Proxy CDS Curves for Individual Corporates Globally

Jin-Chuan Duan*

(September 8, 2017)

Abstract

Corporate credit default swap (CDS) premium is the market price of credit risk posed by a corporate obligor. Although corporate CDS are commonly used for risk benchmarking in accounting and credit risk management, liquid CDS are limited to less than 500 corporate names globally. CDS users must either confine their usage to this limited subset or resort to aggregates derived from the liquid CDS in different industry/rating combinations. This paper offers an intuitive, practical and robust predictive regression model linking liquid USD-denominated CDS premiums of different tenors to a set of obligor-specific attributes, and with the model one can generate proxy CDS curves for corporates without liquid or traded CDS. One key attribute is the actuarial spread that reflects the actuarial value of a CDS contract and is made available by the Credit Research Initiative of National University of Singapore for all exchange-listed firms globally. Other attributes in the predictive regression model include investment vs. speculative grades based on an obligor's credit rating, and some general credit environment variables, among others. This predictive regression, constructed with the historical record on 405 corporate CDS names, enables daily production of proxy CDS curves on around 35,000 exchange-listed corporates globally.

Keywords: Acturial spread, distance to default, credit cycle index, investment grade, high yield, big-data, zero-norm penalty, sequential Monte Carlo.

*Duan is with the National University of Singapore (Business School, Risk Management Institute and Department of Economics). E-mail: bizdjc@nus.edu.sg. The author acknowledges CriAt, a FinTech company specializing in deep credit analytics, for making available its proprietary software for selecting regressors subject to a zero-norm penalty, and also thanks Siyuan Huang of CriAT for her able research assistance.

1 Introduction

Corporate credit default swap (CDS) premium is the market price of credit risk posed by a corporate obligor, reflecting probability of default, recovery rate on the reference debt instrument, additional risk premium demanded by risk averse economic agents, and liquidity condition of the CDS market. CDS are commonly used for risk benchmarking in credit risk management in general and in accounting practice in particular, where the latter pertains to the soon-to-be enforced accounting reporting standards on credit exposures (IFRS-9 for international firms and CECL for US firms). However, liquid CDS are hard to come by and with no more than 500 corporate names globally. CDS users must either confine the usage to this limited subset of liquid CDS or simply resort to some aggregates derived from the liquid CDS in different industry/rating combinations made available by some commercial vendor, say, Markit.

This paper offers an intuitive, practical and robust predictive regression model linking liquid USD-denominated CDS premiums of different tenors to a set of obligor-specific attributes, and with the model one can generate proxy CDS curves for corporates without liquid or traded CDS. Our model is a single equation for all corporate CDS over a long time span, which delivers an R^2 of over 80% and performs robustly for different subgroups of interest. The key to our success is to utilize the actuarial spread (AS) computed by the Credit Research Initiative (CRI) of National University of Singapore, which generates daily updated ASes, among other credit risk measures, for all exchange-listed firms globally, and makes the results freely accessible. AS as a predictor alone is found to deliver an R^2 of 48.6%. Other obligor-specific attributes in the predictive regression include investment vs. speculative grades based on an obligor's credit rating, and some general credit environment variables, among others. This predictive regression model, constructed with the historical record on 405 corporate CDS names, enables daily production of proxy CDS curves on around 35,000 exchange-listed corporates globally, reflecting the CRI's coverage of literally all exchange-listed firms in the world today.

CDS and corporate bond pricing is a much researched topic in the literature, and theoretical pricing models abound; for example, Merton (1974), Longstaff and Schwartz (1995), Duffie and Singleton (1999), Duffie and Lando (2000), Das and Sundaram (2000), and Hull and White (2000), to name just a few. By design, these theoretical models mainly focus on the risk premium arising from risk aversion of economic agents, and are typically stylized in a way to avoid practical complications such as multiple risk drivers, liquidity, or supply-demand imbalance. These theoretical models also come with unknown parameters that need to be estimated, and some of the models go further to rely on latent variable(s), for example, unobserved default intensity. In order to have reasonable empirical performance, the unknown parameters and/or latent variable(s) need to be estimated with market prices on some credit-sensitive instruments such as corporate bonds and/or CDS on the obligor in question. Since the model parameters and/or latent variable(s) are obligor-specific, they cannot be easily ported for use on CDS referencing other obligors. Practically speaking, these theoretical models are limited in application to the pricing of CDS on corporates with traded bonds and/or CDS.

Empirical studies of corporate CDS are even more numerous to cover all. Most studies were

designed to focus on whether CDS are priced according to some theory as opposed to addressing how CDS can be practically priced through a predictive relationship developed on other liquid traded CDS. Ericsson, *et al* (2009), for example, studied the CDS premium in relation to three general theoretical predictors – leverage, volatility and riskless interest rate – on a firm-by-firm basis to find an average R^2 in the order of 60%. When dealing with the three predictors on an individual variable basis, the average R^2 drops to less than 15%. Since the regression is run on a firm-by-firm basis, the coefficients developed on one corporate with traded CDS cannot be used for other corporates without CDS even if one is satisfied with the level of R^2 based on the three-variable model. In short, their study confirms the theoretical prediction by addressing the issue of "why" but offers no practical answer to "how". The relationship of CDS premium vs. corporate bond yield spread (risky bond yield minus riskless bond yield) of the same tenor has been the subject of many empirical studies, for example, Blanco, *et al* (2005) and Zhu (2006) confirmed a long-run parity relationship between the two credit risk measures. Kim, *et al* (2017) further investigated the basis (CDS premium vs. corporate bond yield spread) behavior to see whether basis arbitrage is possible. This line of studies again sheds light on whether a theory holds or arbitrage opportunity exists, but offers no practical answer to predicting CDS for corporates without traded bonds of comparable terms.

Our CDS prediction model utilizes the advancement in modern big-data analytics, particularly the zero-norm penalty regression, which means that one chooses an optimal subset of k regressors among all potential predictive variables. When the number of potential regressors increases to, say, several hundred, the number of possible combinations quickly becomes astronomical, making an exhausted search infeasible even with modern high-power computers. In this paper, we implement the zero-norm penalty regression by a software made available by CriAT, a FinTech firm specializing in deep credit analytics, which utilizes a proprietary sequential Monte Carlo algorithm. Modern penalty regression techniques are typically based on l_1-norm because of their computational efficiency; for example, the Lasso of Tibshirani (1996), the SCAD of Fan (1997) and Fan and Li (2001), and the adaptive Lasso of Zou (2006). However, selecting regressors based on the zero-norm penalty is conceptually more appealing, because it directly addresses the essence of the variable selection problem. Computing speed aside, it also works better because regression coefficients are not distorted by the penalty term (i.e., shrinkage toward zero even being selected). Also interesting to note is the fact that the regression model fit, measured by R^2, is invariant to rotating a group of regressors but the corresponding l_p $(p > 0)$ penalty term is not. Therefore, multicolinearity will interfere with regressor selection based on an l_p $(p > 0)$ penalty, but not with the zero-norm regressor selection.

We consider 28 variables in predicting the observed USD-denominated CDS premiums for 405 reference obligors with tenors from 1 to 5 years on a monthly frequency over the period from August 2001 to February 2017. In addition to US corporates, there are firms from 21 other economies, totaling 141,918 observations in 10 industries according to the Bloomberg Industry Classification System. The 28 variables give rise to 390 potential regressors when interaction terms are considered. Our zero-norm penalty regression selects an optimal combination of 25 regressors among 390 possibilities and many of them are interaction terms, where the optimal number is determined by

applying the BIC. Our predictive regression model delivers an R^2 of 80.70% overall and is found to be stable across different subgroups. It is particularly worth noting that the impact of the 2008 global financial crisis seems to have been absorbed into other variables, leaving the predictive regression model to perform well over the time period which presumably has a structural break, even though the dummy variable reflecting the post-crisis period has not been chosen as part of the predictive model.

2 Constructing proxy CDS curves

Our approach to constructing a practical way of generating proxy CDS curves on five specific tenors (1, 2, 3, 4 and 5 years) is entirely empirical but guided with economic intuition. We first gather a substantial sample of USD-denominated CDS market premiums, spanning over 15 years on a monthly frequency for as many corporate names as we can obtain. Next, we move on to identifying a set of attributes that are concurrently available and intuitively related to the market price of CDS. By considering the interaction terms of these attributes, we obtain a very large set of potential explanatory variables, which in fact equals 390. Finally, we rely on a zero-norm penalty based variable selection technique to identify a subset of 25 regressors (including the intercept term) that can best predict CDS market premiums robustly.

2.1 The CDS data

The CDS market premiums are the Bloomberg computed CDS averages with end-of-day set to 6:00pm EST (New York time). We focus on USD-denominated CDS and extract data from Bloomberg on a monthly frequency starting in August 2001 all the way to February 2017. The 405 corporate names in our extracted USD-denominated CDS sample include beyond US firms (309 out of 405) to cover firms from 21 other economies with Canadian firms being the second largest group (20 out of 405). The firms in the sample covers all 10 industries according the Bloomberg Industry Classification System with Financial being the largest containing 73 firms and Diversified being the smallest having 4 firms. The five CDS tenors are fairly equally distributed where 354 firms with 1-year, 319 with 2-year, 356 with 3-year, 314 with 4-year and 404 with 5-year. On and after the 2008 global financial crisis (defined as September end, 2008), the sample contains 395 firms, whereas before it has 244 firms. The CDS sample contains 141,918 observations in total with 118,559 being investment-grade and the rest being the high-yield. Some descriptive statistics on our CDS data with and without the natural log transformation are provided in Tables 1 and 4.

Our sample also contains 92 data points on CDS referencing subordinated debt, and all are 5-year tenor with Shinshei Bank, a Japanese financial institution, as the reference entity. The data on this subordinated debt CDS fall in the period from April 2006 to December 2013. The sample suggests that a great majority of CDS only references senior unsecured debt.

The aforementioned categorical data characteristics will be used along with some other more granular attributes concerning individual corporate names in constructing our proxy CDS model.

Table 1: Single-regressor R^2 and summary statistics of the regressors

	R^2	Mean	Std	Max	Min
CDS(bps)		150.0377	328.3817	9592.2010	1.2350
logCDS		4.2253	1.1715	9.1687	0.2111
Regressor					
logAS	0.4860	2.1944	1.8994	9.6152	-11.9240
logASlevel	0.4718	2.3773	1.7463	8.3655	-10.2444
logAStrend	0.0333	-0.1830	0.6830	2.4696	-8.1830
DTDlevel	0.3819	5.5199	3.0600	20.1084	-1.1757
DTDtrend	0.0268	0.1090	1.3557	6.1350	-7.0466
SIGMA	0.3769	0.0794	0.0555	0.9492	0.0233
SIZElevel	0.2616	3.5422	1.4390	8.1375	-2.2648
SIZEtrend	0.0216	-0.0037	0.1794	1.6448	-1.8962
TL/TA	0.0534	0.6690	0.1789	2.0325	0.1206
NI/TAlevel	0.1566	0.0038	0.0059	0.0761	-0.0603
NI/TAtrend	0.0016	0.0000	0.0066	0.1044	-0.1459
CASH/TAlevel	0.0011	0.0910	0.1092	0.9785	0.0000
CASH/TAtrend	0.0003	0.0011	0.0294	0.4829	-0.3337
logIndustryCCI	0.3023	2.8066	0.4200	3.8135	1.1936
logCountryCCI	0.2894	2.8950	0.6612	5.1790	-0.8228
3mRateEcon	0.0155	0.9587	1.7389	23.7700	-0.0800
3mRateUS	0.0556	0.6785	1.3026	5.1239	-0.0203
SwapSpread5vs1	0.0113	1.1097	0.6374	2.7300	-0.3562
VIX	0.1589	21.6953	9.6061	59.8900	10.4200
Tenor-1y	0.0800	0.1800	0.3842	1.0000	0.0000
Tenor-2y	0.0038	0.1440	0.3511	1.0000	0.0000
Tenor-3y	0.0011	0.1884	0.3910	1.0000	0.0000
Tenor-4y	0.0116	0.1462	0.3533	1.0000	0.0000
isHY	0.2745	0.1646	0.3708	1.0000	0.0000
isSub	0.0012	0.0006	0.0255	1.0000	0.0000
isUS	0.0000	0.8585	0.3485	1.0000	0.0000
isFinancial	0.0092	0.1467	0.3538	1.0000	0.0000
isAfterCrisis	0.0404	0.7871	0.4093	1.0000	0.0000

And some of the categorical features indeed play a prominent role in explaining CDS market premiums, and can help predict CDS values when their market prices are unavailable.

2.2 The variables used to predict CDS curves

Variables that capture financial conditions of individual corporate obligors and reflect general economic environment are natural candidates for predicting CDS premiums. With the availability of the Credit Research Initiative (CRI) database, a CDS-like credit risk measure, known as actuarial spread (AS), constructed with physical default probability (PD) term structure is readily available on a daily basis on all exchange-listed firms worldwide. Also available on the CRI database are (1) a suite of daily series of credit cycle indices (CCIs) capable of reflecting the credit environment in general or for different industries, and (2) distance-to-default (DTD) estimates for individual firms which loosely speaking is an asset volatility adjusted leverage measure. In the following, we will briefly describe the CRI database, AS, CCI and DTD.

The CRI, launched in 2009 at the National University of Singapore in response to the 2008 global financial crisis, was conceived as a *public good* undertaking to contribute to credit rating reform. The CRI makes freely available its PDs and other credit risk measures. The CRI-PDs are computed with the forward-intensity model of Duan, *et al* (2012), which was designed for obtaining the PD term structure while factoring in the censoring effect arising from other corporate exits such as M&As. The CRI coverage includes practically all exchange-listed firms globally, and its PDs (1 month to 5 years) and ASes (1 year to 5 years) are updated daily for about 35,000 currently active firms. Historical time series are also available on 65,000 plus firms including those delisted firms due to bankruptcies, M&As and other reasons.[1]

The PD term structure is useful in many applications. A 5-year CDS is, for example, sort of a complex average of PDs over the life of the contract mixed with recovery rate, risk premium demanded by market participants, and market liquidity. Duan (2014) showed how AS can be constructed from the PD term structure to mimic CDS of any tenor except for leaving out risk premium and market liquidity. In short, AS is the actuarial value of CDS, which is obviously the closest risk measure to CDS without committing to a specific CDS pricing model. Since the CRI also makes freely available the daily updated ASes (1- to 5-year tenors) on all exchange-listed firms globally, AS becomes our natural candidate for predicting CDS. This choice is clearly supported by the individual R^2 reported later where logAS is shown to have an R^2 of 48.6% on a single-regressor basis in explaining logCDS, the highest among all variables considered.

In addition to logAS, we also consider its transforms – the level and trend of logAS – in a way similar to the CRI default predictor treatment. The 12-month moving average of logAS is considered logASlevel whereas logAS minus logASlevel is referred to as logAStrend. These three variables are obviously linearly dependent by design, but we include all three in the set of 28 potential regressors. Choosing a subset of regressors subject to a zero-norm penalty will naturally

[1]For the technical details on how these PDs and ASes are computed, readers are referred to NUS-RMI Credit Research Initiative Technical Report Version: 2017 Update 1 available at http://www.rmicri.org.

avoid picking all three because having all three does not increase explanatory power but add to the penalty.

In Duan and Miao (2016), a suite of credit cycle indices (CCIs) were used to describe the credit environment. The country CCI at a particular time point is in our deployment the median AS value for a corresponding tenor where the median is taken over all exchange-listed firms domiciled in that country at that time point. Likewise, industry CCIs are the median ASes for the 10 industries globally according to the Bloomberg Industry Classification System. Our CCIs differ from those of Duan and Miao (2016) in two aspects. First, we use AS in stead of PD because our interest is on CDS where the AS has been constructed with the CDS convention in mind. Second, we use the original median series instead of further subjecting 10 industry CCIs to orthogonalization. In our case, the CCIs are simply used as regressors and the correlations among the CCIs do not affect our regressor selection because the selection technique deployed relies on the zero-norm penalty. Naturally, CDS pricing is expected to reflect the credit environment in general as well as those in different industries, and CCIs are simply used as credit environment indicators.

DTD based on the structural credit risk model of Merton (1977) is a commonly adopted measure in credit analysis. Although the concept is standard, its implementation can be challenging due to the fact that the underlying firm asset value in the call option like theoretical setup of Merton (1977) is a latent stochastic process. The Moody's KMV approach has been widely adopted in both academic and commercial applications, which relies on an iterative scheme to estimate the unknown model parameters, the latent firm asset value, and finally the DTD. However, the Moody's KMV approach has statistical shortcomings because it fails to properly account for the Jacobian arising from the call option pricing function, and thus causes some biases. The Moody's KMV approach also specifies a default point formula (100% short-term debt plus 50% long-term debt), which serves as the strike price in the call option analogy. Interestingly, the missing Jacobian also places an implementation limitation whenever the default point formula justifiably needs an expansion to include other liabilities subject to an unknown haircut. Adding to the Moody's KMV default point formula is evidently important for financial institutions where a large portion of corporate liabilities is classified as neither short-term nor long-term; for example, deposits in banks and policy obligations in insurance companies. The CRI database generates DTDs using the maximum likelihood method of Duan (1994, 2000) modified in Duan, *et al* (2012) to accommodate other corporate liabilities.[2]

In addition, we consider common drivers such as interest rate, interest rate term spread and VIX, and individual firm attributes like funding liquidity, leverage, profitability, size, and idiosyncratic equity return volatility. These variables along with the categorical CDS characteristics described earlier are summarized in Table 1. Also reported in Table 1 are the individual R^2 when each of these variables is used as a single regressor along with an intercept term. The results suggest that logAS is the best logCDS predictor with a R^2 of 48.6% and closely followed by logASlevel at 47.2%,

[2]Readers who are interested in technical details are referred to these papers and/or NUS-RMI Credit Research Initiative Technical Report Version: 2017 Update 1. For a more friendly exposition and direct evidence on estimation consequences of different estimation methods, readers are referred to Duan and Wang (2012).

and many of these 28 potential regressors have a decent R^2. It is worth noting that the categorical CDS characteristic like isSub (equals 1 if the CDS references a subordinated debt) has a minuscule R^2 at 0.12%, but our later results show that it still plays a meaningful role when combined with other regressors. Table 2 provides correlations among selected regressors. It is clear from this table that some regressors are highly correlated, for example, logASlevel and DTDlevel.

2.3 Linear regression subject to a zero-norm penalty

In a general classical linear regression setting, one attempts to relate a dependent variable to k regressors where there are n data points. When there are too many potential regressors vis-a-vis the number of data points, in-sample over-fitting is expected and removing some regressors becomes both conceptually sensible and practically necessary. There has been a long-standing interest in designing theoretically sound and practically implementable methods in selecting regressors. In order to have a precise discussion, we state the regressor selection problem as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)'$, and $\boldsymbol{X}$ denotes the n observations of p regressors, i.e., $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)$ with $\boldsymbol{x}_i = (x_1, \cdots, x_n)'$, of which the first vector may represent the intercept term. $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)'$ is the k-dimensional regression coefficients, and $\boldsymbol{\epsilon}$ is n-dimensional *i.i.d.* normally distributed errors with mean 0 and variance σ^2. The task is to select $k_s \leq k$ regressors meeting some criterion, or alternatively, to set some β's to zero.

The classic way of performing such a task is a greedy-search technique that starts with one regressor with the highest R^2, finds the second regressor that delivers the highest R^2 in explaining the residual from the one-regressor model, and then repeats the search sequentially until the stopping criterion is reached. The greedy-search technique is known to be suboptimal because a combination of, say, two regressors may deliver a better predictive power while they individually do not produce top explanatory power. In principle, one could exhaust all possible combinations to find the ideal subset of k_s regressors. Practically speaking, however, it is not feasible when the number of potential regressors is large; for example in this paper, we identified 25 regressors out of 390 potential explanatory variables that means 1.7566×10^{39} possible combinations in total.

The modern way of performing regressor selection is through an l_1-norm penalty, commonly known as Lasso, advanced by Tibshirani (1996) and subsequently improved by, for example, SCAD of Fan (1997) and Fan and Li (2001), and the adaptive Lasso by Zou (2006). The Lasso and its variants have found great popularity in big-data applications these days due to their simplicity and computational efficiency. However, regressor selection based on the l_1-norm penalty is not most conceptually appealing albeit its practicality. This is because regression coefficients will be distorted by the penalty term (i.e., shrinkage toward zero even being selected). Even though the SCAD and adaptive Lasso do have the oracle property[3], i.e, distortion disappears when the sample size approaches infinity, it is mostly a feature that bears limited practical relevance because in

[3]See Fan and Li(2001) for a formal definition of the oracle property.

Table 2: Correlations for a subset of regressors

Regressor	logAS	3mRateUS	SwapSpread5vs1	VIX	logIndustryCCI	DTDlevel
logAS	1.0000	-0.0138	0.1198	0.2545	0.6132	-0.7982
logASlevel	0.9332	-0.0507	0.1672	0.1843	0.5514	-0.8483
logAStrend	0.3948	0.0912	-0.0945	0.2364	0.2954	-0.0508
DTDlevel	-0.7982	0.0953	-0.2267	-0.2505	-0.4080	1.0000
DTDtrend	-0.2395	-0.0943	0.1976	-0.3644	-0.1819	-0.0525
SIGMA	0.5451	-0.0837	0.2136	0.3317	0.3730	-0.5937
SIZElevel	-0.3456	0.1368	-0.0455	-0.0611	-0.0592	0.4050
SIZEtrend	-0.1712	-0.0166	0.0974	-0.0949	-0.0381	0.0152
TL/TA	0.2806	-0.0186	0.0001	-0.0118	0.0301	-0.3167
NI/TAlevel	-0.4129	0.0550	-0.0966	-0.0300	-0.1531	0.4847
NI/TAtrend	-0.0246	0.0017	0.0391	-0.0551	-0.0305	-0.0123
CASH/TAlevel	0.0409	-0.0314	-0.0097	-0.0351	-0.0422	-0.0704
CASH/TAtrend	-0.0043	-0.0125	0.0827	0.0191	0.0389	-0.0400
logIndustryCCI	0.6132	-0.0326	0.1640	0.4785	1.0000	-0.4080
logCountryCCI	0.5783	-0.0677	0.2243	0.4248	0.8026	-0.3261
3mRateEcon	0.0012	0.7492	-0.3832	-0.1539	0.0148	0.0274
3mRateUS	-0.0138	1.0000	-0.5229	-0.2236	-0.0326	0.0953
SwapSpread5vs1	0.1198	-0.5229	1.0000	0.1435	0.1640	-0.2267
VIX	0.2545	-0.2236	0.1435	1.0000	0.4785	-0.2505
Tenor-1y	-0.3572	-0.0039	-0.0006	0.0126	-0.5033	0.0136
Tenor-2y	-0.0669	-0.1505	0.0589	0.0505	-0.0965	-0.0048
Tenor-3y	0.0180	0.0030	-0.0046	0.0036	0.0580	0.0208
Tenor-4y	0.0946	-0.1553	0.0626	0.0477	0.1466	0.0053
isHY	0.3487	-0.0578	0.0113	0.0335	0.0753	-0.3871
isSub	0.0329	0.0117	-0.0086	0.0012	0.0304	-0.0424
isUS	0.0155	-0.1046	0.0437	0.0155	-0.0927	0.0274
isFinancial	0.1956	0.0256	-0.0150	-0.0152	0.2763	-0.2489
isAfterCrisis	-0.0014	-0.8511	0.2946	0.2139	-0.0082	-0.0717

applications the sample size vis-a-vis the number of regressors is unlikely large enough. A more practical concern is perhaps the issue of multicollinearity which analysts inevitably encounter in practice. To understand this point, let us rotate a group of mutually independent regressors to become linearly dependent regressors, knowing that such rotation will not alter the regression model fit, measured by R^2. However, the l_1 norm of the regression coefficients is not invariant to a rotation, and hence the rotation will change the model's l_1 penalty, giving rise to a different penalized estimation outcome. In short, multicollinearity may lead to an undesirable variable selection outcome when a l_1-norm based method is deployed. This concern is not a pure theoretical possibility, because the situation repeatedly occurred in this author's many practical applications in the area of credit analysis.

In principle and probably without contention, a more appealing and direct approach to regressor selection is to pick a fixed number of regressors, say, k_s, where the selection is optimally conducted by minimizing squared residual errors, i.e., the l_2 norm, over all possible combinations. As to k_s, it can be determined, for example, by applying the BIC. Such a variable selection approach is known as applying a zero-norm penalty in the sense of David Donoho, a definition commonly adopted in scientific computing and big-data analytics. It can be viewed as the zero norm because the standard l_p norm approaches this zero-norm when p goes to zero even though such a limiting l_0 "norm" is, strictly speaking, not a proper norm for its missing homogeneity. The penalized regression subject to the zero-norm regularization can be formally stated as

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_{l_2}^2 \\ &s.t.\ \ ||\boldsymbol{\beta}||_{l_0} \leq k_s \leq k \end{aligned} \tag{2}$$

where $||\cdot||_{l_2}$ is the l_2-norm and $||\cdot||_{l_0}$ is the zero-norm, which counts the number of non-zero entries in $\boldsymbol{\beta}$. Also worth noting is the fact that the above minimization problem is equivalent to $\arg\min_{\boldsymbol{\beta}} \left\{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_{l_2}^2 + \lambda||\boldsymbol{\beta}||_{l_0}\right\}$ where the solution is a step function of λ with the jumps corresponding to different values of k_s. This zero-norm penalized regression problem is known to be NP-hard. But the benefit is that this variable selection approach is free of the distortion caused by interference of the l_2-norm objective with the l_1-norm penalty in the presence of multicollinearity. What preventing its adoption in practice is the computational challenge in dealing with extremely large possible combinations that we alluded to earlier. Typical solutions are by approximating the l_0 norm with a penalty function very close to it, for example, Dicker, *et al* (2013). Here we are able to implement the zero-norm regressor selection without approximating the penalty function through a sequential Monte Carlo algorithm developed by CriAT, a FinTech firm specializing in deep credit analytics.

Specifically, we apply CriAT's proprietary software on a randomly selected subsample of 3,000 CDS observations along with their attributes mentioned earlier. In order to ensure CDS referencing subordinated debt are in the subsample, we have included all 92 observations in this category as explained earlier, and the remaining 2908 CDS data are randomly sampled from the remainder. Using a smaller subsample to identify the optimal combination of regressors for a given k_s can significantly speed up finding the zero-norm solution. Once the optimal combination under a k_s is

identified, we then apply the same set of regressors on the entire data set with 141,918 observations, and use the BIC to determine the optimal k_s.[4] Cross-comparing the R^2 of the remainder sample (i.e, the whole sample excluding the subsample of 3,000 observations) with the subsample's R^2 is also a good way of checking whether the in-sample and out-of-sample performances are comparable. Thanks to the CriAT software, we were able to obtain the optimal zero-norm solution using the BIC within several hours on a standard desktop computer, which singles out 25 regressors from 390 potential variables (including all meaningful interaction terms).[5] Note that the intercept term is treated as a potential regressor and the final result suggests that the intercept has been chosen.

3 Performance of the proxy CDS curves

Our predictive regression model selected with the zero-norm penalty and by applying the BIC has 25 regressors (the intercept term included). These 25 regressors are selected with an R^2 of 81.97% using a random subsample of 3,000 observations. When we apply the same set of selected regressors and fix at the previously estimated coefficients to the remainder sample of 138,918 observations, the R^2 only drops slightly to 80.50%, suggesting no over-fitting. The model is then estimated to the whole sample of size 141,918 using the same set of 25 selected regressors to yield an R^2 of 80.70%, much higher than the largest single-regressor R^2 of 48.6% using logAS as reported in Table 1. As Table 3 shows, the regression coefficients change little from the subsample to the whole sample, implying that this prediction regression model is very stable.

We have argued earlier that AS is conceptually a variable closest to its corresponding CDS, which was confirmed by the single-regressor R^2 reported in Table 1. Thus, it is reasonable to expect logAS or its close variant, i.e., logASlevel, to be among the selected regressors, but not both. Interestingly, logASlevel instead of logAS has appeared in the final set, and in fact has appeared four times with one being its own square and the other three interacting with other variables (3mRateEcon, VIX and logIndustryCCI). Using the coefficients based on the whole sample reported in Table 3, logCDS is identified to respond to logASlevel with a variable coefficient, i.e., 0.0096×logASlevel - 0.0314×3mRateEcon - 0.0023×VIX + 0.1044×logIndustryCCI, implying that CDS's relation to AS depends positively on its own AS value and the credit cycle index for the industry that the obligor is in, but inversely on the VIX index and the interest rate of the economy that the obligor is in. Likewise, CDS responds to the US interest rate and USD swap spread (5-year minus 1-year) with variable coefficients, and these variable coefficients depend on binary indicators such as whether the corporate is a financial firm and/or high-yield.

[4] BIC(k_s) is defined as $n \ln\left(\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{U}^*(k_s)}\hat{\boldsymbol{\beta}}(\boldsymbol{U}^*(k_s))||_{l_2}^2\right) + (k_s+1)\ln n$ where $\boldsymbol{X}_{\boldsymbol{U}^*(k_s)}$ represents the chosen regressors with $\boldsymbol{U}^*(k_s)$ being a vector containing 0 and 1 indicating the locations of the chosen regressors, and $\hat{\boldsymbol{\beta}}(\boldsymbol{U}^*(k_s))$ is the optimal regression coefficients corresponding the chosen regressors. The optimal k_s regressors chosen are not unique in terms of permutations but unique in the sense of combinations.

[5] With the 28 explanatory variables, there are 29 potential regressors including the intercept term. If all interaction terms are considered, the maximum number of potential regressors is increased to 435 ($= 29 \times 30/2$). However, some of the terms are redundant when the intercept and/or a dummy variable is involved; for example, squaring a dummy variable yields exactly the same dummy variable, and the product of the intercept with the 28 original variables produces the same set of 28 variables. After trimming the redundant regressors, the total count drops to 390.

It is particularly worth noting that "isAfterCrisis" is nowhere to be seen in the selected regressors. We can interpret it as the impact of the 2008 global financial crisis has been absorbed into other variables. It also suggests that the predictive regression model is stable even over the time period which presumably has a structural break, a fact of which will be confirmed later by comparing two prediction plots generated by our model that does not include the crisis period indicator.

To examine whether the predictive regression model exhibits any bias behavior in different subgroups, we present a set of plots in Figures 1 and 2. The goood performance for the whole sample (14,918 observations for 405 corporates with five tenors over the entire sample period on the monthly frequency) is shown at the top-left plot of Figure 1 where the horizontal and vertical axes are respectively the logarithm of the predicted and observed CDS premiums in basis points; for example, 100 basis points equals 4.605. The plot for the CDS subsample referencing subordinate debts is at the top-right, showing a good performance. Since this subordinate debt group only has 92 observations and all for Shinshei Bank over the entire sample period. The whole sample result literally also represents the senior-debt group. One can see no discernable bias for different tenors either, for which we have plotted 1-year and 5-year CDS contracts but skipped the remaining three groups to conserve space. The difference in credit quality of the reference obligor (investment vs. speculative grade) does not seem to make any material difference as reflected in the bottom two plots in Figure 1.

Further comparisons (US vs. non-US and financial vs. non-financial reference obligors) are shown in Figure 2 where their performances are equally good. Finally, we compare the data before and after the 2008 global financial crisis with the post-crisis period defined as starting from the end of September 2008. Again, one cannot find any material difference pertaining to the potential structural break in the global financial system. The R^2 results for different groups along with their sample sizes are also summarized in Table 4, which corroborate the findings from viewing the plots. All R^2's for different groups are in excess of 71.95%.

4 Concluding Remarks

We have developed a robust predictive regression model for estimating CDS premiums for corporates who do not have traded or liquid CDS contracts. This predictive model has many applications for credit risk management in general and accounting practice in particular as far as the soon-to-be enforced IFRS-9 (for firms outside the US) and CECL (for US firms) financial reporting requirements are concerned. Our approach appears to be entirely empirical, but actually utilizes a critical theoretical result in connection with the actuarial spread model of Duan (2014), because it is this critical variable that makes the predictive model successful. Along the same line, one can in principle use a high-quality CDS theoretical pricing model to incorporate the additional premium due to risk aversion, and thus develops another measurement that can better predict the observed CDS premium. However, even a good CDS pricing model likely needs empirical fine-tuning in a way similar to our approach.

Our empirical model can be viewed as a concrete demonstration of modern big-data analytics in action. Critical to our success is the zero-norm penalty regression technique, which enables the identification of 25 regressors among 390 possibilities arising from 28 potential variables and their interaction terms. Although the 28 potential variables are picked due to their availability and economic intuition, identifying the optimal set of regressors facing the astronomical number of possible combinations would not have been possible without such a big-data analytical tool. Many other financial applications can obviously benefit from a similar approach.

References

[1] Blanco R., S. Brennan and I.W. Marsh, 2005, An Empirical Analysis of the Dynamic Relationship between Investment-Grade Bonds and Credit Default Swaps, *Journal of Finance* 60, 2255-2281.

[2] Das, S.R., and R.K. Sundaram, 2000, A Discrete-Time Approach to Arbitrage-Free Pricing of Credit Derivatives, *Management Science* 46, 46-62.

[3] Dicker, L., B. Huang, and X. Lin, 2013, Variable Selection and Estimation with the Seemless-L_0 Penalty, *Statistica Sinica* 23, 929-962.

[4] Duan, J.C., 2014, Actuarial Par Spread and Empirical Pricing of CDS by Decomposition, *Global Credit Review* 4, 51-65.

[5] Duan, J.C., 2000, Correction: "Maximum Likelihood Estimation Using Price Data of the Derivative Contract", *Mathematical Finance*, 10, 461-462.

[6] Duan, J.C., 1994, Maximum Likelihood Estimation Using Price Data of the Derivative Contract, *Mathematical Finance* 4(2), 155-167.

[7] Duan, J.C., and W. Miao, 2016, Default Correlations and Large-Portfolio Credit Analysis, *Journal of Business and Economic Statistics* 34(4), 536-546.

[8] Duan, J.-C., and T. Wang, 2012, Measuring Distance-to-Default for Financial and Non-Financial Firms, *Global Credit Review* 2, 95-108.

[9] Duan, J.C., J. Sun and T. Wang, 2012, Multiperiod Corporate Default Prediction – A Forward Intensity Approach, *Journal of Econometrics* 170(1), 191-209.

[10] Duffie, D., and D. Lando, 2000, Term Structures of Credit Spreads with Incomplete Accounting Information, *Econometrica* 69, 633-664.

[11] Duffie, D., and K.J. Singleton, 1999, Modeling term Structures of Defaultable Bonds, *Review of Financial Studies* 12, 687-720.

[12] Ericsson, J., K. Jacobs, and R. Oviedo, 2009, The Determinants of Credit Default Swap Premia, *Journal of Financial and Quantitative Analysis* 44(1), 109-132.

[13] Fan, J., 1997, Comments on "Wavelets in Statistics: a Review" by A. Antoniadis, *Journal of the Italian Statistical Society* 6(20, 131-138.

[14] Fan, J. and R. Li, 2001, Variable Selectionvia Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, 96, 1348-1360.

[15] Hull, J.C., and A.D. White, 2000, Valuing Credit Default Swaps I: No Counterparty Default Risk, *Journal of Derivatives* 8, 29-40.

[16] Kim, G.H., H. Li, and W. Zhang, 2017, The CDS-Bond Basis Arbitrage and the Cross Section of Corporate Bond Returns, *Journal of Futures Markets* 37(8), 836-861.

[17] Longstaff, F.A., and E.S. Schwartz, 1995, A Simple Approach to Valuing Risky Fixed and Floating Rate Debt, *Journal of Finance* 50, 789-819.

[18] Merton, R., 1974, On the Pricing of Corporate Debt: the Risk Structure of Interest Rates, *Journal of Finance* 29, 449-470.

[19] NUS-RMI Staff, 2017, NUS-RMI Credit Research Initiative Technical Report Version: 2017 Update 1, The Credit Research Initiative, Risk Management Institute, National University of Singapore (http://rmicri.org).

[20] Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Ser. B*, 58(1), 267-288.

[21] Zhu, H., 2006, An Empirical Comparison of Credit Spreads between the Bond Market and the Credit Default Swap Market, *Journal of Financial Services Research* 29, 211-235.

[22] Zou, H., 2006, The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, 101(476), 1418-1429.

Table 3: Regressor selection results based on the subsample of 3,000 observations and then applied to the whole sample

	Random Subsample			Whole Sample		
Regressor	**Estimate**	**Std Error**	***t*-Stat**	**Estimate**	**Std Error**	***t*-Stat**
Intercept	2.5412	0.1486	17.10	2.4980	0.0222	112.31
3mRateEcon	0.1923	0.0146	13.13	0.1827	0.0022	84.45
3mRateUS	-0.3675	0.0276	-13.33	-0.3646	0.0041	-88.72
VIX	0.0512	0.0074	6.92	0.0464	0.0011	42.11
logCountryCCI	0.7070	0.0461	15.35	0.7376	0.0070	104.81
SIZElevel	-0.5658	0.0315	-17.95	-0.5288	0.0046	-114.66
logASlevel^2	0.0107	0.0018	5.86	0.0096	0.0003	32.98
logASlevel*3mRateEcon	-0.0349	0.0034	-10.19	-0.0314	0.0005	-64.14
logASlevel*VIX	-0.0043	0.0008	-5.33	-0.0023	0.0001	-20.65
logASlevel*logIndustryCCI	0.1224	0.0081	15.12	0.1044	0.0012	87.11
3mRateUS*VIX	0.0065	0.0015	4.43	0.0064	0.0002	28.34
3mRateUS*isFinancial	-0.1687	0.0173	-9.78	-0.1310	0.0027	-48.97
3mRateUS*isHY	0.1655	0.0252	6.56	0.1331	0.0036	37.27
SwapSpread5vs1^2	-0.1906	0.0101	-18.84	-0.1843	0.0015	-123.37
SwapSpread5vs1*DTDlevel	0.0274	0.0041	6.74	0.0227	0.0006	38.61
SwapSpread5vs1*Tenor3y	-0.0827	0.0192	-4.30	-0.0570	0.0028	-20.71
SwapSpread5vs1*isHY	0.1950	0.0385	5.06	0.1494	0.0057	26.29
VIX*logCountryCCI	-0.0204	0.0023	-9.05	-0.0198	0.0004	-54.02
VIX*logIndustryCCI	0.0111	0.0018	6.08	0.0103	0.0003	36.88
TL/TA*isHY	0.2890	0.0728	3.97	0.4454	0.0106	41.98
logIndustryCCI*SIZEtrend	-0.2006	0.0178	-11.25	-0.2035	0.0027	-76.00
logIndustryCCI*isUS	-0.0679	0.0127	-5.35	-0.0668	0.0018	-36.63
CASH/TAlevel*isSub	16.4441	1.4083	11.68	16.3133	1.0263	15.89
SIZElevel^2	0.0378	0.0039	9.67	0.0340	0.0006	60.40
SIZElevel*SIGMA	0.9321	0.0930	10.02	0.9713	0.0134	72.42
R^2	81.97%			80.70%		
Sample Size	3,000			141,918		
BIC	-3,884.09			-188,238.57		

Table 4: R^2 of the proxy CDS model for the whole sample and various subcategories

	R^2	# of Reference Corporates	# of Data
Whole sample	80.70%	405	141,918
US	81.33%	309	121,840
Non-US	76.89%	96	20,078
Financial	81.92%	73	20,818
Non-Financial	80.20%	332	121,100
Investment grade	73.70%	370	118,559
High yield	71.95%	138	23,359
Senior debt	80.67%	404	141,826
Subordinated debt	93.66%	1	92
Before crisis(2008-09)	74.85%	244	17,696
After crisis(2008-09)	81.16%	395	124,222
Tenor(1 year)	78.84%	354	25,548
Tenor(2 years)	79.35%	319	20,432
Tenor(3 years)	78.29%	356	26,740
Tenor(4 years)	77.46%	314	20,750
Tenor(5 years)	77.57%	404	48,448

Figure 1: Performance of the proxy CDS model in predicting market price of CDS for the whole sample and different subcategories

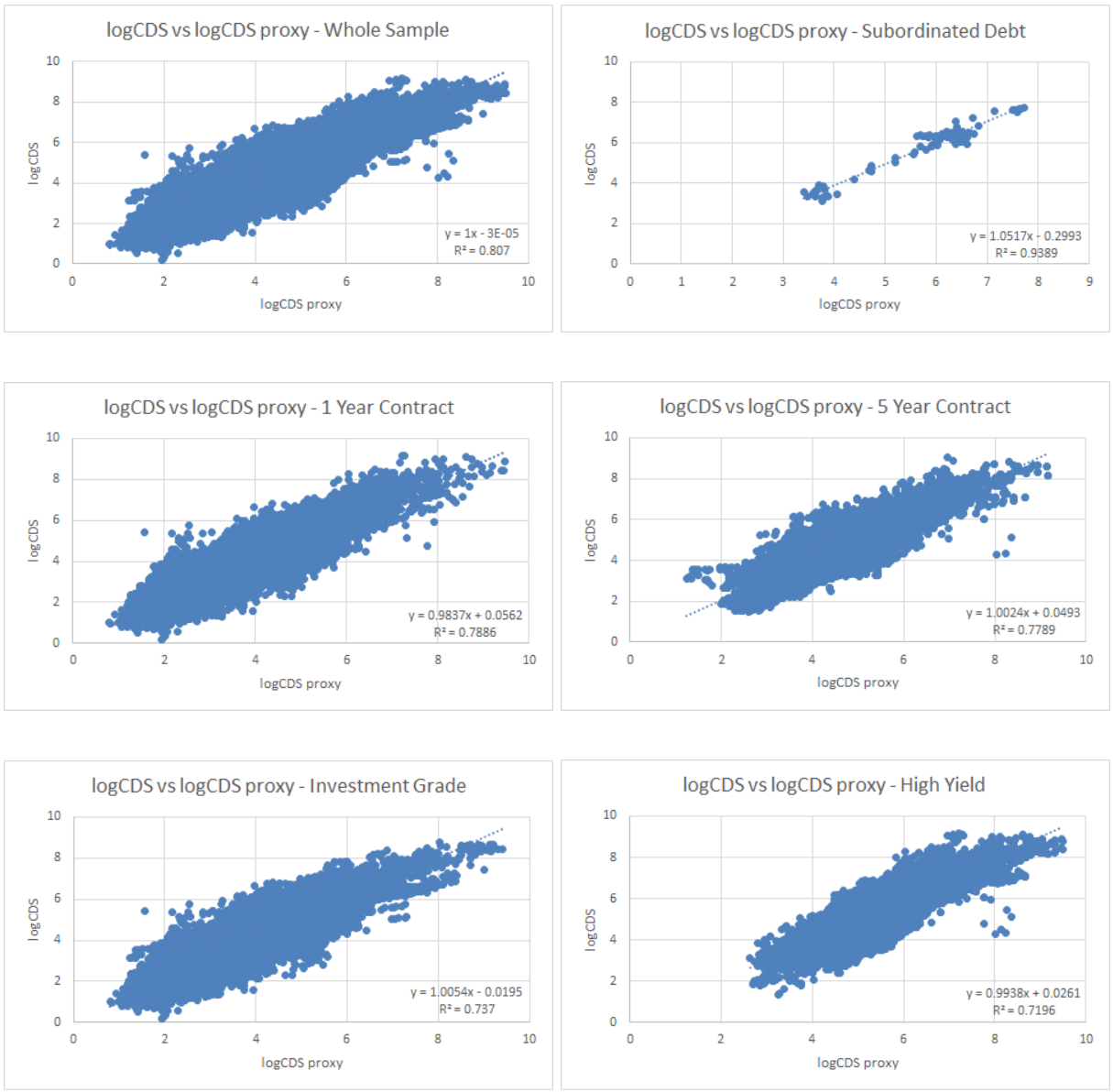

Figure 2: Performance of the proxy CDS model in predicting market price of CDS for more sub-categories