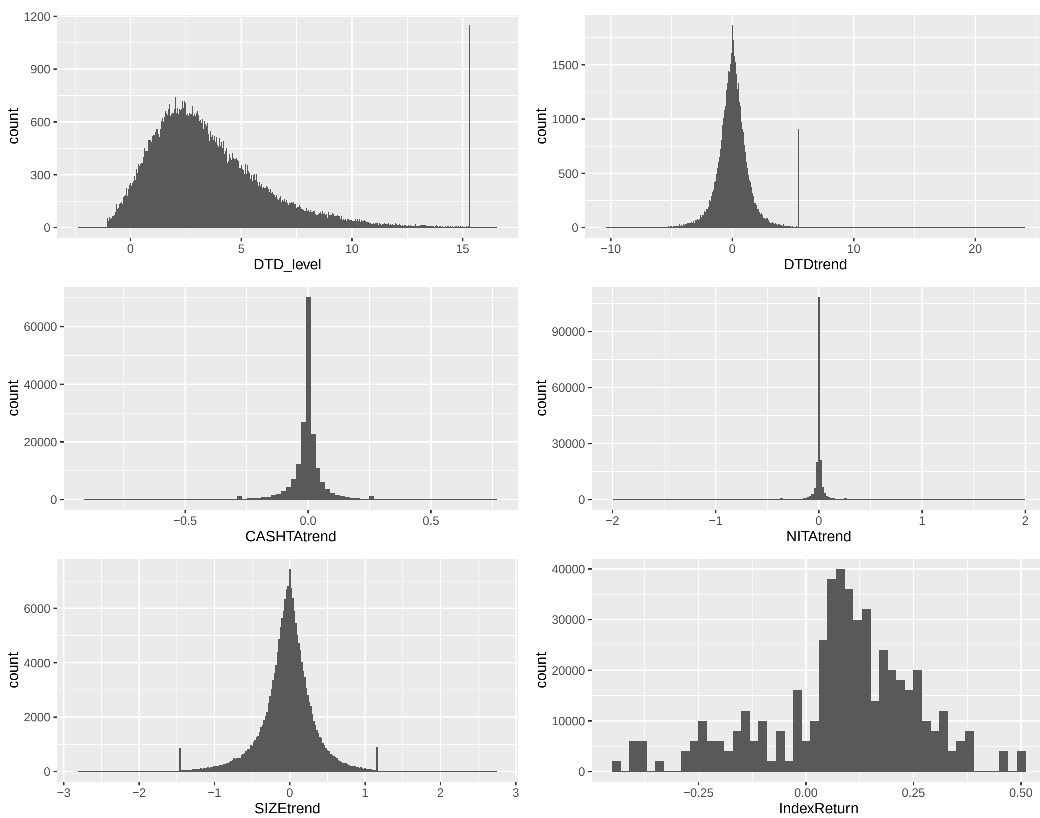
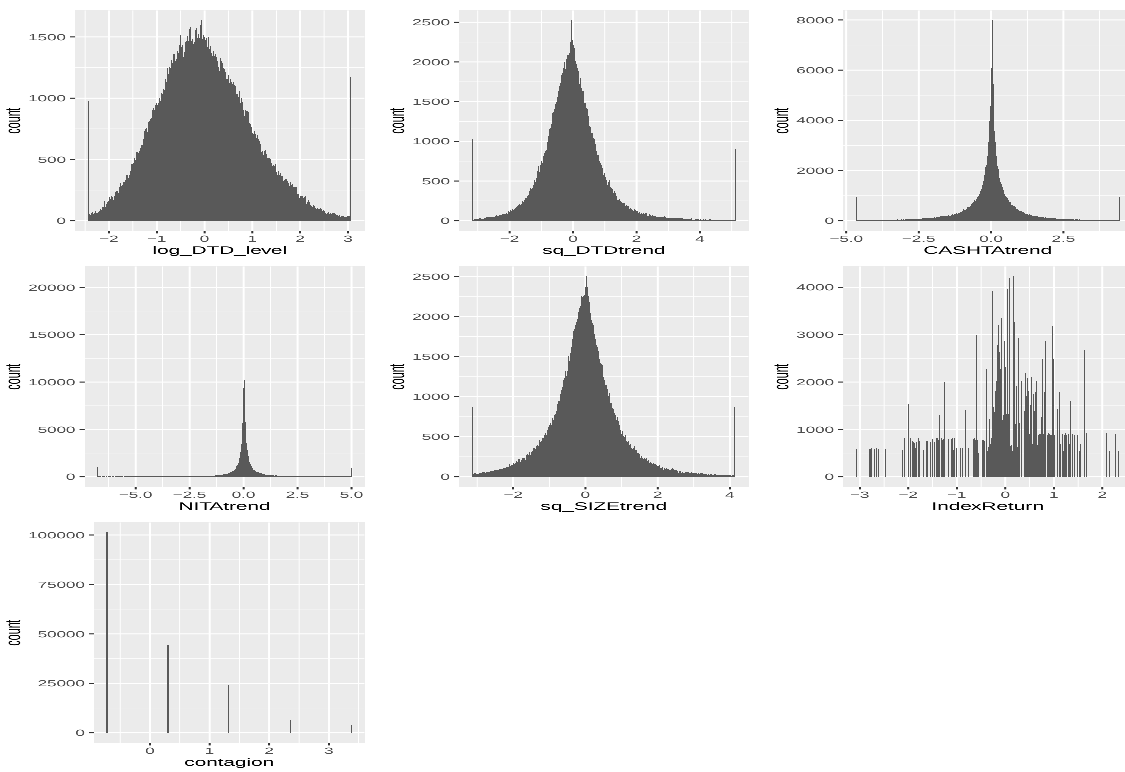


Figure 2: Frequency Plots of Variables after Transformation



We use functional transformations to get the distributions close to normal.

The tests were conducted by segmenting the data in two sets: The first set is used to calibrate the parameters of the proposed model. The parameters were then used to generate default probabilities for the second set. The accuracy of the prediction was tested by using realized defaults in the second set. Specific steps are given below.

1. The testing process is iterative. For each t, starting at $t_0 = 156$, the dataset was separated into 1 to t (fitting dataset) and t to t+12 (testing dataset).

2. The firms in the testing dataset are ranked in decreasing order of their conditional default probabilities and bucketed into ten deciles.

3. Defaults in each decile bucket were noted and a cumulative coverage of the defaults is reported next to the decile. As an example, if for $t = 190$, 16 defaults occur in the next year (that is, period 191 to 202) 10 from the firms listed in the top decile of risk, 5 from the next decile and 1 from the fifth decile, then the first decile is allocated number 10, second decile number 15, and the fifth decile number 16, the rest are allocated 16 for this iteration.

4. This process is iteratively continued for each year beginning at $t = 156$ to $t = 238$, and the numbers in each decile are added.

5. Finally, these numbers are averaged and reported for each method, and the cumulative percentage coverage is reported with each decile.

As is apparent, a better predictive method is likely to have higher percentages of defaults allocated to higher deciles.

Solving for MLE's using the default intensity model (DIM) is a non-convex optimization problem and Duan et al. (2012) use a Sequential Monte Carlo (SMC) algorithm to obtain β parameters of DIM. However, the algorithm takes about 40 hours to run on a 4GB RAM Core i5 system. Nelder-Mead Search optimization routine is used as an approximation for SMC. The optimization routine performs well under reasonable initialization conditions and takes about 20-25 minutes to complete one iteration. The Nelder-Mead optimisation algorithm is used as a part of the 'Optim' package in base R.

6.4 Results

The three models were calibrated using six transformed variables: log of distance to default levels, the square of DTD trend, square of size trend, cash to total assets trend, net income to total assets trend and stock market return.

Results of the calibrations are given in Table 8 and 9. Our calibrated betas are similar in magnitude and sign to betas calibrated through logistic regression. However, the betas estimated by Default Intensity MLE (DIM) are different. This might be due to limitations of the Nelder-Mead algorithm, which was run for 1500 iterations. Moreover, note that in terms of the accuracy statistics, our calibration performs better than either of the other models. Our calibration predicts about 90% of the defaults in the first decile, where as other models are at 85%. However, these statistics should be viewed with due caution as we have only 35 defaults to test our fitted-model on.

6.5 Inclusion of Contagion Effect

According to corporate default literature (see, e.g., Giesecke et. al. 2011), default data shows contagion effect in the sense that occurrences of a default increases the probability of occurrences

Table 7: Accuracy Table with and without Transform

Decile	Without Transform	With Transform
1	0.842	0.895
2	0.921	0.974
3	0.974	0.974
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1

Our calibration methodology performs better when the variables are transformed to a normal distribution.

Table 8: Combined Beta Table

Decile	Our Calibration	DIM	Logit
Constant	−9.251	−6.739	−9.344
log_DTD_level	−1.330	−0.425	−1.837
sq_DTDtrend	−0.199	0.320	−1.267
CASHTAtrend	−0.035	0.006	−0.045
NITAtrend	−0.417	−0.108	−0.060
sq_SIZEtrend	−1.477	−0.615	−0.565
IndexReturn	−0.342	−0.089	−0.218

Table 9: Combined Accuracy Table

Decile	Our Calibration	DIM	Logit
1	0.895	0.842	0.763
2	0.974	0.947	0.921
3	0.974	0.974	0.947
4	1	0.974	0.947
5	1	0.974	0.947
6	1	0.974	0.974
7	1	0.974	1
8	1	1	1
9	1	1	1

of other defaults in the near time horizon. In order to capture this effect, we use aggregate number of defaults in the last period, suitable normalized as a contagion factor in our calibration. The results after including this new variable are shown in Table 10 and 11.

On this limited data, unlike other models, our calibration model performs marginally better (92.1% accuracy) after including the contagion variable.

Table 10: Combined Beta Table with Contagion

Variable	Our Calibration	DIM	Logit
Constant	-9.806	-6.811	-9.145
log_DTD_level	-1.281	-0.322	-1.587
sq_DTDtrend	-0.174	0.072	-1.235
CASHTAtrend	-0.033	0.181	-0.042
NITAtrend	-0.410	0.223	-0.061
sq_SIZEtrend	-1.462	-0.755	-0.582
IndexReturn	0.021	0.007	-0.198
Contagion	1.117	0.194	0.046

Table 11: Combined Accuracy Table with Contagion

Decile	Our Calibration	DIM	Logit
1	0.921	0.868	0.763
2	0.974	0.947	0.921
3	0.974	0.974	0.947
4	1	1	0.947
5	1	1	0.974
6	1	1	0.974
7	1	1	1
8	1	1	1
9	1	1	1

Note that adding the contagion variable marginally improves the fit of the calibration; however, the forward intensity and logit fits remain the same.

7 Conclusion

In this paper we considered the popular default intensity based as well as Logit models that have been used in the past to model corporate defaults. We developed an approximate closed form estimator for parameters - we showed that each parameter maybe approximated by a weighted average of the corresponding covariate observed just before default occurrences. This provides great deal of insight into factors that drive these estimators. We further evaluated the performance of this estimator in a reasonable asymptotic regime. We showed both theoretically and numerically that the proposed estimators perform about as well as those obtained by using far more computationally intensive maximum likelihood methods when the underlying default generating model is correctly specified. Realistically, the default generating mechanism is unknown, and any proposed model is misspecified. In this case we argued that the proposed estimators are as effective as those obtained using the maximum likelihood method. Further, at least on limited corporate default data available to us, we observed that the proposed estimator performs at least as well as MLE under Logit and default intensity models. This supports the use of the proposed estimator for ascertaining and ranking default probabilities.

A Outline of proof of Theorem 4.1

In this section, we provide an outline of the proof of Theorems 4.1 and 4.2. The proofs of all the intermediate lemmas are given in Appendix A.1.
To prove (4.10) in Theorem 4.1, observe that

$$\|\tilde{\beta} - \hat{\beta}(\gamma)\|_2 \leq \|\hat{\beta}(\gamma) - \beta^*(\gamma)\|_2 + \|\beta^*(\gamma) - \tilde{\beta}\|_2, \tag{A.1}$$

where

$$\beta^*(\gamma) = \Sigma^{-1} \frac{E(\hat{V}_\gamma)}{E(\hat{D}_\gamma)}.$$

We develop order upper bounds for the two terms in (A.1). The following observation is essential

$$\Big\| \frac{E\hat{V}_\gamma}{E\hat{D}_\gamma} - \Sigma \cdot \tilde{\beta} \Big\|_2 = O(\gamma) \tag{A.2}$$

From (A.2) it follows that

$$\|\beta^*(\gamma) - \tilde{\beta}\|_2 = O(\gamma), \tag{A.3}$$

since

$$\begin{aligned} \|\beta^*(\gamma) - \tilde{\beta}\|_2 &= \Big\| \Sigma^{-1} \left(\frac{E\hat{V}_\gamma}{E\hat{D}_\gamma} - \Sigma \cdot \tilde{\beta} \right) \Big\|_2 \\ &\leq \|\Sigma^{-1}\| \Big\| \frac{E\hat{V}_\gamma}{E\hat{D}_\gamma} - \Sigma \cdot \tilde{\beta} \Big\|_2. \end{aligned}$$

To see (A.2), we need the following componentwise result:

Lemma A.1.

$$\Big| \frac{E\hat{V}_\gamma^{(i)}}{E\hat{D}_\gamma} - (\Sigma \cdot \tilde{\beta})_i \Big| = O(\gamma)\ \forall i \in \{1, 2, \cdots, d\} \tag{A.4}$$

Then (A.2) follows as

$$\begin{aligned} \Big\| \frac{E\hat{V}_\gamma}{E\hat{D}_\gamma} - \Sigma \cdot \tilde{\beta} \Big\|_2 &\leq \sum_{i=1}^{d} \Big| \frac{E\hat{V}_\gamma^{(i)}}{E\hat{D}_\gamma} - (\Sigma \cdot \tilde{\beta})_i \Big| \\ &= O(\gamma). \end{aligned}$$

Lemma A.2 stated below helps upper bound the second term in the RHS of (A.1).

Lemma A.2.

$$\|\hat{\beta}(\gamma) - \beta^*(\gamma)\|_2^2 = O(\gamma^{\zeta+\delta-1}) + O(\gamma^{\zeta}). \tag{A.5}$$

Since $\zeta \in (0,1)$, Equation (4.10) of Theorem 4.1 then follows.

To see (4.11) in Theorem 1, define

$$\alpha^*(\gamma) = \log\left(\frac{\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E((\exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t))}{\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} ED_{i,t}}\right). \tag{A.6}$$

Then, (4.11) follows from Lemmas A.3 and A.4 below.

Lemma A.3.

$$\|\alpha - \alpha^*(\gamma)\|_2^2 = O(\gamma^\zeta) + O(\gamma^{\zeta+\delta-1}). \tag{A.7}$$

Lemma A.4.

$$\|\hat{\alpha}(\gamma) - \alpha^*(\gamma)\|_2^2 = O(\gamma^\zeta) + O(\gamma^{\zeta+\delta-1}). \tag{A.8}$$

Proof of Theorem 4.2: We use Theorem 4.1 to outline the key steps in proof of Theorem 4.2. Note that,

$$\begin{aligned}
&\frac{\exp(\tilde{\beta}^\intercal V_{i,t} - \alpha) - \exp(\hat{\beta}^\intercal(\gamma)V_{i,t} - \hat{\alpha}(\gamma))}{\exp(\tilde{\beta}^\intercal V_{i,t} - \alpha)} = 1 - \exp((\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma)) \\
&= (\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma)) + \frac{1}{2}((\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma))^2 + \text{ rem. terms}
\end{aligned} \tag{A.9}$$

Squaring (A.9),

$$V_{i,t}^\intercal(\tilde{\beta} - \hat{\beta}(\gamma))(\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma))^2 + ((\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma))^3 + \text{ rem. terms} \tag{A.10}$$

As in the proof of Lemma A.2, we ignore the terms of order larger than 2, as they are asymptotically negligible (more detailed analysis would be added in an updated version), and focus on

$$V_{i,t}^\intercal(\tilde{\beta} - \hat{\beta}(\gamma))(\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t} + (\alpha - \hat{\alpha}(\gamma))^2 + 2(\alpha - \hat{\alpha}(\gamma))(\tilde{\beta} - \hat{\beta}(\gamma))^\intercal V_{i,t}. \tag{A.11}$$

Note that using (4.11) from Theorem 1, $(\alpha - \hat{\alpha}(\gamma))^2$ decays as $O(\gamma^\zeta) + O(\gamma^{\delta+\zeta-1})$. For the first term, we use the independence assumption of Remark 7 to get

$$\|\hat{\beta}(\gamma) - \tilde{\beta}\|_2^2\|\Sigma\| = O(\gamma^\zeta) + O(\gamma^{\delta+\zeta-1}),$$

where Σ is the covariance matrix of $V_{i,t}$. Finally, the last term can be bounded similarly using the Cauchy -Schwarz inequality.

□

A.1 Proofs of key lemmas

We now prove the lemmas used in the proof of Theorem 4.1. The proofs of further intermediate lemmas are presented later in Section A.2.

Proof of Lemma A.1 : To keep the notation simple, we prove the lemma for the one dimensional case. The proof is extended to dimension $d \geq 2$ by essentially following the same steps.

We assume that $\{V_{i,t}\}$ follows the evolution:

$$V_{i,t} = \rho V_{i,t-1} + N_{i,t},$$

with $|\rho| < 1$ and $N_{i,t}$ denote the iid standard Gaussian noise. In the one dimensional case, note that $(\Sigma\tilde{\beta})_1 = \frac{\tilde{\beta}}{1-\rho^2}$.
To prove Lemma A.1, observe that

$$ED_{i,t+1} = E\prod_{j=0}^{t-1}(1 - p(\gamma, V_{i,j}))p(\gamma, V_{i,t})$$

$$EV_{i,t}D_{i,t+1} = E\prod_{j=0}^{t-1}(1 - p(\gamma, V_{i,j}))V_{i,t}p(\gamma, V_{i,t}),$$

where recall that

$$p(\gamma, V_{i,j}) = c\gamma\exp(\tilde{\beta}V_{i,j})(1 + H(\gamma, V_{i,j})).$$

Hence,

$$E\hat{D}_\gamma = \frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E\prod_{j=0}^{t-1}(1 - p(\gamma, V_{i,j}))p(\gamma, V_{i,t})$$

$$E\hat{V}_\gamma = \frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E\prod_{j=0}^{t-1}(1 - p(\gamma, V_{i,j}))V_{i,t}p(\gamma, V_{i,t}).$$

It can be seen that

$$\sum_{t=0}^{T(\gamma)-1}\prod_{j=0}^{t-1}(1 - x_j)x_t = \sum_{k=1}^{T(\gamma)}(-1)^{k+1}\sum_{i_k=k-1}^{T(\gamma)-1}\cdots\sum_{i_1=0}^{i_2-1} x_{i_1}\cdots x_{i_k}.$$

Using the above, and removing $m(\gamma)$ as all firms are homogeneous:

$$E\hat{D}_\gamma = \frac{1}{\gamma T(\gamma)}E\left(\sum_{k=1}^{T(\gamma)}(-1)^{k+1}\sum_{j_k=k-1}^{T(\gamma)-1}\cdots\sum_{j_1=0}^{i_2-1} p(\gamma, V_{i,j_1})\cdots p(\gamma, V_{i,j_k})\right) \tag{A.12}$$

$$E\hat{V}_\gamma = \frac{1}{\gamma T(\gamma)}E\left(\sum_{k=1}^{T(\gamma)}(-1)^{k+1}\sum_{j_k=k-1}^{T(\gamma)-1}\cdots\sum_{j_1=0}^{i_2-1} V_{i,j_k}p(\gamma, V_{i,j_1})\cdots p(\gamma, V_{i,j_k})\right). \tag{A.13}$$

The proof follows once we show that

$$E\hat{V}_\gamma - \frac{\tilde{\beta}}{1-\rho^2}E\hat{D}_\gamma = O(\gamma)$$

while $E\hat{D}_\gamma$ is greater than or equal to a positive constant as $\gamma \to 0$.

To this end, a few additional results are needed. Suppose that $\{Y_t\}_{t\geq 0}$ is an AR(1) stationary process with the evolution

$$Y_t = \rho Y_{t-1} + N_t,$$

with $Y_0 \sim N\left(0, \frac{1}{1-\rho^2}\right)$ and N_t is an iid standard Gaussian process. Let $j_1 < \ldots < j_k$, and $\mathcal{C}_2^k$ be the set of all their pair-wise combinations. Then, the following are true:

Lemma A.5.

$$E\left(\exp\left(\tilde{\beta}\sum_{r=1}^{k} Y_{j_r}\right)\right) = \exp\left(\frac{\tilde{\beta}^2}{1-\rho^2}\left(\frac{1}{2}k + \sum_{i_m, i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m - i_n)}\right)\right) \tag{A.14}$$

and

Lemma A.6.

$$E\left(Y_{j_k}\exp\left(\tilde{\beta}\sum_{r=1}^{k} Y_{j_r}\right)\right) = \frac{\tilde{\beta}}{1-\rho^2}\left(1+\sum_{r=1}^{k-1}\rho^{(j_k - j_r)}\right)\exp\left(\frac{\tilde{\beta}^2}{1-\rho^2}\left(\frac{1}{2}k + \sum_{i_m, i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m - i_n)}\right)\right) \tag{A.15}$$

Using Lemmas A.5, A.6 and the definition of $p(\gamma, V_{i,t})$, it can be seen that

$$E\hat{V}_\gamma - \frac{\tilde{\beta}}{1-\rho^2}E\hat{D}_\gamma$$

equals

$$\frac{1}{\gamma T(\gamma)}\sum_{k=2}^{T(\gamma)}(-1)^{k+1}c^k\gamma^k\sum_{j_k=k-1}^{T(\gamma)-1}\cdots\sum_{j_1=0}^{j_2-1}\left(\sum_{r=1}^{k-1}\rho^{(j_k-j_r)}\right)\exp\left(\frac{\tilde{\beta}^2}{1-\rho^2}\left(\frac{1}{2}k + \sum_{i_m, i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m - i_n)}\right)\right)+O(\gamma). \tag{A.16}$$

Let

$$S_k \triangleq \sum_{j_k=k-1}^{T(\gamma)-1}\cdots\sum_{j_1=0}^{j_2-1}\left(\sum_{r=1}^{k-1}\rho^{(j_k-j_r)}\right)\exp\left(\frac{\tilde{\beta}^2}{1-\rho^2}\left(\frac{1}{2}k + \sum_{i_m, i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m - i_n)}\right)\right).$$

Then,

$$E\hat{V}_\gamma - \frac{\tilde{\beta}}{1-\rho^2}E\hat{D}_\gamma = \frac{1}{\gamma T(\gamma)}\sum_{k=2}^{T(\gamma)}(-1)^{k+1}\gamma^k S_k + O(\gamma). \tag{A.17}$$

To proceed further we need an upper bound on the exponential component of S_k which is independent of the choice of $j_1, \ldots j_k$, but dependent on k. The following lemma is useful to this end.

Lemma A.7. *Let $j_1 < \ldots < j_k$ and let $\mathcal{C}_2^k$ be the set of all their pair-wise combinations. Then,*

$$\sum_{i_m, i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m - i_n)} \leq \frac{k}{1-|\rho|} \tag{A.18}$$

Then,

$$S_k \leq \exp\left(\frac{\tilde{\beta}^2}{1-\rho^2}\left(\frac{1}{2}+\frac{1}{1-|\rho|}\right)k\right)\sum_{j_k=k-1}^{T(\gamma)-1}\cdots\sum_{j_1=0}^{j_2-1}\left(\sum_{r=1}^{k-1}|\rho|^{(j_k-j_r)}\right) \tag{A.19}$$

Now,

$$\sum_{j_k=k-1}^{T(\gamma)-1} \cdots \sum_{j_1=0}^{j_2-1} \left(\sum_{r=1}^{k-1} |\rho|^{(j_k-j_r)} \right) = \sum_{j_k=k-1}^{T(\gamma)-1} \sum_{j_{k-1}=k-2}^{j_k-1} |\rho|^{(j_k-j_{k-1})} \sum_{j_{k-2}=k-3}^{j_{k-1}-1} \cdots \sum_{j_1=0}^{j_2-1} \left(1 + \sum_{r=1}^{k-2} |\rho|^{j_{k-1}-j_r} \right)$$
$$\leq k \sum_{j_k=k-1}^{T(\gamma)-1} \sum_{j_{k-1}=k-2}^{j_k-1} |\rho|^{(j_k-j_{k-1})} j_{k-1}^{k-2}$$
$$\leq kT(\gamma)^{k-1}$$

Thus,

$$E\hat{V}_\gamma - \frac{\tilde{\beta}}{1-\rho^2} E\hat{D}_\gamma \leq \frac{1}{\gamma T(\gamma)} \sum_{k=2}^{T(\gamma)} \tilde{c}^k k \gamma^k T(\gamma)^{k-1} + O(\gamma), \tag{A.20}$$

for an appropriate constant $\tilde{c}$. Further, it is easy to see that since $T(\gamma) = \gamma^{-\zeta}$, the RHS above is $O(\gamma)$.

Similarly, $E\hat{D}_\gamma$ equals

$$\frac{1}{\gamma T(\gamma)} \sum_{k=1}^{T(\gamma)} (-1)^{k+1} \gamma^k \sum_{j_k=k-1}^{T(\gamma)-1} \cdots \sum_{j_1=0}^{j_2-1} \exp\left(\frac{\beta^2}{1-\rho^2} \left(\frac{1}{2}k + \sum_{i_m,i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m-i_n)} \right) \right) + O(\gamma). \tag{A.21}$$

This equals $c \exp(\frac{\beta^2}{2(1-\rho^2)})$ plus

$$\frac{1}{\gamma T(\gamma)} \sum_{k=2}^{T(\gamma)} (-1)^{k+1} \gamma^k \sum_{j_k=k-1}^{T(\gamma)-1} \cdots \sum_{j_1=0}^{j_2-1} \exp\left(\frac{\beta^2}{1-\rho^2} \left(\frac{1}{2}k + \sum_{i_m,i_n \in \mathcal{C}_2^k : i_m > i_n} \rho^{(i_m-i_n)} \right) \right) + O(\gamma). \tag{A.22}$$

As before, (A.22) is $O(\gamma)$, and the result follows. $\square$

Proof of Lemma A.2 :
Establishing Lemma A.2 reduces to estimating the expectation of the square of the error

$$\frac{\hat{V}_\gamma^{(1)}}{\hat{D}_\gamma} - \frac{E\hat{V}_\gamma^{(1)}}{E\hat{D}_\gamma},$$

where $\hat{V}_\gamma^{(1)}$ denotes the first component of vector $\hat{V}_\gamma$. To estimate this, consider the function $f(x,y) = x/y$. From Taylor's series expansion around (x_0, y_0), we have

$$f(x,y) = f(x_0,y_0) + (x-x_0)\frac{\partial f(x_0,y_0)}{\partial x} + (y-y_0)\frac{\partial f(x_0,y_0)}{\partial y} + \ldots.$$

Thus,

$$\frac{\hat{V}_\gamma^{(1)}}{\hat{D}_\gamma} = \frac{E\hat{V}_\gamma^{(1)}}{E\hat{D}_\gamma} + \frac{1}{E\hat{D}_\gamma}(\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)}) - \frac{E\hat{V}_\gamma^{(1)}}{(E\hat{D}_\gamma)^2}(\hat{D}_\gamma - E\hat{D}_\gamma) + \text{ rem. terms},$$

where rem. terms above denotes the remainder higher order terms that can be shown to be asymptotically negligible compared to the dominant terms above (the supporting analysis would be added in a more elaborate version of this paper).

Hence,

$$E\left(\frac{\hat{V}_\gamma^{(1)}}{\hat{D}_\gamma} - \frac{E\hat{V}_\gamma^{(1)}}{E\hat{D}_\gamma}\right)^2$$

$$= \frac{1}{(E\hat{D}_\gamma)^2}E(\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})^2 + \frac{(E\hat{V}_\gamma^{(1)})^2}{(E\hat{D}_\gamma)^4}E(\hat{D}_\gamma - E\hat{D}_\gamma)^2 - 2\frac{E\hat{V}_\gamma^{(1)}}{(E\hat{D}_\gamma)^3}E\left((\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})(\hat{D}_\gamma - E\hat{D}_\gamma)\right) + \text{rem. terms.} \tag{A.23}$$

We now get a handle on the terms above. The essence is captured in analyzing

$$E(\hat{D}_\gamma - E\hat{D}_\gamma)^2,$$

since similar bounds will follow for the other terms. We may rewrite $(\hat{D}_\gamma - E\hat{D}_\gamma)$ as

$$\frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}(D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t)) + \frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}(E(D_{i,t+1}|\mathcal{F}_t) - ED_{i,t+1}),$$

then the error $E(\hat{D}_\gamma - E\hat{D}_\gamma)^2$ is the expectation of

$$\left(\frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}(D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t)) + \frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}(E(D_{i,t+1}|\mathcal{F}_t) - ED_{i,t+1})\right)^2 \tag{A.24}$$

Note that the cross terms where one term has the form $D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t)$, have an expectation zero, and thus, the expectation of (A.24) equals

$$\frac{1}{\gamma^2 T^2(\gamma)m^2(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}E(D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t))^2 \quad + \tag{A.25}$$

$$\frac{1}{\gamma^2 T^2(\gamma)m^2(\gamma)}\sum_{t_1=0}^{T(\gamma)-1}\sum_{t_2=0}^{T(\gamma)-1}\sum_{i_1=1}^{m(\gamma)}\sum_{i_2=1}^{m(\gamma)}E\left((E(D_{i_1,t_1+1}|\mathcal{F}_{t_1}) - ED_{i_1,t_1+1})(E(D_{i_2,t_2+1}|\mathcal{F}_{t_2}) - ED_{i_2,t_2+1})\right). \tag{A.26}$$

Note that the expectation in (A.25) can be bounded above by a constant times γ for each i and t. Then,

$$\frac{1}{\gamma^2 T^2(\gamma)m^2(\gamma)}\sum_{t=0}^{T(\gamma)-1}\sum_{i=1}^{m(\gamma)}E(D_{i,t+1} - E(D_{i,t+1}|\mathcal{F}_t))^2 = \Theta(\gamma^{\zeta+\delta-1}) \tag{A.27}$$

We now place a bound on (A.26). First fix an i and t and consider

$$\begin{aligned}&E\left(E\left(D_{i_1,t_1+1}|\mathcal{F}_{t_1}\right) - ED_{i_1,t_1+1}\right)\left(E\left(D_{i_2,t_2+1}|\mathcal{F}_{t_2}\right) - ED_{i_2,t_2+1}\right)\\&= E\left(E\left(D_{i_1,t_1+1}|\mathcal{F}_{t_1}\right)E\left(D_{i_2,t_2+1}|\mathcal{F}_{t_2}\right)\right) - E\left(D_{i_1,t_1+1}\right)E\left(D_{i_2,t_2+1}\right).\end{aligned} \tag{A.28}$$

We note that

$$E(D_{i,t+1}|\mathcal{F}_t) = p(\gamma, V_{i,t})\mathbb{I}(\tau_i \geq t) \leq c\gamma\exp(\tilde{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t})) \text{ a.s. } \forall i, t$$

Bounding (A.28) hence boils down to evaluating

$$\gamma^2 c^2 \left(E(\exp(\tilde{\beta}^\intercal V_{i,t_1}) \exp(\tilde{\beta}^\intercal V_{j,t_2})) - E(\exp(\tilde{\beta}^\intercal V_{i,t}))^2 \right) + O(\gamma^3), \tag{A.29}$$

along with some significant error terms. By stationarity,

$$E(\exp(\tilde{\beta}^\intercal V_{i,t}))^2 = \exp\left(\tilde{\beta}^\intercal \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal \tilde{\beta} \right) \forall i, t. \tag{A.30}$$

To analyze the first term of (A.29), we need to get a handle on $E((V_{i,t_1} + V_{j,t_2})(V_{i,t_1} + V_{j,t_2})^\intercal)$, done by the following lemma. Define

$$\mathbf{I}^\star \triangleq \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{A.31}$$

where I is the identity matrix of dimension d_1, and $\mathbf{I}^\star \in \Re^{(d_1+d_2)\times(d_1+d_2)}$. Recall that $V_{i,t}$ has the time evolution

$$V_{i,t} = M \cdot V_{i,t-1} + \phi_{i,t}, \tag{A.32}$$

Lemma A.8. *Let $i, j \in \{1, 2, \ldots m(\gamma)\}$, $i \neq j$, $t_1, t_2 \in \{1, 2, \ldots T(\gamma)\}$, $t_1 \leq t_2$. Then in the setup of (A.32),*

$$\begin{aligned} &E((V_{i,t_1} + V_{j,t_2})(V_{i,t_1} + V_{j,t_2})^\intercal) = \\ &2\sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal + M^{t_2-t_1}. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star (M^l)^\intercal + \left(\sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^\intercal \right) \cdot (M^{t_2-t_1})^\intercal. \end{aligned} \tag{A.33}$$

By Lemma A.8,

$$E(\exp(\tilde{\beta}^\intercal V_{i,t_1}) \exp(\tilde{\beta}^\intercal V_{j,t_2})) = \exp\left(\tilde{\beta}^\intercal \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal \tilde{\beta} + \tilde{\beta}^\intercal M^{t_2-t_1}. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^\intercal \tilde{\beta} \right). \tag{A.34}$$

The next lemma is required to bound (A.34). Fix $t_2 - t_1$ to be equal to $k \in \{1, 2, \ldots T(\gamma) - 1\}$. Then,

Lemma A.9.

$$\tilde{\beta}^\intercal M^k. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^\intercal \tilde{\beta} \leq \frac{\|\tilde{\beta}\|_2^2}{1 - \|M\|^2} \|M\|^k. \tag{A.35}$$

Using (A.35) we can upper bound the sum

$$\frac{1}{\gamma^2 \cdot T^2(\gamma) m^2(\gamma)} \sum_{i_1=0}^{m(\gamma)} \sum_{i_2=0}^{m(\gamma)} \sum_{t_1=0}^{T(\gamma)} \sum_{t_2=0}^{T(\gamma)} \left(E\left(E\left(D_{i_1,t_1+1} | \mathcal{F}_{t_1} \right) E\left(D_{i_2,t_2+1} | \mathcal{F}_{t_2} \right) \right) - E\left(D_{i_1,t_1+1} \right) E\left(D_{i_2,t_2+1} \right) \right).$$

We first note that we can use (A.34) and (A.30) to rewrite this as

$$\frac{\exp\left(\tilde{\beta}^T \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^T \tilde{\beta} \right)}{T^2(\gamma) m^2(\gamma)} \sum_{i_1=0}^{m(\gamma)} \sum_{i_2=0}^{m(\gamma)} \sum_{t_1=0}^{T(\gamma)} \sum_{t_2=0}^{T(\gamma)} \left(\exp\left(\tilde{\beta}^\intercal M^{t_2-t_1}. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^\intercal \tilde{\beta} \right) - 1 \right) \tag{A.36}$$

plus remainder terms. We note an observation: In asymptotic analysis, we need only consider the case where $i_1 \neq i_2$ and $t_1 \neq t_2$. Suppose $t_2 = t_1$. In such a case $M^{t_2-t_1} = \mathbf{I}$. Further, the number of cases in which this occurs is exactly equal to $T(\gamma)$, and the expression inside the summation, $\exp(\tilde{\beta}^\intercal \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star (M^l)^\intercal \tilde{\beta}) - 1$ is bounded. To see this, consider the following

$$\begin{aligned}\langle \tilde{\beta}. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^T \tilde{\beta} \rangle &\leq \|\tilde{\beta}\|_2 \Big\| \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^T \tilde{\beta} \Big\| \\ &\leq \frac{\|\tilde{\beta}\|_2^2}{1 - \|M\|^2}\end{aligned}$$

Hence, by summing over all i_1, i_2 a decay rate of $T(\gamma)^{-1}$ will be obtained. Now consider $i_1 = i_2$. In such a case, it can be seen using (A.58), that the only change is replacement of $\mathbf{I}^\star$ by $\mathbf{I}$. In both cases the bound in (A.35) will still hold and hence, an identical analysis will suffice. We henceforth only consider cases where $i_1 \neq i_2$ and $t_1 \neq t_2$. We first evaluate the outer sum over all $m(\gamma)(m(\gamma)-1)$ cases where $i_1 \neq i_2$. By using the fact that all firms are statistically identical, (A.36) can be upper bounded by

$$\frac{\exp\left(\tilde{\beta}^T \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal \tilde{\beta}\right)}{T^2(\gamma)} \sum_{t_1=0}^{T(\gamma)} \sum_{t_2=0}^{T(\gamma)} \left(\exp\left(\tilde{\beta}^\intercal M^{t_2 - t_1}. \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^\star \cdot (M^l)^\intercal \tilde{\beta} \right) - 1 \right).$$

Next, we note that the number of times where $t_2 - t_1 = k$ is exactly $T(\gamma) - k$. Using Lemma A.9 we can upper bound the above by

$$\frac{\exp\left(\tilde{\beta}^\intercal \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal \tilde{\beta}\right)}{T(\gamma)} \sum_{k=0}^{\infty} \left(\exp\left(\frac{\|\tilde{\beta}\|^2}{1 - \|M\|^2} \|M\|^k \right) - 1 \right). \tag{A.37}$$

We now need the following bound on exponential sums

Lemma A.10. *Let $\theta \in \Re$, $\rho \in (0,1)$. Then,*

$$\sum_{k=0}^{\infty} \left(\exp(\theta \cdot \rho^k) - 1 \right) < \infty. \tag{A.38}$$

With $\theta = \frac{\|\tilde{\beta}\|^2}{1-\|M\|^2}$ and $\rho = \|M\|$, Lemma A.10 means that the summation in (A.37) is finite and thus,

$$\frac{\exp\left(\tilde{\beta}^\intercal \cdot \sum_{l=0}^{\infty} M^l \cdot (M^l)^\intercal \tilde{\beta}\right)}{T(\gamma)} \sum_{k=0}^{\infty} \left(\exp\left(\frac{\|\tilde{\beta}\|^2}{1 - \|M\|^2} \|M\|^k \right) - 1 \right) = O(\gamma^\zeta). \tag{A.39}$$

We now get a handle on the error. Consider the significant part of the Taylor series (A.23),

$$\frac{1}{(E\hat{D}_\gamma)^2} E(\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})^2 + \frac{(E\hat{V}_\gamma^{(1)})^2}{(E\hat{D}_\gamma)^4} E(\hat{D}_\gamma - E\hat{D}_\gamma)^2 - 2\frac{E\hat{V}_\gamma^{(1)}}{(E\hat{D}_\gamma)^3} E\left((\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})(\hat{D}_\gamma - E\hat{D}_\gamma) \right)$$

From Lemma A.1, we can rewrite the above as

$$\frac{1}{(E\hat{D}_\gamma)^2} E(\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})^2 + \frac{(\Sigma\tilde{\beta})_1^2}{(E\hat{D}_\gamma)^2} E(\hat{D}_\gamma - E\hat{D}_\gamma)^2 - 2\frac{(\Sigma\tilde{\beta})_1}{(E\hat{D}_\gamma)^2} E\left((\hat{V}_\gamma^{(1)} - E\hat{V}_\gamma^{(1)})(\hat{D}_\gamma - E\hat{D}_\gamma) \right) + O(\gamma) \tag{A.40}$$

As in (A.26), the significant sources of error are the errors in $ED_{i,t+1}$ and $EV_{i,t}D_{i,t+1}$. Similar to Lemma A.1, it can be shown that for all i, t,

$$\left|\frac{EV_{i,t}^{(1)}D_{i,t+1}}{ED_{i,t+1}} - (\Sigma\beta)_1\right| = O(\gamma) \tag{A.41}$$

Let

$$ED_{i,t+1} = c\gamma E\exp(\tilde{\beta}^\intercal V_{i,t}) + e_1(\gamma)$$

and

$$EV_{i,t}^{(1)}D_{i,t+1} = c\gamma EV_{i,t}^{(1)}\exp(\tilde{\beta}^\intercal V_{i,t}) + e_2(\gamma),$$

where $e_1(\gamma)$ and $e_2(\gamma)$ are error terms. By (A.41),

$$e_2(\gamma) = (\Sigma\tilde{\beta})_1 e_1(\gamma) + O(\gamma). \tag{A.42}$$

Then, from Lemma A.1 and (A.40) the total error is

$$\frac{(\Sigma\tilde{\beta})_1 e_2(\gamma) + (\Sigma\tilde{\beta})_1^2 e_1(\gamma) - (\Sigma\tilde{\beta})_1((\Sigma\tilde{\beta})_1 e_1(\gamma) + e_2(\gamma))}{(ED_\gamma)^2} + O(\gamma) = O(\gamma)$$

Lemma A.2 is hence proved.

□

Proof of Lemma A.3 :
First note that

$$\begin{aligned}\alpha(\gamma) - \alpha^*(\gamma) &= \log\left(\frac{\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E(\exp\tilde{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t)}{\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E((\exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t})))\mathbb{I}(\tau_i \geq t)}\right)\\ &= \log\left(1 + \frac{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E(\exp\tilde{\beta}^\intercal V_{i,t} - \exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t})\mathbb{I}(\tau_i \geq t)}{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E((\exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t})))\mathbb{I}(\tau_i \geq t)}\right)\end{aligned} \tag{A.43}$$

We know that

$$\frac{x}{1+x} \leq \log(1+x) \leq x.$$

Hence,

$$\frac{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E(\exp\tilde{\beta}^\intercal V_{i,t} - \exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t})\mathbb{I}(\tau_i \geq t)}{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E((\exp\tilde{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t))} \leq \alpha(\gamma) - \alpha^*(\gamma) \tag{A.44}$$

and

$$\alpha(\gamma) - \alpha^*(\gamma) \leq \frac{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E(\exp\tilde{\beta}^\intercal V_{i,t} - \exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t)}{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=1}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1} E((\exp\hat{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t))} \tag{A.45}$$

Consider first the denominator of the LHS of (A.44). Note that for a fixed i and t, we can use the bound

$$|(\exp\tilde{\beta}^\intercal V_{i,t})(1+H(\gamma,V_{i,t}))\mathbb{I}(\tau_i \geq t))| \leq \exp\tilde{\beta}^\intercal V_{i,t} + \gamma\exp\tilde{\beta}^\intercal V_{i,t}$$

Also,

$$\exp \tilde{\beta}^\intercal V_{i,t}(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t) \to \exp \tilde{\beta}^\intercal V_{i,t} \text{ a.s.}$$

Hence, by the dominated convergence theorem

$$E((\exp \tilde{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t)) \to E((\exp \tilde{\beta}^\intercal V_{i,t})) > 0$$

Hence $\frac{1}{T(\gamma)m(\gamma)} \sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E((\exp \hat{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t))$ is $O(1)$. The denominator of (A.45) can be written as

$$\frac{1}{T(\gamma)m(\gamma)} \sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E(\exp((\hat{\beta} - \tilde{\beta})^\intercal V_{i,t}) \exp(\tilde{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t))$$

Using Fatou's Lemma and similar arguments as above, it can be shown that

$$E((\exp \hat{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t)) = \Omega(1)$$

for every i and t. The asymptotic $\|\alpha(\gamma) - \alpha^*(\gamma)\|_2^2$ is thus governed by

$$\left(\frac{1}{T(\gamma)m(\gamma)} \sum_{i=1}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E(\exp \tilde{\beta}^\intercal V_{i,t} - \exp \hat{\beta}^\intercal V_{i,t})(1 + H(\gamma, V_{i,t}))\mathbb{I}(\tau_i \geq t) \right)^2. \tag{A.46}$$

Additionally, from Assumption 1, we know that $|H(\gamma, V_{i,t})| \leq C\gamma \exp(\tilde{\beta}^\intercal V_{i,t})$. Hence, the key expression to analyse is

$$E(\exp \tilde{\beta}^\intercal V_{i,t} - \exp \hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i \geq t). \tag{A.47}$$

Consider the Taylor series

$$e^x = e^y + \sum_{r=1}^{\infty} \frac{(x - y)^r}{r!} e^y.$$

Setting $x = \hat{\beta}^\intercal V_{i,t}$ and $y = \tilde{\beta}^\intercal V_{i,t}$, (A.47) becomes

$$E\left((\hat{\beta} - \tilde{\beta})^\intercal V_{i,t} \exp(\tilde{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i \geq t) + \sum_{r=2}^{\infty} \frac{((\hat{\beta} - \tilde{\beta})^\intercal V_{i,t})^r}{r!} \exp(\tilde{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i \geq t) \right).$$

As before, the first term contributes significantly while the remaining terms can be shown to be asymptotically negligible. Using the Cauchy-Schwarz inequality on the first term, and the first part of Theorem 4.1, (A.46) is $O(\gamma^\zeta) + O(\gamma^{\zeta+\delta-1})$. □

Proof of Lemma A.4 :
Notice that

$$\begin{aligned}
\alpha^*(\gamma) - \hat{\alpha}(\gamma) = \log &\left(\frac{\frac{1}{\gamma T(\gamma)m(\gamma)} \sum_{i=0}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} D_{i,t+1}}{\frac{1}{\gamma T(\gamma)m(\gamma)} \sum_{i=0}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} ED_{i,t}} \right) \\
&- \log \left(\frac{\frac{1}{T(\gamma)m(\gamma)} \sum_{i=0}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} \exp(\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i \geq t)}{\frac{1}{T(\gamma)m(\gamma)} \sum_{i=0}^{m(\gamma)} \sum_{t=0}^{T(\gamma)-1} E((\exp \hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i \geq t))} \right)
\end{aligned} \tag{A.48}$$

As in the proof of Lemma A.3, the asymptotic behaviour of the MSE of (A.48) is governed by

$$E\left(\frac{\frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}D_{i,t+1}-ED_{i,t+1}}{\frac{1}{\gamma T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}ED_{i,t+1}}\right)^2 \tag{A.49}$$

and

$$E\left(\frac{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\exp(\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t)-E((\exp\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t))}{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}E((\exp\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t))}\right)^2. \tag{A.50}$$

Note that from the proof of Lemma A.2, (A.49) is $O(\gamma^\zeta)+O(\gamma^{\zeta+\delta-1})$. Using arguments similar to the proof of Lemma A.2 it can be seen that

$$E\left(\frac{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}\exp(\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t)-E((\exp\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t))}{\frac{1}{T(\gamma)m(\gamma)}\sum_{i=0}^{m(\gamma)}\sum_{t=0}^{T(\gamma)-1}E((\exp\hat{\beta}^\intercal V_{i,t})\mathbb{I}(\tau_i\geq t))}\right)^2 = O(\gamma^\zeta)+O(\gamma^{\zeta+\delta-1})$$

This proves Lemma A.4.

□

A.2 Proofs of intermediate lemmas

We first state a well known fact: Let $X\in\Re^d$ be a Gaussian random vector with a covariance matrix $\Sigma=(\sigma_{i,j})$. Then for any $\mathbf{r}\in\Re^d$,

$$E\exp(\mathbf{r}^\intercal X)=\exp\left(\frac{1}{2}\mathbf{r}^\intercal\Sigma\mathbf{r}\right)=\exp\left(\frac{1}{2}\sum_{i,j}r_ir_j\sigma_{i,j}\right) \tag{A.51}$$

Proof of Lemma A.5 :
Let

$$Y=(\ Y_{j_1},Y_{j_2},\dots Y_{j_k}\ )^\intercal$$

and $\mathbf{e}_k$ be the all 1's vector of length k. Let $\Sigma_Y=\{\sigma_Y(m,n)\}$ be the variance covariance matrix of Y, where

$$\sigma_Y(m,n)=EY_{j_m}Y_{j_n}.$$

Note that we may re-write

$$\sum_{r=1}^{k}Y_{j_r}=\mathbf{e}_k^\intercal Y \tag{A.52}$$

Then, by (A.51),

$$E\left(\exp(\tilde{\beta}\mathbf{e}_k^\intercal Y)\right)=\exp\left(\frac{\tilde{\beta}^2}{2}\sum_{m,n}\sigma_Y(m,n)\right). \tag{A.53}$$

To complete the proof, we must evaluate $\sigma_Y(m,n)$. By definition,

$$Y_t=\sum_{r=-\infty}^{t}\rho^{t-r}N_{t-r}$$

Suppose $j_m < j_n$. Then,

$$\begin{aligned} Y_{j_n} &= \sum_{r=-\infty}^{j_m} \rho^{j_n - r} N_{j_n - r} + \sum_{r=j_m+1}^{j_n} \rho^{j_n - r} N_{j_n - r} \\ &= \rho^{j_m - j_n} Y_{j_m} + \sum_{r=j_m+1}^{j_n} \rho^{j_n - r} N_{j_n - r}. \end{aligned}$$

Note that the second term of the summation is independent of Y_{j_m}. Further, by stationarity,

$$Y_t \sim N\left(0, \frac{1}{1-\rho^2}\right) \forall t.$$

From the above and by symmetry,

$$\sigma_Y(m,n) = \frac{1}{1-\rho^2} \rho^{|j_m - j_n|}. \tag{A.54}$$

Lemma A.5 follows from (A.53) and (A.54). □

Proof of Lemma A.6:
Let $j_1 < j_2 \cdots < j_k$. Then, Let

$$\mathbf{e}^* = (\tilde{\beta}\mathbf{e}_{k-1}, \beta_k),$$

where $\tilde{\beta}\mathbf{e}_{k-1}$ denotes a $k-1$ dimensional column vector with each entry equalling $\tilde{\beta}$. Let $\mathcal{S}_2$ be the set of all pairwise combinations of $j_1, j_2, \ldots, j_{k-1}$. From (A.51),

$$E\left(\exp(\mathbf{e}^{*\intercal} Y)\right) = \exp\left(\frac{1}{1-\rho^2}\left(\frac{\tilde{\beta}^2}{2}(k-1) + \tilde{\beta}^2 \sum_{i_m, i_n \in \mathcal{S}_2 : i_m > i_n} \rho^{(i_m - i_n)} + \beta_k \tilde{\beta} \sum_{r=1}^{k-1} \rho^{(j_k - j_r)} + \frac{1}{2}\beta_k^2\right)\right) \tag{A.55}$$

Differentiating both sides of (A.55) with respect to β_k, interchanging the derivative and the expectation using the dominated convergence theorem and then setting β_k to $\tilde{\beta}$, we get the desired result.

□

Proof of Lemma A.7 :
Let m, n be such that $k > m > n > 0$. Since $i_1 < i_2 \cdots < i_k$, $i_m - i_n > m - n$. We now note that a term of the form $i_m - i_{m-r}$ occurs exactly $k - r$ times in the summation and further, that $\rho^{(i_m - i_{m-r})}$ is upper bounded by $|\rho^r|$. Then the summation in (A.18) is upper bounded by

$$\sum_{r=1}^{k} (k-r)|\rho|^r,$$

This gives Lemma A.7

□

Proof of Lemma A.8 :
We may rewrite $V_{i,t_1} + V_{j,t_2}$ as

$$V_{i,t_1} + V_{j,t_2} = \sum_{k=-\infty}^{t_1} M^{t_1-k}\phi_{i,k} + \sum_{k=-\infty}^{t_2} M^{t_2-k}\phi_{j,k}$$
$$= \sum_{k=-\infty}^{t_1} M^{t_1-k}\left(\phi_{i,k} + M^{t_2-t_1}\phi_{j,k}\right) + \sum_{k=t_1+1}^{t_2} M^{t_2-k}\phi_{j,k}. \tag{A.56}$$

This is a sum of independent random vectors, whose covariance matrix is the sum of covariance matrices of its individual components. In order to analyse further we must consider the variance covariance matrices of $\phi_{i,t}$ and $\phi_{j,t}$. Let $i \neq j$. Then,

$$\begin{aligned} E(\phi_{i,t}\phi_{j,t}^\intercal) &= E\left(\begin{pmatrix} \epsilon_t \\ \psi_{i,t} \end{pmatrix}(\epsilon_t, \psi_{j,t})\right) \\ &= E\begin{pmatrix} \epsilon_t\epsilon_t^\intercal & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \\ . &= \mathbf{I}^\star \end{aligned} \tag{A.57}$$

Similarly, if $i = j$, then
$$E(\phi_{i,t}\phi_{j,t}^\intercal) = \mathbf{I}. \tag{A.58}$$

First observe that,

$$E(\left(\phi_{i,k} + M^{t_2-t_1}\phi_{j,k}\right)\left(\phi_{i,k} + M^{t_2-t_1}\phi_{j,k}\right)^\intercal) = \mathbf{I} + M^{t_2-t_1}\mathbf{I}^\star + \mathbf{I}^\star(M^{t_2-t_1})^\intercal + M^{t_2-t_1}\cdot(M^{t_2-t_1})^\intercal.$$

This implies that

$$\begin{aligned} &E\left(M^{t_1-k}\left(\left(\phi_{i,k} + M^{t_2-t_1}\phi_{j,k}\right)\left(\phi_{i,k} + M^{t_2-t_1}\phi_{j,k}\right)^\intercal\right)(M^{t_1-k})^\intercal\right) \\ &= M^{t_1-k}(\mathbf{I} + M^{t_2-t_1}\mathbf{I}^\star + \mathbf{I}^\star(M^{t_2-t_1})^\intercal + M^{t_2-t_1}\cdot(M^{t_2-t_1})^\intercal)(M^{t_1-k})^\intercal \\ &= M^{t_1-k}\cdot(M^{t_1-k})^\intercal + M^{t_2-k}\cdot(M^{t_2-k})^\intercal + M^{t_2-k}\cdot\mathbf{I}^\star(M^{t_1-k})^\intercal + M^{t_1-k}\cdot\mathbf{I}^\star(M^{t_2-k})^\intercal. \end{aligned}$$

We now use the fact that the $\phi_{i,t}$'s are independent over time to get the total covariance matrix, given by

$$\sum_{k=-\infty}^{t_1} M^{t_1-k}\cdot(M^{t_1-k})^\intercal + M^{t_2-k}\cdot(M^{t_2-k})^\intercal + M^{t_2-k}\cdot\mathbf{I}^\star(M^{t_1-k})^\intercal + M^{t_1-k}\cdot\mathbf{I}^\star(M^{t_2-k})^\intercal. \tag{A.59}$$

The covariance matrix of the second term of (A.56) can be shown to be

$$\sum_{k=t_1+1}^{t_2} M^{t_2-k}\cdot(M^{t_2-k})^\intercal = \sum_{l=0}^{\infty} M^l\cdot(M^l)^\intercal - \sum_{l=t_2-t_1}^{\infty} M^l\cdot(M^l)^\intercal. \tag{A.60}$$

Lemma A.8 follows.

□

Proof of Lemma A.9 :

$$\begin{aligned}\langle\tilde{\beta}, M^k.\sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^{\star} \cdot (M^l)^{\intercal}\tilde{\beta}\rangle &\leq \|\tilde{\beta}\|_2 \Big\|M^k \cdot \sum_{l=0}^{\infty} M^l \cdot \mathbf{I}^{\star} \cdot (M^l)^{\intercal}\tilde{\beta}\Big\| \\ &\leq \|\tilde{\beta}\|_2 \|M\|^k \sum_{l=0}^{\infty} \|M\|^l \|\mathbf{I}^{\star}\| \|M\|^l \|\tilde{\beta}\|_2 \\ &\leq \frac{\|\tilde{\beta}\|_2^2}{1-\|M\|^2}\|M\|^k,\end{aligned}$$

where the first inequality is a result of the Cauchy - Schwarz inequality, and the second a result of the definition of matrix norm.

$\square$

Proof of Lemma A.10 :
It is sufficient to show that

$$\exp\left(\theta \cdot \rho^k\right) - 1 < 2\theta\rho^k.$$

for each large enough k. To this end, note that if $x > 0$, $f(x) = \frac{\exp(x)-1}{x}$ is an increasing function of x and hence $u_k \triangleq \frac{\exp\left(\theta\cdot\rho^k\right)-1}{\theta\cdot\rho^k}$ decreases in k, such that $\lim_{k\to\infty} u_k = 1$. Hence, $\exists k_1 : \forall k \geq k_1$, $u_k < 2$. Then, we have $\forall k \geq k_1$, $\exp\left(\theta \cdot \rho^k\right) - 1 < 2\theta\rho^k$. Then,

$$\sum_{k=K}^{\infty}\left(\exp\left(\theta \cdot \rho^k\right) - 1\right) \to 0,$$

as $K \to \infty$ and the overall sum is bounded.

$\square$

References

[1] Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), pp.589-609.

[2] Bharath, S. T., & Shumway, T. 2008. Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3), 1339-1369.

[3] Bagley, S. C., White, H. and Golomb, B. A. 2001. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain.*Journal of clinical epidemiology* 54.10, pp 979-985.

[4] Chava, S. and Jarrow, R.A., 2004. Bankruptcy prediction with industry effects. *Review of Finance*, 8(4), pp.537-569.

[5] Duffie, D., Saita, L. and Wang, K., 2007. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3), pp.635-665.

[6] Duffie, D., Eckner, A., Horel, G., Saita, L. 2009. Frailty correlated default. The Journal of Finance, 64(5), 2089-2123.

[7] Duan, J.C., Sun, J. and Wang, T., 2012. Multiperiod corporate default prediction: A forward intensity approach. *Journal of Econometrics*, 170(1), pp.191-209.

[8] Duan, J. C. , and A. Fulop. 2013. Multiperiod corporate default prediction with the partially-conditioned forward intensity.

[9] F. A. Graybill. Matrices with Applications in Statistics. Wadsworth Inc., Belmont, 1983.

[10] Eymen, E., Giesecke, K. and Goldberg, L. R. 2010. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics* 1, pp 642-665.

[11] K Giesecke, FA Longstaff, S Schaefer, I Strebulaev. 2011. Corporate bond default risk: A 150-year perspective. *Journal of Financial Economics* 102 (2), 233-250.

[12] Giesecke, K. and Goldberg, L. R. 2004. Forecasting default in the face of uncertainty. *The Journal of Derivatives* 12, no. 1: 11-25.

[13] Juneja, S. (2017). Dynamic Portfolio Credit Risk and Large Deviations. *In Econophysics and Sociophysics: Recent Progress and Future Directions* (pp. 41-58). Springer, Cham.

[14] Merton, R. C. 1974. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449-470.

[15] Milton, J. C., Shankar, V. N. and Mannering, F. L. 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis & Prevention* 40.1, pp 260-266.

[16] Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pp.109-131.

[17] Shumway, T., 2001. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), pp.101-124.