MIXED RESPONSE MODEL IN CREDIT RISK MODELING

HAO BAI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

December 2017

Abstract

Statistical methods are motivated by the desire of learning from data to solve problems in the real world. The credit risk management area of the banking book in the financial industry is a field extensively applying statistical knowledge to solve the problems and continually innovating new statistical methods. In credit risk area, a fundamental assumption of the probability of default (PD) rates for a portfolio is that the PD rates are monotonic increasing as the borrower's creditworthiness worsen. However, since the banks' internal data are not big enough, the empirical realized PD rates often violate this assumption. For the same reason, the violation of the assumption for the PD transition matrix also happens often. These violations will cause a severe problem if we directly calibrate the risk models based this non-smoothed empirical observed PD rates. We propose a smoothing algorithm for the observed PD rates and PD transition matrix by using Constrained Maximum a Posteriori (CMAP) method to solve these problems. The results from the proposed smoothing method are validated by simulation and real default data showing that

CMAP method can provide the smoothed and consistent PD rates. We also propose a new approach in this dissertation to estimate the correlated mixed response variable which is often found in credit risk area. The proposed approach simultaneously estimates the mixed response regression and estimates the correlation among the response variables. Moreover, we extend this methodology to the high dimensional mixed response regression models by using the pairwise composite likelihood method. The simulation results show that the proposed method can provide accurate coefficients and correlation for mixed response variables model.

Keywords: Smoothing, Constrained Maximum Likelihood Estimation, Maximum a Posteriori, Composite Likelihood, Newton-Raphson Method, Mixed Response

Acknowledgements

First and foremost, I would like to express my highest appreciation to my supervisor, Professor Xin Gao. Without her valuable guidance, constant encouragement and extensive knowledge, this dissertation would not be possible. I truly respect her brilliant insights and enthusiasm on this research.

I would like to extend my sincere appreciation to Professor Yuehua Wu, and Professor Dong Liang as members of my supervisory committee. My appreciation also goes to all faculty members, staffs and fellow graduate students in the Department of Mathematics and Statistics at York. I also would like to thank my classmates Qiong Li, Xiaoying Sun, my co-worker Bill Yang and Professor Xingmin Ji for their excellent suggestions on my dissertation.

Last but not least, I sincerely thank my parents for their continuous support. A special thank goes to my beloved wife for her patience, kindness, wisdom and endless support. This dissertation is dedicated to my lovely family who bring me happiness every day.

Table of Contents

List of Tables

List of Figures

1 Introduction

Statistical methods are motivated by the desire of learning from data to solve problems in the real world. The credit risk management area for the banking book in the financial industry is a field extensively applying statistical knowledge to solve the real problems and continually innovating new statistical methods. In this area, as information technology develops and historical credit data grows exponentially, it becomes possible to develop and apply more advanced statistical algorithms to accurately assess the risk faced by financial institutions and maximize profit under appropriate risk control. Banks experience loss as a result of lending to an individual or company that may default. Losses incurred in a particular year are volatile from year to year depending on the frequency and severity of credit events, even if a portfolio is assumed consistent over time. To precisely measure credit risk, financial institutions usually split credit risk into two categories: Expected Loss and Unexpected Loss. The Expected Loss is the reasonable expectation of the average level of credit losses in the portfolio and is treated as a cost component of the banking business and managed through risk management strategies. Unexpected Loss is the loss above the expected level which the banks know will occur sometime in the future, but cannot know their timing and severity in advance. To control credit risk, the banks need to hold adequate capital against loss caused by these credit risks and make sure that the capital adequately compensates for risks incurred. As a profit-seeking institution, however, banks have a great incentive to minimize the capital they hold and free up economic resources to efficiently invest the capital to make more profit. Consequently, precisely identifying, measuring, monitoring and controlling the credit risk is very important for financial institutions and calls for the use of the advanced, efficient and robust statistical methods.

The Expected Loss of a portfolio is viewed as the probability of default (PD) of the obligors within a specific time frame (typically 12 months), multiplied by the loss given default (LGD) rate and then multiplied by the outstanding exposure at default (EAD).

$$\text{EL} = \text{PD} * \text{LGD} * \text{EAD}$$

Since these three risk factors are random variables, banks know in advance neither the exact number of defaults within a portfolio during a specific time frame nor the exact amount outstanding nor the actual loss rate. However, banks can leverage sound statistical methods to calibrate these three parameters for a portfolio either through the top-down approach or through the bottom-up approach. The top-down approach estimates the risk parameters of PD, LGD and EAD from pre-defined risk segments for a portfolio. The bottom-up approaches directly calibrate the PD, LGD and EAD parameters based on the risk attributes of the individual account(facility). The expected loss of the portfolio then can be calculated based on these risk parameters.

As an introduction, here we briefly introduce the statistical methods involved in the credit risk field under the framework of Expected Loss.

The probability of default measures the possibility of default of a borrower taking a loan or credit within a specific time range. There are different types of PD for the different perspectives of risk management, such as through the cycle (TTC) PD and point in time (PIT) PD, conditional (stressed) PD and marginal (unstressed) PD. Accordingly, different statistical methodologies are applied to calibrate these different forms of PD rates. Usually, PD models for retail customers are built on the borrower's characteristics and their financial status, while PD models for commercial borrowers are built on their financial statement information and macroeconomic variables such as GDP growth, employment rate, and bond yield rate, etc. There are many research papers related to PD models which aim to predict the PD rate for a borrower. The pioneer PD model is Altman's Z-score model [1] which is actually a linear equation of the risk factors. As the PD rate measures the probability of a binary event of default or non-default, the logistic, ordinal or probit regression are widely used in the industry [2]. The decision tree [3] and Cox proportional hazards model [4, 5] are also applied to predict the borrower's PD rate. In addition, some complicated algorithms are adopted by the banks to calibrate the PD rate such as the one factor model [6], which considers the correlation between individual borrowers risk factors and the systematic (market) risk factors, or Markov chain Monte Carlo (MCMC) [7] algorithm, which calibrates the PD rates of no default or very low default risk ratings.

Loss given default (LGD) assesses the percentage of the exposure at the time of default that cannot be recovered from a borrower. Since the empirical distribution of the LGD rates at the facility level is bi-mode and bounded between [0,1] by definition (in the real data, the bound is not rigorously between [0,1] because of the existence of recovery cost and legal fees in the recovery process), the linear regression and fractional response regression are widely used in the industry to predict LGD [8, 9, 10]. The joint beta regression (inflated beta regression) model was proposed by Calabrese [11], which considers the dependent variables as a mixture of the continuous variable on (0, 1) represented by the beta distribution and a discrete probability mass at the boundaries of 0 and 1 (or we can set the bound to 1.5 as the LGD cap) represented by the Bernoulli distribution to model the LGD rate. In addition, the mixed effect model based on Vasicek's single factor framework [12] is applied to LGD calibration, which assumes LGD rate is dependent on both the observable risk factors and the unobservable systematic risk factor [13]. In the one factor models settings, the unobservable systematic risk factor works as a random effect term, and the idiosyncratic risk factor works as the residual term.

Exposure at default (EAD) measures the borrower 's outstanding exposure at the time of default. It has received much less attention than PD and LGD in the literature because it is less critical in the EL calculation by nature, and because of the difficulties that exist in modeling EAD because the distribution of EAD being comparatively random. When a practitioner estimates the EAD for a facility with an explicitly authorized limit amount, there are three typical link functions used to connect the EAD to the authorized limit: usage given default (UGD), credit conversion factor (CCF) and additional utilization factor (AUF). The EAD calibration methodologies focus on predicting link factors from the risk drivers of the borrowers. In the literature, Barakova and Parthasarathy introduce additional utilization factor (AUF) model [14], Asarnow and Marker [15] described the relationship between credit quality and utilization for the lower UGD. Araten and Jacobs [16, 17]proposed the UGD model with risk factors such as the commitment type and the commitment size. Moral [18] studied SME (small and medium enterprise) portfolio by optimization method using the various loss functions. Jacobs [19], Qi [20] and Taplin et al [21] proposed and investigated the credit conversion factor (CCF) model.

Currently, regulators ask the financial institutions to implement the expected credit losses (ECL) models under the IFRS9 framework, which requests the banks recognize the lifetime losses of a borrower more quickly compared to the expectation of losses for one-year time range only under the Advanced Internal Rating Based (AIRB) approach. To reach the expected lifetime credit losses, the risk practitioners apply forward PD models. One approach of forward PD rate models is through the rational PD transition matrix [22] to obtain the future time's PD rates for the borrowers. Under this circumstance, the rational PD transition matrix plays a critical role in the new ECL models and the rational PD rates for each risk rating grade, the components of the transition matrix, become very important. A fundamental assumption of the probability of default (PD) rates for a portfolio is that the PD rates are monotonic increasing as the borrower 's creditworthiness worsens. However, since the banks' internal data are not big enough, the empirical realized PD rates often violate this assumption. For the same reason, the violation of the assumption for the PD transition matrix also happens often. Theses violation will cause a severe problem if we directly use this empirical PD rate to calibrate the risk models. Accordingly, we propose a smoothing algorithm for the observed PD rates and PD transition matrix by using Constrained Maximum a Posteriori (CMAP) method to solve these problems in Chapter 2.

When we assess a borrower's risk attributes, different measurements are utilized including both binary (categorical) and continuous outcomes at the same time. For example, for a mortgage loan, it may have binary risk measures of default status and prepayment status, as well as continuous measures of prepayment amount and loss amount if it may default. These measures may or may not share common risk factors. The traditional approaches of building each model for each risk measure consider neither the correlation nor the trending movement among risk measures. Another practical problem in risk management is the correlation between the PD rates and LGD rates which is investigated by many researchers [23]. Most approaches calculate the correlation based on the observed PD rates and LGD rates, or based on the PD rates and the LGD rates separately calibrated in advance. We propose a new approach to solve this kind of correlated mixed response variable problems. It simultaneously solves the mixed response regression and estimates the correlation among the response variables. Moreover, we extend this methodology to the high dimensional mixed response regression models by using the pairwise composite likelihood method in Chapter 3.

2 Probability of Default Smoothing through Constrained Maximum a Posteriori

In this chapter, we focus on the smoothing algorithm using the Constrained Maximum a Posteriori(CMAP) method to estimate rational PD rates and PD transition matrices in the credit risk area for observed non-monotonic PD rates and PD transition matrices.

2.1 Introduction

It is very important that a regression relationship is monotonic in a specific range in many applications, such as in the credit risk field people assume that the probability of default rates is monotonic increasing as the credit quality of associated risk rating grades become worse, and in the actuarial science the mortality rate follows the specific bathtub curve shape in specific population. However, the empirical distribution of the internal data is usually not rigorously follow the expected distribution because the data samples from only internal data are not big enough to represent the population. This issue of non-monotonous empirical observation will cause severe problems when people want to build the statistical models further based on the empirical distribution to predict some interested relationship. Monotone smoothing methodology is used to solve the non-monotonic observed distribution problem. Many existing monotonic smoothing methodologies either may not result in desirable pattern as the industry expectation or are very complicated to implement. In this chapter, we propose a simple and efficient monotonic smoother for the probability of default (PD) ratio in credit risk field based on Constrained Maximum a Posteriori (CMAP) method leveraging the monotonic increasing attribute of logarithm function. The CMAP method can incorporate the prior knowledge of the parameters which gives us more reasonable estimation to the reality compared to maximum likelihood method. Specifically, in PDs estimation, there is a merit of CMAP method: the estimated PDs can incorporate the historical economic cycle with the prior knowledge of PDs for different rating grades. This property makes the estimated PDs not only monotonic in the single dimension of risk rating grades but also consider the second dimension of economic cycle represented by the cohort of the fiscal year. It will let the financial institutions build the forward-looking PD term structure model and plan their capital more rationally. The testing of monotonicity is demonstrated via the $Standard$ & $Poor's$ ($S\&P'$s) historical real default data and a toy data (small changed real default data from an anonymous financial institution) study.

In credit risk area, financial institutions including banks and non-bank financial companies usually follow or map their own risk rating system to the $Moody's$ $Investors\ Service's$ ratings system or $S\&P'$s rating system to assign a rating grade to a borrower or a instrument to represent the borrower or instrument's creditworthiness. Such as Moody's Investors Service's rating system, securities are assigned a rating from Aaa to C, with Aaa be the highest credit quality and C be the lowest. Accordingly, the PD rate is expected as monotonic increasing from lowest to highest for the rating grade from Aaa to C. This PD pattern has been proved by the long-term historical data industry-wide, as shown by the historical default rates span from the year 1981 to the year 2015 sourced from $S\&P'$s, which is illustrated in Table 2.1. . However, when the financial institutions calibrate the credit models based on the internal data requested by the regulator, they often encounter the problem of the non-monotonic empirical default rates and no default information for some rating grades in some time periods. Therefore, the smoothing algorithm must be applied to the overserved default rates based on the sparse internal default data to provide a smoothed PD rates before they were inputted to next modeling process.

Monotonic smoothing can be done by a variety of different algorithms, such as random smoothing, random walk model, moving average, and simple (linear or seasonal) exponential smoothing. All these techniques for monotonic smoothing have been proposed in the literature. Most of them are combined the monotone and smoothing algorithm together. In terms of monotonicity, a simple estimator can be modeled by isotonic (monotonic) regression through the Pool Adjacent Violators (PAV) algorithm ([24] and [25]). In terms of smoothness, smoother such as smoothing spline [26], kernel [27], regression splines [28], or local polynomial [29], penalized splines with monotonicity constraints [30] and isotonic regression combined with the local average for monotonic smooth of scatterplots [31] can be used.

Specifically, in the perspective of PD rates smoothing, Tasche [32] proposed a smoothing algorithm using quasi-moment matching (QMM) method which is a numerical solution of two-dimensional non-linear equations system. It assumes that both default and performing events in each rating grade are following the normal distribution and share the same variance but with different means and with a constraint that the mean value of maintaining the performing status is higher than the mean value of downgrade to default, which is empirically rational in the industry. For each rating grade at a specific time point, it is assumed that the PD rate follows the logistic regression since there are only two statuses of performing and default for a borrower, and the accuracy ratio is defined thereafter. The QMM algorithm numerically solves the equations to achieve the maximum accuracy ratio to get the estimated PD rates. The proper solutions from QMM method need a meaningful initial guess.

A Constrained Maximum Likelihood Estimation (CMLE) proposed by Yang [33] can also provide the smoothed PD rates by maximizing the likelihood. However, the estimations from this MLE based method often have the drawback of overfitting problem since it always offers us the parameters with the maximum likelihood fitting the training data. In addition, another drawback of MEL method is that it does not incorporate any prior knowledge of PD rates we have learned from the historical credit data. Maximum a Posteriori (MAP) estimation can improve this deficiency caused by MLE method. In this dissertation, we propose the Constrained Maximum a Posteriori (CMAP) algorithm to smooth the observed PD rates and use the QMM and CMLE as the benchmark.

The proposed smoothing CMAP algorithm and benchmark method CMLE are implemented in both R and SAS which call the general purpose optimization function of optim in R and proc nlmixed in SAS, which use the Newton−Raphson method to reach the optimum values. The benchmark smoothing method of QMM is already implemented as an R package called LDPD. Since Newton−Raphson method is leveraged to find the parameters estimation (roots) at the optimized value of the log-likelihood function for both smoothing algorithm and mixed response variable models. We briefly introduce the numerical algorithm here.

Newton−Raphson Method

Newton−Raphson method [63] is a numerical method which is widely used to find the optimized value of a real-valued function. The process is repeated until a sufficiently more accurate value is reached or the pre-defined threshold of accuracy reached. The iterative scheme is constructed by replacing initial value deduct the Jacobian matrix J multiply the inverse of the Hessian matrix H of the function $l_c(\boldsymbol{\eta})$. Let $\hat{\boldsymbol{\eta}}_t$ be the results from t_{th} iteration, then

$$\hat{\boldsymbol{\eta}}_{t+1} = \hat{\boldsymbol{\eta}}_t - H^{-1}(\hat{\boldsymbol{\eta}}_t)J(\hat{\boldsymbol{\eta}}_t),$$

where

$$J(\hat{\boldsymbol{\eta}}_t) = \left.\frac{\partial l_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\hat{\boldsymbol{\eta}}_t},$$
$$H(\hat{\boldsymbol{\eta}}_t) = \left.\frac{\partial^2 l_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\right|_{\hat{\boldsymbol{\eta}}_t}.$$

Newton−Raphson method can fail to get the optimum value of the parameters in three situations: the iteration is not converging to the global maximum value, the iteration is converging too slow, and there is no first-order derivative of the parameters. Thus, we need be careful to fix Newton−Raphson method in these situations. In our PD rates smoothing, we applied the constrained Newton−Raphson method, which use the theory of Lagrange multipliers to state first order necessary optimality conditions of the constraints.

2.2 Constrained Maximum Likelihood Smoothing

A typical credit data is organized this way: first, for a portfolio with N loans at the beginning of a fiscal year, a bank splits these N loans into K rating grades (segments) according to their credit quality. Assume that each loan in the same rating grade $R_i, (i = 1, ..., K)$ has the same probability of default (PD) p_i, N_i loans are assigned to rating R_i at the beginning of year, thus $N = \sum_1^K N_i$. There may have n default loans observed during a specific fiscal year and n_i defaulted loans observed from rating R_i during the fiscal year, then $n = \sum_1^K n_i$.

With the assumption that the credit quality is worse as the rating R_i downgrading, for a sequence of PDs $p_i, i \in [1, K]$, the rational PD rates should follow the monotonic increasing pattern and capped by 1 and floored with 0,

$$0 \leq p_1 \leq p_2 \leq ... \leq p_K \leq 1. \tag{2.1}$$

The borrower within each rating grade R_i has only 2 statuses will be at the end of year: Performing or Default, associated with the probability of $1 - p_i$ and p_i respectively. Obviously, this is a Bernoulli Distribution for each borrower in the rating grade R_i, and the default count n_i follows the binomial distribution $B(N_i, p_i)$, with the following probability mass function,

$$\binom{N_i}{n_i} p_i^{n_i} (1 - p_i)^{N_i - n_i}.$$

According to Miller and Freund [34], the log-likelihood function for rating R_i to be maximized of binomial distribution (Bernoulli trials) is given as

$$\begin{aligned} l_i(p_i) &= \log \binom{N_i}{n_i} p_i^{n_i} (1 - p_i)^{N_i - n_i} \\ &= \log \binom{N_i}{n_i} + n_i \log p_i + (N_i - n_i) \log(1 - p_i), \end{aligned} \tag{2.2}$$

and the MLE of p_i will be at the value which makes $\frac{\partial l_i}{\partial p_i} = 0$,

$$\hat{p}_i = \frac{n_i}{N_i},$$

which is the exact realized (empirical) default rate for the risk rating R_i.

Under the assumption that the default rates from each risk rating are mutually independent, the likelihood function of PDs for the portfolio will be

$$\prod_{i=1}^{K} \binom{N_i}{n_i} p_i^{n_i} (1 - p_i)^{N_i - n_i},$$

and the log-likelihood function l is

$$\sum_{i=1}^{K} \left[\log \binom{N_i}{n_i} + n_i \log p_i + (N_i - n_i) \log(1 - p_i) \right]. \tag{2.3}$$

By simply setting the score function to 0: $\frac{\partial l}{\partial p} = 0$, we will get the MLE of p_i for the portfolio as

$$\hat{p}_i = \frac{n_i}{N_i}, \qquad i = 1, ..., K.$$

Still, the MLE of p_i is the exact observed (empirical) default rates for each rating grade R_i of the portfolio.

In practice, however, we often encounter two prevalent obstacles needed to be resolved before further analysis with the empirical PD: one is the non-monotonic realized PDs ($\frac{n_i}{N_i} \geq \frac{n_{i+1}}{N_{i+1}}$), another is 0 default rates for some rating grade because there is no default in these rating grades during the observation time. Thus, when we calibrate the regression model based on these observed default rates to predict credit behaviors for the further use (such as AIRB, CCAR, IFRS9, etc.), it will generate the unexpected results such as distorted capital. Accordingly, the practitioners have to conduct smoothing process to remove the sample bias from the internal samples and force the realized PDs to be rational as the expected as a monotonic increasing and with floor 0 and cap 1.

Constrained MLE (CMLE) method proposed by Yang [33] leverages the monotonic increasing property of logarithm function. Given K performing risk rating grades, let N_i be the number of the total number in the rating R_i at the beginning of the observation time, and n_i be the number of default for the rating R_i in the observed data set (internal data set of a portfolio). Then the log-likelihood of the estimated PD rates for this portfolio at the end of the year is the same as equation 2.3. However, CMLE runs with the constraint of $p_{i+1} = p_i + e^{\varepsilon_i}$ where $\varepsilon_i \geq 0$. Wherein ε_i is set as the expected increment value of the default rates for the adjacent rating grades of the portfolio.

$$
\begin{aligned}
\max_{p_i} \quad & \sum_{i=1}^{K} \left[\log \binom{N_i}{n_i} + n_i \log p_i + (N_i - n_i) \log(1 - p_i) \right] \\
\text{s.t.} \quad & 0 \leq p_i \leq 1, \\
& \frac{p_{i+1}}{p_i} \geq e^{\varepsilon_i}, \\
& 1 \leq i \leq K, \\
& \varepsilon_i \geq 0.
\end{aligned} \tag{2.4}
$$

With this constrained setup, once we set the appropriate value for ε_i, we can obtain any monotonic increasing pattern with different expected default rate velocity of p_i among rating grades. The typical way to set the ε_i is through the regression of external PD rates and internal PD rates to get the upper or lower bound of ε_i. When $\varepsilon_i = 0$, then we get the exact same constrains of non-decreasing PD rates pattern (equation 2.1),

$$0 \leq p_1 \leq p_2 \leq ... \leq p_K \leq 1.$$

Although the CMLE method can provide a good quality of the monotonic increasing PDs, the estimation does not incorporate any prior knowledge of historical default information about the PD rates we learned from the past experience. Meanwhile, the well-known drawback of MLE method is too sensitive to the training data means the variance of the estimated parameter is high. So, statistician usually adds regularization to MLE method such as reducing variance through introducing bias into the estimate to overcome this drawback. Maximum a posteriori (MAP) is one of the modifications to the MLE method in which the regularization is considered assuming that the estimated parameters themselves are also drawn from a random process in addition to the data. Accordingly, we propose the CMAP method to smooth the empirical default rate in this dissertation.

2.3 Constrained Maximum a Posteriori Smoothing

Maximum a posteriori (MAP) estimator of the parameters is to maximize the entire posteriori distribution which is calculated from the likelihood function and in fact the estimated parameters are the mode of the posteriori distribution if the posteriori distribution is single mode distributed. If the prior distribution is the uniform distribution (equal weights everywhere) or the observed sample data infinitely large, then there is no difference between the MLE and the MAP estimates.

It is well known that if the prior knowledge about the parameters is strong, then the observed data sample will have relatively small impact on the parameter's estimation, (i.e., low variance but high bias), otherwise the estimated parameter will close to MLE's outcomes (i.e., low bias but high variance). We consider p_i to be a random variable from the prior distribution of $\pi(p_i)$, Bayesian approaches incorporate this prior belief about p_i into a posteriori probability $P(p_i|X)$, where X is the observations. Here X is the counts of instruments/borrowers in each risk rating at the beginning of the year and the default counts of each risk rating overserved during the fiscal year. Let i represents the specific risk rating and there is total K risk ratings in the portfolio,

$$P(p_i|X) = \frac{P(X|p_i)\pi(p_i)}{P(X)}.$$

The maximum a posteriori (MAP) estimate is then defined as

$$\hat{p}_i^{MAP} = \arg\max_{p_i} P(p_i|X).$$

Note that because $P(X)$ does not depend on p_i, we have

$$\begin{aligned}
\hat{p}_i^{MAP} &= \arg\max_{p_i} P(p_i|X) \\
&= \arg\max_{p_i} \frac{P(X|p_i)\pi(p_i)}{p(X)} \\
&= \arg\max_{p_i} P(X|p_i)\pi(p_i) \\
&= \arg\max_{p_i} \prod_{X_i \in X} P(X_i|p_i)\pi(p_i) \\
&= \arg\max_{p_i} \sum_{X_i \in X} \log P(X_i|p_i) + \log \pi(p_i). \qquad (2.5)
\end{aligned}$$

Equation 2.5 shows that the MAP estimation comes from injecting our prior beliefs about parameter into the MLE process. However, different from MLE method, the MAP estimation need the assumption of appropriate prior distribution of the parameters. The appropriate choice of the prior distribution can pull the estimation closer to our expectation and thus greatly simplify the MAP estimation process.

2.3.1 Prior Distribution of PD

It is well known that the conjugate distribution of the binomial distribution is the beta distribution with hyperparameters α and β, where α and β are the positive shape parameters respectively of the prior beta distribution, which can be estimated by the distribution of the empirically observed default rates. The posteriori distribution from conjugate beta prior will be another beta distribution with new shape parameters will be $\alpha + C_D$ and $\beta + C_P$, where C_D and C_P are the count of default and count of performing respectively. We applied the conjugate beta prior to our empirical data, however, since the empirical default rate is close to 0 for some rating grades, the posteriori beta distribution cannot provide the estimation for these low default ratings. Another deficiency of posteriori beta distribution is that the results from posteriori beta distribution directly are not monotonic. The estimated PD rates for $S\&P's$ data by posteriori beta distribution is shown in Appendix B. As a result, we need to consider other approaches to fulfill our task.

A practical and rational assumption of the prior distribution of $p_i, i \in [1, K]$ is that they are independent of each other and each p_i is normally distributed along the time frame [35]. In our research, we assume that the PD rates of each rating grade p_i is normally distributed with the mean of long-run average of the probability of default (LRA) and the standard deviation can be calculated by the difference between empirically observed default rates and LRA which can leverage both the observed the default rates and the estimated long run average. The difference between the observed empirical PD and the LRA reflects the economic environment changes and the credit cycle effect on the internal portfolio.

We assume that the real PD rates p_{it} of the default rate for rating grade R_i at fiscal year t is normally distributed with $N(\mu_{p_i}, s_i)$, where both μ_{p_i} and s_i can be derived theoretically from the historical data if the observation history is long enough to cover several credit cycles. If the observed credit history is not long enough, the practitioners usually calibrate the mean value of p_{it} first by combining the short-term internal data with the long-term external data and assuming this combination are distributed as a bivariate normal, then obtain the standard deviation using the empirical observed default rate p_{it} and estimated mean value of default rate μ_{p_i}. Thus, the mean value of μ_{p_i} plays a significant role in calibrating the prior distribution of p_i. We have following common approaches in calibrating μ_{p_i}: the empirical average (mode) or MLE based long-run average (LRA) [35]. The common choices of LRA are as follows,

1. The empirical (weighted) average of observed default rates LRA^{avg};
2. The average default rates attaining the mode that matches the simple average of the observed default rates (LRA^{mod});
3. Long run PDs by MLE approach (LRA^{MLE1}) based on the internal default data only;
4. Long run PDs by MLE approach (LRA^{MLE2}) based on both the internal and external default data together.

All four choices are believed to be consistent estimators [35]. Among these four methods, LRA^{avg} is an unbiased estimator, LRA^{MLE1} is asymptotically efficient, and LRA^{MLE2} is used if the internal data is considered without enough historical information . LRA^{mod} is considered as a conservative estimator of LRA by the Canadian financial institution's regulator. However, it is seldom used in the Advanced Internal Rating-Based (AIRB) framework. Financial institutions using AIRB framework build their own risk rating models and calibrate the PD parameters based on the default information of each risk rating grade from their internal risk rating models.

LRA^{avg} is the most straightforward method and the (weighted) average of the historical default rates (DR) of each rating grade. The weight w can be portfolio count of n_i or portfolio size each year. For an observed M year historic data,

$$L\hat{R}A_i^{avg} = \frac{1}{M}\sum_{j=1}^{M}(w_j)DR_i, \quad i \in (1, K). \tag{2.6}$$

The MLE based LRA is based on Vasicek's one-factor model [12], which is derived from the asset pricing framework of Merton's structure model. Here we give a brief introduction to this approach since it is widely used in the PD parameters estimation.

Under the Merton's framework, a borrower's probability of default at time t p_t is considered as a normalized function of the borrower's asset value, in fact, it is represented by a latent variable of either or not the borrower will default. If p_t is less than a specified (threshold) default point (DP), the borrow becomes the default. At the meanwhile, assume that the firm's default risk is correlated because firm values are correlated via the common dependence on the systematic (market) factor (economy). Then we can assume that the correlation between firms' asset values arises because of correlation between the individual firm's asset value and the common systematic factor.

Now we assume that the correlation structure follows Gaussian Copula, which supposes a correlation between each firm's value and the systematic factor is the same and equal to R. The coefficient R measures the sensitivity of individual borrower's risk to the systematic PD which is uniform across borrowers, $R = Corr(p_{it}, P_t)$. Therefore, the parameter R^2 is the pair-wise correlation of asset values among borrowers, which governs the cross-sectional dependency of credit risk. The PD rate p_t is driven by both the systematic PD rate P_t and the borrower's specific PD risk ε_t,

$$p_t = R \times P_t + \sqrt{1 - R^2} \times \varepsilon_t. \tag{2.7}$$

where systematic factor P_t and individual risk ε_t are assumed to conform the standard normal distribution, with having P_t and ε_t independent. Because p_t is assumed to be a standard normal distribution, given the condition that the borrower will default if the borrower's p_t less than DP, then the unconditional probability of default for a borrower is simply $\Phi(\text{DP})$. Assume there is the threshold DP related to the default, with the conditions of an infinite number of borrowers and the borrower's credit quality are homogeneous in the portfolio. Then the cumulative probability of default of borrower i at time t conditional on the realized systematic PD risk P_t is expressed as:

$$P\big[p_{it} < DP|P_t\big] = \Phi\Big(\frac{1}{\sqrt{1-R^2}}(DP - R \cdot P_t)\Big). \tag{2.8}$$

Since PD for borrower i is assumed independent with borrower j for $i \neq j$, under the condition of homogeneous portfolio, the probability of the realized default number of k out of the portfolio number n at the beginning of the time is

$$P(n_t, k_t) = \binom{n_t}{k_t}\Big(P\big[p_t < DP|P_t\big]\Big)^{k_t}\Big(1 - P\big[p_t < DP|P_t\big]\Big)^{n_t - k_t}.$$

Then, the cumulative probability of the number of defaults can be obtained by integrating out the systematic factor P_t,

$$F(\frac{k_t}{n_t}) = \sum_{i=0}^{k_t}\int_{-\infty}^{\infty} P(n_t, k_t) dP_t.$$

In practice, the financial institutions have a cohort observed default rates for a sequence of risk rating R_i, where $i = 1, \ldots, K$ at each time point t in T years, $t = 1, ..., T$. Let $\theta_t = \frac{k_t}{n_t}$ be the realized default rate in the portfolio at the observation year end, given the infinite number of n within the portfolio, the cumulative distribution of θ_t will asymptotically be

$$\lim_{n\to\infty} F(\theta_t) = \Phi\Bigg[\frac{1}{R}\big(\sqrt{1-R^2}\cdot\Phi^{-1}(\theta_t) - DP\big)\Bigg].$$

Accordingly, $\Phi^{-1}(\theta_t)$ is normally distributed with mean and standard deviation equal to $\frac{DP}{\sqrt{1-R^2}}$ and $\frac{R}{\sqrt{1-R^2}}$ respectively. The realized default rate θ_t is actually a transformation of P_t, which is

$$\Phi^{-1}(\theta_t) = \frac{1}{\sqrt{1-R^2}}(DP - R \cdot P_t). \tag{2.9}$$

Miu and Ozdemir [35] proposed the Generalized Least Squares (GLS) estimator to reach the two types of MLE estimators of LRA: internal observed default data only and combine both short-term internal data and longer-term external observed default data. If there is no time dependence (serial correlation), the internal data only estimator will be

$$\begin{aligned} DP^{MLE} &= \frac{\sqrt{1-R^2}}{T}\sum_{t=1}^{T}\Phi^{-1}(\theta_t), \\ LRA^{MLE} &= \Phi(DP^{MLE}). \end{aligned} \tag{2.10}$$

The second approach of combining the short-term internal data and the longer external default data can be applied if the financial institutions think that their internal observed time is not long enough to obtain the reliable PD estimation. This approach first need map the risk rating grades of the internal and external to make sure the pairwise default data referring to the same credit quality bands.

From equation 2.9, we can derive that $\Phi^{-1}(\theta)$ for both internal and external default rates follows normal distribution. Now, we assume that the external $\Phi^{-1}(\theta_x)$ and internal $\Phi^{-1}(\theta)$ follow a bivariate normal distribution with correlation ρ. Following the MLE approach of the bivariate normal distribution, we reach the estimator of both marginal external LRA LRA_X^{MLE} and the internal LRA of LRA^{MLE} which is conditional on the external LRA,

$$\begin{aligned} LRA_X^{MLE} = \quad & \Phi\Big(\frac{\sqrt{1-R_X^2}}{T_X}\sum_{t=1}^{T_X}\Phi^{-1}(\theta_{t,X})\Big), \\ LRA^{MLE} = \quad & \Phi\Big(\frac{\sqrt{1-R^2}}{T}\sum_{t=1}^{T}\Phi^{-1}(\theta_t) + \frac{R\cdot\rho}{R_x}\Big[T\cdot LRA_X^{MLE} \\ & -\sqrt{1-R^2}\sum_{t=1}^{T}\Phi^{-1}(\theta_{t,X})\Big]\Big). \end{aligned} \tag{2.11}$$

Once we have Long run PD estimation for p_i, the mean value of p_{it}, the standard deviation of p_{it} can be calculated directly by the formula

$$s_i = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(p_i-\mu_i)^2}. \tag{2.12}$$

Therefore, the prior distribution for the PD rate for rating R_i at time t is

$$\pi(p_i) = \frac{1}{\sqrt{2\pi}s_i}e^{-\frac{(p_i-\mu_{p_i})^2}{2s_i^2}}. \tag{2.13}$$

2.4 Redefining the CMAP Method

Recall that equation 2.6 of the MAP method and equation 2.4 show the CMLE for PD rates calibration, with the belief of prior distribution for the PD rate p_i, we reach the CMAP algorithm for PD estimation.

$$
\begin{aligned}
\hat{p_i} = \arg\max_{p_i} \sum_{X_i \in X} \log P(X_i|p_i) + \log \pi(p_i) \\
\text{subject to.} \quad & 0 \leq p_i \leq 1 \\
& \frac{p_i}{p_{i-1}} \geq exp(\varepsilon_i) \qquad (2.14) \\
& 1 \leq i \leq K \\
& \varepsilon_i \geq 0
\end{aligned}
$$

where,

$$
\begin{aligned}
\log P(X_i|p_i) &= \sum_{i=1}^{K} \left[\log \binom{N_i}{n_i} + n_i \log p_i + (N_i - n_i) \log(1 - p_i) \right], \\
\log \pi(p_i) &= \sum_{i=1}^{K} \log \pi_i(p_i), \\
\pi_i(p_i) &= \frac{1}{\sqrt{2\pi s_i}} e^{-\frac{(p_i - \mu_{p_i})^2}{2s_i^2}}.
\end{aligned}
$$

2.5 Empirical Data Analysis

2.5.1 $S\&P's$ Default Data

All three smoothing algorithms of QMM, CMLE and CMAP need the input of total count at the beginning of the observation time and default numbers during the observation period for each rating grade. We conduct the empirical data analysis on the $S\&P's$ rated corporate entities spaning from the year 2009 to the year 2013, from which the required data can be extracted from $S\&P's$ annual report. Out of this time range, $S\&P's$ only publish the default rates but without the number of issuer and default. Data from another rating agency $Moody's$ is less suitable because $Moody's$ did not provide issuer numbers at rating grade level and estimate default rates in a way that makes it impossible to infer exact grade-level numbers of defaults. To get the reliable mean values of each rating grades and have a first impression of the long-term industry-wised default rate, we illustrate the summary statistic of $S\&P's$ observed default data span from the year 1981 to the year 2015 in Table 2.1.

Table 2.1: Empirical default rate distribution from $S\&P$: 1981-2015.

Rating grade	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC-C
Average	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%	0.2%	0.3%	0.5%	0.8%	1.3%	2.2%	6.4%	9.1%	23.7%
Median	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	0.7%	1.7%	5.3%	7.4%	23.1%
Standard	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%	0.2%	0.3%	0.4%	0.4%	0.9%	0.8%	1.7%	2.1%	4.9%	7.6%	11.8%
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Maximum	0.0%	0.0%	0.4%	0.4%	0.6%	0.4%	0.8%	1.1%	1.4%	1.3%	3.7%	3.1%	7.0%	8.7%	17.2%	32.4%	49.5%

Sources: $S\&P$ (2016, tables 9 From year 1981 to year 2015).

We can note from Tables 2.1 that the average default rates were monotone increasing over a long-time period as the associated risk rating's credit quality worsened, which prove that our assumption about the monotonic increasing property of the PD rate is correct. Since the time frame of Table 2.1 is long enough, we assume that the distribution of $S\&P's$ data follow the same mean and standard deviation and will work as the input value of the prior distributions in the proposed CMAP method.

Table 2.2 below presents the observed credit data by $S\&P's$ span from the year 2009 to the year 2013. Nevertheless, Table 2.2 reveals that the empirical default rates for each fiscal year are volatile and usually violate the monotone assumption. As well as the high rating grades (above AA) usually has no default observed, and the very low default rate is realized in a good economic environment. Thus, the empirical default rates must be smoothed before further use.

The real data analysis will base on the data in Table 2.2.

Table 2.2: Default data from $S\&P$: 2009-2013.

Rating grade	2009			2010			2011			2012			2013		
	Rated	Defaults	DR	Rated	Defaults	DR	Rated	Defaults	DR	Rated	Defaults	DR	Rated	Defaults	DR
AAA	81	0	0.00%	72	0	0.00%	51	0	0.00%	24	0	0.00%	21	0	0.00%
AA+	37	0	0.00%	25	0	0.00%	36	0	0.00%	51	0	0.00%	48	0	0.00%
AA	188	0	0.00%	143	0	0.00%	120	0	0.00%	61	0	0.00%	63	0	0.00%
AA-	245	0	0.00%	209	0	0.00%	207	0	0.00%	238	0	0.00%	210	0	0.00%
A+	340	1	0.29%	353	0	0.00%	357	0	0.00%	337	0	0.00%	325	0	0.00%
A	510	2	0.39%	474	0	0.00%	470	0	0.00%	445	0	0.00%	419	0	0.00%
A-	546	0	0.00%	528	0	0.00%	560	0	0.00%	548	0	0.00%	542	0	0.00%
BBB+	498	2	0.40%	457	0	0.00%	473	0	0.00%	523	0	0.00%	515	0	0.00%
BBB	541	1	0.18%	583	0	0.00%	549	0	0.00%	589	0	0.00%	641	0	0.00%
BBB-	459	5	1.09%	430	0	0.00%	508	1	0.20%	525	0	0.00%	533	0	0.00%
BB+	266	0	0.00%	254	2	0.79%	260	0	0.00%	311	0	0.00%	324	0	0.00%
BB	295	3	1.02%	276	1	0.36%	319	0	0.00%	333	0	0.00%	359	0	0.00%
BB-	441	4	0.91%	379	2	0.53%	403	0	0.00%	403	3	0.74%	395	0	0.00%
B+	438	24	5.48%	393	0	0.00%	509	2	0.39%	526	3	0.57%	538	0	0.00%
B	482	48	9.96%	436	3	0.69%	586	7	1.19%	646	9	1.39%	739	0	0.00%
B-	303	52	17.16%	290	6	2.07%	301	12	3.99%	299	10	3.34%	352	3	0.85%
CCC-C	190	92	48.42%	220	49	22.27%	138	22	15.94%	154	41	26.62%	158	2	1.27%
All	5860	234	3.99%	5522	63	1.14%	5847	44	0.75%	6013	66	1.10%	6182	5	0.08%

Note: $S\&P's$ corporate ratings, defaults and default rates from 2009 to 2013. Sources: $S\&P$ (2010, Tables 51 to 53), $S\&P$ (2011, Tables 50 to 52), $S\&P$ (2012, Tables 50 to 52), $S\&P$ (2013, Tables 50 to 52), $S\&P$ (2014, tables 50 to 52).

Tables 2.3 below illustrates the smoothed PD rate by three methodologies: QMM, CMLE and CMAP. All three methods provide the monotonic increasing PD rates as the rating grades' creditworthiness worsen, however, QMM cannot estimate smoothed PD parameters in the situation of very low default (good economic) situation such as the year of 2013. This is a critical drawback of QMM.

Table 2.3: Comparison of the smoothed PD rates: QMM, CMLE and CMAP.

Rating grade	QMM					CMLE					CMAP				
	2009	2010	2011	2012	2013	2009	2010	2011	2012	2013	2009	2010	2011	2012	2013
AAA	0.0078	0.0009	0.0001	0.0013	-	0.0005	0.0001	0.0000	0.0001	0.0000	0.0003	0.0001	0.0000	0.0001	0.0000
AA+	0.0083	0.0016	0.0002	0.0021	-	0.0007	0.0001	0.0000	0.0001	0.0000	0.0004	0.0001	0.0000	0.0001	0.0000
AA	0.0087	0.0023	0.0004	0.0031	-	0.0009	0.0001	0.0000	0.0001	0.0000	0.0005	0.0002	0.0001	0.0001	0.0000
AA-	0.0091	0.0030	0.0008	0.0042	-	0.0012	0.0002	0.0001	0.0001	0.0000	0.0007	0.0002	0.0001	0.0002	0.0000
A+	0.0093	0.0037	0.0012	0.0051	-	0.0016	0.0003	0.0001	0.0002	0.0000	0.0010	0.0003	0.0001	0.0002	0.0000
A	0.0095	0.0044	0.0016	0.0058	-	0.0022	0.0004	0.0001	0.0002	0.0000	0.0013	0.0004	0.0001	0.0003	0.0000
A-	0.0096	0.0050	0.0020	0.0064	-	0.0023	0.0005	0.0002	0.0003	0.0000	0.0018	0.0006	0.0002	0.0004	0.0000
BBB+	0.0098	0.0058	0.0026	0.0074	-	0.0031	0.0007	0.0002	0.0004	0.0000	0.0025	0.0008	0.0003	0.0005	0.0000
BBB	0.0100	0.0072	0.0038	0.0088	-	0.0043	0.0009	0.0003	0.0006	0.0000	0.0033	0.0011	0.0004	0.0007	0.0000
BBB-	0.0103	0.0090	0.0055	0.0106	-	0.0057	0.0012	0.0004	0.0008	0.0000	0.0045	0.0015	0.0005	0.0010	0.0000
BB+	0.0106	0.0114	0.0084	0.0130	-	0.0060	0.0017	0.0004	0.0011	0.0000	0.0049	0.0020	0.0005	0.0013	0.0000
BB	0.0109	0.0151	0.0139	0.0164	-	0.0081	0.0022	0.0004	0.0015	0.0000	0.0067	0.0027	0.0005	0.0018	0.0000
BB-	0.0113	0.0205	0.0237	0.0210	-	0.0091	0.0025	0.0005	0.0020	0.0000	0.0090	0.0028	0.0006	0.0024	0.0000
B+	0.0117	0.0278	0.0412	0.0281	-	0.0548	0.0027	0.0038	0.0057	0.0001	0.0485	0.0030	0.0043	0.0061	0.0001
B	0.0121	0.0381	0.0706	0.0375	-	0.0996	0.0069	0.0119	0.0139	0.0008	0.0970	0.0073	0.0124	0.0144	0.0008
B-	0.0126	0.0496	0.1123	0.0466	-	0.1716	0.0260	0.0399	0.0334	0.0067	0.1658	0.0263	0.0410	0.0345	0.0070
CCC-C	0.0131	0.0662	0.1872	0.0662	1.0000	0.4842	0.2171	0.1594	0.2662	0.0127	0.4631	0.2189	0.1646	0.2638	0.0141

Table 2.4 lists the comparison of the statistical performance of these three methodologies. We calculate the -2 log likelihood statistic for the estimated PD rates. As the PD rate of each risk rating is a binomial distribution, the formula for log-likelihood calculation is as

$$l(p) = \sum_{i=1}^{n} \left(\log \binom{N_i}{k_i} + k_i \log p + (N_i - k_i) \log(1-p) \right).$$

Table 2.4: Comparison of the performance of three methodologies.

Method	-2 Log likelihood				
	2009	2010	2011	2012	2013
CMLE	51.33	29.89	20.87	27.38	7.07
CMAP	53.71	30.03	21.03	27.55	7.10
QMM	958.34	167.31	159.83	179.18	N/A

Table 2.4 shows that the CMLE method has the highest likelihood score, the QMM method has the lowest likelihood score. The likelihood score of the CMAP method is higher QMM method but a little bit lower log-likelihood than CMLE.

MSE statistic is illustrated in Tables 2.5 below calculated by the estimated PD rates against the realized PD rates for each observation year. QMM perform well in the prevailing economic situation (the year 2010-2012) but works worse in the credit downturn and credit peak stages (the year 2009 and year 2013), which means that the QMM method is not sensitive to economic environment change. However, both CMLE and CMAP work well during the whole credit cycle.

Table 2.5: Comparison of MSE of the estimation.

	Y2009	Y2010	Y2011	Y2012	Y2013
CMLE	0.01%	0.01%	0.00%	0.00%	0.00%
CMAP	0.06%	0.01%	0.00%	0.00%	0.00%
QMM	25.74%	2.77%	1.17%	4.22%	100.02%

Considering the QMM method cannot smooth the default rates in the very low default (credit peak, year 2013) environment and the estimated PD rates are not sensitive to the severe credit environment (credit peak and downturn), next, we will conduct a comparison of the CMLE and CMAP only through a toy default data.

2.5.2 Real Default Data

The $S\&P$'s default data is a typical representative for the industry-wide portfolio which has longer observed default history and large numbers of borrowers but not a good representative of the default data from a specific financial institution which has short observed default history and small numbers of borrowers. To show the performance of the proposed CMAP algorithm on the portfolio from a financial institution, we further compare the proposed CMAP and CMLE method on a real default data which sources from an anonymous financial institution. We kept the company name secret for the confidential reason and modified the raw data a little bit (around 5% of the real data) without any impact on the methodologies comparison. We continually compare log-likelihood score and MSE statistics as the quantitative measurement. In addition, we compare the estimated PD rates at each rating grade level along the time to validate the consistency and stability of the estimated PD rates, which have the significant business meaning since both regulatory capital and economic capital calculation are based the PDs (with LGD and EAD). The volatile PDs will cause the volatile capital calculation and generates other issues for the bank's risk management and operations.

The real default data spans from the year 2003 to the year 2014 which includes the financial crisis period (the year 2007 to the year 2009). Its rating grade follows

Moody's standard which has 20 risk rating levels from Aaa to Ca. There are around 5000 firms within the portfolio and this number is minor changed every year caused by the new borrowers added in, some borrowers dropped out, and default may happen within the portfolio. Total 97 defaults concentrate in the speculative grade (Ba1 and below). The default rates for the year 2009 are the highest because of the severe financial crisis. There is no information for the risk rating Aa2 since it is prevalent that there is no default information at all for some investment grade of an internal portfolio. Overall, the real data is a typical representative of the financial institution's internal default data which does not have monotonic increasing PD rates, the default rate is sparse since the number of default is small, and some ratings have no default-information during the observation period. Consequently, the real data is a typical credit data for us to test the proposed smoothing methodology.

The input values of the total counts and default numbers for each rating and fiscal year for the smoothing algorithms are presented in Appendix C. To make the dissertation easily to be read, Table 2.6 below shows the empirical default rates from the real default data. Figure 2.1 is the empirical PD curve of the real data which illustrates all the properties of the data we mentioned above.

Figure 2.1 shows that the realized default rates of the toy data in each fiscal year are not monotonic increasing as the risk ratings creditworthiness worsens, most

Table 2.6: Realized default rate from empirical default data.

Rating	Empirical Default Rate by Years											
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Aa1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Aa2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Aa3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	N/A	N/A	N/A
A3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Baa1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0061	0.0000
Baa2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Baa3	0.0000	0.0000	0.0000	0.0013	0.0014	0.0016	0.0019	0.0022	0.0026	0.0029	0.0032	0.0036
Ba1	0.0011	0.0012	0.0027	0.0015	0.0016	0.0017	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Ba2	0.0011	0.0012	0.0014	0.0016	0.0036	0.0042	0.0025	0.0028	0.0000	0.0000	0.0000	0.0000
Ba3	0.0019	0.0020	0.0000	0.0047	0.0050	0.0054	0.0144	0.0095	0.0071	0.0119	0.0044	0.0049
B1	0.0065	0.0047	0.0050	0.0055	0.0030	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B2	0.0079	0.0085	0.0095	0.0071	0.0120	0.0174	0.0145	0.0106	0.0116	0.0064	0.0070	0.0000
B3	0.0000	0.0000	0.0000	0.0000	0.0263	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Caa1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Caa2	0.0000	0.0000	0.0000	0.0000	0.0625	0.1111	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Caa3	0.0000	0.0000	1.0000	0.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ca	0.0000	0.0000	0.0000	0.0000	0.0952	0.1250	0.0833	0.1000	0.0000	0.0000	0.0000	N/A

Note: 1. The data is generated by randomly change the 5% of the real default data; 2. Some ratings have no information in particular fiscal years; then it will generate the empirical PD of *N/A*.

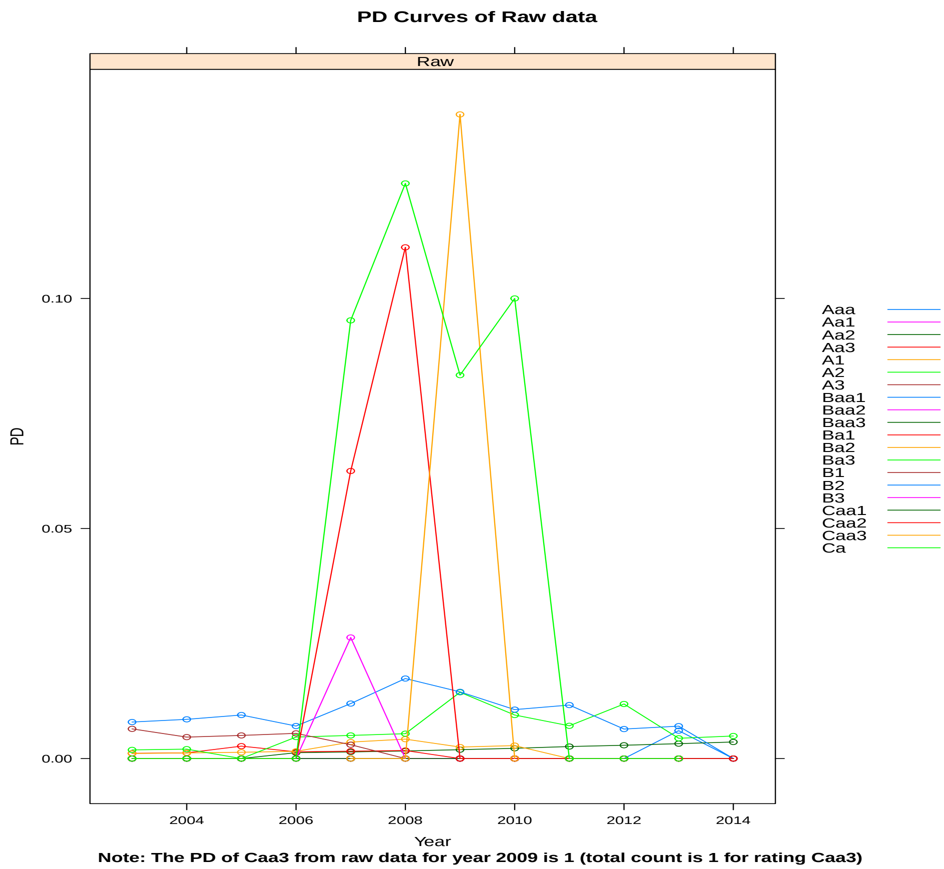


Figure 2.1: The time series plot of the realized default rates.

realized default rates are 0 for the investment grades and rate AA2 has no any information at all , both situations need to be assigned with a reasonable PD rates.

Smoothing Results of Real Default Data

The input parameters of the prior normal distribution in the CMAP algorithm is tuned by the empirical mean and standard deviation of the real data for each rating grade to make the methodology applied in the dissertation consistent. Table 2.7 and 2.8 illustrate the smoothed PD rates by the CMLE and CMAP method respectively. Figure 2.2 and 2.3 are the PD curves for the associated estimated PD rates by each method accordingly. We note that both work well in default rate's monotonic smoothing perspective. The smoothed PD rates are granularly increasing as the credit quality becomes worse at each observation time. Meanwhile, both can calibrate the PD rates estimations for the risk ratings with non-default information such as no information at all or 0 default rates. Accordingly, both methods can be used for PD rates smoothing.

Table 2.7: Smoothed PD rates by CMLE method.

Rating	Smoothed PD Rates by MLE Method											
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001
Aa1	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0003	0.0001
Aa2	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0002	0.0002	0.0002	0.0004	0.0001
Aa3	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0002	0.0003	0.0005	0.0002
A1	0.0002	0.0002	0.0002	0.0003	0.0004	0.0004	0.0005	0.0004	0.0003	0.0004	0.0006	0.0002
A2	0.0003	0.0003	0.0003	0.0004	0.0005	0.0005	0.0006	0.0005	0.0004	0.0005	0.0009	0.0003
A3	0.0004	0.0004	0.0004	0.0005	0.0007	0.0007	0.0008	0.0007	0.0006	0.0007	0.0012	0.0004
Baa1	0.0005	0.0005	0.0006	0.0007	0.0009	0.0010	0.0011	0.0009	0.0008	0.0009	0.0016	0.0006
Baa2	0.0007	0.0007	0.0007	0.0009	0.0012	0.0013	0.0015	0.0012	0.0011	0.0012	0.0017	0.0008
Baa3	0.0009	0.0009	0.0010	0.0013	0.0017	0.0018	0.0021	0.0016	0.0014	0.0017	0.0017	0.0011
Ba1	0.0013	0.0012	0.0014	0.0017	0.0022	0.0024	0.0028	0.0022	0.0015	0.0019	0.0018	0.0011
Ba2	0.0017	0.0017	0.0018	0.0023	0.0030	0.0032	0.0038	0.0029	0.0020	0.0025	0.0019	0.0012
Ba3	0.0023	0.0022	0.0025	0.0031	0.0041	0.0043	0.0051	0.0040	0.0027	0.0034	0.0026	0.0016
B1	0.0031	0.0030	0.0034	0.0042	0.0055	0.0058	0.0053	0.0042	0.0037	0.0036	0.0027	0.0017
B2	0.0042	0.0041	0.0045	0.0053	0.0075	0.0079	0.0072	0.0056	0.0050	0.0048	0.0037	0.0017
B3	0.0044	0.0043	0.0047	0.0055	0.0101	0.0083	0.0076	0.0059	0.0053	0.0051	0.0038	0.0018
Caa1	0.0047	0.0045	0.0050	0.0058	0.0106	0.0089	0.0079	0.0062	0.0055	0.0053	0.0040	0.0019
Caa2	0.0049	0.0047	0.0052	0.0061	0.0586	0.0739	0.0169	0.0065	0.0058	0.0056	0.0042	0.0020
Caa3	0.0051	0.0050	0.0055	0.0064	0.0616	0.0776	0.1411	0.0119	0.0061	0.0058	0.0044	0.0022
Ca	0.0054	0.0052	0.0058	0.0067	0.0952	0.1250	0.1482	0.0989	0.0064	0.0061	0.0047	0.0147

The time series plot shown in Figure 2.2 illustrates the monotonic increasing property of the estimated PD rates by CMLE method which also assigns the PDs to all rating grades (even the rating grades without default realized or without iussers information during the observation time).

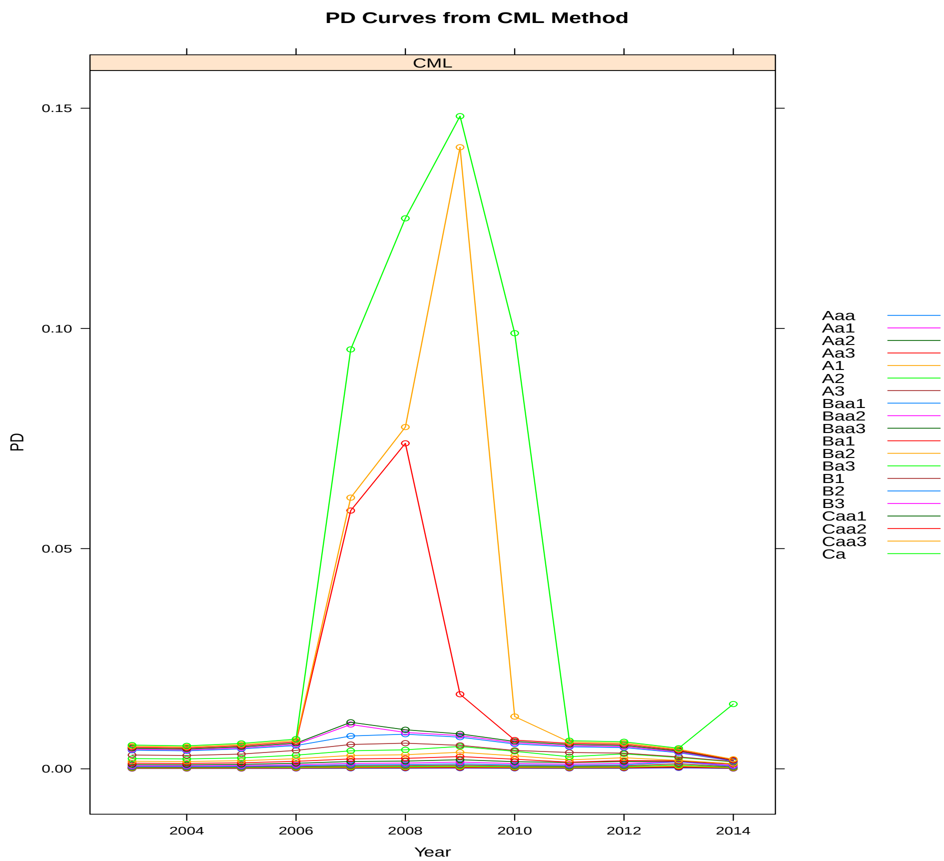


Figure 2.2: The time series plot of the estimated PDs by CMLE method.

Table 2.8: Smoothed PD rates by CMAP method.

Rating	Smoothed PD Rates by MAP Method											
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Aa1	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001
Aa2	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0003	0.0001
Aa3	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0004	0.0002
A1	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0005	0.0003
A2	0.0003	0.0003	0.0003	0.0003	0.0004	0.0004	0.0004	0.0005	0.0004	0.0004	0.0007	0.0004
A3	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	0.0006	0.0006	0.0005	0.0006	0.0009	0.0005
Baa1	0.0005	0.0005	0.0006	0.0006	0.0007	0.0007	0.0008	0.0008	0.0007	0.0008	0.0012	0.0007
Baa2	0.0007	0.0007	0.0008	0.0009	0.0010	0.0010	0.0010	0.0011	0.0010	0.0011	0.0013	0.0009
Baa3	0.0010	0.0010	0.0010	0.0012	0.0013	0.0013	0.0014	0.0015	0.0013	0.0015	0.0013	0.0012
Ba1	0.0013	0.0013	0.0014	0.0015	0.0017	0.0017	0.0019	0.0016	0.0014	0.0016	0.0014	0.0013
Ba2	0.0018	0.0018	0.0019	0.0020	0.0023	0.0023	0.0025	0.0022	0.0019	0.0021	0.0019	0.0017
Ba3	0.0024	0.0024	0.0025	0.0028	0.0032	0.0031	0.0034	0.0029	0.0026	0.0028	0.0025	0.0023
B1	0.0033	0.0032	0.0034	0.0037	0.0043	0.0042	0.0040	0.0036	0.0035	0.0031	0.0032	0.0025
B2	0.0044	0.0043	0.0046	0.0050	0.0058	0.0057	0.0054	0.0049	0.0047	0.0042	0.0043	0.0033
B3	0.0046	0.0046	0.0048	0.0053	0.0066	0.0060	0.0057	0.0051	0.0049	0.0044	0.0045	0.0035
Caa1	0.0049	0.0048	0.0050	0.0056	0.0069	0.0063	0.0060	0.0054	0.0052	0.0046	0.0048	0.0037
Caa2	0.0132	0.0132	0.0132	0.0132	0.0329	0.0373	0.0087	0.0106	0.0132	0.0132	0.0132	0.0132
Caa3	0.0164	0.0189	0.0215	0.0240	0.0345	0.0392	0.0729	0.0111	0.0266	0.0292	0.0318	0.0370
Ca	0.0172	0.0199	0.0226	0.0252	0.0649	0.0714	0.0765	0.0576	0.0279	0.0306	0.0334	0.0389

The proposed CMAP method has the same smoothing properties as the CMLE method of the monotonic increasing estimated PDs for each observation time. The time series plot of the estimated PD rates by CMAP method is shown in Figure 2.3.

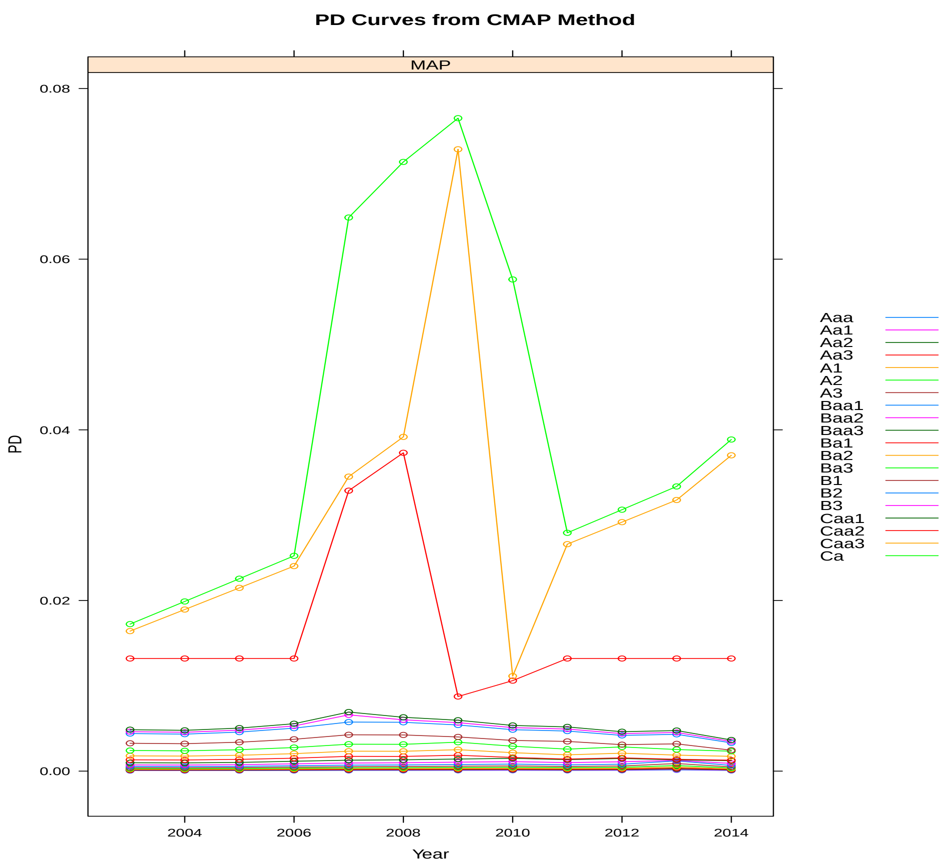


Figure 2.3: The time series plot of the estimated PDs by CMAP method.

The log-likelihood score is shown in Table 2.9 and MSE statistics is presented in Table 2.10 to compare the statistical performance between two methodologies. Still, the CMLE has a little higher log likelihood statistic than CMAP. The total -2 log-likelihood statistics is 1235 for CMAP method and 1286 for CMAP method respectively.

Table 2.9: Comparison of Log Likelihood of CMLE and CMAP.

-2 Log Likelihood													
Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
CMLE	123.6	110.6	110.2	122.2	161.9	147.5	137.4	98.9	67.9	67.6	56.8	30.6	1235.2
CMAP	127.1	114.1	113.6	125.8	167.6	153.9	143.9	103.2	71.1	71.1	60.4	34.5	1286.3

Table 2.10: Comparison of MSE of the estimation.

	Y2003	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009	Y2010	Y2011	Y2012	Y2013	Y2014
CMLE	0.01%	0.01%	0.02%	0.02%	0.42%	0.77%	74.24%	0.03%	0.03%	0.03%	0.01%	0.03%
CMAP	0.08%	0.10%	0.12%	0.15%	0.35%	1.01%	86.00%	0.22%	0.18%	0.21%	0.24%	0.31%

Table 2.11 and Table 2.12 is the comparison of the weighted MSE for two methodologies. The weight is defined as the risk rating grades frequency in the portfolio. CMLE method has a little lower weighted MSE in general. However, the weighted MSE of the prediction power shows that the CMAP method has lower weighted than

CMLE method at some observation time,

$$MSE_{weighted} = \sum_{t=1}^{T} \sum_{i=1}^{K} w_i * MSE_{it}.$$

Table 2.11: Comparison of weighted MSE of the estimation.

	Y2003	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009	Y2010	Y2011	Y2012	Y2013	Y2014
CMLE	0.14%	0.13%	0.16%	0.10%	0.22%	0.33%	1.31%	0.23%	0.22%	0.25%	0.17%	0.13%
CMAP	0.15%	0.14%	0.17%	0.13%	0.34%	0.52%	1.37%	0.31%	0.24%	0.27%	0.19%	0.14%

Table 2.12: Prediction power: weighted MSE of the estimation.

	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009	Y2010	Y2011	Y2012	Y2013	Y2014
CMLE	0.153%	0.202%	0.145%	0.557%	0.427%	1.086%	0.408%	0.569%	0.359%	0.253%	0.220%
CMAP	0.176%	0.225%	0.185%	0.487%	0.538%	1.094%	0.333%	0.409%	0.388%	0.291%	0.291%

It is as we expected that a little higher log-likelihood score and a little lower MSE from CMLE method than the proposed CMAP since the nature of CMLE method is to find the optimized value at the time point (rating year) and does not consider the time (credit cycle) effect. However, the CMAP method is leveraging the prior knowledge (prior distribution of the PD rates for each risk rating) of the PD rates. To show the advantages of the proposed CMAP method, we next compare the estimated PD rates at the risk rating level as shown in Figure 2.4. A heat map of difference between two methodologies presented in Table 2.13. From Table 2.13, we can get the conclusion that the CMLE method is more sensitive to the economic cycle than CMAP. The estimated PD rates for the upper medium grade above (A3 above) from the two methods are very close. For the rating grades highly speculative above (Baa1 B3), the CMLE estimates are higher in the regular economic environment (year 2000 to year 2006) but are lower in downturn credit environment (year 2008 to year 2009), while the rating grades of substantial risks below (Caa1 below) has the reverse behavior which is higher in the peak credit environment and lower in the downturn period.

Figure 2.4: Comparison of the estimations from CMAP and CMLE.

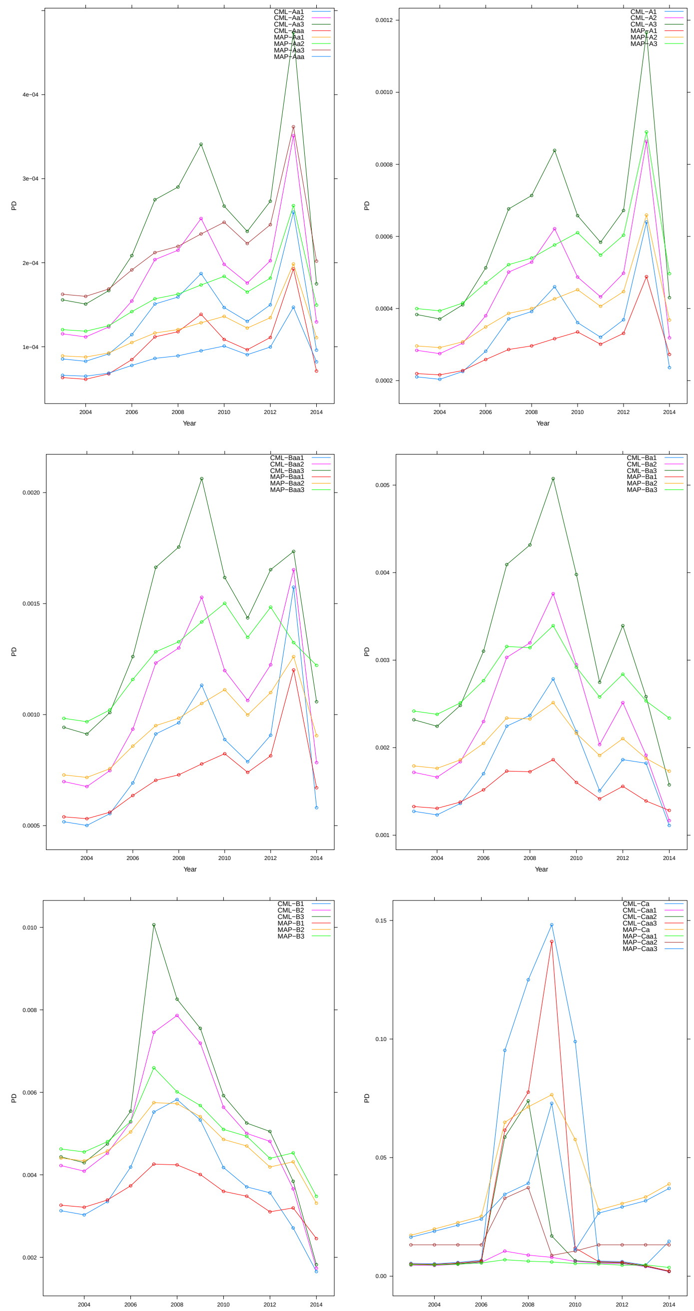


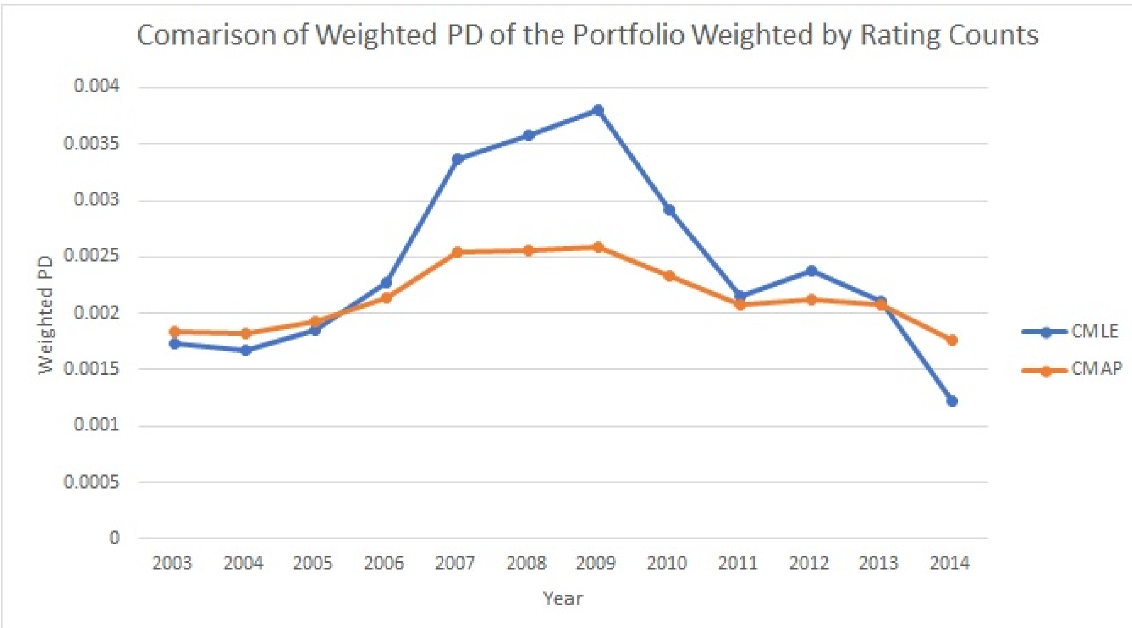
Table 2.13: Heat map of the difference of the estimated PD rates.

Date	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	0.0003%	0.0004%	0.0001%	-0.0007%	-0.0026%	-0.0029%	-0.0043%	-0.0008%	-0.0006%	-0.0011%	-0.0046%	0.0011%
Aa1	0.0004%	0.0005%	0.0001%	-0.0009%	-0.0035%	-0.0039%	-0.0059%	-0.0011%	-0.0008%	-0.0015%	-0.0062%	0.0015%
Aa2	0.0005%	0.0007%	0.0001%	-0.0013%	-0.0047%	-0.0052%	-0.0079%	-0.0014%	-0.0011%	-0.0021%	-0.0083%	0.0020%
Aa3	0.0007%	0.0009%	0.0002%	-0.0017%	-0.0063%	-0.0071%	-0.0107%	-0.0019%	-0.0014%	-0.0028%	-0.0112%	0.0027%
A1	0.0009%	0.0012%	0.0003%	-0.0023%	-0.0085%	-0.0095%	-0.0144%	-0.0026%	-0.0020%	-0.0038%	-0.0152%	0.0037%
A2	0.0012%	0.0017%	0.0004%	-0.0031%	-0.0115%	-0.0129%	-0.0195%	-0.0035%	-0.0026%	-0.0051%	-0.0205%	0.0049%
A3	0.0016%	0.0023%	0.0005%	-0.0042%	-0.0155%	-0.0174%	-0.0263%	-0.0047%	-0.0036%	-0.0069%	-0.0276%	0.0067%
Baa1	0.0022%	0.0030%	0.0006%	-0.0057%	-0.0209%	-0.0234%	-0.0354%	-0.0064%	-0.0048%	-0.0093%	-0.0373%	0.0090%
Baa2	0.0030%	0.0041%	0.0009%	-0.0076%	-0.0282%	-0.0316%	-0.0478%	-0.0086%	-0.0065%	-0.0125%	-0.0391%	0.0121%
Baa3	0.0040%	0.0056%	0.0012%	-0.0103%	-0.0380%	-0.0427%	-0.0646%	-0.0116%	-0.0088%	-0.0169%	-0.0411%	0.0164%
Ba1	0.0055%	0.0075%	0.0016%	-0.0185%	-0.0514%	-0.0644%	-0.0922%	-0.0581%	-0.0092%	-0.0305%	-0.0431%	0.0172%
Ba2	0.0074%	0.0101%	0.0021%	-0.0249%	-0.0693%	-0.0870%	-0.1245%	-0.0784%	-0.0124%	-0.0411%	-0.0036%	0.0565%
Ba3	0.0099%	0.0137%	0.0029%	-0.0337%	-0.0936%	-0.1174%	-0.1680%	-0.1059%	-0.0168%	-0.0555%	-0.0049%	0.0763%
B1	0.0134%	0.0184%	0.0039%	-0.0454%	-0.1263%	-0.1585%	-0.1320%	-0.0578%	-0.0226%	-0.0460%	0.0486%	0.0801%
B2	0.0181%	0.0249%	0.0053%	-0.0240%	-0.1705%	-0.2139%	-0.1782%	-0.0780%	-0.0305%	-0.0621%	0.0656%	0.1577%
B3	0.0190%	0.0261%	0.0055%	-0.0252%	-0.3468%	-0.2246%	-0.1871%	-0.0819%	-0.0321%	-0.0652%	0.0689%	0.1656%
Caa1	0.0200%	0.0274%	0.0058%	-0.0265%	-0.3641%	-0.2557%	-0.1964%	-0.0860%	-0.0337%	-0.0685%	0.0723%	0.1739%
Caa2	0.8311%	0.8468%	0.7966%	0.7089%	-2.5771%	-3.6601%	-0.8192%	0.4080%	0.7408%	0.7633%	0.8965%	1.1193%
Caa3	1.1277%	1.3969%	1.5981%	1.7615%	-2.7059%	-3.8431%	-6.8264%	-0.0733%	2.0517%	2.3334%	2.7329%	3.4852%
Ca	1.1841%	1.4667%	1.6780%	1.8495%	-3.0356%	-5.3602%	-7.1677%	-4.1291%	2.1543%	2.4501%	2.8696%	2.4161%

Comment: The yellow color means the difference within ±1 base point (0.01%), the green color means the estimated PD rate from CMLE is less than CMAP by 1 more bps, and the red color means the estimated PD rate from CMLE is higher than CMAP by 1 more bps.

Next, we compare the impact of the CMAP and CMLE methods to expected loss calculation, which is the most critical concern of the financial institutions. Assume that the portfolio is same and independent on LGD and EAD, we calculate the portfolio level weighted PD, weighting schema is calculated by the frequency weight of each risk rating grade in the portfolio over the time. As shown in Figure 2.5, the weighted PD from CMAP method is more consistent and stable, which is expected by the financial institution and makes them manage the risk more operability and rational.

Figure 2.5: Comparison of the weighted PD of the portfolio.



2.6 Conclusion and Future Work

In this chapter, we propose a smoothing algorithm to smooth the empirical PD rate. We investigate and compare the performance proposed CMAP method against QMM and CMLE methods on the historical $S\&P$'s data and a real data. The results shows that all three methods of QMM, CMLE and CMAP can work in empirical PD rates monotonic smoothing. However, QMM does not work well as the other two methods in the peak credit period (very low default situation) and the results from QMM has the lowest likelihood score and MSE statistics in the performance measurement. Compared to CMLE method, the proposed CMAP has little worse statistical performance regarding to the likelihood score and MSE statistics, and very close regarding the weighted MSE statistics. Nevertheless, the CMLE method is susceptible to the credit environment change, which may come from it's drawback of very sensitive to the sample data. CMAP considers the prior knowledge of the historical PD rate which assumes the PD rates are normally distributed with a mean of the Long Run Default Rates (LRA) and can leverage the external PD rate data which is close to the systematic credit data. Thus CMAP can provide the less sensitive estimators than CMLE. Because the PD rates change as the economic cycle change, the sensitive PD estimation will cause the dramatic change in the financial institution's capital calculation year over year, which can cause the financial institutions' operating difficulty. Accordingly, from the perspective of stability and consistency of the capital calculation, CMAP method is preferable than CMLE since it provides the less sensitive estimated PD rates with very close statistical performance.

Future work with this approach would include applying this method to the actuarial field to smooth the mortality rate curve which has the hump shape or bathtub shape and is not monotonic increasing or decreasing. Also, this approach can be applied to the smooth the yield rates in the fixed income pricing and market risk calculation.

3 PD Transition Probabilities Matrices Smoothing through Constrained Maximum a Posteriori

3.1 PD Transition Probabilities Matrices

In Chapter 2, we propose a smoothing algorithm for empirical PD rates by using constrained a maximum posterior method, which can be used for calculating the consistent and reliable regulatory and economic capital by the financial institutions. When estimating the expected credit loss (ECL) for a portfolio during its lifetime, we need the term structure of default probabilities, which is defined as the instantaneous or the cumulative probability of default at each time point t in the future for an instrument of a borrower. PD transition probabilities show the probability of a company migrating from one rating category to another during a certain period of time. The empirical transition matrices are based on internal historical data. Let p_t

denote the probability of default during period $t : PD(t - 1, t) = p_t$. This is called the conditional or marginal default probability since it is the probability that the firm defaults at time t given that it has survived until time $t - 1$. By it's nature, PD transition matrices for n periods of time can be calculated by as the n^{th} power of the one year matrix.

PD rating transition probability matrices are a crucial parameter in the applications of credit risk area for credit portfolio loss estimation in diversified portfolios pooled by rating, risk management of revolver loan commitments, and counterparty credit risk management of the smaller trading counterparties do not have any outstanding public debt or CDS. It also plays an important role in pricing and investment decisions in the rating (pool level) portfolios. Without a doubt, owing to scarce historical data, temporal observations and censored observation data, the bank's internal empirical transition probability matrices do not show behaviors as expected from the industry's viewpoint of the term structure of the transition probabilities. In this chapter, we present a constrained maximum a posteriori estimation methodology for smoothing the empirical PD transition probabilities matrices using an optimization algorithm that leads to consistency with empirical observations and desired theoretical properties. It is sensitive to the risk measurements and it serves as a reliable and consistent capital management.

The empirical estimator of the transition probabilities p_{ij}, the probability of a firm upgrade or downgrade from rating i to rating j, can be simply obtained by dividing the transition numbers by the total number of rating i at the beginning of the observation time. Let N_i be the count of firms with rating i at the beginning of observation time and N_{ij} be the count of the firms migrating from rating i to rating j during the observation period, then PD transition probability $P_{ij} = n_{ij}/N_{ij}$. The major credit rating agencies such as Moody's and $S\&P'$s publish the transition matrices every year for different sectors. Assume that there is $K+1$ risk ratings in a portfolio, where the $(K+1)^{th}$ rating represent the default rating, then the PD rating transition probability matrix is given by using P_{ij} as the i^{th} row and j^{th} column element, that is $P_{ij} = Pr(j|i) = n_{ij}/N_{ij}$, e.g.,

$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} & \dots & P_{1,K+1} \\ P_{2,1} & P_{2,2} & \dots & P_{2,j} & \dots & P_{2,K+1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots & P_{i,K+1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}.$$

Since the total of transition probability from a rating i to all other ratings must be 1 in a theoretical situation, so that

$$\sum_{j=1}^{K+1} P_{ij} = 1 \quad where\, i = 1, \dots K + 1.$$

As one of widely used credit loss modeling approaches in the measure of expected credit losses (ECL), the PD transition matrix of the risk rating migration which reflects the credit quality change of the instrument or an obligor for the desired transition horizon. The internal realized transition matrix is sometimes not reliable to be directly used because the transition process suffers from the rarity of the observed events and the presence of censored observations during the observation time. A rational and desired PD transition matrix should hold the following properties[61]:

- The default probability is monotonically increasing as the creditworthiness worsens, i.e., the PD rates of the worse creditworthiness is higher than PD rates of a better creditworthiness rating.

- The sum of transition probabilities of each risk rating is 1.

- The transition probabilities must be monotonic, i.e., a transition probability to a farther state must be less than a transition probability to a nearer state, and the highest transition probabilities is at the diagonal (the most borrowers will stay at the same rating grade in the next term).

- The monotonically decreasing transition probabilities is away from the diagonal but not need to be symmetric, i.e., the transition probability of a one-notch upgrade need not have the same probability as a one-notch downgrade.

- The estimated transition matrix need consider the issue of censored data, i.e., the information about the survival time of an issuer is missing due to leave the company or in a given rating state both prior to the estimation window and after the end of the window is discarded.

3.2 Proposed Smoothing Algorithm

We propose a new PD transition matrix smoothing methodology which generates the smoothed estimates corresponding with a well-defined transition matrix and with the maximum likelihood or small discrepancies. The process includes two steps while the first step is smoothing the empirical default rates for each rating grade and the second step is smoothing the empirical transition probabilities with constraints which force the estimated transition probabilities fit our expectation listed above. Let's say there are K performing risk ratings and $K+1$ transition states where the $(K+1)$th state represents the default status. Let p_{ij} be the transition probability from rating i to rating j where i and j is from 1 to K. Let N_i be the count of the issuers with rating i initially at the beginning of the observation period, and n_{ij} be the migrated count of issuers from rating i to rating j at the end of the observation period. As the transition probabilities for each risk rating i is a categorical distribution, the conjugate prior distribution for it is a Dirichlet distribution with parameters s_i and α_{ij} which can be tuned by the historical data. Let X be the transition probabilities and X_i be the transition probabilities of ith risk rating. The estimates of the transition probabilities are the optimal solutions for the following likelihood function with constraints of the transition probabilities,

$$
\begin{aligned}
\hat{p_{ij}} = &\arg\max_{p_{ij}} \sum_{X_i \in X} \log P(X_i|p_{ij}) + \log \pi(p_{ij}) \\
\text{s.t.} \qquad & 0 \leq p_{ij} \leq 1, \\
& \frac{p_{iK}}{p_{i-1,K}} \geq exp(c_i), \\
& \sum_{j=1}^{K} p_{ij} = 1, \\
& \frac{p_{ij}}{p_{i,j-1}} \geq exp(b_{ij}) \quad if \quad j \leq i, \\
& \frac{p_{ij}}{p_{i,j-1}} \leq exp(b_{ij}) \quad if \quad j \geq i, \\
& c_i \geq 0, \\
& b_{ij} \geq 0, \\
& 1 \leq i \leq K, \\
& 1 \leq j \leq K.
\end{aligned}
\tag{3.1}
$$

where

$$\log P(X_i|p_{ij}) = \sum_{i=1}^{K}\Big(\sum_{j=1}^{K} n_{ij}\log(p_{ij})\Big),$$
$$\log \pi(p_{ij}) = \sum_{i=1}^{K}\log \pi_i(p_{ij}),$$
$$\pi_i(p_i) = \prod_{i=1}^{K}\left[\frac{1}{B(s_i)}\prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1}\right].$$

To solve this optimization problem, we propose a two-step progressive method which first to calibrate the smoothing default rates p_{iK} for each risk rating i to make the default rates monotonic increasing as the rating's creditworthiness worsen, then plug these calibrated default rates into (3.1) to calibrate the transition probability of p_{ij} for the transition matrix. The second step can be solved in three different ways: the first one ignores the prior distribution then the equation (3.1) is exactly same as the CMLE; the second solution is to calibrate the conjugate prior Dirichlet distribution's parameters of α_i and α_{ij} if we have enough historical transition matrix data for the portfolio; the third solution is to apply the empirical Bayes model.

Here is a briefly introducation of the empirical Bayes model [62], the estimated transition probability $\hat{p}_{ij}$ can be calculated by

$$\hat{p}_{ij} = \frac{\frac{n_{ij}+M_jQ_j}{(n_{i\cdot}+M_j)+\varepsilon}}{1+K\varepsilon}, \tag{3.2}$$

where $n_{i\cdot} = \sum_{j=1}^{K} n_{ij}$ and M_j are weights chosen to be large when the observed transition probabilities are not significantly different from the marginal frequencies and M_j is chosen to be near to 0 when the raw transition frequencies are significantly different from the marginal ones. The formulas for M_j are

$$M_j = \frac{1}{max\{.0001, \frac{[\sum_{j=1}^{K} \frac{(n_{ij} - n_{i.}Q_j)^2}{n_{i.}Q_j} - K + 1]}{n_{i.} - K + 1}\}}, \tag{3.3}$$

where the Q_js are the marginal probabilities of rating j, modified using ε so that they are all positive, that is

$$Q_j = \frac{n_j + N\varepsilon}{N(1 + K\varepsilon)}. \tag{3.4}$$

The empirical Bayes model is generated by Dirichlet prior distribution. The detail derivation of the empirical Bayes model please refer [36].

A corporate transition matrix of the financial institution from S&P (source: table 52, year 2011, [55]) is used to illustrate the proposed methodology shown in Table 3.1 below, which shows that raw transition matrix for this portfolio which is non-smoothing and the sum of transition matrix is not 1 since the 'NR' rating exist ('NR' rating means S&P does not rate a particular obligation as a matter of policy) , also the estimated transition probabilities are not monotonic decreasing away from the diagonal.

Table 3.1: Corporate transition matrix - financial institutions (%).

Rating	Issuers	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC/C	D	NR
AAA	30	36.67	53.33	3.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6.67
AA+	11	0	90.91	9.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA	47	0	0	25.53	59.57	6.38	0	0	2.13	0	0	0	0	0	0	0	0	0	0	6.38
AA-	67	0	0	1.49	52.24	32.84	4.48	1.49	1.49	0	0	0	0	0	0	0	0	0	0	5.97
A+	149	0	0	0	7.38	58.39	22.15	0.67	1.34	0.67	0	0	0	0	0	0	0	0	0	9.4
A	138	0	0	0	0	11.59	53.62	22.46	6.52	0.72	0.72	0	0	0	0	0	0	0	0	4.35
A-	127	0	0	0	0	0.79	14.96	48.82	13.39	11.81	3.15	3.15	1.57	0	0	0	0	0	0	2.36
BBB+	99	0	0	0	0	0	0	13.13	67.68	8.08	5.05	1.01	1.01	0	0	0	0	0	0	4.04
BBB	119	0	0	0	0	0	0	0.84	6.72	73.95	6.72	3.36	2.52	0.84	0	0	0	0	0	5.04
BBB-	102	0	0	0	0	0	1.96	0	0.98	17.65	64.71	2.94	1.96	0	0	0	0	0	0.98	8.82
BB+	45	0	0	0	0	0	0	0	2.22	4.44	20	48.89	2.22	4.44	4.44	0	0	2.22	0	11.11
BB	58	0	0	0	0	0	0	0	0	1.72	5.17	24.14	43.1	5.17	0	0	1.72	5.17	0	13.79
BB-	61	0	0	0	0	0	0	0	0	1.64	0	4.92	13.11	49.18	6.56	9.84	1.64	0	0	13.11
B+	51	0	0	0	0	0	0	0	0	0	0	0	7.84	17.65	58.82	3.92	3.92	0	0	7.84
B	60	0	0	0	0	0	0	0	0	0	0	0	0	1.67	21.67	51.67	5	3.33	3.33	13.33
B-	40	0	0	0	0	0	0	0	0	0	0	0	0	2.5	2.5	35	42.5	5	2.5	10
CCC/C	13	0	0	0	0	0	0	0	0	0	0	0	0	0	7.69	0	30.77	38.46	7.69	15.38

Table 3.2 below is the count information of issuers' migration in the transition matrix which is the input values for the proposed methodology and from which the numbers in table 3.1 are derived.

Table 3.2: Corporate transition matrix - financial institutions (count).

Rating	Issuers	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC/C	D	NR
AAA	30	11	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
AA+	11	0	10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA	47	0	0	12	28	3	0	0	1	0	0	0	0	0	0	0	0	0	0	3
AA-	67	0	0	1	35	22	3	1	1	0	0	0	0	0	0	0	0	0	0	4
A+	149	0	0	0	11	87	33	1	2	1	0	0	0	0	0	0	0	0	0	14
A	138	0	0	0	0	16	74	31	9	1	1	0	0	0	0	0	0	0	0	6
A-	127	0	0	0	0	1	19	62	17	15	4	4	2	0	0	0	0	0	0	3
BBB+	99	0	0	0	0	0	0	13	67	8	5	1	1	0	0	0	0	0	0	4
BBB	119	0	0	0	0	0	0	1	8	88	8	4	3	1	0	0	0	0	0	6
BBB-	102	0	0	0	0	0	2	0	1	18	66	3	2	0	0	0	0	0	1	9
BB+	45	0	0	0	0	0	0	0	1	2	9	22	1	2	2	0	0	1	0	5
BB	58	0	0	0	0	0	0	0	0	1	3	14	25	3	0	0	1	3	0	8
BB-	61	0	0	0	0	0	0	0	0	1	0	3	8	30	4	6	1	0	0	8
B+	51	0	0	0	0	0	0	0	0	0	0	0	4	9	30	2	2	0	0	4
B	60	0	0	0	0	0	0	0	0	0	0	0	0	1	13	31	3	2	2	8
B-	40	0	0	0	0	0	0	0	0	0	0	0	0	1	1	14	17	2	1	4
CCC/C	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	5	1	2

The smoothing process is conducted this way; we first smooth the default rates DR for this transition matrix by our proposed methodology as described in 2.14. And thus we will get the constraint for the summation of transition probability for each non-default risk rating i, $\sum_{i=1}^{K} p_{ij} = 1 - DR_i$. Then we continue to run an optimization process with all constraints of 3.1 to get the estimates for the transition probabilities between the non-zero observed transition probabilities of each risk rating. We can not estimate the transition probabilities out of range of non-zero observed transition probabilities for each risk rating since we have any information about them, so the transition probabilities for them are still 0. The smoothed transition matrix by the proposed method is shown in table 3.3. It has all the properties of a well-behaved transition matrix as we expected and can be used for calibrating the PD term structure.

Table 3.3: Smoothed corporate transition matrix - financial institutions

Rating	Issuers	Default	NR	Default Rate	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC/C
AAA	30	0	2	0.00045	0.4820	0.4816	0.0360	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA+	11	0	0	0.00047	0	0.9090	0.0915	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA	47	0	3	0.00050	0	0	0.1200	0.6500	0.1100	0.0410	0.0400	0.0385	0	0	0	0	0	0	0	0	0
AA-	67	0	4	0.00053	0	0	0.0600	0.5250	0.3250	0.0447	0.0300	0.0148	0	0	0	0	0	0	0	0	0
A+	149	0	14	0.00056	0	0	0	0.0810	0.0637	.0241.241	0.0180	0.0150	0.0074	0	0	0	0	0	0	0	0
A	138	0	6	0.00059	0	0	0	0	0.1210	0.5600	0.2350	0.0683	0.0076	0.0075	0	0	0	0	0	0	0
A-	127	0	3	0.00062	0	0	0	0	0.0500	0.1460	0.4770	0.1321	0.1150	0.0299	0.0299	0.0165	0	0	0	0	0
BBB+	99	0	4	0.00065	0	0	0	0	0	0	0.1400	0.7000	0.0800	0.0560	0.0120	0.0113	0	0	0	0	0
BBB	119	0	6	0.00088	0	0	0	0	0	0	0.0300	0.0800	0.8000	0.0750	0.0050	0.0050	0.0041	0	0	0	0
BBB-	102	1	9	0.00119	0	0	0	0	0	0.0180	0.0180	0.0200	0.2000	0.7350	0.0040	0.0038	0	0	0	0	0
BB+	45	0	5	0.00125	0	0	0	0	0	0	0	0.0260	0.0550	0.2400	0.5900	0.0290	0.0160	0.0160	0.0090	0.0090	0.0087
BB	58	0	8	0.00169	0	0	0	0	0	0	0	-	0.0200	0.0600	0.2900	0.5100	0.0620	0.0150	0.0140	0.0140	0.0133
BB-	61	0	8	0.00228	0	0	0	0	0	0	0	0	0.0090	0.0090	0.0560	0.1500	0.5650	0.0960	0.0940	0.0190	0
B+	51	0	4	0.00240	0	0	0	0	0	0	0	0	0	0	0	0.0850	0.1910	0.6360	0.0430	0.4260	0
B	60	2	8	0.00740	0	0	0	0	0	0	0	0	0	0	0	0	0.0200	0.2400	0.6310	0.0610	0.0410
B-	40	1	4	0.00777	0	0	0	0	0	0	0	0	0	0	0	0	0.0280	0.0280	0.3970	0.4820	0.0570
CCC/C	13	1	2	0.04407	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0480	0.0480	0.3820	0.4780

3.3 Conclusion and Future Work

In this chapter, following the previous chapters empirical PD rate smoothing algorithm, we propose a smoothing algorithm to smooth the empirical PD transition matrics to satisfy the industry's expectation. The well-behaved PD transition matrix should have the properties of monotonic increasing default probabilities, the sum of PD transition probabilities for each rating is 1 and decreasing transition probabilities as the transition state away from the diagonal (stay at the same rating). However, the empirical internal PD transition matrix usually does not hold these properties. We propose a two-step smoothing methodology for the observed transition matrix, which applies the CMAP on the observed default rates first then conduct an optimization process to obtain the well-behaved transition matrix with maximum likelihood score. We apply the proposed algorithm to the $S\&P$'s data; the result shows that the proposed algorithm can provide a reliable PD transition matrix.

4 Mixed Response Model with Pairwise Composite Likelihood Method

In this chapter, we focus on solving the regression problem of the high dimensional correlated mixed response data which contains both continuous and discrete response variables using the pairwise composite likelihood method. This kind of regression widely exists in economic, risk management, finance, biomedical, and psychological health study. An example of risk management is that the data of a real estate mortgage portfolio which may include both discrete outcomes of default, performing and prepayment, also contains the continuous outcomes of loss amount and prepayment amount as well, and the outcomes share the universal (or partial) risk drivers.

4.1 Introduction

Mixed response model is that the dependent variables include both continuous and discrete (binary or categorical) variables and the dependent variables are correlated. Many existing approaches have been proposed to solve the bivariate mixed response problems or multivariate continuous (or discrete) response, such as multivariate multiple regression [37] deals with the multiple continuous response problems, Multiple-Discrete Choice Models [38] is applied to solve the multiple categorical (binary) response problems. Some research in multiple mixed response model [64, 39]. We adopt the pairwise composite likelihood approach for jointly analyzing high dimensional mixed responses of different types (e.g., binary (categorical) and continuous data). Our approach is based on a latent Gaussian process to represent the discrete response variables and then model the dependence among the responses variables assuming that the latent variables and the continuous variables are multivariate normal distributed. Our proposed method can be used to estimate the high dimensional regression problem of mixed response with either common explanatory variables for all response or different explanatory variables for each response variable.

There are several ways to solve the mixed responses model in the literature. Olkin and Tate [40] proposed the general location model by decomposing the joint distribution of the continuous and categorical variables into a marginal multinomial distribution and a conditional multivariate normal distribution respectively to the categorical variables and the continuous variables given the categorical variables. Yang et al. [41] extend this method to mixed Poisson and continuous responses through a likelihood-based approach. Another approach for mixed response model adopts an opposite factorization way against the general location model which decompose the joint distribution to a multivariate marginal distribution for the continuous responses and a conditional distribution of the categorical variables given the continuous variables. Cox and Wermuth [42] empirically compared the choice between these two methods and the independency and conditional independency as well. Moreover, Heckman [43] proposed a general model for simultaneously analyzing two mixed correlated responses which take into account the correlation between errors in the model for responses. Ryan [44] and Fitzmaurice and Laird [45] extend and used this approach for clustering the outcomes from discrete and continuous response.

In high dimensional correlated mixed response data, the main obstacle lies in both the tremendous computational complexity from the high dimensions and the fact that continuous and categorical response has the intrinsically different measure scale. In this dissertation, we propose to apply the pairwise composite likelihood method to solving the regression models composed with mixed response variable through leveraging the latent variable representing the discrete response by constraining the parameters of the latent model proposed by Dunson [46] for identifiability without restrictions on the correlation. Thus, the dimension of the response is reduced to two and the scales of all response convert to the same. This approach simplified the complexity of high dimension mixed unequal scales response regression problem to three types of two-dimensional response variables: both continuous response, both discrete response and mixed continuous and discrete response.

Another significant merit of applying pairwise likelihood approach is that, without any change in the methodology, our proposed method can estimate three types high dimensional regression problems: high dimension continuous response (also traditionally called multivariate multilevel model), high dimension discrete response (multilevel discrete choice model), and high dimension mixed discrete and continuous response models.

A simulation study is conducted to examine the proposed methodology and compare the proposed methodology against the marginal regression with respect to their estimation outcomes and the performance. The results show the proposed method can provide accurate and consistent estimator for high dimension mixed correlated response data. The proposed methodology is implemented by using R code; the Newton-Raphson method is applied to reach the optimum value of estimators. In this dissertation,the pairwise composite likelihood method is conducted to reduce the dimension of the mixed response variables models. We briefly introduce the methodology here.

Pairwise Composite Likelihood Estimation

Composite likelihood methods are first introduced by Lindsay [47], which is motivated by the issue of computational feasibility arising in the likelihood method with high-dimensional data analysis. The idea of projecting high-dimensional complicated likelihood functions to low-dimensional computationally feasible likelihood objects by multiplying a collection of component likelihoods, which is methodologically appealing. The individual component can be a conditional density, marginal density, or pairwise density. The estimating equation obtained from the derivative of the composite log-likelihood is an unbiased estimating equation. The literature on both theoretical and practical applications for inference based on composite likelihood continues to expand fast as shown in the overview by Varin, Reid and Firth [48]. They are extensions of the Fisher's likelihood theory, one of the most influential approaches in statistics. Composite likelihood inherits many functional properties of inference based on the full likelihood function but is more easily implemented on high-dimensional data sets.

The bivariate pairwise composite likelihood is one special case of a composite likelihood, in which the pseudo-likelihood is defined as the product of the bivariate likelihood of all possible pairs of observations. Accordingly, it only captures the bivariate relationships among the variables within the random vector. Therefore, it can only be used to make inference for parameters that identical in the bivariate and multivariate densities. For example, the mean parameter and the pairwise correlation take the same values in the bivariate and multivariate densities for multivariate normal random variables.

The maximum composite likelihood estimators can be reached at the maximum of the composite likelihood. When deriving the statistical inference, composite likelihood needs use Godambe information [49] to replace Fisher information, and the Godambe information is asymptotic efficiency computed under some regularity conditions compared to Fisher information I. The calculation of Godambe information G of the parameters $\boldsymbol{\eta}$ for the log composite likelihood function $cl(\boldsymbol{\eta})$ needs sensitivity or Hessian matrix H and variability matrix J. The parameters $\boldsymbol{\eta}$ can be obtained by solving the composited score function $U(\boldsymbol{\eta})$. If the composite likelihood function is a true log-likelihood function, $G = H = I$. Otherwise, the Godambe information is calculated by

$$G(\hat{\boldsymbol{\eta}}) = H(\hat{\boldsymbol{\eta}})J^{-1}(\hat{\boldsymbol{\eta}})H(\hat{\boldsymbol{\eta}}),$$

where

$$J(\hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} U(\hat{\boldsymbol{\eta}})U(\hat{\boldsymbol{\eta}})^T,$$

$$H(\hat{\boldsymbol{\eta}}) = \left.\frac{\partial^2 cl(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\right|_{\hat{\boldsymbol{\eta}}}.$$

Latent Variables

The proposed mixed response model leverages the latent variable concept to deal with the different measure scale issues of the discrete outcomes. Before diving into the proposed mixed response model, we briefly review the latent variable here first. Suppose M_i^* is the value of the latent variable for the observation i ($i = 1, ..., I$), and M_i is the corresponding observable binary variate. Under the latent variable assumption, there is a threshold value t for a binary observation such that

$$M_i = \begin{cases} 1, & \text{if } M_i^* \geq t, \\ 0, & \text{if } M_i^* \leq t, \end{cases}$$

where 0 and 1 are arbitrary binary codings for M_i. The example here involves binary outcomes but the idea is easily generalized to multiple categorical and ordinal outcomes. Now consider the following linear regression model on M_i^*,

$$M_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where X_i represents the explanatory variables for the i_{th} observation unit. Assume that ε_i are *iid* and $\varepsilon_i \sim N(0, \sigma)$, the logistic model or cumulative probit model for the binary outcomes is naturally derived according to the normal distribution, as described, for example, by Anderson and Sophia [50]. Since this dissertation use probit model, we describe it below,

$$\begin{aligned} P(M_i = 1|X_i) &= 1 - \Phi(\frac{t - \beta_0 - \beta_1 X_i}{\sigma}), \\ P(M_i = 0|X_i) &= \Phi(\frac{t - \beta_0 - \beta_1 X_i}{\sigma}). \end{aligned}$$

The latent variables have been used in many mixed response models to analyze the regression problems of multiple noncommensurate outcomes. Julie et al. [51] apply the latent variable with partial likelihood to solve the bivariate mixed response model with binary outcome. And Paul [44] applies the latent variable for mixed response model with categorical (binary or multi-category) outcomes. Both methodologies assume that the joint distribution of the continuous outcomes and the latent variables behind the binary (ordinal) outcomes are a bivariate normal distribution. Moreover, Green [38] applies the latent variable concept to the discrete choice model which deals with the multiple binary (or multiple levels) outcomes using the assumption that the latent variables from the discrete outcomes and the continuous response are from a multivariate normal distribution. In this dissertation, we continue to follow the assumption that the latent variables from the discrete outcomes are from a multivariate normal distribution to construct the joint model for mixed responses.

To focus on the proposed methodology of applying pairwise composite likelihood method to the high dimensional correlated mixed response, the discrete variables in this dissertation are all binary response, but the proposed method is easy to extend to categorical (ordinal) discrete outcomes.

4.2 Mixed Response with Composite Likelihood

4.2.1 Model Setup

Suppose there are n observations in a dataset and each observation has q correlated response variables, in which m response variables are binary outcomes, and $q-m$ response variables are continuous outcomes. The response variables may or may not share the common explanatory variables and the design matrices for each response variable may or may not be balanced. We let $\boldsymbol{M}$ represents the binary response and $\boldsymbol{Y}$ represents the continuous response variable. To solve this kind of high dimensional mixed response regression problem, we assume that there are unknown latent variable $\boldsymbol{M}^*$ where $\boldsymbol{M} = \text{sgn}(\boldsymbol{M}^* > 0)$ is hidden behind the binary variable, and the joint distribution of the continuous response variables and latent variables are normally distributed,

$$\begin{pmatrix} M^{(1)*} \\ M^{(2)*} \\ \cdots \\ M^{(m)*} \\ Y^{(m+1)} \\ Y^{(m+2)} \\ \cdots \\ Y^{(q)} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \\ \cdots \\ \mu^{(m)} \\ \mu^{(m+1)} \\ \mu^{(m+2)} \\ \cdots \\ \mu^{(q)} \end{pmatrix}, \quad \sum \right),$$

where $\sum$ is a symmetric matrix as

$$\begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1m}\sigma_1\sigma_m & \cdots & \cdots & \cdots & \rho_{1q}\sigma_1\sigma_q \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2m}\sigma_2\sigma_m & \cdots & \cdots & \cdots & \rho_{2q}\sigma_2\sigma_q \\ & & \cdots & & & & & \\ \rho_{1m}\sigma_1\sigma_m & \rho_{2m}\sigma_2\sigma_m & \cdots & \sigma_m^2 & \cdots & \cdots & \cdots & \rho_{mq}\sigma_m\sigma_q \\ \rho_{1,m+1}\sigma_1\sigma_{m+1} & \rho_{2,m+1}\sigma_2\sigma_{m+1} & \cdots & \rho_{m,m+1}\sigma_m\sigma_{m+1} & \sigma_{m+1}^2 & \cdots & \cdots & \rho_{m+1,q}\sigma_{m+1}\sigma_q \\ \rho_{1,m+2}\sigma_1\sigma_{m+2} & \rho_{2,m+2}\sigma_2\sigma_{m+2} & \cdots & \cdots & \cdots & \sigma_{m+2}^2 & \cdots & \rho_{m+2,q}\sigma_{m+2}\sigma_q \\ & & & & & & & \\ \cdots & \cdots & \cdots & & & & \cdots & \\ \rho_{1q}\sigma_1\sigma_q & \rho_{2q}\sigma_2\sigma_q & \cdots & \rho_{mq}\sigma_m\sigma_q & \cdots & \cdots & \cdots & \sigma_q^2 \end{pmatrix}.$$

It is assumed that all response variables of the continuous outcomes and the hidden latent variables behind the binary outcomes follow the multivariate normal distribution. Given that the measurement scale of the binary response is unmeasurable, the dimension of response in the regression may be high, and the computation complexity of the direct solution, it is natural and convenient to apply the pairwise composite likelihood method to reduce the dimension of the equation system from q to 2. Consequently, the likelihood function needs to be instead of the composite pairwise likelihood function to obtain the estimators of the regressions. For the simplicity perspective, we use the notation of $Z^{(i)}$ to replace either $M^{(i)*}$ if the i^{th} response variable is binary or $Y^{(i)}$ if the i^{th} response variable is continuous, as shown below,

$$\begin{pmatrix} Z^{(1)} \\ Z^{(2)} \\ \cdots \\ Z^{(m)} \\ Z^{(m+1)} \\ Z^{(m+2)} \\ \cdots \\ Z^{(q)} \end{pmatrix} = \begin{pmatrix} M^{(1)*} \\ M^{(2)*} \\ \cdots \\ M^{(m)*} \\ Y^{(m+1)} \\ Y^{(m+2)} \\ \cdots \\ Y^{(q)} \end{pmatrix}.$$

The parameters need to be estimated for any pair of response variables $(Z^{(j)}, Z^{(k)}$ $\boldsymbol{\eta}_{j,k}$, which includes $\boldsymbol{\beta}_j, \boldsymbol{\beta}_k, \sigma_j, \sigma_k, \rho_{jk}$. According to the properties of the multivariate normal distribution, the pair of response variables $(Z^{(j)}, Z^{(k)})$ follows the bivariate normal distribution. Thus, the regression of each pair of response variables $(Z^{(j)}, Z^{(k)})$ is shown as below,

$$\begin{aligned} z_i^{(j)} &= \boldsymbol{x}_i^{(j)}\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_i^{(j)}, \quad i = 1 \cdots n \\ z_i^{(k)} &= \boldsymbol{x}_i^{(k)}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_i^{(k)}, \quad i = 1 \cdots n \end{aligned}$$

where, $\boldsymbol{\epsilon}_i^{(j)}$s is iid and $\boldsymbol{\epsilon}_i^{(k)}$s is iid respectively as well, however, they are distributed as a bivariate normal with correlation ρ_{jk} and mean 0,

$$\begin{pmatrix} \epsilon_i^{(j)} \\ \\ \epsilon_i^{(k)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_j^2, & \rho_{jk}\sigma_j\sigma_k \\ \\ \rho_{jk}\sigma_j\sigma_k, & \sigma_k^2 \end{pmatrix} \right).$$

Let

$$\boldsymbol{z}_i^{(j,k)} = \begin{pmatrix} \boldsymbol{z}_i^{(j)} \\ \\ \boldsymbol{z}_i^{(k)} \end{pmatrix}, \boldsymbol{\mu}_i^{(j,k)} = \begin{pmatrix} \boldsymbol{x}_i^{(j)}\boldsymbol{\beta}_j \\ \\ \boldsymbol{x}_i^{(k)}\boldsymbol{\beta}_k \end{pmatrix}, \Sigma^{(j,k)} = \begin{pmatrix} \sigma_j^2, & \rho_{j,k}\sigma_j\sigma_k \\ \\ \rho_{j,k}\sigma_j\sigma_k, & \sigma_k^2 \end{pmatrix}.$$

The joint density function of $(\boldsymbol{z}_i^{(j)}, \boldsymbol{z}_i^{(k)})$ will be

$$\begin{aligned} & h(\boldsymbol{z}_i^{(j)}, \boldsymbol{z}_i^{(k)}; \boldsymbol{x}_i, \boldsymbol{\eta}_{j,k}) && (4.1) \\ = \; & (2\pi)^n \Sigma^{(j,k)\frac{n}{2}} \exp\big(-\frac{1}{2}(\boldsymbol{z}_i^{(j,k)} - \mu^{(j,k)})^T \Sigma^{(j,k)-1}(\boldsymbol{z}_i^{(j,k)} - \mu^{(j,k)})\big). \end{aligned}$$

The likelihood function for the pair of variables of $\boldsymbol{z}_j^{(*)}$ and $\boldsymbol{z}_k^{(*)}$ is given by

$$L_{j,k}(\boldsymbol{\eta}_{j,k}) \;\; = \;\; \prod_{i=1}^{n} h_(\boldsymbol{z}_i^{(j)}, \boldsymbol{z}_i^{(k)}; \boldsymbol{x}_i, \boldsymbol{\eta}_{j,k}).$$

The parameters estimation process via Fisher scoreing can be obtained by solving the score equation

$$U_{j,k}(\boldsymbol{\eta}_{j,k}) = \Sigma_{i=1}^{n} h_(\boldsymbol{z}_i^{(j)}, \boldsymbol{z}_i^{(k)}; \boldsymbol{x}_i, \boldsymbol{\eta}_{j,k})^{-1} \frac{\partial}{\partial \boldsymbol{\eta}_{j,k}} h_(\boldsymbol{z}_i^{(j)}, \boldsymbol{z}_i^{(k)}; \boldsymbol{x}_i, \boldsymbol{\eta}_{j,k}).$$

Accordingly, let $\boldsymbol{\eta}_{j,k} = (\boldsymbol{\beta}_j^{*T}, \boldsymbol{\beta}_{\boldsymbol{k}}^{*T}, \boldsymbol{\rho}_{jk}, \boldsymbol{\sigma}_j, \boldsymbol{\sigma}_k)$, then $\boldsymbol{\eta}$ be the superset of parameters $\boldsymbol{\eta}_{j,k}$, then the pairwise composite likelihood function of these q response variables model is given by

$$CL(\boldsymbol{\eta}) = \prod_{j=1}^{q-1}\prod_{k=j+1}^{q} L_{j,k}(\boldsymbol{\eta}), \tag{4.2}$$

and the log-likelihood function and the score function are given by

$$\begin{aligned} cl(\boldsymbol{\eta}) &= \log CL(\boldsymbol{\eta}) \\ &= \sum_{j=1}^{q-1}\sum_{k=j+1}^{q} l_{j,k}(\boldsymbol{\eta}_{jk}), \qquad (4.3) \\ S(\boldsymbol{\eta}) &= \frac{\partial cl(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \\ &= \sum_{j=1}^{q-1}\sum_{k=j+1}^{q} \frac{\partial l_{j,k}(\boldsymbol{\eta}_{j,k})}{\partial \boldsymbol{\eta}_{j,k}}. \qquad (4.4) \end{aligned}$$

Now the problem is to solve the score functions derived from three types of bivariate response regression: both continuous, both binary and mixed with one continuous and one binary response. If we can solve the problem combined with these three types of bivariate response regression simultaneously, we can easily extend the solution to the problems of correlated high dimensional mixed response regression since the components of composite pairwise likelihood functions are exactly same.

To illustrate our proposed methodology, we use the simplest data set that includes all three types of combinations of bivariate response variable, which is a data set includes n observations and each observation has two binary outcomes of $M_i^{(1)}$ and $M_i^{(2)}$ and two continuous outcomes variables of $Y_i^{(3)}$ and $Y_i^{(4)}$, where $i = 1, ..., n$. let

$M_i^{(1)*}$ and $M_i^{(2)*}$ are the unknown latent variables from binary response $M_i^{(1)}$ and $M_i^{(2)}$. The binary response $M^{(j)}$, $j = 1, 2$, is assumed that $M^{(j)} = \text{sgn}(M_j^* > 0)$. The 4 response variables are correlated and they either share common covariates or have their own scalar covariate separately. By the assumption, $(M^{(1)*}, M^{(2)*}, Y^{(3)}, Y^{(4)})^T$ follow a multivariate normal distribution. To simplify the notation, we use Z to represent these four response variables,

$$\begin{pmatrix} Z^{(1)} \\ Z^{(2)} \\ Z^{(3)} \\ Z^{(4)} \end{pmatrix} = \begin{pmatrix} M^{(1)*} \\ M^{(2)*} \\ Y^{(3)} \\ Y^{(4)} \end{pmatrix}.$$

Thus, the regression model for each response variables will be

$$\begin{aligned} z_i^{(1)} &= \boldsymbol{x}_i^{(1)} \beta_1^* + \epsilon_i^{(1)}, \quad i = 1 \cdots n \\ z_i^{(2)} &= \boldsymbol{x}_i^{(2)} \beta_2^* + \epsilon_i^{(2)}, \\ z_i^{(3)} &= \boldsymbol{x}_i^{(3)} \beta_3 + \epsilon_i^{(3)}, \\ z_i^{(4)} &= \boldsymbol{x}_i^{(4)} \beta_4 + \epsilon_i^{(4)}, \end{aligned}$$

where $\epsilon_i^{(1)}$s, $\epsilon_i^{(2)}$s, $\epsilon_i^{(3)}$s, and $\epsilon_i^{(4)}$s are iid respectively and they are normally distributed as

$$\begin{pmatrix} \epsilon_i^{(1)} \\ \epsilon_i^{(2)} \\ \epsilon_i^{(3)} \\ \epsilon_i^{(4)} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix} \right). \quad (4.5)$$

Let $g(\mu) = X^T\boldsymbol{\beta} + \varepsilon$, $X = (X_1, ..., X_4)^T$ is the design matrix, $\beta_1^{*T}, \beta_2^{*T}, \beta_3, \beta_4$ are the unknown parameter vectors, and $g(\cdot)$ is a known monotone link function. Common choices of g include logit, probit, or complement log-log functions. The parameters need to be estimated $\boldsymbol{\eta} = (\beta_1^{*T}, \beta_2^{*T}, \beta_3^T, \beta_4^T, \rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}, \sigma_1, \sigma_2, \sigma_3, \sigma_4)^T$.

Accordingly, the pairwise composite likelihood function of these four response variables model is given by

$$\begin{aligned} CL(\boldsymbol{\eta}) &= \prod_{j=1}^{3}\prod_{k=2}^{4} L_{j,k}(\boldsymbol{\eta}) \\ &= L_{1,2}(\boldsymbol{\eta})L_{1,3}(\boldsymbol{\eta})L_{1,4}(\boldsymbol{\eta})L_{2,3}(\boldsymbol{\eta})L_{2,4}(\boldsymbol{\eta})L_{3,4}(\boldsymbol{\eta}), \end{aligned} \quad (4.6)$$

and the log-likelihood function and the score function are given by

$$
\begin{aligned}
cl(\boldsymbol{\eta}) &= logCL(\boldsymbol{\eta}) \\
&= logL_{1,2}(\boldsymbol{\eta}) + logL_{1,3}(\boldsymbol{\eta}) + logL_{1,4}(\boldsymbol{\eta}) + logL_{2,3}(\boldsymbol{\eta}) + logL_{2,4}(\boldsymbol{\eta}) + logL_{3,4}(\boldsymbol{\eta}) \\
&= l_{1,2}(\boldsymbol{\eta}) + l_{1,3}(\boldsymbol{\eta}) + l_{1,4}(\boldsymbol{\eta}) + l_{2,3}(\boldsymbol{\eta}) + l_{2,4}(\boldsymbol{\eta}) + l_{3,4}(\boldsymbol{\eta}),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial cl(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\partial l_{1,2}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + \frac{\partial l_{1,3}(\boldsymbol{\eta})}{\partial \eta} + \frac{\partial l_{1,4}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + \frac{\partial l_{2,3}(\eta)}{\partial \boldsymbol{\eta}} + \frac{\partial l_{2,4}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + \frac{\partial l_{3,4}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \\
&= U_{1,2}(\boldsymbol{\eta}) + U_{1,3}(\boldsymbol{\eta}) + U_{1,4}(\boldsymbol{\eta}) + U_{2,3}(\boldsymbol{\eta}) + U_{2,4}(\boldsymbol{\eta}) + U_{3,4}(\boldsymbol{\eta}). \qquad (4.7)
\end{aligned}
$$

Please note, as mentioned by many researchers, such as Julie et al. [51], the unknown latent variance is not uniquely identifiable since it is unobservable. However, we can rescale the latent variable by dividing its' standard deviation to produce a new latent variable which has the variance of 1. Thus, in our equation, both σ_1 and σ_2 are constant 1 and do not need to be estimated by our methodology.

Next, we will introduce the detailed process for solving these three cases of bivariate regressions respectively, they are the fundamental building blocks of pairwise composite likelihood method for high dimensional mixed response regression.

4.2.2 Case 1: Both Continuous Outcomes

The first case is both continuous outcomes. If j = 3 and k = 4 in our simplest four response variables model setup, then the bivariate regression of this combination is the case of both continuous response variables. In fact, it is an extension of simple linear regression which has just one continuous response variable and is the simplest multiple level linear regression. According to the assumption of multivariate normal distribution assumption of the response variables, the response variables $Z^{(3)}$ and $Z^{(4)}$ follow the bivariate normal distribution $\mathcal{N}_2$ with the correlation of ρ_{34} and standard deviation of σ_3 and σ_4 respectively. The parameters need to be estimated in the regression will be $\boldsymbol{\eta}_{34} = \{\beta_3, \beta_4, \sigma_3, \sigma_4, \rho_{34}\}$. Remember that,

$$\begin{aligned} z_i^{(3)} &= \boldsymbol{x}_i^{(3)}\beta_3 + \epsilon_i^{(3)}, \quad i = 1 \cdots n \\ z_i^{(4)} &= \boldsymbol{x}_i^{(4)}\beta_4 + \epsilon_i^{(4)}, \quad i = 1 \cdots n \end{aligned}$$

where $\epsilon_i^{(3)}$s and $\epsilon_i^{(4)}$s are iid respectively. However, they are bivariate normally distributed as

$$\begin{pmatrix} \epsilon_i^{(3)} \\ \epsilon_i^{(4)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_3^2, & \rho_{34}\sigma_3\sigma_4 \\ \rho_{34}\sigma_3\sigma_4, & \sigma_4^2 \end{pmatrix} \right).$$

Let

$$\begin{aligned}\mu_i^{(3)} &= \boldsymbol{x}_i^{(3)}\boldsymbol{\beta}_3,\\ \mu_i^{(4)} &= \boldsymbol{x}_i^{(4)}\boldsymbol{\beta}_4.\end{aligned}$$

Then the joint pdf of $Z^{(3)}$ and $Z^{(4)}$ is

$$\begin{aligned}&f_{3,4}(z_i^{(3)}, z_i^{(4)})\\ =&\frac{1}{2\pi\sigma_3\sigma_4\sqrt{1-\rho_{34}^2}}\exp\Big(-\frac{1}{2(1-\rho_{34}^2)}\Big[\frac{(z_i^{(3)}-\mu_i^{(3)})^2}{\sigma_3^2}\\ &+\frac{(z_i^{(4)}-\mu_i^{(4)})^2}{\sigma_4^2}-\frac{2\rho_{34}(z_i^{(3)}-\mu_i^{(3)})(z_i^{(4)}-\mu_i^{(4)})}{\sigma_3\sigma_4}\Big]\Big).\end{aligned} \tag{4.8}$$

Our goal is to estimate the unknown parameters from the regression of this pair of correlated response variables, and then cooperate with the likelihood function from other pairs response variables to estimate all the unknown parameters in our 4 mixed responses variable model. Thus, we need to derive the likelihood function for it and solve the score function to maximize the likelihood function.

To simplify the equation 4.8, let

$$\Lambda = \frac{1}{(1-\rho_{34}^2)}\left[\frac{(z_i^{(3)}-\mu_i^{(3)})^2}{\sigma_3^2}+\frac{(z_i^{(4)}-\mu_i^{(4)})^2}{\sigma_4^2}-\frac{2\rho_{34}(z_i^{(3)}-\mu_i^{(3)})(z_i^{(4)}-\mu_i^{(4)})}{\sigma_3\sigma_4}\right].$$

The log-likelihood function and score function for the bivariate regression model of the continuous response variables $(Z^{(3)}, Z^{(4)})$ are

$$\begin{aligned}
l_{34} &= \log \prod_{i=1}^{n} f_{3,4}(z_i^{(3)}, z_i^{(4)}) \\
&= \log \prod_{i=1}^{n} \frac{1}{2\pi\sigma_3\sigma_4\sqrt{1-\rho_{34}^2}} \exp(-\frac{1}{2}\Lambda) \\
&= -\sum_{i=1}^{n} \frac{1}{2}\Big(2\log(2\pi) + \log(\sigma_3^2) + \log(\sigma_4^2) + \log(1-\rho_{34}^2) + \Lambda\Big), \\
S_{34} &= \frac{\partial l_{3,4}(\boldsymbol{\eta}_{34})}{\partial \boldsymbol{\eta}_{34}}.
\end{aligned}$$

The estimators of $\boldsymbol{\eta}_{34}$ can be obtained by setting the score function above equal to 0, and the maximum log-likelihood can be reached by Newton-Raphson method numerically. We list the required input for Newton-Raphson method of first and second derivative of l_{34} w.r.t $\boldsymbol{\eta}_{34}$ in Appendix E.

4.2.3 Case 2: Both Binary Outcomes

The second case is both binary outcomes, in our four responses variables model setup, it must be $j = 1$ and $k = 2$. We apply the latent variables concept to the binary response variables $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$. As mentioned previously, the standard deviation of both binary response variables are not identifiable and are set to 1 by rescaling mechanic. Thus, they do not need to be estimated. Therefore, the parameters need to be estimated in this kind of bivariate regression is $\boldsymbol{\eta}_{12} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho_{12})$. In this dissertation, we keep using the notation of $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ for simplicity reasons and assume it is already rescaled to the standard deviation of 1. According to the assumption of multivariate normal distributed of the response variables, the latent variables $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ follow the bivariate normal distribution with the correlation of ρ_{12}, which is also the covariance between the response variables $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$.

Traditionally, economists name the model with both binary outcomes model as the bivariate binary choice model as in literature [38], which is the extension of single discrete choice model, the basic building block of the discrete choice models used to measure the discrete consumer choices, such as whether to vote for a political candidate, whether to purchase a specific brand car, etc. The bivariate binary choice model is used to model the bivariate choice simultaneously, such as choose the university and professor simultaneously from a graduate students perspective. Let's first review the model setup for the bivariate binary regression based on normality assumptions,

$$\begin{aligned} \boldsymbol{z}_i^{(1)} &= \boldsymbol{x}_i^{(1)}\boldsymbol{\beta}_1^* + \epsilon_i^{(1)}, \quad i = 1 \cdots n \\ \boldsymbol{z}_i^{(2)} &= \boldsymbol{x}_i^{(2)}\boldsymbol{\beta}_2^* + \epsilon_i^{(2)}, \quad i = 1 \cdots n \end{aligned}$$

where $\epsilon_i^{(1)}$s and $\epsilon_i^{(2)}$s are iid respectively. However, they are bivariate normally distributed as

$$\begin{pmatrix} \epsilon_i^{(1)} \\ \epsilon_i^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, & \rho_{12} \\ \rho_{12}, & 1 \end{pmatrix} \right),$$

and the mean value of the $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ as below,

$$\begin{aligned} \mu_i^{(1)} &= \boldsymbol{x}_i^{(1)} \boldsymbol{\beta}_1^*, \\ \mu_i^{(2)} &= \boldsymbol{x}_i^{(2)} \boldsymbol{\beta}_2^*. \end{aligned}$$

For the binary outcomes, traditionally researchers use the logit or probit link function to build the linear relationship between the independent variables and dependent variables. Considered the context of the mixed response in our model is assumed normally distributed, it is convenient for the binary response variable adopting the probit link than logit link, which can bring much more convenient in perspective of mathematical complexity. Then, the associated probability for the joint event of $(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$ is

$$\begin{aligned} P(z_i^{(1)} = 1, z_i^{(2)} = 1 | X^{(1)}, X^{(2)}) &= \Phi_2(\mu_i^{(1)}, \mu_i^{(2)}, \rho_{12}), \\ P(z_i^{(1)} = 1, z_i^{(2)} = 0 | X^{(1)}, X^{(2)}) &= \Phi_2(\mu_i^{(1)}, -\mu_i^{(2)}, -\rho_{12}), \\ P(z_i^{(1)} = 0, z_i^{(2)} = 1 | X^{(1)}, X^{(2)}) &= \Phi_2(-\mu_i^{(1)}, \mu_i^{(2)}, -\rho_{12}), \\ P(z_i^{(1)} = 0, z_i^{(2)} = 0 | X^{(1)}, X^{(2)}) &= \Phi_2(-\mu_i^{(1)}, -\mu_i^{(2)}, \rho_{12}), \end{aligned}$$

where Φ_2 denotes the bivariate normal cdf. And the above equations can be rewritten as

$$\begin{aligned} P(Z^{(1)} = z_1, Z^{(2)} = z_2 | X_1, X_2) &= \Phi_2((2z_1 - 1)\mu_1, (2z_2 - 1)\mu_2, (2z_1 - 1)(2z_2 - 1)\rho_{12}) \\ &= \Phi_2(s_1\mu_1, s_2\mu_2, s_1 s_2 \rho_{12}), \end{aligned}$$

where $s_1 = (2z_1 - 1)$ and $s_2 = (2z_2 - 1)$.

The likelihood function and the log-likelihood function of this model is

$$\begin{aligned} L_{12} &= \prod_{i=1}^{n} \Phi_2(s_1\mu_1, s_2\mu_2, s_1 s_2 \rho_{12}), \\ l_{12} &= \sum_{i=1}^{n} \log \Phi_2(s_1\mu_1, s_2\mu_2, s_1 s_2 \rho_{12}). \end{aligned}$$

Green [38] provides the detail solution for the bivariate discrete choice regression which is exactly same as our second case of bivariate binary response regression. Still, we will use the Newton-Raphson method numerically to solve this regression model. The thorough description of the solution process to the case of both binary outcomes and the derivatives of the score function is shown in Appendix F,

4.2.4 Case 3: Mixed Binary and Continuous Outcomes

Now, only the case of mixed bivariate response regression model left. When one response variable $\boldsymbol{Z}^{(j)}$ is binary variable and another response $\boldsymbol{Z}^{(k)}$ is continuous variable, in our four response variables model setup, it must be $j = 1$ or 2 and $k = 3$ or 4. Same as in the case 2, the standard devastation of the rescaled $Z^{(j)}$ is set to 1 and does not need to be estimated. Accordingly, the parameters need to be estimated in the mixed bivariate regression are $\boldsymbol{\eta}_{jk} = (\boldsymbol{\beta}_j^*, \boldsymbol{\beta}_k, \sigma_k, \rho_{jk})$. The joint distribution of the latent variables $Z^{(j)}$ and the continuous variable $Z^{(k)}$ follow the bivariate normal distribution with the correlation ρ_{jk}.

To model the correlated mixed continuous and binary outcomes simultaneously, we apply the reverse factorization method against the general location model proposed by Olkin and Tate [40] for the joint bivariate normal distribution: a marginal model for the continuous outcome and a probit model for the binary variable which conditions on the continuous outcome. This factorization method provides a mathematically convenient way to model two correlated outcomes. There are advantages of using the latent variable formulation that it leads to a set of intuitively appealing covariates in the conditional model and provides a moment structure which can be easily extended to allow for splitting the binary outcomes. In addition, the coefficients of the conditional model are functions of the variance and correlation parameters from the underlying latent variable model.

Recall that when $j = 1, 2$ and $k = 3, 4$,

$$\begin{aligned} z_i^{(j)} &= \boldsymbol{x}_i^{(j)} \boldsymbol{\beta}_j^* + \epsilon_i^{(j)}, \quad i = 1 \cdots n \\ z_i^{(k)} &= \boldsymbol{x}_i^{(k)} \boldsymbol{\beta}_k + \epsilon_i^{(k)}, \quad i = 1 \cdots n \end{aligned}$$

where $\epsilon_i^{(j)}$s and $\epsilon_i^{(k)}$s are iid respectively. However, they are bivariate normally distributed as

$$\begin{pmatrix} \epsilon_i^{(j)} \\ \epsilon_i^{(k)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, & \rho_{jk}\sigma_k \\ \rho_{jk}\sigma_k, & \sigma_k^2 \end{pmatrix} \right),$$

and the mean value of the $Z^{(j)}$ and $Z^{(k)}$ as below,

$$\begin{aligned} \mu_i^{(j)} &= \boldsymbol{x}_i^{(j)} \boldsymbol{\beta}_j^*, \\ \mu_i^{(k)} &= \boldsymbol{x}_i^{(k)} \boldsymbol{\beta}_k^*. \end{aligned}$$

To solve this bivariate mixed binary and continuous outcomes model, we decompose the joint distribution function of binary and continuous variable to the marginal density function of $P(z_i^{(k)} | X, \boldsymbol{\eta}_{jk})$ and conditional distribution of $P(z_i^{(j)} | z_i^{(k)}, X, \boldsymbol{\eta}_{jk})$, which assumes that the hidden latent variable corresponding to binary outcomes is normally distributed and the marginal density of the continuous outcomes is also normally distributed. Let ϕ and Φ be the pdf and cdf of the standard normal distribution,

1. $Z^{(k)}$ is continuous response

$$P(z_i^{(k)}|X, \boldsymbol{\eta}_{jk}) = \frac{1}{\sigma_k}\phi\bigg(\frac{(z_i^{(k)} - \mu_i^{(k)})}{\sigma_k}\bigg).$$

2. $Z^{(j)}$ is binary response

$$P(z_i^{(j)}|z_i^{(k)}, X, \boldsymbol{\eta}_{jk}) = \Phi\bigg(-\frac{\mu_i^{(j)} + \rho_{jk}\frac{\sigma_j}{\sigma_k}(z_i^{(k)} - \mu_i^{(k)})}{\sigma_j\sqrt{1-\rho_{jk}^2}}\bigg).$$

The likelihood function, the log-likelihood function and the score function for this mixed response model will be

$$\begin{aligned}
L_{jk} &= \prod_{i=1}^{n}\big[1(z_i^{(j)} = 0)\Phi(z_i^{(j)}|z_i^{(k)}) + 1(z_i^{(j)} = 1)(1 - \Phi(z_i^{(j)}|z_i^{(k)}))\big]P(z_i^{(k)}), \\
l_{jk} &= \sum_{i=1}^{n}\log\Big(\big[1(z_i^{(j)} = 0)\Phi(z_i^{(j)}|z_i^{(k)}) + 1(z_i^{(j)} = 1)(1 - \Phi(z_i^{(j)}|z_i^{(k)}))\big]P(z_i^{(k)})\Big), \\
\frac{\partial l_{jk}(\boldsymbol{\eta}_{jk})}{\partial \boldsymbol{\eta}_{jk}} &= l_{jk}^{-1}\int\Big[\frac{\partial \log P(\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}} + \frac{\partial \log P(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}}\Big]P_{j,k}(\boldsymbol{z}^{(j)}, \boldsymbol{z}^{(k)})d\boldsymbol{z}^{(j)} \\
&= \frac{\int\Big[\frac{\partial \log P(\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}} + \frac{\partial \log P(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}}\Big]P_{j,k}(\boldsymbol{z}^{(j)}, \boldsymbol{z}^{(k)})d\boldsymbol{z}_j^*}{\int P_{j,k}(\boldsymbol{z}^{(j)}, \boldsymbol{z}^{(k)})d\boldsymbol{z}^{(j)}} \\
&= \frac{\partial \log\big[1(\boldsymbol{z}^{(j)} = 0)\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)}) + 1(\boldsymbol{z}^{(j)} = 1)(1 - \Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)}))\big]}{\partial \boldsymbol{\eta}_{jk}} \\
&\quad + \frac{\partial \log P(\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}} \\
&= \frac{\partial}{\partial \boldsymbol{\eta}_{jk}}\log P(\boldsymbol{z}^{(k)}) + \Big[\frac{1(\boldsymbol{z}^{(j)} = 0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)} = 1)}{1 - \Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big]\frac{\partial \Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}}.
\end{aligned}$$

To simplify the calculation related formulas, we define 2 new equations as below,

$$
\begin{aligned}
c_k &= \frac{(\boldsymbol{z}^{(k)} - \boldsymbol{\mu}_k)}{\sigma_k}, \\
d_j &= -\frac{(\boldsymbol{\mu}_j + \rho_{jk}\frac{\sigma_j}{\sigma_k}(\boldsymbol{z}^{(k)} - \boldsymbol{\mu}_k))}{\sigma_j\sqrt{1-\rho_{jk}^2}}.
\end{aligned}
$$

Then, the log-likelihood function will be written as

$$
\begin{aligned}
l_{jk} &= \frac{\partial}{\partial \boldsymbol{\eta}_{jk}} \log P(\boldsymbol{z}^{(k)}) + \left[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\right]\frac{\partial \Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}{\partial \boldsymbol{\eta}_{jk}} \\
&= \frac{\partial}{\partial \boldsymbol{\eta}_{jk}} \log\left[\frac{\phi(c_k)}{\sigma_k}\right] + \left[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\right]\frac{\partial \Phi(d_j)}{\partial \boldsymbol{\eta}_{jk}}. \qquad (4.9)
\end{aligned}
$$

The estimation parameters can be estimated by solving the score function above. Julie et al. [51] have also presented the detail solutions for this kind of factorization. We illustrate the detail first and second derivatives needed for Newton-Raphson method in Appendix G.

4.3 Simulation Study

In our simulation study, we assume that a random sample of two binary outcomes and two continuous outcomes are drawn from four variates normal distribution with an arbitrary correlations matrix. Each outcome has its own regression parameters such as coefficients and standard deviation respectively. Our goal is to investigate the performance of the proposed method. Also, we use the marginal regression parameters as the benchmark model and compare the accuracy of the estimated parameters from both methods. In addition, we provide relevant test statistics for testing hypothesis about the regression model coefficients, the standard deviation, and the correlation.

4.3.1 Simulation Data

We generate the data samples with 4 outcomes: 2 binary outcomes (from the hidden latent variables) and 2 continuous outcomes from a multi-variate normal distribution with $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ and covariance matrix of Σ. The covariance matrix was compound symmetric with the variances for 2 binary variables equal to 1. Each response variable has its' own associated covariates $X_i = (x_0, x_1, ..., x_k)$, where $x_0 = 1$ and x_1- x_k be the scalar value, and $i = (i = 1, 2, \ldots, n)$. For each outcome, the mean parameter $\mu_i = X_i^T \boldsymbol{\beta}_i$, where the covariates were simulated from the multivariate normal distribution with variances equal to 1 and the off-diagonal covariance is arbitrary chosen. We took sample sizes n = 300, 1000, and 3000 for different test scenarios, and 3 covariates for each regression model corresponding to each outcome.

We generate different simulation scenarios such as low correlation, low observation, medium and high observation numbers, either common covariates for all the regression or different covariates for different response regressions. These scenarios is through set different numbers of observations, correlation matrix, regression coefficients, and standard deviations of each response variable. Since the proposed method can work with common shared covariate for some response, we also simulate a scenario which all the response variables share a common covariate vectors. Also, the simulated scenarios include shared regression coefficients parameters. Each scenario's results are calculated by the average of minimum of 300 times simulation. The mean value and standard deviation from the Monte Carlo replications will be presented as the simulation results.

4.3.2 Simulation Results

Table 4.1-4.6 give the simulation results for different scenarios. Overall, the proposed composite pairwise likelihood method can estimate the parameters accurately and efficiently. The absolute percentage of the difference between the estimated values and actual values is less then 1% mostly and the standard deviation from simulation is small. Compared with marginal linear regression for the continuous outcomes and logistic regression (probit model) for binary outcomes, the proposed method provides the significantly accurate not only in the fitting coefficient of linear equations but also the correlations between the outcomes and the standard deviations of the residuals. In the simulation results, we use 'Alg' to represent proposed method, Δ_{Marg} and Δ_{Alg} represent the difference between the marginal solution to the true value and the difference between the proposed method to the true value respectively.

Scenario 1: 300 observations, 1000 replications with lower correlation between each other outcomes

Table 4.1: Simulation results of scenario 1, K = 300.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.325	0.121	0.202	0.075	62.7%	0.9%
β_{11}	0.1	0.164	0.127	0.101	0.077	63.9%	1.4%
β_{12}	0.2	0.331	0.121	0.204	0.072	65.3%	1.8%
β_{20}	0.2	0.329	0.125	0.203	0.077	64.3%	1.4%
β_{21}	0.3	0.495	0.128	0.304	0.075	65.0%	1.4%
β_{22}	0.1	0.162	0.125	0.1	0.074	62.5%	0.4%
β_{30}	0.5	0.501	0.06	0.501	0.06	0.1%	0.1%
β_{31}	8	7.999	0.057	8	0.052	0.0%	0.0%
β_{32}	10	10.001	0.057	10.001	0.053	0.0%	0.0%
β_{40}	0.4	0.4	0.057	0.4	0.057	0.0%	0.0%
β_{41}	5	4.997	0.056	4.998	0.053	0.1%	0.0%
β_{42}	8	8	0.058	8.001	0.055	0.0%	0.0%
σ_1	1						
σ_2	1						
σ_3	1	0.996	0.04	0.994	0.04	0.4%	0.6%
σ_4	1	0.994	0.04	0.993	0.04	0.6%	0.7%
ρ_{12}	0.2	0.12	0.057	0.196	0.092	40.1%	2.0%
ρ_{13}	0.3	0.231	0.055	0.296	0.071	23.0%	1.2%
ρ_{14}	0.1	0.077	0.058	0.1	0.074	22.8%	0.3%
ρ_{23}	0.4	0.309	0.051	0.4	0.065	22.8%	0.1%
ρ_{24}	0.2	0.153	0.056	0.198	0.073	23.7%	0.8%
ρ_{34}	0.4	0.396	0.049	0.398	0.049	1.0%	0.4%

Scenario 2: 300 observations, 1000 replications with lower correlation between each other outcomes, all the outcomes share the common covariates.

Table 4.2: Simulation results of scenario 2, K = 300.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.33	0.122	0.205	0.075	65.2%	2.5%
β_{11}	0.1	0.166	0.128	0.102	0.079	65.8%	2.5%
β_{12}	0.2	0.32	0.128	0.198	0.078	60.2%	1.1%
β_{20}	0.1	0.158	0.121	0.098	0.075	58.5%	1.5%
β_{21}	0.2	0.328	0.13	0.203	0.08	63.9%	1.5%
β_{22}	0.1	0.168	0.121	0.104	0.075	67.6%	3.9%
β_{30}	0.5	0.5	0.057	0.5	0.057	0.1%	0.1%
β_{31}	8	7.999	0.056	7.999	0.056	0.0%	0.0%
β_{32}	10	10.001	0.058	10.001	0.058	0.0%	0.0%
β_{40}	4	4	0.059	4	0.059	0.0%	0.0%
β_{41}	3	3	0.059	3	0.059	0.0%	0.0%
β_{42}	1	0.999	0.056	0.999	0.056	0.1%	0.1%
σ_1	1						
σ_2	1						
σ_3	1	0.996	0.041	0.994	0.041	0.4%	0.6%
σ_4	1	0.996	0.04	0.995	0.04	0.4%	0.5%
ρ_{12}	0.2	0.127	0.057	0.201	0.09	36.4%	0.3%
ρ_{13}	0.3	0.238	0.055	0.304	0.07	20.8%	1.3%
ρ_{14}	0.3	0.236	0.055	0.302	0.07	21.2%	0.8%
ρ_{23}	0.3	0.237	0.054	0.301	0.069	21.1%	0.3%
ρ_{24}	0.2	0.158	0.056	0.201	0.071	21.2%	0.3%
ρ_{34}	0.3	0.299	0.054	0.299	0.054	0.3%	0.3%

Scenario 3: 3000 observations, 1000 replications with lower correlation between each other outcomes

Table 4.3: Simulation results of scenario 3, K = 3000.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.324	0.038	0.201	0.023	62.0%	0.7%
β_{11}	0.1	0.164	0.037	0.102	0.022	64.1%	1.6%
β_{12}	0.2	0.324	0.037	0.201	0.022	62.1%	0.4%
β_{20}	0.2	0.325	0.038	0.201	0.023	62.6%	0.5%
β_{21}	0.3	0.485	0.041	0.299	0.024	61.7%	0.4%
β_{22}	0.1	0.161	0.038	0.1	0.023	61.5%	0.4%
β_{30}	0.5	0.5	0.018	0.5	0.018	0.1%	0.1%
β_{31}	8	8	0.018	8	0.016	0.0%	0.0%
β_{32}	10	10	0.018	10	0.016	0.0%	0.0%
β_{40}	0.4	0.4	0.017	0.4	0.017	0.1%	0.1%
β_{41}	5	5	0.018	5	0.017	0.0%	0.0%
β_{42}	8	8	0.019	8	0.018	0.0%	0.0%
σ_1	1						
σ_2	1						
σ_3	1	0.999	0.013	0.999	0.013	0.1%	0.1%
σ_4	1	1	0.013	1	0.013	0.0%	0.0%
ρ_{12}	0.2	0.123	0.018	0.201	0.029	38.4%	0.3%
ρ_{13}	0.3	0.236	0.017	0.301	0.022	21.5%	0.2%
ρ_{14}	0.1	0.079	0.018	0.1	0.023	21.4%	0.3%
ρ_{23}	0.4	0.311	0.016	0.401	0.02	22.2%	0.3%
ρ_{24}	0.2	0.156	0.017	0.201	0.022	21.9%	0.6%
ρ_{34}	0.4	0.4	0.016	0.4	0.016	0.0%	0.1%

Scenario 4: 1000 observations, 1000 replications with lower correlation between each other outcomes

Table 4.4: Simulation results of scenario 4, K = 1000.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.323	0.063	0.201	0.039	61.6%	0.4%
β_{11}	0.1	0.163	0.069	0.101	0.042	63.5%	0.9%
β_{12}	0.2	0.327	0.068	0.203	0.04	63.6%	1.4%
β_{20}	0.2	0.322	0.066	0.199	0.041	60.9%	0.6%
β_{21}	0.3	0.489	0.069	0.301	0.04	62.9%	0.3%
β_{22}	0.1	0.161	0.065	0.099	0.038	60.7%	1.4%
β_{30}	0.5	0.5	0.031	0.5	0.031	0.0%	0.0%
β_{31}	8	8.001	0.032	8.001	0.029	0.0%	0.0%
β_{32}	10	10	0.031	10	0.029	0.0%	0.0%
β_{40}	0.4	0.4	0.032	0.4	0.032	0.1%	0.1%
β_{41}	5	5	0.032	5	0.03	0.0%	0.0%
β_{42}	8	7.999	0.031	7.999	0.029	0.0%	0.0%
σ_1	1						
σ_2	1						
σ_3	1	0.999	0.023	0.999	0.023	0.1%	0.1%
σ_4	1	0.998	0.022	0.998	0.022	0.2%	0.2%
ρ_{12}	0.2	0.123	0.031	0.2	0.05	38.6%	0.0%
ρ_{13}	0.3	0.235	0.03	0.301	0.038	21.5%	0.3%
ρ_{14}	0.1	0.079	0.032	0.101	0.04	21.4%	0.6%
ρ_{23}	0.4	0.31	0.028	0.4	0.036	22.4%	0.0%
ρ_{24}	0.2	0.155	0.032	0.2	0.041	22.5%	0.0%
ρ_{34}	0.4	0.4	0.028	0.4	0.028	0.1%	0.1%

Scenario 5: 300 observations, 1000 replications with medium level correlation between each other outcomes

Table 4.5: Simulation results of scenario 5, K = 300.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.325	0.065	0.202	0.04	62.3%	0.8%
β_{11}	0.1	0.164	0.067	0.101	0.038	64.0%	1.1%
β_{12}	0.2	0.328	0.067	0.204	0.038	64.2%	1.8%
β_{20}	0.2	0.323	0.066	0.199	0.041	61.3%	0.4%
β_{21}	0.3	0.49	0.07	0.301	0.04	63.2%	0.5%
β_{22}	0.1	0.161	0.064	0.099	0.037	61.3%	1.0%
β_{30}	0.5	0.5	0.031	0.5	0.031	0.0%	0.0%
β_{31}	8	8.001	0.032	8.001	0.029	0.0%	0.0%
β_{32}	10	10	0.032	10	0.029	0.0%	0.0%
β_{40}	0.5	0.501	0.031	0.501	0.031	0.2%	0.2%
β_{41}	8	8.001	0.032	8	0.029	0.0%	0.0%
β_{42}	10	9.999	0.031	9.999	0.028	0.0%	0.0%
σ_1	1						
σ_2	1						
σ_3	1	0.999	0.022	0.999	0.022	0.1%	0.1%
σ_4	1	0.998	0.022	0.998	0.022	0.2%	0.2%
ρ_{12}	0.3	0.186	0.032	0.301	0.049	37.8%	0.4%
ρ_{13}	0.3	0.235	0.03	0.301	0.038	21.5%	0.4%
ρ_{14}	0.51	0.4	0.026	0.511	0.033	21.6%	0.1%
ρ_{23}	0.4	0.31	0.028	0.4	0.036	22.4%	0.1%
ρ_{24}	0.4	0.31	0.029	0.4	0.038	22.4%	0.0%
ρ_{34}	0.4	0.4	0.028	0.401	0.028	0.0%	0.2%

Scenario 6: 300 observations, 1000 replications with lower correlation between each other outcomes, the parameter for both continuous response are same and the parameter for both binary response are same too.

Table 4.6: Simulation results of scenario 6, K = 300.

Para	TRUE	Marginal	Std. of Marg	Alg	Std of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.324	0.063	0.201	0.039	61.8%	0.5%
β_{11}	0.1	0.164	0.069	0.101	0.041	63.9%	1.1%
β_{12}	0.2	0.329	0.068	0.204	0.04	64.6%	1.9%
β_{20}	0.2	0.322	0.066	0.2	0.041	60.9%	0.0%
β_{21}	0.1	0.161	0.065	0.099	0.039	60.8%	0.6%
β_{22}	0.2	0.322	0.065	0.199	0.039	60.9%	0.5%
β_{30}	0.5	0.5	0.031	0.5	0.031	0.0%	0.0%
β_{31}	8	8.001	0.032	8.001	0.03	0.0%	0.0%
β_{32}	10	10	0.032	10	0.03	0.0%	0.0%
β_{40}	0.5	0.501	0.032	0.501	0.032	0.1%	0.1%
β_{41}	8	8	0.032	8	0.031	0.0%	0.0%
β_{42}	10	9.999	0.031	9.999	0.03	0.0%	0.0%
σ_1	1						
σ_2	1						
σ_3	1	0.999	0.022	0.999	0.022	0.1%	0.1%
σ_4	1	0.998	0.022	0.998	0.022	0.2%	0.2%
ρ_{12}	0.2	0.124	0.032	0.199	0.051	38.2%	0.4%
ρ_{13}	0.3	0.235	0.03	0.301	0.038	21.6%	0.2%
ρ_{14}	0.3	0.235	0.03	0.3	0.038	21.7%	0.1%
ρ_{23}	0.3	0.235	0.029	0.301	0.038	21.5%	0.2%
ρ_{24}	0.2	0.157	0.031	0.2	0.04	21.5%	0.2%
ρ_{34}	0.3	0.3	0.03	0.301	0.03	0.1%	0.2%

Table 4.7− 4.10 below show the results from the single run of the simulated data. Still, the estimated value from proposed composite likelihood simultaneously is more accurate than the marginal method which predict the regression parameters from each regression respectively. Moreover, the model performance from the proposed model are much better than marginal solution. The statistics we are using are AUROC for binary outcomes, the RMSE, MAE and pseudo R_Squared for continuous outcomes. Figure 3.1−3.4 are the ROC curve for the simulated data to intuitively compare the area under ROC curve. We randomly simulate 2 data set, Data I has 300 observations and Data II has 3000 records to compare the impact from the observation count.

The first simulated data has 300 observations. The estimated parameter is improved by the proposed method than marginal method.

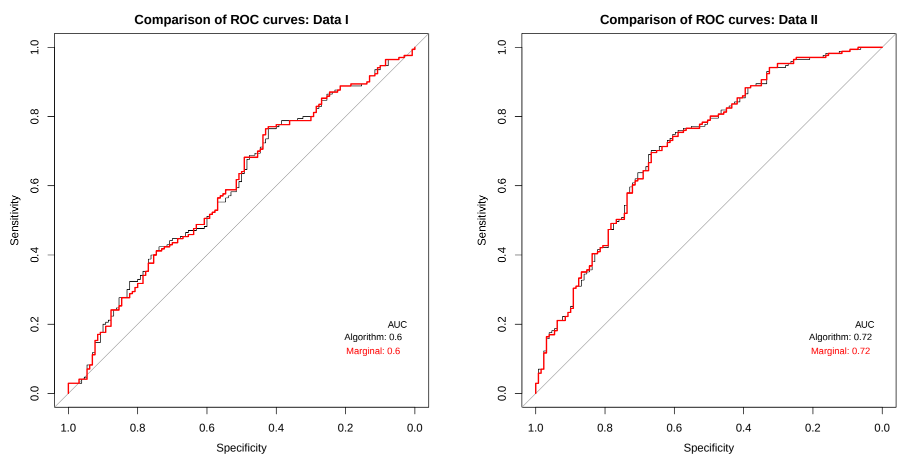
Table 4.7: Simulation results from data I, K = 300.

	TRUE	Marginal	Algorithm	sd.of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.28	0.18	0.07	40.5%	12.1%
β_{11}	0.1	0.08	0.06	0.07	23.4%	42.5%
β_{12}	0.2	0.34	0.21	0.08	71.2%	5.8%
β_{20}	0.2	0.39	0.23	0.08	96.7%	16.7%
β_{21}	0.3	0.82	0.48	0.08	172.3%	60.3%
β_{22}	0.1	0.18	0.12	0.07	75.8%	18.0%
β_{30}	0.5	0.53	0.53	0.06	6.2%	6.0%
β_{31}	8	7.96	7.96	0.05	0.5%	0.5%
β_{32}	10	9.91	9.94	0.05	0.9%	0.6%
β_{40}	0.4	0.43	0.43	0.06	7.7%	7.4%
β_{41}	5	4.96	4.99	0.06	0.7%	0.2%
β_{42}	8	8.01	8.03	0.06	0.1%	0.3%
σ_1	1					
σ_2	1					
σ_3	1	0.98	0.97	0.04	2.4%	2.5%
σ_4	1	1.06	1.06	0.05	5.8%	5.7%
ρ_{12}	0.2	0.14	0.25	0.09	27.6%	23.0%
ρ_{13}	0.3	0.19	0.24	0.08	38.1%	20.8%
ρ_{14}	0.1	0.03	0.05	0.07	65.4%	49.9%
ρ_{23}	0.4	0.35	0.47	0.06	12.6%	17.2%
ρ_{24}	0.2	0.15	0.21	0.07	23.0%	3.4%
ρ_{34}	0.4	0.38	0.39	0.06	4.3%	2.0%

Table 4.8: Model performance from data I.

	AUC_1	AUC_2	R_squared_3	RMSE_3	MAE_3	R_squared_4	RMSE_4	MAE_4
Perf_Marginal	60%	72%	1.00	0.10	0.08	1.00	0.05	0.04
Perf_Algorithm	60%	72%	1.00	0.08	0.06	1.00	0.04	0.04

Figure 4.1: ROC curve from data I.



(a) ROC of M1

(b) ROC of M2

The second simulated data has 3000 observations. The estimated parameter from single run is much improved by the proposed method than marginal method in the perspective of estimation accuracy.

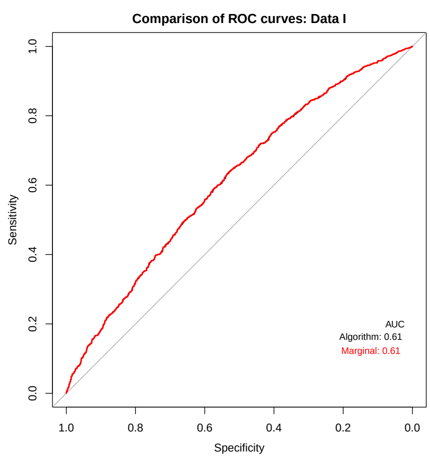
Table 4.9: Simulation results from data II, K = 3000.

	TRUE	Marginal	Algorithm	sd.of Alg	Δ_{Marg}	Δ_{Alg}
β_{10}	0.2	0.31	0.19	0.02	54.3%	4.3%
β_{11}	0.1	0.1	0.06	0.02	2.2%	38.1%
β_{12}	0.2	0.4	0.25	0.02	99.3%	24.6%
β_{20}	0.2	0.33	0.2	0.02	63.2%	0.8%
β_{21}	0.3	0.47	0.29	0.02	57.4%	3.0%
β_{22}	0.1	0.15	0.09	0.02	49.2%	10.7%
β_{30}	0.5	0.49	0.49	0.02	2.4%	2.4%
β_{31}	8	8	8	0.02	0.0%	0.0%
β_{32}	10	9.98	9.98	0.02	0.2%	0.2%
β_{40}	0.4	0.4	0.4	0.02	0.7%	0.6%
β_{41}	5	4.97	4.96	0.02	0.7%	0.7%
β_{42}	8	7.99	7.99	0.02	0.1%	0.1%
σ_1	1					
σ_2	1					
σ_3	1	1	1	0.01	0.3%	0.3%
σ_4	1	1.01	1.01	0.01	1.3%	1.3%
ρ_{12}	0.2	0.13	0.2	0.03	36.6%	2.3%
ρ_{13}	0.3	0.25	0.32	0.02	16.6%	7.3%
ρ_{14}	0.1	0.08	0.1	0.02	22.3%	1.9%
ρ_{23}	0.4	0.32	0.41	0.02	20.7%	2.1%
ρ_{24}	0.2	0.15	0.19	0.02	25.4%	4.8%
ρ_{34}	0.4	0.38	0.38	0.02	5.1%	5.1%

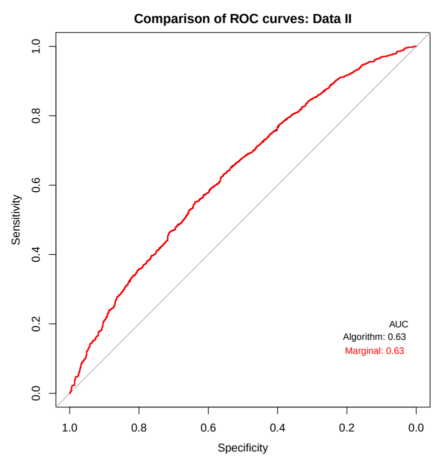
Table 4.10: Model performance from data II.

	AUC_1	AUC_2	R_squared_3	RMSE_3	MAE_3	R_squared_4	RMSE_4	MAE_4
Perf_Marginal	61%	63%	1.00	0.02	0.02	1.00	0.03	0.03
Perf_Algorithm	61%	63%	1.00	0.02	0.02	1.00	0.04	0.03

Figure 4.2: ROC curve from data II.



(a) ROC of M1



(b) ROC of M2

Both single run simulation shows that the performance from the proposed model can be improved by the observation size increase but the performance from marginal regressions keeps same. One reason may come from the proposed method solve the equations simultaneously depending on the correlation between data. The more the data size will make the proposed method catch the correlation structure more accurately. At the same time, we note that the AUROC is almost same for both marginal regression and the proposed model, which because we truncate the simulated latent variables to a binary outcome. When we compare the simulated true latent value to the actual predicted value, we found that the proposed model can provide much more accurate estimates than marginal regression.

4.4 Inference of the Composite Likelihood Estimation

The inference for composite likelihood estimation can be found in many literature [46, 60], we express them here for our special case of the proposed composite likelihood method.

Theorem 4.4.1 *Under the regularity conditions stated in (D1)-(D6) (see Appendix D), the score function from pairwise composite likelihood (4.4) is unbiased. Therefore, the estimator form this score function is also unbiased.*

Proof. Since the equations (4.4) of the score function $S(\boldsymbol{\eta})$ of mixed response model are linear combinations of valid likelihood score function $cl_{j,k}(\boldsymbol{\eta})$ associated with the event probabilities forming the composite log-likelihood function, they immediately satisfy the requirement of being unbiased because the linear combination of the unbiased estimator is still unbiased.

Theorem 4.4.2 *Under the regularity conditions stated in (D1)-(D6) (see Appendix D), assume the true parameter is θ_0, the estimator from maximum the log likelihood function (4.2) $\hat{\theta}_{CMLE}$ is asymptotically normally distributed:*

$$(\hat{\theta}_{CMLE} - \theta_0) \xrightarrow{d} MVN_K(0, G^{-1}).$$

Proof. The asymptotically normality of CMLE is proved in many literature, we just give the proof of our specific case used in this dissertation.

$$Let \quad \hat{\theta}_{CMLE} = \max_{\theta} \; cl(\theta).$$

Define the score function of θ_{CMLE}, the first derivative of the composite log likelihood $\frac{\partial cl(\theta)}{\partial \theta}$, then $E(\frac{\partial cl(\theta)}{\partial \theta}) = 0$ for all θ. Let θ_0 be the true unknown parameter vector value, according to the Central Limit Theorem (CLT), we have

$$\sqrt{N}\frac{\partial cl(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}_K(0, J),$$

where $J = Var[\frac{\partial cl(\theta_0)}{\partial \theta}]$, and $\mathcal{N}_K$ stands for the multivariate normal distribution of K dimensions.

Since $\hat{\theta}_{CMLE}$ is the CML estimator, expanding $\frac{\partial cl(\hat{\theta}_{CMLE})}{\partial \theta}$ around the true value of θ_0 in a first-order Taylor series, we obtain

$$0 = \frac{\partial cl(\hat{\theta}_{CMLE})}{\partial \theta} = \frac{\partial cl(\theta_0)}{\partial \theta} + \frac{\partial^2 cl(\theta_0)}{\partial \theta^2}(\hat{\theta}_{CMLE} - \theta_0).$$

Then we get

$$\hat{\theta}_{CMLE} - \theta_0 = \left(- \frac{\partial^2 cl(\theta_0)}{\partial \theta^2}\right)^{-1} \frac{\partial cl(\theta_0)}{\partial \theta}.$$

From the law of large numbers (LLN), we also have that -$\frac{\partial^2 cl(\theta_0)}{\partial \theta^2} \xrightarrow{d} H$, which is the sample mean of $\frac{\partial^2 cl(\theta_0)}{\partial \theta^2}$ converges to the population mean for the quantity H, $E\Big(- \frac{\partial^2 cl(\theta_0)}{\partial \theta^2}\Big)$.

Now applying Slutsky's theorem, and assuming non-singularity of J and H, we prove the following limiting distribution:

$$(\hat{\theta}_{CMLE} - \theta_0) \xrightarrow{d} \mathcal{N}_K(0, H^{-1}JH^{-1}).$$

Let $G = HJ^{-1}H$, then

$$(\hat{\theta}_{CMLE} - \theta_0) \xrightarrow{d} \mathcal{N}_K(0, G^{-1}).$$

Theorem 4.4.3 *Under the regularity conditions stated in (D1)-(D6) (see Appendix D), With probability tending to 1, as $n \to \infty$, the estimator from maximum the log likelihood function (4.2) θ is consistent to θ_0 satisfies $\sqrt{N}(\theta - \theta_0) = O_p(1)$.*

Proof.

we know that

$$cl(\theta) - cl(\theta_0) = cl'(\theta_0)(\theta - \theta_0) + (\theta - \theta_0)\ cl''(\theta_0^*)(\theta - \theta_0).$$

Let θ lie on a ball of $Cn^{-\frac{1}{2}}\mu$, where μ is a unit vector.

$$\frac{1}{\sqrt{n}} cl'(\theta_0) = O_p(1),$$

$$cl'(\theta_0) = O_p(\sqrt{n}),$$

$$cl'(\theta_0)(\theta - \theta_0) = O_p(\sqrt{n}\frac{1}{\sqrt{n}}) = O_p(1)),$$

$$cl''(\theta_0^*) = O_p(n),$$

$$(\theta - \theta_0)\ cl''(\theta_0^*)(\theta - \theta_0) = O_p(n \times \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{n}}) = O_p(1).$$

Choose C big enough; the second term dominates the first term. Therefore the local maximum is not on the boundary. Accordingly, there exists a local maximum that is inside the ball.

4.5 Conclusions and Future work

we propose a solution to high dimensional mixed response variables regressions using the pairwise composite likelihood method, which includes both binary (or categorical) and continuous variables. We leverage the pairwise composite likelihood method to reduce the dimensions and simplify the problem to three types of bivariate normal distribution problems: two continuous response problems, two discrete response problems and mixed binary (categorical) and continuous response problems. The simulation study shows that the proposed pairwise composite likelihood method can accurately estimate the parameters of coefficients, correlation and standard deviations. One advantage of the proposed method is that it can be easily extended to all kinds of high dimensional response problems: continuous response variables problem, discrete response variables problems, and mixed response variables problems. A disadvantage of this method is that the calibration process takes longer time than the traditional multi-level multivariate regression for the continuous variable problems. The possible future work is that it can be extended to the high dimensional mixed response variables with mixed independent variables containing both fixed effects and random effects.

5 Conclusions and Future Work

In this thesis, we focus on solving the smoothing problems of the non-monotonic increasing empirical PD rates in chapter 2 and solving the smoothing the not well-behaved PD transition matrices in the credit risk area in chapter 3. In chapter 4, we propose using the pairwise composite likelihood method to solve the high dimensional mixed response problem. PD rates is a critical parameter in the capital calculation and risk modeling, and a rational PD rates for any portfolio in the industry-wide follows monotonic increasing property as the associated risk ratings creditworthiness worsen. However, the observed empirical PD rates usually do not hold this property due to many kinds of data issues. In chapter 2, we investigate monotonic smoothing algorithm for the probability of default and transition matrix using the Constrained Maximum a Posterior (CMAP) methodology for smoothing the empirical non-monotonic default rates and transition matrices. We also propose a solution to the high dimensional mixed response variables regression model through the pairwise composite likelihood methodology. We propose a smoothing algorithm to smooth the empirical PD rate. We investigate and compare the performance proposed CMAP method against QMM and CMLE methods on the historical $S\&P$s data and real data. The results show that all three methods of QMM, CMLE and CMAP can work in empirical PD rates monotonic smoothing. However, QMM does not work well as the other two methods in the peak credit period (very low default situation) and the results from QMM has the lowest likelihood score and MSE statistics in the performance measurement. Compared to CMLE method, the proposed CMAP has little worse statistical performance regarding the likelihood score and MSE statistics, and very close regarding the weighted MSE statistics. Nevertheless, the CMLE method is susceptible to the credit environment change, which may come from it's drawback of very sensitive to the sample data. CMAP considers the prior knowledge of the historical PD rate which assumes the PD rates are normally distributed with a mean of the Long Run Default Rates (LRA) and can leverage the external PD rate data which is close to the systematic credit data. Thus CMAP can provide the less sensitive estimators than CMLE. Because the PD rates change as the economic cycle change, the sensitive PD estimation will cause the dramatic change in the financial institution's capital calculation year over year, which can cause the financial institutions' operating difficulty. Accordingly, from the perspective of stability and consistency of the capital calculation, CMAP method is preferable than

CMLE since it provides the less sensitive estimated PD rates with very close statistical performance. Future work with this approach would include applying this method to the actuarial field to smooth the mortality rate curve which has the hump shape or bathtub shape and is not monotonic increasing or decreasing. Also, this approach can be applied to the smooth the yield rates in the fixed income pricing and market risk calculation.

In chapter 3, following the previous chapters empirical PD rate smoothing algorithm, we propose a smoothing algorithm to smooth the empirical PD transition matrics to satisfy the industrys expectation. The well-behaved PD transition matrix should have the properties of monotonic increasing default probabilities, the sum of PD transition probabilities for each rating is 1 and decreasing transition probabilities as the transition state away from the diagonal (stay at the same rating). However, the empirical internal PD transition matrix usually does not hold these properties. We propose a two-step smoothing methodology for the observed transition matrix, which applies the CMAP on the observed default rates first then conduct an optimization process to obtain the well-behaved transition matrix with maximum likelihood score. We apply the proposed algorithm to the $S\&P$s data; the result shows that the proposed algorithm can provide a reliable PD transition matrix.

In chapter 4, we propose a solution to high dimensional mixed response variables regressions using the pairwise composite likelihood method, which includes both binary (or categorical) and continuous variables. We leverage the pairwise composite likelihood method to reduce the dimensions and simplify the problem to three types of bivariate normal distribution problems: two continuous response problems, two discrete response problems and mixed binary (categorical) and continuous response problems. The simulation study shows that the proposed pairwise composite likelihood method can accurately estimate the parameters of coefficients, correlation and standard deviations. One advantage of the proposed method is that it can be easily extended to all kinds of high dimensional response problems: continuous response variables problem, discrete response variables problems, and mixed response variables problems. A disadvantage of this method is that the calibration process takes longer time than the traditional multi-level multivariate regression for the continuous variable problems. The possible future work is that it can be extended to the high dimensional mixed response variables with mixed independent variables containing both fixed effects and random effects.

Bibliography

[1] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finances*, XXIII(4):189209, 1968.

[2] D. Memic. Assessing credit default using logistic regression and multiple discriminant analysis: Empirical evidence from bosnia and herzegovina. *Interdisciplinary Description of Complex Systems*, 13(1):128–153, 2015.

[3] G. Loeffler and P. N. Posch. *Credit Risk Modeling Using Excel and VBA*. West Sussex, England: Wiley Finance, 2007.

[4] R. Cao, J. M. Vilar, and A. Devia. Modelling consumer credit risk via survival analysis. *SORT*, 33(1):3–30, 2009.

[5] S. Figini and D. Fantazzini. Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11:29–54, 2009.

[6] B. H. Yang. Estimating long-run probability of default, asset correlation and portfolio-level probability of default using vasicek models. *journal of risk model validation*, 7(4):3–19, 2013.

[7] M. Kruger. Validation and monitoring of pd models for low default portfolios using proc mcmc. *SAS*, 2015.

[8] T. Bellotti and J. Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28:171–182, 2012.

[9] G. Gupton and R. Stein. *LossCalcTM: Model for predicting loss given default (LGD)*. Moodys KMV, 2002.

[10] J. A. Bastos. Forecasting bank loans loss-given-default. *Journal of Banking and Finance*, 34:2510–2517, 2010.

[11] R. Calabrese. Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied stochastic models in business and industry*, 30(2):99–114, 2012.

[12] O. Vasicek. Moodys kmv corporation working paper: Probability of loss on loan portfolio, 1987.

[13] A. Hamerle, M. Knapp, and N. Wildenauer. *Modeling loss given default: A point in time-approach, in The Basel II Risk Parameters*. Springer, Berlin, 2006.

[14] I. Barakova and H. Parthasarathy. modeling corporate ead at default. *SSRN Working Paper*, 2013.

[15] E. Asarnow and J. Marker. Historical performance of the u.s. corporate loan market: 1988-1993. *Commercial Lending Review*, 10(2):13–32, 1995.

[16] M. Araten and Jr. M. Jacobs. Loan equivalents for revolving credits and advised lines. *The RMA Journal*, pages 34–49, May, 2001.

[17] Jr. M. Jacobs. An empirical study of exposure at default. *Manuscript, Office of the Comptroller of the Currency, U.S. Department of the Treasury*, 2008.

[18] G. Moral. Ead estimates for facilities with explicit limits in engelmann, b. and rauhmeier, r., eds. the basel ii risk parameters: Estimation, validation, and stress testing. *Springer, New York*, pages 197–242, 2006.

[19] Jr. M. Jacobs. An empirical study of exposure at default. *Journal of Advanced Studies in Finance*, 1:31–59, 2010.

[20] M. Qi. *Exposure at Default of Unsecured Credit Cards. Economics Working Paper*. Office of the Comptroller of the Currency, 2009.

[21] R. Taplin, M. T. To and J. Hee. Modeling exposure at default, credit conversion factors and the basel ii accord. *The Journal of Credit Risk*, 3:75–84, 2007.

[22] Moody's. *IFRS 9 Impairment Modeling Regulatory Perspectives and Modeling Approaches*. Moody's Analytics Knowledge Services, 2016.

[23] M. Ermolova and H. Penikas. Pd − lgd correlation study, evidence from the russian corporate bond market. 2015.

[24] H. D. Brunk. On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.*, 29:437–454, 1958.

[25] R.E. Barlow, D.J. Bartholomew, J. M. Bremner, and H.D. Brunk. *Statistical Inference Under Order Restrictions: The theory and Applications of Isotonic Regression*. Wiley, New York, 1972.

[26] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

[27] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[28] R. L. Eubank. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York, 1988.

[29] J. Fan and I. Gijbels. *Local Polynomial Modeling and its Applications.* Chapman and Hall, London, 1996.

[30] S.N. Wood. Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15:1126–1133, 1994.

[31] J. Friedman and R. Tibshirani. The monotone smoothing of scatterplots. *Technometrics*, 26:243–250, 1984.

[32] D. Tasche. The art of probability-of-default curve calibration. *Journal of Credit Risk*, 9:63–103, 2013.

[33] B. H. Yang. Smoothing algorithms by constrained maximum likelihood. 2017.

[34] A. J. Richard, I. Miller, and J. Freund. *Miller and Freund's Probability and Statistics for Engineers.* Chapman and Hall, London, 2015.

[35] P. Miu and B. Ozdemir. Estimating and validating long-run probability of default with respect to basel ii requirements. *Journal of Risk Model Validation*, 2(2), 2008.

[36] A. O'Hagan. Kendall's advanced theory of statistics. *Bayes Inference*, 2B, 1994.

[37] Jerome H. F. and L. Bremiman. Preidcting multivariate response in multiple linear regression. *Journal of the Royal Statistics*, 59(1):3–54, 1997.

[38] W. Green. *Econometric Analysis, 7th ed.* Prentice Hall, 2012.

[39] A. Teixeira-Pinto and S. T. Normand. Correlated bivariate continuous and binary outcomes: Issues and applications. *Stat Med*, 28(13):1753–1773, 2009.

[40] L. Olkin and R. F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat*, 32:448–456, 1961.

[41] Y. Yang, J. Kang, K. Mao, and J. Zhang. Regression models for mixed poisson and continuous longitudinal data. *Stat. Med*, 26:3782–3800, 1961.

[42] D. R. Cox and N. Wermuth. Response models for mixed binary and quantitative variable. *Biometrika*, 79:441–461, 1992.

[43] D. Hedeker and R. D. Gibbons. A random effects ordinal regression for multilevel analysis. *Biometrics*, 50:933–944, 1994.

[44] P. Catalano and L. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Stat. Asso*, 50(3):1078–1095, 1992.

[45] G. M. Fitzmaurice and N. M Laird. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90:845–852, 1995.

[46] D. B. Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society*, 62(2):355–366, 2000.

[47] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:220–239, 1988.

[48] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:542, 2011.

[49] V. Godambe. An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist*, 31:1208–1211, 1960.

[50] A. Skrondal and S. Rabe-Hesketh. *Latent Variable Modelling: A Survey*, volume 34. Blackwell, Oxford, 2007.

[51] J. S. Najita, Y. Li, and P. J. Catalano. A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *J R Stat Soc Ser C Appl Stat*, 58:555–573, 2009.

[52] P. J. Catalano. Bivariate modeling of clustered continuous and ordered categorica outcomes. *Statics in Medicine*, 16:883–900, 1997.

[53] M. S. Azadbakhsh, X. Gao, and H. Jankowski. Multiple comparisons using

composite likelihood in clustered data. *International Journal of Biostatistics*, 2016.

[54] S & P. Default, transition, and recovery: 2009 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2010.

[55] S & P. Default, transition, and recovery: 2010 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2011.

[56] S & P. Default, transition, and recovery: 2011 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2012.

[57] S & P. Default, transition, and recovery: 2012 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2013.

[58] S & P. Default, transition, and recovery: 2013 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2014.

[59] S & P. Default, transition, and recovery: 2015 annual global corporate default study and rating transitions. *Report, Standard* & *Poor's*, March 2016.

[60] Chandra R. Bhat. The composite marginal likelihood (cml) inference approach with applications to discrete and mixed dependent variable models.

[61] A. M. Berd. Dynamic estimation of credit rating transition probabilities. *The Journal of Investment Strategies*, Feb 2, 2005.

[62] W DuMouchel and M Schonlau. A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities. *KDD*, 189-193, 1998.

[63] K. E. Atkinson. *An Introduction to Numerical Analysis.* John Wiley and Sons Inc, 1989.

[64] E. Bahrami Samani, M. Ganjali, and S. Eftekhari Sankhy. A latent variable model for mixed continuous and ordinal responses with nonignorable missing responses: Assessing the local influence via covariance structure. *The Indian Journal of Statistics*, 72:38–57, 2010.

A Notations

To formally study the problem mentioned in the introduction, we introduce notations used in this dissertation here. In this dissertation, the following are some notations used in the following Chapters. A^T denotes the transpose of a matrix A. $\boldsymbol{v}^T$, v_j and $\|\boldsymbol{v}\|$ denote the transpose, j^{th} component and the L_2 norm of a vector $\boldsymbol{v}$, respectively. Let $\boldsymbol{v} = (v_1, v_2, \ldots, v_p)^T$ be a $p \times 1$ vector, $A = (a_{ij}) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)$ be a $q \times p$ matrix where a_{ij}'s are the elements of A and $\boldsymbol{a}_j$'s are the column vectors of A, and $\mathcal{B} = \{i_1, i_2, \ldots, i_k\}$ be an index set with $1 \leq i_1 \leq \ldots \leq i_k \leq p$. Let $|\mathcal{B}|$ denote the size of $\mathcal{B}$ which is equal to k. Denote $\boldsymbol{v}_{[\mathcal{B}]} = (v_{i_1}, \ldots, v_{i_k})^T$, $A_{[\mathcal{B}]} = (\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_k})$. Let $I_S(t)$ be the indicator function such that $I_S(t) = 1$ if $t \in S$ and $I_S(t) = 0$ otherwise, $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise. The latent variable behind observed variable M is represnted by M^*. Denote the sign function as sgn. Denote the logarithm function as log and exponential function as exp or e respectively. Denote the inverse function of $f(x)$ as $f^{-1}(x)$. Let $f'(x)$ and $f''(x)$ denote the first and second order derivatives of a univariate function, $f(x)$ with respect to the scalar x, and let

$\partial f(\boldsymbol{v})/\partial \boldsymbol{v}$ and $\partial^2 f(\boldsymbol{v})/(\partial \boldsymbol{v} \partial \boldsymbol{v}^T)$ denote the first and second order derivative with respect to the vector $\boldsymbol{v}$. $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function (pdf) and cumulative distribution function (cdf) of a standard normal distribution respectively. N_2 and $\Phi_2(\cdot)$ denote the probability density function (pdf) and cumulative distribution function (cdf) of a standard bivariate normal distribution respectively

B Results from Beta Priori for $S\&P$'s Data

Table B.1: CMAP Results from Beta Prior for $S\&P$ data.

Rating	2009	2010	2011	2012	2013
AAA	N/A	N/A	N/A	N/A	N/A
AA+	N/A	N/A	N/A	N/A	N/A
AA	0.0001	0.0001	0.0001	0.0001	0.0001
AA-	0.0001	0.0001	0.0001	0.0001	0.0001
A+	0.0019	0.0002	0.0002	0.0002	0.0002
A	0.0024	0.0002	0.0002	0.0002	0.0002
A-	0.0002	0.0002	0.0002	0.0002	0.0002
BBB+	0.0034	0.0004	0.0003	0.0003	0.0003
BBB	0.0020	0.0006	0.0006	0.0006	0.0005
BBB-	0.0089	0.0007	0.0021	0.0006	0.0006
BB+	0.0010	0.0073	0.0011	0.0009	0.0009
BB	0.0094	0.0048	0.0020	0.0019	0.0018
BB-	0.0093	0.0060	0.0013	0.0079	0.0013
B+	0.0491	0.0025	0.0055	0.0071	0.0019
B	0.0894	0.0098	0.0138	0.0155	0.0020
B-	0.1445	0.0233	0.0405	0.0348	0.0114
CCC-C	0.3226	0.1845	0.1444	0.2118	0.0281

Clearly, the posteriori estimator from beta distribution is not monotonic in each year, thus, it cannot be used as the smoothing solution directly in this specific area.

C Default Count of Real Data

Table C.1: Real data: the total count at the beginning of the year.

Rating	Total Count in the Portfolio by Years											
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	22	20	18	16	16	15	14	13	12	11	10	9
Aa1	18	17	16	15	14	13	12	11	10	9	8	7
Aa2	0	0	0	0	0	0	0	0	0	0	0	0
Aa3	3	3	3	3	2	2	2	3	3	3	2	1
A1	14	13	13	12	11	10	9	8	7	7	7	6
A2	7	5	4	3	2	2	2	2	1	0	0	0
A3	71	66	63	60	57	53	56	45	41	38	39	36
Baa1	329	310	288	269	251	234	218	203	190	177	165	151
Baa2	399	377	350	324	301	280	259	235	212	193	173	154
Baa3	1091	983	882	786	695	608	519	449	383	346	309	278
Ba1	876	808	742	684	626	573	523	476	435	361	301	249
Ba2	907	817	732	636	559	474	400	355	316	296	269	240
Ba3	532	489	456	428	398	370	347	317	281	253	227	204
B1	465	429	398	365	333	302	274	251	230	213	203	194
B2	378	352	317	283	251	230	207	188	172	156	142	127
B3	56	52	47	42	38	32	29	25	19	15	13	13
Caa1	51	47	28	40	36	31	44	25	24	22	20	18
Caa2	1	1	1	1	16	9	5	3	1	1	1	1
Caa3	0	0	0	0	1	1	1	1	0	0	0	0
Ca	8	7	6	5	21	16	12	10	4	3	2	0

Table C.2: Real data: the default count during the observation year.

Rating	Total Default in the Portfolio by Years											
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Aaa	0	0	0	0	0	0	0	0	0	0	0	0
Aa1	0	0	0	0	0	0	0	0	0	0	0	0
Aa2	0	0	0	0	0	0	0	0	0	0	0	0
Aa3	0	0	0	0	0	0	0	0	0	0	0	0
A1	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	0	0	0	0	0	0	0	0	0	0
Baa1	0	0	0	0	0	0	0	0	0	0	1	0
Baa2	0	0	0	0	0	0	0	0	0	0	0	0
Baa3	0	0	0	1	1	1	1	1	1	1	1	1
Ba1	1	1	2	1	1	1	0	0	0	0	0	0
Ba2	1	1	1	1	2	2	1	1	0	0	0	0
Ba3	1	1	0	2	2	2	5	3	2	3	1	1
B1	3	2	2	2	1	0	0	0	0	0	0	0
B2	3	3	3	2	3	4	3	2	2	1	1	0
B3	0	0	0	0	1	0	0	0	0	0	0	0
Caa1	0	0	0	0	0	0	0	0	0	0	0	0
Caa2	0	0	0	0	1	1	0	0	0	0	0	0
Caa3	0	0	0	0	0	0	1	0	0	0	0	0
Ca	0	0	0	0	2	2	1	1	0	0	0	0

Comparison of Realized and Estimated Default Rates

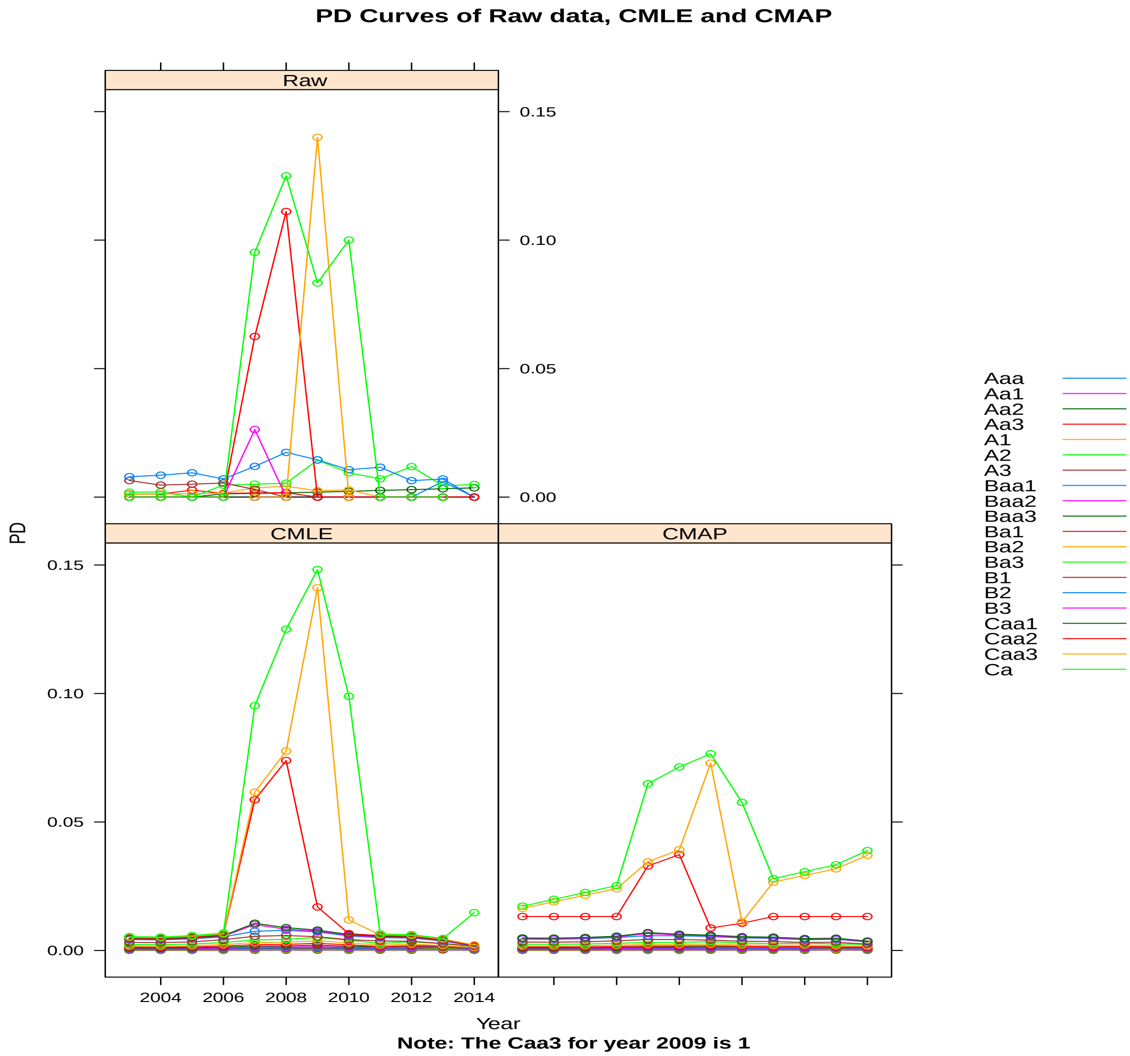


Figure C.1: Please note that the realized default rate for Caa3 at year 2000 is 1, we modify it to .2 for showing at this image.

Comparison of Realized and Estimated Default Rates by Rating Grades

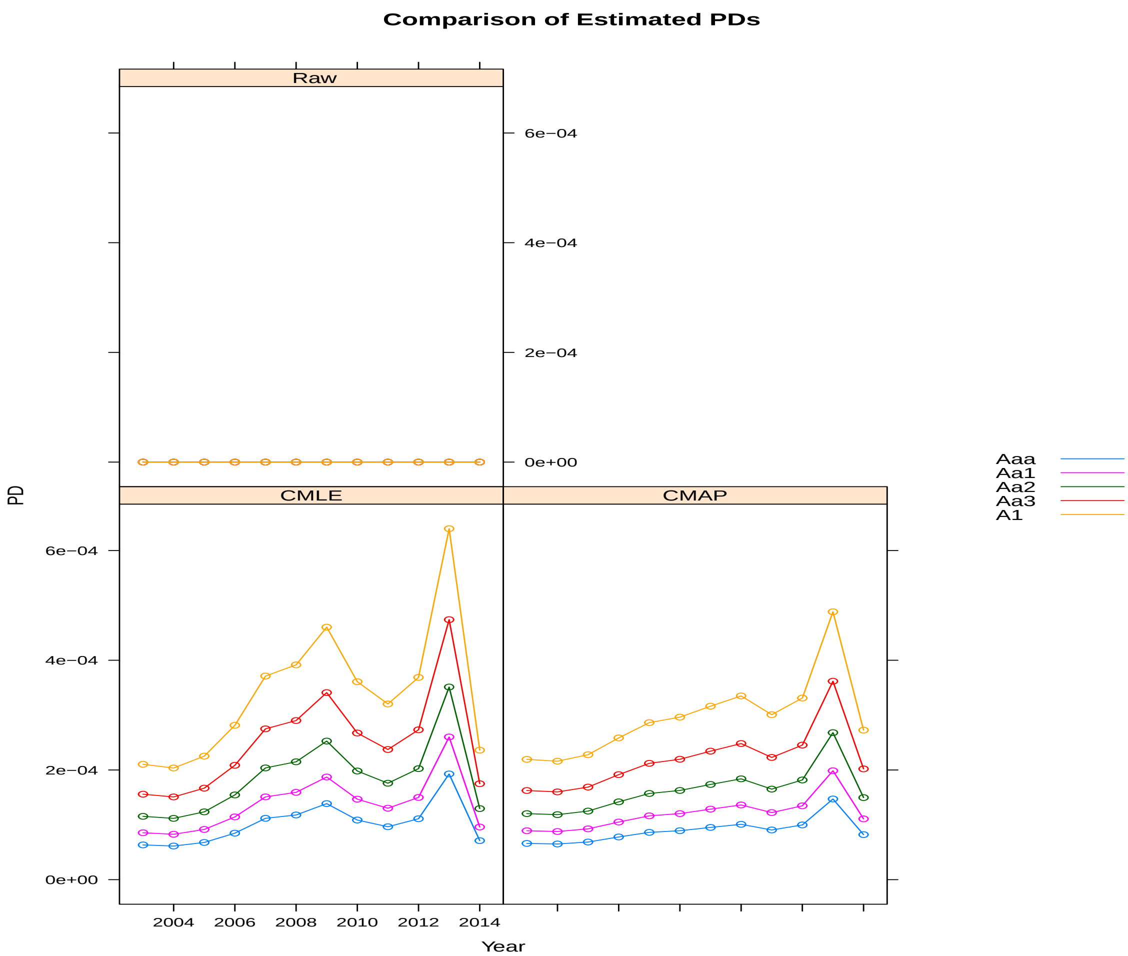


Figure C.2: For the best risk rating grades, both CMLE and CMAP give the estimated PDs although the realized default rates are 0 for these rating grades. And CMAP gives higher estimation for these ratings in the worse economic environment (year 2008).

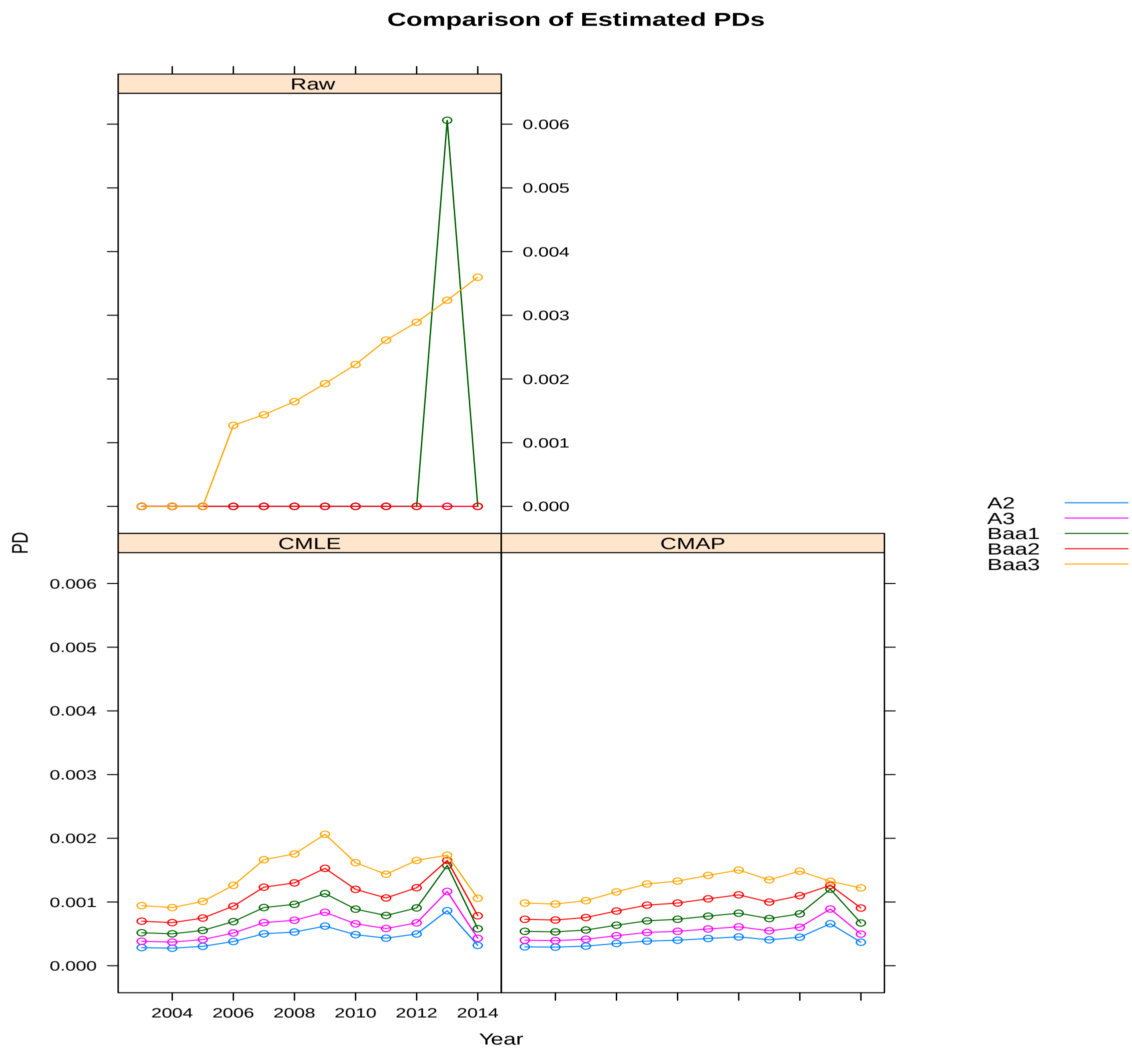


Figure C.3: For the rating just above investment grade, the estimated PDs from CMAP is close to the estimated PDs from CMLE, but higher in the worse economic environment (year 2008).

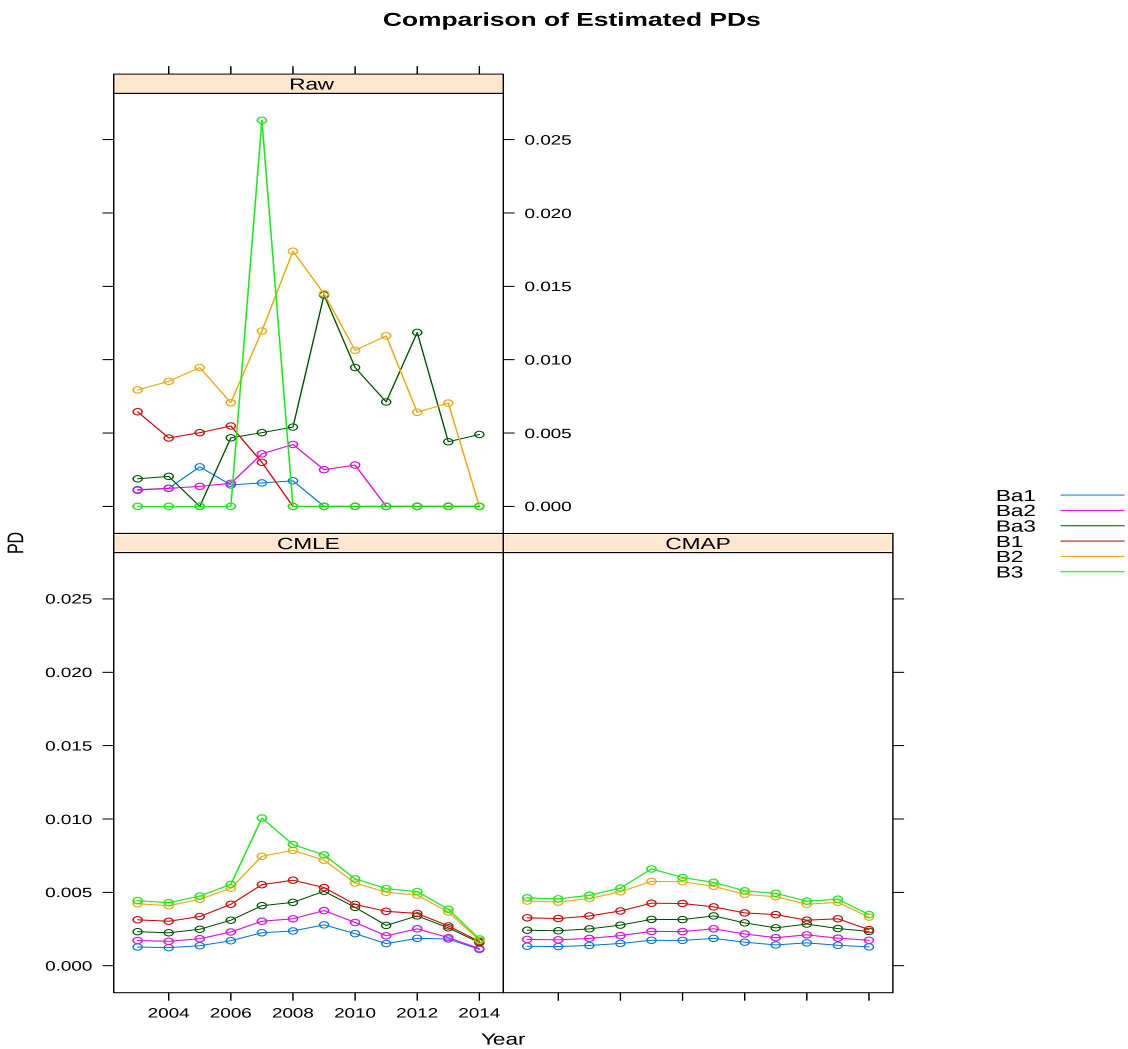


Figure C.4: For the speculative rating grades, compared to CMLE method, the CMAP method generate the estimated PDs is higher in the better economic environment and lower in the worse economic environment (year 2008).

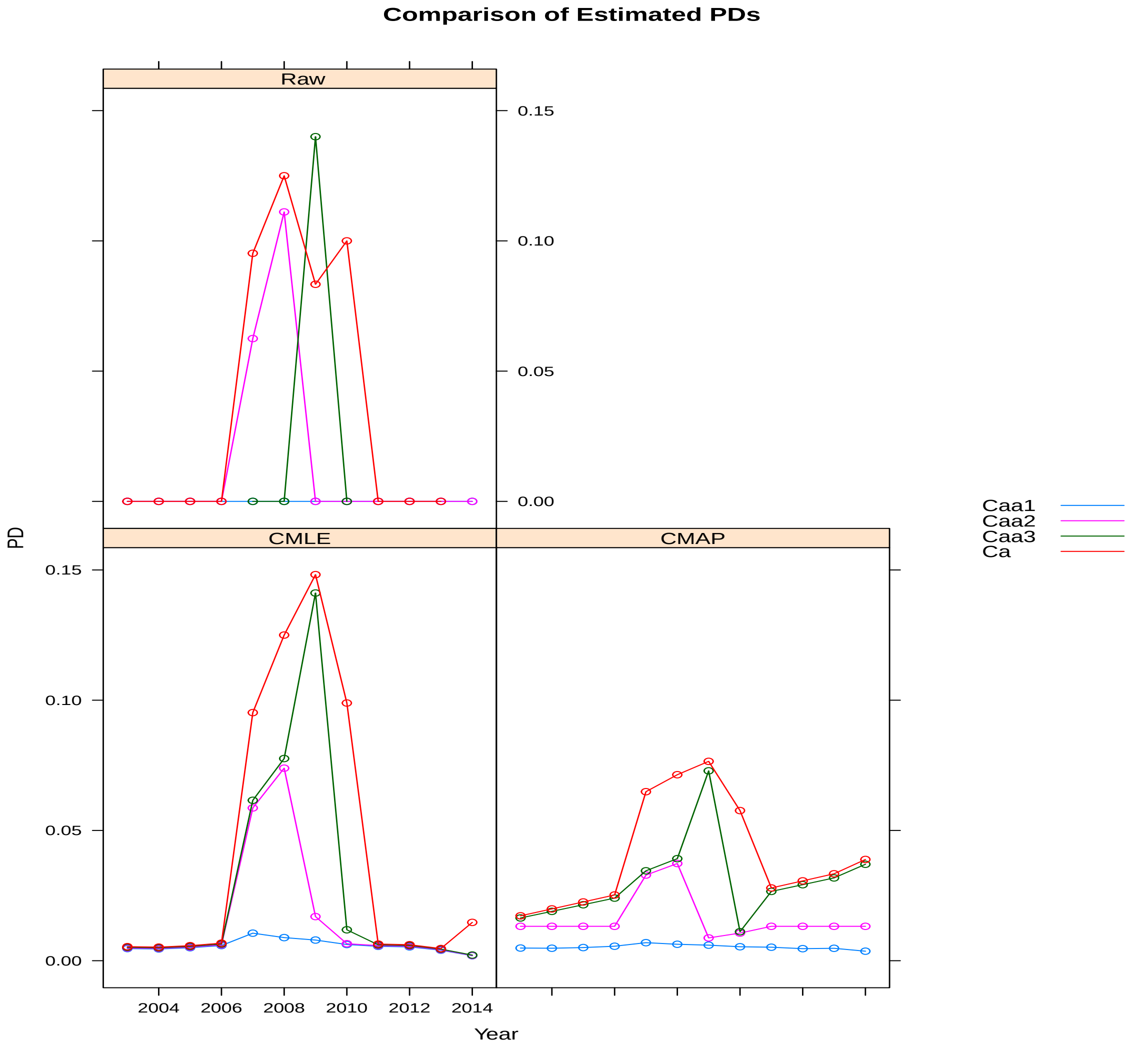


Figure C.5: For the worst risk rating grades, the CMAP's estimation is higher is higher in the better economic environment and lower in the worse economic environment (year 2008).

D Regularity Conditions in Composite Likelihood

(D1). The marginal density function of x , $f(x;\theta)$ is distinct for different values of x, i.e. if $\theta_1 \neq \theta_2$ then $P(f(x;\theta_1) \neq f(x;\theta_2)) > 0$, for all observations of X.

(D2). The marginal densities of x have common support for all θ.

(D3). The true value θ_0 is an interior point of Ω, the space of possible values of the parameter θ.

(D4). Let I and ∇I denote the index and partial derivative operator, respectively, as in the standard multi-index notation from multivariable calculus. The marginal density $\log f$ is three times continuously differentiable in a closed ball around θ_0. Moreover, there exists a constant C and an integrable function $L(x)$ such that

$$|\nabla I \nabla_i \theta \log f(x;\theta)| \leq L(x),$$

for all $||\theta - \theta_0|| \leq C$, all $|I| = 2$, and any $i = 1, ..., N$. Here, $||\cdot||_2$ denotes the Euclidean norm.

(D5). $J(\theta_0)$ is well-defined (i.e. exists and is finite) and invertible.

(D6). $H(\theta_0)$ is well-defined (i.e. exists and is finite) and (strictly) positive definite

(Also, this regularity conditions can be seen in [54]).

E Derivatives for Both Continuous Response Variables

$$
\begin{aligned}
\frac{\partial l_{34}}{\partial \beta_3} &= \frac{1}{1-\rho_{34}^2}\Big[\frac{\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)}}{\sigma_3^2} - \frac{\rho_{34}(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]' X_3 \\
\frac{\partial l_{34}}{\partial \beta_4} &= \frac{1}{1-\rho_{34}^2}\Big[\frac{\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)}}{\sigma_4^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3\sigma_4}\Big]' X_4 \\
\frac{\partial l_{34}}{\partial \sigma_3} &= -\frac{1}{\sigma_3}\Big(1 - \frac{1}{1-\rho_{34}^2}\Big[\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]\Big) \\
\frac{\partial l_{34}}{\partial \sigma_4} &= -\frac{1}{\sigma_4}\Big(1 - \frac{1}{1-\rho_{34}^2}\Big[\frac{(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_4^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]\Big) \\
\frac{\partial l_{34}}{\partial \rho_{34}} &= \frac{\rho_{34}}{1-\rho_{34}^2}\Big[1 - S + \frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\rho_{34}\sigma_3\sigma_4}\Big] \\
\frac{\partial^2 l_{34}}{\partial \beta_3 \partial \beta_3{}'} &= -\frac{1}{(1-\rho_{34}^2)\sigma_3^2} X_3 X_3' \\
\frac{\partial^2 l_{34}}{\partial \beta_4 \partial \beta_4{}'} &= -\frac{1}{(1-\rho_{34}^2)\sigma_4^2} X_4 X_4' \\
\frac{\partial^2 l_{34}}{\partial \sigma_3^2} &= \frac{1}{\sigma_3^2}\Bigg[\Big(1 - \frac{1}{1-\rho_{34}^2}\Big[\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]\Big) \\
&\quad + \frac{1}{1-\rho_{34}^2}\Big(-\frac{2(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^2} + \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big)\Bigg]
\end{aligned}
$$

$$\frac{\partial^2 l_{34}}{\partial\sigma_4^2} = \frac{1}{\sigma_4^2}\Bigg[\Big(1 - \frac{1}{1-\rho_{34}^2}\Big[\frac{(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_4^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]\Big)$$

$$+ \frac{1}{1-\rho_{34}^2}\Big(-\frac{2(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_4^2} + \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big)\Bigg]$$

$$\frac{\partial^2 l_{34}}{\partial\rho_{34}^2} = \Big(\frac{1}{1-\rho_{34}^2} + \frac{2\rho_{34}^2}{(1-\rho_{34}^2)^2}\Big)\Big[1 - S + \frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\rho_{34}\sigma_3\sigma_4}\Big]$$

$$- \frac{\rho_{34}}{1-\rho_{34}^2}\Big[\frac{2}{1-\rho_{34}^2}\Big(\rho_{34}S - \frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big)$$

$$+ \frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\rho_{34}^2\sigma_3\sigma_4}\Big]$$

$$\frac{\partial^2 l_{34}}{\partial\beta_3\partial\beta_4} = \frac{1}{(1-\rho_{34}^2)}\frac{\rho_{34}}{\sigma_3\sigma_4}X_3X_4'$$

$$\frac{\partial^2 l_{34}}{\partial\beta_3\partial\sigma_3} = -\frac{1}{1-\rho_{34}^2}\Big[\frac{2(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^3} - \frac{\rho_{34}(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3^2\sigma_4}\Big]'X_3$$

$$\frac{\partial^2 l_{34}}{\partial\beta_3\partial\sigma_4} = \frac{1}{1-\rho_{34}^2}\Big[\frac{\rho_{34}(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4^2}\Big]'X_3$$

$$\frac{\partial^2 l_{34}}{\partial\beta_3\partial\rho_{34}} = \frac{2\rho_{34}}{(1-\rho_{34}^2)^2}\Big[\frac{\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)}}{\sigma_3^2} - \frac{\rho_{34}(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]'X_3$$

$$- \frac{1}{(1-\rho_{34}^2)}\frac{(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})'}{\sigma_3\sigma_4}X_3$$

$$\frac{\partial^2 l_{34}}{\partial\beta_4\partial\sigma_3} = \frac{1}{1-\rho_{34}^2}\Big[\frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^2\sigma_4}\Big]'X_4$$

$$\frac{\partial^2 l_{34}}{\partial\beta_4\partial\sigma_4} = -\frac{1}{1-\rho_{34}^2}\Big[\frac{2(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_4^3} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3\sigma_4^2}\Big]'X_4$$

$$\frac{\partial^2 l_{34}}{\partial\beta_4\partial\rho_{34}} = \frac{2\rho_{34}}{(1-\rho_{34}^2)^2}\Big[\frac{\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)}}{\sigma_4^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3\sigma_4}\Big]'X_4$$

$$- \frac{1}{(1-\rho_{34}^2)}\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'}{\sigma_3\sigma_4}X_4$$

$$\frac{\partial^2 l_{34}}{\partial\sigma_3\partial\sigma_4} = \frac{1}{(1-\rho_{34}^2)\sigma_3}\Big[\frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4^2}\Big]$$

$$\frac{\partial^2 l_{34}}{\partial\sigma_3\partial\rho_{34}} = \frac{2\rho_{34}}{(1-\rho_{34}^2)^2\sigma_3}\Big[\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})}{\sigma_3^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]$$

$$- \frac{1}{(1-\rho_{34}^2)\sigma_3}\Big[\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]$$

$$\frac{\partial^2 l_{34}}{\partial\sigma_4\partial\rho_{34}} = \frac{2\rho_{34}}{(1-\rho_{34}^2)^2\sigma_4}\Big[\frac{(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_4^2} - \frac{\rho_{34}(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]$$

$$- \frac{1}{(1-\rho_{34}^2)\sigma_4}\Big[\frac{(\boldsymbol{z}^{(3)}-\boldsymbol{\mu}^{(3)})'(\boldsymbol{z}^{(4)}-\boldsymbol{\mu}^{(4)})}{\sigma_3\sigma_4}\Big]$$

F Derivatives for Both Binary Response Variables

$$
\begin{aligned}
\mu^{(1)} &= X^{(1)T} * \beta_1^* \\
\mu^{(2)} &= X^{(2)T} * \beta_2^* \\
s_1 &= 2\boldsymbol{z}^{(1)} - 1 \\
s_2 &= 2\boldsymbol{z}^{(2)} - 1 \\
\tilde{\rho_{12}} &= s_1 s_2 \rho_{12} \\
q_1 &= s_1 \mu^{(1)} \\
q_2 &= s_2 \mu^{(2)} \\
\tilde{\Sigma_{12}} &= \begin{pmatrix} 1 & \tilde{\rho_{12}} \\ \tilde{\rho_{12}} & 1 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
S_i &= \frac{q_i}{\sigma_i} \\
\delta_i &= \frac{1}{\sqrt{1-\tilde{\rho_{12}}^2}} \\
Q_1 &= \delta_i(S_2 - \tilde{\rho_{12}} S_1) \\
Q_2 &= \delta_i(S_1 - \tilde{\rho_{12}} S_2) \\
W_1 &= \phi_1(S_1)\Phi_1(Q_1) \\
W_2 &= \phi_1(S_2)\Phi_1(Q_2)
\end{aligned}
$$

And we can derive that

$$\delta_i \phi(S_1)\phi(Q_1) = \delta_i \phi(S_2)\phi(Q_2) = \phi_2(S_1, S_2, \tilde{\rho_{12}}).$$

To simplify our derivatives, we use the following notations,

$$\phi_2^b = \phi_2(S_1, S_2, \tilde{\rho_{12}}),$$

$$\Phi_2^b = \Phi_2(S_1, S_2, \tilde{\rho_{12}}).$$

Then, the score function will be

$$\begin{aligned}\frac{\partial l_{1,2}(\eta)}{\partial \eta} &= \frac{1}{\Phi_2(\frac{s_1 X_1^T \beta_1^*}{\sigma_1}, \frac{s_2 X_2^T \beta_2^*}{\sigma_2}, \tilde{\rho_{12}})} \frac{\partial \Phi_2(\frac{s_1 X_1^T \beta_1^*}{\sigma_1}, \frac{s_2 X_2^T \beta_2^*}{\sigma_2}, \tilde{\rho_{12}})}{\partial \eta} \\ &= \frac{1}{\Phi_2(\frac{q_1^*}{\sigma_1}, \frac{q_2^*}{\sigma_2}, \tilde{\rho_{12}})} \frac{\partial \Phi_2(\frac{q_1^*}{\sigma_1}, \frac{q_2^*}{\sigma_2}, \tilde{\rho_{12}})}{\partial \eta} \\ &= \frac{1}{\Phi_2(S_1, S_2, \tilde{\rho_{12}})} \frac{\partial \Phi_2(S_1, S_2, \tilde{\rho_{12}})}{\partial \eta} \\ &= \frac{1}{\Phi_2^b} \frac{\partial \Phi_2^b}{\partial \eta}.\end{aligned}$$

And the detail first and second derivatives of the score function are

$$\begin{aligned}\frac{\partial l_{12}}{\partial \beta_1^*} &= \frac{s_1}{\sigma_1} \frac{W_1}{\Phi_2^b} \cdot X_1, \\ \frac{\partial l_{12}}{\partial \beta_2^*} &= \frac{s_2}{\sigma_2} \frac{W_2}{\Phi_2^b} \cdot X_2, \\ \frac{\partial l_{12}}{\partial \sigma_1} &= -\frac{\partial l_{12}}{\partial \beta_1^*} \frac{\beta_1}{\sigma_1}, \\ \frac{\partial l_{12}}{\partial \sigma_2} &= -\frac{\partial l_{12}}{\partial \beta_2^*} \frac{\beta_2}{\sigma_2}, \\ \frac{\partial l_{12}}{\partial \rho_{12}} &= \frac{s_1 s_2 \phi_2^b}{\Phi_2^b}, \\ \frac{\partial^2 l_{12}}{\partial \beta_1^{*2}} &= -\frac{X_1 X_1'}{\sigma_1^2} \Big[\frac{S_1 W_1}{\Phi_2^b} + \frac{\tilde{\rho_{12}} \phi_2^b}{\Phi_2^b} + \frac{W_1^2}{\Phi_2^2} \Big], \\ \frac{\partial^2 l_{12}}{\partial \beta_2^{*2}} &= -\frac{X_2 X_2'}{\sigma_2^2} \Big[\frac{S_2 W_2}{\Phi_2^b} + \frac{\tilde{\rho_{12}} \phi_2^b}{\Phi_2^b} + \frac{W_2^2}{\Phi_2^2} \Big], \\ \frac{\partial^2 l_{12}}{\partial \beta_1^* \partial \beta_2^{*\prime}} &= \frac{s_1 s_2 X_1 X_2'}{\sigma_1 \sigma_2} \Big[\frac{\phi_2^b}{\Phi_2^b} - \frac{W_1 W_2}{\Phi_2^{b2}} \Big],\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l_{12}}{\partial \beta_1^* \partial \rho_{12}} &= \frac{s_2 X_1}{\sigma_1} \frac{\phi_2^b}{\Phi_2^b}\Big[\tilde{\rho_{12}}\delta_i Q_1 - S_1 - \frac{W_1}{\Phi_2^b}\Big], \\
\frac{\partial^2 l_{12}}{\partial \beta_2^* \partial \rho_{12}} &= \frac{s_1 X_2}{\sigma_2} \frac{\phi_2^b}{\Phi_2^b}\Big[\tilde{\rho_{12}}\delta_i Q_2 - S_2 - \frac{W_2}{\Phi_2^b}\Big], \\
\frac{\partial^2 l_{12}}{\partial \beta_1^* \partial \sigma_1} &= \frac{s_1}{\sigma_1}\Big[- \frac{W_1}{\sigma_1 \Phi_2^b} + \frac{S_1 W_1 q_1 + \phi_2^b \tilde{\rho_{12}} q_1}{\Phi_2^b \sigma_1^2} + \frac{W_1^2 q_1}{\Phi_2^{b2}\sigma_1^2}\Big] X_1, \\
\frac{\partial^2 l_{12}}{\partial \beta_2^* \partial \sigma_2} &= \frac{s_2}{\sigma_2}\Big[- \frac{W_2}{\sigma_2 \Phi_2^b} + \frac{S_2 W_2 q_2 + \phi_2^b \tilde{\rho_{12}} q_2}{\Phi_2^b \sigma_2^2} + \frac{W_2^2 q_2}{\Phi_2^{b2}\sigma_2^2}\Big] X_2, \\
\frac{\partial^2 l_{12}}{\partial \beta_1^* \partial \sigma_2} &= \frac{s_1 q_2}{\sigma_1 \sigma_2^2}\Big[- \frac{\phi_2^b}{\Phi_2^b} + \frac{W_1 W_2}{\Phi_2^{b2}}\Big] X_1, \\
\frac{\partial^2 l_{12}}{\partial \beta_2^* \partial \sigma_1} &= \frac{s_2 q_1}{\sigma_2 \sigma_1^2}\Big[- \frac{\phi_2^b}{\Phi_2^b} + \frac{W_1 W_2}{\Phi_2^{b2}}\Big] X_2, \\
\frac{\partial^2 l_{12}}{\partial {\sigma_1}^2} &= \frac{2 q_1}{\sigma_1^3}\frac{W_1}{\Phi_2^b} - \frac{q_1}{\sigma_1^2}\Big[\frac{S_1 W_1 q_1 + \phi_2^b \tilde{\rho_{12}} q_1}{\Phi_2^b \sigma_1^2} + \frac{W_1^2 q_1}{\Phi_2^{b2}\sigma_1^2}\Big], \\
\frac{\partial^2 l_{12}}{\partial {\sigma_2}^2} &= \frac{2 q_2}{\sigma_2^3}\frac{W_2}{\Phi_2^b} - \frac{q_2}{\sigma_2^2}\Big[\frac{S_2 W_2 q_2 + \phi_2^b \tilde{\rho_{12}} q_2}{\Phi_2^b \sigma_2^2} + \frac{W_2^2 q_2}{\Phi_2^{b2}\sigma_2^2}\Big], \\
\frac{\partial^2 l_{12}}{\partial \rho_{12} \partial \sigma_1} &= -\frac{s_2 X_1 \beta_1}{\sigma_1^2}\frac{\phi_2^b}{\Phi_2^b}\Big(\delta_i \tilde{\rho_{12}} Q_1 - S_1 - \frac{W_1}{\Phi_2^b}\Big), \\
\frac{\partial^2 l_{12}}{\partial \sigma_1 \partial \sigma_2} &= -\frac{q_1 q_2}{\sigma_1^2 \sigma_2^2}\Big(- \frac{\phi_2^b}{\Phi_2^b} + \frac{W_1 W_2}{\Phi_2^{b2}}\Big), \\
\frac{\partial^2 l_{12}}{\partial \rho_{12} \partial \sigma_2} &= -\frac{s_1 X_2 \beta_2}{\sigma_2^2}\frac{\phi_2^b}{\Phi_2^b}\Big(\delta_i \tilde{\rho_{12}} Q_2 - S_2 - \frac{W_2}{\Phi_2^b}\Big), \\
\frac{\partial^2 l_{12}}{\partial {\rho_{12}}^2} &= \frac{\phi_2^b}{\Phi_2^b}\Big[\delta_i^2 \tilde{\rho_{12}}\big(1 - \delta_i^2 (S_1^2 + S_2^2 - 2\tilde{\rho_{12}} S_1 S_2)\big) + \delta_i^2 S_1 S_2 - \frac{\phi_2^b}{\Phi_2^b}\Big].
\end{aligned}
$$

G Derivatives for Mixed Response Variables

$$\begin{aligned}
\frac{\partial l_{jk}}{\partial \boldsymbol{\eta}_{jk}} &= \frac{\partial}{\partial \boldsymbol{\eta}_{jk}} \log P(c_k) + \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big]\frac{\partial \Phi(d_j)}{\partial \boldsymbol{\eta}_{jk}} \\
&= -\frac{1}{\sigma_k}\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}} - c_k\frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}} + \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big]\phi(d_j)\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}} \\
\frac{\partial^2 l_{jk}}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T} &= \Big(\frac{1}{\sigma_k^2}\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial \sigma_k}{\partial \eta^T} - \frac{1}{\sigma_k}\frac{\partial^2 \sigma_k}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T} - \frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}^T} - c_k\frac{\partial^2 c_k}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T}\Big) \\
&\quad + \Big[-\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi^2(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{\big(1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})\big)^2}\Big]\Big[\phi(d_j)^2\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial d_j}{\partial {\boldsymbol{\eta}_{jk}}^T}\Big] \\
&\quad + \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big] \\
&\quad \times\Big[-d_j\phi(d_j)\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}^T} + \phi(d_j)\frac{\partial^2 d_j}{\partial \boldsymbol{\eta}_{jk}\partial {\boldsymbol{\eta}_{jk}}^T}\Big]
\end{aligned}$$

To get the first and second order derivatives, we need to use the following derivatives function from the normal density function,

$$
\begin{aligned}
\Phi\prime(d_j) &= \phi(d_j), \\
\phi\prime(c_k) &= -c_k\phi(c_k), \\
\phi\prime\prime(c_k) &= (c_k^2-1)\phi(c_k),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}_{jk}} \log P(c_k) &= \frac{\partial \log(\frac{1}{\sigma_k}\phi(c_k))}{\partial \boldsymbol{\eta}_{jk}} \\
&= \frac{\partial}{\partial \boldsymbol{\eta}_{jk}} \log \phi(c_k) - \frac{\log \sigma_k}{\partial \boldsymbol{\eta}_{jk}} \\
&= \frac{\frac{\partial \phi(c_k)}{\partial \boldsymbol{\eta}_{jk}}}{\phi(c_k)} - \frac{\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}}}{\sigma_k} \\
&= \frac{-c_k\phi(c_k)\frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}}}{\phi(c_k)} - \frac{\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}}}{\sigma_k} \\
&= -c_k\frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}} - \frac{1}{\sigma_k}\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \log P(c_k)}{\partial \boldsymbol{\eta}_{jk} \partial \boldsymbol{\eta}_{jk}^T} &= \frac{1}{\sigma_k^2}\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial \sigma_k}{\partial \boldsymbol{\eta}_{jk}^T} - \frac{1}{\sigma_k}\frac{\partial^2 \sigma_k}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T} - \frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial c_k}{\partial \boldsymbol{\eta}_{jk}^T} - c_k\frac{\partial^2 c_k}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T}, \\
\frac{\partial \Phi(d_j)}{\partial \boldsymbol{\eta}_{jk}} &= \phi(d_j)\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}}, \\
\frac{\partial^2 \Phi(d_j)}{\partial \boldsymbol{\eta}_{jk}\partial \boldsymbol{\eta}_{jk}^T} &= -d_j\phi(d_j)\frac{\partial d_j}{\partial \boldsymbol{\eta}_{jk}}\frac{\partial d_j}{\partial {\boldsymbol{\eta}_{jk}}^T} + \phi(d_j)\frac{\partial^2 d_j}{\partial {\boldsymbol{\eta}_{jk}}^T \partial {\boldsymbol{\eta}_{jk}}^T}.
\end{aligned}
$$

Accordingly, the detail first derivatives for $\boldsymbol{\eta}_{jk}$ is as below,

$$
\begin{aligned}
\frac{\partial c_k}{\partial \beta_k} &= -\frac{1}{\sigma_k} X^{(k)}, \\
\frac{\partial c_k}{\partial \sigma_k} &= -\frac{1}{\sigma_k} c_k, \\
\frac{\partial d_j}{\partial \beta_k} &= \frac{\rho_{jk} X^{(k)}}{\sigma_k \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial d_j}{\partial \beta_j} &= -\frac{X^{(j)}}{\sigma_j \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial d_j}{\partial \sigma_k} &= \frac{\rho_{jk} c_k}{\sigma_k \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial d_j}{\partial \sigma_j} &= \frac{X^{(j)} \beta_j}{\sigma_j^2 \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial d_j}{\partial \rho_{jk}} &= -\frac{c_k}{\sqrt{1-\rho_{jk}^2}} - \frac{X^{(j)} \beta_j \rho_{jk}}{\sigma_j (1-\rho_{jk}^2)^{\frac{3}{2}}} - \frac{\rho_{jk}^2 c_k}{(1-\rho_{jk}^2)^{\frac{3}{2}}},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l_{jk}}{\partial \beta_k} &= \frac{c_k}{\sigma_k} X^{(k)} + \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big] \frac{\phi(d_j)\rho_{jk} X^{(k)}}{\sigma_k \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial l_{jk}}{\partial \beta_j} &= \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big] \Big[-\frac{\phi(d_j) X^{(j)}}{\sigma_j \sqrt{1-\rho_{jk}^2}}\Big], \\
\frac{\partial l_{jk}}{\partial \sigma_k} &= -\frac{1}{\sigma_k} + \frac{c_k^2}{\sigma_k} + \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big] \frac{\phi(d_j)\rho_{jk} c_k}{\sigma_k \sqrt{1-\rho_{jk}^2}}, \\
\frac{\partial l_{jk}}{\partial \sigma_j} &= \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big] \phi(d_j) \frac{X^{(j)} \beta_j}{\sigma_j^2 \sqrt{1-\rho_{jk}^2}},
\end{aligned}
$$

$$\frac{\partial l_{jk}}{\partial \rho_{jk}} = \Big[\frac{1(\boldsymbol{z}^{(j)}=0)}{\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})} - \frac{1(\boldsymbol{z}^{(j)}=1)}{1-\Phi(\boldsymbol{z}^{(j)}|\boldsymbol{z}^{(k)})}\Big]\phi(d_j)$$
$$\times \Big[-\frac{c_k}{\sqrt{1-\rho_{jk}^2}} - \frac{X^{(j)}\beta_j\rho_{jk}}{\sigma_j(1-\rho_{jk}^2)^{\frac{3}{2}}} - \frac{\rho_{jk}^2 c_k}{(1-\rho_{jk}^2)^{\frac{3}{2}}}\Big].$$

and the detail second derivatives are

$$\frac{\partial^2 c_k}{\partial \sigma_k^2} = \frac{2}{\sigma_k^2} c_k,$$
$$\frac{\partial^2 c_k}{\partial \beta_k \partial \sigma_k} = \frac{1}{\sigma_k^2} x_k,$$
$$\frac{\partial^2 d_j}{\partial \rho_{jk} \partial \sigma_j} = \frac{\rho_{jk}(X^{(j)}\beta_j)}{\sigma_j^2(1-\rho_{jk}^2)^{\frac{3}{2}}},$$
$$\frac{\partial^2 d_j}{\partial \rho_{jk} \partial \sigma_k} = \frac{c_k}{\sigma_k}\Big[\frac{1}{\sqrt{1-\rho_{j,k}^2}} + \rho_{j,k}^2(1-\rho_{j,k}^2)^{-\frac{3}{2}}\Big],$$
$$\frac{\partial^2 d_j}{\partial \beta_j \partial \sigma_j} = \frac{X^{(j)}}{\sigma_j^2\sqrt{1-\rho_{jk}^2}},$$
$$\frac{\partial^2 d_j}{\partial \beta_j \partial \rho_{jk}} = -\frac{X^{(j)}\rho_{jk}}{\sigma_j(1-\rho_{jk}^2)^{\frac{3}{2}}},$$
$$\frac{\partial^2 d_j}{\partial \beta_k \partial \rho_{jk}} = \frac{X^{(k)}}{\sigma_k}\Big[\frac{1}{\sqrt{1-\rho_{jk}^2}} + \rho_{jk}^2(1-\rho_{jk}^2)^{-\frac{3}{2}}\Big],$$

$$\frac{\partial^2 d_j}{\partial \sigma_k \partial \beta_k} = -\frac{\rho_{jk} X^{(k)}}{\sigma_k^2 \sqrt{1-\rho_{jk}^2}},$$

$$\frac{\partial^2 d_j}{\partial \sigma_k^2} = -\frac{2\rho_{jk} c_k}{\sigma_k^2 \sqrt{1-\rho_{jk}^2}},$$

$$\frac{\partial^2 d_j}{\partial \sigma_j^2} = -\frac{2X^{(j)T}\beta_j}{\sigma_j^3 \sqrt{1-\rho_{jk}^2}},$$

$$\frac{\partial^2 d_j}{\partial \rho_{jk}^2} = -\frac{c_k \rho_{jk}}{(1-\rho_{jk}^2)^{\frac{3}{2}}} - \frac{X^{(j)}\beta_j}{\sigma_j}\Big[\frac{1}{(1-\rho_{jk}^2)^{\frac{3}{2}}} + \frac{3\rho_{jk}^2}{(1-\rho_{jk}^2)^{\frac{5}{2}}}\Big]$$
$$- c_k\Big[\frac{2\rho_{jk}}{(1-\rho_{jk}^2)^{\frac{3}{2}}} + \frac{3\rho_{jk}^3}{(1-\rho_{jk}^2)^{\frac{5}{2}}}\Big].$$