A NEW METHOD FOR MULTI-CLASS CLASSIFICATION WITH MULTIPLE DATA SOURCES, WITH APPLICATION TO ABDOMINAL PAIN DIAGNOSIS

by

Shen Ling

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
May 2022

Contents

List of Tables

List of Figures

Abstract

In this thesis, we deal with two extremely challenging issues that arise in a medical diagnosis problem. Namely multi-class classification and integration of data from multiple sources. Both are issues that arise in a wide variety of data analysis problems. We present simple but effective methods for dealing with these issues that significantly improve performance in an abdominal pain emergency diagnosis problem, and are widely applicable wherever these issues arise.

For integrating data from multiple sources, such as various medical tests that might be ordered for a patient, our method involves fitting separate predictors on the different sources of data, then performing a linear combination of these predictors. We show that in common cases, this method performs asymptotically better than analysing a single source of data. We also show that the method performs well compared to the popular multiple imputation approach. This very straightforward approach is applicable to a wide range of problems.

For the multi-class classification, we develop a hierarchical tree clustering of the diagnoses, thus reducing the multiclass classification to a series of binary classifications. The hierarchical tree is created using a mixture of data-driven methods based on posterior predictive probability and expert knowledge. We use a statistical learning method to combine the outputs of the binary classifications into an overall output. We find that this works better than multiplying the probabilities from the binary classifiers, which can be misled by the conditional classifiers whose conditions are not met.

Acknowledgements

I would like to express my gratitude to my primary supervisors, Prof. Hong Gu, Prof. Chris Field and Prof. Toby Kenney who guided me throughout this research and offered deep insight into the study. I would also like to thank Dr. Michael Butler and Nidhin Nandhakumar for their help on abdominal pain data processing. Data support from the Charles V Keating Emergency Department and the Department of Emergency Medicine at Dalhousie University is also gratefully acknowledged. I would also like to thank my mother who supported me.

Chapter 1

Introduction

Emergency departments are extremely busy and as a result, misdiagnoses are common, sometimes with dangerous consequences for patients. Machine learning tools for assisting physician diagnoses could be very helpful in preventing some of these misdiagnoses. This thesis looks at the particular case of abdominal pathology, and builds a classifier for predicting diagnoses for patients presenting with abdominal pain. Abdominal pain is the most common symptom among patients presenting at emergency departments. There are two major statistical difficulties encountered in this project. These are common statistical problems that arise in a number of problems in many different fields, so the methods developed in this thesis are expected to have broad applicability.

The first challenge is the integration of different sources of data. All patients arriving at the emergency department are assessed at triage, where a number of basic predictors are recorded. However, further tests are ordered by the treating physician based on their assessment of which tests are needed. This means that different clinical variables will be available for each patient, and we need to fit models that make use of all variables available for each patient.

The second challenge is multi-class classification. A number of popular classification methods are either only available for binary classification problems, or have better performance for binary classification problems. In the medical diagnosis problem, the diagnoses can naturally be hierarchically clustered based on similarity of symptoms. This allows us to arrange the diagnoses in a tree structure, reducing the multiclass classification problem into a sequence of binary classifiers.

1.1 Background

Abdominal pain is a common medical condition, accounting for roughly a quarter of all presentations to the emergency department (ED). The symptomatologys associated with abdominal pathologies are unpleasant for the patient, and requires special attention in their

evaluation (Clark and Kruse, 1990). It is challenging for the clinician to identify the source of abdominal pain since it requires a thorough understanding of the pathogenesis of the various abdominal diseases that cause pain, and the pathways over which it is transmitted. In fact, in the busy environment of the emergency department (ED), up to 15% of all patients are misdiagnosed (Burroughs et al., 2005).

Initial complaints of abdominal pain can be very nonspecific and only evolve to more disease-specific symptoms over time (Macaluso and McNamara, 2012). This increases the difficulty of an accurate identification of the cause of acute abdominal pain. The first step in the diagnostic pathway is clinical evaluation. In daily practice, a preliminary diagnosis will be made based on medical history, physical examination, and, in some cases, laboratory parameters. After clinical assessment, the decision can be made to perform additional diagnostic investigations such as plain radiography, ultrasound, and computed tomography (CT) to increase the certainty of the diagnosis (Gans et al., 2015). In addition to these difficulties related to the diagnosis, there are concerns about misdiagnosis related to interruptions during the course of a shift of the attending clinician (Monteiro et al., 2015).

While machine learning techniques have been widely applied in medicine, (e.g. (Shavlik et al., 1990)) fully diagnosing patients involves integrating information from a large number of sources in a way that exceeds the capabilies of current methods. We therefore aim to develop a tool to assist the physician in making the diagnosis by providing a list of the most plausible diagnoses. The physician can then examine the list to see whether there are any likely possibilities that may have been overlooked. In addition to the list of most plausible diagnoses, we will estimate the probability of each diagnosis, to further help the physician confirm their diagnosis.

1.2 Data and Challenge

We analyse a dataset consisting of 116,008 presentations to emergency departments in Nova Scotia during the period January 2010 to February 2015. We have restricted attention to 39 abdominal pathology diagnoses, excluding rare diagnoses for which there is not sufficient data. The objective of the analysis is to predict the most likely diagnoses for each patient.

There are two major challenges in this dataset, that arise in a large number of statistical problems. The first is the large number of classes. Methods such as neural networks, random forest and boosted trees can be directly applied to multi-class classifications. However, the

performance is often better for binary classifications. Therefore, by restructuring the problem as a series of binary classifiers, taking into account the relations between diagnoses, we can substantially improve the accuracy of the model.

The second major statistical challenge is that results from various medical tests are available only for certain cases, specifically for those patients for whom the physician ordered that test. This is in one sense a data integration problem, where data are pooled from multiple sources with different test results available for each patient. However, it can also be viewed as a missing data problem, where the missing variables are arranged in blocks, and each block is either present or missing. There are a number of available methods for handling missing data, but these methods do not take advantage of the block-missing structure in this dataset. Because the decision of what tests to order is based on the patient's symptoms and suspected diagnosis, this is considered missing not at random. In these cases, by incorporating the non-random nature of the missing data into the model, it is possible to improve prediction. However, in practice we want our diagnosis system to provide preliminary diagnoses before the physician has finished ordering tests. We therefore want to avoid using the information about which tests have been ordered in our prediction, so that our method will generalise to this preliminary use.

1.3 Contribution

We propose a model combination method to deal with block missing data. We fit two models: one using the complete cases, and one using the incomplete cases. For each patient, each model gives a prediction. We then form the final prediction as a linear combination of these two predictions, where the coefficients of the linear combination are based on the sample sizes and the bias in the missing data. This coefficient can be found in general by cross-validation. In the linear regression case, we also develop a plug-in estimator, which is more accurate and easier to analyse theoretically. We prove that for linear regression, our combination method is asymptotically better than the complete case method. It makes weak or no distribution assumptions and our main interest lies in predictions. The model combination method can be applied to both classification and regression problems using either linear or nonlinear methods. We have developed the theory for linear regression but our method works well empirically in other situations.

For the multi-class classification problem, we build a hierarchical tree of diagnoses using a similarity measure based on the average predicted probability of each diagnosis. We use expert knowledge to decide on the threshold for the clustering. We then combine the diagnoses at this threshold and repeat the clustering.

We put the model combination method and the hierarchical tree together to build a method for predicting the diagnosis of emergency department patients. Our method provides a list of most likely diagnoses with the corresponding probabilities to assist the physician in making a final diagnosis. The classifier is able to give preliminary predictions from the triage results, then refine the results to incorporate new test results as they become available. This adaptive method with the ability to update predictions as new test results become available means that our method can be applied in practice to assist the physician at every step of the decision-making process.

We develop an automated diagnosis system using hierarchical tree structures and model combination techniques to help reduce misdiagnosis. Our proposed method employs a hierarchical classification technique that mimics a typical triage process conducted by a trained physician. The model could provide preliminary triage results into multiple different pathologies with only a handful of variables, then it could refine the prediction given extra test results. This unique property makes it very feasible for practical application and potentially assisting physicians in every step of their decision-making process. For each patient, at each stage (considering different medical tests), the model needs to output a posterior probability vector for the diagnoses based on the current available variables.

Chapter 2

Literature Review on Missing Data and Class Binarization

In this chapter, we review different missing data mechanisms and their corresponding approaches for data analysis. We also review approaches to multi-class classification problems.

2.1 Missing Data Mechanisms

Rubin (1976) formalized the concept of missing-data mechanisms by treating the missing-data indicators as random variables and assigning them a distribution. Specifically, let $Z = (Z_{ij})$ denote a rectangular $n \times p$ data set; the ith row is $Z_i = (Z_{i1}, ..., Z_{ip})$, where Z_{ij} is the jth observation for subject i. The missingness pattern of this dataset can be represented by the missing indicator matrix $M = (M_{ij})$ with the ith row $M_i = (M_{i1}, ..., M_{ip})$, such that M_{ij} is 1 if Z_{ij} is missing and M_{ij} is 0 if Z_{ij} is present. We use the notation

$$Z_i = (Z_{i1}, ..., Z_{ip}) \sim f(Z_i|\theta)$$

$$M_i = (M_{i1}, ..., M_{ip}) \sim f(M_i|\phi)$$

where $(Z_i, M_i), i = 1, ..., n$ are assumed to be independent and identically distributed. In Rubin (1976), the joint distribution is factored as

$$f(Z_i, M_i|\theta, \phi) = f(Z_i|\theta) f(M_i|Z_i, \phi)$$

where $f(Z_i|\theta)$ represents the model for the data without missing values, $f(M_i|Z_i, \phi)$ models the missing data mechanism, and (θ, ϕ) denotes unknown parameters.

Three broad types of missingness mechanisms, moving from the simplest to the most general, are:

2.1.1 MCAR: Missing completely at random

When missingness M is independent of the data Z, missing or observed, that is, if

$$f(M_i|Z_i, \phi) = f(M_i|\phi)$$

the data are called missing completely at random (MCAR). With the exception of some planned missing-data designs, MCAR is a strong assumption. Missingness often depends on the observed and/or unobserved data.

2.1.2 MAR: Missing at random

Let $Z_{obs,i}$ denote the observed component of Z_i and $Z_{mis,i}$ the missing component. A less restrictive assumption is that missingness depends only on the observed values $Z_{obs,i}$, and not on the missing values $Z_{mis,i}$. That is,

$$f(M_i|Z_i, \phi) = f(M_i|Z_{obs,i}, \phi)$$

The missing-data mechanism is then called missing at random (MAR). For example in a medical dataset, blood test results might be missing if the doctor did not consider them necessary based on the available data for that patient.

The observed data consist of the values of the variables (Z_{obs}, M) and the distribution of the observed data is obtained by integrating Z_{mis} out of the joint density of $Z = (Z_{obs}, Z_{mis})$ and M. That is, for unit i,

$$\begin{aligned} f(Z_{obs,i}, M_i|\theta, \phi) &= \int f(Z_{obs,i}, Z_{mis,i}, M_i|\theta, \phi)\mathrm{d}Z_{mis,i} \\ &= \int f(Z_{obs,i}, Z_{mis,i}|\theta) f(M_i|Z_{obs,i}, Z_{mis,i}, \phi)\mathrm{d}Z_{mis,i} \end{aligned}$$

Under MAR, $M_i|Z_i = M_i|Z_{obs,i}$, so

$$\begin{aligned} f(Z_{obs,i}, M_i|\theta, \phi) &= f(M_i|Z_{obs,i}, \phi) \int f(Z_{obs,i}, Z_{mis,i}|\theta)\mathrm{d}Z_{mis,i} \\ &\propto \int f(Z_{obs,i}, Z_{mis,i}|\theta)\mathrm{d}Z_{mis,i} \\ &= f(Z_{obs,i}|\theta) \end{aligned}$$

The full likelihood of θ and ϕ is:

$$L_{full}(\theta, \phi|Z_{obs}, M) \propto \prod_{i=1}^{n} f(Z_{obs,i}, M_i|\theta, \phi)$$

This missing-data mechanism is also called ignorable, because when it is MAR and the parameter space for (θ, ϕ) is a Cartesian product space, the Likelihood-based inferences for θ can be based on

$$L_{ign}(\theta|Z_{obs}) \propto \prod_{i=1}^{n} f(Z_{obs,i}|\theta)$$

Because the ignorable likelihood only depends on observed data Z_{obs}, we don't need to build a model for M. If assuming missingness is MCAR or MAR, the missing-data mechanism can be ignored and we only need to model the observed data Z_{obs} to derive likelihood-based inferences for θ.

2.1.3 MNAR: Missing not at random

The mechanism is called missing not at random (MNAR) if the distribution of M depends on the missing values in the data matrix Z.

A common example is that people with higher income are less likely to reveal their income. That is, the non-response probability for the income variable depends on values that can be missing. An MNAR mechanism is often referred to as non-ignorable missingness because the missing-data mechanism cannot be ignored for the inference. In other words, the valid likelihood-based inferences require specification of the missing data mechanism. Analysing MNAR data involves making assumptions about the missing pattern based on the particular problem. In this chapter, we focus on reviewing methods for the MCAR and MAR datasets.

2.2 Handling Missing Data

A good method for handling missing data should

- Minimize bias: Missing data introduces bias into parameter estimates; a good method should make that bias as small as possible.
- Maximize the use of available information: Avoid discarding any data, and use the available data to produce parameter estimates that are efficient (i.e., have minimum sampling variability).
- Yield optimal estimates of uncertainty: We want accurate estimates of standard errors, confidence intervals and p-values.

- Ideally accomplish all of the above without making unnecessarily restrictive assumptions about the missing data mechanism.

The so-called conventional methods are deficient in one or more of these goals, but Maximum likelihood (ML) and multiple imputation (MI) do very well at satisfying these criteria. Our idea of model combination also satisfies these for the particular problem of block missingness (block missing at random).

2.2.1 Conventional approaches

The most common approaches to deal with MCAR and MAR data are

1. Complete case analysis: listwise deletion. The analysis is only run on cases which have a complete set of data. An observation with a missing value in any variable would be removed entirely (Baraldi and Enders, 2010).

2. Available case analysis: pairwise deletion. The analysis uses cases that contain some missing data. We only choose to omit cases with a missing value on the variables we are interested in, but not for other cases (Baraldi and Enders, 2010).

3. Single imputation that the missing value is replaced by a value (Myers, 2000):

    - LOCF (Last Observation Carried Forward). Impute the missing data with the value of the last observation with available data.
    - Mean imputation. Impute the missing data using the mean of the non-missing values.
    - Hot-deck imputation (local imputation/Nearest neighbour). A missing case is replaced with a case with similar characteristics.
    - (Stochastic) regression imputation. Build a regression model with baseline characteristics as predictors of the outcome cases using the available data. Then use the model to predict the outcome for cases with missing values.

Broadly speaking, there are two general model-based approaches for handling nonignorable missingness (MNAR) that can be distinguished: selection models and pattern-mixture models. These methods are based on the likelihood for a model and can be used in maximum likelihood (ML) or fully Bayes modeling.

2.2.2 Selection models

The joint distribution of Z_i and M_i can be written as the product of the marginal distribution of Z_i and the conditional distribution of M_i given Z_i:

$$f(Z_i, M_i) = f(Z_i)f(M_i|Z_i) \tag{2.1}$$

This is a natural way of factoring the model, with $f(Z_i)$ the model for the data in the absence of missing values, and $f(M_i|Z_i)$ the model for the missing-data mechanism that determines what parts of Z are observed. Equation 2.1 is sometimes called a selection model factorization of the joint distribution of (Z_i, M_i) because of connections with the econometric literature on selection bias (Heckman, 1976).

If the MAR assumption is plausible, the selection model formulation leads directly to the ignorable likelihood — the distribution $f(M_i|Z_i)$ for the missing-data mechanism is not needed for likelihood inferences, which can be based solely on the model for $f(Z_i)$.

$$\begin{aligned}
f(Z_{obs,i}, M_i) &= \int f(Z_{obs,i}, Z_{mis,i}, M_i)\mathrm{d}Z_{mis,i} \\
&= \int f(Z_{obs,i}, Z_{mis,i})f(M_i|Z_{obs,i}, Z_{mis,i})\mathrm{d}Z_{mis,i} \\
&= f(M_i|Z_{obs,i}) \int f(Z_{obs,i}, Z_{mis,i})\mathrm{d}Z_{mis,i} \\
&\propto \int f(Z_{obs,i}, Z_{mis,i})\mathrm{d}Z_{mis,i} \\
&= f(Z_{obs,i})
\end{aligned}$$

The selection model factorization does not require full specification of the model for the missing-data mechanism when the data are MAR, but it does if the data are MNAR.

2.2.3 Pattern-mixture models

Another factorization is also possible. Pattern-mixture models (Glynn et al., 1986), (Glynn et al., 1993)) specify the marginal distribution of M_i and the conditional distribution of Z_i given M_i:

$$f(Z_i, M_i) = f(M_i)f(Z_i|M_i)$$

The focus of these models is on the conditional distribution of the response variable given that the data are available. In cases where missing values are not meaningful, rather than unobserved, this model can be more natural. It can also be helpful in cases where we want to use the missingness of variables as an additional predictor.

The distribution of the observed data is obtained as follows:

$$f(Z_{obs,i}, M_i) = \int f(Z_{obs,i}, Z_{mis,i}, M_i)\mathrm{d}Z_{mis,i} \tag{2.2}$$

$$= \int f(M_i) f(Z_{obs,i}, Z_{mis,i}|M_i)\mathrm{d}Z_{mis,i} \tag{2.3}$$

$$= f(M_i) \int f(Z_{obs,i}, Z_{mis,i}|M_i)\mathrm{d}Z_{mis,i} \tag{2.4}$$

$$\propto f(Z_{obs,i}|M_i) \tag{2.5}$$

The pattern-mixture models, avoid specification of the model for the missing-data mechanism in MNAR situations, by directly modelling the conditional distribution of Z given the missing status.

This way of modelling the data has advantages when performing imputation. Missing values $Z_{mis,i}$ should be imputed from their predictive distribution given the observed data including M_i, that is, $f(Z_{mis,i}|Z_{obs,i}, M_i)$. Under MAR this equals $f(Z_{mis,i}|Z_{obs,i})$, which is a conditional distribution derived from the selection model. However, if data are not MAR, the predictive distribution of $Z_{mis,i}$ given $Z_{obs,i}$ and M_i is modeled directly in the pattern-mixture formulation as $f(Z_{mis,i}|Z_{obs,i}, M_i)$.

2.2.4 Other approaches

Other methods for analysing missing data include:

1. Nonresponse weighting (Little and Rubin, 2002, Chapter 3) Roderick et al. (2002). It gives weights for responses based on likelihood of response for complete case analysis, often used in surveys.

2. Multiple imputation (MI), where missing values are replaced by multiple sets of plausible values (Rubin, 1987; Little and Rubin, 2002, Chapter 5) Roderick et al. (2002).

3. Weighted estimating equation (WEE) methods (Lipsitz, Ibrahim and Zhao, 1999) Lipsitz et al. (1999). The contribution to the estimating equation from a complete observation is weighted by the inverse probability of being observed.

Single imputation procedures, such as mean imputation, do not account for the uncertainty in the imputations; once the imputation is completed. Analyses proceed as if the imputed values were the known true values rather than imputed. This will lead to overly precise results and the potential for incorrect conclusions. Maximum likelihood methods are sometimes a viable approach for dealing with missing data (Graham, 2009). However, these methods are primarily available only for certain types of models, such as longitudinal or structural equation models, and can generally be run only using special software such as Amos (SPSS, 2009a) and Lisrel (Scientific Software International, 2006).

2.3 Multiple Imputation

Multiple imputation involves filling in the missing values multiple times, creating multiple "complete" datasets. Described in detail by Schafer and Graham (2002), the missing values are imputed based on the observed values for a given individual and the relations observed in the data for other participants. Multiple imputation procedures, particularly MICE, are very flexible and can be used in a broad range of settings. Because multiple imputation involves creating multiple predictions for each missing value, the analyses of multiply imputed data take into account the uncertainty in the imputations and yield more accurate standard errors estimates.

If there is not much information in the observed data (used in the imputation model) regarding the missing values, the imputations will be very variable, leading to high standard errors in the analyses. In contrast, if the observed data are highly predictive of the missing values the imputations will be more consistent across imputations, resulting in smaller, but still accurate, standard errors (Greenland and Finkle, 1995).

2.3.1 The Chained Equation Approach to Multiple Imputation

Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Schafer (1997) developed various JM techniques for imputation under the

multivariate normal, the log-linear, and the general location model. JM involves specifying a multivariate distribution for the missing data, and drawing imputation from their conditional distributions by Markov chain Monte Carlo (MCMC) techniques. This methodology is attractive if the multivariate normal distribution is a reasonable description of the data. FCS specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, FCS draws imputations by iterating over the conditional densities. A low number of iterations (say $10 - 20$) is often sufficient. This is much more flexible for modelling non-normal variables, but is computationally very expensive.

Multivariate imputation by chained equations (MICE) is a particular multiple imputation technique (Raghunathan et al., 2001), (Van Buuren, 2007)). MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR).

In the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its own distribution (not assumed a joint normal distribution), with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression.

The MICE procedure fits a regression of each variable on all other variables within a chosen subset of variables (for example in a regression problem, we might choose not to use the response variable to impute missing values of predictors). For fitting this regression, the missing values of the predictors are imputed using a basic method, such as mean imputation. We then replace the values of the current variable by the new imputed predictions, and iterate over all other variables. We repeat this until convergence of the imputed values. We assume that there are p variables, of which k are subject to missing data and $p-k$ are complete. The algorithm is summarised in Algorithm 1. The process described in steps 3 and 4 is repeated for several cycles to create one imputed data set. Standard software uses 5 to 20 cycles by default. The imputed values obtained after the last cycle are used as the imputed values for the first imputed data set. The entire process is then repeated M times to produce M imputed data sets. The advantage of this approach over methods that assume a multivariate normal distribution is that different values can be modelled using different distributions in the regression. For example, $\{0, 1\}$ valued variables can be modelled using logistic regression.

Algorithm 1 MICE algorithm for multiple imputation

1: Specify an imputation model for each of the k variables that are subject to missing data.

2: For each of the k variables that are subject to missing data, fill in the missing values with random draws from those subjects with observed values for the variable in question. Note that these initial imputed values do not respect the multivariate relations in the data and will be overwritten by better imputed values in later stages of the algorithm.

3: **for** the first variable that is subject to missing data: **do**

a. Regress this first variable on all the other variables using those subjects with complete data on the first variable and observed or currently imputed values of the other variables.

b. The estimated regression coefficients and their variance-covariance matrix (and the estimated variance of the residual distribution if a linear regression model was fit for a continuous variable) are extracted from the regression model estimated in (a).

c. Using the quantities obtained in (b), randomly perturb the estimated regression coefficients in a way that reflects the degree of uncertainty arising from the data.

d. Using the set of perturbed regression coefficients obtained in (c), the conditional distribution of the first variable is determined for each subject with missing data on that variable.

e. A value of the variable is drawn from this conditional distribution for each subject with missing data on the first variable.

4: Repeat step 3 for each of the variables that is subject to missing data. Steps 3 and 4 form 1 cycle of the imputation process for creating 1 imputed data set.

5: Repeat steps 3 and 4 the desired number of times (suggested 5 to 20 cycles). The final imputed values are used as the imputed values in first imputed data set.

6: Repeat steps 2-5 M times to produce M imputed data sets.

2.3.2 Analyzing Multiply Imputed Data

Once the data have been imputed, each imputed data set is "complete" in the sense that it has no missing values. Analyzing multiply imputed data involves running a standard analysis such as regression on each of the imputed data sets and combining the estimates from each data set to obtain the final result. In order to combine the results across m data sets, first decide on the quantity of interest q to compute, such as a univariate mean, regression coefficient, predicted probability, or first difference (Buuren and Groothuis-Oudshoorn, 2011).

The variance estimates involve both the "within" variance calculated for each dataset individually, as well as the "between" variance that reflects the uncertainty in the imputations—how variable the results are across the imputed datasets. We can combine directly and use as the multiple imputation estimate of this parameter, $\bar{q}$, the average of the m separate estimates, $q_j (j = 1, ..., m)$:

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

The variance of the point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point estimates across the data sets (multiplied by a factor that corrects for the bias because $m < \infty$). Let $SE(q_j)^2$ denote the estimated variance (squared standard error) of q_j from the data set j, and $S_q^2 = \frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{(m-1)}$ be the sample variance across the m point estimates. The standard error of the multiple imputation point estimate is the square root of

$$\frac{1}{m} \sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + 1/m) = SE(\bar{q})^2$$

.

2.4 Block Missing Methodology

Clearly, general methods for handling missing data may deliver sub-optimal performance for block missing data, since they are not specifically designed for block missing data. There are also direct attempts to tackle block missing data. e.g. incomplete Source-Feature Selection (iSFS) model proposed by Xiang et al. (2014) which partition the whole data set into

multiple groups according to the availability of data sources to make both feature-level and source-level analysis for feature selection. A hybrid approach proposed by Li et al. (2014) focus on the inferences of regression coefficients. When the observations for some covariates are complete, they use a Bayesian inference, and for the parameter with missing data, they employ a frequentist method. Xue and Qu (2021) proposed a Multiple Blockwise Imputation (MBI) approach that creates multiple predictions for each missing value from multiple modalities in order to utilize more observed information from incomplete case groups than traditional imputation methods. However, block level imputations when the number of features is high could still result in regressing many sub imputation models, rendering this approach very computationally inefficient.

2.5 Multiclass Classification

The majority of classification methods were designed to solve binary classification problems. Some of these methods can be naturally extended to the multi-class case. Others need special formulations to be able to solve the latter case. The first category of algorithms include decision trees, neural networks, k-Nearest Neighbor and Naive Bayes classifiers. The second category include Support Vector Machines (SVM). Even for methods in the first category, where there is a natural extension to a multiclass classification, this extension can prove less accurate in some cases than the original binary classification method. It is therefore often desirable to estimate a multi-class classification by aggregating the results of a number of binary classifiers.

Class binarization strategies reduce a k-class problem into a series of binary problems for classification. Two of the most common strategies in the literature are one-versus-one (OVO) and one-versus-all (OVA) approaches (Rifkin and Klautau, 2004), also named one-against-one (OAO) and one-against-all (OAA) (Lorena et al., 2008).

In the OVO or OAO strategy, for every pair of classes, a binary classifier is trained on the subset of the data consisting of observations from those two classes. The outputs of the $\binom{k}{2}$ classifiers are combined for prediction.

In the OVA or OAA strategy, for each of the k classes, a binary classifier is trained on the whole data set, with the chosen class as one class, and the other k-1 classes combined to make a single alternative class. These classifiers are then combined to give the overall classification. The OVA strategy requires fewer classifiers to be trained, but the training

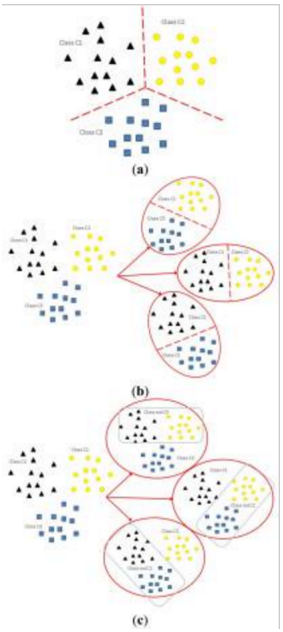

Figure 2.1: Taken from (Ng et al., 2014), three different classification strategies illustrated with a 3-class problem (a) Simple multi-class classification; (b) One-versus-one (OVO) class binarization; (c) One-versus-all (OVA) class binarization.

data for each of these classifiers consists of the whole data set, so is larger, and is often unbalanced, whereas the OVO strategy involves fitting $\binom{k}{2}$ classifiers, but each classifier is trained on a smaller and more balanced dataset.

The differences among the formulations of multi-class classification are illustrated graphically in Figure 2.1.

Chapter 3

Model Combination Method

In this chapter, we propose a model combination method to deal with block missing data. Our method involves fitting separate models for the complete and incomplete cases, then taking a linear combination of the predictions from the two models. Since both models should give approximately unbiased predictions, the coefficients of the linear combinations should sum to 1. We let α be the coefficient of the partial data model, and therefore $1-\alpha$ be the coefficient of the complete-case model. There are two data-driven approaches for choosing α from the data. Cross validation is a traditional approach for choosing a tuning parameter related to the bias-variance trade-off. It is computationally intensive and difficult to analyze from a theoretical perspective. For linear and logistic regression, we are able to analytically solve for the optimal value of α, based on the variance structure of the data. This allows a plug-in estimator where we replace the variances in this formula with estimates from the data. This requires less computation and is theoretically more tractable.

We are dealing with estimation of a prediction function f that predicts our response variable Y from a vector X of p_1 covariates. Our training data is divided into two parts, as shown in Figure 3.1. The first part, termed as full data, contains n_1 observations of X and Y. The second, partial data, contains $n_2 = n - n_1$ observations, but only of the first p_2 covariates. We will use the following notation:

$$\begin{aligned}
\mathbf{Y}_1^T &= (\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}) \\
\mathbf{Y}_2^T &= (\mathbf{y}_{n_1+1}, \ldots, \mathbf{y}_n) \\
\mathbf{X}_1^T &= (\mathbf{X}_{11}, \mathbf{X}_{12})^T = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}) \\
\mathbf{X}_{21}^T &= (\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n)
\end{aligned}$$

It is straightforward to estimate the predictor function f from either part of the training data on its own. We can use any linear or non-linear predictive models including linear or generalized linear models, random forest, support vector machine, neural networks, etc. Thus we have two estimators $\widehat{f}_1$ from the first part of the training data, $(\mathbf{Y}_1, \mathbf{X}_1)$, and $\widehat{f}_2$ from

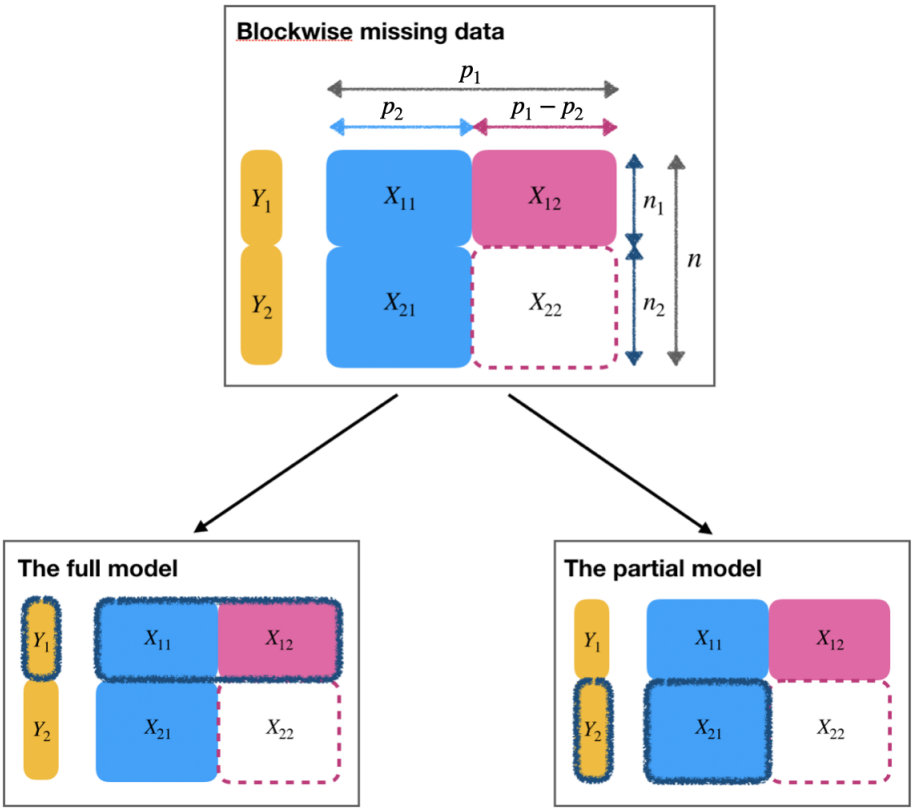

Figure 3.1: An illustration of a blockwise missing data problem with the two blocks of variables represented in pink and blue, and the blank region represents the missing part of the data. The data is partitioned into the full model and the partial model, as highlighted by the dark blue boxes. The model combination method is the weighted mixture of these two models.

the second part of the training data, $(\mathbf{Y}_2, \mathbf{X}_{21})$. For prediction of future partially-observed data points, we also fit a model $\hat{f}_0$ using all training observations, but only the variables present in all observations. For prediction of fully observed data points, we will use the following combined estimation which is a simple weighted average of two estimators:

$$\hat{f}(x) = (1-\alpha)\widehat{f_1}(x) + \alpha\widehat{f_2}(x) \tag{3.1}$$

for some suitably chosen scalar α, which we will estimate from the data. We will present two approaches to estimating α.

The plug-in approach for linear regression is introduced in Section 3.1 and for a more general loss function is given in Section 3.2. We present the leave-one-out cross-validation estimate for α in Section 3.3. Finally we prove some asymptotic result for the model combination method for linear regression models in Section 3.4.

3.1 Model Combination for Linear regression models

We first consider the linear regression model case. That is, we assume that

$$Y = X\beta + \epsilon \tag{3.2}$$

for some vector β of coefficients, where $\epsilon \sim N(0, \sigma_\epsilon^2)$. We will restrict $\widehat{f_1}$ and $\widehat{f_2}$ to be linear functions, estimated by least squares. We let $\widehat{f_1}(x) = x^T\widehat{\boldsymbol{\beta_1}}$ and $\widehat{f_2}(x) = x^T\widehat{\boldsymbol{\beta_2}}$, where the ith element of $\widehat{\beta_2}$: $(\widehat{\beta_2})_i = 0$ for $i > p_2$. Now our combined estimator (3.1) can be rewritten as

$$\widehat{f_\alpha}(\mathbf{x}) = \mathbf{x}^T\widehat{\boldsymbol{\beta_\alpha}} \tag{3.3}$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = (1-\alpha)\widehat{\boldsymbol{\beta}}_{\mathbf{1}} + \alpha\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$.

We will choose α in an attempt to minimise the expected squared error loss:

$$R(\alpha) = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left((1-\alpha)\widehat{f_1}(\mathbf{x}_0) + \alpha\widehat{f_2}(\mathbf{x}_0)\right) - f(\mathbf{x}_0)\right]^2 \tag{3.4}$$

$$= (1-\alpha)^2 A + \alpha^2 B \tag{3.5}$$

where $A = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\widehat{f_1}(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2$ and $B = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\widehat{f_2}(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2$ are the expected squared error for the full model and the partial model respectively. Because we are assuming that the two parts of training data are independent, we have that $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$ are

independent, and $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ is unbiased, so the cross term from (3.4) vanishes to give (3.5).

Under the assumption that both A and B exist and are finite, the minimum of $R(\alpha)$ is easily seen to be achieved for

$$\alpha^* = \frac{A}{A+B}. \tag{3.6}$$

From standard linear regression theory, assuming the existence of $(X^TX)^{-1}$, we know that the conditional expected squared error

$$\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{Y|X}\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2\right) = \sigma_\epsilon^2 \operatorname{tr}((X^TX)^{-1}\mathbb{E}_{\mathbf{x}_0}(\mathbf{x}_0\mathbf{x}_0^T))$$

Although the estimates of the expected squared error with expectation over both training and test data in linear regression has been widely discussed, it seems to have always been the case that estimators for the above conditional expected squared error have been used with an assumption that $(X^TX)^{-1}$ exists. The difficulty of further taking the expectation over the distribution of X is that in general $\mathbb{E}_X(X^TX)^{-1}$ doesn't exist. However to develop an unbiased plug-in estimator and to better study the asymptotic behaviour of our proposed estimator, we need to analyze the asymptotic behaviour of the quantities A and B. We will develop the theory for a bounded estimator under the normal assumptions for the predictor variables as following.

Theorem 3.1.1. *Assume the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X \sim N(X\beta, \sigma_\epsilon^2)$, where β is a $p_1 \times 1$ regression coefficient vector. Suppose the theoretical optimal predictor from the full data is $f_1(x) = x\beta$ and let $f_2(x) = x\beta_2$ be the theoretical optimal predictor subject to the constraint that the ith element of β_2: $(\beta_2)_i = 0$ for all $i > p_2$. Let the least square estimators be $\widehat{f}_1(x) = x^T\widehat{\boldsymbol{\beta}_{\mathbf{1}}}$ and $\widehat{f}_2(x) = x^T\widehat{\boldsymbol{\beta}_{\mathbf{2}}}$, where the ith element of $\widehat{\beta}_2$: $(\widehat{\beta}_2)_i = 0$ for $i > p_2, (p_2 < p_1)$.*

As the sample sizes n_1 and n_2 tend to ∞,

$$A = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1}\right) = \sigma_\epsilon^2\left(\frac{p_1}{n_1} + \frac{p_1^2+p_1}{n_1^2} + o\Big(\frac{1}{n_1^2}\Big)\right), \tag{3.7}$$

$$B = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2}\right) = \left(\sigma_\eta^2 - \sigma_\epsilon^2\right) + \sigma_\eta^2\left(\frac{p_2}{n_2} + \frac{p_2^2+p_2}{n_2^2} + o\Big(\frac{1}{n_2^2}\Big)\right), \tag{3.8}$$

where C_n is a sequence of constants, bounded below, satisfying $C_n = o(n)$, and σ_η^2 is the theoretical minimum MSE for a linear predictor using only the first p_2 predictors. That is, $\sigma_\eta^2 - \sigma_\epsilon^2 = (\boldsymbol{\beta} - \boldsymbol{\beta_2})^T \mathbb{E}(\mathbf{x_0}\mathbf{x_0}^T)(\boldsymbol{\beta} - \boldsymbol{\beta_2})$.

The proof of **Theorem 3.1.1** is given in Section 3.4.1.

Remark 1. *When $\beta \neq \beta_2$, we have that $\sigma_\eta^2 - \sigma_\epsilon^2 > 0$, so that B is bounded below by a constant, while $A \to 0$ as $n_1 \to \infty$. Thus $\alpha_* \to 0$ as $n_1 \to \infty$.*

Following **Theorem 3.1.1**, we now re-define a bounded version of the risk in the following

Corollary 3.1.1.1. *Under the same model assumptions as in **Theorem 3.1.1**, for any fixed value of α, we have the risk*

$$R(\alpha) = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f_1}(\mathbf{x}_0) + \alpha\widehat{f_2}(\mathbf{x}_0)\right) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1\wedge n_2}\right] \tag{3.9}$$

$$= (1-\alpha)^2 A + \alpha^2 B + o((n_1 \wedge n_2)^{-2}) \tag{3.10}$$

where $C_{n_1\wedge n_2}$ is a sequence of constants, bounded below, satisfying $C_{n_1\wedge n_2} = o(n_1 \wedge n_2)$ and A and B are given by (3.7) and (3.8).

A plug-in estimator of α^* can be easily obtained as $\widehat{\alpha^*} = \dfrac{\hat{A}}{\hat{A} + \hat{B}}$, where $\hat{A}$ and $\hat{B}$ can be obtained by plugging estimators $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_\eta^2$ for σ_ϵ^2 and σ_η^2 in (3.7) and (3.8).

The unbiased estimators for σ_ϵ^2 and σ_η^2 are:

$$\widehat{\sigma_\epsilon^2} = \frac{1}{n_1 - p_1}\sum_{i=1}^{n_1}(\mathbf{y}_i - \widehat{f_1}(\mathbf{x}_i))^2$$

$$\widehat{\sigma_\eta^2} = \frac{1}{n_2 - p_2}\sum_{i=1+n_1}^{n}(\mathbf{y}_i - \widehat{f_2}(\mathbf{x}_i))^2$$

We will show that with this estimator for α, our combination method performs asymptotically better than using only the complete case data to estimate β.

3.2 General Loss Function with $O\left(n^{-\frac{1}{2}}\right)$ Convergence

In this section, we assume that observations in the training data X_1 and X_2 are drawn i.i.d. from some arbitrary distribution, and the conditional distributions of $Y|X$ for different data

points are independent and follow the same type of distribution. We have a general loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that measures how close our estimates $\hat{y}$ are to the true value, y. We do not require the loss function to be minimised when $\hat{y} = y$, so we can use the loss function to measure a transformed estimator. For example in a binary classification problem, we could let

$$L(y, \hat{y}) = -\left[y \log\left(\frac{\exp(\hat{y})}{1 + \exp(\hat{y})} \right) + (1 - y) \log\left(\frac{1}{1 + \exp(\hat{y})} \right) \right] \tag{3.11}$$

which is the negative log-likelihood of y under the assumption that the logistic transform of the estimated $P(y = 1)$ is $\hat{y}$.

We will assume that f_1 and f_2 are chosen from some class $\mathcal{M}$ of integrable functions. We will let $f_1(x)$ be the function that minimises $\mathbb{E}(L(Y, f(X)))$, subject to $f_1 \in \mathcal{M}$. Similarly, we will let $f_2(x)$ be the function that minimises $\mathbb{E}(L(Y, f(X)))$, subject to $f_2 \in \mathcal{M}$ and for any two points x and x' with the same elements on the first p_2 dimensions, we have $f_2(x) = f_2(x')$. We let $\widehat{f_1}(x)$ and $\widehat{f_2}(x)$ be the estimated functions from the full data (Y_1, X_1) and partial data (Y_2, X_{21}) corresponding to their theoretical optimum $f_1(x)$ and $f_2(x)$ respectively, and define $E_1(x) = \widehat{f_1}(x) - f_1(x)$, $E_2(x) = \widehat{f_2}(x) - f_2(x)$ and let $\eta(x) = f_2(x) - f_1(x)$. It follows that $\widehat{f_\alpha}(x) = f_1(x) + E_1(x) + \alpha(\eta(x) + E_2(x) - E_1(x))$, and so if E_1 and α are sufficiently small, and $\mathcal{M}$ is closed under linear combinations, we have

$$\begin{aligned}\mathbb{E}(L(\mathbf{y_0}, \widehat{f_\alpha}(\mathbf{x_0})) = \mathbb{E}(L(\mathbf{y_0}, f_1(\mathbf{x_0})) + \frac{1}{2}\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0})) \left[E_1(\mathbf{x_0}) + \alpha\left(\eta(\mathbf{x_0}) + E_2(\mathbf{x_0}) - E_1(\mathbf{x_0}) \right) \right]^2 \right) \\ + o(\mathbb{E}(E_1(\mathbf{x_0})^2)) + o(\alpha^2)\end{aligned}$$

where the first derivative in the Taylor expansion vanishes because of the optimality of f_1. From this, we see that the asymptotically optimal value of α is

$$\alpha^* = \frac{-\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0})) E_1(\mathbf{x_0}) \left(\eta(\mathbf{x_0}) + E_2(\mathbf{x_0}) - E_1(\mathbf{x_0}) \right) \right)}{\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0})) \left(\eta(\mathbf{x_0}) + E_2(\mathbf{x_0}) - E_1(\mathbf{x_0}) \right)^2 \right)}$$

Since E_1 and E_2 are estimated from different data sets, we can assume they are independent, so the covariance is zero, making the cross terms in the previous expression negligible. We will also assume that E_1 and E_2 converge in L^2-norm to zero, while η is a non-zero constant function, so that asymptotically

$$\alpha^* = \frac{\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0})) E_1(\mathbf{x_0})^2 \right)}{\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0})) \eta(\mathbf{x_0})^2 \right)} \tag{3.12}$$

This is similar to the linear case, where under the squared-error loss, $\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0}))$ is a constant, so the numerator becomes $\mathbb{E}\left((\hat{f}_1(\mathbf{x_0}) - \mathbb{E}(Y|\mathbf{x_0}))^2\right)$, which we defined as A. The denominator is approximately equal to $\mathbb{E}\left((\hat{f}_2(\mathbf{x_0}) - \mathbb{E}(Y|\mathbf{x_0}))^2\right)$, which we defined as B. Since A is asymptotically smaller than B, we have simplified the expression $\frac{A}{A+B}$ from Section 3.1, to $\frac{A}{B}$ in (3.12). To obtain our plug-in estimator in the linear case, we used the fact that the estimation error A is asymptotically proportional to the irreducible error. If we make the similar assumption here

$$\mathbb{E}(L(\mathbf{y_0}, f_1(\mathbf{x_0}) + E_1(\mathbf{x_0})) = \mathbb{E}(L(\mathbf{y_0}, f_1(\mathbf{x_0}))\left(1 + \frac{p_1}{n_1}\right) + o_p\left(\frac{1}{n_1}\right)$$

then we can estimate $\mathbb{E}(L(\mathbf{y_0}, f_1(\mathbf{x_0}))$ from the validation loss function. That is, we divide our training data into a sub-training sample and a validation sample. We let n_{train} be the number of samples in this sub-training set, and n_{val} be the number of validation samples. Then since $\hat{f}_1 \to f_1$, we have an estimator of $\mathbb{E}(L(\mathbf{y_0}, f_1(\mathbf{x_0}))$ as

$$\frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} L(y_i, \hat{f}_{1,\text{train}}(x_i)),$$

and thus an estimator for $\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0}))E_1(\mathbf{x_0})^2\right)$ as $\frac{2p_1}{n_{\text{train}} n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} L(y_i, \hat{f}_{1,\text{train}}(x_i))$. For the denominator, we estimate the derivative from first principles. That is, we use

$$\mathbb{E}\left(L(y_0, f_1(\mathbf{x_0}) + {n_{\text{val}}}^{-\frac{3}{2}}(\eta(\mathbf{x_0}))) - L(y_0, f_1(\mathbf{x_0}))\right) \approx \frac{{n_{\text{val}}}^{-3}}{2} \mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0}))\eta(\mathbf{x_0})^2\right)$$

so that

$$\frac{2n_{\text{val}}^3}{n_{\text{train}}} \sum_{j=1}^{n_{\text{train}}} \left(L(y_j, \hat{f}_{1,\text{train}}(x_j) + {n_{\text{val}}}^{-\frac{3}{2}}(\hat{f}_2(x_j) - \hat{f}_{1,\text{train}}(x_j))) - L(y_j, \hat{f}_{1,\text{train}}(x_j))\right)$$

is an estimator for

$$\mathbb{E}\left(\frac{\partial^2}{\partial \hat{y}^2} L(\mathbf{y_0}, f_1(\mathbf{x_0}))\eta(\mathbf{x_0})^2\right)$$

This gives the estimator

$$\widehat{\alpha^*} = \frac{\frac{p_1}{n_{\text{train}} n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} L(y_i, \hat{f}_{1,\text{train}}(x_i))}{\frac{n_{\text{val}}^3}{n_{\text{train}}} \sum_{j=1}^{n_{\text{train}}} \left(L(y_j, \hat{f}_{1,\text{train}}(x_j) + {n_{\text{val}}}^{-\frac{3}{2}}(\hat{f}_2(x_j) - \hat{f}_{1,\text{train}}(x_j))) - L(y_j, \hat{f}_{1,\text{train}}(x_j))\right)} \tag{3.13}$$

3.3 Cross validation estimation

For cases where we do not want to make assumptions about the convergence of $\widehat{f}_1$, an alternative approach is to use cross-validation to estimate α. We use leave-one-out cross-validation (LOOCV) to estimate the expected loss. That is, we let $\widehat{f_{1,(i)}}$ be the estimated predictor from the training data with the ith observation removed, where $i \in \{1, 2, \cdots, n_1\}$. We let $\widehat{y_{1,(i)}} = \widehat{f_{1,(i)}}(\mathbf{x_i})$. We also have the partial model which are fitted on data (Y_2, X_{21}) and predict on each observation in X_1 $\widehat{y_{2,i}} = \widehat{f}_2(\mathbf{x_i})$. We then select $\hat{\alpha}_{\text{cv}}$ as

$$\hat{\alpha}_{\text{cv}} = \arg\min_{\alpha \in [0,1]} \sum_{i=1}^{n_1} L(y_i, (1-\alpha)\widehat{y_{1,(i)}} + \alpha \widehat{y_{2,i}})$$

The procedure is written out in detail in Algorithm 2.

Algorithm 2 CV estimation

1: Fit a model $\hat{f}_2$ on the partial data $(\mathbf{x_{n_1+1}}, y_{n_1+1}), \ldots, (\mathbf{x_{n_1+n_2}}, y_{n_1+n_2})$.
2: **for** $i \in 1, \ldots, n_1$ **do**
3: Fit a model $\widehat{f_{1,(i)}}$ on the data $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_{i-1}}, y_{i-1}), (\mathbf{x_{i+1}}, y_{i+1}), \ldots, (\mathbf{x_{n_1}}, y_{n_1})$.
4: Let $\widehat{y_{1,(i)}} = \widehat{f_{1,(i)}}(\mathbf{x_i})$.
5: Let $\widehat{y_{2,i}} = \widehat{f}_2(\mathbf{x_i})$.
6: Fit $\hat{\alpha}_{\text{cv}}$ to minimise: $\sum_{i=1}^{n_1} L(y_i, (1-\alpha)\widehat{y_{1,(i)}} + \alpha \widehat{y_{2,i}})$.

For linear regression with the squared error loss function, we don't need to perform the leave-one-out procedure to get $\hat{\alpha}_{\text{cv}}$. More specifically, we have

$$\sum_{i=1}^{n_1} L(y_i, (1-\alpha)\widehat{y_{1,(i)}} + \alpha \widehat{y_{2,i}}) = \sum_{i=1}^{n_1} \left(y_i - ((1-\alpha)\widehat{y_{1,(i)}} + \alpha \widehat{y_{2,i}}) \right)^2 \tag{3.14}$$

$$= \sum_{i=1}^{n_1} \left((1-\alpha)(y_i - \widehat{y_{1,(i)}}) + \alpha(y_i - \widehat{y_{2,i}}) \right)^2 \tag{3.15}$$

$$= \sum_{i=1}^{n_1} \left((1-\alpha) \left(\frac{y_i - \widehat{y_{1,i}}}{1 - A_{ii}} \right) + \alpha(y_i - \widehat{y_{2,i}}) \right)^2 \tag{3.16}$$

where $y_i - \widehat{y_{1,(i)}} = \dfrac{y_i - \widehat{y_{1,i}}}{1 - A_{ii}}$ (see e.g. p. 257 in Wood 2017), $\widehat{y_{1,i}}$ is the predicted value for the ith observation with the model fitted on all data (Y_1, X_1), and A_{ii} is the ith diagnal element of the projection matrix $A = X_1(X_1^T X_1)^{-1} X_1^T$. Let $e_{i1} = y_i - \widehat{y_{1,i}}$, $\tilde{e}_{i1} = \dfrac{e_{i1}}{1 - A_{ii}}$,

and $e_{i2} = y_i - \widehat{y_{2,i}}$. A closed form solution can be given as $\hat{\alpha}_{\text{cv}} = \frac{\sum_{i=1}^{n_1} \tilde{e}_{i1}(\tilde{e}_{i1} - e_{i2})}{\sum_{i=1}^{n_1}(\tilde{e}_{i1} - e_{i2})^2}$. A similar argument as that in Wood (2017, page 260) can lead to a more stable GCV estimate if we replace A_{ii} by $tr(A)/n_1 = p_1/n_1$, thus $\hat{\alpha}_{\text{gcv}} = \frac{\sum_{i=1}^{n_1} e_{i1}(e_{i1} - \frac{n_1 - p_1}{n_1} e_{i2})}{\sum_{i=1}^{n_1}(e_{i1} - \frac{n_1 - p_1}{n_1} e_{i2})^2}$.

For logistic regression with the loss function given as the binomial negative log-likelihood as defined in (3.11) on the logistic-transformed probability or equivalently the linear predictor $\hat{y}$, we have

$$\begin{aligned}
&\sum_{i=1}^{n_1} L(y_i, (1-\alpha)\widehat{y_{1,(i)}} + \alpha\widehat{y_{2,i}}) \\
&= -\sum_{i=1}^{n_1} \left[y_i \log\left(\frac{\exp((1-\alpha)\widehat{y_{1,(i)}} + \alpha\widehat{y_{2,i}})}{1 + \exp((1-\alpha)\widehat{y_{1,(i)}} + \alpha\widehat{y_{2,i}})} \right) + (1 - y_i) \log\left(\frac{1}{1 + \exp((1-\alpha)\widehat{y_{1,(i)}} + \alpha\widehat{y_{2,i}})} \right) \right]
\end{aligned}$$

Since $\widehat{y_{1,(i)}} \approx \widehat{y_{1,i}} - (z_i - \widehat{y_{1,i}})A_{ii}/(1 - A_{ii})$, where z_i is the working response for the ith observation in the final stage of the IRLS fitting for logistic regression on full data (Y_1, X_1) and A_{ii} is the ith diagnal element of the corresponding weighted least square fitting projection matrix (Page 262, Wood 2017). $\widehat{y_{1,(i)}}$ can be easily approximated by fitting logistic regression once on the full data. The linear predictor now is $\widehat{y_{1,(i)}} + \alpha(\widehat{y_{2,i}} - \widehat{y_{1,(i)}})$. we can find $\hat{\alpha}_{\text{cv}}$ by fitting a no-intercept logistic regression with off-set given by $\widehat{y_{1,(i)}}$. Following the similar argument for GCV, we can replace $A_{ii}/(1 - A_{ii})$ by $p_1/(n_1 - p_1)$ for a more stable GCV estimate.

Since α is constrained to $[0, 1]$, we set any negative values of $\hat{\alpha}_{\text{cv}}$ to 0, and any values of $\hat{\alpha}_{\text{cv}}$ greater than 1 to 1 in the above procedures for estimating α.

Generally for nonlinear predictive models, such as neural networks, GBM or SVM etc., there is no short cut to get the CV estimation as for linear regression and logistic regression cases. However, for random forest, if we can extract the out-of-bag (OOB) prediction for each observation in the one time fitting on the full data (Y_1, X_1), then we can use the OOB prediction for $\widehat{y_{1,(i)}}$.

3.4 Theory

In this section, we show that under natural conditions, for the linear regression model, our model combination method is asymptotically more accurate than the full-data model.

The following Theorems are proved in Section 3.4.1:

Theorem 3.4.1. *Under the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X \sim N(X\beta,\sigma_\epsilon^2)$, where β is a $p_1 \times 1$ regression coefficient vector. Suppose the theoretical optimal predictor from the full data is $f_1(x) = x\beta$ and let $f_2(x) = x\beta_2$ be the theoretical optimal predictor subject to the constraint that the ith element of β_2: $(\beta_2)_i = 0$ for all $i > p_2$. Let the least square estimators be $\widehat{f}_1(x) = x^T\widehat{\boldsymbol{\beta_1}}$ and $\widehat{f}_2(x) = x^T\widehat{\boldsymbol{\beta_2}}$, where the ith element of $\widehat{\beta}_2$: $(\widehat{\beta}_2)_i = 0$ for $i > p_2, (p_2 < p_1)$. Let $\alpha^* = \dfrac{A}{A+B}$ where $A = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1}\right)$ and $B = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2}\right)$ and C_n be a sequence of constants, bounded below, satisfying $C_n = o(n)$. Let $\widehat{\alpha}^* = \dfrac{\widehat{A}}{\widehat{A}+\widehat{B}}$ where $\hat{A} = \hat{\sigma}_\epsilon^2\left(\dfrac{p_1}{n_1} + \dfrac{p_1^2+p_1}{n_1^2}\right)$, $\hat{B} = \left(\widehat{\sigma_\eta^2} - \widehat{\sigma_\epsilon^2}\right) + \widehat{\sigma_\eta^2}\left(\dfrac{p_2}{n_2} + \dfrac{p_2^2+p_2}{n_2^2}\right)$, $\widehat{\sigma_\epsilon^2} = \dfrac{1}{n_1-p_1}\sum_{i=1}^{n_1}(\mathbf{y}_i - \widehat{f}_1(\mathbf{x}_i))^2$ and $\widehat{\sigma_\eta^2} = \dfrac{1}{n_2-p_2}\sum_{i=1+n_1}^{n}(\mathbf{y}_i - \widehat{f}_2(\mathbf{x}_i))^2$. With fixed p_1 and p_2 and $\sigma_\eta^2 - \sigma_\epsilon^2 > 0$, we have*

$$
\begin{aligned}
&(1) \quad 0 \le \alpha^* = \frac{A}{A+B} < 1 \\
&(2) \quad \lim_{n_1\to\infty}\alpha^* = 0, \quad \alpha^* = O({n_1}^{-1}) \\
&(3) \quad \mathbb{E}\left((\widehat{\alpha}^* - \alpha^*)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})
\end{aligned}
$$

From Theorem 3.4.1 (3), it is easily seen that the convergence rate of $\widehat{\alpha}^* - \alpha^* \to 0$ is $O_p(n_1^{-3/2}) + O_p(n_1^{-1}n_2^{-1/2})$, which is faster than the convergence rate of $\alpha^* \to 0$. The convergence rate of $\alpha^* \to 0$ is $O({n_1}^{-1})$ which is faster than the convergence rate of $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ which is $O_p({n_1}^{-1/2})$. It is also easily seen that the convergence rate of $\widehat{\alpha}^* \to 0$ is also $O_p({n_1}^{-1})$.

Now we are ready to prove the risk of the model combination method with our plug-in estimator of α^* is asymptotically better than the risk of full model which is equivalent to $\alpha^* = 0$. Recall the expected loss for a constant α is defined as

$$
\begin{aligned}
R(\alpha) &= \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0) + \alpha\widehat{f}_2(\mathbf{x}_0)\right) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1\wedge n_2}\right] \\
&= (1-\alpha)^2 A + \alpha^2 B + o((n_1 \wedge n_2)^{-2})
\end{aligned}
$$

Theorem 3.4.2. *Under the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X \sim N(X\beta,\sigma_\epsilon^2)$, where β is a $p_1 \times 1$ regression coefficient vector. Let the least square estimators $\widehat{f}_1(x) = x^T\widehat{\boldsymbol{\beta_1}}$ and $\widehat{f}_2(x) = x^T\widehat{\boldsymbol{\beta_2}}$,*

where the ith element of $\widehat{\beta}_2$*:* $(\widehat{\beta}_2)_i = 0$ *for* $i > p_2, (p_2 < p_1)$*. Let* $\widehat{\alpha}^* = \dfrac{\widehat{A}}{\widehat{A}+\widehat{B}}$ *where* $\hat{A} = \hat{\sigma}_\epsilon^2\left(\dfrac{p_1}{n_1} + \dfrac{p_1^2+p_1}{n_1^2}\right)$*,* $\hat{B} = \left(\widehat{\sigma_\eta^2} - \widehat{\sigma_\epsilon^2}\right) + \widehat{\sigma_\eta^2}\left(\dfrac{p_2}{n_2} + \dfrac{p_2^2+p_2}{n_2^2}\right)$*,* $\widehat{\sigma_\epsilon^2} = \dfrac{1}{n_1 - p_1}\sum_{i=1}^{n_1}(\mathbf{y}_i - \widehat{f}_1(\mathbf{x}_i))^2$ *and* $\widehat{\sigma_\eta^2} = \dfrac{1}{n_2 - p_2}\sum_{i=1+n_1}^{n}(\mathbf{y}_i - \widehat{f}_2(\mathbf{x}_i))^2$*. With fixed* p_1 *and* p_2 *and* $\sigma_\eta^2 - \sigma_\epsilon^2 > 0$*, for sufficient large* n_1 *and* n_2 *and* $n_2 > n_1$*, we have*

$$R_{\text{plugin}} = \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\widehat{\alpha}^*)\widehat{f}_1(\mathbf{x}_0)) + \widehat{\alpha}^*\widehat{f}_2(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1 \wedge n_2}\right] < R(0).$$

3.4.1 Proof of Theorems for Linear Regression Case

Lemma 3.4.3. *If* X_1 *and* X_2 *are i.i.d. standard normally distributed, then the moment generating function of the product* X_1X_2 *is*

$$M_{X_1X_2}(t) = \left(1 - t^2\right)^{-\frac{1}{2}}$$

Proof.

$$\begin{aligned} M_{X_1X_2}(t) = E(e^{tX_1X_2}) = E_{X_1}(E_{X_2|X_1}(e^{tX_1X_2})) = E_{X_1}(M_{X_2}(tX_1)) = E_{X_1}\left(e^{\frac{t^2X_1^2}{2}}\right) \\ = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{t^2x^2}{2} - \frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2(1-t^2)^{-1}}}\,dx = \left(1-t^2\right)^{-\frac{1}{2}} \end{aligned}$$

□

Lemma 3.4.4. *If* $\sum_{i=1}^{p} s_i = \dfrac{1}{2}$ *and* $s_i > 0$ *for all* i*, then* $\prod_{i=1}^{p}(1-s_i) \geqslant \dfrac{1}{2}$*.*

Proof.

$$\begin{aligned} \prod_{i=1}^{p}(1-s_i) &= 1 - \sum_{i=1}^{p} s_i + \sum_{i\neq j} s_is_j - \sum_{i,j,k \text{ distinct}} s_is_js_k + \cdots \\ &= 1 - \frac{1}{2} + \sum_{i\neq j} s_is_j\left(1 - \sum_{k\neq i,j} s_k\right) + \cdots \\ &\geqslant \frac{1}{2} \end{aligned}$$

□

Lemma 3.4.5. *If Y follows the linear model $Y = X\beta + E$, where E are i.i.d. normal variables with mean 0 and variance σ^2, and $\mathbf{x_0}$ follows a random multivariate normal distribution with mean 0 and variance Σ, then for fixed X, and a constant*

$$C > \sigma^2 \operatorname{tr}((X^TX)^{-1}\Sigma)$$

$$E_{\mathbf{x}_0}E_{Y|X}\left(\left(\left[\mathbf{x}_0^T(\widehat{\boldsymbol{\beta}}-\beta)(\widehat{\boldsymbol{\beta}}-\beta)^T\mathbf{x}_0\right]-C\right)_+\right) \leqslant 8(1+\sqrt{2})Ce^{-\sqrt{\frac{C}{2\sigma^2\operatorname{tr}((X^TX)^{-1}\Sigma)}}}$$

Proof. Let $Z = X\Sigma^{-\frac{1}{2}}$ and $\beta_Z = \Sigma^{\frac{1}{2}}\beta$. Then our linear model is $Y = Z\beta_Z + E$, and we have $\operatorname{tr}\left((X^TX)^{-1}\Sigma\right) = \operatorname{tr}\left(\Sigma^{\frac{1}{2}}(X^TX)^{-1}\Sigma^{\frac{1}{2}}\right) = \operatorname{tr}\left((Z^TZ)^{-1}\right)$. Let $\mathbf{z_0} = \Sigma^{-\frac{1}{2}}\mathbf{x_0}$, so $\operatorname{Var}(\mathbf{z_0}) = \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = I$.

Let $M = (Z^TZ)^{-1}Z^T$. We have $\widehat{\boldsymbol{\beta_Z}} = MY$ and $E_{(Y|Z)}\widehat{\boldsymbol{\beta_Z}} = \beta_Z$. Thus we want to bound

$$E_{\mathbf{x}_0}E_{Y|Z}\left(\left(\left[\mathbf{z}_0^T(MY-\beta_Z)(MY-\beta_Z)^T\mathbf{z}_0\right]-C\right)_+\right) = E_{\mathbf{z}_0}E_{Y|Z}\left(\left((\mathbf{z}_0^TME)^2-C\right)_+\right)$$

E and $\mathbf{z_0}$ are independent, $E \sim N(0,\sigma^2I_n)$, and $\mathbf{z_0} \sim N(0,I_p)$. Let $M = UDV^T$ be the singular value decomposition of M. Since U and V are orthogonal matrices, we have that $\mathbf{u} = U^T\mathbf{z_0} \sim N(0,I_p)$ and $\mathbf{v} = \sigma^{-1}V^TE \sim N(0,I_n)$. Thus we want to bound

$$E_{\mathbf{u}}E_{\mathbf{v}}(((\sigma\mathbf{u}^TD\mathbf{v})^2-C)_+) = \int_C^\infty P\left((\sigma\mathbf{u}^TD\mathbf{v})^2 > a\right)\,da = \int_C^\infty 2P\left(\mathbf{u}^TD\mathbf{v} > \frac{\sqrt{a}}{\sigma}\right)\,da$$

Where we have used the fact that the distribution of $\mathbf{u}^TD\mathbf{v}$ is symmetric about 0, so that $P\left(\mathbf{u}^TD\mathbf{v} < -\frac{\sqrt{a}}{\sigma}\right) = P\left(\mathbf{u}^TD\mathbf{v} > \frac{\sqrt{a}}{\sigma}\right)$.

Now $\mathbf{u}^TD\mathbf{v} = \sum_{i=1}^p d_iu_iv_i$, where u_i and v_i are independent standard normal random variables. Therefore, the moment generating function of $\mathbf{u}^TD\mathbf{v}$ is

$$M(t) = \prod_{i=1}^p\left(1-d_i^2t^2\right)^{-\frac{1}{2}}$$

The Chernoff bound therefore gives, for $t > 0$,

$$P\left(\mathbf{u}^TD\mathbf{v} > \frac{\sqrt{a}}{\sigma}\right) \leqslant M(t)e^{-t\frac{\sqrt{a}}{\sigma}} = \frac{e^{-\sqrt{\frac{a}{2\sigma^2\sum_{i=1}^p d_i^2}}}}{\sqrt{\prod_{i=1}^p\left(1-\frac{d_i^2}{2\sum_{i=1}^p d_i^2}\right)}}$$

where we have set $t = \left(2\sum_{i=1}^p d_i^2\right)^{-1/2}$. Thus,

$$P\left(\mathbf{u}^TD\mathbf{v} > \frac{\sqrt{a}}{\sigma}\right) \leqslant \sqrt{2}e^{-\sqrt{\frac{a}{2\sigma^2\sum_{i=1}^p d_i^2}}}$$

This gives

$$\begin{aligned}
E_{\mathbf{u}}E_{\mathbf{v}}(((\mathbf{u}^T D\mathbf{v})^2 - C)_+) &= \int_C^\infty 2P\left(\mathbf{u}^T D\mathbf{v} > \frac{\sqrt{a}}{\sigma}\right) da \\
&\leqslant \int_C^\infty 2\sqrt{2}e^{-\sqrt{\frac{a}{2\sigma^2\sum_{i=1}^p d_i^2}}}\, da \\
&= 2\sqrt{2}\int_{\sqrt{C}}^\infty 2re^{-\frac{r}{\sqrt{2\sigma^2\sum_{i=1}^p d_i^2}}}\, dr \\
&= 4\sqrt{2}\left(\left[-r\sigma\sqrt{2\sum_{i=1}^p d_i^2}\, e^{-\frac{r}{\sigma\sqrt{2\sum_{i=1}^p d_i^2}}}\right]_{\sqrt{C}}^\infty + \sigma\sqrt{2\sum_{i=1}^p d_i^2}\int_{\sqrt{C}}^\infty e^{-\frac{r}{\sigma\sqrt{2\sum_{i=1}^p d_i^2}}}\, dr\right) \\
&= 4\sqrt{2}\left(\sigma\sqrt{2C\sum_{i=1}^p d_i^2}\, e^{-\sqrt{\frac{C}{2\sigma^2\sum_{i=1}^p d_i^2}}} + 2\sigma^2\left(\sum_{i=1}^p d_i^2\right)e^{-\sqrt{\frac{C}{2\sigma^2\sum_{i=1}^p d_i^2}}}\right) \\
&\leqslant 8(1+\sqrt{2})Ce^{-\sqrt{\frac{C}{2\sigma^2\sum_{i=1}^p d_i^2}}}
\end{aligned}$$

The last inequality is obtained by substituting $\sum_i d_i^2 = \mathrm{tr}((Z^TZ)^{-1}) = \mathrm{tr}((X^TX)^{-1}\Sigma)$.

□

Lemma 3.4.6. *For any $a > 0$ and n satisfying $an > 1/2 > a^2$*

(i) If $nX \sim \chi_n^2$, then $P(|X-1| > a) \leqslant 2e^{-\frac{a^2 n}{8}}$.

(ii) If X is the mean of n products of pairs of independent standard normal distributions, i.e. $X = \frac{1}{n}\sum_{i=1}^n (Z_{i1}Z_{i2})$ where Z_{ij} are all i.i.d. standard Normal, then $P(|X| > a) \leqslant 2e^{-\frac{a^2 n}{4}}$

Proof. (i) The moment-generating function of the chi-square distribution with n degrees of freedom is $M(t) = (1-2t)^{-\frac{n}{2}}$, so X has moment generating function $M(t) = \left(1 - 2\frac{t}{n}\right)^{-\frac{n}{2}}$. We use the Chernoff bounds

$$\begin{aligned}
P(X \geqslant 1+a) &\leqslant M_x(t)e^{-(1+a)t} && \text{for } t = \frac{an}{4} > 0 \\
P(X \leqslant 1-a) &\leqslant M_x(t)e^{-(1-a)t} && \text{for } t = -\frac{an}{4} < 0
\end{aligned}$$

We have that $-\frac{\log(1-x)}{x} - x$ is continuous and has only one local minimum in the interval $\left[0, \frac{1}{2}\right]$, so in particular, $-\frac{\log(1-x)}{x} \leqslant 1+x$ or equivalently $(1-x)^{-\frac{1}{x}} \leqslant e^{1+x}$ for $0 \leqslant x \leqslant \frac{1}{2}$. Therefore,

$$M\left(\frac{an}{4}\right) = \left(1-\frac{a}{2}\right)^{-\frac{n}{2}} = \left(\left(1-\frac{a}{2}\right)^{-\frac{2}{a}}\right)^{\frac{an}{4}} \leqslant e^{\left(1+\frac{a}{2}\right)\frac{an}{4}}$$

Substituting this in the Chernoff bound gives

$$P\left(X \geqslant 1+a\right) \leqslant e^{\left(1+\frac{a}{2}\right)\frac{an}{4}} e^{-(1+a)\frac{an}{4}} = e^{-\frac{a^2 n}{8}}$$

Similarly, setting $t = -\dfrac{an}{4}$, $\dfrac{\log(1+x)}{x} \geqslant 1-x$, so $(1+x)^{\frac{1}{x}} \geqslant e^{1-x}$ for $0 \leqslant x \leqslant \dfrac{1}{2}$, meaning

$$M\left(-\frac{an}{4}\right) = \left(1+\frac{a}{2}\right)^{-\frac{n}{2}} = \left(\left(1+\frac{a}{2}\right)^{\frac{2}{a}}\right)^{-\frac{an}{4}} \leqslant e^{-\left(1-\frac{a}{2}\right)\frac{an}{4}}$$

Thus, the Chernoff bound gives

$$P\left(X \leqslant 1-a\right) \leqslant e^{-\left(1-\frac{a}{2}\right)\frac{an}{4}} e^{(1-a)\frac{an}{4}} = e^{-\frac{a^2 n}{8}}$$

(ii) By Lemma 3.4.3, the moment generating function for the average of n products of i.i.d. standard normal distributions is $M(t) = \left(1-\dfrac{t^2}{n^2}\right)^{-\frac{n}{2}}$ setting $t = an$, gives

$$M(t) = \left(1-a^2\right)^{-\frac{n}{2}} \leqslant e^{\left(1+a^2\right)\frac{a^2 n}{2}}$$

so the Chernoff bound gives

$$P\left(X \geqslant a\right) \leqslant e^{\left(1+a^2\right)\frac{a^2 n}{2}} e^{-a^2 n} = e^{-\frac{a^2 n}{2}\left(1-a^2\right)} \leqslant e^{-\frac{a^2 n}{4}}$$

Since the product of independent standard normal distributions is symmetric, this gives the results.

□

Lemma 3.4.7. *If X is an $n \times p$ matrix whose rows are distributed according to i.i.d. multivariate normal distributions with mean 0 and invertible covariance matrix Σ, then*

$$P\left(\left\|\Sigma^{-\frac{1}{2}} \frac{X^T X}{n} \Sigma^{-\frac{1}{2}} - I\right\|_\infty > A\right) \leqslant 2p^2 e^{-\frac{A^2 n}{8}}$$

Proof. Let $Z = X\Sigma^{-\frac{1}{2}}$. The elements of Z are i.i.d. standard normal, and we want to bound

$$P\left(\left\|\frac{Z^T Z}{n} - I\right\|_\infty > A\right) \leqslant \sum_{i,j=1}^{p} P\left(\left|\left(\frac{Z^T Z}{n} - I\right)_{ij}\right| > A\right)$$

For $i = j$, we have that $\left(\dfrac{Z^T Z}{n} - I\right)_{ij} = \dfrac{1}{n}\sum_{k=1}^{n} Z_{ki}^2 - 1$, which is a scaled, centred chi-square distribution with n degrees of freedom, so by Lemma 3.4.6(i), we have

$$P\left(\left|\left(\frac{Z^T Z}{n} - I\right)_{ij}\right| > A\right) \leqslant 2e^{-\frac{A^2 n}{8}}$$

For $i \neq j$, we have $\left(\frac{Z^TZ}{n} - I\right)_{ij} = \frac{1}{n}\sum_{k=1}^{n} Z_{ki}Z_{kj}$ is a mean of n i.i.d. products of independent standard normal distributions. By Lemma 3.4.6(ii), we have

$$P\left(\left|\left(\frac{Z^TZ}{n} - I\right)_{ij}\right| > A\right) \leqslant 2e^{-\frac{A^2n}{4}}$$

Putting these together gives

$$P\left(\left\|\Sigma^{-\frac{1}{2}}\frac{X^TX}{n}\Sigma^{-\frac{1}{2}} - I\right\|_\infty > A\right) \leqslant 2p^2e^{-\frac{A^2n}{8}}$$

□

Lemma 3.4.8. *For a $p \times p$ matrix M, if $\|M - I\|_\infty \leqslant c$ for a constant $c < \frac{1}{p}$, then* $\mathrm{tr}(M^{-1}) \leqslant p - 1 + \frac{1}{1-pc}$

Proof. Let $Q = M - I$. Since $\|Q\|_\infty \leqslant c$, we have that $\|Q^k\|_\infty \leqslant p^{k-1}c^k$, and we have $M^{-1} = (I+Q)^{-1} = I - Q + Q^2 - Q^3 + \cdots$, so in particular

$$\mathrm{tr}(M^{-1}) = \mathrm{tr}(I - Q + Q^2 - Q^3 + \cdots) \leqslant p(1 + c + pc^2 + \cdots) = p - 1 + \frac{1}{1-pc}$$

□

Theorem 3.4.9. *Under the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X \sim N(X\beta, \sigma_\epsilon^2)$, where β is a $p_1 \times 1$ regression coefficient vector. Suppose the theoretical optimal predictor from the full data is $f_1(x) = x\beta$ and let $f_2(x) = x\beta_2$ be the theoretical optimal predictor subject to the constraint that the ith element of β_2: $(\beta_2)_i = 0$ for all $i > p_2$. Let the least square estimators $\widehat{f_1}(x) = x^T\widehat{\boldsymbol{\beta_1}}$ and $\widehat{f_2}(x) = x^T\widehat{\boldsymbol{\beta_2}}$, where the ith element of $\widehat{\beta}_2$: $(\widehat{\beta}_2)_i = 0$ for $i > p_2, (p_2 < p_1)$.*

As the sample sizes n_1 and n_2 tend to ∞,

$$A = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f_1}(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1}\right) = \sigma_\epsilon^2\left(\frac{p_1}{n_1} + \frac{p_1^2 + p_1}{n_1^2} + o\Big(\frac{1}{n_1^2}\Big)\right), \tag{3.17}$$

$$B = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f_2}(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2}\right) = \left(\sigma_\eta^2 - \sigma_\epsilon^2\right) + \sigma_\eta^2\left(\frac{p_2}{n_2} + \frac{p_2^2 + p_2}{n_2^2} + o\Big(\frac{1}{n_2^2}\Big)\right), \tag{3.18}$$

where C_n is a sequence of constants, bounded below, satisfying $C_n = o(n)$, and σ_η^2 is the theoretical minimum MSE for a linear predictor using only the first p_2 predictors. That is, $\sigma_\eta^2 - \sigma_\epsilon^2 = (\boldsymbol{\beta} - \boldsymbol{\beta_2})^T\mathbb{E}(\mathbf{x_0}\mathbf{x_0}^T)(\boldsymbol{\beta} - \boldsymbol{\beta_2})$.

Proof. From standard linear regression theory, we know that the conditional expected MSE is

$$\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{Y|X}\left(\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2\right)=\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)$$

The first step to proving (3.17) is to show that

$$\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2\wedge C_{n_1}\right)=\mathbb{E}_X\left(\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)\wedge C_{n_1}\right)+o(n_1^{-2}) \tag{3.19}$$

We do this by dividing into two cases:

Case 1 $\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty<\frac{n_1^{-0.4}}{p_1}$,

Case 2 $\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty\geqslant\frac{n_1^{-0.4}}{p_1}$.

It is convenient to define a random variable Z as an indicator for Case 1. That is $Z=1$ in Case 1 and $Z=0$ in Case 2.

By Lemma 3.4.7, Case 2 has probability bounded by $2p_1^2e^{-\frac{n_1^{-0.8}n_1}{8p_1^2}}=o(n_1^{-2}C_{n_1}^{-1})$. Since the loss is bounded by C_{n_1}, it follows that the contribution of this case to the total expectation is $o(n_1^{-2})$. Therefore, we only need to prove (3.19) in Case 1.

From the general theory of bounded random variables:

$$E(X\wedge c)=E(X)-E((X-c)_+)$$

we see that it is sufficient to prove that in Case 1,

$$E_{\mathbf{x}_0}E_{Y|X}\left(\left(\left(\widehat{f}_1(\mathbf{x}_0)-E_{(X,Y)}\widehat{f}_1(\mathbf{x}_0)\right)^2-C_{n_1}\right)_+\middle|Z=1\right)=o(n_1^{-2})$$

By Lemma 3.4.8, Case 1 implies

$$\mathrm{tr}\left((X^TX)^{-1}\Sigma\right)=\mathrm{tr}\left(\Sigma^{\frac{1}{2}}(X^TX)^{-1}\Sigma^{\frac{1}{2}}\right)<\frac{p_1-1+\frac{1}{1-n_1^{-0.4}}}{n_1}<\frac{p_1+1}{n_1}$$

so

$$\mathrm{tr}((X^TX)^{-1}\Sigma)<\frac{C_{n_1}}{\sigma_\epsilon^2}$$

Thus, by Lemma 3.4.5:

$$E_{\mathbf{x}_0}E_{Y|X}\left(\left(\left[\mathbf{x}_0^T(\widehat{\boldsymbol{\beta}}-E_{(X,Y)}\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}-E_{(X,Y)}\widehat{\boldsymbol{\beta}})^T\mathbf{x}_0\right]-C_{n_1}\right)_+\middle|Z=1\right)\leqslant 8(1+\sqrt{2})C_{n_1}e^{-\sqrt{\frac{C_{n_1}}{\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)}}}$$
$$\leqslant 8(1+\sqrt{2})C_{n_1}e^{-\sqrt{\frac{C_{n_1}n_1}{\sigma_\epsilon^2(p_1+1)}}}$$
$$=o(n_1^{-2})$$

This proves (3.19). We now need to show that

$$\mathbb{E}_X\left(\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)\wedge C_{n_1}\right)=\sigma_\epsilon^2\left(\frac{p_1}{n_1}+\frac{p_1^2+p_1}{n_1^2}+o\Big(\frac{1}{n_1^2}\Big)\right)$$

We have shown that the probability of Case 2 is $P(Z=0)=o\left(n_1^{-2}C_{n_1}^{-1}\right)$, so we may assume Case 1, in which case, $\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)<C_{n_1}$. Let $Q=\dfrac{\Sigma^{-\frac{1}{2}}(X^TX)\Sigma^{-\frac{1}{2}}}{n_1}-I$. Case 1 assumes that $\|Q\|_\infty<\dfrac{n_1^{-0.4}}{p_1}$. It follows that $\dfrac{\Sigma^{-\frac{1}{2}}(X^TX)\Sigma^{-\frac{1}{2}}}{n_1}=I+Q$ is invertible with inverse $I-Q+Q^2-\cdots$. Therefore

$$\mathrm{tr}(\Sigma^{\frac{1}{2}}(X^TX)^{-1}\Sigma^{\frac{1}{2}})=n_1^{-1}\,\mathrm{tr}((I+Q)^{-1})=n_1^{-1}\,\mathrm{tr}(I-Q+Q^2-Q^3+\cdots)=\frac{p_1}{n_1}-\frac{\mathrm{tr}(Q)-\mathrm{tr}(Q^2)}{n_1}+o(n_1^{-2}).$$

We now have $\mathbb{E}_{X|Z=1}\left(\sigma_\epsilon^2\,\mathrm{tr}((X^TX)^{-1}\Sigma)\wedge C_{n_1}\right)=\sigma_\epsilon^2\mathbb{E}_{X|Z=1}\left(\dfrac{p_1}{n_1}-\dfrac{\mathrm{tr}(Q)-\mathrm{tr}(Q^2)}{n_1}+o(n_1^{-2})\right)$.
Thus we need to show $\mathbb{E}_{X|Z=1}(-\mathrm{tr}(Q)+\mathrm{tr}(Q^2))=\dfrac{p_1^2+p_1}{n_1}+o(n_1^{-1})$.

For the first term, we have $\mathbb{E}_X(Q)=\mathbb{E}_{X|Z=1}(Q)P(Z=1)+\mathbb{E}_{X|Z=0}(Q)P(Z=0)=0$ and $P(Z=0)=o\left(n_1^{-2}C_{n_1}^{-1}\right)$, thus $\mathbb{E}_{X|Z=1}tr(Q)=o(n_1^{-2})$.

Similarly for the second term we have $\mathbb{E}_X(Q^2)=\mathbb{E}_{X|Z=1}(Q^2)P(Z=1)+\mathbb{E}_{X|Z=0}(Q^2)P(Z=0)=\mathbb{E}_{X|Z=1}(Q^2)+o(n_1^{-2})$.

By symmetry of the Q matrix, $\mathrm{tr}(Q^2)=\sum_{i,j}(Q_{ij}^2)$. Since the rows of $X\Sigma^{-\frac{1}{2}}$ are i.i.d. vectors, each follows $N(0,I)$, we have for $i=j$, Q_{ij} follows a centred scaled chi-square distribution, and so has variance $\dfrac{2}{n_1}$. For $i\neq j$, Q_{ij} is a mean of n_1 independent products of two standard normal random variables, so has variance $\dfrac{1}{n_1}$. Thus

$$\mathbb{E}_X tr(Q^2)=\sum_{i,j}\mathrm{Var}(Q_{ij})=p_1\frac{2}{n_1}+p_1(p_1-1)\frac{1}{n_1}=\frac{p_1^2+p_1}{n_1}$$

This proves (3.17).

The proof of (3.18) is similar. We again have two cases

Case 1 $\left\|\Sigma_{11}^{-\frac{1}{2}}\left(\frac{X_{21}^T X_{21}}{n_2}\right)\Sigma_{11}^{-\frac{1}{2}} - I\right\|_\infty < \frac{n_2^{-0.4}}{p_2}$, denoted by $Z = 1$.

Case 2 $\left\|\Sigma_{11}^{-\frac{1}{2}}\left(\frac{X_{21}^T X_{21}}{n_2}\right)\Sigma_{11}^{-\frac{1}{2}} - I\right\|_\infty \geqslant \frac{n_2^{-0.4}}{p_2}$, denoted by $Z = 0$.

where Σ_{11} is the $p_2 \times p_2$ block of the covariance matrix corresponding to the variables in X_{21},

Since $P(Z = 0) = o(n_2^{-2}C_{n_2}^{-1})$, the contribution of Case 2 to the total expectation is $o(n_2^{-2})$. Thus

$$\begin{aligned}
B &= \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2}\right) \\
&= \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2} | Z = 1\right) + o(n_2^{-2}) \\
&= \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 | Z = 1\right) \\
&- \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 - C_{n_2}\right)_+ | Z = 1\right) + o(n_2^{-2})
\end{aligned}$$

The same proof shows that the second term of the above is $o(n_2^{-2})$. The first term above is equal to $\mathbb{E}_{X_{21}}\left(\left(\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{Y_2|X_{21}}\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2\right) \wedge C_{n_2}\right)$. Since $\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{Y_2|X_{21}}\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 = \sigma_\eta^2 \operatorname{tr}((X_{21}^T X_{21})^{-1}\Sigma_{11}) + (\sigma_\eta^2 - \sigma_\epsilon^2)$, we have

$$B = \mathbb{E}_{X_{21}}\left(\sigma_\eta^2 \operatorname{tr}((X_{21}^T X_{21})^{-1}\Sigma_{11}) \wedge (C_{n_2} - (\sigma_\eta^2 - \sigma_\epsilon^2))\right) + (\sigma_\eta^2 - \sigma_\epsilon^2) + o(n_2^{-2})$$

and a similar proof leads to

$$\mathbb{E}_{X_{21}}\left(\sigma_\eta^2 \operatorname{tr}((X_{21}^T X_{21})^{-1}\Sigma_{11}) \wedge (C_{n_2} - (\sigma_\eta^2 - \sigma_\epsilon^2))\right) = \sigma_\eta^2\left(\frac{p_2}{n_2} + \frac{p_2^2 + p_2}{n_2^2}\right) + o(n_2^{-2})$$

which completes the proof. □

Corollary 3.4.9.1. *Under the same model assumptions as in* ***Theorem 3.1.1****, for any fixed value of* α*, we have the risk*

$$R(\alpha) = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0) + \alpha\widehat{f}_2(\mathbf{x}_0)\right) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1 \wedge n_2}\right] \tag{3.20}$$

$$= (1-\alpha)^2 A + \alpha^2 B + o((n_1 \wedge n_2)^{-2}) \tag{3.21}$$

Where $C_{n_1 \wedge n_2}$ *is a sequence of constants, bounded below, satisfying* $C_{n_1 \wedge n_2} = o(n_1 \wedge n_2)$ *and* A *and* B *are given by (3.7) and (3.8).*

Proof. Similar to the proof for **Theorem 3.1.1**, we define an indicator variable as

Case 1 $\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty < \frac{n_1^{-0.4}}{p_1}$ and $\left\|\Sigma_{11}^{-\frac{1}{2}}\left(\frac{X_{21}^TX_{21}}{n_2}\right)\Sigma_{11}^{-\frac{1}{2}}-I\right\|_\infty < \frac{n_2^{-0.4}}{p_2}$, denoted by $Z=1$.

Case 2 $\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty \geqslant \frac{n_1^{-0.4}}{p_1}$ or $\left\|\Sigma_{11}^{-\frac{1}{2}}\left(\frac{X_{21}^TX_{21}}{n_2}\right)\Sigma_{11}^{-\frac{1}{2}}-I\right\|_\infty \geqslant \frac{n_2^{-0.4}}{p_2}$, denoted by $Z=0$.

From the proof for **Theorem 3.1.1**, we have $P(Z=0) \leq o(n_1^{-2}C_{n_1}^{-1}) + o(n_2^{-2}C_{n_2}^{-1})$, thus the contribution of Case 2 to the total expectation is $o(n_1^{-2})+o(n_2^{-2}) = o((n_1\wedge n_2)^{-2})$. Thus

$$
\begin{aligned}
R(\alpha) =&\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0)+\alpha\widehat{f}_2(\mathbf{x}_0)\right)-f(\mathbf{x}_0)\right)^2\wedge C_{n_1\wedge n_2}\right]\\
=&\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0)+\alpha\widehat{f}_2(\mathbf{x}_0)\right)-f(\mathbf{x}_0)\right)^2\wedge C_{n_1\wedge n_2}|Z=1\right]+o((n_1\wedge n_2)^{-2})\\
=&\mathbb{E}_X\left(\left(\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(Y|X)}\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0)+\alpha\widehat{f}_2(\mathbf{x}_0)\right)-f(\mathbf{x}_0)\right)^2\right)\wedge C_{n_1\wedge n_2}\right)+o((n_1\wedge n_2)^{-2})\\
=&\mathbb{E}_X\left(\left((1-\alpha)^2\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(Y|X)}\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)+\alpha^2\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(Y|X)}\left(\widehat{f}_2(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2\right)\wedge C_{n_1\wedge n_2}\right)\\
&+o((n_1\wedge n_2)^{-2})\\
=&(1-\alpha)^2A+\alpha^2B+o((n_1\wedge n_2)^{-2})
\end{aligned}
$$

The fourth line is because $\widehat{f}_1(\mathbf{x_0})-f(\mathbf{x_0})$ and $\widehat{f}_2(\mathbf{x_0})-f(\mathbf{x_0})$ are conditionally independent given X, because $\widehat{f}_1$ and $\widehat{f}_2$ are estimated from independent data sets. The last equality is true because for any non-negative random variables W_1 and W_2 and any constant C, we have

$$\mathbb{E}\left(W_1\wedge\frac{C}{2}+W_2\wedge\frac{C}{2}\right)\leq\mathbb{E}((W_1+W_2)\wedge C)\leq\mathbb{E}(W_1\wedge C+W_2\wedge C)$$

Setting $W_1=(1-\alpha)^2\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(Y|X)}\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)$ and $W_2=\alpha^2\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(Y|X)}\left(\widehat{f}_2(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2$ the proof of Theorem 3.1.1 gives that both sides of this inequality are equal to $(1-\alpha)^2A+\alpha^2B+o((n_1\wedge n_2)^{-2})$.

□

Theorem 3.4.10. *Assume the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X\sim N(X\beta,\sigma_\epsilon^2)$, where β is a $p_1\times 1$ regression coefficient vector. Suppose the theoretical optimal predictor from the*

full data is $f_1(x) = x\beta$ *and let* $f_2(x) = x\beta_2$ *be the theoretical optimal predictor subject to the constraint that the* i*th element of* β_2*:* $(\beta_2)_i = 0$ *for all* $i > p_2$*. Let the least square estimators* $\widehat{f}_1(x) = x^T\widehat{\boldsymbol{\beta_1}}$ *and* $\widehat{f}_2(x) = x^T\widehat{\boldsymbol{\beta_2}}$*, where the* i*th element of* $\widehat{\beta}_2$*:* $(\widehat{\beta}_2)_i = 0$ *for* $i > p_2, (p_2 < p_1)$*. Let* $\alpha^* = \dfrac{A}{A+B}$ *where* $A = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1}\right)$ *and* $B = \mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_2}\right)$ *and* C_n *be a sequence of constants, bounded below, satisfying* $C_n = o(n)$*. Let* $\widehat{\alpha}^* = \dfrac{\widehat{A}}{\widehat{A}+\widehat{B}}$ *where* $\hat{A} = \hat{\sigma}_\epsilon^2\left(\dfrac{p_1}{n_1} + \dfrac{p_1^2+p_1}{n_1^2}\right)$*,* $\hat{B} = \left(\widehat{\sigma_\eta^2} - \widehat{\sigma_\epsilon^2}\right) + \widehat{\sigma_\eta^2}\left(\dfrac{p_2}{n_2} + \dfrac{p_2^2+p_2}{n_2^2}\right)$*,* $\widehat{\sigma_\epsilon^2} = \dfrac{1}{n_1 - p_1}\sum_{i=1}^{n_1}(\mathbf{y}_i - \widehat{f}_1(\mathbf{x}_i))^2$ *and* $\widehat{\sigma_\eta^2} = \dfrac{1}{n_2 - p_2}\sum_{i=1+n_1}^{n}(\mathbf{y}_i - \widehat{f}_2(\mathbf{x}_i))^2$*. With fixed* p_1 *and* p_2 *and* $\sigma_\eta^2 - \sigma_\epsilon^2 > 0$*, we have*

$$\begin{aligned}
&(1) \quad 0 \le \alpha^* = \frac{A}{A+B} < 1\\
&(2) \quad \lim_{n_1\to\infty}\alpha^* = 0, \quad \alpha^* = O({n_1}^{-1})\\
&(3) \quad \mathbb{E}\left((\widehat{\alpha}^* - \alpha^*)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})
\end{aligned}$$

Proof. (1) By definition, $A \ge 0$ and $B > 0$, so $0 \le \alpha^* = \dfrac{A}{A+B} < 1$
(2) By Theorem 3.1.1, $\lim\limits_{n_1\to\infty} A = 0$ and $B \ge \sigma_\eta^2 - \sigma_\epsilon^2$, thus $\lim\limits_{n_1\to\infty}\alpha^* = 0$.

From $n_1\alpha^* = \dfrac{\sigma_\epsilon^2\left(p_1 + \frac{p_1^2+p_1}{n_1} + o(n_1^{-1})\right)}{\left(\sigma_\eta^2 - \sigma_\epsilon^2\right) + O(n_1^{-1}) + O(n_2^{-1})} = O(1)$, we have $\alpha^* = O({n_1}^{-1})$.

(3)

$$\widehat{\alpha}^* - \alpha^* = \frac{\hat{A}}{\hat{A}+\hat{B}} - \frac{A}{A+B} = \frac{\hat{A}B - A\hat{B}}{(A+B)(\hat{A}+\hat{B})}$$

We therefore want to prove that

$$\mathbb{E}\left(\left(\frac{\hat{A}B - A\hat{B}}{(A+B)(\hat{A}+\hat{B})}\right)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})$$

Since $A \geqslant 0$ and $\hat{A} \geqslant 0$, it is sufficient to prove

$$\mathbb{E}\left(\left(\frac{\hat{A}B - A\hat{B}}{B\hat{B}}\right)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})$$

From standard regression theory, we have that $\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}$ and $\frac{\hat{\sigma}_\eta^2}{\sigma_\eta^2}$ follow scaled chi-squared distributions with degrees of freedom n_1-p_1-1 and n_2-p_2-1 respectively. Let $a = \frac{\sigma_\eta^2}{\sigma_\eta^2-\sigma_\epsilon^2}$ and $b = a-1 = \frac{\sigma_\epsilon^2}{\sigma_\eta^2-\sigma_\epsilon^2}$. By Lemma 3.4.6(i), $P\left(\left|\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}-1\right| > \frac{1}{5b}\right) \leqslant 2e^{-\frac{n_1-p_1-1}{200b^2}}$ and $P\left(\left|\frac{\hat{\sigma}_\eta^2}{\sigma_\eta^2}-1\right| > \frac{1}{5a}\right) \leqslant 2e^{-\frac{n_2-p_2-1}{200a^2}}$. Since $(\hat{\alpha}^*-\alpha^*)^2$ is bounded by 1, the effect of the low probability events on the expectation is small. If $\left|\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}-1\right| < \frac{1}{5b}$ and $\left|\frac{\hat{\sigma}_\eta^2}{\sigma_\eta^2}-1\right| < \frac{1}{5a}$, then

$$\left|\frac{\widehat{\sigma_\eta^2}-\hat{\sigma_\epsilon^2}}{\sigma_\eta^2-\sigma_\epsilon^2}-1\right| = \left|a\left(\frac{\widehat{\sigma_\eta^2}}{\sigma_\eta^2}-1\right)-b\left(\frac{\widehat{\sigma_\epsilon^2}}{\sigma_\epsilon^2}-1\right)\right| \leqslant a\left|\left(\frac{\widehat{\sigma_\eta^2}}{\sigma_\eta^2}-1\right)\right| + b\left|\left(\frac{\widehat{\sigma_\epsilon^2}}{\sigma_\epsilon^2}-1\right)\right| \leqslant \frac{2}{5}$$

For large enough n_1, n_2, this implies $\hat{B} > \frac{B}{2}$. Thus, it is sufficient to prove

$$\mathbb{E}\left(\left(\frac{\hat{A}B-A\hat{B}}{B^2}\right)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})$$

or equivalently

$$\mathbb{E}\left(\frac{A^2}{B^2}\left(\left(\frac{\hat{A}}{A}-1\right)-\left(\frac{\hat{B}}{B}-1\right)\right)^2\right) = O(n_1^{-3}) + O(n_1^{-2}n_2^{-1})$$

We know that $\frac{A}{B} = O(n_1^{-1})$, so we only need to show that $\mathbb{E}\left(\left(\frac{\hat{A}}{A}-1\right)^2\right) = O(n_1^{-1})$, $\mathbb{E}\left(\left(\frac{\hat{B}}{B}-1\right)^2\right) = O(n_2^{-1})+O(n_1^{-1})$ and $\mathbb{E}\left(\frac{\hat{B}}{B}-1\right)\left(\frac{\hat{A}}{A}-1\right) = O(n_1^{-1/2})O((n_1\wedge n_2)^{-1/2}) \leq O(n_2^{-1}) + O(n_1^{-1})$.

Since $A = \sigma_\epsilon^2\frac{p_1}{n_1} + O(n_1^{-2})$, we have $\frac{\hat{A}}{A}-1 = \frac{\hat{\sigma}_\epsilon^2\frac{p_1}{n_1}+O(n_1^{-2})}{\sigma_\epsilon^2\frac{p_1}{n_1}+O(n_1^{-2})}-1 = \frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}\left(1-O(n_1^{-1})\right)-1$. This means $\mathbb{E}\left(\left(\frac{\hat{A}}{A}-1\right)^2\right) = \mathbb{E}\left(\left(\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}\left(1-O(n_1^{-1})\right)-1\right)^2\right)$. Since $\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}$ follows a scaled chi-squared distribution, $\mathbb{E}\left(\left(\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2}\left(1-O(n_1^{-1})\right)-1\right)^2\right) = O(n_1^{-1})$ as required.

Similarly,

$$\mathbb{E}\left(\left(\frac{\hat{B}}{B}-1\right)^2\right)=\mathbb{E}\left(\left(\frac{\widehat{\sigma_\eta^2}\left(1+\frac{p_2}{n_2}+\frac{p_2^2+p_2}{n_2^2}\right)-\hat{\sigma_\epsilon^2}}{\sigma_\eta^2\left(1+\frac{p_2}{n_2}+\frac{p_2^2+p_2}{n_2^2}\right)-\sigma_\epsilon^2}\left(1-O(n_2^{-1})\right)-1\right)^2\right)=\mathbb{E}\left(\left(\frac{\widehat{\sigma_\eta^2}-\hat{\sigma_\epsilon^2}}{\sigma_\eta^2-\sigma_\epsilon^2}-1+O(n_2^{-1})\right)^2\right)$$

and from

$$\frac{\widehat{\sigma_\eta^2}-\hat{\sigma_\epsilon^2}}{\sigma_\eta^2-\sigma_\epsilon^2}-1=a\left(\frac{\widehat{\sigma_\eta^2}}{\sigma_\eta^2}-1\right)-b\left(\frac{\widehat{\sigma_\epsilon^2}}{\sigma_\epsilon^2}-1\right)$$

we get

$$\mathbb{E}\left(\left(\frac{\widehat{\sigma_\eta^2}-\hat{\sigma_\epsilon^2}}{\sigma_\eta^2-\sigma_\epsilon^2}-1\right)^2\right)=a^2\mathbb{E}\left(\left(\frac{\widehat{\sigma_\eta^2}}{\sigma_\eta^2}-1\right)^2\right)+b^2\mathbb{E}\left(\left(\frac{\widehat{\sigma_\epsilon^2}}{\sigma_\epsilon^2}-1\right)\right)=O(n_2^{-1})+O(n_1^{-1})$$

The Cauchy-Schwartz inequality gives that $\mathbb{E}(AB)\leqslant\sqrt{\mathbb{E}(A^2)\mathbb{E}(B^2)}$, thus

$$\mathbb{E}\left(\frac{\hat{B}}{B}-1\right)\left(\frac{\hat{A}}{A}-1\right)\leq\sqrt{\mathbb{E}\left(\left(\frac{\hat{B}}{B}-1\right)^2\right)\mathbb{E}\left(\left(\frac{\hat{A}}{A}-1\right)^2\right)}=O(n_1^{-1/2})O((n_1\wedge n_2)^{-1/2})$$

□

Now we are ready to prove the risk of the model combination method with our plug-in estimator of α^* is asymptotically better than the risk of full model which is equivalent to $\alpha^*=0$. Recall the expected loss for a constant α is defined as

$$\begin{aligned}R(\alpha)&=\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left[\left(\left((1-\alpha)\widehat{f}_1(\mathbf{x}_0)+\alpha\widehat{f}_2(\mathbf{x}_0)\right)-f(\mathbf{x}_0)\right)^2\wedge C_{n_1\wedge n_2}\right]\\&=(1-\alpha)^2A+\alpha^2B+o((n_1\wedge n_2)^{-2})\end{aligned}$$

Lemma 3.4.11. *Under the standard linear regression assumptions, if Z is a $\{0,1\}$-valued function of X, with $Z=1\Rightarrow\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty<\frac{n_1^{-0.4}}{p_1}$ then*

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(\widehat{f}_1(\mathbf{x}_0))-f(\mathbf{x}_0)\right)^4Z\right]=o(n_1^{-1})$$

Proof. For fixed $\hat{f}_1$, we have that the conditional distribution of $\hat{f}_1(\mathbf{x_0})-f(\mathbf{x_0})$ given $\hat{\beta}$ is normal with mean 0 and variance $(\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)$, so

$$\mathbb{E}_{\mathbf{x}_0}\left[\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^4\right]=3\left((\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)\right)^2=3\,\mathrm{tr}\left(\Sigma^{\frac{1}{2}}(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\Sigma^{\frac{1}{2}}\right)$$

Recall that $\hat{\beta} - \beta = (X^T X)^{-1} X^T E$ is normal with variance $\sigma_\epsilon^2 (X^T X)^{-1}$. This gives

$$\begin{aligned}\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^4 Z\right] &= 3\mathbb{E}_{(X,Y)}\left(\left((\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)\right)^2 Z\right) \\ &= 3\mathbb{E}_X\left(Z\mathbb{E}_{Y|X}\left((\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)\right)^2\right)\end{aligned}$$

Let $v = A\Sigma^{\frac{1}{2}}(\hat{\beta}-\beta)$ for some orthogonal matrix A. We have that $v^T v = (\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)$, and v is multivariate normal with variance $\sigma_\epsilon^2 A\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma^{\frac{1}{2}}A^T$. If we choose A so that $A\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma^{\frac{1}{2}}A^T$ is a diagonal matrix Λ, then

$$\begin{aligned}\mathbb{E}((v^T v)^2) = \mathbb{E}\left(\sum_{i,j=1}^p v_i^2 v_j^2\right) &= \sum_{i,j=1}^p \mathbb{E}\left(v_i^2 v_j^2\right) = \sigma_\epsilon^4\Big(\sum_{i\neq j}\lambda_i\lambda_j + \sum_{i=1}^p 3\lambda_i^2\Big) \\ &= \sigma_\epsilon^4\left(\left[\operatorname{tr}\left(\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma^{\frac{1}{2}}\right)\right]^2 + 2\operatorname{tr}\left(\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma(X^T X)^{-1}\Sigma^{\frac{1}{2}}\right)\right)\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[Z\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^4\right] \\ &= 3\sigma_\epsilon^4\mathbb{E}_X\left(Z\left(\left[\operatorname{tr}\left(\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma^{\frac{1}{2}}\right)\right]^2 + 2\operatorname{tr}\left(\Sigma^{\frac{1}{2}}(X^T X)^{-1}\Sigma(X^T X)^{-1}\Sigma^{\frac{1}{2}}\right)\right)\right)\end{aligned}$$

Recall that if $Z = 1$, then $\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I\right\|_\infty < \frac{n_1^{-0.4}}{p_1}$ so $\operatorname{tr}\left(n_1\Sigma^{\frac{1}{2}}\left(X^T X\right)^{-1}\Sigma^{\frac{1}{2}}\right) \leqslant p_1 + 1$ by Lemma 3.4.8. Also,

$$\begin{aligned}\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-1}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I &= \left(\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I\right)\left(\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} + I\right) \\ &= \left(\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I\right)\left(\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I\right) \\ &\quad + 2\left(\Sigma^{-\frac{1}{2}}\left(\frac{X^T X}{n_1}\right)\Sigma^{-\frac{1}{2}} - I\right)\end{aligned}$$

So

$$\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-1}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty \leqslant p_1\left\|\left(\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right)\right\|_\infty^2 + 2\left\|\Sigma^{-\frac{1}{2}}\left(\frac{X^TX}{n_1}\right)\Sigma^{-\frac{1}{2}}-I\right\|_\infty \leqslant \frac{n_1^{-0.8}}{p_1}+2\frac{n_1^{-0.4}}{p_1}$$

Therefore by Lemma 3.4.8, $\operatorname{tr}\left(n_1^2\Sigma^{\frac{1}{2}}\left(X^TX\right)^{-1}\Sigma\left(X^TX\right)^{-1}\Sigma^{\frac{1}{2}}\right)\leqslant p_1+1$ so

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(\widehat{f}_1(\mathbf{x}_0))-f(\mathbf{x}_0)\right)^4 Z\right]=o(n_1^{-1})$$

□

Theorem 3.4.12. *Assume the standard linear regression model assumptions that the rows in data X and test data $\mathbf{x}_0$ are i.i.d. from $N_{p_1}(0,\Sigma)$, and $Y|X\sim N(X\beta,\sigma_\epsilon^2)$, where β is a $p_1\times 1$ regression coefficient vector. Let the least square estimators $\widehat{f}_1(x)=x^T\widehat{\boldsymbol{\beta_1}}$ and $\widehat{f}_2(x)=x^T\widehat{\boldsymbol{\beta_2}}$, where the ith element of $\widehat{\beta}_2$: $(\widehat{\beta}_2)_i=0$ for $i>p_2,(p_2<p_1)$. Let $\widehat{\alpha}^*=\dfrac{\widehat{A}}{\widehat{A}+\widehat{B}}$ where $\hat{A}=\hat{\sigma}_\epsilon^2\left(\dfrac{p_1}{n_1}+\dfrac{p_1^2+p_1}{n_1^2}\right)$, $\hat{B}=\left(\widehat{\sigma_\eta^2}-\widehat{\sigma_\epsilon^2}\right)+\widehat{\sigma_\eta^2}\left(\dfrac{p_2}{n_2}+\dfrac{p_2^2+p_2}{n_2^2}\right)$, $\widehat{\sigma_\epsilon^2}=\dfrac{1}{n_1-p_1}\displaystyle\sum_{i=1}^{n_1}(\mathbf{y}_i-\widehat{f}_1(\mathbf{x}_i))^2$ and $\widehat{\sigma_\eta^2}=\dfrac{1}{n_2-p_2}\displaystyle\sum_{i=1+n_1}^{n}(\mathbf{y}_i-\widehat{f}_2(\mathbf{x}_i))^2$. With fixed p_1 and p_2 and $\sigma_\eta^2-\sigma_\epsilon^2>0$, for sufficient large n_1 and n_2 and $n_2>n_1$, we have*

$$R_{\text{plugin}}=\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\widehat{\alpha}^*)\widehat{f}_1(\mathbf{x}_0))+\widehat{\alpha}^*\widehat{f}_2(\mathbf{x}_0))-f(\mathbf{x}_0)\right)^2\wedge C_{n_1\wedge n_2}\right]<R(0).$$

Proof. It is equivalent to show

$$\mathbb{E}\left(R(\widehat{\alpha}^*)\right)-R(\alpha^*)<R(0)-R(\alpha^*).$$

where $\alpha^*=\dfrac{A}{A+B}$, where $A=\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X,Y)}\left(\left(\widehat{f}_1(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2\wedge C_{n_1}\right)$, $B=\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{(X_{21},Y_2)}\left(\left(\widehat{f}_2(\mathbf{x}_0)-f(\mathbf{x}_0)\right)^2\wedge C_{n_2}\right)$ and C_n is a sequence of constants, bounded below, satisfying $C_n=o(n)$.

By Theorem 3.4.9 and Theorem 3.4.10, we have

$$\begin{cases} A = O({n_1}^{-1}) \\ B = (\sigma_\eta^2 - \sigma_\epsilon^2) + O({n_2}^{-1}) \\ \alpha^* = O({n_1}^{-1}) \end{cases} \tag{3.22}$$

Therefore, the RHS is

$$R(0) - R(\alpha^*) = (A + o((n_1 \wedge n_2)^{-2}) - ((1-\alpha^*)^2 A + (\alpha^*)^2 B + o((n_1 \wedge n_2)^{-2})) \tag{3.23}$$

$$= \alpha^* \Big[(2-\alpha^*)A - \alpha^* B \Big] + o((n_1 \wedge n_2)^{-2})) = O(n_1^{-2}). \tag{3.24}$$

Next, consider the LHS, we first define an indicator variable as

Case 1 $\left\| \Sigma^{-\frac{1}{2}} \left(\frac{X^T X}{n_1} \right) \Sigma^{-\frac{1}{2}} - I \right\|_\infty < \frac{n_1^{-0.4}}{p_1}$ and $\left\| \Sigma_{11}^{-\frac{1}{2}} \left(\frac{X_{21}^T X_{21}}{n_2} \right) \Sigma_{11}^{-\frac{1}{2}} - I \right\|_\infty < \frac{n_2^{-0.4}}{p_2}$, denoted by $Z = 1$.

Case 2 $\left\| \Sigma^{-\frac{1}{2}} \left(\frac{X^T X}{n_1} \right) \Sigma^{-\frac{1}{2}} - I \right\|_\infty \geqslant \frac{n_1^{-0.4}}{p_1}$ or $\left\| \Sigma_{11}^{-\frac{1}{2}} \left(\frac{X_{21}^T X_{21}}{n_2} \right) \Sigma_{11}^{-\frac{1}{2}} - I \right\|_\infty \geqslant \frac{n_2^{-0.4}}{p_2}$, denoted by $Z = 0$.

The LHS is

$$
\begin{aligned}
&R_{\text{plugin}} - R(\alpha^*) \\
&= \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\widehat{\alpha}^*)\widehat{f}_1(\mathbf{x}_0)) + \widehat{\alpha}^*\widehat{f}_2(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1\wedge n_2}\right] \\
&- \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\alpha^*)\widehat{f}_1(\mathbf{x}_0)) + \alpha^*\widehat{f}_2(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 \wedge C_{n_1\wedge n_2}\right] \\
&= \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\widehat{\alpha}^*)\widehat{f}_1(\mathbf{x}_0)) + \widehat{\alpha}^*\widehat{f}_2(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 Z\right] \\
&- \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(((1-\alpha^*)\widehat{f}_1(\mathbf{x}_0)) + \alpha^*\widehat{f}_2(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 Z\right] + o((n_1\wedge n_2)^{-2}) \\
&= \mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\Bigg[(\widehat{\alpha}^*-\alpha^*)\Bigg((\alpha^*+\widehat{\alpha}^*-2)\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 \\
&\qquad + 2(1-\alpha^*-\widehat{\alpha}^*)\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right) \\
&\qquad\qquad + (\alpha^*+\widehat{\alpha}^*)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2\Bigg) Z\Bigg] + o((n_1\wedge n_2)^{-2}) \\
&= 2\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\Bigg[(\widehat{\alpha}^*-\alpha^*)\Bigg(-\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 + \left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right) \\
&\qquad\qquad + \alpha^*\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2\Bigg) Z\Bigg] + o(n_1^{-2})
\end{aligned}
$$

Thus, it is sufficient to show that

$$
\begin{aligned}
\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[(\widehat{\alpha}^*-\alpha^*)\left(\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^2 Z\right)\right] &= o(n_1^{-2}) \\
\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[(\widehat{\alpha}^*-\alpha^*)\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right) Z\right)\right] &= o(n_1^{-2}) \\
\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[(\widehat{\alpha}^*-\alpha^*)\left(\alpha^*\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^2\right) Z\right] &= o(n_1^{-2})
\end{aligned}
$$

The Cauchy-Schwartz inequality gives that $\mathbb{E}(AB) \leqslant \sqrt{\mathbb{E}(A^2)\mathbb{E}(B^2)}$, so it is sufficient to prove

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left((\widehat{\alpha}^* - \alpha^*)^2\right) = O(n_1^{-3}) \tag{3.25}$$

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\left(\widehat{f}_1(\mathbf{x}_0)) - f(\mathbf{x}_0)\right)^4 Z\right] = o(n_1^{-1}) \tag{3.26}$$

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[\alpha^{*2}\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)^4 Z\right] = o(n_1^{-1}) \tag{3.27}$$

$$\mathbb{E}_{(X,Y)}\mathbb{E}_{\mathbf{x}_0}\left[(\widehat{\alpha}^* - \alpha^*)\left(\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)Z\right)\right] = o(n_1^{-2}) \tag{3.28}$$

(3.25) is from Theorem 3.4.10; (3.26) is Lemma 3.4.11. (3.27) follows from the fact that $\alpha^* = O(n_1^{-1})$. For (3.28), because $\hat{f}_1$ and $\hat{f}_2$ are estimated on different data, they are independent. Furthermore, since $\hat{\alpha}^*$ is estimated from the mean squared residuals of the regression, and $\hat{f}_1$ and $\hat{f}_2$ are estimated from the mean of the regression, they are conditionally independent given X. Therefore, for any X,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x_0},Y|X}\left((\hat{\alpha}^* - \alpha^*)\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\right) \\
&= \mathbb{E}_{\mathbf{x_0},Y|X}(\hat{\alpha}^* - \alpha^*)\mathbb{E}_{\mathbf{x_0},Y|X}\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\mathbb{E}_{\mathbf{x_0},Y|X}\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)
\end{aligned}$$

Also, because $\hat{f}_1$ is an unbiased estimator, for any X, we have $\mathbb{E}_{\mathbf{x_0},Y|X}\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right) = 0$. Thus, we have

$$\mathbb{E}_X\left(Z\mathbb{E}_{\mathbf{x_0},Y|X}(\hat{\alpha}^* - \alpha^*)\mathbb{E}_{\mathbf{x_0},Y|X}\left(\widehat{f}_1(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\mathbb{E}_{\mathbf{x_0},Y|X}\left(\widehat{f}_2(\mathbf{x}_0) - f(\mathbf{x}_0)\right)\right) = 0$$

□

These asymptotic results showed that the model combination method in linear regression models with the plug-in estimate for α gives more accurate prediction than the full data model. More thorough comparisons will be needed to verify the model combination methods for finite samples in both linear models and nonlinear models, missing at random and missing not at random, which will be the content of next Chapter.

Chapter 4

Comparisons of Model Combination Methods and Multiple Imputation Methods in Simulations and Real Data Analyses for Two Blocks of Predictors

In this chapter we compare the model combination methods proposed in Chapter 3 with the state-of-the-art multiple imputation method, and complete case approach for block missing data with two blocks of predictors according to the predictive performance of the methods. We simulate a range of scenarios for missing at random (MAR), including true models following linear regression models, logistic regression models and nonlinear regression models. The details of the simulation design and results for these scenarios are included in Section 4.1, Section 4.2 and Section 4.3 respectively. We also demonstrate the performance of these methods for missing not at random (MNAR) in both regression and logistic regression simulations in Section 4.4. Finally we compare the predictive performance of these methods in two real data applications in Section 4.5.

4.1 Linear regression

We compare the performance of our method and other methods for dealing with missing data on a simulation involving linear regression on multivariate Gaussian predictors with one block of ten variables that are available in all samples, and another block of ten variables that are available in only some samples. We study the effect of several different factors on the performance of various methods: the level of correlation between predictors; the number of true predictors in the missing block; and the number of complete and partial observations.

4.1.1 Simulation design

In order to simulate different levels of correlations between the predictor variables, we use the following method to simulate the covariance matrix for a total of 20 predictor variables. The covariance matrix is simulated from $B = R^T R$, where R is a 20x20 upper triangular

matrix. We fix the diagonal entries of R as 1 for the high correlation case; 15 for the medium correlation case; and 30 for the low correlation case. We simulate the non-zero off-diagonal entries of R as i.i.d. following a mixture of uniform distributions: Unif$(-1, 1)$, Unif$(-10, 10)$, Unif$(-0.5, 0.5)$ and Unif$(10, 20)$ all with probability 1/4. We normalize B though $\Sigma = \text{diag}(B)^{-1/2} B \text{diag}(B)^{-1/2}$ to have all ones on the diagonal.

Having generated a random covariance matrix Σ for each scenario of high, medium and low correlations, we simulate the predictor matrix X with each row i.i.d. following a multivariate normal distribution with mean 0, and covariance matrix given by Σ. We first simulate the complete X matrix, and later remove the observations to generate a block missing structure. We designate the first 10 predictors as the first block that are present in all observations, and the last 10 predictors as the second block with some observations removed.

We study five scenarios for sample sizes (n_1, n_2), where n_1 denotes the number of complete observations and n_2 denotes the number of partial observations: (A) $n_1 = 40$, $n_2 = 60$; (B) $n_1 = 80$, $n_2 = 20$; (C) $n_1 = 40$, $n_2 = 960$; (D) $n_1 = 80$, $n_2 = 920$; (E) $n_1 = 200$, $n_2 = 800$. The scenarios (A) and (B) are for small sample sizes with missing blocks relatively large or small respectively. The scenarios (C) (D) and (E) are for larger samples with partial observations dominant, which reflect more realistic cases for most of the real data applications. These scenarios are illustrated in Figure 4.1.

We simulate the response variable $\mathbf{y} = \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ following a Gaussian distribution and $\boldsymbol{\beta}$ given by one of the following three different sets of regression coefficients:

(I) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \cdots, 0)$ where the first 5 elements of $\boldsymbol{\beta}$ are all set to equal to one, and the remaining 15 elements are all zero. In this case, the response is related only to the first block of variables which are always present.

(II) $\boldsymbol{\beta} = (0, \cdots, 0, 1, 1, 1, 1, 1)$, where the first 15 elements of $\boldsymbol{\beta}$ are all zero, only the last 5 elements of β are all set to equal to one. In this case, the response is related only to the second block of variables with missing observations.

(III) $\boldsymbol{\beta} = (1, 1, 0, \cdots, 0, 1, 1, 1)$, where the middle 15 elements of $\boldsymbol{\beta}$ are all zero. In this case, the response is related to two variables that are always present from the first block and three variables from the second block.

The random noise is $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n)$. We fix Signal-to-Noise ratio as $\text{SNR} = \dfrac{\text{Var(Signal)}}{\text{Var(Noise)}} = \dfrac{\text{Var}(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})}{\text{Var}(\boldsymbol{\epsilon})} = 1$, thus $\sigma_\epsilon^2 = \text{Var}(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}) = \boldsymbol{\beta}^T \text{Var}(\mathbf{x})\boldsymbol{\beta}$.

For each of these above $3 \times 5 \times 3$ scenarios, we simulate 100 replicate data sets. For

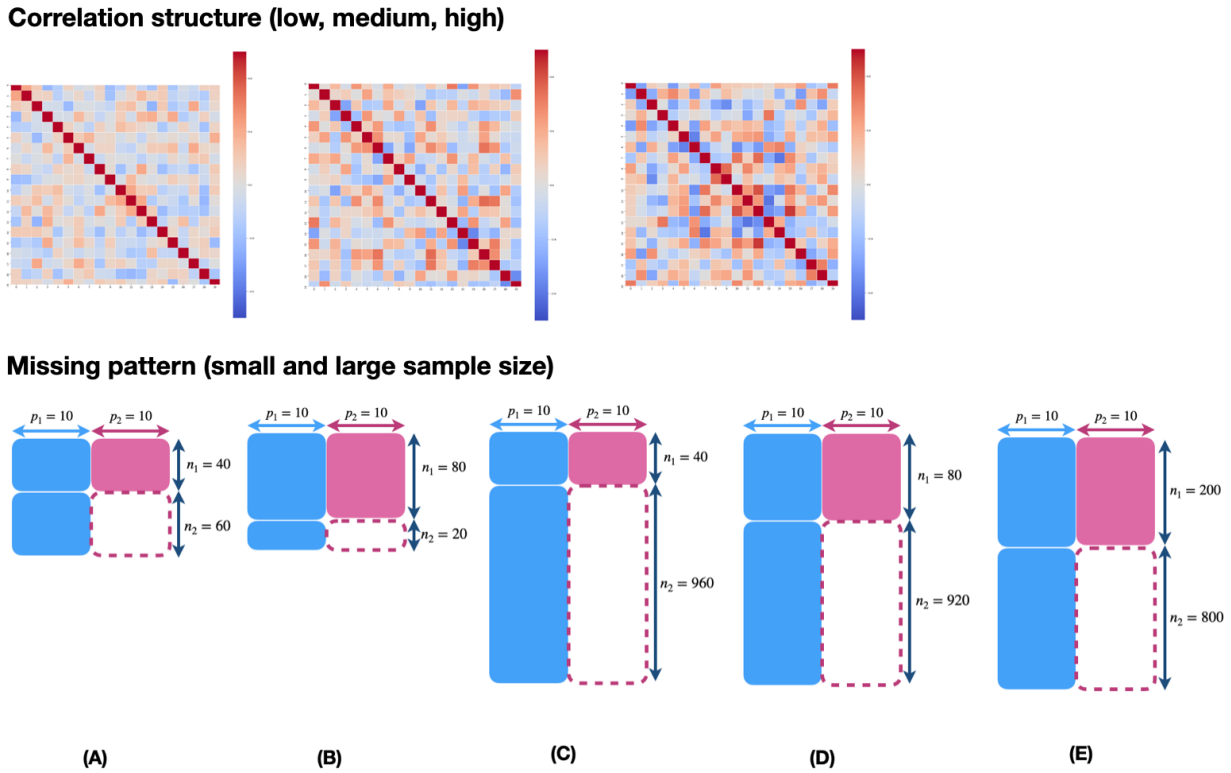

Figure 4.1: 5 data structure scenarios for the simulations: (A) and (B): 100 observations in total with $n_1 = 40$ and $n_1 = 80$. (C), (D) and (E): 1000 observations in total with $n_1 = 40$, $n_1 = 80$ and $n_1 = 200$, respectively.

each data set, we fit the following five models: (1) M_0 fits models using all $n = n_1 + n_2$ observations with only the first block of predictors. The fitted regression coefficients have the first 10 dimensions equal to the regression coefficients of this partial model and the last 10 dimensions all equal to 0. (2) M_1 fits models using the n_1 full observations with both blocks of variables, i.e. based on data (Y_1, X_1). (3) M_α(CV) fits models based on the model combination method described in Section 3.3, using cross-validation to estimate the combination parameter α. (4) M_α(plugin) fits models based on the model combination method described in Section 3.1, using the plug-in method to estimate the combination parameter α. (5) MI fits models based on the multiple imputation method described as follows.

We apply MICE (Multivariate Imputation by Chained Equations) to do multiple imputation in the simulation. MICE imputes incomplete multivariate data by chained equations. The Bayesian linear regression methods as implemented in the MICE package in R is used for imputation to create 20 imputation replicates for the missing data for each data set.

The quantities of interest in our problem are the regression coefficients $\hat{\boldsymbol{\beta}}$. For the ith imputed replicate, we perform a standard linear regression to estimate the coefficient $\hat{\beta}_{\text{MI}}^{(i)}$. For the pooled estimate, we take the average $\hat{\beta}_{\text{MI}}^{\text{ave}} = \dfrac{\hat{\beta}_{\text{MI}}^{(1)} + \cdots + \hat{\beta}_{\text{MI}}^{(20)}}{20}$ as our overall estimate.

4.1.2 Performance assessment

We assess the performance of the fitted regression models using mean squared error (MSE) for two different prediction scenarios. In the first scenario, we consider the complete case prediction where the predictors are fully observed for future observations. In the second scenario, we consider the incomplete case prediction where only the first block of variables for the future observations are available.

For predictions on complete cases, the MSE is defined by $MSE = \mathbb{E}(\hat{Y} - f(X))^2$. For the simulation, because we know the true distribution of $\mathbf{x}$, instead of using a test data set, we can directly calculate the MSE over the distribution. To make the notation clear, the observed value, true value and prediction value are denoted by y, $f(\mathbf{x})$ and $\hat{y} = \hat{f}(\mathbf{x})$ respectively , where $y = f(\mathbf{x}) + \epsilon$, $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$ and $\hat{y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$. Therefore, the MSE can be

expressed as

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}[(\hat{y} - f(\mathbf{x}))]^2 &= \mathbb{E}_{\mathbf{x}}[(\mathbf{x}^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\boldsymbol{\beta})^2] \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbb{E}(\mathbf{x}\mathbf{x}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \operatorname{Var}(\mathbf{x})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the regression coefficient vector estimated from the data, $\boldsymbol{\beta}$ is the true coefficient vector, and $\operatorname{Var}(\mathbf{x})$ is the covariance matrix used for the simulation. The MSEs can be easily calculated from the simulation results for each scenario.

For predictions on incomplete cases, models M_1, M_α(CV) and M_α(plugin) are all not applicable, since these models all need the full predictor variables. The comparisons will be focused on comparing the performance of model M_0 with the MI method. For the MI method, a full predictor vector of dimension 20 will be fitted based on each of the 20 replicate imputed training data sets, the mean of the 20 estimated coefficient vectors will be the final estimated regression coefficient $\hat{\boldsymbol{\beta}}_{MI}$. In principle, the predictions will be made also on the imputed test data with the imputation model based on the training data. To make the assessment comparible to other model results, we calculate the theoretical MSE as follows. From the available full data, suppose the imputation model based on the multivariate regression of X_{12} on X_{11} resulted in a coefficient matrix $\hat{A}$, then for each test case with available block of variables $\mathbf{x}_1$, the imputed block $(\hat{\mathbf{x}}_2)^T = \mathbf{x}_1^T\hat{A}$. Thus the MSE can be calculated as:

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}}[(\hat{y} - f(\mathbf{x}))^2] \\ =&\mathbb{E}_{\mathbf{x}}[((\mathbf{x}_1^T, \hat{\mathbf{x}}_2^T)\hat{\boldsymbol{\beta}}_{MI} - f(\mathbf{x}))^2] \\ =&\mathbb{E}_{\mathbf{x}}[(\mathbf{x}^T \begin{bmatrix} I & \hat{A} \\ 0 & 0 \end{bmatrix} \hat{\boldsymbol{\beta}}_{MI} - \mathbf{x}^T\boldsymbol{\beta})^2] \\ =&(\hat{\boldsymbol{\beta}}_{MI,inc} - \boldsymbol{\beta})^T \operatorname{Var}(\mathbf{x})(\hat{\boldsymbol{\beta}}_{MI,inc} - \boldsymbol{\beta})\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{MI,inc} = \begin{bmatrix} I & \hat{A} \\ 0 & 0 \end{bmatrix} \hat{\boldsymbol{\beta}}_{MI}$.

For each scenario, we calculate the average RMSE and the associated SE over 100 replicate data sets for all methods.

4.1.3 Linear Regression Simulation Results

Table 4.1 and Table 4.2 show the average of RMSE and standard error (SE) over 100 replicate data sets for complete case prediction and incomplete case prediction respectively. In order to compare the performance of model combination methods (M_α(CV) and M_α(plugin)) versus MI method, we conduct a one sample t-test on the RMSEs on the 100 replicate data sets to compare M_α(CV) versus MI and M_α(plugin) versus MI.

From Table 4.1, our method has significantly smaller average RMSE than that of MI in nearly all scenarios. For model scenario (I), the partial model has smaller RMSE. But there is no significant difference between our method and the partial model. The model combination method significantly outperforms MI in this case. Our method has not only better RMSE, but also smaller SE than MI.

Another finding is that both the model combination method and MI perform better when the variables are more correlated. This makes sense, because when the predictors are correlated, the variables in the first block form strong surrogates for the true predictors, so the blocks with missing data can produce better models than in the situations with lower correlation. It might seem, intuitively that MI should show more improvement with higher correlation, because this means that the imputed values are more accurate. However, the higher accuracy of the imputed values also means that the M_2 model (defined in Chapter 3 for $\hat{f}_2$) is more accurate. Therefore, our method has similar improvement in the high correlation case.

We also see that the difference between the methods gets smaller as the number of complete cases increases. This makes sense, since with enough complete cases, our method, MI and the complete-case model M_1 will all converge to the true coefficients. On the other hand, the number of incomplete observations causes a similar reduction in RMSE for our method and MI, meaning that the difference between the methods' performances is similar in scenarios (A) and (C), and in scenarios (B) and (D).

The plug-in estimation and the cross validation estimation for α have comparable results to each other under most scenarios. For a few large-sample scenarios with low correlation, M_α(CV) performs significantly worse than M_α(plugin) and MI.

From Table 4.2, our model more consistently outperforms MI at prediction for incomplete test data. Even when the number of complete cases gets larger, MI still underperforms. This is because, in order to make predictions from incomplete data, MI must impute the missing

variables, so the predictions depend on both the imputation model and the fitted model. This involves estimating many more parameters, so leads to larger model variance.

Table 4.1: Theoretical RMSE and SE of the linear regression model on complete cases under MAR for the partial model M_0, the full model M_1, the model combination method M_α estimated by CV estimation and Plug-in estimation and the multiple imputation method MI. The corresponding SEs are given in parentheses.

High correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.739	2.336	0.957	0.857	1.414	1.431	1.992	1.314	1.190	1.404	1.392	2.514	1.405	1.277	1.734
	(0.019)	(0.055)	(0.025)	(0.019)	(0.034)	(0.010)	(0.048)	(0.023)	(0.017)	(0.036)	(0.013)	(0.061)	(0.021)	(0.015)	(0.043)
(B)	0.744	1.299	1.141	1.095	1.094	1.432	1.139	1.017	0.969	1.003	1.409	1.462	1.244	1.163	1.302
	(0.017)	(0.023)	(0.017)	(0.018)	(0.020)	(0.009)	(0.019)	(0.016)	(0.015)	(0.017)	(0.014)	(0.028)	(0.022)	(0.021)	(0.024)
(C)	0.226	2.230	0.382	0.223	1.086	1.226	2.053	1.140	1.017	1.204	1.084	2.621	1.118	0.998	1.418
	(0.005)	(0.059)	(0.027)	(0.005)	(0.033)	(0.001)	(0.048)	(0.018)	(0.013)	(0.026)	(0.002)	(0.062)	(0.019)	(0.008)	(0.034)
(D)	0.226	1.272	0.312	0.226	0.764	1.228	1.127	0.894	0.838	0.863	1.086	1.507	0.963	0.881	1.002
	(0.005)	(0.024)	(0.015)	(0.005)	(0.019)	(0.001)	(0.023)	(0.014)	(0.014)	(0.018)	(0.002)	(0.030)	(0.013)	(0.011)	(0.024)
(E)	0.225	0.731	0.268	0.230	0.495	1.228	0.641	0.563	0.592	0.556	1.085	0.827	0.670	0.720	0.644
	(0.006)	(0.013)	(0.007)	(0.006)	(0.011)	(0.001)	(0.012)	(0.010)	(0.010)	(0.012)	(0.002)	(0.015)	(0.012)	(0.013)	(0.013)

Medium correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.865	2.648	1.179	1.051	1.793	1.859	2.213	1.608	1.469	1.777	1.807	3.104	1.794	1.636	2.292
	(0.022)	(0.056)	(0.028)	(0.021)	(0.050)	(0.012)	(0.053)	(0.025)	(0.019)	(0.040)	(0.014)	(0.073)	(0.029)	(0.019)	(0.059)
(B)	0.864	1.505	1.337	1.272	1.294	1.845	1.279	1.162	1.109	1.203	1.799	1.630	1.431	1.358	1.461
	(0.020)	(0.028)	(0.022)	(0.020)	(0.025)	(0.010)	(0.024)	(0.021)	(0.020)	(0.020)	(0.016)	(0.031)	(0.025)	(0.024)	(0.028)
(C)	0.261	2.817	0.516	0.257	1.582	1.636	2.265	1.420	1.282	1.628	1.482	3.068	1.468	1.318	1.981
	(0.007)	(0.069)	(0.041)	(0.006)	(0.057)	(0.001)	(0.050)	(0.021)	(0.017)	(0.036)	(0.001)	(0.067)	(0.024)	(0.013)	(0.054)
(D)	0.271	1.534	0.355	0.274	0.915	1.639	1.274	1.056	0.993	1.153	1.482	1.675	1.204	1.112	1.215
	(0.006)	(0.031)	(0.018)	(0.005)	(0.030)	(0.001)	(0.025)	(0.024)	(0.018)	(0.001)	(0.031)	(0.017)	(0.017)	(0.014)	(0.028)
(E)	0.258	0.872	0.319	0.269	0.555	1.635	0.709	0.670	0.648	0.829	1.482	0.952	0.838	0.791	0.835
	(0.006)	(0.013)	(0.008)	(0.006)	(0.016)	(0.001)	(0.012)	(0.011)	(0.011)	(0.011)	(0.001)	(0.015)	(0.014)	(0.013)	(0.014)

Low correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.814	2.555	1.119	0.993	1.747	2.186	2.269	1.714	1.588	1.856	2.016	2.850	1.881	1.721	2.222
	(0.023)	(0.059)	(0.028)	(0.023)	(0.446)	(0.010)	(0.054)	(0.029)	(0.027)	(0.039)	(0.014)	(0.064)	(0.031)	(0.022)	(0.054)
(B)	0.824	1.453	1.269	1.201	1.331	2.188	1.303	1.230	1.176	1.234	2.010	1.681	1.529	1.440	1.585
	(0.023)	(0.025)	(0.023)	(0.022)	(0.024)	(0.012)	(0.023)	(0.020)	(0.020)	(0.023)	(0.013)	(0.030)	(0.028)	(0.024)	(0.029)
(C)	0.241	2.654	0.465	0.240	1.535	1.959	2.428	1.586	1.466	1.780	1.713	2.861	1.623	1.443	2.021
	(0.006)	(0.070)	(0.029)	(0.006)	(0.043)	(0.001)	(0.058)	(0.027)	(0.023)	(0.043)	(0.001)	(0.063)	(0.024)	(0.017)	(0.053)
(D)	0.257	1.509	0.331	0.257	0.948	1.956	1.284	1.132	1.066	1.067	1.711	1.662	1.260	1.187	1.244
	(0.006)	(0.026)	(0.016)	(0.006)	(0.019)	(0.001)	(0.027)	(0.020)	(0.020)	(0.023)	(0.001)	(0.030)	(0.019)	(0.018)	(0.024)
(E)	0.244	0.839	0.296	0.252	0.600	1.958	0.762	0.722	0.700	0.661	1.711	0.966	0.883	0.848	0.780
	(0.006)	(0.014)	(0.008)	(0.006)	(0.011)	(0.001)	(0.014)	(0.014)	(0.013)	(0.012)	(0.001)	(0.015)	(0.013)	(0.013)	(0.014)

Note: The underlined entry denotes the best performance for that particular setting, i.e. the smallest value in that row. The red and blue colour indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

4.2 Logistic regression

4.2.1 Simulation design

For Logistic regression, we use the same procedure to simulate the predictor matrices X as in Section 4.1.1, but we are more focused on large sample cases, thus we only simulate scenarios

Table 4.2: Theoretical RMSE and SE of the linear regression model on incomplete cases under MAR for the partial model M_0 (the model combination method M_α is equivalent to the partial model) and the multiple imputation method MI. The corresponding SEs are given in parentheses.

High correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(A)	0.777(0.019)	0.981(0.027)	1.431(0.010)	2.275(0.027)	1.392(0.013)	2.100(0.027)
(B)	0.707(0.017)	0.726(0.018)	1.432(0.009)	2.175(0.019)	1.409(0.014)	2.041(0.025)
(C)	0.226(0.005)	0.643(0.021)	1.226(0.001)	2.167(0.018)	1.084(0.002)	1.930(0.022)
(D)	0.226(0.005)	0.405(0.014)	1.228(0.001)	2.091(0.013)	1.086(0.002)	1.847(0.012)
(E)	0.222(0.006)	0.286(0.007)	1.228(0.001)	2.077(0.008)	1.085(0.002)	1.822(0.008)

Medium correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(A)	0.865(0.022)	1.006(0.027)	1.859(0.012)	2.495(0.026)	1.807(0.014)	2.337(0.032)
(B)	0.875(0.020)	0.897(0.021)	1.845(0.010)	2.358(0.022)	1.799(0.016)	2.248(0.027)
(C)	0.267(0.007)	0.666(0.023)	1.636(0.001)	2.361(0.023)	1.482(0.001)	2.164(0.025)
(D)	0.257(0.006)	0.339(0.008)	1.639(0.001)	2.281(0.011)	1.482(0.001)	2.020(0.014)
(E)	0.256(0.006)	0.278(0.007)	1.635(0.001)	2.268(0.007)	1.482(0.001)	1.993(0.008)

Low correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(A)	0.866(0.023)	1.021(0.025)	2.186(0.010)	2.664(0.028)	2.016(0.014)	2.444(0.030)
(B)	0.819(0.023)	0.829(0.022)	2.188(0.012)	2.521(0.019)	2.010(0.013)	2.257(0.023)
(C)	0.248(0.006)	0.630(0.023)	1.959(0.001)	2.470(0.018)	1.713(0.001)	2.197(0.020)
(D)	0.256(0.006)	0.342(0.009)	1.956(0.001)	2.363(0.011)	1.711(0.001)	2.058(0.011)
(E)	0.246(0.006)	0.255(0.006)	1.958(0.001)	2.323(0.007)	1.711(0.001)	2.016(0.007)

Note: The underlined entry denotes the best performance for that particular setting, i.e. the smallest value on that row. The red and blue colour indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

(C), (D) and (E) in this simulation. We generate the response variable y following the logistic model $y|x \sim \text{Bernoulli}(p(x))$, where $\text{logit}(p(x)) = x^T\beta$. We use the same β setting as for linear regression and there is no intercept ($\beta_0 = 0$). For scenario (I), the first 5 elements of the coefficient vector β are all set equal to one, and the remaining 15 elements are all zero. For scenario (II), the first 15 elements of β are all zero, and the last 5 elements of β are all set equal to one. For scenario (III), the middle 15 elements of β are all zero, while the first two and last three elements of β are set to one. Therefore, $x^T\beta$ also follows a Gaussian

distribution with mean 0, thus each of the two classes contains around 50% of observations. We calculate the results for model combination methods based on the plug-in method from Section 3.2 and the GCV method from Section 3.3.

We assess the performance using negative log likelihood (NLL) as a loss function on the test data. We simulate 1000 complete test observations for calculating test NLL for each simulation. For incomplete case prediction, we evaluate test NLL using the predictions from only the first block of predictors on the same test data set.

4.2.2 Results

Table 4.3 shows the average test data NLL values over 100 replicates and the standard error of the mean NLL.

For complete case predictions (Table 4.3), the model combination method has a smaller test NLL than any other method under most cases for model scenarios (II) and (III). As expected, M_0 has smaller NLL for model scenario (I). There is no significant difference between the model combination method and the partial model in this case. Note that in real applications, unless we know *a priori* that the second block predictors are not important, we usually include them in the model. In terms of standard error, the variance of the model combination method is always between that of the partial model M_0 and the full model M_1, as we expected. This shows that the model combination method is practical and able to produce better and robust results.

The model combination method M_α(CV) has significantly smaller NLL than that of MI under most cases for model scenarios (II) and (III). M_α(plugin) gives better results when y is not directly related to the missing predictors and worse results when y is directly related to the missing predictors, even significantly worse than MI for model scenarios (II). While M_α(CV) shows relatively better results in three cases, using LOOCV can be very time consuming and sometimes unstable. In logistic simulation, we use GCV instead of LOOCV as mentioned in Section 3.3 to obtain more stable and efficient results.

The difference between the model combination method and MI gets larger as the number of complete cases increases. Our method converges to the true coefficient vector faster than MI (scenario (E)).

For the incomplete case, the model combination method (which is equivalent to the partial model) achieves smaller test NLL and SE, as shown in Table 4.4. This is particularly

important for real-world data sets, where predictions will be made on a large number of incomplete cases. This also shows that the model combination method is more flexible in both complete and incomplete cases.

In summary, the overall performance of the model combination method is more stable and accurate on block missing data for both linear regression and logistic regression.

Table 4.3: Test NLL and SE of the logistic regression model on complete cases under MAR for the partial model M_0, the full model M_1, the combined model M_α estimated by CV estimation and Plug-in estimation and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	430.708	10376.635	439.196	431.296	432.414	631.329	11915.137	626.928	631.831	625.987	523.918	10693.76	522.863	524.106	520.805
	(1.844)	(240.374)	(2.444)	(1.866)	(1.884)	(1.212)	(255.8)	(2.338)	(1.231)	(1.271)	(1.646)	(213.984)	(1.97)	(1.646)	(1.712)
(D)	430.07	3448.811	434.332	430.396	430.635	630.654	1486.93	589.675	631.238	615.733	523.784	1932.907	516.868	524.29	516.898
	(1.685)	(399.684)	(2.214)	(1.696)	(1.714)	(1.205)	(220.563)	(2.512)	(1.211)	(1.317)	(1.529)	(283.406)	(1.704)	(1.598)	(1.582)
(E)	429.895	506.632	433.014	433.839	431.233	629.252	573.369	536.595	538.591	588.923	525.846	541.266	494.536	496.886	506.473
	(1.501)	(4.385)	(1.548)	(1.552)	(1.542)	(1.27)	(3.838)	(2.172)	(2.291)	(1.446)	(1.916)	(5.57)	(2.567)	(2.589)	(1.962)

Medium correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	416.481	10162.391	422.284	417.085	417.424	641.656	11646.389	633.399	641.85	635.628	530.73	10687.321	526.95	530.901	528.316
	(2.052)	(200.133)	(2.497)	(2.041)	(2.075)	(1.173)	(267.903)	(2.292)	(1.169)	(1.279)	(1.9)	(241.41)	(2.23)	(1.925)	(1.923)
(D)	421.227	3116.709	424.368	421.27	422.135	643.845	1661.215	596.299	644.467	627.783	534.172	1965.506	525.052	530.039	527.741
	(1.86)	(374.096)	(1.996)	(1.67)	(1.891)	(1.176)	(259.559)	(3.049)	(1.18)	(1.358)	(1.551)	(289.517)	(1.761)	(1.564)	(1.587)
(E)	419.684	499.332	422.768	423.549	420.933	641.905	570.689	536.227	538.659	599.624	528.025	539.56	493.259	494.869	510.287
	(1.846)	(4.759)	(1.951)	(1.956)	(1.848)	(1.13)	(4.141)	(2.235)	(2.173)	(1.165)	(1.783)	(4.294)	(2.337)	(2.358)	(1.894)

Low correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	418.52	10081.796	427.478	419.097	419.292	673.717	11165.787	653.193	673.968	667.4	560.895	11178.427	545.353	561.083	558.549
	(1.608)	(241.096)	(2.425)	(1.594)	(1.609)	(0.726)	(205.895)	(3.036)	(0.733)	(0.954)	(1.397)	(212.121)	(2.089)	(1.401)	(1.413)
(D)	419.032	2817.976	425.421	419.711	419.905	673.438	2057.823	596.867	673.795	654.302	558.948	2013.118	542.098	559.458	548.735
	(1.859)	(350.312)	(3.43)	(1.865)	(1.895)	(0.923)	(296.214)	(2.724)	(0.906)	(1.032)	(1.579)	(282.178)	(2.803)	(1.615)	(1.606)
(E)	423.473	501.327	426.545	427.523	425.232	674.232	542.47	514.772	517.87	618.168	559.583	530.501	496.208	498.698	530.276
	(1.741)	(4.556)	(1.876)	(1.969)	(1.782)	(0.887)	(4.014)	(2.236)	(2.21)	(1.017)	(1.337)	(4.122)	(2.077)	(2.227)	(1.403)

Note: The underlined entry denotes the best performance for that particular setting, i.e. the smallest value on that row. The red and blue colour indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

4.3 Non-linear target function

4.3.1 Simulation design

We now examine the performance of different methods where there is a non-linear relation between the predictors and the response variable. We focus on large sample cases (D) and (E) in this simulation. We generate two types of predictor variables. First simulate predictor matrices X with multivariate Gaussian distribution using the same procedure as described in Section 4.1.1.

Table 4.4: Test NLL and SE of the logistic regression model on incomplete cases under MAR for the partial model M_0 (the model combination method M_α is equivalent to the partial model) and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(C)	430.708(1.844)	438.382(2.414)	631.329(1.212)	640.701(2.334)	523.918(1.646)	533.668(2.640)
(D)	430.07(1.685)	434.235(2.384)	630.654(1.205)	638.794(2.967)	523.784(1.529)	532.527(2.923)
(E)	429.895(1.501)	432.52(1.513)	629.252(1.27)	638.596(1.451)	525.846(1.916)	533.467(2.406)

Medium correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(C)	416.481(2.052)	418.35(2.061)	641.656(1.173)	643.603(1.148)	530.73(1.9)	532.865(1.928)
(D)	421.227(1.86)	422.709(1.892)	643.845(1.176)	646.281(1.18)	534.172(1.551)	535.475(1.565)
(E)	419.684(1.846)	421.344(1.907)	641.905(1.13)	648.598(1.302)	528.025(1.783)	532.342(1.852)

Low correlation

	(I) Y related to first block		(II) Y related to second block		(III) Y related to both blocks	
	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI	$M_0(M_\alpha)$	MI
(C)	418.52(1.608)	420.024(1.609)	673.717(0.726)	675.358(0.761)	560.895(1.397)	562.495(1.44)
(D)	419.032(1.859)	419.773(1.883)	673.438(0.923)	675.106(0.954)	558.948(1.579)	560.542(1.612)
(E)	423.473(1.741)	425.068(1.752)	674.232(0.887)	682.067(1.094)	559.583(1.337)	564.528(1.39)

Note: The underlined entry denotes the best performance for that particular setting, i.e. the smallest value on that row. The red and blue colour indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

In practice we seldom have normally distributed data, so it is necessary to examine the non-linear model performances under non-normal predictors. To generate dependent X predictors with non-normal distributions, we use the methods from the previous simulations to generate multivariate normal predictors, then perform a univariate transformation on each predictor to make its marginal distribution chi-squared with one degree of freedom. That is, we use the following procedure to generate the predictor matrix X''.

- Step1. Generate $X \sim N(\mathbf{0}, \Sigma)$.
- Step2. Let $X_i' = F_i(X_i) \sim \text{Unif}([0,1])$, where F_i is the marginal distribution of function X_i. That is, F_i is the c.d.f. of a normal distribution.
- Step3. Let $X_i'' = F^{-1}_{\chi^2_1}(X_i') \sim \chi^2_1$, where $F^{-1}_{\chi^2_1}$ is quantile function of a chi-squared distribution with one degree of freedom.

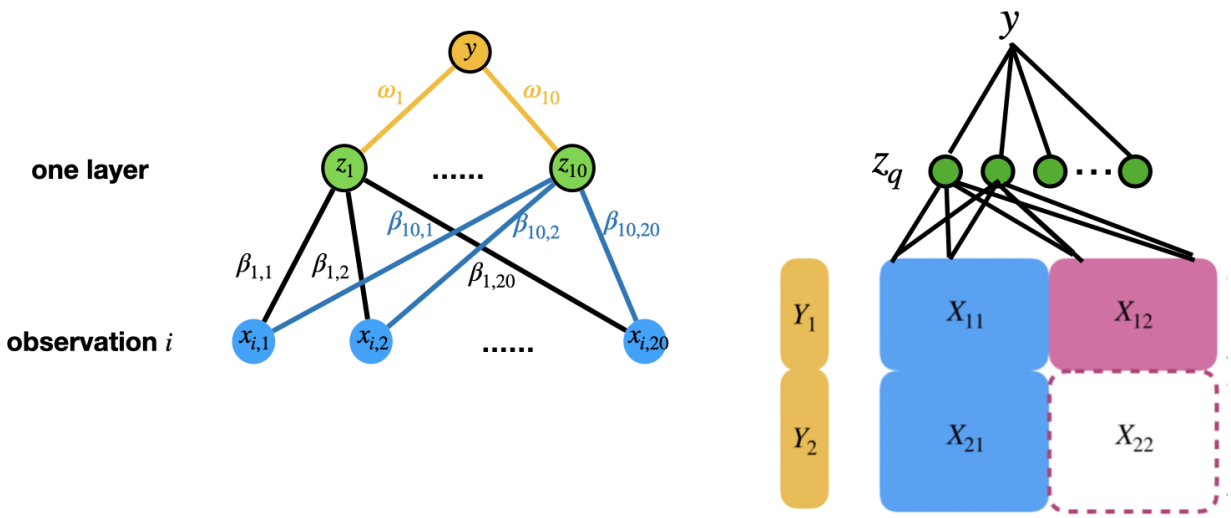

Figure 4.2: Neural network model with one layer

We design the neural network model as non-linear target function shown in Figure 4.2, $y_i = f(x_i) + \epsilon_i = \sum_{q=1}^{10} \omega_q z_{iq} + \epsilon_i = \sum_{q=1}^{10} \omega_q \sigma(x_i^T \beta_q) + \epsilon_i$ where $i = 1, ..., n$ and we set $q = 10$ for the additional layer $z_{iq} = \sigma(x_i^T \beta_q)$ with $\sigma(x_i^T \beta_q) = \frac{1}{1 + e^{-(x_i^T \beta_q)}}$. β is a 10×20 matrix with entries generated i.i.d. from a random normal distribution with mean 0 and standard deviation 5. For scenario (I), the last ten values of every β_q are set to zero. For scenario (II), the first ten values of every β_q are set to zero. ω is a 10×1 vector where all 10 elements are set equal to one.

We perform simulations with three different Signal-to-Noise-Ratios (SNR). To make the results more comparable across these scenarios, we fix the irreducible error $\text{Var}(\epsilon_i) = 1$ for all scenarios and rescale the conditional mean. That is, we set $y_i = c \cdot f(x_i) + \epsilon_i$, so the SNR is $\text{Var}(c \cdot f(x_i)) = c^2 \text{Var}(f(x_i))$. In each scenario, we set c to achieve the desired SNR (0.5, 1 or 2). We compute $\text{Var}(f(x_i))$ for each scenario by simulating 10000 observations x_i, and computing the variance of the resulting signal $f(x_i)$.

In summary, we use the following steps to generate data. For each SNR scenario:

- Step1. Generate 10000 predictors $X \sim N(\mathbf{0}, \mathbf{\Sigma})$ or $\chi^2_{\mathbf{1}}$.
- Step2. Calculate truth $Y^* = f(X)$ using X from Step1, and then estimate signal variance by $Var(Y^*)$.
- Step3. Calculate $c = \sqrt{SNR/Var(Y^*)}$ for SNR = 0.5/1/2.

For each replicate:

- Step4. Re-generate 1000 predictors X by the same procedure in the Step1.
- Step5. Generate noisy response Y by $y = c \cdot f(x) + \epsilon$ with $\epsilon \sim_{iid} N(0,1)$ where $f(x) = \sum_{q=1}^{10} \omega_q \sigma(x^T \beta_q)$.
- Step6. Mask values in X to make data consistent with specific missing patterns.

We use random forest for fitting all models M_0, M_1, M_2, M_α and MI. We use the "$tuning\_rf()$" function from the sklean package, to perform a grid search to select the tuning parameters for RF in this simulation. For a set of data by a given scenario, the 5-fold CV is used to search for the tuning parameters with the best score. More specifically, we fix the number of trees = 1000 (more trees), the fraction of features to consider when looking for the best split = 0.2, minimum leaf size = 10, and 5-fold CV is used to select the best "maximum depth" (maximum number of levels in each decision tree) parameter from the grid (6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28) for each scenario.

We calculate the results for model combination methods based on 5 fold CV.

We assess the performance using average MSE and SE over 100 replicates test data sets. For each scenario, we simulate 1000 complete test observations. For incomplete cases prediction, we evaluate average test MSE using the predictions from only the first block of predictors on the same test data set.

4.3.2 Results

The results for the normal predictor neural network model simulation are given in Tables 4.5, 4.6 and 4.7. The model combination method performs significantly better than MI when SNR=2 for model scenario (II). There are no significant differences between the performance of the model combination method and MI for the majority of scenarios for SNR=0.5 and 1. However, the model combination method outperforms MI in a significant majority of scenarios, even if the individual scenarios are not significant based on 100 replicates. When there is stronger signal, the model combination performs significantly better than MI for more cases. For the non-normal χ_1^2 predictor neural network model simulations in Table 4.8, the model combination method performs significantly better than MI in most scenarios

under strong SNR. Since the imputation could be biased if the predictors do not follow a multivariate normal distribution, the MI performs worse in this case.

Table 4.5: Test MSE and SE of neural network non-linear model with normal predictors, SNR = 0.5 on complete cases under MAR for the partial model M_0, the full model M_1, the model combination method M_α and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.101(0.049)	1.199(0.051)	1.102(0.049)	1.113(0.049)	1.153(0.052)	1.083(0.05)	1.153(0.052)	1.125(0.052)	1.094(0.05)	1.163(0.053)	1.086(0.05)	1.121(0.052)
(E)	1.155(0.049)	1.228(0.053)	1.157(0.049)	1.168(0.05)	1.307(0.06)	1.218(0.054)	1.218(0.055)	1.28(0.058)	1.138(0.05)	1.190(0.052)	1.13(0.049)	1.162(0.05)

Medium correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.182(0.054)	1.333(0.058)	1.186(0.054)	1.214(0.054)	1.168(0.05)	1.127(0.048)	1.141(0.049)	1.168(0.05)	1.221(0.053)	1.237(0.055)	1.221(0.054)	1.263(0.055)
(E)	1.073(0.046)	1.178(0.05)	1.078(0.046)	1.099(0.047)	1.239(0.055)	1.15(0.051)	1.141(0.05)	1.139(0.051)	1.226(0.059)	1.196(0.057)	1.208(0.058)	1.203(0.058)

Low correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.125(0.052)	1.34(0.058)	1.125(0.052)	1.213(0.054)	1.393(0.061)	1.337(0.058)	1.333(0.058)	1.348(0.059)	1.286(0.054)	1.282(0.052)	1.259(0.052)	1.26(0.052)
(E)	1.065(0.05)	1.142(0.051)	1.067(0.049)	1.078(0.05)	1.382(0.06)	1.117(0.05)	1.117(0.05)	1.192(0.054)	1.230(0.054)	1.236(0.053)	1.201(0.052)	1.203(0.052)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

Table 4.6: Test MSE and SE of neural network non-linear model with normal predictors, SNR = 1 on complete cases under MAR for the partial model M_0, the full model M_1, the model combination method M_α and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.07(0.045)	1.213(0.051)	1.071(0.045)	1.11(0.048)	1.473(0.06)	1.582(0.065)	1.47(0.059)	1.472(0.059)	1.097(0.048)	1.32(0.06)	1.096(0.049)	1.128(0.051)
(E)	1.225(0.055)	1.347(0.062)	1.225(0.055)	1.301(0.059)	1.212(0.058)	1.086(0.051)	1.069(0.051)	1.139(0.056)	1.190(0.053)	1.183(0.054)	1.190(0.053)	1.162(0.052)

Medium correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.201(0.054)	1.419(0.061)	1.213(0.054)	1.337(0.058)	1.322(0.059)	1.237(0.055)	1.346(0.058)	1.248(0.056)	1.613(0.066)	1.258(0.052)	1.258(0.052)	1.405(0.057)
(E)	1.164(0.051)	1.362(0.057)	1.165(0.051)	1.256(0.054)	1.37(0.057)	1.279(0.052)	1.276(0.052)	1.341(0.056)	1.453(0.062)	1.3(0.056)	1.299(0.056)	1.312(0.056)

Low correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.198(0.05)	1.511(0.063)	1.209(0.05)	1.387(0.058)	1.62(0.067)	1.404(0.058)	1.404(0.058)	1.464(0.061)	1.51(0.068)	1.512(0.066)	1.51(0.068)	1.45(0.064)
(E)	1.231(0.054)	1.424(0.063)	1.232(0.054)	1.374(0.061)	1.834(0.08)	1.404(0.061)	1.404(0.061)	1.511(0.068)	1.51(0.066)	1.367(0.063)	1.373(0.062)	1.388(0.062)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

Table 4.7: Test MSE and SE of neural network non-linear model with normal predictors, SNR = 2 on complete cases under MAR for the partial model M_0, the full model M_1, the model combination method M_α and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.521(0.065)	1.86(0.077)	1.543(0.066)	1.902(0.079)	1.731(0.079)	1.955(0.082)	1.715(0.078)	1.848(0.085)	1.474(0.066)	1.484(0.067)	1.438(0.065)	1.738(0.077)
(E)	1.125(0.055)	1.277(0.059)	1.154(0.054)	1.17(0.054)	2.34(0.107)	2.11(0.094)	2.106(0.095)	2.211(0.098)	1.298(0.063)	1.358(0.066)	1.256(0.061)	1.3(0.063)

Medium correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.221(0.057)	1.742(0.075)	1.256(0.058)	1.365(0.062)	2.414(0.102)	2.195(0.096)	2.155(0.093)	2.191(0.093)	2.48(0.104)	2.413(0.1)	2.311(0.096)	2.191(0.096)
(E)	1.321(0.058)	1.443(0.063)	1.324(0.057)	1.473(0.065)	2.067(0.089)	1.652(0.07)	1.662(0.07)	1.746(0.074)	2.285(0.097)	1.889(0.076)	1.901(0.077)	1.97(0.082)

Low correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	2.012(0.093)	2.248(0.104)	2.044(0.093)	2.31(0.108)	3.0(0.118)	2.24(0.092)	2.24(0.092)	2.589(0.102)	1.897(0.079)	2.555(0.099)	1.909(0.078)	2.027(0.081)
(E)	1.479(0.061)	1.922(0.078)	1.479(0.061)	1.533(0.063)	3.839(0.151)	2.291(0.091)	2.291(0.091)	2.727(0.109)	2.419(0.109)	2.272(0.094)	2.253(0.099)	2.209(0.096)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

Table 4.8: Test MSE and SE of neural network non-linear model with χ^2_1 predictors, SNR = 2 on complete cases under MAR for the partial model M_0, the full model M_1, the model combination method M_α and the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	1.741(0.089)	2.658(0.016)	1.774(0.009)	2.428(0.015)	3.0(0.02)	3.005(0.017)	2.928(0.016)	2.978(0.017)	2.984(0.041)	3.376(0.043)	2.874(0.041)	3.013(0.042)
(E)	1.613(0.008)	2.065(0.011)	1.633(0.008)	1.855(0.01)	2.87(0.016)	2.207(0.011)	2.207(0.011)	2.477(0.013)	3.168(0.043)	2.933(0.034)	2.872(0.035)	2.915(0.037)

Medium correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	2.17(0.012)	3.048(0.019)	2.267(0.014)	2.81(0.018)	3.115(0.019)	2.703(0.015)	2.68(0.017)	2.876(0.017)	2.95(0.039)	3.245(0.036)	2.808(0.038)	3.074(0.038)
(E)	1.898(0.011)	2.463(0.017)	1.948(0.012)	2.35(0.016)	2.491(0.015)	2.2(0.014)	2.2(0.014)	2.368(0.015)	3.071(0.051)	3.392(0.058)	3.095(0.052)	3.133(0.053)

Low correlation

	(I) Y related to first block				(II) Y related to second block				(III) Y related to both blocks			
	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI	M_0	M_1	M_α	MI
(D)	2.351(0.02)	3.512(0.031)	2.411(0.02)	3.508(0.031)	3.66(0.026)	3.075(0.02)	3.075(0.02)	3.189(0.021)	3.498(0.076)	3.661(0.079)	3.49(0.077)	3.716(0.079)
(E)	2.308(0.016)	3.16(0.024)	2.409(0.017)	3.186(0.022)	3.539(0.022)	2.676(0.017)	2.676(0.017)	2.879(0.018)	4.01(0.075)	3.828(0.072)	3.745(0.073)	3.753(0.064)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

4.4 Simulations for Missing Not at Random (MNAR)

4.4.1 Simulation design

One of the assumptions behind our method is that the data are missing at random (MAR). However, data which are missing not at random (MNAR) are very common, so it is important to see how robust our method is to this assumption violation. We use the same settings as for

the linear regression and logistic regression simulations above. However, instead of randomly selecting a certain number of observations that are missing the second block of predictors, we randomly select whether the second block is missing with probability depending on the response variable Y. In logistic regression, we set 60% missing when Y= yes and 40% missing when Y=no. In linear regression, we partition the Y value into 2 parts: values larger than the median; and values less than the median, and setting 60% missing in the top half and 40% missing in the bottom half. We still use average theoretical RMSE for linear regression and average test NLL for logistic regression over 100 replicate data sets. The sizes of the test data are all 1000.

4.4.2 Results

The simulation results show that the model combination method under MNAR for both linear and logistic regressions result in similar outcomes as in the provided by MAR case. Table 4.9 shows the theoretical RMSE and SE of linear prediction on complete cases under MNAR. The model combination method has significantly smaller average RMSE than that of MI in nearly all scenarios. The model combination method performs significantly worse than MI only when the variables are less correlated in large sample scenario (E) and model scenarios (II) and (III). Table 4.10 shows the test NLL and SE of logistic prediction on complete cases under MNAR. The model combination method $M_\alpha(CV)$ has significantly smaller NLL than that of MI in most cases for model scenarios (II) and (III). As expected, M_0 has smaller NLL for model scenario (I). There is no significant difference between the model combination method and the partial model or the model combination method and the MI method in this case. Therefore, we conclude that the model combination method is still applicable under MNAR.

In summary, given that we can't tell whether the missing data are MAR or MNAR, when there are over 50% cases with large block of missing, we recommend using the model combination method.

Table 4.9: Theoretical RMSE and SE of the linear regression model on complete cases under MNAR for the partial model M_0, the full model M_1, the combined model M_α estimated by CV estimation and Plug-in estimation, the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.757	1.998	1.051	0.916	1.384	1.434	1.786	1.314	1.222	1.385	1.367	2.434	1.405	1.259	1.732
	(0.016)	(0.047)	(0.025)	(0.015)	(0.033)	(0.009)	(0.036)	(0.023)	(0.016)	(0.027)	(0.014)	(0.062)	(0.021)	(0.017)	(0.042)
(B)	0.764	1.242	1.188	1.110	1.121	1.429	1.068	1.022	0.986	0.984	1.426	1.443	1.217	1.160	1.330
	(0.018)	(0.023)	(0.020)	(0.017)	(0.020)	(0.009)	(0.020)	(0.017)	(0.017)	(0.021)	(0.015)	(0.031)	(0.025)	(0.022)	(0.031)
(C)	0.222	2.413	1.097	0.245	1.062	1.228	2.195	1.254	0.997	1.220	1.086	2.886	1.401	0.966	1.415
	(0.005)	(0.052)	(0.005)	(0.022)	(0.033)	(0.001)	(0.055)	(0.041)	(0.012)	(0.027)	(0.002)	(0.060)	(0.046)	(0.010)	(0.034)
(D)	0.224	1.289	0.737	0.274	0.726	1.227	1.175	0.983	0.812	0.830	1.090	1.546	1.192	0.867	0.992
	(0.006)	(0.024)	(0.022)	(0.006)	(0.019)	(0.001)	(0.023)	(0.020)	(0.014)	(0.019)	(0.002)	(0.031)	3(0.027)	(0.011)	(0.023)
(E)	0.221	0.716	0.452	0.366	0.493	1.227	0.633	0.610	0.586	0.573	1.086	0.847	0.779	0.705	0.693
	(0.006)	(0.013)	(0.012)	(0.005)	(0.012)	(0.001)	(0.012)	(0.012)	(0.012)	(0.010)	(0.02)	(0.015)	(0.014)	(0.011)	(0.013)

Medium correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.857	2.407	1.494	1.064	1.808	1.847	2.014	1.602	1.491	1.764	1.812	2.614	1.763	1.572	2.082
	(0.020)	(0.058)	(0.034)	(0.019)	(0.045)	(0.011)	(0.046)	(0.027)	(0.021)	(0.043)	(0.017)	(0.059)	(0.035)	(0.024)	(0.051)
(B)	0.889	1.460	1.301	1.223	1.299	1.844	1.203	1.202	1.144	1.245	1.794	1.615	1.460	1.409	1.549
	(0.019)	(0.024)	(0.020)	(0.019)	(0.023)	(0.010)	(0.021)	(0.019)	(0.019)	(0.022)	(0.015)	(0.026)	(0.022)	(0.020)	(0.029)
(C)	0.258	2.893	1.380	0.289	1.432	1.638	2.329	1.546	1.234	1.515	1.482	3.173	1.957	1.295	1.958
	(0.006)	(0.065)	(0.047)	(0.006)	(0.042)	(0.001)	(0.045)	(0.037)	(0.018)	(0.032)	(0.002)	(0.056)	(0.048)	(0.013)	(0.060)
(D)	0.261	1.569	0.652	0.315	0.852	1.639	1.286	1.124	0.972	1.140	1.484	1.745	1.105	1.070	1.167
	(0.006)	(0.027)	(0.024)	(0.006)	(0.026)	(0.001)	(0.026)	(0.024)	(0.017)	(0.023)	(0.002)	(0.028)	(0.024)	(0.014)	(0.027)
(E)	0.258	0.848	0.572	0.431	0.572	1.637	0.703	0.687	0.670	0.842	1.482	0.939	0.886	0.828	0.856
	(0.007)	(0.014)	(0.012)	(0.007)	(0.008)	(0.001)	(0.013)	(0.012)	(0.012)	(0.013)	(0.001)	(0.015)	(0.013)	(0.011)	(0.018)

Low correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(A)	0.852	2.378	1.126	0.977	1.805	2.161	2.215	1.791	1.682	1.979	2.020	2.595	1.905	1.737	2.212
	(0.017)	(0.056)	(0.034)	(0.016)	(0.045)	(0.011)	(0.044)	(0.028)	(0.024)	(0.039)	(0.015)	(0.050)	(0.031)	(0.023)	(0.045)
(B)	0.815	1.428	1.271	1.283	1.374	2.171	1.237	1.279	1.203	1.258	2.001	1.594	1.502	1.448	1.572
	(0.019)	(0.026)	(0.020)	(0.018)	(0.024)	(0.010)	(0.022)	(0.021)	(0.022)	(0.023)	(0.011)	(0.027)	(0.022)	(0.021)	(0.029)
(C)	0.248	2.834	1.311	0.271	1.476	1.960	2.491	1.574	1.420	1.655	1.711	3.280	1.855	1.423	1.990
	(0.006)	(0.064)	(0.059)	(0.006)	(0.047)	(0.001)	(0.051)	(0.040)	(0.020)	(0.036)	(0.001)	(0.072)	(0.052)	(0.017)	(0.048)
(D)	0.256	1.541	1.019	0.309	0.941	1.960	1.346	1.097	1.043	1.026	1.710	1.672	1.191	1.139	1.170
	(0.006)	(0.025)	(0.024)	(0.006)	(0.022)	(0.001)	(0.024)	(0.023)	(0.019)	(0.022)	(0.001)	(0.030)	(0.028)	(0.017)	(0.026)
(E)	0.246	0.832	0.652	0.399	0.633	1.958	0.730	0.723	0.710	0.657	1.713	0.896	0.859	0.821	0.763
	(0.005)	(0.014)	(0.012)	(0.005)	(0.014)	(0.001)	(0.013)	(0.012)	(0.012)	(0.011)	(0.001)	(0.016)	(0.015)	(0.014)	(0.013)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

4.5 Real data Application

4.5.1 Public school data

This application is an example of how the blockwise missing problem can occur naturally when integrating data from multiple resources. In New York State, the grades 3-8 mathematics assessments measure the higher learning standards and reflect students' progress toward college and career readiness. We build a model to predict the average Math Score for a school from its demographic and school survey data, shown in Figure 4.3. The mean scale Math

Table 4.10: Test NLL and SE of the logistic regression model on complete cases under MNAR for the partial model M_0, the full model M_1, the combined model M_α estimated by CV estimation and Plug-in estimation, the multiple imputation method MI. The corresponding SE are given in parentheses.

High correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	431.101	9725.985	436.115	431.35	431.998	630.438	11828.015	626.349	628.39	624.693	522.055	10854.348	523.587	522.192	519.478
	(1.756)	(253.274)	(2.223)	(1.753)	(1.79)	(1.215)	(245.236)	(2.160)	(1.473)	(1.26)	(1.457)	(203.676)	(1.767)	(1.475)	(1.506)
(D)	430.092	3229.793	433.38	432.115	431.039	628.78	1786.443	587.150	623.771	614.39	520.905	2340.815	516.483	520.942	514.029
	(1.877)	(401.059)	(2.042)	(1.963)	(1.92)	(1.324)	(263.344)	(2.682)	(1.955)	(1.372)	(1.554)	(309.124)	(1.823)	(1.624)	(1.589)
(E)	431.701	524.821	437.301	438.662	433.046	629.149	584.614	540.005	626.119	590.951	521.875	541.643	489.755	521.235	502.669
	(1.728)	(3.856)	(1.734)	(2.036)	(1.726)	(1.168)	(3.916)	(2.115)	(1.791)	(1.243)	(1.542)	(4.271)	(1.965)	(1.829)	(1.513)

Medium correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	420.968	10130.977	428.4489	421.307	422.012	640.929	11569.196	633.281	638.844	634.847	527.408	10849.455	527.053	527.008	524.417
	(1.478)	(192.461)	9(2.21)	(1.485)	(1.513)	(0.918)	(223.227)	(2.317)	(1.677)	(1.15)	(1.788)	(197.553)	(2.086)	(1.843)	(1.907)
(D)	418.502	2671.105	424.445	419.551	419.693	642.169	1976.365	594.055	637.938	627.171	531.009	2507.309	523.706	530.566	524.246
	(1.473)	(326.686)	(2.945)	(1.54)	(1.509)	(1.074)	(299.942)	(2.587)	(1.641)	(1.25)	(1.624)	(348.096)	(1.873)	(1.721)	(1.686)
(E)	420.285	510.549	425.797	429.554	421.865	642.44	581.612	538.921	637.849	600.766	526.916	548.385	493.572	526.186	509.731
	(1.951)	(4.482)	(2.145)	(2.559)	(2.008)	(1.052)	(4.416)	(2.333)	(1.804)	(1.132)	(1.452)	(3.939)	(2.021)	(1.832)	(1.545)

Low correlation

	(I) Y related to first block					(II) Y related to second block					(III) Y related to both blocks				
	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI	M_0	M_1	M_α(CV)	M_α(plugin)	MI
(C)	421.689	10177.69	429.37	422.005	422.618	672.568	11088.448	651.720	666.479	667.738	558.443	10903.738	552.992	557.795	555.099
	(1.856)	(239.005)	(2.375)	(1.851)	(1.877)	(0.694)	(254.815)	(2.196)	(1.432)	(0.869)	(1.712)	(199.633)	(2.207)	(1.825)	(1.811)
(D)	421.68	4175.722	428.258	423.037	422.764	672.436	2143.496	594.246	662.96	652.282	559.813	2144.011	544.384	558.317	549.75
	(1.879)	(424.98)	(2.211)	(1.926)	(1.899)	(0.795)	(308.093)	(2.518)	(2.055)	(0.934)	(1.425)	(314.089)	(2.396)	(1.564)	(1.475)
(E)	419.662	508.443	425.101	426.047	421.069	672.974	558.295	521.793	661.501	618.65	561.177	553.632	502.248	561.151	531.818
	(1.771)	(4.434)	(1.996)	(2.162)	(1.803)	(0.729)	(4.58)	(2.452)	(2.036)	(0.83)	(1.452)	(4.477)	(2.372)	(1.743)	(1.445)

Note: The underlined entry denotes the best performance for that particular setting. The red and blue colours indicate cases where the M_α model performs significantly better and worse respectively than MI based on the one sample t-test.

Score (range 148-423) is calculated from all students in grades 3-8 for each school. Data are from two resources: New York City Department of Education NYCDE (2017) and New York State Education Department NYSED (2017). NYSED provides demographic predictors for all schools in New York State such as total student enrollment in school, percentage of White, Black, Hispanic and Asian, percentage of female, percentage of English Language Learners and percentage of poverty. NYCDE provides extra information from the school survey. These predictors are only available for schools in New York City. This information includes percentage of disability, Collaborative Teachers Score, Effective School Leadership Score, Rigorous Instruction Score, Supportive Environment Score, Strong Family-Community Ties Score and Trust Score. The test results are correlated with both student characteristics (demographic predictors) and the learning environment, school evaluations, student assessments and other scores (school survey predictors), which utilizes feedback from students, teachers, and parents received through the annual NYC School Survey. The data has 3107 rows and 16 predictors where 948 rows are New York City schools with all 16 predictors and 2159 rows

are other schools in New York State with only 8 predictors. The predictors and response are shown in Figure 4.4.

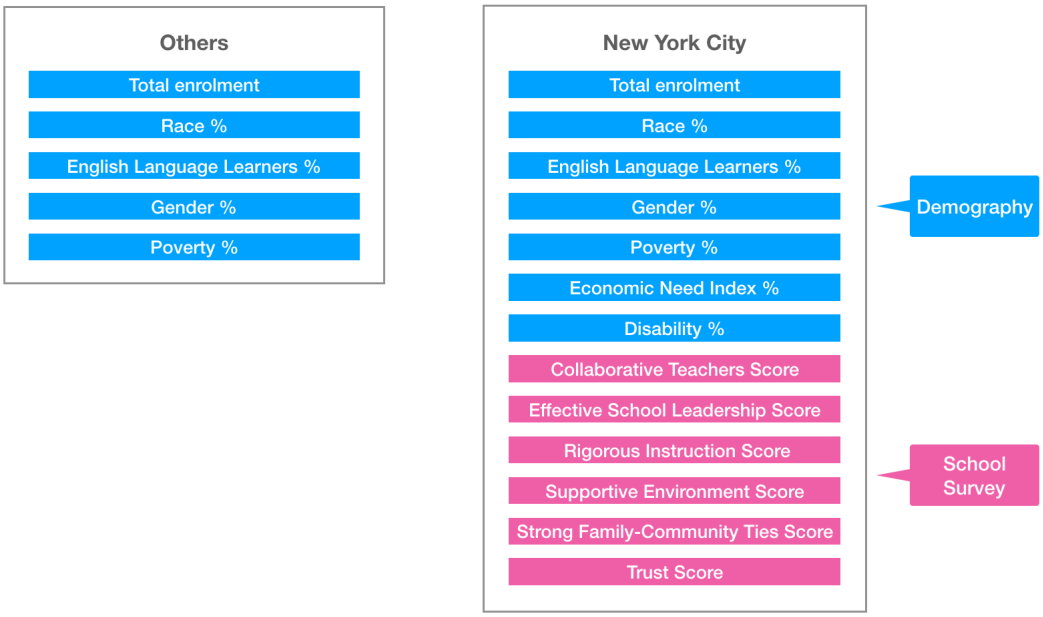

Figure 4.3: Predictors of demography and school survey.

We use stratified random train-test splitting taking 75% of complete and incomplete data as training and the remaining 25% as testing. We still use LOOCV to estimate α and 20 imputation replicates in MI. The Math Score is the response variable and its scores range from 148 to 423. We centered and scaled it between -2.3 and 3.3.

Because of the nonlinear relationship between predictors and the response, we fit the regression model including both first and second order polynomials of the predictors with their interaction terms. We perform ANOVA with F-test in order to determine whether interaction terms are helpful to polynomial regression models. Model 1 has only the first and second order terms. Model 2 adds interaction terms. We find that there is compelling evidence that the polynomial with interaction terms is better than the polynomial without, pvalue $= 2.2 \times 10^{-16}$ for partial model M_0 and pvalue $= 0.0031$ for full model M_1. The polynomial regression result is also better than the linear model and the GAM (Generalized Additive Models) model. Here is our regression model:

```
lmFit=lm(Math ~ (Total+White+Black+Hispanic+Asian+Female
                 +English.Language.Learners+Poverty
                 +Disability+Economic.Need.Index
```

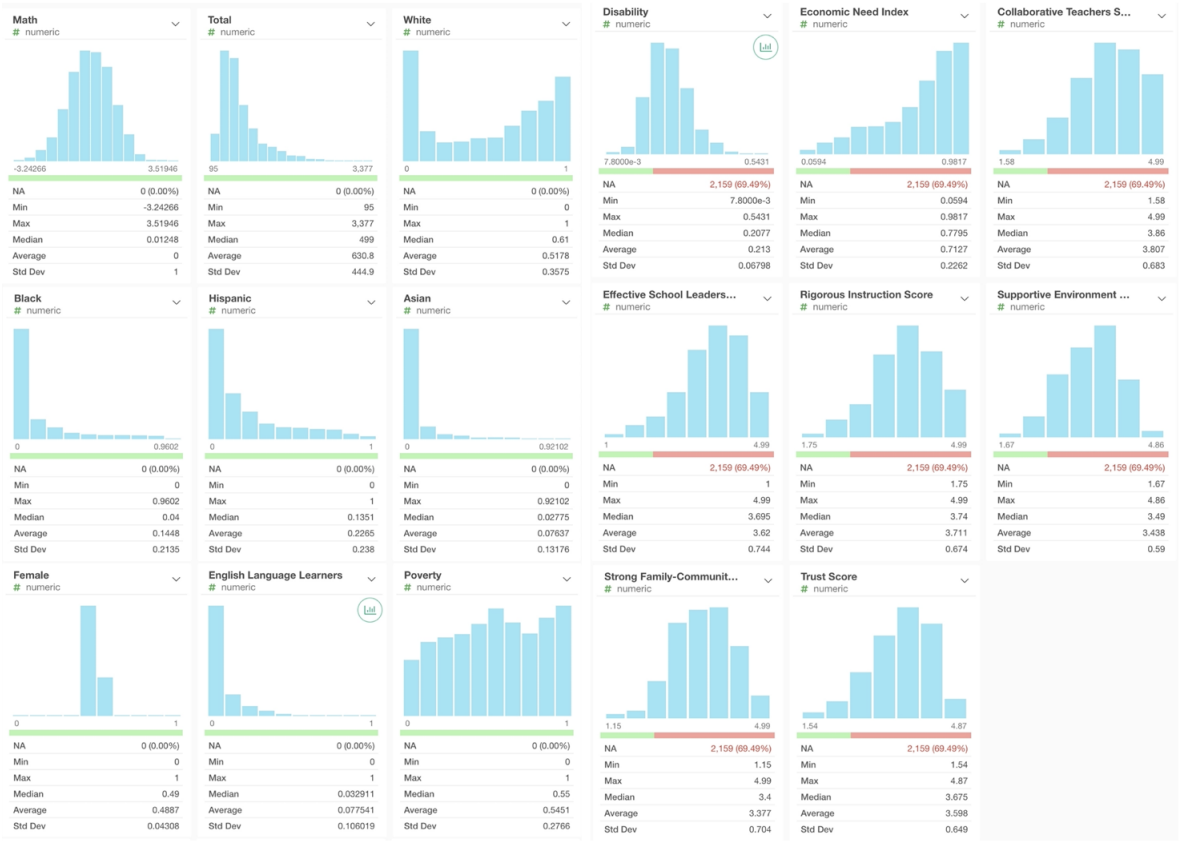

Figure 4.4: Histogram for the predictors and response

```
+Collaborative.Teachers.Score
+Effective.School.Leadership.Score
+Rigorous.Instruction.Score
+Supportive.Environment.Score
+Strong.Family.Community.Ties.Score
+Trust.Score)^2
+I(Total^2)+I(White^2)+I(Black^2)+I(Hispanic^2)
+I(Asian^2)+I(Female^2)+I(English.Language.Learners^2)
+I(Poverty^2)+I(Disability^2)+I(Economic.Need.Index^2)
+I(Collaborative.Teachers.Score^2)
+I(Effective.School.Leadership.Score^2)
+I(Rigorous.Instruction.Score^2)
+I(Supportive.Environment.Score^2)
+I(Strong.Family.Community.Ties.Score^2)
```

Table 4.11: Test RMSE of public school data in New York State on all cases, complete cases and incomplete cases for the partial model M_0, the full model M_1, the model combination method M_α (CV estimation) and the multiple imputation method MI.

	M_0	M_1	M_α	MI
Testing data (complete cases)	0.744	0.914	0.612	0.960
Testing data (incomplete cases)	0.605		0.605	0.630
Testing data (all cases)	0.640	0.703	0.607	0.614

```
+I(Trust.Score^2), data = TRAIN)
```

More predictors are introduced in the models by including the quadratic terms and interactions, thus there are 44 predictors in M_0 and 152 predictors in M_1.

Polynomial regression is technically a special case of linear regression, so our theory from Section 3.4 can be applied if MAR can be assumed. However, the MAR assumption is doubtful here because obviously missingness happens on all schools not in New York city. Table 4.11 illustrates the test RMSE results. The partial model M_0 RMSE is derived by using only first block of predictors, i.e. demographic predictors. The full model M_1 RMSE for all cases is derived by using M_1 for complete case prediction and using M_0 for incomplete case prediction. The model combination method M_α RMSE for all cases is calculated by replacing the full model by the combined model for complete cases. The MI RMSE is obtained using the multiple imputation method, with the same polynomial regression model. The estimation of the model combination parameter is 0.709. Table 4.11 demonstrates that model combination prediction performs better than MI, the partial model M_0 or the full model M_1.

4.5.2 Abdominal pain diagnosis example

We have already demonstrated how model combination method performed in a regression application. Now we apply the method to a subset of the abdominal pain diagnosis data set in order to evaluate the model combination method for two blocks of predictors with block missing in a classification application.

Abdominal pain can represent a spectrum of conditions from benign and self-limited diseases to surgical emergencies. Evaluating abdominal pain requires an approach that relies on the likelihood of disease, triage assessment and laboratory tests. Therefore, it is difficult to distinguish some diagnoses. This data was abstracted from the Emergency Department

Table 4.12: Test classification accuracy of abdominal pain data on all cases (complete cases and incomplete cases) for the partial model M_0, the full model M_1, the model combination method M_α (CV estimation) and the multiple imputation method MI.

Diagnosis	M_0	M_1	M_α	MI
Bleeding in early pregency vs. Incomplete abortion	0.853	0.853	0.857	0.855
Biliary Colic vs. Pancreatitis	0.706	0.729	0.731	0.727
Hematuria vs. Other urologic	0.830	0.831	0.832	0.818

Information System (EDIS) where the triage assessment is the first block of predictors including gender, DBP (diastolic blood pressure), temperature, patients' complaint, etc. The complete blood count - a set of laboratory tests - is the second block of predictors including WBC (white blood cells), RBC (Red blood cells), Hemoglobin, etc. All patients have triage assessment, but only some of the patients have complete blood count results. We describe the details of this data in Section 5.1.1.

We choose three pairs of diagnoses which are difficult to distinguish: Bleeding in early pregency and Incomplete abortion from pelvic pathology; Biliary Colic and Pancreatitis from general abdominal pathology; Hematuria and Other urologic from renal system pathology. The sample sizes for these pairs are 2607, 1538 and 2142 respectively. For each pair, about half of the patients have CBC results. We construct the M_0 model for all observations with triage variables and the M_1 model for the observations with all variables (triage and CBC). We use stratified random train-test splitting taking 75% of complete and incomplete data as training and the remaining 25% as testing. We use GBM to fit all models and use LOOCV to estimate α. We do 20 imputation replicates in the MI method.

The comparison in terms of test classification accuracy is summarized in Table 4.12. The classification accuracy results based on all testing data include both complete cases and incomplete cases as in Section 4.5.1. The partial model M_0 accuracy is derived by using only the first block of predictors. The full model M_1 accuracy for all cases is derived by using M_1 to predict the complete cases and M_0 to predict the incomplete cases. The model combination method M_α accuracy for all cases is calculated by replacing the full model by the combined model for complete cases. The MI method accuracy is obtained using the multiple imputation method. In this dataset, the missing pattern is not at random, because tests are ordered by doctors in cases where they believe the results will help them to diagnose the patients. The results of four methods presented in Table 4.12 show that the model

combination method performs better than either MI, the partial model or the full model on test data.

More detailed application on the abdominal pain diagnosis problem by using the model combination method with multi-block variables will be given in the next chapter.

Chapter 5

Application on the Abdominal Pain Diagnosis Problem

In this chapter, we apply the model combination method developed in this thesis to analyse a real data set containing abdominal pain patients at emergency departments in Nova Scotia. This data set presents a number of challenges beyond the standard block missing problem. Firstly, this is a multiclass classification problem with 39 classes, which is challenging for standard classification methods, so we develop a hierarchical tree structure over 39 diagnoses based on a combination of the data and medical knowledge and thus divide a classification problem for a large number of classes into many classification problems, each with a low number of classes.

Secondly, there are multiple blocks of predictors, and patients can have different combinations of blocks. The model combination method in Chapter 3 was designed to handle a single block of missing predictors, so to apply it in this situation requires us to extend the method to handle multiple missing blocks. After these modifications, our method is able to make informative and accurate predictions of the probability of each diagnosis. Using this, we are able to create a short list of most plausible diagnoses. This list is usually fairly short, and has high coverage of the true diagnosis.

5.1 Abdominal Pain Data Analysis

5.1.1 Data Introduction

The present study is a retrospective review of 116,008 presentations to the Emergency departments (ED) in Nova Scotia. These presentations were diagnosed with one of 39 ICD-9 physician diagnoses related to abdominal pathology. The data set was abstracted from the Emergency Department Information System (EDIS) and includes four blocks of variables: triage variables, complete blood count (CBC) variables, liver enzyme variables and radiology variables with 42 predictors having both numeric and text values. The triage variables are available for all patients, but the other blocks are available only if ordered by the physician.

For a particular patient, any combination of these blocks might be available.

The variables in each block are as follows:

- Triage variables: patient age, patient gender, presenting ICD-9 complaint, Canadian Triage and Acuity Scale (CTAS) Score, vitals on presentation (blood pressure, oxygen saturation, respiratory rate, pulse rate, pain scale score, Glasgow Coma Scale (GCS) score, temperature), time of year and time of day. Bullard et al. (2008)

- CBC (complete blood count) variables: RBC (red blood cells), WBC (white blood cells), Hgb (hemoglobin), Plt (platelet count) and white blood cell differential (neutrophils, lymphocytes, monocytes, eosinophils and basophils), which are the relative proportion of each leukocyte type in the blood.

- Liver enzyme variables: alanine aminotransferase (ALT) and aspartate aminotransferase (AST), gamma-glutamyl transpeptidase (GGT), bilirubin and alkaline phosphatase (ALP).

- Radiology variables: This consists of unstructured text giving radiologist reports from various diagnostic imaging technologies: computed tomography (CT), ultrasound, magnetic resonance imaging (MRI) and X-ray (radiography). In order to use the text reports in our classifier, we need to convert them to numerical variables. Our collaborators in the Computer Science department developed a method based on Conditional Random Fields (Wallach, 2004) and structured perceptron (McDonald et al., 2010) to convert the free text data into 1620 random variables with 4 levels: 0 = not mentioned; -1 = explicitly described as not observed; 0.5 = observed but not critical; and 1 = critical and observed.

The data was divided into training, validation and testing sets. The training set contains 71,396 cases from January 2010 to June 2013 (three and a half years), the validation set 16,669 cases from July 2013 to February 2014 (eight months), and the testing set 25,441 cases from March 2014 to February 2015 (one year).

5.1.2 Data exploration

From Figure 5.1, We see that the majority of patients are female, and this is true for nearly all ages. We also see that patients tend to be younger, with the number of patients decreasing

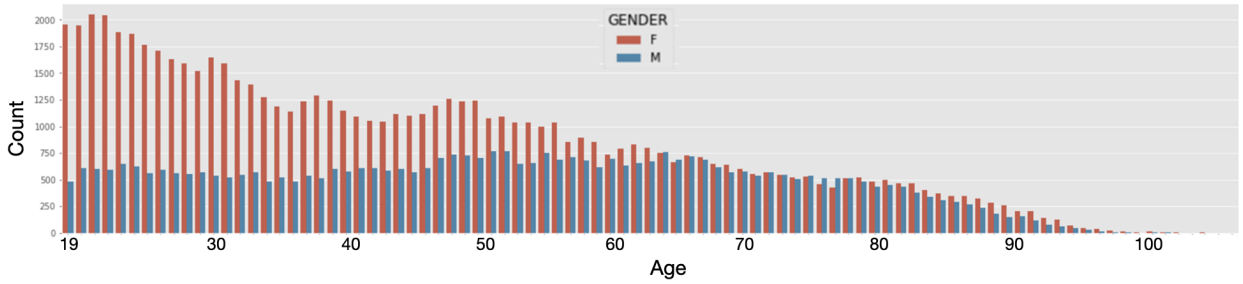

Figure 5.1: Gender & Age

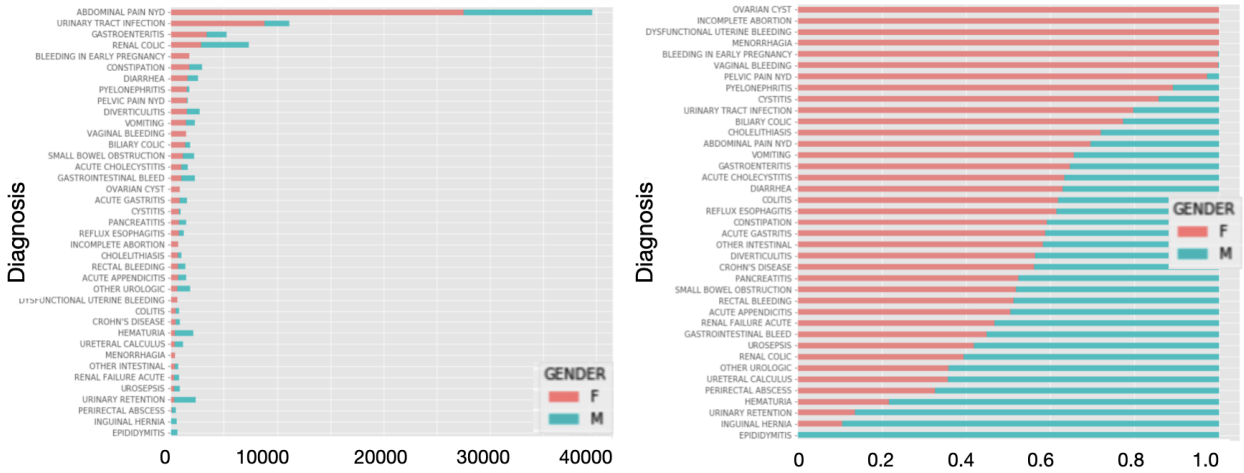

Figure 5.2: Gender & Diagnosis

with age for both male and female patients.

Figure 5.2 and 5.3 summarises the diagnoses present in the training and test data set by gender and Age. The frequency of different diagnoses varies a lot with gender, with several diagnoses only affecting one gender. The most common diagnosis for both genders was Abdominal Pain not yet diagnosed (NYD). The diagnoses which are most female-dominated are reproductive system diseases, pelvic disease, pyelonephritis and cystitis. The most male-dominated diagnoses are Epididymitis, Inguinal Hernia, Urinary Retention, Hematuria and Perirectal Abscess. The frequency of various diagnoses also varies with age. While Abdominal Pain NYD and Urinary Tract Infection are the most common diagnoses for all ages, gastroenteritis is among the most common for young patients, and renal colic is one of the most common conditions in middle-aged patients. For elderly patients, Constipation, Gastrointestinal Bleeding, Hematuria, and Urinary Retention were among the most common diagnoses.

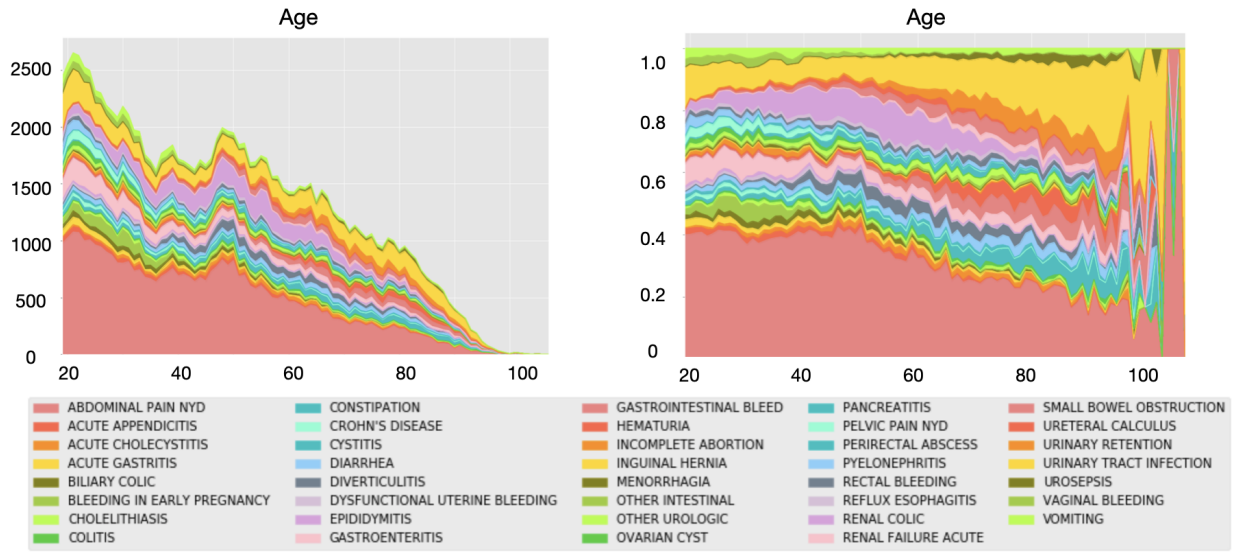

Figure 5.3: Age & Diagnosis

We next examine the predictor variables. Figure 5.4 shows the distribution of temperature, CTAS and pain for each diagnosis. CTAS (Severity of the condition) is an overall assessment of the patient's condition. It is on an integer scale between 1 and 5, with lower numbers indicating a more serious condition. Nearly all the patients in the dataset have CTAS in the range 2-4. CTAS is clearly associated with Diagnosis, with Urinary Sepsis, Gastrointestinal Bleeding, Ureteral Stones having lower average CTAS scores than other diagnoses. CTAS is also negatively correlated with temperature and pain score. Hyperthermia is common for Urosepsis and Pyelonephritis, but rare for other diagnoses. Most patients have a pain score in the range 5-8. However, Ureteral Calculus, Renal Colic and Crohn's Disease have higher pain scores, and other diagnoses such as Diarrhea, Rectal Bleeding, Acute Renal Failure, Urinary Sepsis, and Gastrointestinal Bleeding, Vaginal Bleeding, Bleeding in Early Pregnancy, Rectal Bleeding and Vomiting often have lower pain scores.

Figure 5.5 shows the distribution of Glasgow Coma Scale (GCS), and blood pressure measurements for each diagnosis. The GCS, a score from 3 to 15, is a neurologic assessment of a patient's level of consciousness. It was designed to differentiate between coma and other states of impaired consciousness. A GCS value of 15 indicates a fully awake patient, while a GCS value of 3 indicates deep coma or a brain-dead state. In this data, we found that the vast majority of patients in the data set have GCS of 15. However, for a few diagnoses,

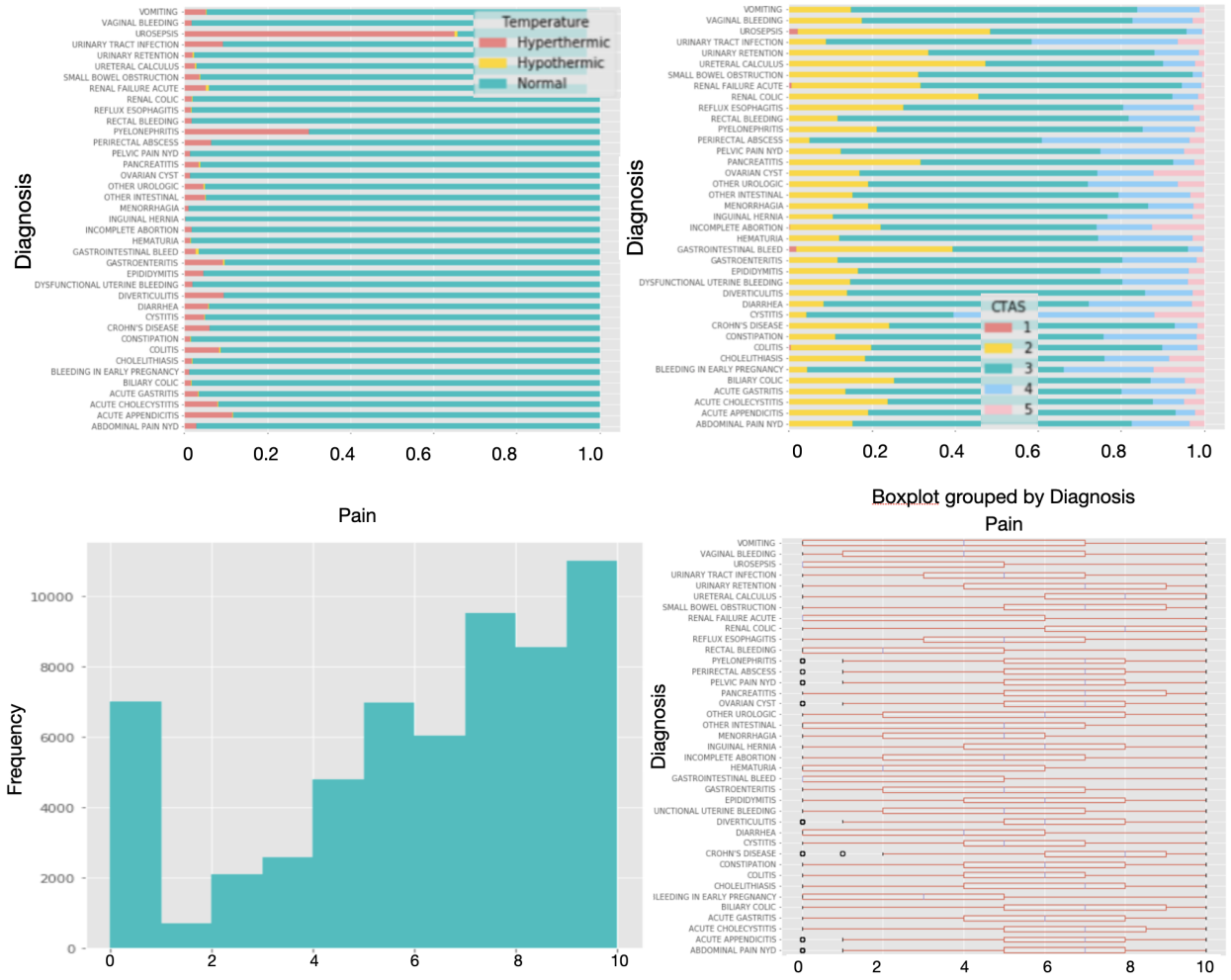

Figure 5.4: Temperature, CTAS (Severity of the condition) and Pain

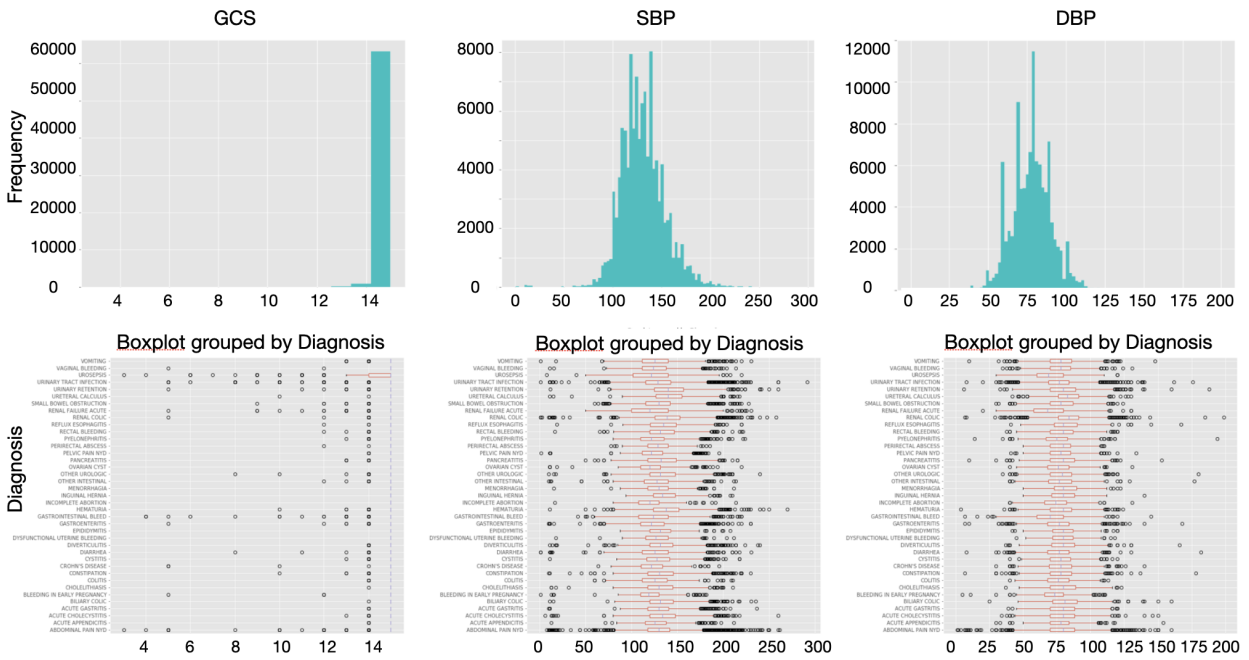

Figure 5.5: Glasgow Coma Scale (GCS), SBP and DBP

such as Urosepsis and Abdominal Pain NYD, lower GCS scores are occasionally observed. Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) are two measurements of a patient's blood pressure. From Figure 5.5, we see that both are approximately normally distributed, and the distributions are fairly similar across diagnoses. However, there are a few diagnoses with lower average SBP and DBP, namely, Urosepsis, Acute Renal Failure, Gastrointestinal Bleeding and Bleeding in Early Pregnancy.

Figure 5.6 shows the distribution of Respiration Rate (RR), Heart Rate (HR) and Blood Sugar (Glucose) for all patients. RR (respiration rate) is the measure of how often an individual breathes in a minute. The value is usually between 15-20. Urosepsis may feature increased RR values because it causes inflammation in the lungs which results in more rapid breathing. For most other diagnoses, RR has a fairly similar distribution - mostly in the 15-20 range with a few large outliers. Glucose has a skewed distribution, which varies somewhat between diagnoses, with Perirectal Abcess patients having high Glucose, and diagnoses relating to pregnancy having lower Glucose levels. HR (heart rate) measures the rate at which the heart beats. There are also about 1700 out of 116000 missing values for this variable in the data. These are cases where it is impossible to record the patient's heart rate. We will treat these as missing at random.

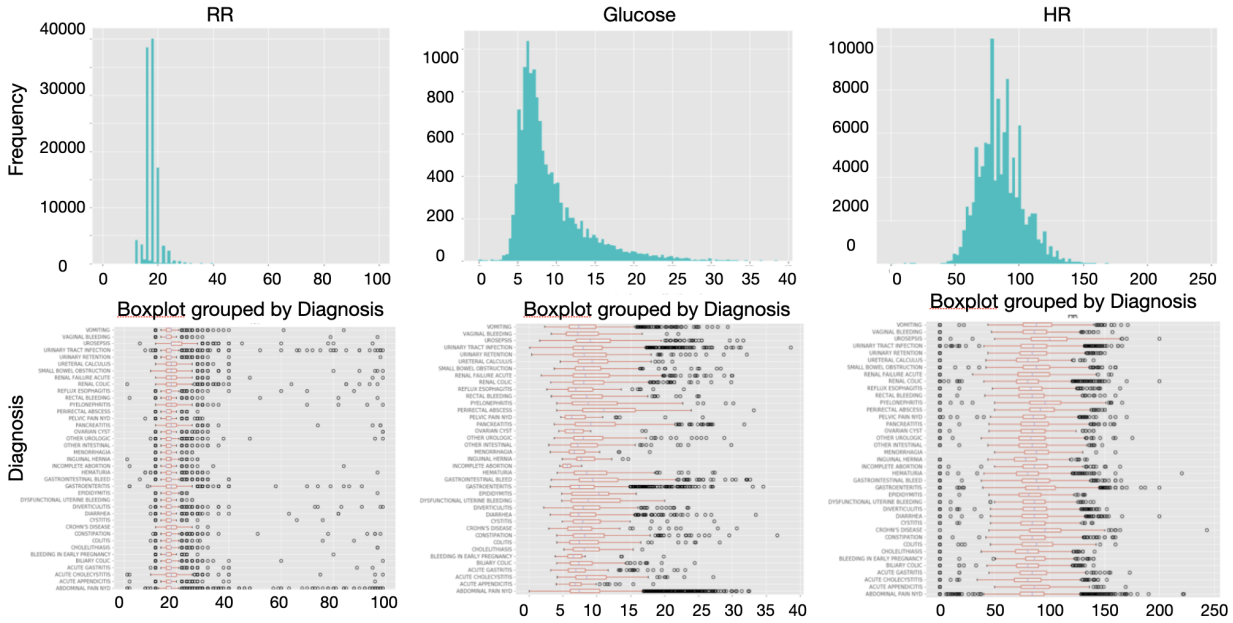

Figure 5.6: Respiration rate Glucose and Heart rate

Figure 5.7 shows the distributions of the main variables in the complete blood count. These are often used by the physician to diagnose a number of conditions. From the figure we see that Red Blood Cells (RBC) has a fairly normal distribution, that is similar for most diagnoses, but is often lower for Gastrointestinal Bleeding or Acute Renal Failure. White Blood Cells (WBC) has a slightly skewed distribution, with some outliers, and is similar for most diagnoses, but is higher for Urosepsis, Pyelonephritis, Perirectal Abcess, Pancreatitis, Epididymitis, Diverticulitis, Crohn's Disease, Cholitis, Acute Cholecystitis and Acute Appendicitis. Hemoglobin is fairly normally distributed with similar distributions for most diagnoses, but reduced levels for Acute Renal Failure and Gastrointestinal Bleeding.

Figure 5.8 shows the distribution of other variables in the Complete Blood Count. These all have skewed distributions with some outliers. The platelet count was lower in patients with Urosepsis, and higher in patients with Crohn's Disease. Lymphocytes and Neutrophils are percentages, so are constrained to sum to at most 100. Indeed, in most cases, the total of lymphocytes and neutrophils is very close to 100. The distribution of these percentages varies between diagnoses. For example, Urosepsis patients mostly had, on average, a lower percentage of lymphocytes and a higher percentage of neutrophils than other patients, while patients with gynocological conditions tended to have higher lymphocyte percentages and lower neutrophil percentages.

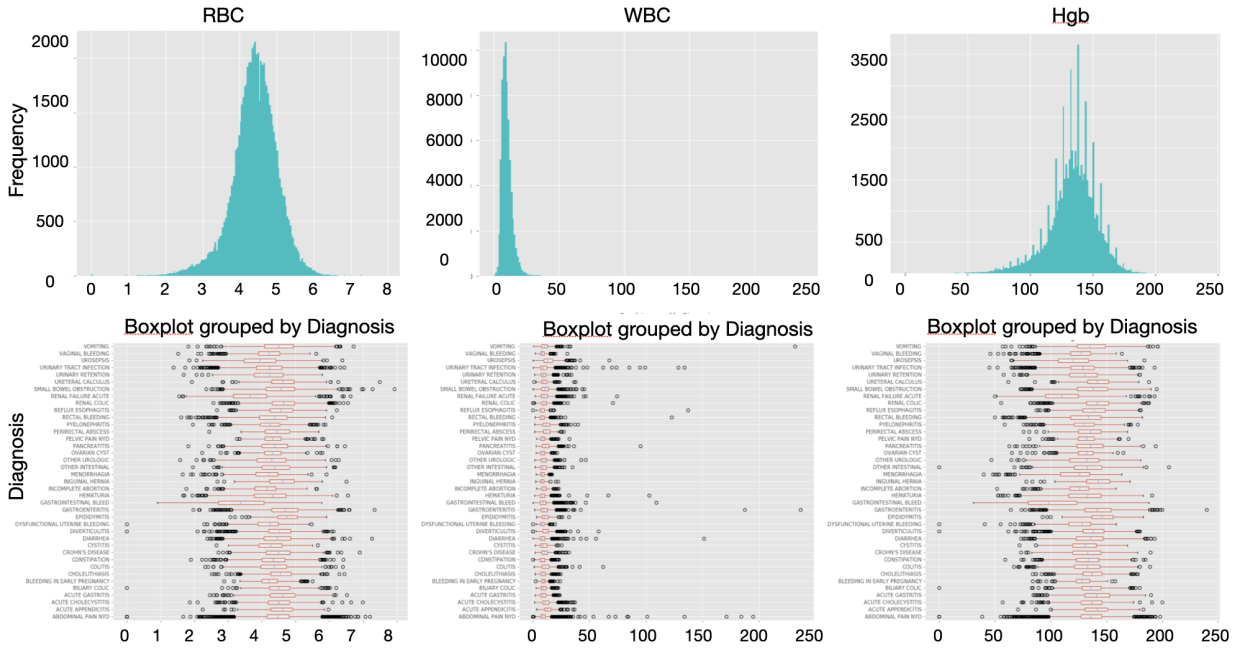

Figure 5.7: RBC, WBC and Hemoglobin

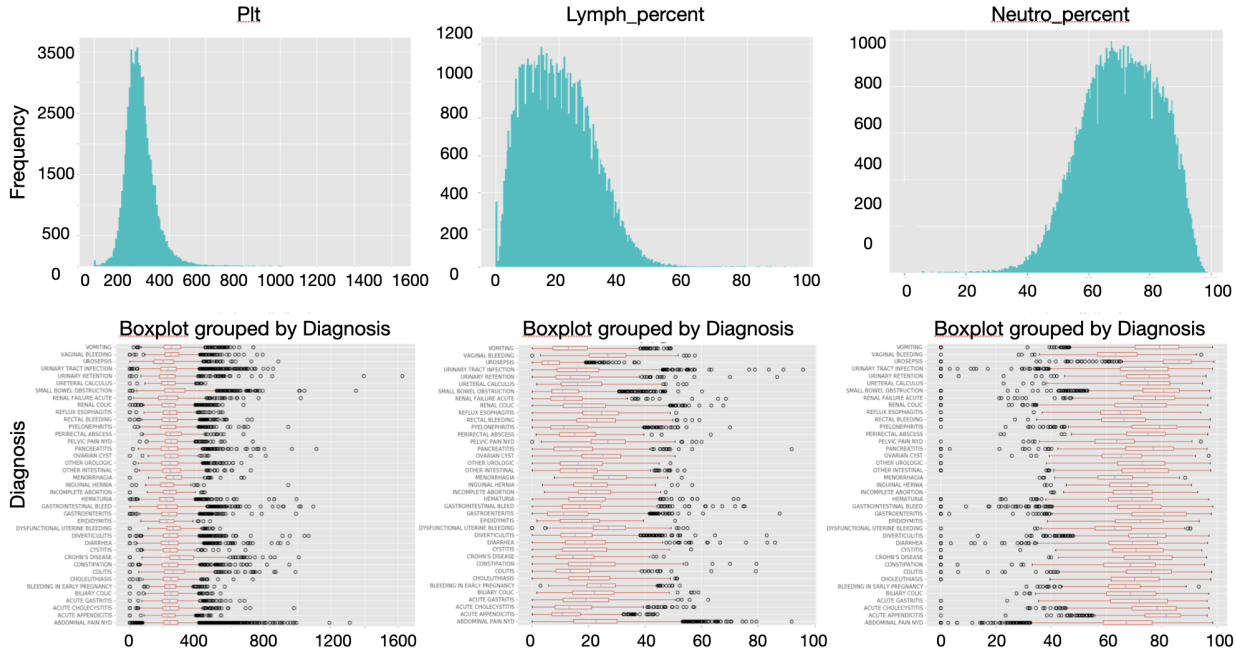

Figure 5.8: Platelet, Lymphocytes and Neutro percent

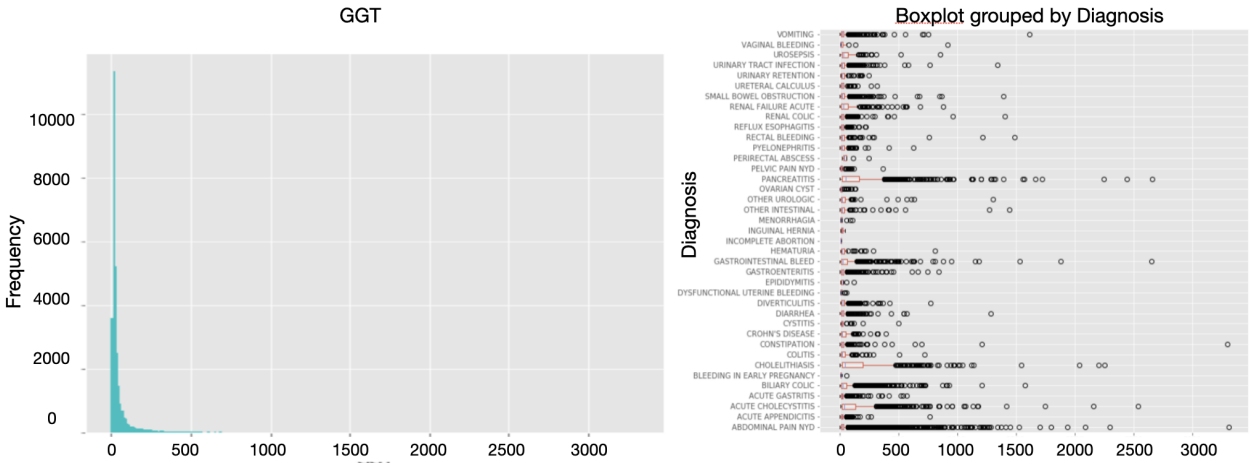

Figure 5.9: Glutamyl transpeptidase (GGT)

We next look at the liver enzyme variables. Figure 5.9 shows the distribution of Glutamyl transpeptidase (GGT). GGT is a liver enzyme that can be increased as a result of certain diseases. GGT is an enzyme found primarily in the liver and pancreas, and it is also found in the bile ducts and the small intestine. The data shows that high GGT levels are present in patients with Pancreatitis, Acute Cholecystitis, and Gallstones. Pathologically elevated GGT levels are seen in obstructive diseases of the biliary tract, acute and chronic alcoholic hepatitis, drug-induced hepatitis. Alcoholics may have decreased GGT levels after they stop drinking. Pathologically elevated GGT can also indicate pancreatic tumors, prostate tumors and other types of cancers. GGT is abnormal when it is greater than 50 U/L. Except abdominal pain NYD, the top 10 diagnosis with abnormal GGT are gastrointestinal pathology and urinary tract infection, shown in Figure 5.10.

In Liver function test results, normal total bilirubin is 3.42-20 μmol/L. Physiological elevated levels of bilirubin can be observed in neonatal jaundice, while pathologically elevated levels of bilirubin may be observed in a number of conditions including biliary tract obstruction, viral hepatitis A and other types of viral hepatitis, cholestasis hepatitis, acute alcoholic hepatitis and inherited abnormal bilirubin metabolism such as Gilbert syndrome. Total bilirubin is abnormal when it is greater than 20 or less than 3.4. Figure 5.10 shows the top 10 diagnosis with abnormal total bilirubin.

Normal Alkaline Phosphatase in females is 50-135 U/L, and in males is 45-125 U/L. Pathologically elevated levels of the enzyme are seen in skeletal diseases such as rickets, bone malignancies, and malignant bone metastases. It can also be seen with hepatic disease such as

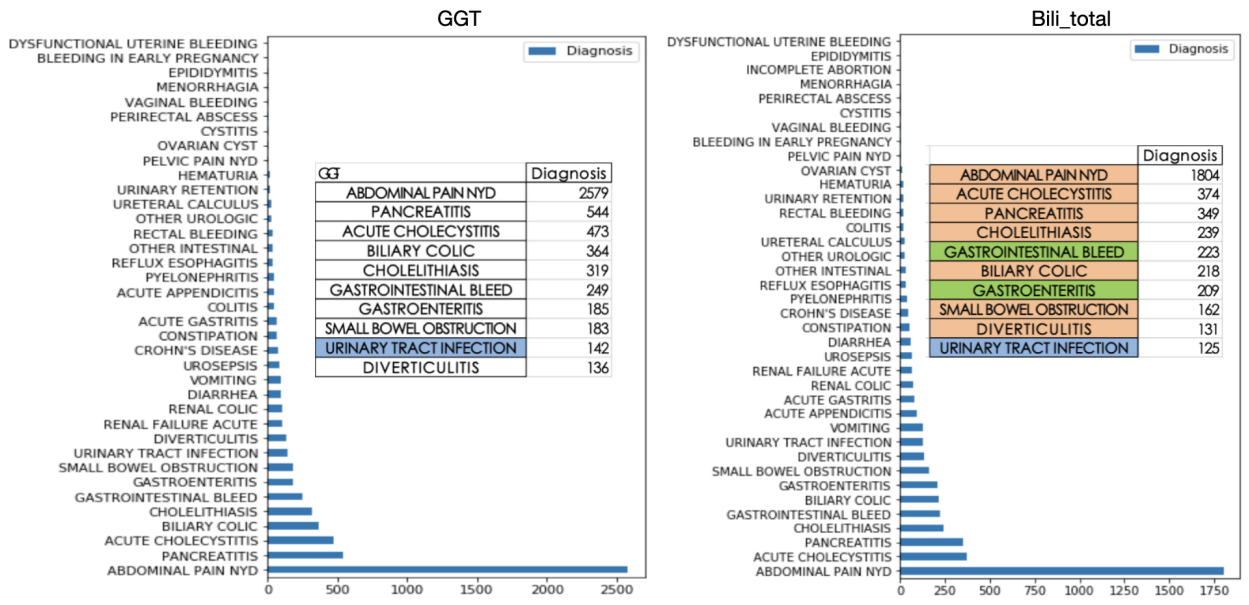

Figure 5.10: Left: Abnormal GGT diagnosis list (GGT>50 U/L). Right: Abnormal total bilirubin diagnosis list (Bili Total<3.4 or Bili Total>20).

extrahepatic bile duct obstruction, liver cancer, cirrhosis, and capillary cholangiohepatitis. Other causes include hyperparathyroidism or thyroid insufficiency in children. Pathology decreased levels of alkaline phosphatase are seen in severe chronic nephritis and anemia. In Figure 5.11, Except abdominal pain NYD, the top 10 diagnosis with abnormal Alk in females are gastrointestinal pathology and urinary tract infection and in males are gastrointestinal pathology and acute renal failure.

We also plot the correlation for each block variables in Figure 5.12. In triage variables, CTAS was highly negatively correlated with pain rating data and DBP was highly positively correlated with SBP. In CBC variables, each cell count indicator is highly positively correlated with its percentage indicator. Lymphocytes are highly negatively correlated with Neutro percent. In liver enzyme variables, ALT and AST have high positive correlation. GGT correlated strongly with alkaline phosphatase.

5.2 The Hierarchical Tree Structure

We deal with the multi-class classification problem by arranging the diagnoses into a hierarchical tree structure, so that the multi-class classification can be achieved by a sequence of conditional binary or ternary classification problems at each node of the tree. That is, at

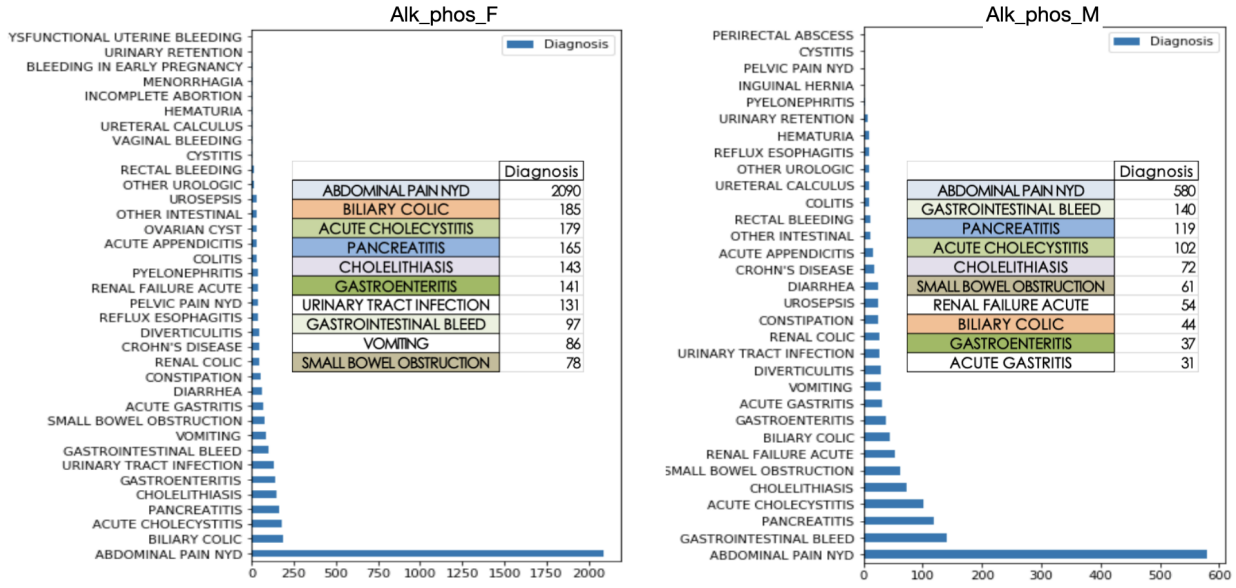

Figure 5.11: Left: Female abnormal diagnosis list (Alk phos<50 or > 135U/L). Right: Male abnormal diagnosis list (Alk phos<45 or >125U/L)

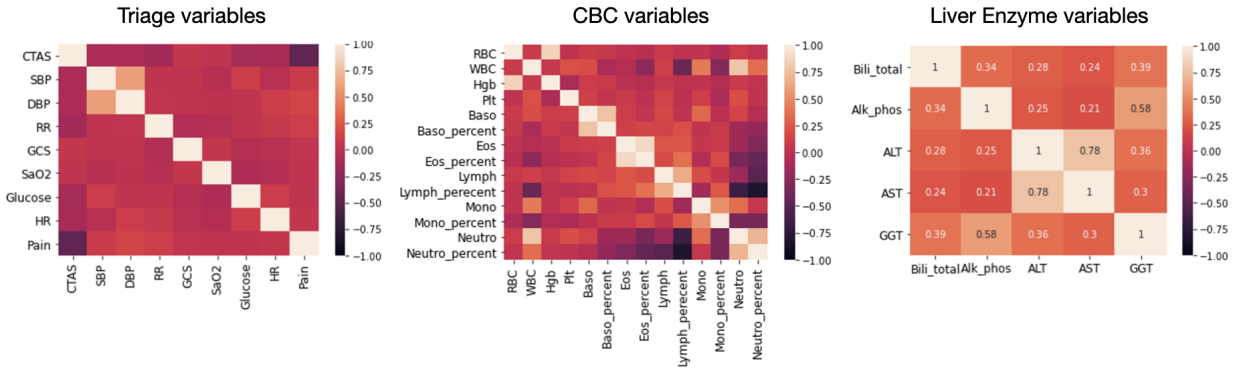

Figure 5.12: Correlation for triage variables, CBC variables and Liver enzyme variables

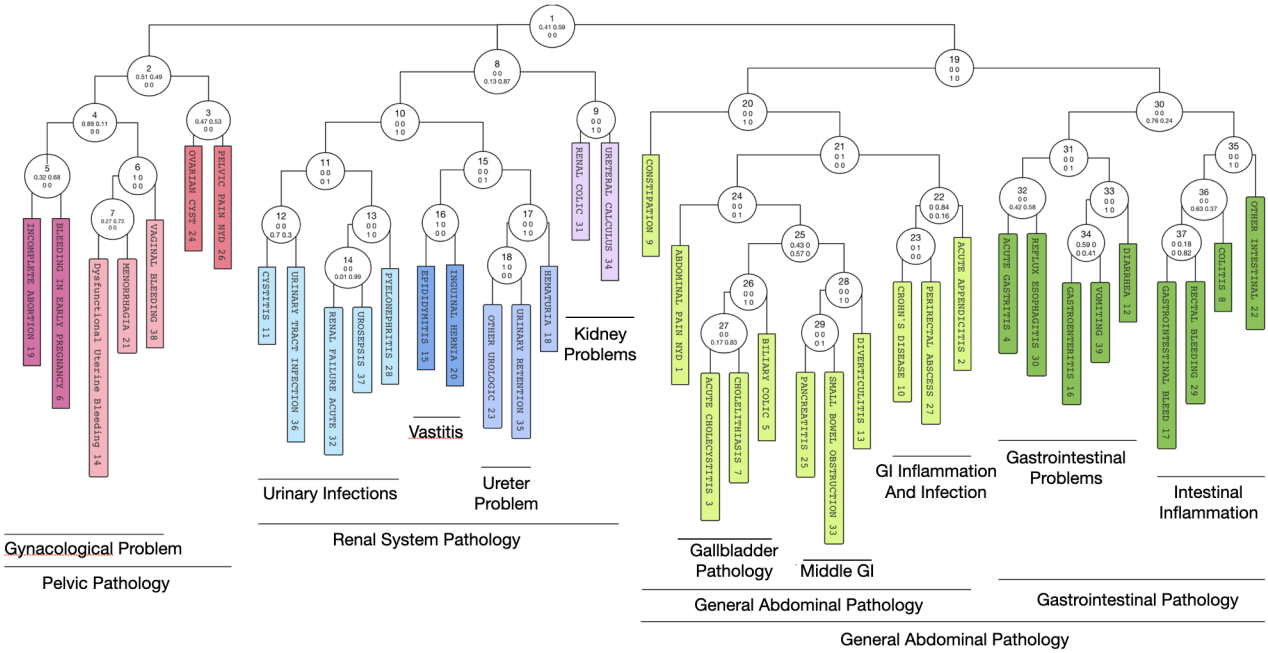

Figure 5.13: The tree-structured model framework

each node of the tree, we fit a classifier to predict which branch below that node the patient should follow, conditional on the patient being below that node.

5.2.1 Constructing Hierarchical Tree

In order for this tree structure to work well, it should be relatively easy to separate the higher nodes of the tree. Thus, we want to cluster the diagnoses that are most difficult to distinguish together at each level of the tree.

We construct the tree layer by layer. We start with each diagnosis in its own group. We then use the hierarchical clustering analysis (hclust function in R) to cluster the diagnosis groups based on a similarity measure described in Section 5.2.2. (We recalculate the similarity measure for each layer.) We consult with medical experts to decide where to cut the clustering in each layer. The diagnosis groups in each cluster are then merged into a single diagnosis group in the next layer and the process is repeated.

5.2.2 Similarity Measure based on Posterior Predictive Probability

The hierarchical clustering method is based on a similarity measure between the diagnoses. Our aim is to build a tree in which the diagnosis groups at higher nodes are easy to distinguish. To do this, we cluster diagnoses that are most difficult to distinguish. For each layer, we start by fitting a gradient boosting algorithm (GBM) classifier on the triage variables in the training data with a multi-class response having one class for each group of diagnoses in the current layer. We use this method to predict the posterior probability of each diagnosis for the validation data set. For the ith patient in the validation set, and the jth diagnosis, let

$$Y_{ij} = \begin{cases} 1 & \text{if the ith patient has diagnosis j} \\ 0 & \text{otherwise} \end{cases}$$

and let Z_{ij} be the posterior probability assigned to diagnosis j by the gbm predictor for this validation sample. We define our similarity matrix S by

$$S_{jk} = \frac{n_{val} \sum_i Z_{ij} Y_{ik}}{\sum_i Z_{ij} \sum_i Y_{ik}}$$

where n_{val} is the total number of observations in the validation data set. The resulting matrix is not symmetric, so we symmetrize it by averaging the matrix with its transpose. Note that this measure is different from the percentage of mis-assigned patients from the kth true diagnosis to the jth diagnosis.

The details of constructing the hierarchical tree are given in the following section.

5.2.3 Details of each Layer

Figures 5.14-5.22 show the heatmap of our similarity measure and the fitted dendrogram for the hierarchical clustering for layers 1-9 respectively. We describe the clustering performed at each layer.

1. The bottom layer shows all the 39 classes in the problem, in Figure 5.14.

2. Layer 2

   From Figure 5.14, there are three clear clusters that can be created:

   - Gynacological Problems(6+19)

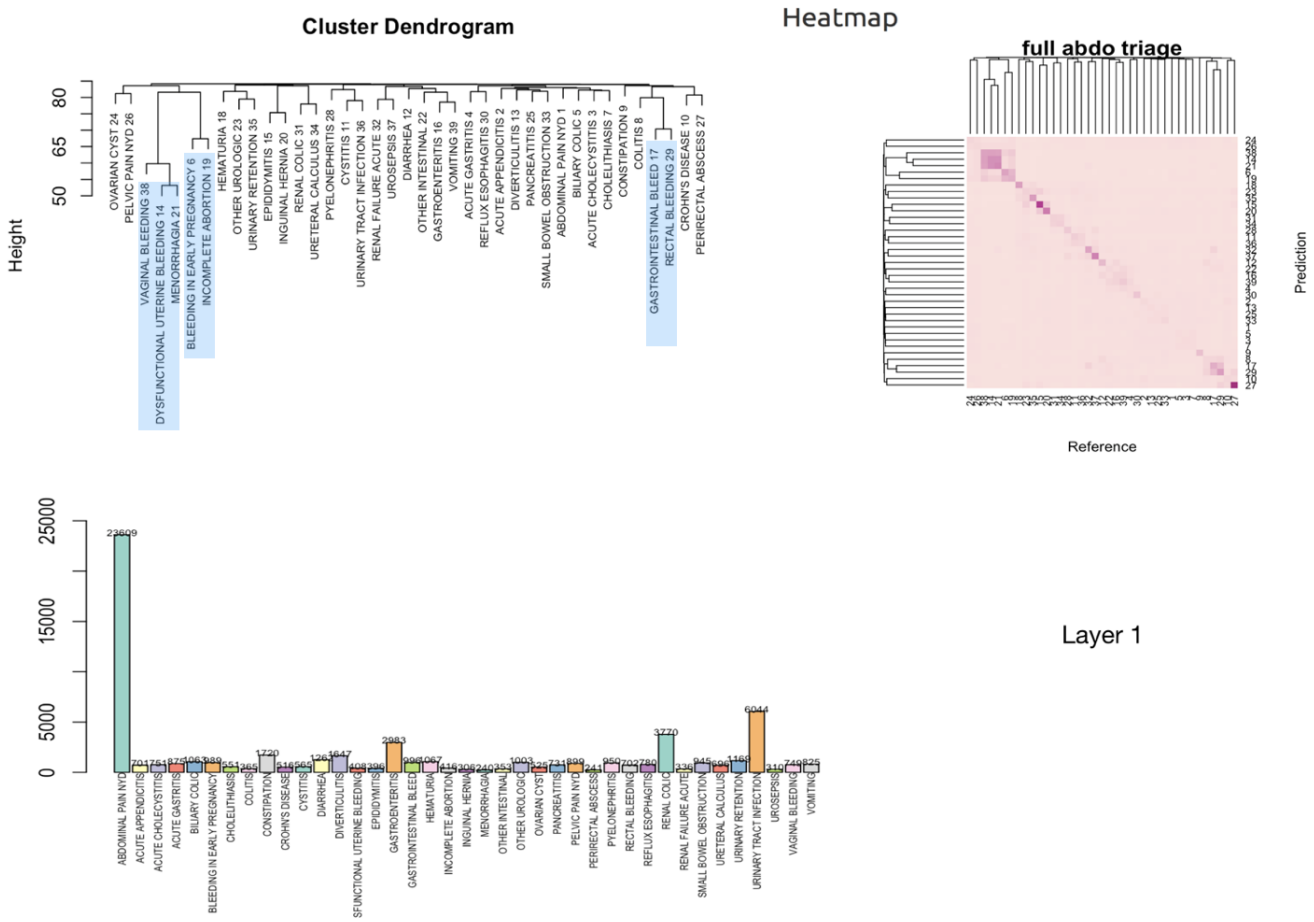

Figure 5.14: Dendrogram and Heatmap of the 1st layer

Table 5.1: Class Reference

Class	Diagnosis
1	ABDOMINAL PAIN NYD
2	ACUTE APPENDICITIS
3	ACUTE CHOLECYSTITIS
4	ACUTE GASTRITIS
5	BILIARY COLIC
6	BLEEDING IN EARLY PREGNANCY
7	CHOLELITHIASIS
8	COLITIS
9	CONSTIPATION
10	CROHN'S DISEASE
11	CYSTITIS
12	DIARRHEA
13	DIVERTICULITIS
14	DYSFUNCTIONAL UTERINE BLEEDING
15	EPIDIDYMITIS
16	GASTROENTERITIS
17	GASTROINTESTINAL BLEED
18	HEMATURIA
19	INCOMPLETE ABORTION
20	INGUINAL HERNIA
21	MENORRHAGIA
22	OTHER INTESTINAL
23	OTHER UROLOGIC
24	OVARIAN CYST
25	PANCREATITIS
26	PELVIC PAIN NYD
27	PERIRECTAL ABSCESS
28	PYELONEPHRITIS
29	RECTAL BLEEDING
30	REFLUX ESOPHAGITIS
31	RENAL COLIC
32	RENAL FAILURE ACUTE
33	SMALL BOWEL OBSTRUCTION
34	URETERAL CALCULUS
35	URINARY RETENTION
36	URINARY TRACT INFECTION
37	UROSEPSIS
38	VAGINAL BLEEDING
39	VOMITING

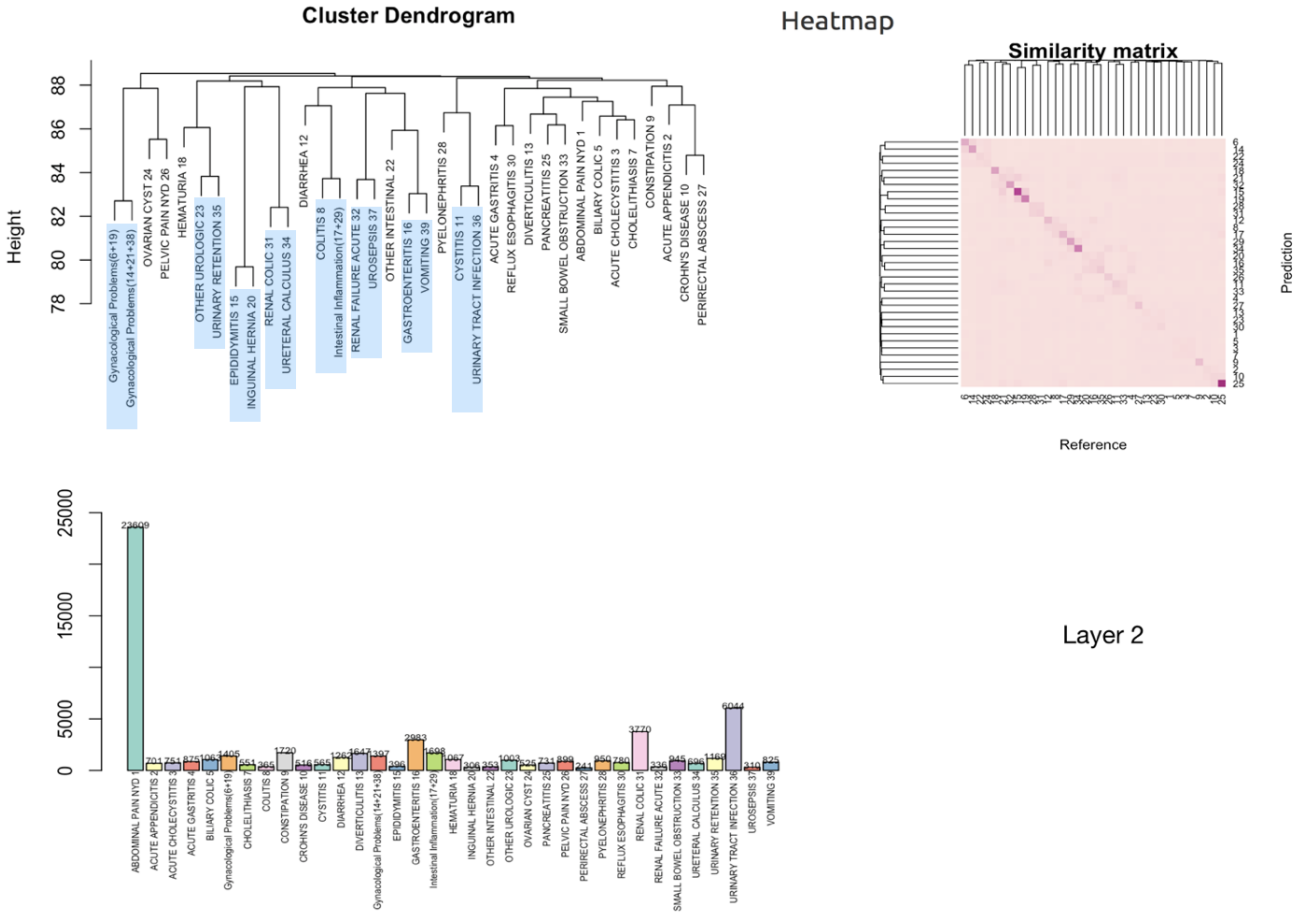

Figure 5.15: Dendrogram and Heatmap of the 2nd layer

- Gynacological Problems(14+21+38)
- Intestinal Inflammation(17+29)

These are very clear-cut, based on our similarity measure, and were confirmed to make sense medically. We therefore combine diagnoses within these groups to form the 2nd layer.

3. Layer 3

   In the dendrogram in Figure 5.15, the following small clusters are clear-cut in the second layer. These clusters were confirmed to make medical sense.

   - Gynacological Problems(6+19+14+21+38)
   - Intestinal Inflammation(8+17+29)

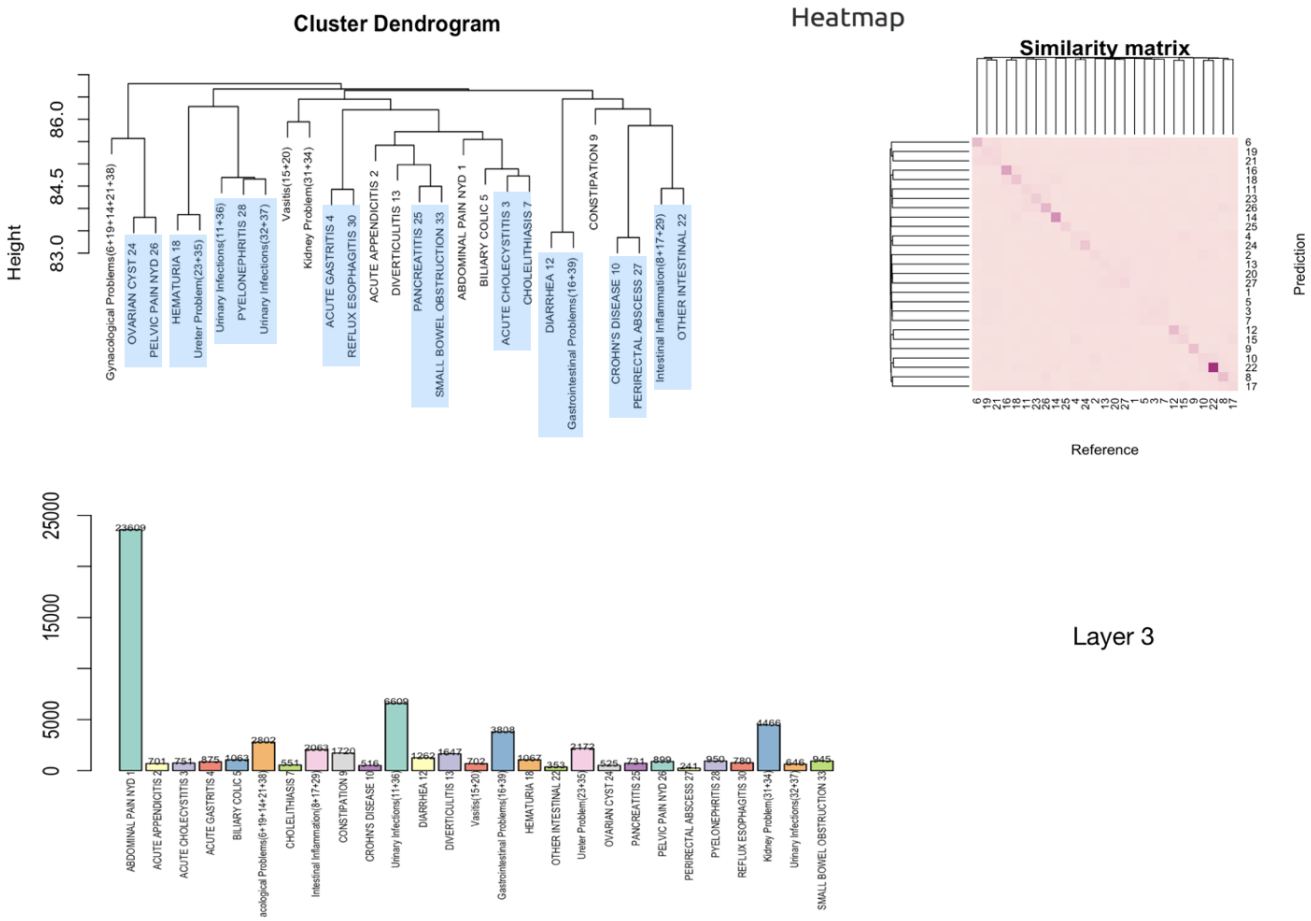

Figure 5.16: Dendrogram and Heatmap of the 3rd layer

- Urinary Infections(11+36)
- Vasitis(15+20)
- Gastrointestinal Problems(16+39)
- Ureter Problems(23+35)
- Kidney Problems(31+34)
- Urinary Infections(32+37)

4. Layer 4

   In the dendrogram in Figure 5.16, we see several more clear-cut clusters, which our medical collaborators confirm are medically reasonable.

   - Pelvic pathology(24+26)

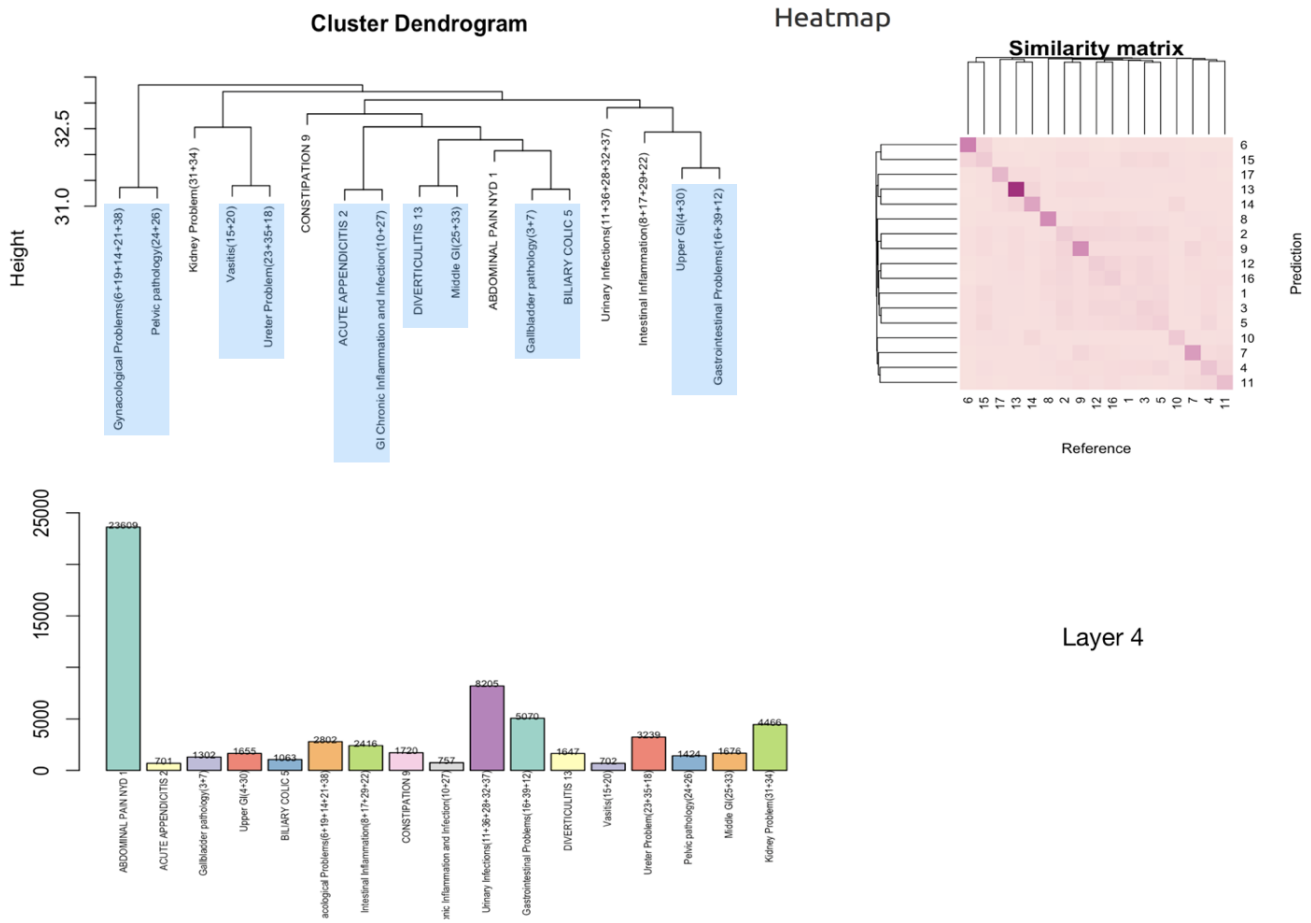

Figure 5.17: Dendrogram and Heatmap of the 4th layer

- Ureter Problems(23+35+18)
- Urinary Infections(11+36+28+32+37)
- Gastrointestinal Problems(16+39+12)
- GI Chronic Inflammation and Infection(10+27)
- Intestinal Inflammation(8+17+29+22)
- Upper GI(4+30)
- Middle GI(25+33)
- Gallbladder pathology(3+7)

5. Layer 5

In Figure 5.17 the following clusters are clear-cut and medically reasonable:

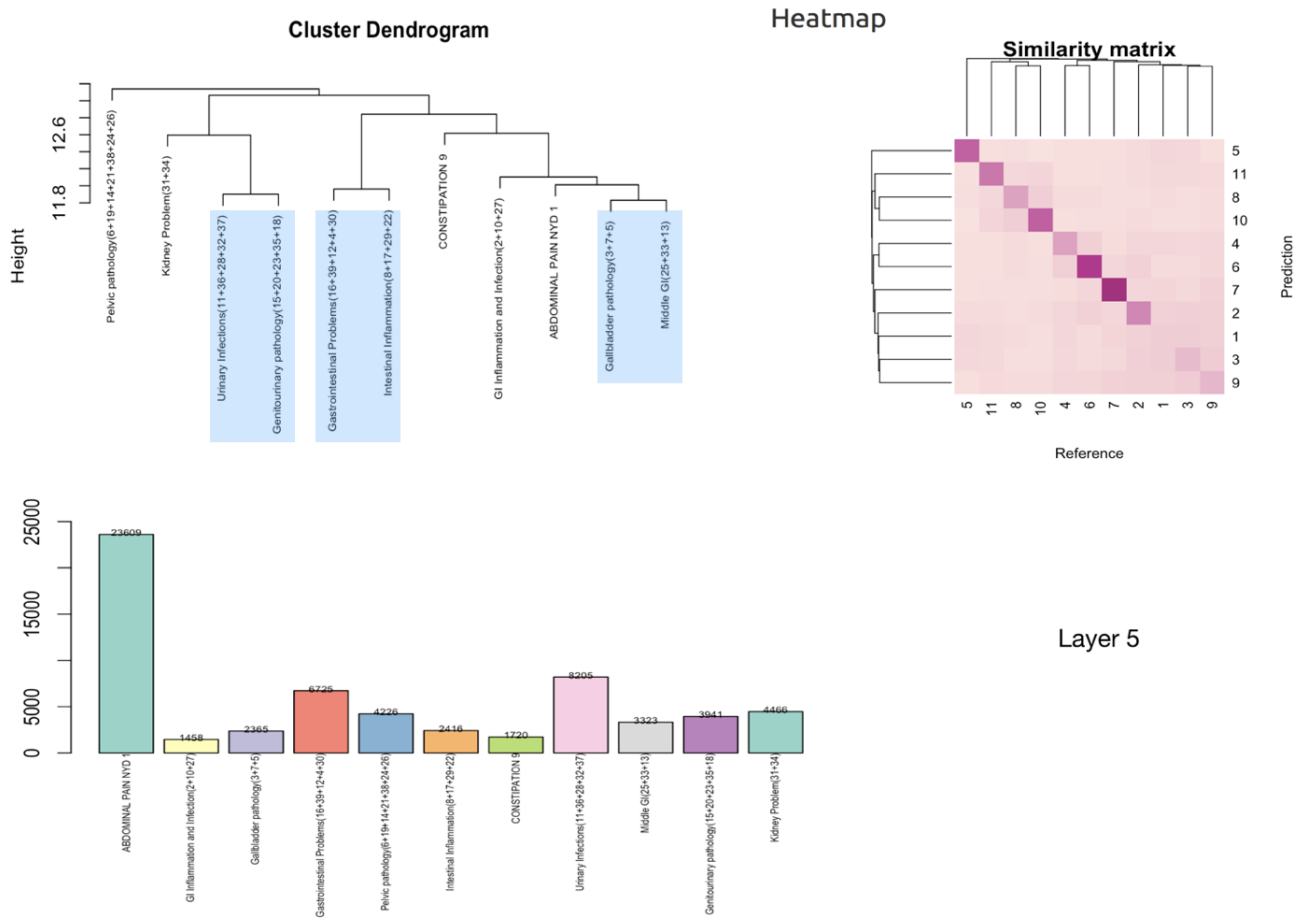

Figure 5.18: Dendrogram and Heatmap of the 5th layer with 11 classes

- Gynecological problem(6+19+14+21+38) & pelvic pathology(24+26) = pelvic pathology(6+19+14+21+38+24+26)
- Vasitis(15+20) & Ureter problems(23+35+18) = Genitourinary pathology(15+20+23+35+18)
- Acute appendicitis 2 & GI Chronic Inflammation and Infection(10+27) = GI Inflammation and Infection(10+27+2)
- Diverticulitis 13 & Middle GI(25+33) = Middle GI(13+25+33)
- Gallbladder pathology(3+7) & Biliary colic 5 = Gallbladder pathology(3+7+5)
- Gastrointestinal Problems(16+39+12) & Upper GI(4+30) = Gastrointestinal Problems(16+39+12+4+30)

6. Layer 6

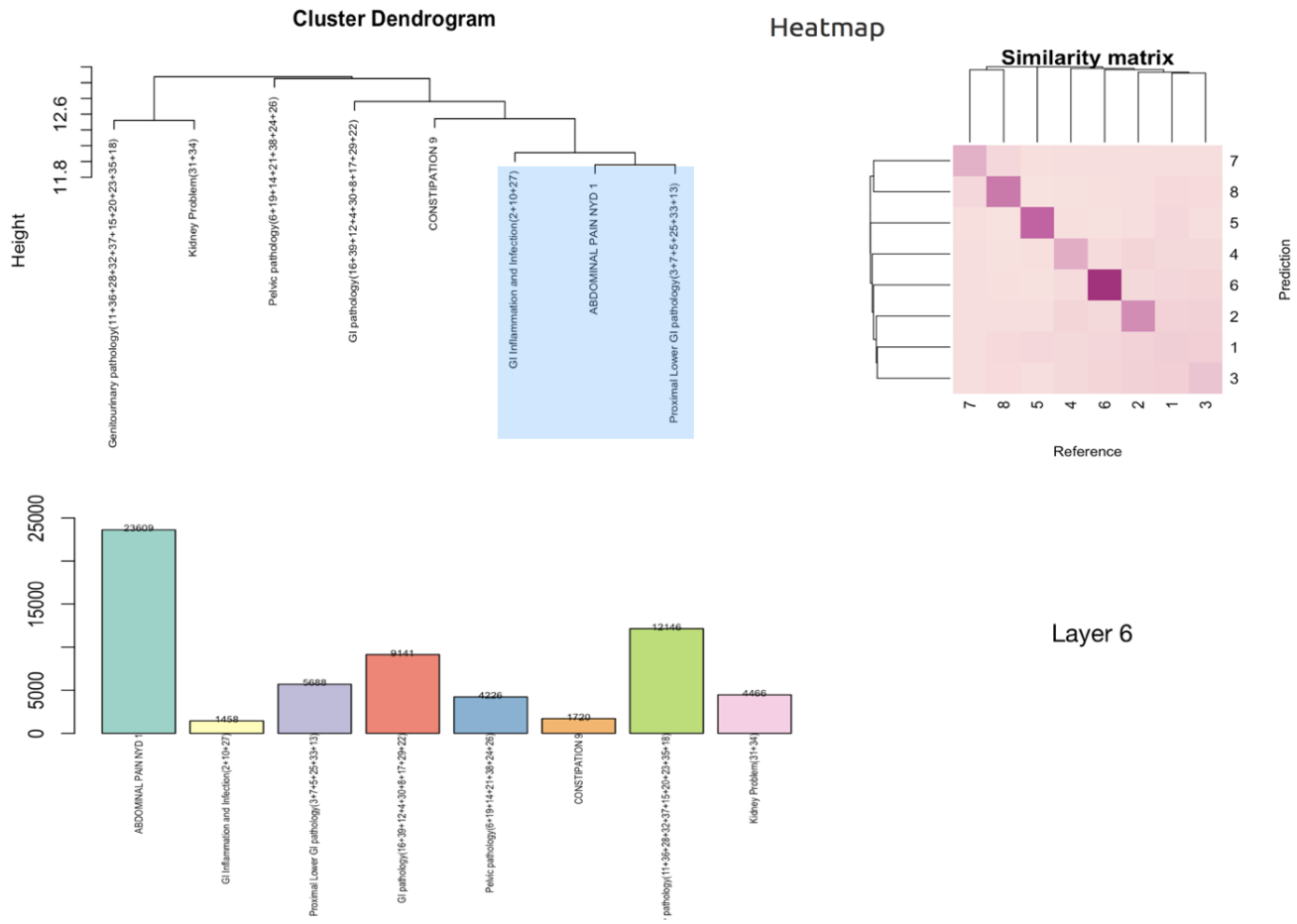

Figure 5.19: Dendrogram and Heatmap of the 6th layer

In Figure 5.18, we see the following clear-cut and medically reasonable clusters:

- Urinary infections(11+36+28+32+37) + genitourinary problems(15+20+23+35+18) = Genitourinary pathology(11+36+28+32+37+15+20+23+35+18)
- Gastrointestinal problems(16+39+12+4+30) + intestinal inflammation(8+17+29+22) = GI pathology(16+39+12+4+30+8+17+29+22)
- Gallbladder pathology(3+7+5) + middle GI(25+33+13) = Proximal Lower GI pathology(3+7+5+25+33+13)

7. Layer 7

In Figure 5.19, the clusters are slightly less clear-cut. However, the following cluster seems the most clear-cut and medically reasonable. Given that the clusters are not

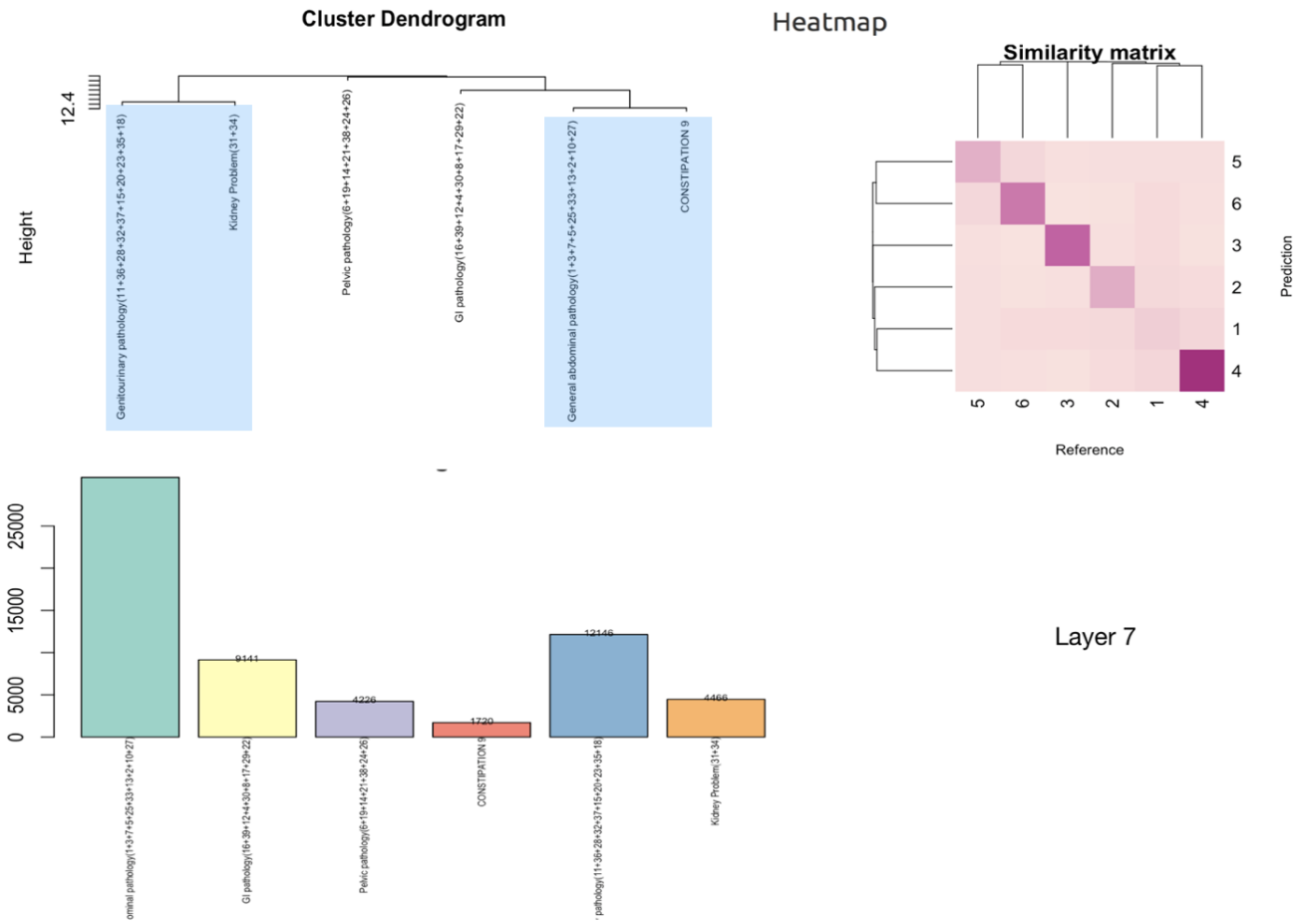

Figure 5.20: Dendrogram and Heatmap of the 7th layer with 6 classes

very clear-cut in this dendrogram, we only form a single cluster in this layer.

- Abdominal Pain NYD (1) + Proximal lower GI pathology (3+7+5+25+33+13) + GI inflammation and infection (2+10+27) = General abdominal pathology (1+3+7+5+25+33+13+2+10+27)

8. Layer 8

The dendrogram in Figure 5.20 shows two clear-cut clusters, which are not medically unreasonable:

- General abdominal pathology(1+3+7+5+25+33+13+2+10+27) + Constipation 9 = General abdominal pathology(1+3+7+5+25+33+13+2+10+27+9)

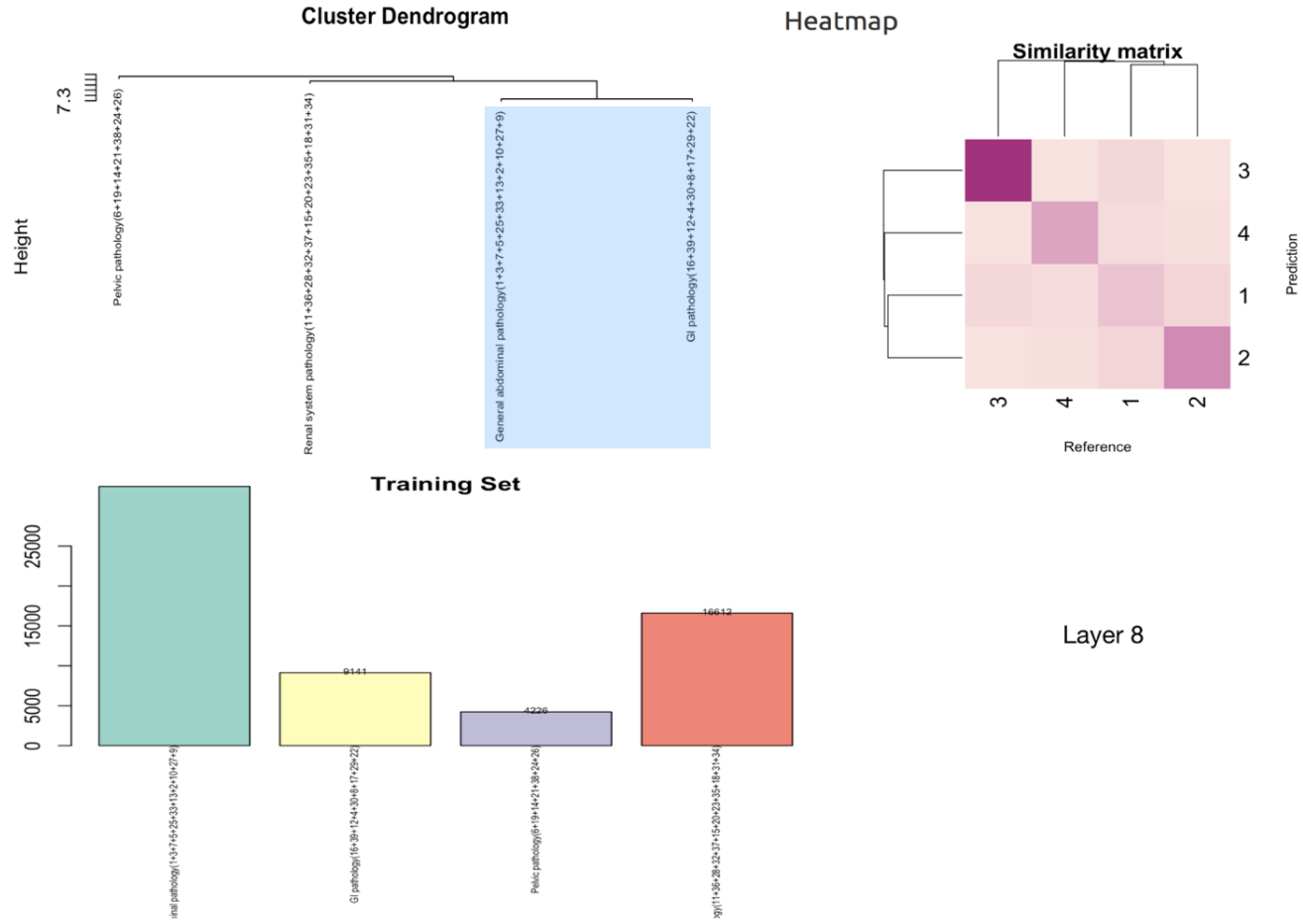

Figure 5.21: Dendrogram and Heatmap of the 8th layer

- Genitourinary pathology(11+36+28+32+37+15+20+23+35+18) + Kidney problems(31+34) = Renal system pathology(11+36+28+32+37+15+20+23+35+18+31+34)

9. Layer 9

The dendrogram in Figure 5.21 shows one clear-cut cluster, which is not medically unreasonable.

- General abdominal pathology(1+3+7+5+25+33+13+2+10+27+9) + GI pathology(16+39+12+4+30+8+17+29+22) = General abdominal pathology (1+3+7+5+25+33+13+2+10+27+9+16+39+12+4+30+8+17+29+22)

Finally, we have 3 classes at the very top level (Figure 5.22),

- General abdominal pathology
  (1+3+7+5+25+33+13+2+10+27+9+16+39+12+4+30+8+17+29+22)
- Pelvic pathology(6+19+14+21+38+24+26)
- Renal system pathology(11+36+28+32+37+15+20+23+35+18+31+34)

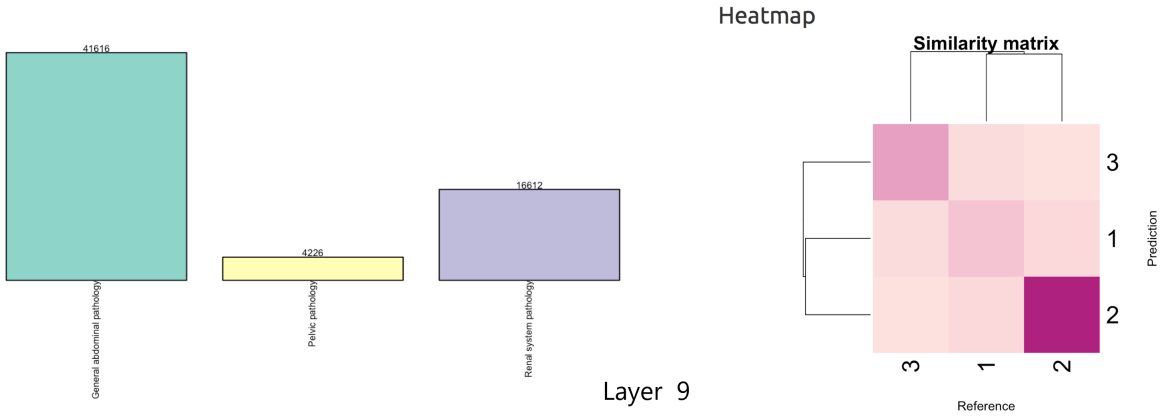

Figure 5.22: Dendrogram and Heatmap of the 9th layer

As shown in Figure 5.22, these are very easily distinguished clusters, and it does not make medical sense to perform any further clustering on them. Combining all these 9 layers of clusterings, we obtain the tree which provides our model framework in Figure 5.13.

5.3 Model Development

In this section, we describe four different methods that we will fit on the data. The first method is a competitive baseline off-the-shelf method based on GBM. The second method uses our hierarchical tree to split the 39-class classification problem into many binary or ternary classification problems, each fitted by GBM. The third method only uses the model combination idea, without using the hierarchical tree, to directly combine 39-class classifiers. The fourth method combines the hierarchical tree structure with the model combination

Table 5.2: LightGBM parameters

Parameter	Value
boosting type	gbdt
num boost round	200
num leaves	120
max depth	10
num iterations	200
early stopping rounds	20
min data in leaf	40
learning rate	0.01
feature fraction	0.9
bagging fraction	0.9
bagging freq	5
min gain to split	0.2

method so that model combination method is applied to combine binary or ternary classifiers on each node of the hierarchical tree.

5.3.1 Benchmark Model Prediction

we use LightGBM as our benchmark GBM off-the-shelf method. GBM is able to deal with multi-class classification and missing data (Ke et al., 2017). And a preliminary analysis of the triage variables showed that it is competitive with other off-the-shelf methods. We use the LightGBM package for fitting the GBM. This package has been shown to achieve high prediction accuracy and efficient computation, which is an important practical consideration for such a large data set. The optimized tuning parameters are in Table 5.2.

There are 4 boosting types in Light-GBM: 'gbdt', 'rf', 'dart' and 'doss'. Among them, 'gbdt' has more stable results, and is most widely used. *learning_rate* is the step length of the change of the gradient. The default value is 0.1, and it is usually set in the range 0.05-2. In the experiment, when using smaller learning rate 0.01, the accuracy improves to the best. *num_leaves* is one of the most important parameters to control the calculation complexity of the model. Higher *num_leaves* could lead to higher accuracy, but also may leads to overfitting. An appropriate value should be determined according to the size of the dataset. This parameter is usually set to be less than $2^{(max\_depth)}$, in which *max_depth* is another important parameter related to the complexity of the model. *max_depth* controls the

maximum depth of each tree. Higher values may lead to overfitting. *num_iterations* controls the number of iterations (the number of trees to be built). More trees may increase the preciseness and the training time, and may lead to overfitting. If validation metrics did not get improved after the last round, *early_stoppong_rounds* would stop training. According to experience, this parameter is usually set to be 10% of the parameter *num_iteration*. Default value of $min\_data\_in\_leaf$ is 20. This parameter can be used to deal with overfitting. $feature\_fraction$ means the feature score or sub-feature processing column sampling. Light-GBM randomly selects feature subsets on each iteration (tree). For example, if it is set to be 0.6, Light-GBM will select 60% of the features before training each tree. In our model, we select all features before training each tree. By $bagging\_fraction$, we could set the percentage of rows used in the iteration of each tree. This means that random rows will be selected to match each learner (tree). This not only improves the generalization ability, but also improves the training speed.

5.3.2 Model using Hierarchical Tree Structure combined with GBM classifier

Next, we use our hierarchical tree structure to train the classifier. For each node in our tree, we fit a GBM classifier to predict the conditional probability of an observation lying on each branch, given that it is below this node. For a given diagnosis, the probability assigned to that diagnosis is the product of the probabilities of the branches leading to that diagnosis.

5.3.3 Model Combination Method applied on 39-class classifiers

Unlike the model combination for regression models, where the linear combination operation most naturally works on the linear predictors from different models, the model combination for this classification problem can work in different ways. The most direct way is to use a weighted mean of the predicted probabilities similar to that in Random forest. Alternatively we can combine the models using the logistic transformed probabilities, as detailed in Chapter 3. We have tested both methods in the application and found generally the latter idea works better. In addition, the optimization procedure to find the model combination weights also works easily because the problem is naturally transformed into a logistic regression problem.

In this dataset, there are four blocks of variables: Triage, CBC, Liver enzymes, and Radiology. Our model combination method was designed to handle only a single block of missing variables. We need to extend our method to handle three blocks of missing variables.

(Triage variables are never missing) With three missing blocks, there are eight models that need to be fitted. In this dataset, we note that there are very few patients who have radiology data, but who do not have both CBC and liver enzyme data. We therefore do not fit models for these patients. This leaves five models, as shown in Figure 5.23. We label these models M_0, M_1, M_2, M_3 and M_4.

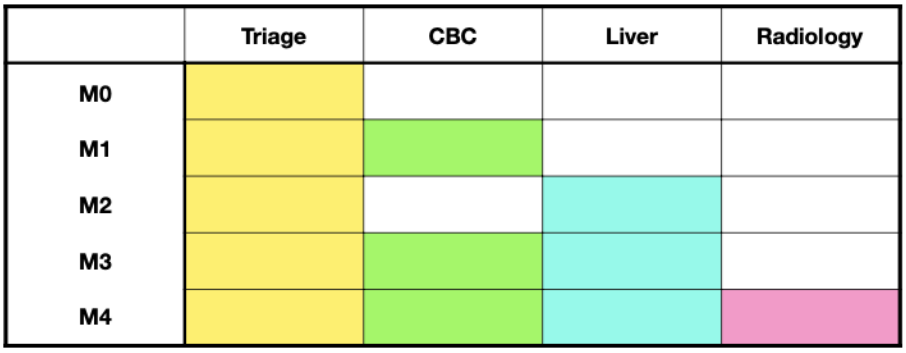

Figure 5.23: The data is partitioned into different models based on four groups of variables, triage, CBC, Liver enzyme and radiology. M_0 model fit on triage variables; M_1 model fit on triage and CBC variables; M_2 model fit on triage and liver enzyme variables; M_3 model fit on triage, CBC and liver enzyme variables; M_4 model fit on triage, CBC, liver enzyme and radiology variables;

For each of the five models, M_0, M_1, M_2, M_3 and M_4, we fit a Random Forest (RF) model on the corresponding data to predict the probability of each of the 39 classes. We select the tuning parameters of the RF classifier using 5-fold cross-validation on the training data. We select best fit tuning parameters from the following candidates: the number of trees = 200, the number of features to consider when looking for the best split is $\sqrt{n\_features}$, the minimum leaf size = (1, 5, 10), and the maximum number of levels in each decision tree = (10, 20, 30, 40).

In addition to the need to estimate coefficient α_i for the models M_i to obtain a combined classifier for patients with complete data, it is also necessary to obtain coefficients for classifiers from various combinations of partial data. For example, to predict the diagnosis for a patient with triage and CBC, we take a linear combination of the predictions from M_0 and M_1. We will let α_{ij} denote the coefficient of model M_j in the classifier for patients with the blocks of variables for model M_i. For example, for a patient

with CBC and liver enzyme variables, the logistic transformed probability of diagnosis k is $\alpha_{30}\hat{f}_{0,k}(x)+\alpha_{31}\hat{f}_{1,k}(x)+\alpha_{32}\hat{f}_{2,k}(x)+\alpha_{33}\hat{f}_{3,k}(x)$, where $\hat{f}_{j,k}$ is the predicted logistic-transformed probability of diagnosis k using the model M_j. There is no α_{34} coefficient, because the radiology data are not available for this patient, so M_4 cannot be used to make prediction for this patient. We have a total of 13 coefficients to choose, but with the constraints $\sum_j \alpha_{ij} = 1$ for $i = 1, 2, 3$ and 4, there are a total of 9 coefficients to be estimated.

For each i we estimate the α_{ij} to minimise the cross-entropy loss on the validation data for model M_i,

$$\text{argmin}_\alpha \sum_{i=1} -\log \mathbb{P}\left(Y = Y_i \mid \alpha\right)$$

subject to the constraints that $\alpha_{ij} \geq 0$ and $\sum_j \alpha_{ij} = 1$.

5.3.4 Model using both Tree Structure and Model Combination method

Finally, we apply both the hierarchical tree structure and the model combination method to predict the probability of each diagnosis.

The difficulty with freely estimating all the parameters α_{ij} is that for a particular i, the validation set might be relatively small, and we may need to fit as many as four free model parameters. This can lead to unstable estimates, and worse prediction accuracy. To avoid this, we impose some consistency between the estimated α_{ij}.

We define all 13 α_{ij} as functions of 3 coefficients, α, β and γ. $1-\alpha$ represents the relative contribution of the model including CBC variables. That is, the contribution of M_1 versus M_0 and of M_3 versus M_2. $1-\beta$ is the relative contribution of models including the liver enzyme variables. $1-\gamma$ is the contribution of models including the radiology variables. So the combined model for a full set of predictors is

$$\begin{aligned} Q = &\gamma\beta\alpha M_0 \\ &+ \gamma\beta(1-\alpha)M_1 \\ &+ \gamma(1-\beta)\alpha M_2 \\ &+ \gamma(1-\beta)(1-\alpha)M_3 \\ &+ (1-\gamma)M_4 \end{aligned}$$

The idea is that we first apply our model combination to combine the models M_0 and M_1 and to combine the models M_2 and M_3, using the same combination coefficient α. We let

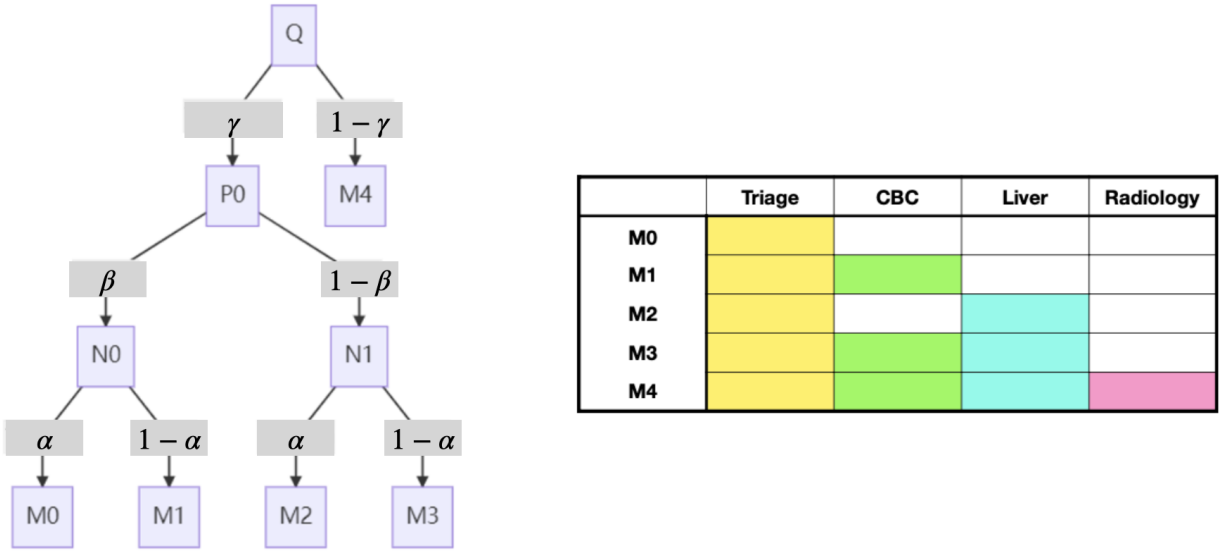

Figure 5.24: Stagewise linear combination weighting

N_0 and N_1 denote the resulting combined models. Next, we find the coefficient β to combine the models N_0 and N_1. Finally, we find the coefficient γ to combine the combination of N_0 and N_1 with the model M_4. We demonstrate this process in Figure 5.24.

To determine the stagewise linear combination weight we first apply the previously presented method to calculate α for combining M_0 and M_1. We call this combined model N_0.

$$N_0 = \alpha M_0 + (1 - \alpha) M_1 \tag{5.1}$$

For fitting this α, we fit M_0 and M_1 on all observations that have CBC variables, even if they also have liver enzyme variables or radiology variables. Because these observations have been used to fit α, it is reasonable to use the same α for combining the models M_2 and M_3. This gives a new model N_1 for all observations with Liver enzymes.

$$N_1 = \alpha M_2 + (1 - \alpha) M_3 \tag{5.2}$$

We can now use our method to choose a coefficient β to combine N_0 and N_1.

$$P_0 = \beta N_0 + (1 - \beta) N_1 \tag{5.3}$$

After combining N_0 and N_1, let the result be P_0. Finally, we combine P_0 with M_4 in the usual way:

$$Q = \gamma P_0 + (1 - \gamma) M_4 \tag{5.4}$$

The training steps for each node are listed below (Figure 5.25). The combination parameter α can be estimated as follows:

- Step one: use 5-fold CV to choose the best random forest (RF) parameters on the training data

- Step two: Train models M_0, M_1, M_2, M_3 and M_4. We get the corresponding logit of predicted probability as $\hat{f}_0$, $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ and $\hat{f}_4$

- Step three: Solve α on the validation set of size n_1 by minimizing the negative log likelihood loss:
$$\text{argmin}_\alpha \sum_{i=1}^{n_1} L(y, \alpha\hat{f}_0 + (1-\alpha)\hat{f}_1)$$
which is a logistic regression with off-set without intercept.

Let N_0 denote the combination $\alpha M_0 + (1-\alpha)M_1$. This can be solved by performing logistic regression with one predictor $M_1 - M_0$ and an offset M_0 on the validation set which consists of all patients with CBC variables. We use the same value of α to combine M_2 and M_3. Let N_1 denote the combination $\alpha M_2 + (1-\alpha)M_3$.

- Step four: solve β on the validation set of size n_2 by minimizing the negative log likelihood loss:
$$\text{argmin}_\beta \sum_{i=1}^{n_2} L(y, \beta\hat{f}_{N_0} + (1-\beta)\hat{f}_{N_1})$$
.

- Step five: solve γ on the validation set n_3 by minimizing the negative log likelihood loss:
$$\text{argmin}_\gamma \sum_{i=1}^{n_3} L(y, \gamma\hat{f}_{P_0} + (1-\gamma)\hat{f}_4)$$
This procedure is illustrated in Figure 5.25.

The combination model gives a posterior probability for each branch of the tree. In theory, the probability of each diagnosis should be the product of all conditional probabilities of branches above it. However, for branches not above the true diagnosis, these predicted conditional probabilities are not meaningful, so the predictions can be strange. This can lead to some strange results, where a particular diagnosis is given a high probability because the predicted conditional probability conditioning on a false group of diagnoses is very

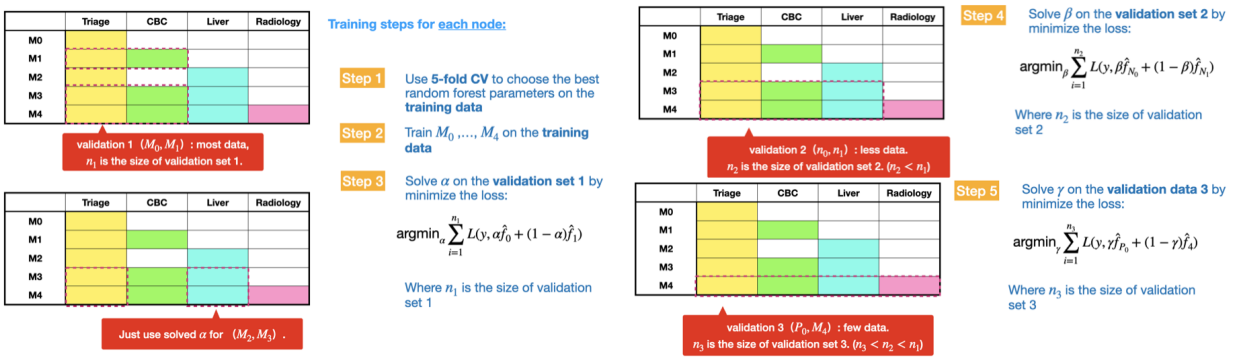

Figure 5.25: Model combination procedure

high. To stabilise our estimates, instead of multiplying the conditional probabilities, we use a Multilayer Perceptron to learn an estimated final probability vector from the vector of conditional probabilities at each node. A Multilayer Perceptron is a neural network with non-linear input-to-output mapping. It has input and output layers, and multiple hidden layers with many neurons stacked together (Bishop and Nasrabadi, 2006). Our purpose is to fit a flexible model to better estimate the final posterior probabilities of 39 classes. Neural network models (here Multilayer Perceptron model) with 39 output nodes is a convenient method to achieve this purpose. More specifically, after we get all combined posterior probabilities on 37 nodes, we use the output of these 37 combined posterior probability as input and use Multilayer Perceptron to learn how to best combine the input predictions to make a better output prediction. Practically we find it greatly improves the final prediction. In Multilayer Perceptron we set three hidden layers with 100, 50 and 100 neurons separately. The activation function is sigmoid. Learning rate is adaptive which is constant as long as training loss keeps decreasing. Maximum number of iterations is 1000.

5.4 Prediction Results

We are comparing methods that estimate the probability of each diagnosis. The aim of each method is to assign as much probability as possible to the true diagnosis. The natural way to assess the performance of methods is with the average probability assigned to the true diagnosis on test data. It is also possible to compare log probability assigned to the true diagnosis, which is the log-likelihood. However, the log-likelihood can be influenced by outliers where the method assigns a very low probability to the true diagnosis.

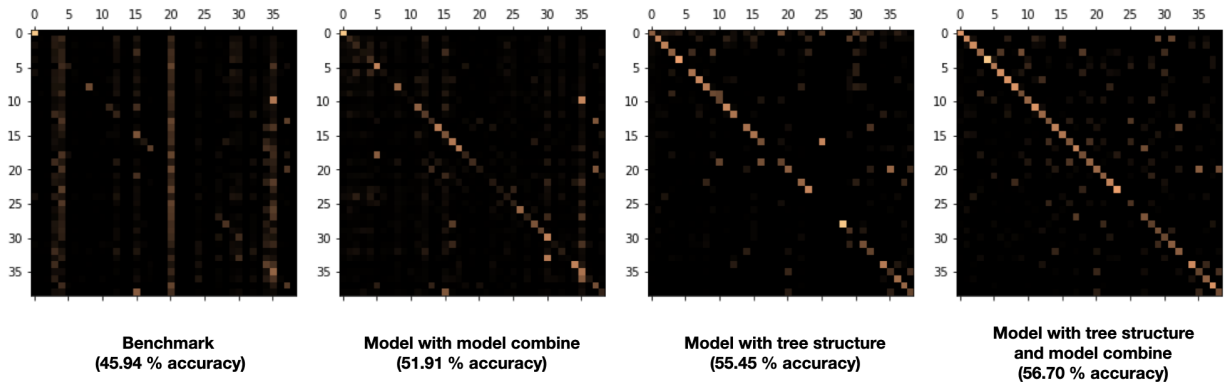

Figure 5.26: Accuracy comparison

The comparison results for all the methods described in Section 5.3 are shown in Figure 5.26. The average predicted probability of the true diagnosis is 45.94% under benchmark. It is 51.91% under the model combination method applied directly to the 39-class classification problem. It is 55.45% using the hierarchical tree structure without model combination. It is 56.7% using the hierarchical tree structure with model combination.

Based on these results, we see that the model combination and hierarchical tree methods have greatly improved the accuracy of our classification.

The objective of this research is to provide a list of the most plausible diagnoses for each patient, to allow the physician to check that they have not overlooked any plausible diagnoses. In practice, the busy physician will only be able to examine several most likely diagnoses, so it is important that the true diagnosis is fairly near the top of the list. We therefore examine how often the true diagnosis appears in each position of the list. Table 5.3 and Figure 5.27 show the frequency with which the true diagnosis is among the n diagnoses with highest predicted probability for $n = 1, ..., 8$. We see that by this measure, the model combination and hierarchical tree methods have improved our prediction substantially. For our method, the true diagnosis is within the 3 most probable diagnoses over 80% of the time, and within the 5 most probable diagnoses over 90% of the time. Given that the literature estimates a misdiagnosis rate of around 15% or higher, this level of accuracy means that our method is likely to be of practical value in helping physicians to diagnose patients.

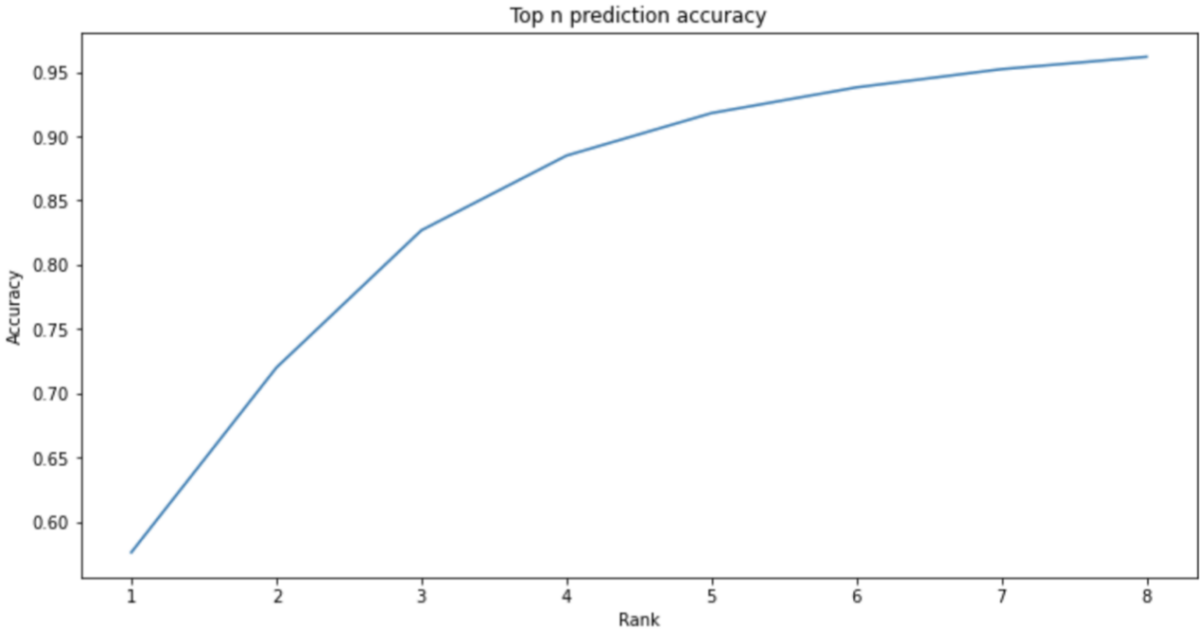

Figure 5.27: Cumulative prediction accuracy

5.5 Performance Evaluation

Our method estimates a probability for each diagnosis. This is potentially very helpful for the physician, indicating how plausible each diagnosis is, so the physician is not distracted by implausible diagnoses. However, the predicted probability is only valuable if it is a reliable estimate of the true probability. In this section, we therefore aim to assess the reliability of these predicted probabilities. We do this via the following visualisations. For each true diagnosis, we order all patients in the test data in decreasing order of the estimated probability of that diagnosis. We then plot both the cumulative average estimated probability

Table 5.3: Cumulative prediction accuracy table

Rank	Accuracy
1	0.567
2	0.720
3	0.817
4	0.876
5	0.912
6	0.923
7	0.951
8	0.960

of that diagnosis, and the cumulative proportion of patients who have that diagnosis. If the two curves are close together, it indicates that the estimated probabilities are a good assessment of the true probabilities. That is, for those patients for whom the estimated probability of diagnosis i is close to p, the proportion of those patients who actually have diagnosis i is also close to p. On the other hand, if the curves are far apart, it indicates that the estimated probabilities are unreliable, and may mislead the physician. Note that because these are cumulative averages, there is much more variance on the left side of the curve, so the curves usually do not match so well on that side. The shape of the curves indicates the power of the method to distinguish between the diagnoses. Ideally, the left-hand part of the curves would consist of a horizontal line of height 1, and the right hand part would be a decreasing straight line. This indicates a method that assigns probability 1 to all cases where diagnosis i is the true diagnosis, and 0 to all cases where it is not.

These plots are shown for all diagnoses in Figures 5.28. We see that in general, the probabilities estimated by our method are fairly reliable. For most diagnoses the curves start high, indicating a small number of patients confidently predicted, but then fall quickly, indicating much more uncertainty about the diagnosis.

Figure 5.29 shows histograms of the predicted probability of each diagnosis for patients for whom it is the true diagnosis, and patients for whom it is the false diagnosis. If our method is good at distinguishing between diagnoses, we should expect to see a good separation between the two histograms. If the histograms overlap, it indicates that this diagnosis is difficult to predict. We see that there is some separation. In many cases, most predicted probabilities are close to zero, indicating a rare diagnosis. In some cases, the data are sufficient to increase the probability of the diagnosis, but in many cases, we are unable to overcome the low prior probability of the diagnosis.

5.6 Producing a shortlist of diagnoses

In order to more efficiently present our predictions to the physician, instead of providing the posterior probability for each diagnosis, it is more convenient to only provide the physician

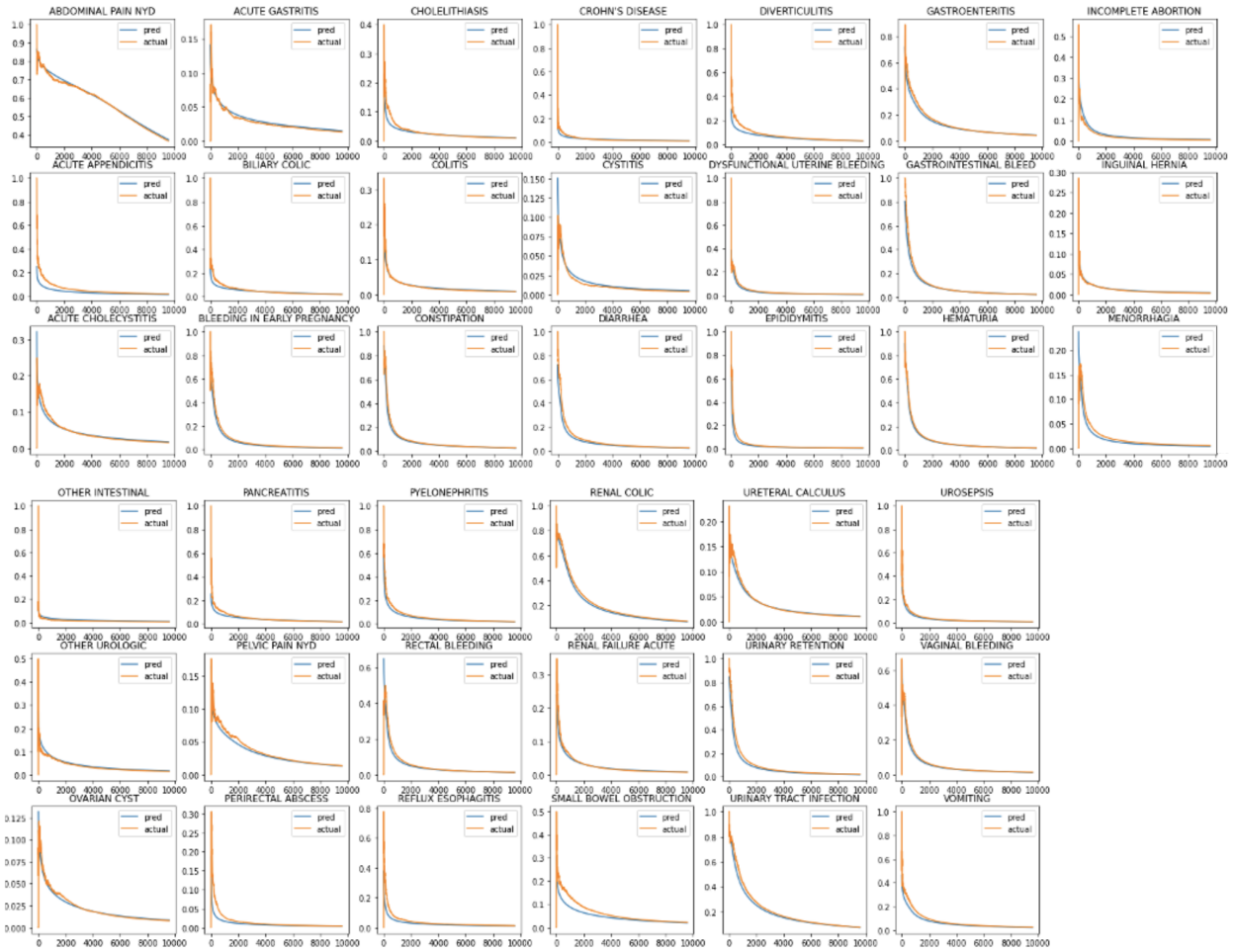

Figure 5.28: The running cumulative average prediction probability for each disease, versus the running cumulative average of actual probability.

Figure 5.29: The histogram of prediction probability between true cases and false cases for each disease.

with the most plausible diagnoses. To do this, we rank the diagnoses from highest to lowest posterior probability, then present the physician with the top-ranked diagnoses. A natural choice is to present enough diagnoses to ensure that the total predicted probability for the listed diagnoses exceeds a certain threshold (e.g. 95%). We can think of this as being like a 95% confidence interval. The length of this shortlist will vary between patients: in some cases only a single diagnosis will be given, for very clear-cut cases; in other cases where the diagnosis is very unclear, the list may contain a large number of diagnoses.

Table 5.4 shows the number of diagnoses in this list for each patient from the test data, and the corresponding coverage. The average number of diagnoses in the list is 7, but the length varies, with quite a few easy cases providing a short list, and some difficult cases producing a much longer list. Because the length of the list must be an integer, we cannot perfectly achieve 95% predicted coverage, but we see that the actual predicted coverage is close to the target percentage. We also see that the true coverage of the list is slightly higher than the predicted coverage, indicating that our method slightly overestimates the probability of unlikely diagnoses. This means that our list is slightly conservative, including more diagnoses than might be necessary.

This approach is based entirely on the predicted probability, with no consideration of the cost of different misdiagnoses. For future work, we should consider modifying the ranking to incorporate the clinical importance of different diagnoses. For example, it may be better to rank a less likely but more severe diagnosis above a more likely but less dangerous diagnosis, because of the greater cost of overlooking the severe diagnosis.

5.7 Conclusion

From the three evaluations of our model, we are confident in its ability to solve the huge difficulties of block missing and unbalanced multi-class classification. With the use of hierarchical tree structure, stagewise model combination and MLP ensemble techniques, it provides high prediction accuracy for possible diagnosis.

We have tried many different machine learning models to explore the optimal potential using hierarchical tree structure and model combination technique. Table 5.5 shows the prediction accuracy from different kinds of models and what techniques they employed. The results indicate that using hierarchical tree structure, stagewise model combination and multilayer perceptron (MLP) ensemble leads to the highest accuracy. The top 1 accuracy is

Table 5.4: The number of most likely diagnoses for each patient for their true diagnosis to be included with an estimated probability of 0.95

Estimated No. to cover 95%	Frequency	Average Posterior coverage	Actual coverage probability
1	552	0.955	0.960
2	1426	0.954	0.971
3	2114	0.959	0.964
4	2730	0.959	0.966
5	2966	0.957	0.960
6	3199	0.956	0.962
7	3222	0.956	0.962
8	2224	0.954	0.963
9	2886	0.955	0.962
10	1609	0.953	0.966
11	1083	0.954	0.972
12	680	0.953	0.982
13	427	0.953	0.977
14	195	0.953	0.969
15	89	0.953	0.978
16	31	0.953	0.968
17	5	0.953	1.000
18	2	0.954	1.000
19	1	0.954	1.000

56.7%, and top 5 accuracy is up to 91.2%, which illustrates a significant improvement of the prediction accuracy.

Table 5.5: A summary of the different approaches attempted

Approaches	Top 1	Top 3	Top 5
Tree + Extratree ensemble, without model combination	51.0 %	80.5 %	90.8 %
Tree + original model combination	52.4 %	80.5 %	90.8 %
Tree + original model combination + MLP ensemble	47.6 %	77.3 %	83.6 %
Tree + stagewise model combination + MLP ensemble	56.7%	81.7%	91.2%

To evaluate the performance of our classification model, in addition to cumulative prediction accuracy, we apply three different measurements. First, we compare the running cumulative average prediction probability for each disease, versus the running cumulative average of actual probability. This indicates how well the model is performing on that particular diagnosis. If the classification model is accurate, the curve will drop slowly and exhibit no systematic difference. This provides a means of indicating the diagnoses for which the model prediction is inaccurate. The second way to evaluate whether the estimated posterior

probability is close to the truth, is by comparing the histogram of prediction probability between true cases and false cases for each disease. The bigger deviations or less overlaps between two histograms indicate better classification performance. These two measurements indicate that our model exhibits stable performance across all diseases without significant deviations. Third, instead of using the first k assignments to assess the accuracy, we report how many diagnosis are needed to cover 95% posterior probability. Having a flexible length for the list of plausible diagnoses ensures that we only present plausible diagnoses to the physician. The length of this list is an indication of how confident we are about the predicted diagnoses. A long list indicates uncertainty over the true diagnosis.

For future work we will incorporate the clinical importance of different diagnoses into the ranking. It may be better to rank a less likely but more severe diagnosis above a more likely but less dangerous diagnosis because of the greater cost of overlooking the severe diagnosis.

Chapter 6

Discussion

This thesis has two major contributions. First, we provide a model combination method to deal with cases where large blocks of the data are missing. We study the asymptotic behaviour of the method for linear regression, and show that using the incomplete data improves prediction accuracy over complete case analysis. In our experiments, the model combination method improved the prediction accuracy on both simulations and real data examples. Some advantages of our method: (1) All subjects, so long as at least one of the block of predictors is available, can be used for the model combination method, and all predictors can contribute to the model; (2) the difficulty of guessing unknowns is bypassed, as the model combination method is only based on the data available; (3) computation time is shorter for Plug-in estimation. In future work, we will assess more thoroughly how well the model combination method works on problems with multiple missing blocks.

The second contribution is applying this model combination method and a hierarchical tree structure on diagnoses to develop an automated diagnosis assistance tool for emergency department abdominal pathology data. Our automated diagnosis tool will provide clinicians with the most likely diagnoses with high accuracy. In future work, we will not only output a posterior probability vector for the diagnoses based on the currently available variables, but plan on using our model to predict how much a particular additional test results would improve our ability to diagnose a particular patient. This can be used to help the physician decide which additional tests (if any) to order. We also plan to extend our method to all diagnoses in the emergency department, not just ones associated with abdominal pathology.

Bibliography

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Burroughs, T. E., Waterman, A. D., Gallagher, T. H., Waterman, B., Adams, D., Jeffe, D. B., Dunagan, W. C., Garbutt, J., Cohen, M. M., Cira, J., et al. (2005). Patient concerns about medical errors in emergency departments. *Academic Emergency Medicine*, 12(1):57–64.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Clark, V. L. and Kruse, J. A. (1990). Clinical methods: the history, physical, and laboratory examinations. *JAMA*, 264(21):2808–2809.

Gans, S. L., Pols, M. A., Stoker, J., Boermeester, M. A., expert steering group, et al. (2015). Guideline for the diagnostic pathway in patients with acute abdominal pain. *Digestive surgery*, 32(1):23–31.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing inferences from self-selected samples*, pages 115–142. Springer.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88(423):984–993.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576.

Greenland, S. and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

Li, Z., Li, Q., Han, C.-P., and Li, B. (2014). A hybrid approach for regression analysis with block missing data. *Computational Statistics & Data Analysis*, 75:239–247.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160.

Lorena, A. C., De Carvalho, A. C., and Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30(1-4):19–37.

Macaluso, C. R. and McNamara, R. M. (2012). Evaluation and management of acute abdominal pain in the emergency department. *International journal of general medicine*, 5:789.

McDonald, R., Hall, K., and Mann, G. (2010). Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 456–464.

Monteiro, S. D., Sherbino, J. D., Ilgen, J. S., Dore, K. L., Wood, T. J., Young, M. E., Bandiera, G., Blouin, D., Gaissmaier, W., Norman, G. R., et al. (2015). Disrupting diagnostic reasoning: do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians? *Academic Medicine*, 90(4):511–517.

Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug information journal: DIJ/Drug Information Association*, 34(2):525–533.

Ng, S. S., Tse, P. W., and Tsui, K. L. (2014). A one-versus-all class binarization strategy for bearing diagnostics of concurrent defects. *Sensors*, 14(1):1295–1321.

NYCDE (2017). *New York City Department of Education Demographic data*. https://infohub.nyced.org/reports-and-policies/citywide-information-and-data/information-and-data-overview.

NYSED (2017). *New York State Education Department student data*. https://data.nysed.gov.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *Journal of machine learning research*, 5(Jan):101–141.

Roderick, J., Little, A., and Rubin, D. B. (2002). *Statistical analysis with missing data*. J. Wiley.

Shavlik, J. W., Dietterich, T., and Dietterich, T. G. (1990). *Readings in machine learning*. Morgan Kaufmann.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

Wallach, H. M. (2004). Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., Initiative, A. D. N., et al. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206.

Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927.