Alternative Methods via Random Forest to Identify Interactions in a General Framework and Variable Importance in the Context of Value-Added Models

by

Arturo Valdivia

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2013 by the
Graduate Supervisory Committee:

Randall Eubank, Chair
Dennis Young
Mark Reiser
Ming-Hung Kao
Jennifer Broatch

ARIZONA STATE UNIVERSITY

December 2013

ABSTRACT

This work presents two complementary studies that propose heuristic methods to capture characteristics of data using the ensemble learning method of random forest. The first study is motivated by the problem in education of determining teacher effectiveness in student achievement. Value-added models (VAMs), constructed as linear mixed models, use students' test scores as outcome variables and teachers' contributions as random effects to ascribe changes in student performance to the teachers who have taught them. The VAMs teacher score is the empirical best linear unbiased predictor (EBLUP). This approach is limited by the adequacy of the assumed model specification with respect to the unknown underlying model. In that regard, this study proposes alternative ways to rank teacher effects that are not dependent on a given model by introducing two variable importance measures (VIMs), the *node-proportion* and the *covariate-proportion*. These VIMs are novel because they take into account the final configuration of the terminal nodes in the constitutive trees in a random forest. In a simulation study, under a variety of conditions, true rankings of teacher effects are compared with estimated rankings obtained using three sources: the newly proposed VIMs, existing VIMs, and EBLUPs from the assumed linear model specification. The newly proposed VIMs outperform all others in various scenarios where the model was misspecified.

The second study develops two novel interaction measures. These measures could be used within but are not restricted to the VAM framework. The *distribution-based* measure is constructed to identify interactions in a general setting where a model specification is not assumed in advance. In turn, the *mean-based* measure is built to estimate interactions when the model specification is assumed to be linear. Both measures are unique in their construction; they take into account not only the outcome values, but also the internal structure of the trees in a random forest. In a separate

simulation study, under a variety of conditions, the proposed measures are found to identify and estimate second-order interactions.

To Darío

ACKNOWLEDGEMENTS

This manuscript represents not only the completion of the dissertation work but also the culmination of a journey that started years ago.

I would like to start thanking Randy Eubank, my dissertation advisor. I have been extremely fortunate to culminate this journey under his mentorship. Week after week and month after month, his dedication, advice, guidance, and encouragement have been unparalleled. His moral support and complete commitment to this project have forever changed my conception of the selflessness and dedication a professor could have towards a student, and it will have a long lasting effect in the manner I approach my professional and academic career.

I also thank the dissertation committee members, Dennis Young, Mark Reiser, Ming-Hung Kao, and Jennifer Broatch, for their insightful comments and recommendations at different stages of the dissertation.

I would also like to thank Sharon Lohr. With a strong sense of what my academic path should be, she has shaped it from the moment I started my graduate studies in statistics, and she has had an important contribution to the present work. In addition, she facilitated funding for conference expenses and my research during three summers. Furthermore, George Runger and Yan Yang had insightful comments and recommendations in the early stages of research related to this dissertation.

I am thankful with Debbie Olson for her role as administrative lifeguard, predisposition to guide me through the thick and thin of the graduate school administrative maze, and exceptional ability to put down fires from the first day to the last.

I am grateful to those outstanding individuals with whom I had the opportunity to work in parallel to my graduate studies. Lattie Coor, Sybill Francis, Alan Brown, Wendy

Wolfersteig, Marjorie Kaplan, Paul Holley, Cassandra Larsen, Benah Parker, Miguel Montiel, Maria Moratto, and many others. In their different roles of chairman, director, supervisor, or co-worker, they have influenced my professional career and development. Moreover, I thank my classmates and friends that have made my stint in graduate school enjoyable and unforgettable.

I thank my parents, Beatriz, Alberto, and Darío, and younger siblings, David and Montserrat. They are a constant reminder of what is dear to me and a source of peace and joy that keeps me sane in moments of weakness and despair. My elder brother, Walter, has read a number of drafts of this manuscript at different stages of the process and provided useful comments. More importantly, he has been and still is a ubiquitous influence in my life, a counselor of sorts. Whether living in the same household or thousands of miles apart, I am privileged to have continual interaction and enjoy thought-provoking discussions with him about important topics that often transcend academia.

I also thank part of my extended family, Julia Steinberg, Marko Svetina, and Olena Tsurska, not only for their feedback on the drafts of this work but also for their limitless encouragement and support.

Finally, I would like to thank Dubravka Svetina, my companion in life. Her contribution and influence on this dissertation has extended well beyond her marital promises. Through countless days and nights, she has morphed into what was needed most: advisor, mentor, assistant, proofreader, English teacher, and/or therapist. To simply say that this work would not have been possible without her help will understate her real contribution. I am extremely fortunate to share my life with somebody who I admire for her high moral principles, noble and passionate heart, personal and professional accomplishments, and highly competitive spirit that challenges me and

continuously makes me a better person. Now that the arrival of a new family member is impending, I look forward to sharing with them the new challenges and rewards in life.

Research related to this work was partially supported by the National Science Foundation under grant DRL-0909630.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Much of the debate about educational reform has centered on teacher and school accountability. Attempts to measure teacher's influence on student achievement have been of interest in the scientific community for several decades (Hanushek, 1971; Bryk and Weisberg, 1976; Hanushek, 1979). As a result, programs such as the Tennessee Value-Added Assessment System (Sanders et al., 1997) have been implemented in specific states or school districts to account for school and/or teacher effects since the early 1990s. However, since the most recent reauthorization of The Elementary and Secondary Education Act, The No Child Left Behind Act of 2001, a major emphasis has been placed on setting standards that each teacher must meet in order to be considered highly qualified. As a consequence, many states and school districts have adopted or are in the process of adopting programs intended to measure teachers' effects on student achievement.

Somewhat more generally, decision-makers are actively pursuing an environment where teacher accountability could be used as an instrument for promotion and retention. In the 2012 State of the Union address, President Obama stated, "Teachers matter. So instead of bashing them, or defending the status quo, let's offer schools a deal. Give them the resources to keep good teachers on the job, and reward the best ones. And in return, grant schools flexibility: to teach with creativity and passion; to stop teaching to the test; and to replace teachers who just aren't helping kids learn." This position is not new in the research community. Goldhaber and Hansen (2010), among others, propose linking administrative decisions related to teachers (for example, remuneration and tenure) to measures of their contributions toward student learning, arguing that this approach agrees with administrative decision-making based on employees' level of productivity.

Since teacher effectiveness cannot be measured directly, many researchers have used value-added models (VAMs) as an indirect approach to assessing effectiveness. These models attempt to ascribe changes in student achievement to their corresponding teachers and/or schools. A number of models have been proposed and are currently used for this purpose. Most of these models, henceforth called traditional VAMs, are either special cases or extensions of a general mixed model described in McCaffrey et al. (2004) (e.g., McCaffrey and Lockwood (2011); Mariano et al. (2010)). In these models, students' test scores are used as outcome variables while the contributions of teachers are treated as random effects. Hence, the value-added score for a teacher is obtained as the predicted value of the random effect.

The appropriateness of the use of VAMs in education is an ongoing debate (Stewart, 2006; Rothstein, 2009, 2010; Briggs and Domingue, 2011; Kinsler, 2012). This debate could be approached in different ways. For example, we could question the validity of VAMs in education as an appropriate instrument for decision making since we cannot make causal inferences due to the lack of randomness in student assignment (i.e., does teachers' effectiveness cause students' progress?), a problem inherent to most observational studies. Alternatively, we could address limitations in the current VAMs by proposing alternatives to improve them. For example, Karl et al. (2011) address problems associated with missing data not at random, often encountered in educational settings.

This study contributes to the understanding of VAMs and seeks to enhance their usefulness in assessing teacher's performance by addressing issues related to the rigid model structure they impose on data. Specifically, limitations arise because the linear model structure only includes covariates that are explicitly included in the model; typically, few interactions are considered and nonlinearity is generally only addressed through quadratic terms. While it might be possible that a certain teacher is more effective with a certain group of students, such situation can only be taken into account in

the VAM if the corresponding interaction is modeled in advance. Otherwise, the VAM would be misspecified.

To address VAM limitations, we work with the ensemble learning method of random forest proposed by Breiman (2001). The advantage of using this approach is that no structure is predetermined, as opposed to the traditional VAMs. Therefore, in principle, any possible effect would be taken into account and discovered.

The use of data mining and statistical learning methodology is not entirely uncommon in educational research (Baker and Yacef, 2009). For example, applications of data mining methods in education have been employed for predicting student outcomes such as graduation (Mendez et al., 2008) and retention (Chong et al., 2010), and some work has been done to determine the relative contribution of different learning methods (Beck and Mostow, 2008). However, the use of data mining techniques to measure teacher effectiveness in students' outcomes has yet to receive research attention.

This dissertation examines two different, yet related topics. The first part compares the information about teacher effects obtained using traditional VAM methodology and data mining techniques. For the former, we use two linear mixed models: the covariate adjustment model and the gain score model. For the latter, we work with random forest. In the linear mixed models, the teacher effects are obtained using the empirical best linear unbiased predictors (EBLUPs). In random forest, we are not aware of any existing methodology that directly quantifies teacher effects. Therefore, we work with variable importance measures (VIMs) to rank teacher effects. In particular, we develop and propose two new VIMs based on the final configuration of terminal nodes in the regression trees that compose the random forest: the *node-proportion* and the *covariate-proportion* VIMs. In a simulation study, under a variety of conditions, true rankings of teacher effects are compared with estimated rankings obtained using three

sources: the newly proposed VIMs, existing VIMs, and the EBLUPs from the assumed linear model specification.

The second part of this dissertation introduces two novel methods to assess interactions using characteristics of the random forest. The *distribution-based* measure is constructed to identify interactions in a general setting where a model specification is not assumed in advance. The *mean-based* measure is built to estimate interactions when the model specification is assumed to be linear. Both measures are unique in their construction; they take into account not only the outcome values, but also the internal structure of the trees in random forest. Specifically, given two variables, we consider the frequency of appearance of those variables in different branches in each tree as well as their relative position with respect to each other and with respect to the root node. The distribution-based interaction measure is used to identify potential interactions and is based in the final configuration of the splitting variables in the trees. By contrast, the mean-based interaction measure obtains interaction estimates using the structure of the tree to assign weights to a linear combination of relevant node outcome means.

Chapter 2 presents the background literature for linear models, VAMs, and data mining methods used in this study. Chapter 3 introduces the proposed VIMs and a simulation study that compares the proposed VIMs with existent VIMs and those obtained using the EBLUPs. Chapter 4 presents the proposed interaction measures and a simulation study to determine the measures' accuracy. Concluding remarks and discussion of the limitations of the study as well as future research directions are presented in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we introduce a general framework that encompasses linear models, VAMs, and data mining models. In addition, different variable importance measures and interaction detection techniques that derive from the data mining literature are examined and computational considerations are discussed.

2.1 A General Framework

Let $\mathbf{X} = (X_1, \ldots, X_P) \in \mathbb{R}^P$ be a real valued random P-vector of continuous or categorical random variables and $Y \in \mathbb{R}$ a real valued random variable. We assume that Y is determined by

$$Y = F(\mathbf{X}) + \epsilon, \tag{2.1}$$

where F is a real valued function and $\epsilon \sim N(0, \sigma^2)$.

Given realizations of $(\mathbf{X}, Y)$, $\{\mathbf{x}_i, y_i\}_{i=1}^N$, the interest is in using this information to obtain a function $\hat{F}(\cdot)$ that predicts Y given values of $\mathbf{X}$. We are also interested in some cases in concomitant parameters estimators that may be obtained in the process of evaluating $\hat{F}(\cdot)$.

We consider two different approaches for inference about F. In the first, we assume that F has a given structure; i.e.,

$$Y = F(\mathbf{X}; \Theta) + \epsilon, \tag{2.2}$$

for Θ a vector of unkhown model parameters. Then, we use data that are realized from (2.2) to estimate the value of Θ. Analysis based on traditional VAMs derived from this perspective is described in Section 2.2. Our second approach is nonparametric in that it presumes no *a priori* structure for F and instead employs predetermined procedures and algorithms to estimate certain of its features. A large number of data mining methods, including random forest, use this tactic and are described in Section 2.3.

2.2 Linear Mixed Models in the Context of VAMs

VAMs in education are models that measure the additional value a teacher (school or program) contributes to student achievement. In this study we center on teacher contributions, but extensions to school or program contributions are straightforward (McCaffrey et al., 2004).

In terms of (2.1), VAMs assume that the model specification is linear in the parameters. Moreover, the VAMs considered here assume teacher effects are random. This assumption, although not uncommon, is not pervasive. Some research in VAMs treats teacher effects as being fixed. Since the choice is an assumption, in principle, both formulations are plausible (Demidenko, 2004, p. 55).

In this study we choose to analyze teacher influence as random rather than fixed effects based on three justifications. First, the fixed effects approach would be preferable if the number of teachers is small and the number of students per teacher is large (Demidenko, 2004, p. 55); in educational studies we often have the opposite situation. Second, it seems more natural to assume that teachers are assigned to schools based on a population of teachers with some specific distribution (Searle, 1971, p. 7). Third, by allowing teacher effects to be random, a positive correlation among students in the same class can be introduced in the model through the teacher effects variance-covariance matrix.

Our interest here is not to argue that random effects are more adequate than fixed effects, but rather to provide some arguments for our choice. In any case, the real focus of this study is on the comparison of data mining techniques with linear model methods and the use of random effects is satisfactory for that purpose. Similarly, for the sake of model specificity, all other covariates in the models are considered fixed (e.g., gender, age, ethnicity, free and reduced lunch status).

The linear mixed model specification in the context of VAMs could allow for multivariate teacher effects for each teacher, i.e., different teacher effects assigned to each teacher based on course topics and/or time periods, for example, the model specification used in Mariano et al. (2010) and McCaffrey and Lockwood (2011). Then, it is common to assume that random effects corresponding to the same teacher are correlated while random effects corresponding to different teachers are not. Because our objective is to draw comparisons between VAMs and alternative methods, we present a simplified model specification: the case of univariate teacher effects, i.e. a single teacher effect is assigned to each teacher. A linear mixed model in this context can be represented by

$$\mathbf{y} = \mathbf{U}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{y} = (\mathbf{y_1}^\top, \ldots, \mathbf{y}_N^\top)^\top$ is an Nh-vector and $\mathbf{y}_i = (y_{i1}, \ldots, y_{ih})^\top$ is an h-vector (with h being, e.g., the number of years) of readings from subject i for $i = 1, \ldots, N$. The $Nh \times (p_1 + 1)$ matrix $\mathbf{U} = [\mathbf{U}_1^\top, \ldots, \mathbf{U}_N^\top]^\top$, with $h \times (p_1 + 1)$ matrices $\mathbf{U}_i$ for $i = 1 \ldots, N$, represents the student-level covariates. The $Nh \times p_2$ matrix $\mathbf{Z} = [\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_N^\top]^\top$ is the design matrix that relates the p_2 teachers to each student, and each $h \times p_2$ matrix $\mathbf{Z}_i$ for $i = 1 \ldots, N$, relates the p_2 teachers to student i. The $(p_1 + 1)$-vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p_1})^\top$ represents the fixed effects and $\mathbf{b} = (b_1, \ldots, b_{p_2})^\top$ is an p_2-vector of teacher (random) effects where the b_j's are iid $N(0, \sigma_\tau^2)$ random variables for $j = 1, \ldots, p_2$. Similarly, the Nh-vector $\boldsymbol{\epsilon} = (\epsilon_1^\top, \ldots, \epsilon_N^\top)^\top$ represents the error terms, with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ih})^\top$ and the ϵ_{ik} are iid $N(0, \sigma^2)$ random variables for all $i = 1, \ldots, N$ and $k = 1, \ldots, h$. The values for σ_τ^2, σ^2, and the vector $\boldsymbol{\beta}$ are unknown.

To relate this formulation to (2.2), we take $\Theta = (\boldsymbol{\beta}^\top, \mathbf{b}^\top)^\top$ and $\mathbf{X} = [\mathbf{U}\ \mathbf{Z}]$ with $\mathbf{U}$ the matrix of covariates and $\mathbf{Z}$ a matrix of indicator variables. Then $F(\mathbf{X}; \Theta) = \mathbf{X}\Theta$.

The observed data $\{(\mathbf{U}_i, \mathbf{Z}_i, \mathbf{y}_i)\}_{i=1}^N$ (or some subset thereof) can be used to obtain estimators $\hat{\sigma_\tau^2}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\beta}}$ of σ_τ^2, σ^2, and $\boldsymbol{\beta}$, respectively. In order to obtain a teacher effect prediction we can use an estimator of the conditional expectation of the random effects,

given the observed outcome values: namely,

$$\hat{\mathbf{b}} = \hat{E}(\mathbf{b}|\mathbf{Y} = \mathbf{y}) = \hat{\sigma}^2_\tau \mathbf{Z}^\top \hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\beta}}) \tag{2.4}$$

provides an estimator of $\mathbf{b}$, where

$$\hat{\mathbf{V}} = \mathbf{Z}(\hat{\sigma}^2_\tau I)\mathbf{Z}^\top + \hat{\sigma}^2\mathbf{I} \tag{2.5}$$

is the estimated covariance matrix of $\mathbf{y}$. While there are different approaches that are used to estimate the unknown parameters of the model, the most common choice is the one we use for this study as well: the restricted maximum likelihood estimate or REML as described in Patterson and Thompson (1971).

We next describe the two generic linear mixed models that are used in this study: the Covariate Adjustment Model (CAM) and the Gain Score Model (GSM).

Covariate Adjustment Model

The CAM applies to a single cohort of students in two contiguous years or grades, $h = 1, 2$, where $h = 1$ corresponds to the first grade of the study. We assume M teachers in grade 2 and the dependent variable is students' scores in year 2. The year 1 scores are treated as a covariate. The model assumes that each student has only one teacher per year and teacher effects are random.

Using the superscript c to characterize the CAM, (2.3) for each i can be expressed as

$$y_{i2} = \delta^c y_{i1} + (\mathbf{u}^c_i)^\top \boldsymbol{\beta}^c + \mathbf{z}_i^\top \mathbf{b}^c + \epsilon^c_i, \quad i = 1, \ldots, N. \tag{2.6}$$

Assuming p covariates, $(\mathbf{u}^c_i) = (u_{i0}, \ldots, u_{ip})^\top$ with $u_{i0} = 1$. Then, in (2.2), we have $\mathbf{u}_i = (y_i, (\mathbf{u}^c_i)^\top)^\top$, $\boldsymbol{\beta} = (\delta^c, (\boldsymbol{\beta}^c)^\top)^\top$, for $\boldsymbol{\beta}^c = (\beta^c_0, \ldots, \beta^c_p)^\top$, the $(p+2)$-vector of fixed effects and δ^c the slope that relates the year 2 scores to the year 1 scores. The vector $\mathbf{z}_i^\top = (z_{i1}, \ldots, z_{iM})$ in 2.6 corresponds to the $\mathbf{Z}_i$ part of the design matrix in (2.3). Since each student has only one teacher per year, the vector $\mathbf{z}_i$ has only one coordinate equal to

one and the rest equal to zero. The teacher effects are represented by the vector $\mathbf{b}^c = (b_1^c, \ldots, b_M^c)$ and we assume that $\mathbf{b}^c \sim \mathbf{N}\Big(\mathbf{0}, (\sigma_\tau^c)^2\mathbf{I}\Big)$. The ϵ_i^c in the model are *iid* $N\Big(0, (\sigma^c)^2\Big)$ random variables that are independent of $\mathbf{b}^c$.

The matrix representation of (2.6) takes the form

$$\mathbf{y}_2 = \delta^c \mathbf{y}_1 + \mathbf{U}\boldsymbol{\beta}^c + \mathbf{Z}\mathbf{b}^c + \boldsymbol{\epsilon}^c. \tag{2.7}$$

The $\mathbf{y}_2 = (y_{12}, \ldots, y_{N2})^\top$, $\mathbf{y}_1 = (y_{11}, \ldots, y_{N1})^\top$, the i-th row of the matrices $\mathbf{U}$ and $\mathbf{Z}$ are $\mathbf{u}_i^\top$ and $\mathbf{z}_i^\top$, respectively, and $\boldsymbol{\epsilon}^c = (\epsilon_1^c, \ldots, \epsilon_N^c)^\top \sim \mathbf{N}\Big(\mathbf{0}, (\sigma^c)^2\mathbf{I}\Big)$.

Because of the simple structure of the CAM, if the N observations are ordered appropriately, it is possible to simplify (2.5). Notice that

$$\hat{\mathbf{V}} = c\hat{o}v(\mathbf{y_2}) = \mathbf{Z}(\hat{\sigma}_\tau^2 I)\mathbf{Z}^\top + \hat{\sigma}^2\mathbf{I} = (\hat{\sigma}_\tau^c)^2\mathbf{A} + (\hat{\sigma}^c)^2 I, \tag{2.8}$$

where $\mathbf{A} = diag(\mathbf{J_1}, \ldots, \mathbf{J_M})$ is a block diagonal matrix and $\mathbf{J_m}$ is a square matrix of ones with dimensions equal to the number of students that have been taught by teacher m, for $m = 1, \ldots, M$. The empirical best linear unbiased predictor (EBLUP) for teacher effects given the observed outcome values in this case is

$$\hat{\mathbf{b}}^c = (\hat{\sigma}_\tau^c)^2\mathbf{Z}'[(\hat{\sigma}_\tau^c)^2\mathbf{A} + (\hat{\sigma}^c)^2\mathbf{I}]^{-1}(\mathbf{y}_2 - \hat{\delta}^c\mathbf{y}_1 - \mathbf{U}\hat{\boldsymbol{\beta}}^c). \tag{2.9}$$

Gain Score Model

The construction of the GSM is similar to the one for the CAM. The main difference is that the dependent variable is now the difference of year 2 and year 1 scores. Specifically, we have $y_i^g = y_{i2} - y_{i1}$ and

$$y_i^g = \mathbf{u}_i'\boldsymbol{\beta}^g + \mathbf{z}_i'\mathbf{b}^g + \epsilon_i^g, \tag{2.10}$$

for $i = 1, \ldots, N$. Here the superscript g denotes that the parameters correspond to the GSM and, apart from the δ^c slope parameter, the model is defined as in (2.6). The matrix representation of (2.10) is

$$\mathbf{y}^g = \mathbf{U}\boldsymbol{\beta}^g + \mathbf{Z}\mathbf{b}^g + \boldsymbol{\epsilon}^g, \tag{2.11}$$

and the EBLUPs for teacher effects given the observed outcome values are

$$\hat{\mathbf{b}}^g = (\hat{\sigma}_\tau^g)^2 \mathbf{Z}'[(\hat{\sigma}_\tau^g)^2 \mathbf{A} + (\hat{\sigma}^g)^2 I]^{-1}(\mathbf{y}^g - \mathbf{U}\hat{\boldsymbol{\beta}}^g). \tag{2.12}$$

Note that the teacher effects predictions $\hat{\mathbf{b}}^c$ and $\hat{\mathbf{b}}^g$ do not have to be equal, since the different model specifications in (2.6) and (2.10) change the interpretations of these effects. Both, the CAM and the GSM have the limitation of considering only two years of data and produce one year of estimated teacher effects. Therefore, a different model has to be used for each year of student data. Although the usefulness of the GSM has been questioned, it has been shown that the information obtained with this model might still be valid (Williams and Zimmerman, 1996).

Notice that (2.9) and (2.12) are central in the analysis of variable importance measures, presented in Chapter 3. These equations also play a role of shrinkage estimators, relative to the least squares estimators for $\mathbf{b}$ in an alternative model, where $\mathbf{b}$ is instead a vector of fixed effects (Robinson, 1991). This happens because the EBLUP for teacher effects takes into account the entire data, while the alternative least square estimators only consider that teacher's own students. Therefore, the EBLUP for $\mathbf{b}$ shrinks toward their mean. As we describe in Section 2.4, a similar shrinkage effect is realized from certain data mining methods.

2.3 Data Mining Methods

In the previous section we assumed that F in (2.1) had certain structure with corresponding parameters and used realized data from the model to estimate (via REML) the unknown parameters. This, in turn, produces a fitted model that can be used for predictions. In this section, we go in a somewhat different direction and introduce data mining methods. The motivation continues to be finding a function $\hat{F} := F_\alpha$ that estimates F in (2.1) and produces model predictions. Here α represents the data mining method or algorithm used. The approach now is to use a set of realizations of (X, Y) to obtain $\hat{F}$ without assuming an underlying parametric structure for F. Once F_α has been

determined, it is often possible to describe this function using a set of parameters, Ξ; i.e., given a new observation $\mathbf{x}$, the predicted outcome is given by $\hat{y} = F_\alpha(\mathbf{x}; \Xi)$. Notice that the resulting parameters Ξ are of a fundamentally different nature than the model parameters Θ in (2.2). In particular, Ξ is dependent on the realizations of the predictor variable $\mathbf{X}$ used to obtain F_α while the vector Θ has no particular meaning here because no underlying structure is assumed for F.

We begin by describing decision trees that are the constitutive elements of various supervised learning methods such as classification and regression trees, CART, (Breiman et al., 1984) and random forest (Breiman, 2001). The latter method is a key component for the developments in Chapters 3 and 4. Accordingly, we conclude the chapter with a description of the random forest algorithm as well as specific characteristics of the method that are relevant for our study.

Decision Trees

Although the label *decision tree* was not adopted at that time, one of the first descriptions of a decision tree as a recursive method was presented by Belson (1959). Both the dependent and independent variables were treated as being binary and the independent variable that was selected for splitting was taken to be the one that seemed the most associated with the dependent variable, where the strength of association was measured by the largest difference between expected and observed frequencies in a contingency table.

Morgan and Sonquist (1963) introduced the first known computer program to obtain predictions using a decision tree. Their original motivation was the lack of methods that could appropriately address the limitations of the application of multivariate statistical techniques to survey data (Morgan and Sonquist, 1963; Sonquist and Morgan, 1964). In particular, the authors contended that most stratified and clustered survey samples had severe limitations in the proper applications of statistical

tests of significance. They argued that this was due to the difficulty presented in the construction of theoretical models that reflected chains of causation when intercorrelation or interaction between predictors was present (Sonquist and Morgan, 1964, p. 139).

The Morgan and Sonquist algorithm is known as the *Automatic Interaction Detector (AID)*. It can be viewed as providing the foundation for several decision tree algorithms that are in use today. The problem addressed by the method corresponds to predictions of a single continuous dependent variable using one or more independent variables or predictors. Although predictors could be continuous in nature, they must be converted to categorical variables in order to be used in the algorithm. Then, for each variable, splitting points that divide the data set in two subsets are determined based on those categories.

To grow a tree, the AID algorithm proceeds as follows. Starting with the entire data set, the data are split into two subsets based on the predictor that minimizes the dependent variable residual sum of squares. Once the data set has been split, the procedure continues recursively on each new subset, until some type of stopping criterion is reached.

The methods described so far focus on categorical predictors with any continuous variable being transformed into a categorical one prior to the analysis. CART is one of the most popular techniques that allows for continuous independent variables. It uses different splitting criteria for classification and regression: namely, the Gini Index and the sum of squared deviations as in AID, respectively.

A number of alternatives to CART have also been proposed. C4.5 (Quinlan, 1993) is similar in nature to CART with differences arising in the splitting criterion. It uses entropy instead of the Gini Index for classification trees and techniques tested empirically rather than cross-validation to estimate error rates.

We next describe decision trees grown using CART, the method used in our study. As before, $\{\mathbf{x}_i, y_i\}_{i=1}^N$ is a set of N observations from (2.1) with $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$ and y_i being the i-th realizations of $\mathbf{X} = (X_1, \ldots, X_P)$ and Y, respectively. The estimators of F are based on decision trees that are obtained from binary recursive partitioning methods or sets of rules that allow the splitting of data into different groups. These rules split the data in terms of the values of one covariate at a time. They are called trees because they can be represented by a collection of nodes and branches corresponding to the splits that were made of the data. When a decision tree is used with a continuous response to estimate F in (2.1), it is generally referred to as a "regression" tree. When Y is categorical the term "classification" tree is used instead.

To build a tree we start with all the observations of a data set in the root node of the tree. Then, one of the covariates is selected along with a splitting point on that covariate. This determines which observations go to the left branch and which ones go to the right branch. We continue this process recursively until some stopping criterion is reached, resulting in a collection of subsets of the data. These subsets are called the leaves or terminal nodes of the tree. By contrast, nodes that split are called non-terminal or internal.

We denote by D and $\bar{D}$ the number of internal and terminal nodes, respectively. For example, in Figure 2.1, the tree has 8 non-terminal nodes (including the root node) and 9 terminal nodes. In addition, η_d, $d = 1, \ldots, D$ and $\bar{\eta}_d$, $d = 1, \ldots, \bar{D}$ represent individual internal and terminal nodes, respectively, for a particular tree and we will refer to a specific node as η or $\bar{\eta}$ when no confusion arises. When the relation between the child nodes and their parent node is needed, we use η^L and η^R for the left and right child nodes of η.

A graphical representation of a regression tree is shown in Figure 2.1. This corresponds to educational data where a cohort of students in two contiguous years is

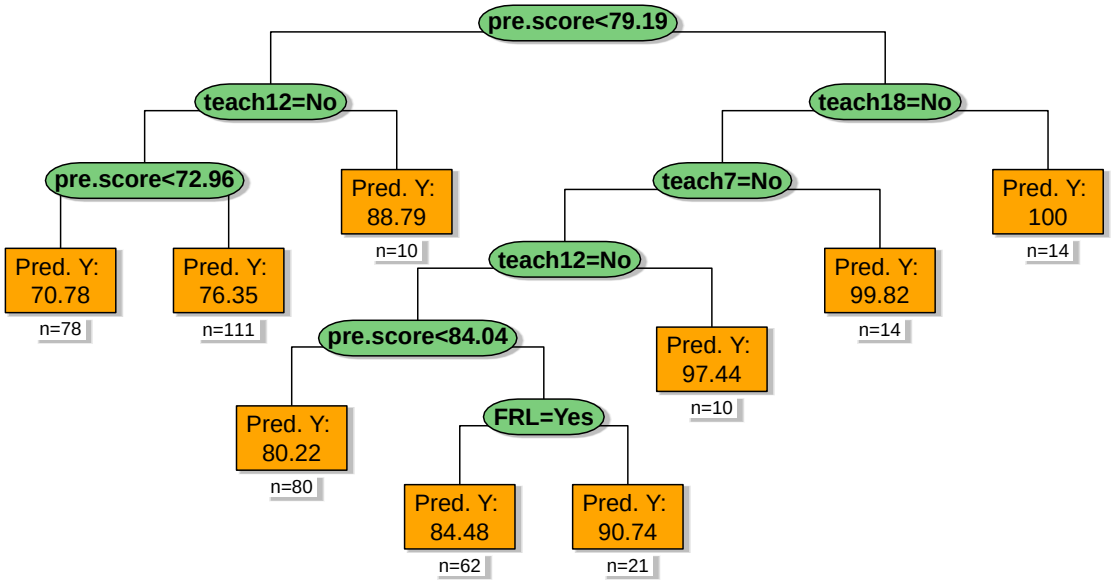

Figure 2.1: An example of a regression tree

considered. The scores obtained in year 2 are the response variable and the covariates or input variables are year 1 scores (pre.scores), students' demographic information such as gender or free or reduced luncheon indicators (FRL), and students' teacher indicators. The green colored node at the top is the root node. The data are split into two groups here with those responses in agreement with the condition inside the green colored node being assigned to the left branch of the tree. For example, students having year-1 scores less than 79.19 are assigned to the left child node while those with values of at least 79.19 are relegated to the right child. The splitting process continues using the predictor variables until it terminates with the terminal nodes that are represented as orange squares.

The terminal nodes of a tree correspond to nonoverlapping regions in the range of the predictor variable $\mathbf{X}$. The predicted value for a response with covariate values in terminal node $\bar{\eta}_d$ is then taken to be some constant c_d. We will use $\bar{\eta}_d$ to represent both the d-th terminal node and the region it defines in the predictor space. With that

convention, a regression tree may be defined in terms of the parameter vector $\Xi = (\bar{\eta}_1, c_1, \ldots, \bar{\eta}_{\bar{D}}, c_{\bar{D}})$ and the associated estimator of F at $\mathbf{X} = \mathbf{x}$ is

$$F_\alpha(\mathbf{x}; \Xi) = \sum_{d=1}^{\bar{D}} c_d I(\mathbf{x} \in \bar{\eta}_d). \tag{2.13}$$

The goal is now to obtain an optimal choice for Ξ. If least squares is used for our optimality criterion, we could consider using

$$\hat{\Xi} = \underset{\Xi}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - F_\alpha(\mathbf{x}_i; \Xi))^2. \tag{2.14}$$

This may lead to a tree with a single observation in every terminal node. Thus, some constraints are required such as an *a priori* choice for the minimum number of observations in a node. Sometimes a penalty term for tree complexity is appended to the least squares criterion in (2.14).

If for example, we restrict the minimum number of observations per terminal node, then the constants are estimated by

$$\hat{c}_d = \bar{y}_d = \frac{\sum_{i=1}^{N} y_i I(\mathbf{x}_i \in \bar{\eta}_d)}{\sum_{i=1}^{N} I(\mathbf{x}_i \in \bar{\eta}_d)}. \tag{2.15}$$

Thus, all that remains is the determination of the terminal nodes. This, in turn, is determined by the splitting algorithm. A so-called greedy method is one that at each step picks the covariate and associated split point that gives the largest reduction in the total error sum of squares. The resulting choice for Ξ provides an approximation to $\hat{\Xi}$ in (2.14).

For example, in the case of the tree in Figure 2.1, the left most terminal node corresponds to the case of observations of students having year 1 score less than 72.96 and a teacher other than number 12. The average value of year 2 score is 70.78 for the response values that correspond to this particular range for the predictors. Thus, if a new observation arrived that had a year 1 score value of 71 with any teacher but number 12, its year 2 score value would be predicted as 70.78.

Limitations of Decision Trees and Alternative Methods

The AID technique suffers several shortcomings in statistical analysis. Doyle (1973) illustrated this in an analysis of the results obtained by Heald (1972) using the AID program. Drawbacks of the method include the need for very large sample sizes, the risk that the tree built based on intercorrelated predictors produces spurious results, bias created from the model-building process, bias obtained from noise, and bias produced by skewed variables.

Doyle (1973) suggested that AID should be used primarily as an exploratory or descriptive method to gain insight about the correct model specification (in terms of possible non-linearities or interactions) and then only when substantial prior information is not available. Doyle and Fenwick (1975) argued that AID could only be useful if it produces a model specification that could be validated by traditional techniques such as regression.

In order to address some of the problems with AID type methods, formal tests of significance have been proposed (e.g. Messenger and Mandell (1972), Kass (1975), and Scott and Knott (1976)). In its original formulation, AID only used a standard t-test to assess the differences between data subsets that are obtained after a split. However, this test gives us less information that we might like. By construction, the between group sum of squares t-statistic is maximized and therefore irrelevant predictors can still generate significant t-statistics.

Messenger and Mandell (1972) proposed an alternative splitting criterion which targets the selection of the splitting variable that maximizes the number of observation in each modal category. The algorithm based on this splitting criterion is called THAID and is described in detail in Morgan and Messenger (1973). Kass (1980) has argued that this criterion is missing a solid statistical foundation and there is limited knowledge about its theoretical behavior.

Kass (1975) developed a statistic to test the null hypothesis that a predictor is completely unrelated to the dependent variable. It was developed for one single predictor with C categories and a continuous dependent variable although the extension to several uncorrelated predictors is straightforward.

Kass (1975) assumes that the predictor of interest is monotonic, so that there are $C-1$ possible splits. Let n represent the number of observations at an arbitrary node in the tree with n^L and n^R being the number of observations in each of the subsets obtained after splitting. If $\bar{y}$, $\bar{y}^L$, and $\bar{y}^R$ denote the dependent variable means for the group and both subsets, respectively, and S^2 is the variance of the dependent variable, the proportion of variance explained by the ℓ-th split is $P_\ell = n^L n^R(\bar{y}^R - \bar{y}^L)^2/n^2S^2$. An optimal split is one for which the explained proportion of variance is

$$K = n \max_{\ell \in \{1,\ldots,(C-1)\}} \{P_\ell\}. \tag{2.16}$$

Let n_i denote the number, $\tilde{Y}_i$ the response mean, and $r_i = n_i/n$ the relative frequency for observations that correspond to the i-th category, for $i = 1, ..., C$. Under the null hypothesis of no correlation between the dependent variable and the predictor, each observation is equally likely to belong to any category. Under this assumption, the probability distribution of the mean $\tilde{Y}_i$ in category i with n_i observations is approximately normal with mean $\tilde{Y}$, the grand response mean, and variance $(1-r_i)S^2/n_i$ and the joint distribution of response means for each category is approximately C-variate normal with mean vector $\tilde{\mathbf{Y}} = (\tilde{Y}, \ldots, \tilde{Y})$ and covariance matrix $S^2\Sigma/N$, where $\Sigma = (\sigma_{ij})$ for

$$\sigma_{ij} = \begin{cases} (n-n_i)/n_i, & \text{if } i = j, \\ -1, & \text{if } i \neq j. \end{cases}$$

If we now take

$$\nu_i = n^{1/2}(r_i/(1-r_i))^{1/2}(\tilde{Y}_i - \tilde{Y})/S$$

for $i = 1, ..., C$, approximately, the joint probability distribution of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_C)$ given $\mathbf{r} = (r_1, \ldots, r_C)$ is multinormal with mean $\mathbf{0}$ and covariance matrix Σ_ν where $\Sigma_\nu = (\sigma^\nu_{ij})$

and

$$\sigma_{ij}^{\nu} = \begin{cases} -(r_i/(1-r_i))^{1/2}(r_j/(1-r_j))^{1/2}, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

If $h(\cdot|\mathbf{r})$ denotes the corresponding normal density function, the null permutation distribution of $K = (\max_i |\nu_i|)^2$ is approximately

$$Prob(K \leq k) = \int \cdots \int_R h(\boldsymbol{\nu}|\mathbf{r}) d\nu_1 \ldots d\nu_{C-1},$$

where the region of integration R depends in general on r_i and k.

Of particular interest for our study are predictors with 2 categories: i.e., C=2. In that case $K = \nu_1^2$ with ν_1 approximately $N(0,1)$ and K approximately χ_1^2 distributed.

Scott and Knott (1976) give an alternative approximation for the distribution of the Kass (1975) test statistic. Under the null hypothesis with a nominal unordered predictor they show that, assuming C/N remains fixed as $C \to \infty$ and $\max r_i \to 0$, then $\left(\sqrt{K} - \sqrt{2(\sum_{i=1}^{C} \sqrt{r_i/C})^2 C/\pi}\right)$ has an approximate $N\left(0, 1 - \frac{2N + (\sum_{i=1}^{C} \sqrt{r_i/C})^2 C}{N\pi}\right)$ distribution while K has approximately the same distribution as $m\chi_{\nu_0}^2$ where

$$m = 1 - \frac{2N + \left(\sum_{i=1}^{C} r_i^{1/2}\right)^2}{N\pi},$$

and $\chi_{\nu_0}^2$ is a chi-squared random variable having

$$\nu_0 = \frac{\left(\sum_{i=1}^{C} r_i^{1/2}\right)^2}{2\left(\pi - 2 + \frac{\left(\sum_{i=1}^{C} r_i^{1/2}\right)^2}{N}\right)}$$

degrees of freedom. Scott and Knott (1976) suggest that this approximation could provide guidance to determine if the results obtained at each node of the tree were statistically significant.

Kass (1980) introduced the *chi − squared* Automatic Interaction Detector (CHAID) that uses significance testing for each split. This algorithm is suitable only for

situations where both the predictors and dependent variable are categorical. For a given group of observations, the chosen predictor to be used as the splitting variable is based on the most significant test statistic obtained among predictors. For a given predictor, this statistic is based on the χ^2 statistic for contingency tables when the number of classes of this predictor has not been reduced, or an approximation of the χ^2 statistic when the contingency table has been reduced. The reduction of the contingency table is obtained in a stepwise manner to approximate the optimal reduction based on all the possible combination of classes in the contingency table to obtain the optimal χ^2 statistic. In addition, CHAID introduces the possibility of having multi-way splits based on the number of remaining classes in the reduced contingency table for the chosen predictor of that split.

An important difference between CART and AID, THAID, or CHAID is that the stopping criteria for growing the tree is the least restrictive in CART. This is because a large tree is pruned based on error rates obtained from cross-validation.

GUIDE (Loh, 2002) is an algorithm developed to eliminate variable selection bias via chi-squared analysis of residuals and bootstrap calibration of significance probabilities. However, these analyses are not often applicable and require that continuous variables be transformed into categorical variables (as in THAID and CHAID).

Hothorn et al. (2006) introduced methodology to construct regression trees via a technique called unbiased recursive partitioning. What they propose is to divide the recursive tasks of selecting a splitting variable and determining the splitting point for that variable into two different steps in the creation of nodes for a tree.

Given a node or subset of observations, the first step is to use this set to test a global null hypothesis of independence between the response and any of the covariates. These tests are constructed using permutation methods (Strasser and Weber, 1999). The global null hypothesis is the intersection of partial null hypotheses for each covariate. The

partial hypotheses are tested via linear statistics that depend on non-random transformations of the covariate and permutations of the subset of responses. The intuition behind this test is that the conditional distributions of these linear statistics are unknown. However, it is possible to estimate them using the conditional distribution of the response variable based on permutations of the subset of response outcomes.

Once the conditional distribution of the linear statistics is estimated for each covariate, decisions about the partial null hypotheses and therefore the global null hypothesis can be made. Since the test statistics for each covariate might not be measured on the same scale, to determine which covariate is chosen as a splitting variable, P-values for the conditional distribution of test statistics for each covariate are used thereby allowing for cross-covariate comparisons.

The second step consists of determining the splitting point in the chosen splitting variable. Hothorn et al. (2006) proposes using a similar testing procedure to determine the optimal splitting point. The recursive process stops when the global null hypothesis cannot be rejected for any resulting partition.

Random Forest

Random forest is a classification or regression tree variant proposed by Breiman (2001). It uses the basic regression tree estimator in conjunction with bootstrap aggregation or bagging (Breiman, 1996).

The bagging premise is that one first generates T bootstrap samples $\mathcal{B}_1, \ldots, \mathcal{B}_T$ by sampling at random with replacement from the original data set. One then computes the estimator of interest for each bootstrap sample and averages the result. In the case of regression trees this translates into the computation of the parameter vector $\hat{\Xi}(\mathcal{B}_t)$ using (2.14) for $t = 1, \ldots, T$ by applying a splitting algorithm to each of the bootstrap samples.

The resulting estimator of F at $\mathbf{X} = \mathbf{x}$ is

$$\hat{F}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} F_{\alpha}\Big(\mathbf{x}; \hat{\Xi}(\mathcal{B}_t)\Big). \tag{2.17}$$

The random forest method builds on this idea with an additional step that randomly reduces the number of variables to be used for splitting at each step in the creation of terminal nodes. Specifically, for each of the T bootstrap samples, a tree is constructed where at each of its current terminal nodes we randomly select $p \leq P$ of the covariates and use the best one of these variables to produce a binary split, until the stopping criterion is attained.

One of the most appealing arguments for the use of random forest is that it has been shown that the addition of more trees, however large, does not produce an overfit of the solution, but approaches the most efficient solution. In practice, the limit is approached with a moderate number of trees.

Random forest has obtained a large empirical success for classification and regression problems. However, very few theoretical results have been presented (Breiman, 2004; Biau et al., 2008). Extensions to random forest that account for specific characteristics of the educational data include, among others, clustering in individual regression trees (Toth and Eltinge, 2011), correlated data in random forest (Strobl et al., 2007)), and measures of variable selection and variable importance scores (Genuer et al., 2010).

2.4 Variable Importance Measures

Variable importance measures (VIMs) or importance scores are measures used to determine the relative contribution that each covariate has in predicting the dependent variable. There are different VIMs that have been proposed in the literature on regression trees and random forests. This section describes several of those measures. As an aside, we mention here that in the VAM setting there are indicator variables that correspond to

the random effects. This fact will be used in Chapter 3 to obtain new methods of assessing effect importance with VIMs.

Let us now assume that the data have been split into two disjoint sets: $\mathcal{L}_N$ and $\mathcal{J}_N$ that we refer to as the learning (or training) and test data, respectively. Breiman et al. (1984) then propose an importance measure for decision trees based on the estimated improvement in squared error loss that a variable has in the internal nodes of the tree. In this regard we denote the estimated test error based on squared error loss at node η by

$$\widehat{err}(\eta) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{J}_N} (y_i - \bar{y}_\eta)^2 I(\mathbf{x}_i \in \eta),$$

where

$$\bar{y}_\eta = \frac{1}{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}_N} I(\mathbf{x}_i \in \eta)} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}_N} y_i I(\mathbf{x}_i \in \eta).$$

That is, $\widehat{err}(\eta)$ is the sum of the squared differences between the values of the outcome variables in the test data set that arrive at node η and the mean of the outcome variables from the training data set in node η. The estimated improvement in squared error loss is defined as

$$\hat{\Delta}(\eta) = \widehat{err}(\eta) - (\widehat{err}(\eta^L) + \widehat{err}(\eta^R)).$$

Now suppose that a decision tree has D internal nodes, $\eta_1, \ldots, \eta_D$. The importance measure for variable X_p is

$$VI_p = \sum_{d=1}^{D} \hat{\Delta}(\eta_d) \phi_p(\eta_d), \tag{2.18}$$

where

$$\phi_p(\eta) = \begin{cases} 1, & \text{if covariate } X_p \text{ is used as the splitting variable in node } \eta, \\ 0, & \text{otherwise.} \end{cases}$$

In words, the relative importance of covariate X_p is the sum of estimated improvements in squared error loss for every node where X_p is used as the splitting variable.

Intuitively, there are two components that influence the importance measure of a variable. First, the covariates that are found by the splitting criterion that are closer to the root of the tree are potentially more important than those covariates that are closer to the leaves. This is because more observations are considered in nodes closer to the root and therefore the estimated improvement in squared error loss tends to be greater. Second, for a particular node, the magnitude of the difference between the child node means determines the importance of a covariate relative to that node.

Notice that a covariate that has a large number of categories will have a larger number of possible splitting points. If this number is large relative to the number of categories of other covariates, the variable importance values tend to be large as well. Therefore, this measure may be biased toward covariates with larger numbers of categories.

Hothorn et al. (2006) shows that their recursive partitioning method is not biased in that it does not favor the selection of covariates with larger numbers of categories. Otherwise, they find that the method performs as well as traditional recursive partitioning methods.

Breiman's VIM idea extends readily to bagging and random forest based estimators. For example, with a random forest with T trees, $t_1, \ldots, t_T$, we use the average of these measures obtained for every single tree: i.e., $VI_p = \frac{1}{T}\sum_{t=1}^{T} VI_p^t$, with VI_p^t the value of VI_p for the tth tree. For each tree, the observations in the training set not used to generate that tree are used as the test data set. Although random forest will correct some bias given the random selection of covariates for every node, variables with larger numbers of categories are still favored in selection, and tend to obtain larger variable importance measures.

For random forests the bootstrapping mechanism that is used to create the estimators can be exploited to obtain measures of a variable's importance. The key to

doing so relies on extracting information from the observations that have not been resampled. With random forest, the training set that is used to construct a tree is the bootstrap sample from the original data set: a sample with replacement that also has N observations, some of which are sampled more than once while others are not sampled at all. Suppose we use T bootstrap samples $\mathcal{B}_1, \ldots, \mathcal{B}_T$. The training data set in the b-th sample is indicated with a superscript as $(x_{i1}^{(b)}, ..., x_{iP}^{(b)})$ for $b = 1, \ldots, T, i = 1, ..., N$. Now, for every tree obtained using a bootstrap sample, the observations in the original data set that are not considered in this sample form the *out-of-bag* samples (OOB). For a particular bootstrap sample $\mathcal{B}$, we use $\mathcal{B}^c$ to denote the corresponding OOB sample.

A random forest specific VIM suggested by Breiman (2001) has been called permutation accuracy importance (PAI). For each covariate X_p, it is obtained as the difference in prediction accuracy between the results of predicting the original OOB data set and its permuted version, where the permutation occurs only for covariate X_p.

Formally, let $F_\alpha\Big(\cdot; \hat{\Xi}(\mathcal{B})\Big)$ be the regression tree estimator produced by the random forest algorithm corresponding to a particular bootstrap sample $\mathcal{B}$. The associated estimated prediction accuracy is

$$\Lambda\Big(\hat{\Xi}(\mathcal{B})\Big) = \frac{1}{|\mathcal{B}^c|} \sum_{i:(\mathbf{x}_i, y_i) \in \mathcal{B}^c} (F_\alpha\Big(\mathbf{x}_i; \hat{\Xi}(\mathcal{B})\Big) - y_i)^2, \tag{2.19}$$

where $|\mathcal{B}^c|$ is the number of observations in the OOB sample corresponding to $\mathcal{B}$. The prediction accuracy for X_p over the entire forest of T trees is the average prediction accuracy of all the trees: i.e.,

$$\Lambda(X_p) = \frac{1}{T} \sum_{t=1}^{T} \Lambda\Big(\hat{\Xi}(\mathcal{B}_t)\Big).$$

We now permute the values of covariate p in the OOB samples to create a new sample for each tree. That is, the data set values are the same for all the observations and covariates, except those corresponding to covariate X_p. Those values are randomly reassigned in a different order. If $\mathbf{x}_1^{(p)}, \ldots, \mathbf{x}_N^{(p)}$ denotes the permuted covariate vectors,

the resulting accuracy is

$$\Lambda^*(X_p) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{B}_t^c|}\sum_{i:(\mathbf{x}_i,y_i)\in\mathcal{B}_t^c}\Big(F_\alpha\Big(\mathbf{x}_i^{(p)};\hat{\Xi}(\mathcal{B}_t)\Big)-y_i\Big)^2.$$

Then, the variable importance measure based on PAI is

$$PAI = \Lambda^*(X_p) - \Lambda(X_p). \tag{2.20}$$

The intuition behind the PAI measure is that if a covariate is important, the permutation should produce a large gap between the prediction accuracy of the original OOB samples and the one obtained from permuting that variable. So, large values of PAI suggest that a covariate has predictive utility.

PAI is not reliable when covariates are of different types (e.g., continuous and categorical, qualitative and quantitative), the variables are on a different scale of measurement, or the variables have different numbers of categories. This happens because a covariate with a larger number of categories relative to other covariates will tend to have a better prediction accuracy and a larger difference with the permuted version. Hence, it will be biased towards covariates with larger numbers of categories. Also, the PAI overestimates the importance of correlated covariates; variables that are not important might be considered much more relevant because they might be highly correlated with other covariates.

The conditional variable importance concept of Strobl et al. (2008) takes into account the correlation among covariates by using a conditional permutation to minimize the effect of correlated variables. The method is built following a similar framework to that used for the PAI measure. First, the OOB prediction accuracy before permutation is obtained as before using (2.19). Then, the procedure creates a grid over the predictor variable space by creating splits on all the covariates other than the one of interest, using the same splitting values and variables that were originally employed to generate the tree. The values of the focal variable are then permuted within this grid and the modified

prediction accuracy is obtained. As with PAI, the difference between the original and modified prediction accuracy is used to reflect the importance of the covariate in a given tree. The average of this quantity over all the trees is the conditional variable importance for the variable.

When testing a null hypothesis of independence between the dependent variable, Y, and a covariate, X_p, if this covariate is correlated with others, the PAI measure would tend to overestimate its importance. The conditional variable importance method approximates a test of conditional independence between the dependent variable and the covariate. Notice, however, that if the covariate is independent of other covariates, PAI and conditional variable importance should produce similar results. In the simulation study described subsequently, the VIMs obtained with PAI were more accurate than those obtained with conditional variable importance.

OOBForest is a methodology introduced by Tuv et al. (2009) to determine variable importance. It uses training samples and sums of squared differences. It takes advantage of OOB samples by using them to select the best splitting attribute on each node with the same information criterion as the one used in the traditional random forest. This method reduces the bias in variable importance and is faster than the random forest algorithm *cForest* implemented in the *party* package in the R language.

pForest Variable Importance is a partial permutation method to measure variable importance. As described in Deng (2011), it basically compares the importance score of X_p for each $p = 1, ..., P$ with the importance score of X'_p, a partially permuted version of X_p. For each tree t, let $VI_t(X_p)$ and $VI_t(X'_p)$ be the importance score for X_p and X'_p, respectively. Then, take $VI(X_p) = (1/T)\sum_{t=1}^{T} I[VI_t(X_p) > VI_t(X'_p)]$. Assuming independence, $\sum_{t=1}^{T} I[VI_t(X_p) > VI_t(X'_p)]$ follows a binomial distribution with probability π_p that $VI_t(X_p) > VI_t(X'_p)$, where π_p has some specified value. The value of *pForest* variable importance is then defined to be the smallest fraction of rows used in the partial

permutation of X_p that are needed to obtain $VI_t(X_p) > VI_t(X'_p)$ with given probability π_p. The smaller this fraction, the higher the variable importance.

A number of additional algorithms have been adapted or created to address variable importance. These include the asymptotic p-value ANOVA F-test or χ^2-test (Loh and Shih, 1997), variable importance using GUIDE (Loh, 2012), the asymptotic p-value of conditional inference test (Hothorn et al., 2006), and the exact p-value of maximally selected Gini Gain (Strobl et al., 2007), among others.

When obtaining VIMs based on random forest, there is a shrinkage effect similar to the one present in the linear mixed model random effects. Intuitively, this occurs because only a subset of variables is considered to select the splitting variable at each node. Therefore, even when a covariate random effect is much larger than others, this covariate can only appear on the nodes where it has been considered for selection.

Similarly to EBLUPs for teacher effects, the VIMs obtained from random forest take into account the entire data, not just the teacher's own students. The shrinkage effect will be influenced by the number of students each teacher has with the consequence that teachers with fewer students do not appear as frequently in the trees. Since each tree is built based on a bootstrap sample, a teacher with few students might have even fewer students or no students at all, in certain trees. Additionally, most VIMs are determined by the number of observations affecting each node in which the covariate is used as a splitting variable and the number of times that covariate appears in the tree. Therefore, teachers with fewer students might not only be considered less important than teachers with more students, but their effect estimates might be less accurate than those for teachers with more students.

We have described several approaches to determine variable importance measures. These approaches have advantages and limitations. The main characteristic of most approaches is that they use the accuracy of predictions in normal and altered conditions

to assess a covariate's importance. While some of these methods have shown empirical success, they are based on strong assumptions about the distribution of the covariates, the independence between covariates, the differences in variable types among covariates, etc. Even the conditional approach cannot fully take into account the possible correlation between covariates.

In summary, although empirical results have shown that the accuracy of random forest predictions is comparable to the best machine learning methods, variable importance measures have not had the same level of success. In Chapter 3 we propose new variable importance measures that derive from a different perspective that does not rely on prediction accuracy.

2.5 Interactions

Interactions have been a topic of study in statistical learning for several decades. In the literature on data mining, diverse methods have been proposed in an attempt to measure or identify interactions. The following section describes several of those methods.

Data Mining Methods for Interaction Detection

The first attempt to capture interactions was via the AID and AID III algorithms. Sonquist and Morgan (1964) and Sonquist et al. (1971), respectively, introduce a series of interaction measures based on the premise that interactions are determined through the subgroups that are affected by the predictors belonging to the same branch in the tree. The authors also propose that interdependence could be assessed with their method; if, for example, there are two candidate predictors that are being evaluated for splitting and one of them is chosen, if the other predictor is no longer relevant (in terms of the dependent variable residual sum of squares) in at least one of the resulting two subsets, these two predictors are deemed to be highly dependent. If a predictor is not considered as the splitting variable for any of the subsets in the tree, this predictor may not matter. Doyle (1973) and Doyle and Fenwick (1975) found limitations with this approach.

Friedman and Popescu (2003, 2008) introduce a method and a test statistic to determine interactions. Assume that F in (2.1) is twice differentiable in $\mathbf{X}$. If variables X_p and X_q interact, then $E_{\mathbf{X}}\left[\frac{\partial^2 F(\mathbf{X})}{\partial X_p \partial X_q}\right]^2 > 0$. To see this observe that if there is no interaction between X_p and X_q, then $F(\mathbf{X})$ could be expressed as the sum of a function that does not depend on X_p and a function that does not depend on X_q and, accordingly, the second order mixed partial derivative will vanish identically. The extension to high order interactions is straightforward. For example, for a third order interaction when variables X_p, X_q, and X_r do not interact we have $E_{\mathbf{X}}\left[\frac{\partial^3 F(\mathbf{X})}{\partial X_p \partial X_q \partial X_r}\right]^2 > 0$.

Friedman and Popescu (2008) then characterize interactions via partial dependence functions. Let $\mathbf{X}_s$ be a subset of predictor variables corresponding to variables with indices $s = \{s_1, \ldots, s_p\} \subset \{1, 2, ..., P\}$ for $p \leq P$ and use $\mathbf{X}_{-s}$ to denote the collection of variables that remain after those in $\mathbf{X}_s$ are removed from $\mathbf{X}$. The partial dependence of $F(\mathbf{X})$ on $\mathbf{X}_s$ is defined by $F_s(\mathbf{X}_s) = E_{\mathbf{X}_{-s}}[F(\mathbf{X})]$. Given the set of N observations used originally to determine and estimator $\hat{F}$ of F, F_s can be estimated by

$$\hat{F}_s(\mathbf{x}_{ks}) = \frac{1}{N}\sum_{i=1}^{N} \hat{F}(\tilde{\mathbf{x}}_i), \tag{2.21}$$

for $k = 1, \ldots, N$. Here, $\mathbf{x}_{ks}$ is the kth realization of variables with indices in s and

$$\tilde{x}_{ij} = \begin{cases} x_{kj}, & j \in s, \\ x_{ij}, & j \notin s, \end{cases}$$

for $j = 1, \ldots, P$.

If two variables X_p and X_q do not interact, the partial dependence of $F(\mathbf{X})$ on $\mathbf{X}_s$ with $s = \{p, q\}$ could be represented as the sum of the partial dependence functions for each variable

$$F_{pq}(X_p, X_q) = F_p(X_p) + F_q(X_q) \tag{2.22}$$

and $F_{pq}(X_p, X_q), F_p(X_p)$, and $F_q(X_q)$ can be estimated using (2.21). Similarly, if X_p does not interact with any other variable, then

$$F(\mathbf{X}) = F_p(X_p) + F_{-p}(\mathbf{X}_{-p}). \tag{2.23}$$

Friedman and Popescu (2008) introduce a statistic based on these partial dependence functions that can be used to test for the presence of an interaction between two variables X_p and X_q. It takes the form

$$H^2_{pq} = \frac{\sum_{i=1}^{N}[\hat{F}_{pq}(x_{ip}, x_{iq}) - \hat{F}_p(x_{ip}) - \hat{F}_q(x_{iq})]^2}{\sum_{i=1}^{N} \hat{F}^2_{pq}(x_{ip}, x_{iq})}, \tag{2.24}$$

where x_{ip} is the i-th realization of variable X_p. It measures the fraction of the variance of $\hat{F}_{pq}(X_p, X_q)$ not accounted by $\hat{F}_p(X_p)$ and $\hat{F}_q(X_q)$. A related statistic for testing the interaction of variable X_p and any other variable is provided by

$$H^2_p = \frac{\sum_{i=1}^{N}[\hat{F}(\mathbf{x}_i) - \hat{F}_p(x_{ip}) - \hat{F}_{-p}(\mathbf{x}_{i,-p})]^2}{\sum_{i=1}^{N} \hat{F}^2(\mathbf{x}_i)} \tag{2.25}$$

where $\mathbf{x}_{i,-p}$ is the i-th realization of variables $\mathbf{X}_{-p}$. Similarly, the third order interaction between variables X_p, X_q, and X_r can be assessed with

$$H^2_{pq} = \sum_{i=1}^{N} \Big[\hat{F}_{pqr}(x_{ip}, x_{iq}, x_{ir}) - \hat{F}_{pq}(x_{ip}, x_{iq}) - \hat{F}_{pr}(x_{ip}, x_{ir}) - \hat{F}_{qr}(x_{iq}, x_{ir}) + \tag{2.26}$$

$$+ \hat{F}_p(x_{ip}) + \hat{F}_q(x_{iq}) + \hat{F}_r(x_{ir})\Big]^2 / \sum_{i=1}^{N} \hat{F}^2_{pqr}(x_{ip}, x_{iq}, x_{ir}). \tag{2.27}$$

If there is no interaction, all three H statistics should be near zero. Larger values of the H statistic correspond to stronger interactions.

Previous Attempts to Identify Interactions Using Random Forest

There have been only a few attempts to identify interactions using random forests. Winham et al. (2012) measure interactions based on the rankings obtained from variable importance measures in the context of genome wide association studies. The authors work with a data set where all variables are categorical, including the dependent variable.

They then consider variable importance methods that are based on the mean decrease in accuracy and Gini importance. Their study selects the k highest ranked variables, where k is determined by the number of variables with causal effects. It uses *Heritability* as the degree of genetic determination of a trait to differentiate marginal effects from interaction effects (see Winham et al. (2012, p. 4)). The authors also compare their results to those obtained using p-values from univariate logistic regression for different numbers of covariates (dimension). They find that the random forest variable importance measures fail to detect interaction effects in high-dimensional data in the absence of a strong marginal component.

Kelly and Okada (2012) study variable interaction measures by using random permutation of OOB samples in a random forest. This method derives directly from the variable importance measure proposed by Breiman (2001). It exploits random data permutations and measures the amount of information that is gained when another variable is present based on the errors obtained when permuting OOB cases in the random forest. It is considered to be a method prototype and is currently developed for random forest classification problems and its extension to regression problems has not been addressed.

2.6 Computational Methods and Software

In order to obtain regression trees, random forest, conditional trees and random forest based on a conditional inference framework, we make use of several packages developed in the programming language R. We conclude this chapter with a brief overview of these software resources.

The *rpart* package (Therneau and Atkinson, 1997) was originally developed for the programming language S, and then adapted to R. The package routines implement several of the procedures developed in Breiman et al. (1984), and, in particular, the regression trees we use in this work. The programs build regression trees using a two

stage method. In the first stage, the tree is built by determining the covariates to be used as splitting criteria and then determining the associated splitting points. This process continues recursively until one of the stopping criteria is reached. These criteria include the maximum number of observations in a terminal node, the number of nodes, and a threshold for the improvement in prediction between the parent nodes and the child nodes. In the second stage, the tree is pruned using cross-validation. The time and computer power needed for the *rpart* routines is generally very small (just a few seconds) even for data with a few hundred covariates and thousands of observations.

The package *randomForest* (Liaw and Wiener, 2002) is based on the original Fortran code introduced by Breiman (2001) and can be used for regression and classification problems. It implements the random forest algorithm for regression as described in Section 2.3 and can obtain variable importance measures as described in Section 2.4.

It is possible to modify a few parameters in the algorithm: the number of trees to be used, the number of variables considered at each node for splitting, different types of criteria to determine how much to grow the trees and the types of results and indicators obtained alongside the final tree. The time and computer power needed for the *randomForest* routines is dependent on these parameters. For example, for a sample of 3600 observations with about 100 covariates and 1000 trees requires about 3gb of RAM and between 80 to 240 seconds of computing time on a modern intel core processor. When the proximity matrix and variable importance measures are obtained, the computing time increases three times or more, depending on the covariates numbers of categories.

The *party* package is built in the environment R and implements the Hothorn et al. (2006) methodology. The function *ctree* carries out unbiased tree algorithms through conditional inference trees. The package introduces the function $cforest$ that builds a random forest based on these unbiased trees (Strobl et al., 2009), where binary

partitioning is used alongside conditional inference procedures to obtain conditional inference trees.

This package has more flexibility than *randomForest*. The user has the ability to change all the parameters that could be modified in *randomForest*, plus additional functions, such as $model - based$ recursive partitioning with the *mob* function.

The time and computer power needed for *party* routines is much higher than for previous packages. In particular, when obtaining variable importance measures with the *varimp* function, the program stores all the permutations necessary for the conditional inference procedures. For example, for a sample of 400 observations with about 40 covariates and 500 trees with the default length, the *varimp* function requires about 16gb of RAM and between 150 to 250 minutes of computing time on a modern intel core processor. Furthermore, since the package only approximates uncorrelated covariates for the variable importance results, when the covariate are effectively uncorrelated, the results are suboptimal in comparison to those found using *randomForest* routines.

CHAPTER 3

VARIABLE IMPORTANCE MEASURES

Random forest has become a popular data mining method that has been found useful in several research fields. Its appeal lies in its predictive accuracy that is comparable to the best machine learning methods. In particular, random forest performs well when the structure of the underlying model is nonlinear, the number of covariates is very large, covariates are highly correlated, and/or complex covariate interactions are present. Random forest can also be used to produce variable importance measures (VIMs) for variable selection purposes.

VIMs have found applications to variable ranking and selection problems in a variety of settings during the last decade. But, as far as we know, this is the first time they have been used to assess relative contributions from random effects such as those from teachers in a VAM related context. In this chapter, we propose new VAM-relevant VIMs that are employed in simulations to compare their performance with the estimated teacher effects one obtains via linear models methodology. Comparisons of these two approaches when the linear model is misspecified are of particular interest and relevance for our particular avenue of research.

3.1 A New Approach to Variable Importance Measures

In this section, we propose two new VIMs that are constructed by taking into account the final configuration of the trees' terminal nodes in a random forest rather than just differences in prediction accuracy. This is the point of departure for our work from other measures that can be found in the literature.

As in the previous chapter, we presume there are N observations in the training data set, $\mathcal{L}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i$ and y_i the values for the predictors and response variables, respectively, for the ith observation. Then, bootstrap samples $\mathcal{B}_t$, $t = 1, \ldots, T$, from $\mathcal{L}_N$ are used to grow the trees in a random forest.

A by-product of regression trees and random forests is the proximity matrix. This is an $N \times N$ symmetric matrix where every cell represents the proportion of ocurrences where the observation corresponding to the row position belongs to the same terminal node as the observation for the column position. If we consider the proximity matrix for a single tree, it is a matrix of zeros and ones, where a coordinate with a *one* indicates that two observations, the first determined by the row position and the second by the column position, share the same terminal node.

To be a bit more precise, suppose there are T trees and the terminal nodes in tree t are $\bar{\eta}_d^t, d = 1, \ldots, \bar{D}_t$. The entry in the ith row and jth column for the proximity matrix for tree t is

$$\sum_{d=1}^{\bar{D}_t} I(\mathbf{x}_i \in \bar{\eta}_d^t) I(\mathbf{x}_j \in \bar{\eta}_d^t). \tag{3.1}$$

This is zero or one depending on whether or not $\mathbf{x}_i$ and $\mathbf{x}_j$ are contained in some common node of the tree. For random forests this requires a bit more explanation. Here the sample of observations used to build the tth tree is $\mathcal{B}_t$, a bootstrap sample of $\mathcal{L}_N$. Nonetheless, the corresponding proximity is built with respect to $\mathcal{L}_N$ using comparisons only between those observations in $\mathcal{L}_n \backslash \mathcal{B}_t^c$; the entries for columns (and rows) of the proximity matrix corresponding to observations in $\mathcal{B}_t^c$ are all set to zero. Note that the set $\mathcal{L}_n \backslash \mathcal{B}_t^c$ contains the same observations as $\mathcal{B}_t$ but without duplicates.

The proximity matrix for the random forest (or any collection of trees) is produced by averaging across trees as

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{d=q}^{\bar{D}_t} I(\mathbf{x}_i \in \bar{\eta}_d^t) I(\mathbf{x}_j \in \bar{\eta}_d^t). \tag{3.2}$$

This formula provides the key insight into the VIMs we will subsequently develop. It suggests means by which one may access information about the actual elements, rather than just their average, that comprise the terminal nodes of a tree and we will exploit this facility in what follows.

Node- and Covariate-Proportions

A typical variable importance measure, for example (2.18) and (2.20), is ultimately based on some measure of predictive ability such as

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i:(\mathbf{x}_i,y_i)\in\mathcal{L}_N\setminus\mathcal{B}_t^c}\left(\sum_{d=1}^{\bar{D}_t} I(\mathbf{x}_i\in\bar{\eta}_d^t)(y_i-\bar{y}_d^t)^2\right) \tag{3.3}$$

with $\bar{y}_d^t$ the mean response for elements whose independent variable values fall in $\bar{\eta}_d^t$. More generally, the squared error and node average could be replaced by other quantities of possible interest. The main point is that the only feature of the terminal nodes that enters into the performance assessment is a single summary measure and importance is gauged by improvement in prediction.

The new VIMs we consider here differ from this standard approach in that they are not concerned with prediction of outcomes but rather with the influence a variable has on the composition of the terminal nodes in a tree. Our measures are tailored for use with binary covariates such as the presence or absence of a particular variable in the model. In particular, for a VAM set-up, the covariates of interest are the teachers and therefore amenable to analysis using binary variables. In the discussion that follows we make the simplifying assumption that $X_1,\ldots,X_p$ are binary in nature. The case of more complex variables represents an avenue for future study.

We define the node-proportion VIM for the covariate X_m as

$$\Psi_m=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_m^t}\left\{\sum_{k:(\mathbf{x}_k,y_k)\in\mathcal{L}_N\setminus\mathcal{B}_t^c}\left[x_{km}\sum_{d=1}^{\bar{D}_t} I(x_k\in\bar{\eta}_d^t)p_m(\bar{\eta}_d^t)\right]\right\}, \tag{3.4}$$

with

$$N_m^t=\sum_{i:(\mathbf{x}_i,y_i)\in\mathcal{L}_N\setminus\mathcal{B}_t^c} x_{im}$$

the number of observations in $\mathcal{L}_N \backslash \mathcal{B}_t^c$ where X_m was realized as 1 (e.g., the number of students taught by the mth teacher) and

$$p_m(\bar{\eta}_d^t) = \frac{\sum\limits_{i:(\mathbf{x}_i, y_i) \in \mathcal{L}_N \backslash \mathcal{B}_t^c} I(\mathbf{x}_i \in \bar{\eta}_d^t) x_{im}}{\sum\limits_{i:(\mathbf{x}_i, y_i) \in \mathcal{L}_N \backslash \mathcal{B}_t^c} I(\mathbf{x}_i \in \bar{\eta}_d^t)}$$

the proportion of observations in the dth terminal node of tree t with $X_m = 1$. The expression between the brackets in (3.4),

$$x_{km} \sum_{d=1}^{\bar{D}_t} I(x_k \in \bar{\eta}_d^t) p_m(\bar{\eta}_d^t), \tag{3.5}$$

can be viewed as variable X_m's marginal importance due to the kth observation in the tth tree. If η_d^t is the terminal node containing $\mathbf{x}_k$ and the realized value of X_m for this observation is zero then (3.5) does not contribute to the value of Ψ_m. When $x_{km} = 1$, the value of (3.5) is determined by the proportion of observations in node η_d^t having $X_m = 1$. Thus, Ψ_m is determined by these marginal importance values averaged over those observations that themselves have $X_m = 1$ and across all trees. It is nonnegative and bounded above by one.

If we focus on the teacher effect scenario, $X_m = 1$ if a student is taught by the mth teacher. So, in that instance Ψ_m is the average over teacher m's students of the average (across trees) proportion of students that were also taught by teacher m in the terminal nodes they inhabit. Values of Ψ_m that are closer to unity will indicate a stronger teacher effect in the sense of producing a (relatively) more homogenous subset of terminal nodes that can be attributed directly to the teacher.

Our other proposed measure is the covariate-proportion

$$\Upsilon_m = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_m^t} \left\{ \sum_{k:(\mathbf{x}_k, y_k) \in \mathcal{L}_N \backslash \mathcal{B}_t^c} \left[x_{km} \sum_{d=1}^{\bar{D}_t} I(x_k \in \bar{\eta}_d^t) q_m(\bar{\eta}_d^t) \right] \right\}. \tag{3.6}$$

with

$$q_m(\bar{\eta}_d^t) = \frac{1}{N_m^t} \left(\sum_{i:(\mathbf{x}_i, y_i) \in \mathcal{L}_N \backslash \mathcal{B}_t^c} I(\mathbf{x}_i \in \bar{\eta}_d^t) x_{im} \right)$$

the proportion of observations in the tth tree with $X_m = 1$ that inhabit the dth terminal node. The expression between the brackets in (3.6) is again an assessment of variable X_m's marginal importance due to the kth observation in the tth tree. Note that the only difference between Ψ_m and Υ_m is the proportion used to build each measure; the former uses the fraction of observations in node η_d^t that have $X_m = 1$ while the latter is the fraction of observations having $X_m = 1$ that inhabit node η_d^t. As with Ψ_m, Υ_m takes values in $[0, 1]$ with values closer to one suggesting that X_m has more influence on the terminal nodes of the trees in a forest.

It is insightful to examine how our two VIM measures relate to the proximity matrix. We illustrate this by examining the covariate-proportion measure. Assume for simplicity that each tree was constructed with the entire training set $\mathcal{L}_N$ rather than bootstrap samples. Then, (3.6) could be expressed as

$$\begin{aligned}\Upsilon_m &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_m}\sum_{k=1}^{N} x_{km}\sum_{d=1}^{\bar{D}_t} I(\mathbf{x}_k \in \bar{\eta}_d^t)\frac{\sum_{i=1}^{N} I(\mathbf{x}_i \in \bar{\eta}_d^t)x_{im}}{N_m} \\ &= \frac{1}{N_m^2}\sum_{i=1}^{N}\sum_{k=1}^{N} x_{im}x_{km}\frac{1}{T}\sum_{t=1}^{T}\sum_{d=1}^{\bar{D}_t} I(\mathbf{x}_i \in \bar{\eta}_d^t)I(\mathbf{x}_j \in \bar{\eta}_d^t).\end{aligned}$$

It is easy to see here that we are constructing a summary measure directly from the proximity matrix. Specifically, we average its entries over all pairs of observations that both have $X_m = 1$ to obtain an empirical assessment of the chance that observations with $X_m = 1$ will cohabit terminal nodes with others having that same quality.

To provide an illustration of the use of our new variable importance measures we considered a specific data set from the simulation that we discuss in detail in Section 4.2. Figure 3.1 shows the node-proportion VIMs for a dataset that was generated using (2.6) with 40 teachers. Figure 3.1a) is a barplot of the absolute values of the true teacher effects that were used to produce the data while Figure 3.1b) gives the corresponding node-proportion VIMs. Observe first that as long as a teacher has taught some students, his/her node-proportion VIM will be larger than zero; however, the teachers with the

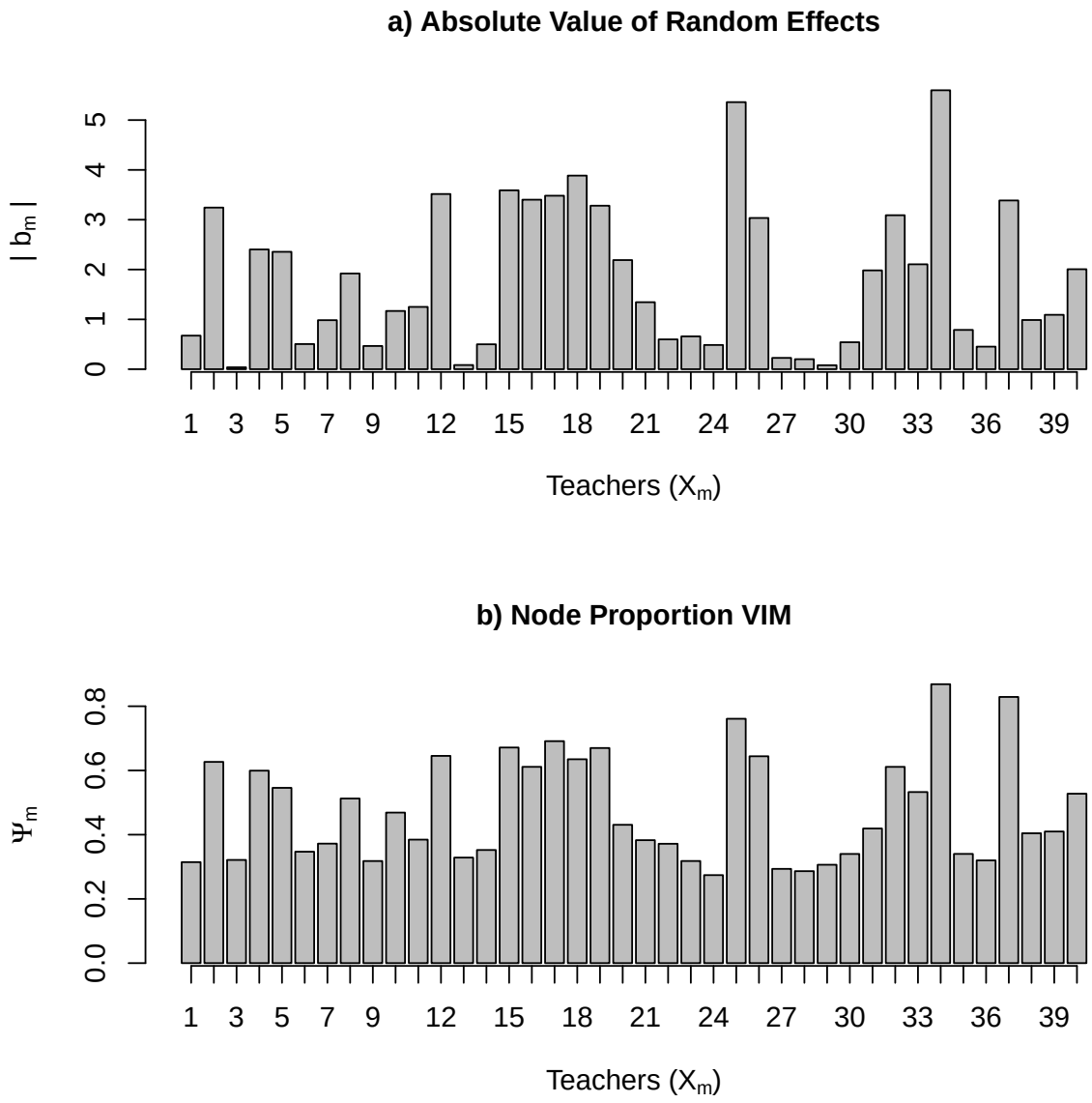

Figure 3.1: a) The absolute value of random effects, $|b_m|$, and b) the node-proportion VIM, Ψ_m, for each teacher (X_m) for data generated with 40 teachers.

smaller random effects (in magnitude) tend to have the lowest node-proportion VIM values. Of course, the node-proportion VIMs do not estimate the magnitude of the teacher effects but rather the relative importance of each effect. Thus, Figures 3.1a) and b) are comparable only in terms of the order they imply for teacher effects.

As in the next section, we can use Spearman's correlation between the absolute value of random effects and VIM values to provide an indication of how adequately the VIMs order the teacher effects. For this example, Spearman's correlation between the absolute value of the random effects and the node-proportion VIM in Figure 3.1 is 0.95.

As another illustration, Figure 3.2 presents the results from data generated using (2.6) with 40 teachers where in this instance only two teacher effects were different from zero: namely, those for teacher 1 and teacher 21. The teacher effect for teacher 1 is positive and about 50% larger in magnitude than the one for teacher effect 21, that is negative. Observe that although the node-proportion measure adequately identifies the teacher effects, it does not differentiate between positive and negative effects. In the next subsection we suggest alternatives for overcoming this limitation.

Comparing VIMs with EBLUPs

We now discuss the comparisons of VIMs with other measures of variable importance derived from the analysis of linear mixed models. These comparisons are relevant in particular when the true relationships that are present in the data set under study include interactions and nonlinear terms. This is because the VIMs based on random forest are capable of capturing these characteristics while the linear mixed model will overlook them unless they are explicitly included in the model specification. First, we explain why this comparison is necessary and then explain the approach that we follow.

In the context of VAMs, both VIMs and EBLUPs provide information about teacher (random) effects. The VIMs determine the relative importance of each teacher

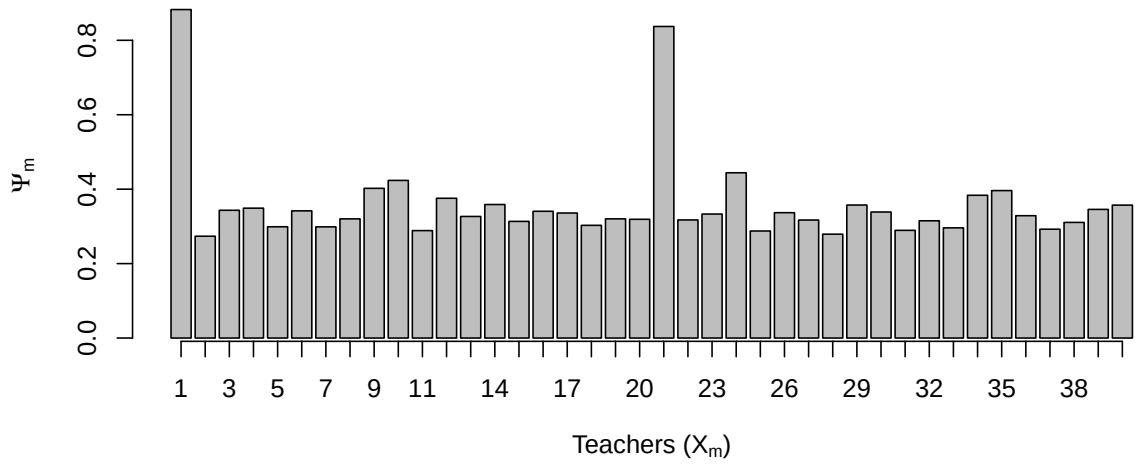

Figure 3.2: The node-proportion VIM for each teacher (X_m) for a data set with 40 teachers and 2 teacher effects: the true teacher effect for teacher 1, $b_1 = 3$, and for teacher 21, $b_{21} = -2$

effect in the model. Simply put, the VIMs produce a ranking of teacher effects. The EBLUPs on the other hand, provide teacher effect predictions.

It is clear that the EBLUPs provide more information than the VIMs. Nevertheless, in the context of VAMs, the first task should be to determine the rankings of teacher effects appropriately. Since the EBLUPs depend on the model specification, while the VIMs do not, when the model is misspecified the VIMs might produce measures that are more appropriate than the EBLUPs. But we can only make such determination if we can adequately compare VIMs and EBLUPs. Specifically, we need to obtain an EBLUP analog of our random forest VIMs.

We will use the ranking obtained from the absolute value of the EBLUPs and refer to it as the VIM for the linear mixed model, or $VIM^{\ell m}$. Observe also that we use the "absolute value" of the EBLUPs, because the ranking of teacher effects obtained using VIMs do not convey information about the direction of the effects.

This is certainly not the only approach to measuring importance in a linear mixed model. An alternative would be to use the t-statistics of the EBLUPs. Since the covariates corresponding to teacher effects are all binary variables, when their class sizes are equal the t-statistics will produce the same ranking as the EBLUPs. Hence, in this case, the two approaches are equivalent.

When we work with simulated data in the next section we will know the true value of all the teacher effects thereby making the comparisons between the absolute value of these effects and all the VIMs feasible. In this case, the importance measure with the ranking that best approaches the ranking of the true teacher effects should be considered preferable. Hence, a statistical measure such as Spearman's rank correlation coefficient will provide a notion of VIM accuracy and that is what we employ for summary purposes.

On the other hand, when we work with empirical data one additional consideration is needed. When the random effects are normally distributed with a mean of zero, the realization of these random effects produces positive and negative values. The VIMs produce only a ranking of influence for the covariates, and this influence could be positive or negative.

To determine the direction of each variable effect, we suggest the use of two methods based on the random forest outcome predictions. The first method is simple; the direction of variable X_m's effect would be positive if the outcome mean for observations affected by variable X_m is at least as large as the overall outcome mean, and negative otherwise. The second method is related to the PAI measure described in (2.19) as we now explain.

From a given bootstrap sample $\mathcal{B}$, the direction of the mth variable's effect could be assessed from the sign of the sum across observations with realized value $X_m = 1$ of simple differences between the predicted outcome values from the original data and the

corresponding values of the same data but without the mth factor contribution; namely,

$$\hat{\delta}^{+}(\mathcal{B}) = 1 \cdot \text{sgn}\Big\{ \sum_{i:x_{im}=1} \Big[F_{\alpha}\Big(x_{i1}, \ldots, x_{i(m-1)}, 1, x_{i(m+1)}, \ldots, x_{iP}; \hat{\Xi}(\mathcal{B})\Big) - F_{\alpha}\Big(x_{i1}, \ldots, x_{i(m-1)}, 0, x_{i(m+1)}, \ldots, x_{iP}; \hat{\Xi}(\mathcal{B})\Big)\Big]\Big\}.$$

The direction of the mth variable effect based on the random forest is then given by

$$\delta^{sign}(X_m) = \text{sgn}\left(\sum_{t=1}^{T} \delta_m^{+}(\mathcal{B}_t)\right).$$

We have suggested only two possibilities and there are undoubtedly many alternative methods that could be used to determine the direction of the covariate effects using VIMs. This type of assessment is important in the context of VAMs where we need to ascertain if a teacher effect is positive or negative. However, the focus of the current work is on determining if the proposed VIMs accurately capture the ranking of teacher effects. Future work will target development of methods to determine effect directionality.

3.2 Simulation Study

We now present the results of a simulation study that was undertaken to determine how our proposed VIMs would perform relative to EBLUPs for ranking effects from random effects models. The specific setting is a mock teacher performance assessment. So, the dependent variable represents a "student score" associated with some random "teacher" effect.

The factor levels that were used in our study were as follows.

1. The number of teachers was taken to be 10, 20, 40, and 100.

2. We divided the teachers into two groups of equal sizes and assigned to each teacher within a group the same number of students, while allowing for teachers in different groups to have different numbers of students. To express this feature in our simulation, we used SpT_{ℓ} to denote the number of students per teacher in group ℓ for $\ell = 1, 2$. The ratios of the number of students per teacher in group 1 to the number of

students per teacher in group 2, SpT_1/SpT_2, that we considered were $\frac{12}{12}$, $\frac{24}{24}$, $\frac{36}{36}$, $\frac{36}{12}$, and $\frac{30}{18}$. For example, if the factor level was $\frac{36}{12}$, the data were generated with 36 students per teacher for all the teachers in group 1 and 12 students per teacher for all the teachers in group 2.

Given that each student was assigned to a single teacher, the sample sizes varied according to the factor combinations of SpT_ℓ and the number of teachers. For example, if the total number of teachers was 20 and the number of students per teacher in group 1 to group 2 ratio was $\frac{36}{12}$, then the data generated contained 480 observations (360+120).

3. The ratio of teacher effect variance (σ^2_τ) to student variance (σ^2) was set at 1, 2, 5, and 20. For example, $\sigma^2_\tau/\sigma^2 = 5$ would indicate that the teacher variance was five times as large as the student variance. For simplicity we chose $\sigma^2 = 1$ so that the teacher effect variance coincides with σ^2_τ/σ^2.

4. The number of trees in random forest was chosen to be 100, 500, 1000, 2000, and 3000. The number of randomly selected variables considered for each split was equal to the highest integer smaller or equal to the total number of covariates square root. The trees were allowed to grow until the terminal nodes had at most five observations.

5. Two model settings were considered: the CAM in (2.6) and the GSM in (2.10). Then, for each setting, a family of four models was examined. The first model for each setting is the baseline presented in (2.6) and (2.10), respectively. For the other cases extensions of the baseline models were employed. In the CAM setting we used

$$y_{i2} = \delta^c y_{i1} + (\boldsymbol{\beta}^c)^\top \mathbf{u}_i + (\mathbf{b}^c)^\top \mathbf{z}_i + \sum_{j=1}^{P}\sum_{k=1}^{K_2} \lambda^c_{jk} u_{ij} z_{ik} + \sum_{j\neq\ell}\sum_{k=1}^{K_2} \lambda^c_{j\ell k} u_{ij} u_{i\ell} z_{ik} + \epsilon^c_{i2} \tag{3.7}$$

while for the GSM setting the extension is

$$y^g_i = (\boldsymbol{\beta}^g)^\top \mathbf{u}_i + (\mathbf{b}^g)^\top \mathbf{z}_i + \sum_{j=1}^{P}\sum_{k=1}^{K_2} \lambda^g_{jk} u_{ij} z_{ik} + \sum_{j\neq\ell}\sum_{k=1}^{K_2} \lambda^g_{j\ell k} u_{ij} u_{i\ell} z_{ik} + \epsilon^g_i. \tag{3.8}$$

The two baseline models allow us to include simulation settings that assume no interaction effects. For the CAM this is true when λ^c_{jk} and $\lambda^c_{j\ell k}$ in (3.7) are zero for all

$j, \ell = 1, ..., P$ and $k = 1, ..., K_2$. Similarly, for the GSM this is true when λ^g_{jk} and $\lambda^g_{j\ell k}$ in (3.8) are zero for all $j, \ell = 1, ..., P$ and $k = 1, ..., K_2$. In these cases, the linear mixed model is correctly specified, and the random effects prediction are the EBLUPS.

For the simulations we consider four covariates associated with fixed effects: the *prescore* is obtained from a normal distribution with mean 75 and variance 21, *gender* is a binary variable obtained from a binomial distribution with probability of success equal to 0.5, *urban* or *rural housing* is another binary variable whose realization is obtained from a binomial distribution with probability of having urban housing 0.4, and *free and reduced lunch program* is binary with probability 0.8 of being part of the program. The associated fixed effects used in the simulations for these variables are .5, .2, 3, and -5, respectively. An overall mean of 50 was used.

The *good teacher - bad teacher* model represents the family of simulations that does not account for interactions effects and keeps most of the assumptions of the baseline model except the one related to the distribution of $\mathbf{b}$ (denoted $\mathbf{b}^c$ for the CAM and $\mathbf{b}^g$ for the GSM). Specifically, these models are constructed with only two teachers having large effects in the model, one positive and one negative, while the rest of the teachers have no effects. This model is used because it is a simple variation of the baseline model where $\mathbf{b}^c$ or $\mathbf{b}^g$ are no longer random. For the simulations, the positive effect is set at $1.5 * \sigma^2_\tau$ and the negative effect at $-1 * \sigma^2_\tau$. Thus, in this case, σ^2_τ is used only for setting the good and bad teacher effects and does not represent the teacher variance. We use the same set of covariates associated with fixed effects and values used in the respective baseline model.

The *simple interaction model* represents the family of models that include second-order interaction effects between one covariate associated with a fixed effect, x_{ij}, and another covariate associated with a random effect, z_{ik}. For the CAM, this is represented from (3.7) by having at least one $\lambda^c_{jk} \neq 0$ for $j = 1, ..., P$ and $k = 1, ..., K_2$. An interaction effect of 10 is considered in the simulations for half of the teachers

randomly determined when a student, taught by one of those selected teachers, lives in the rural area. For the GSM, simple interactions are modeled when at least one $\lambda^g_{jk} \neq 0$ for $j = 1, ..., P$ and $k = 1, ..., K_2$ in (3.8).

The *complex interaction model* is the family of models that include third-order interactions among three covariates, two of them associated with fixed effects and one associated with random effects. For the CAM, this is represented in (3.7) by having at least one $\lambda^c_{j\ell k} \neq 0$ for $j, \ell = 1, ..., P$, $j \neq \ell$, and $k = 1, ..., K_2$. In the simulations, we randomly determined half of the teachers to be susceptible to this interaction effect, and the interaction that was studied corresponded to students living in an urban area, belonging to the free and reduced lunch program, and being taught by one of these teachers. We use an interaction effect value of 20.

In the GSM, a complex interaction occurs when at least one $\lambda^g_{j\ell k} \neq 0$ for $j, \ell = 1, ..., P$, $j \neq \ell$, and $k = 1, ..., K_2$ in (3.8). For the simulations, we have used the same effect values as in the CAM case.

Procedures and Analysis

The fully crossed factorial design would yield a total of 3200 combinations. Initial investigation of the influence of the number of trees suggested that using 1000 trees is adequate. Hence, the results presented here were analyzed based on 1000 trees. In addition, the initial investigation also showed that the combination of 100 teachers and a $\frac{36}{36}$ ratio of number of students per teacher in group 1 to group 2 provided similar results to the combination of 100 teachers and a $\frac{24}{24}$ group 1 to group 2 ratio across every other combination of factors. Thus, only the latter was included in our analysis. The partially crossed factorial design yielded a total of 608 combinations.

Five VIMs were considered: namely,

a) the VIM based on the absolute value of the EBLUPs, $VIM_{\ell m}$, where the subscript "ℓm" stands for *linear model*,

b) the VIM based on the PAI, denoted by VIM_1,

c) the VIM based on the improvement in squared error loss, denoted by VIM_2,

d) the *node proportion*, VIM_Ψ, and

e) the *covariate proportion*, VIM_Υ.

In preliminary results, the VIM based on the conditional recursive partitioning was also considered. However, since the study design did not include correlation among covariates, VIM_1 and conditional variable importance produced similar results, with VIM_1 slightly outperforming the conditional variable importance. In addition, the computing time needed to calculate the conditional variable importance was considerably larger than for the rest of VIMs. As a result the conditional variable importance has not been included in this study.

For every experimental setting, 100 replicates were obtained. Then, the VIMs were compared to the true teacher rankings. For the baseline, simple interactions, and complex interactions models, the Spearman's rank correlation was computed between the absolute value of the *true teacher effects* and the VIMs for each replicate.

For the *good teacher - bad teacher* model, only two teacher effect values were different from zero (the effects of the good and bad teachers), while the VIMs produced rankings for every teacher effect. Spearman's correlation in this situation seemed inadequate to produce relevant comparisons. Hence, we introduced an alternative measure based on the ratio of the average of the true rankings over the average of the estimated rankings for the good and bad teachers. For example, let us assume the estimated rankings from one of the VIMs of the good and bad teachers are (a permutation of) 1 and 4. The true rankings for those teachers must be a permutation of 1 and 2. Our rank based measure would then be $\frac{1+2}{1+4} = .6$. This association measure penalizes heavily when the estimates produce incorrect rankings for the good and bad teachers and is maximized at 1 when the rankings of those two teachers match their true rankings.

For each factor combination, the results of VIM_{Ψ} for each experiment were expressed as the average

$$r(\Psi) = \frac{1}{R}\sum_{j=1}^{R} r_j(\Psi), \tag{3.9}$$

where $R = 100$ is the number of replicates for each factor combination in the study, $r_j(\Psi)$ is the respective rank based comparison measure between the absolute value of the *true teacher effects* and VIM_{Ψ} for the results obtained in replicate j. The averages $r(\ell m), r(\Upsilon), r(1)$, and $r(2)$ are similarly defined.

We determined by means of paired t-tests when the correlation of the proposed measures were statistically better or at least no worse than the correlations for the linear model VIMs (or association in the case of the *good teacher - bad teacher* model). The associated confidence intervals with 95% confidence levels are used in the description of the results.

Results

We now present selected results from the simulation study for the CAM and GSM models. For several factor combinations, when the linear model did not account for complex interactions, VIM_{Ψ} and VIM_{Υ} significantly outperformed $VIM_{\ell m}$, VIM_1, and VIM_2. On the other hand, when the linear model was correctly specified or was misspecified to the extent described by the simple interaction model, $VIM_{\ell m}$ outperformed all the other measures. In this section, we summarize the most relevant findings in the simulation study for each factor combination. This includes experimental settings where our proposed VIMs significantly outperformed $VIM_{\ell m}$ as well as others where they did not. In Section 3.3 we will center our attention on the former and provide the rationale for those results.

Graphical representations will be used to allow us to observe the patterns across different experimental conditions. In each figure that follows we plot the mean correlation for the five studied measures (y-axis). Based on the results we obtained, the following

discussion will be directed toward comparisons among $VIM_{\ell m}$, VIM_{Ψ}, and VIM_1, unless otherwise noted. We use these three measures because VIM_{Ψ} generally outperformed VIM_{Υ}, and VIM_1 outperformed VIM_2.

Four different scenarios were considered for the CAM, each of which will represent a row of graphs within a subsequent figure. The scenarios under consideration include CAM_1 as the baseline model, CAM_2 as the *good teacher-bad teacher* model, CAM_3 as a simple interaction model, and CAM_4 as a complex interaction model.

Figure 3.3 plots the mean correlation for all five VIMs across the number of teachers (x-axis) when 12 students per teacher (left column graphs) or 24 students per teacher (right column graphs) are considered and the ratio of teacher variance over student variance is 2.

For the CAM baseline model, CAM_1, when the number of students per teacher was 12, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. As the number of students per teacher increased, the random forest VIMs tended to improve in performance, although $VIM_{\ell m}$ still outperformed the rest. A similar trend was obtained when the number of students per teacher was 24. In particular, when the number of teachers was 40 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, 0.90 and 0.92 for $VIM_{\ell m}$, 0.80 and 0.87 for VIM_1, and 0.73 and 0.73 for VIM_{Ψ}. The 95% confidence intervals of the difference between the correlations of $VIM_{\ell m}$ and VIM_{Ψ} were (.151, .184) and (.070, .094), respectively. When the number of teachers was 100 and the number of students per teacher was 12 and 24, the mean correlations increased slightly, respectively, to .94 and .96 ($VIM_{\ell m}$), .84 and .91 (VIM_1), and .80 and .89 (VIM_{Ψ}). The 95% confidence intervals of the difference between the correlations of $VIM_{\ell m}$ and VIM_{Ψ} were (.133, .147) and (.067, .077), respectively. When the number of students per teacher was 24, all the

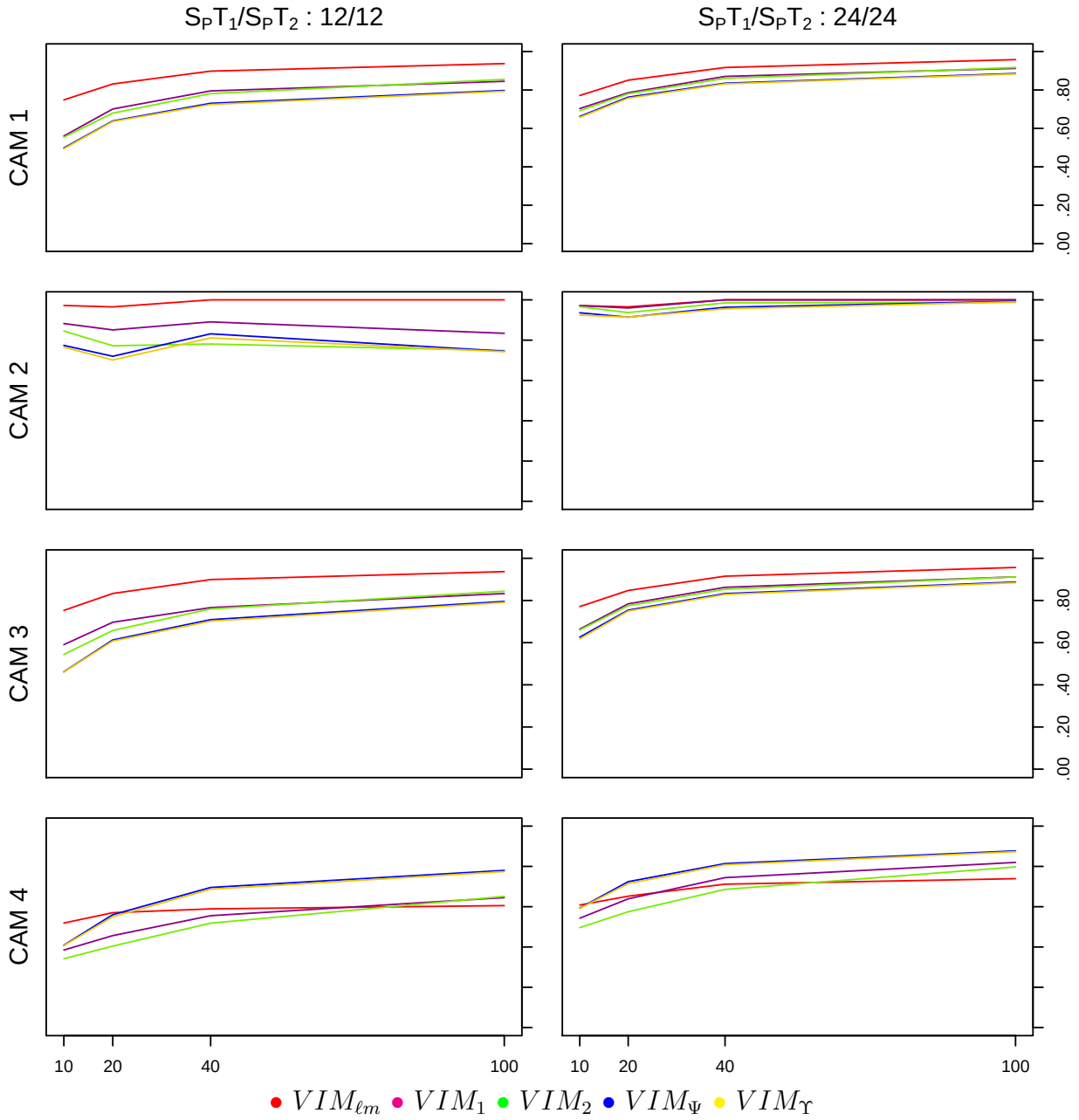

Figure 3.3: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different CAM models, the number of students per teacher is 12 (left column) or 24 (right column), and $\sigma_\tau^2/\sigma^2 = 2$.

VIMs based on random forest improved and reduced appreciably the gap towards $VIM_{\ell m}$ results in comparison to the case with 12 students per teacher.

For the CAM *good teacher - bad teacher model*, CAM_2, when the number of students per teacher was 12, $VIM_{\ell m}$ significantly outperformed the random forest measures for any number of teachers. Among the data mining VIM measures, VIM_1 performed better. The minimum mean association among all measures was .72 (VIM_Ψ) for 20 teachers and 12 students per teacher. In the case of 24 students, all measures yielded high mean association. The random forest measures, VIM_1 in particular, was not significantly worse than $VIM_{\ell m}$ for any number of teachers. In particular, when the number of teachers was 100, none of the VIM measures were significantly worse than $VIM_{\ell m}$. For example, with a 95% confidence level, the interval for the difference between $VIM_{\ell m}$ and VIM_Ψ association was (-.001, .016).

For the CAM *simple interactions model*, CAM_3, the results were similar to the CAM_1 results. For both scenarios, 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. However, as the number of students per teacher increased from 12 to 24, the random forest VIMs tended to improve in performance and the difference in performance between $VIM_{\ell m}$ and the random forest VIMs was reduced. Specifically, when the number of teachers was 40 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, 0.90 and 0.92 for $VIM_{\ell m}$, 0.77 and 0.87 for VIM_1, and 0.71 and 0.83 for VIM_Ψ. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference were (.172, .208) and (.070, .095), respectively. When the number of teachers was 100 and the number of students per teacher was 12 and 24, the mean correlations slightly increased, respectively, to .94 and .97 ($VIM_{\ell m}$), .83 and .91 (VIM_1), and .80 and .89 (VIM_Ψ). The 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference were (.133, .149) and (.064, .074), respectively.

For the CAM *complex interactions* model, CAM_4, Figure 3.3 shows that VIM_{Ψ} and VIM_{Υ} significantly outperformed the remaining measures, including $VIM_{\ell m}$, in nearly all cases. $VIM_{\ell m}$ significantly outperformed the proposed measures only for the case with 10 teachers and 12 students per teacher. In this scenario, the sample mean correlations were .52 for $VIM_{\ell m}$ and .41 for VIM_{Ψ} and the 95% confidence intervals of $VIM_{\ell m}$ and VIM_{Ψ} correlations difference was (.044, .176). For the cases with 10 teachers and 24 students per teacher, and 20 teachers and 12 students per teacher, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. In the remaining scenarios; namely, 20 teachers and 24 students per teacher, and 40 and 100 teachers with either 12 and 24 students per teacher, VIM_{Ψ} and VIM_{Υ} significantly outperformed $VIM_{\ell m}$ and any other VIM measure. As case in point, with 20 teachers and 24 students, the sample mean correlations were 0.72 for VIM_{Ψ} and 0.65 for $VIM_{\ell m}$ and the 95% confidence interval for the Spearman's correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ was (.041, .103). With 40 teachers and 12 or 24 students per teacher the mean correlations were, respectively, 0.70 or 0.81 for VIM_{Ψ} and 0.59 or 0.71 for $VIM_{\ell m}$. The corresponding 95% confidence intervals for the correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ were (.082, .130) and (.085, .119), respectively. With 100 teachers and 12 or 24 students per teacher the mean correlations were, respectively, 0.78 or 0.88 for VIM_{Ψ} and 0.61 or 0.74 for $VIM_{\ell m}$. The corresponding 95% confidence intervals for the correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ were (.160, .189) and (.126, .150), respectively. When comparing the results that were obtained with our proposed measures and other random forest VIMs, VIM_1 was not significantly worse than VIM_{Ψ} only in the case with 10 teachers and 12 students per teacher. In every other scenario, VIM_{Ψ} and VIM_{Υ} outperformed significantly the other two random forest measures, VIM_1 and VIM_2. Finally, notice that an increase in the number of students per teacher from 12 to 24 produced higher correlations for every VIM studied.

Figure 3.4 plots the mean correlation for the five measures across the number of teachers (x-axis) when 12 students per teacher (left column graphs) or 24 students per

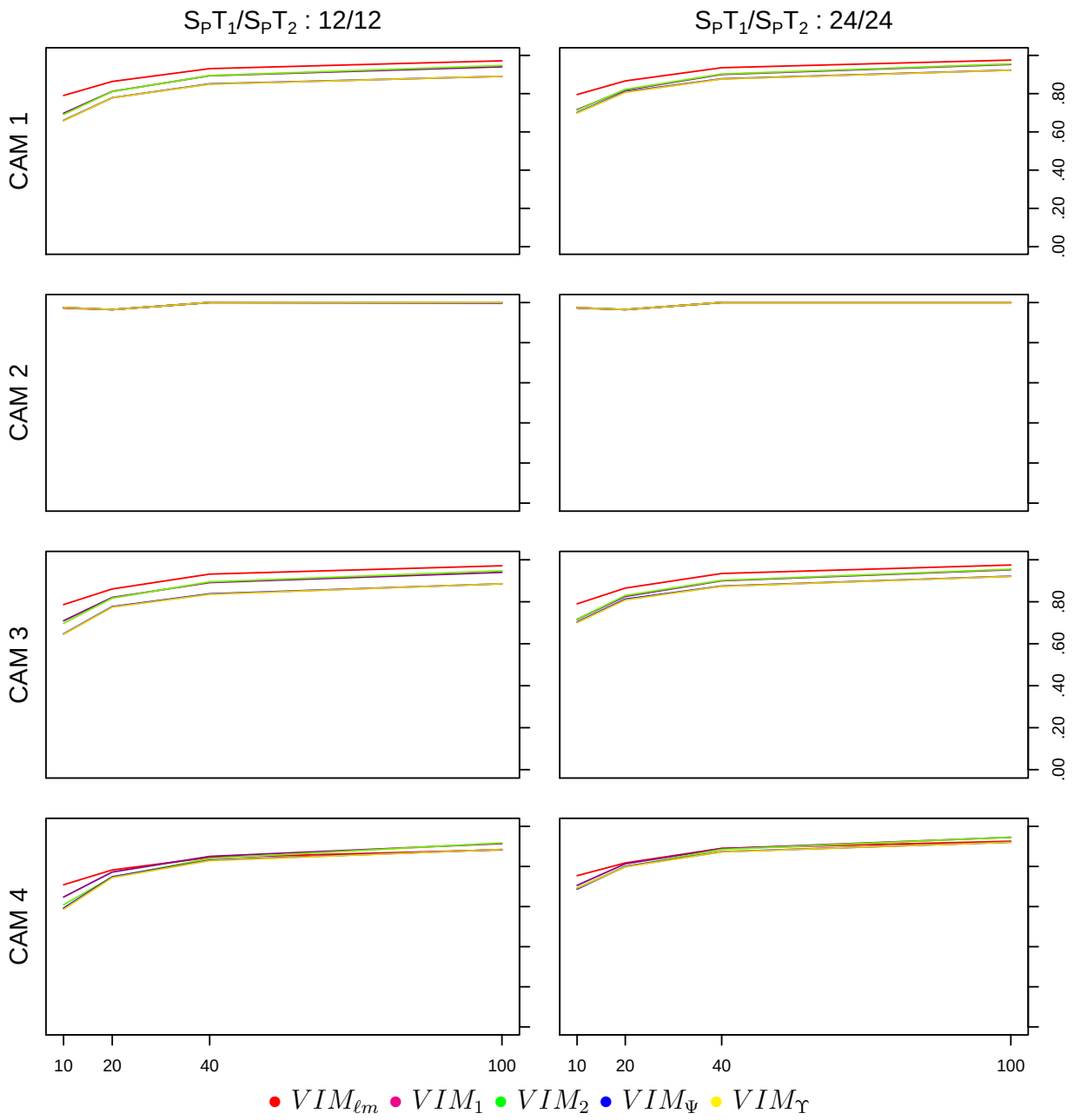

Figure 3.4: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different CAM models, , the number of students per teacher is 12 (left column) or 24 (right column), and $\sigma^2_\tau/\sigma^2 = 5$.

teacher (right column graphs) are considered and the ratio of teacher variance over student variance is 5.

For CAM_1, in scenarios with 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers.

However, when the number of students per teacher increased from 12 to 24, all the VIMs based on random forest improved and reduced the gap with $VIM_{\ell m}$ correlation results. To illustrate this, we mention that when the number of teachers was 40 and the number of students per teacher was 12 or 24, the mean correlations were, respectively, 0.93 or 0.94 for $VIM_{\ell m}$, 0.89 or 0.90 for VIM_1, and 0.85 or 0.88 for VIM_Ψ. The 95% confidence intervals for the difference of $VIM_{\ell m}$ and VIM_Ψ correlations were (.064, .096) or (.041, .075), respectively.

For CAM_2, for every combination of number of teacher and students per teacher, $VIM_{\ell m}$ was not significantly better or worse than any of the random forest VIMs. This happened because all measures adequately identified the *good* and *bad* teacher effects and placed them at the top of the ranking in almost every replicate, obtaining association measures approaching 1 when the number of teachers was 10 or 20, and equal to 1 when the number of teachers was 40 or 100.

For CAM_3, the results were similar to the CAM_1 results. For both scenarios, 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. As the number of students per teacher increased from 12 to 24, the random forest VIMs tended to improve slightly in performance. For example, when the number of teachers was 40 and the number of students per teacher was 12 or 24, the mean correlations were, respectively, 0.932 or 0.934 for $VIM_{\ell m}$ and 0.875 or 0.886 for VIM_Ψ. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference were (.079, .109) and (.044, .076), respectively.

The CAM_4 results in Figure 3.4 for a teacher/student variance ratio of 5 were quite different than those in Figure 3.3 when the ratio of teacher variance over student variance was 2. In this case, the *node proportion*, VIM_Ψ, and *covariate proportion*, VIM_Υ, still performed well; however, they were significantly outperformed by $VIM_{\ell m}$, in cases with 10 teachers with 12 or 24 students per teacher, 20 teachers with 12 students

per teacher, or 40 teachers with 24 students per teacher. In addition, they no longer differed significantly from $VIM_{\ell m}$ in the remaining cases that were studied. On the other hand, VIM_1 was not significantly worse than $VIM_{\ell m}$ for cases with 20 or 40 teachers and any number of students per teacher, and significantly outperformed $VIM_{\ell m}$ in those cases with 100 teachers and 12 or 24 students per teacher. For the latter case, the 95% confidence intervals for the Spearman's correlation difference of VIM_1 and $VIM_{\ell m}$ were (.024, .038) and (.011, .028), respectively. In addition, observe that an increase in the number of teacher or the number of students per teacher produced higher correlations for every VIM studied. Although we will not discuss it in detail here, we merely remark that similar conclusions can be drawn for conditions in which the ratio of teacher variance over student variance was 20. The corresponding figures are presented in Appendix A.

Figure 3.5 plots the mean correlation/association for the five measures across the ratios of the number of students per teacher in group 1 to those in group 2 (x-axis) when teacher variances range from 1 to 20 (columns) with 40 teachers. In conditions with fewer teachers (10 and 20, respectively) the results tended to follow a similar pattern, although correlations/associations across conditions for all measures tended to be lower. Analogous plots for the 10 and 20 teachers cases are provided in Appendix A.

Not surprisingly, for CAM_1, $VIM_{\ell m}$ outperformed significantly the random forest VIMs for all values of teacher variance. As seen in Figure 3.5, as teacher variance increased, the gap between $VIM_{\ell m}$ and the random forest measures decreased. The lowest correlations were found in conditions with group 1 to group 2 ratio equal to 36 to 12, for all measures. An illustration of the typical results is provided by the case where the group 1 to group 2 ratio was 30 to 18. In that instance the sample mean correlations were .94 for $VIM_{\ell m}$ and .83 for VIM_{Ψ} and the 95% confidence intervals for the difference of $VIM_{\ell m}$ and VIM_{Ψ} correlations was (.088, .140).

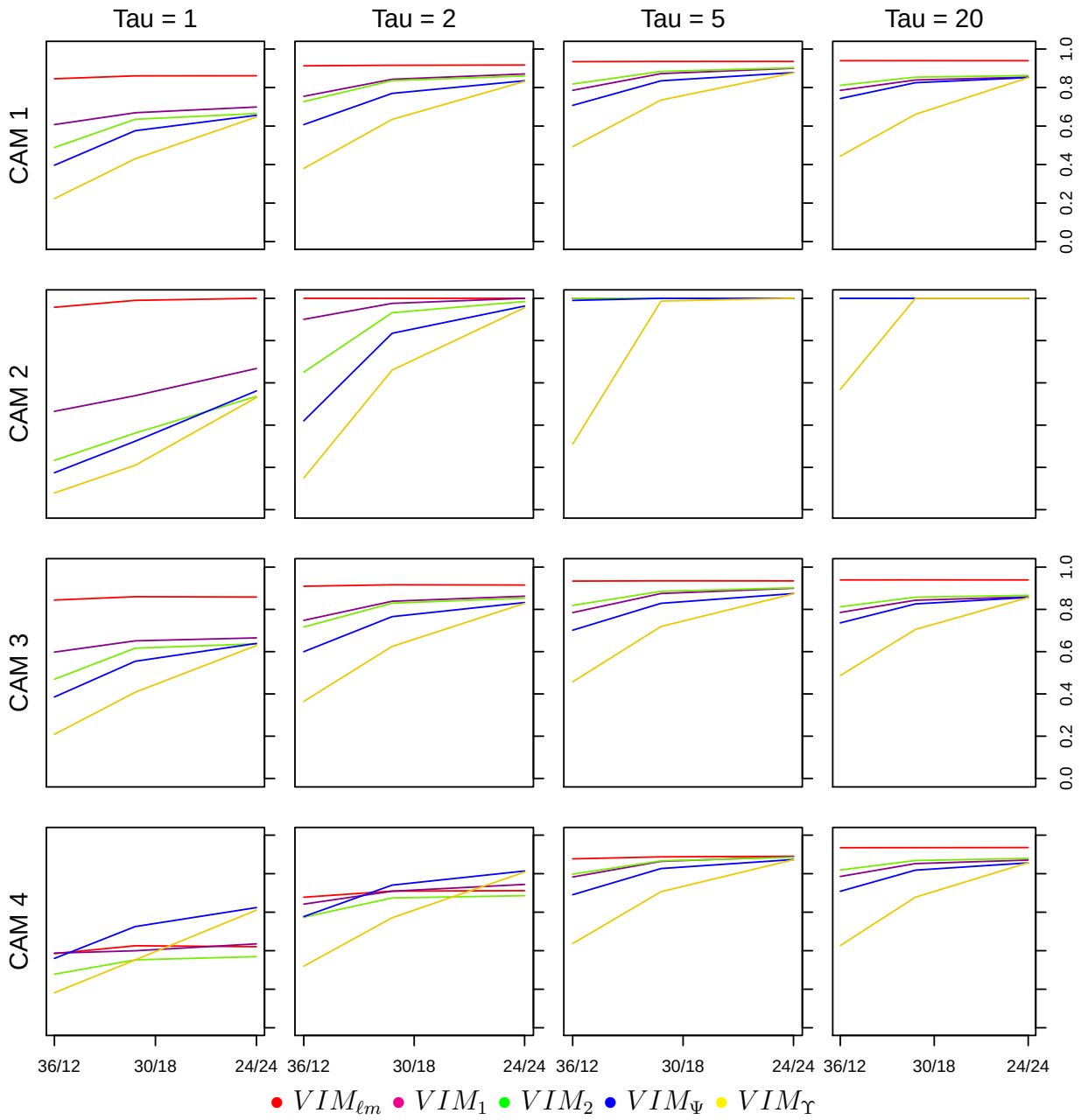

Figure 3.5: Mean correlation/association between the VIMs and the absolute value of true teacher effects when SpT_1/SpT_2 varies for different CAM models and different σ^2_τ/σ^2, when the number of teachers is 40.

In CAM_2, when teacher variances were 1 or 2, $VIM_{\ell m}$ significantly outperformed the random forest methods in every case. When the teacher variance was 1 and the number of students per teacher in group 1 to group 2 ratio were 30 to 18 or 36 to 12, the mean associations were, respectively, 0.99 or 0.96 for $VIM_{\ell m}$ and 0.32 or 0.17 for VIM_{Ψ}. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_{Ψ} association differences were (.610, .723) and (.745, .820), respectively. By contrast, when the teacher variances were 5 or 20, the mean association in most cases was 1 or close to 1 and $VIM_{\ell m}$ were not significantly better than any random forest VIMs.

For CAM_3, Figure 3.5 suggests that results for simple interactions resemble those for the baseline model that was previously discussed. We note here as well that, regardless of the group 1 to group 2 ratio, $VIM_{\ell m}$ outperformed significantly all random forest methods. The performance gap between $VIM_{\ell m}$ and the random forest measures narrowed as the teacher variance increased and the number of students per teacher in group 1 to group 2 ratio became balanced.

For CAM_4, when teacher variance was 1, VIM_{Ψ} significantly outperformed $VIM_{\ell m}$ in conditions where the number of students per teacher in group 1 to group 2 ratios were 30 to 12 or 24 to 24 (balanced case). Specifically, the mean correlations were 0.53 or 0.62 for VIM_{Ψ} and 0.43 or 0.42 for $VIM_{\ell m}$, respectively. The corresponding 95% confidence intervals for the correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ were (.065, .134) and (.176, .230), respectively. When the number of students per teacher in group 1 to group 2 ratio was 36 to 12, VIM_{Ψ} was not significantly better than $VIM_{\ell m}$. Similar conclusions were obtained when the teacher variance was 2; that is, VIM_{Ψ} significantly outperformed $VIM_{\ell m}$ when the number of students per teacher in group 1 to group 2 ratios were 30 to 12 or 24 to 24 although $VIM_{\ell m}$ outperformed signifcantly VIM_{Ψ} when the number of students per teacher in group 1 to group 2 ratio was 36 to 12. When the teacher variance was 5 or 20, $VIM_{\ell m}$ outperformed signifcantly VIM_{Ψ} for any combination of the number of students per teacher in group 1 to group 2 ratio.

Figure 3.6 plots the mean correlation/association for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers (columns) when the number of students per teacher in group 1 to group 2 ratio is $\frac{24}{24}$ (balanced case). Due to the similarity of results, associated graphs for $\frac{12}{12}$ and $\frac{36}{36}$ have been relegated to Appendix A.

As shown in Figure 3.6 for CAM_1, $VIM_{\ell m}$ yielded significantly higher correlations than random forest VIMs in every case that was studied. As the teacher variance increased, the random forest measures closed the gap with $VIM_{\ell m}$; however, in the case of 40 teachers and teacher variance equal to 20, the random forest mean correlations were slightly smaller than in the case of teacher variance equal to 5. Overall, the highest correlations for all measures tended to be found in conditions with 40 teachers and teacher variance of 5. Although not shown in Figure 3.6 the results for cases with 100 teacher were analogous to those with 40 teachers. The results for CAM_3 were similar to those for CAM_1.

For CAM_2, when the teacher variance was 5 or 20, VIM_Ψ was not significantly worse than $VIM_{\ell m}$ for each case that was considered. On the other hand, when the teacher variance was 1 or 2, $VIM_{\ell m}$ outperformed significantly VIM_Ψ in each instance. An illustration of this is when the teacher variance was 1. A 95% confidence interval for the difference in association measures for $VIM_{\ell m}$ and VIM_Ψ was (.295,0.408) in that case

For CAM_4, Figure 3.6 shows that for those cases with teacher variance equal to 1 or 2, VIM_Ψ was not significantly worse or was significantly better than $VIM_{\ell m}$ and VIM_1. This was also the case when the number of teachers was 20 or 100 and teacher variance is 5. In contrast, for the remaining cases with teacher variances of 5 or 20, $VIM_{\ell m}$ was significantly better than VIM_Ψ. All the measures' performances tended to improve as the number of teachers increased. In general, that was also the trend with an

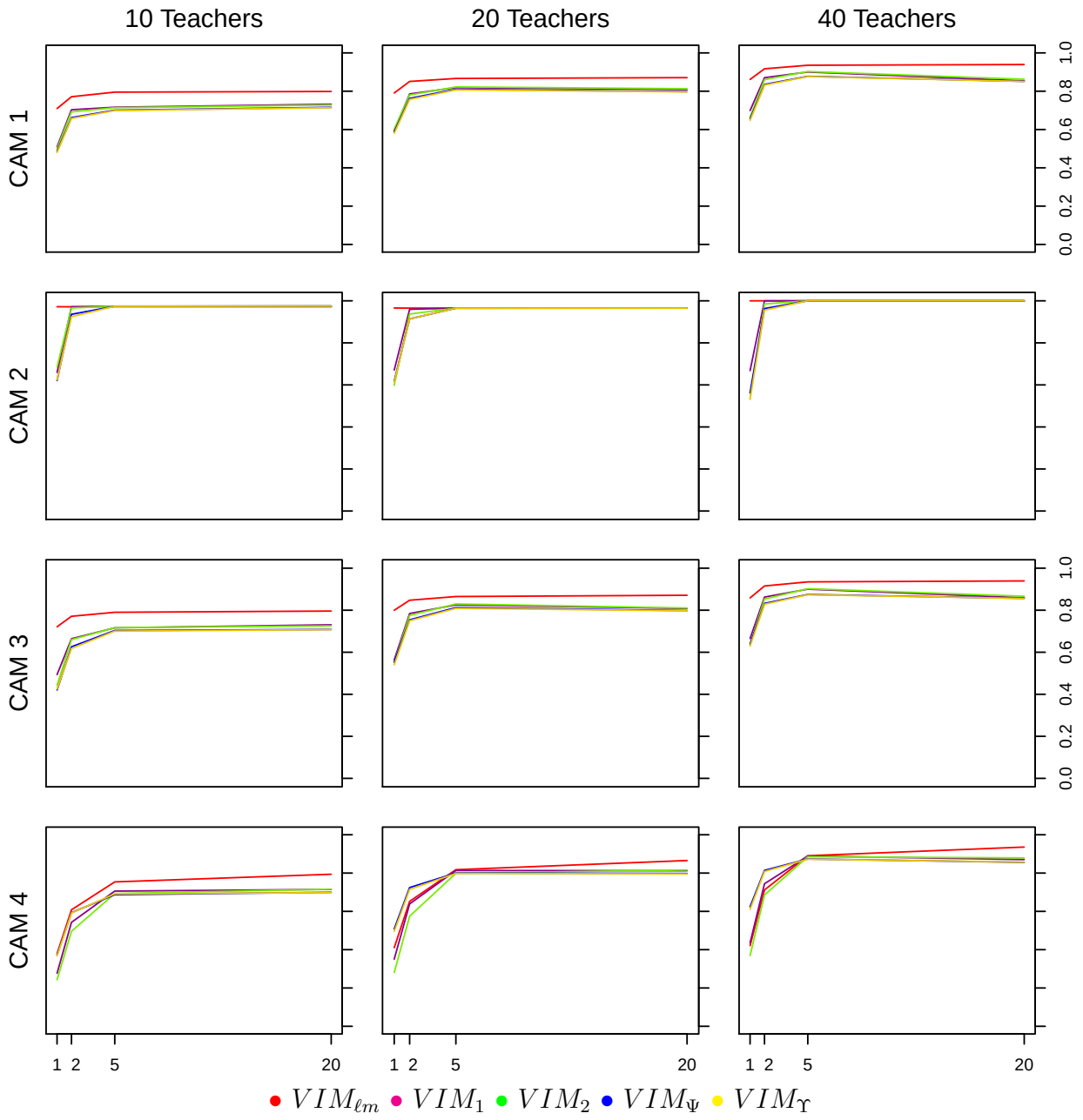

Figure 3.6: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teacher in group 1 to group 2 ratio is 24/24.

increase in teacher variance except for the random forest VIM correlations when the teacher variance increased from 5 to 20.

Figure 3.7 summarizes the mean correlation for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers (columns) when the number of students per teacher in group 1 to group 2 ratio is the unbalanced case of $\frac{36}{12}$. For the CAM_1 and CAM_3 cases, $VIM_{\ell m}$ outperformed significantly VIM_{Ψ} in every case. For CAM_2, when the teacher variance was 1, 2, or 5, $VIM_{\ell m}$ outperformed significantly VIM_{Ψ} for any number of teachers. We also considered when the number of teachers was 100 although it is not shown in the figure. In that instance when the teacher variance was 5, we found that VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. Similarly, when the teacher variance was 20, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. For example, with 20 teachers and teacher variance 20, the mean association for $VIM_{\ell m}$ was .965 and for VIM_{Ψ} was .966 and the 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} association difference was (-.003, .001). For CAM_4, when the teacher variance was 2, 5, or 20, $VIM_{\ell m}$ outperformed significantly VIM_{Ψ} for cases with 10, 20, or 40 teachers. When the number of teachers was 100 and the teacher variance was 2, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. Similarly, when the teacher variance was 1 and the number of teachers 40 or 100, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. As a specific instance, with 40 teachers and teacher variance of 1, the mean correlations for $VIM_{\ell m}$ was .385 and for VIM_{Ψ} was .360 and the 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} correlation difference was (-.010, .061).

Figure 3.7 also shows that in this unbalanced case, for all four CAM specifications, the covariate proportion VIM, VIM_{Υ}, obtained consistently low correlations/associations. This is due to the fact that, given the construction of the measure in (3.6), it is more sensitive to unbalanced situations. Since the unbalancedness is extreme in this case, the *covariate proportion* VIM loses the ability to adequately obtain appropriate rankings.

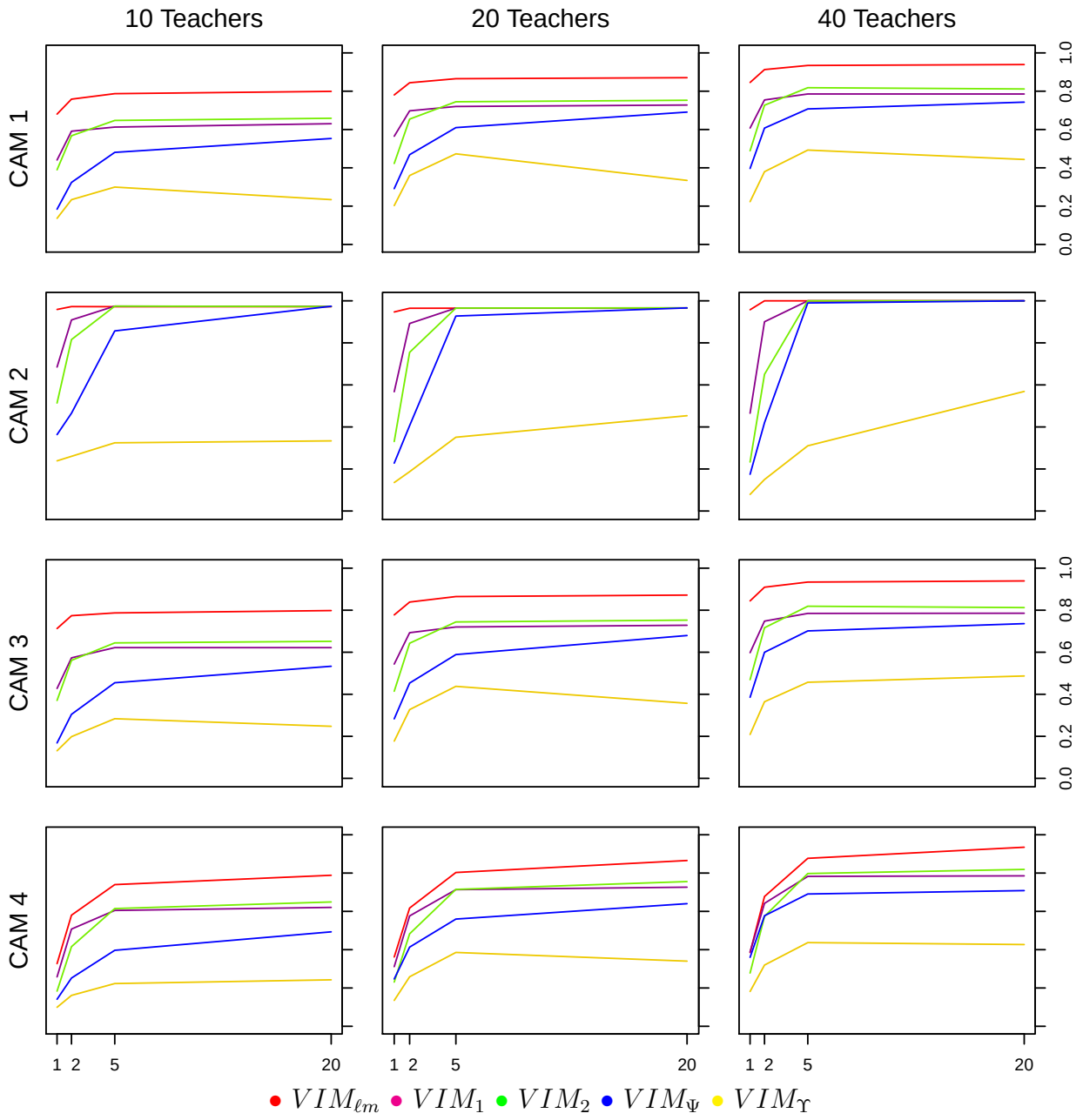

Figure 3.7: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teacher in group 1 to group 2 ratio is 36/12.

Figure 3.8 is a plot of the mean correlations/associations for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers (columns) when the number of students per teacher is the unbalanced $\frac{30}{18}$ case. For CAM_1 and CAM_3, $VIM_{\ell m}$ outperformed significantly VIM_{Ψ} in every case. For CAM_2, when the teacher variance was 1 or 2, $VIM_{\ell m}$ outperformed significantly VIM_{Ψ} for any number of teachers. When the teacher variance was 5 or 20, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. For example, with 10 teachers and teacher variance 5, the mean association for $VIM_{\ell m}$ was .972 and for VIM_{Ψ} was .975. The 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} association difference was (-.007, .002). For CAM_4, when teacher variance was 5 or 20 for any number of teachers or when variance was 1 or 2 for 10 and 20 teachers, $VIM_{\ell m}$ significantly outperformed VIM_{Ψ}. However, when the number of teachers was 40 or 100 and the teacher variance was 1 or 2, it was VIM_{Ψ} that significantly outperformed $VIM_{\ell m}$. In particular, with 40 teachers and a teacher variance of 2, the mean correlations for $VIM_{\ell m}$ was .709 and for VIM_{Ψ} was .741. The 95% confidence interval for VIM_{Ψ} and $VIM_{\ell m}$ correlation difference was (.015, .047). In general, for all CAM scenarios, the mean correlation/association differences between the data mining methods and $VIM_{\ell m}$ presented in Figure 3.8 were smaller when the imbalance of the number of students per teacher was $\frac{30}{18}$ than for $\frac{36}{12}$. Because the unbalanced case described in Figure 3.8 is not as extreme as the one described in Figure 3.7, the *covariate proportion* VIM, VIM_{Υ}, obtained somewhat low correlations/associations, but not as low as in the previous case.

Let us now discuss our findings for the GAM simulation. For this case, four different scenarios are also considered: GSM_1 as the baseline model, GSM_2 as the *good teacher-bad teacher* model, GSM_3 as a simple interaction model, and GSM_4 as a complex interaction model.

Figure 3.9 plots the mean correlation/association for all five VIMs across the number of teachers (x-axis) when 12 students per teacher (left column graphs) or 24

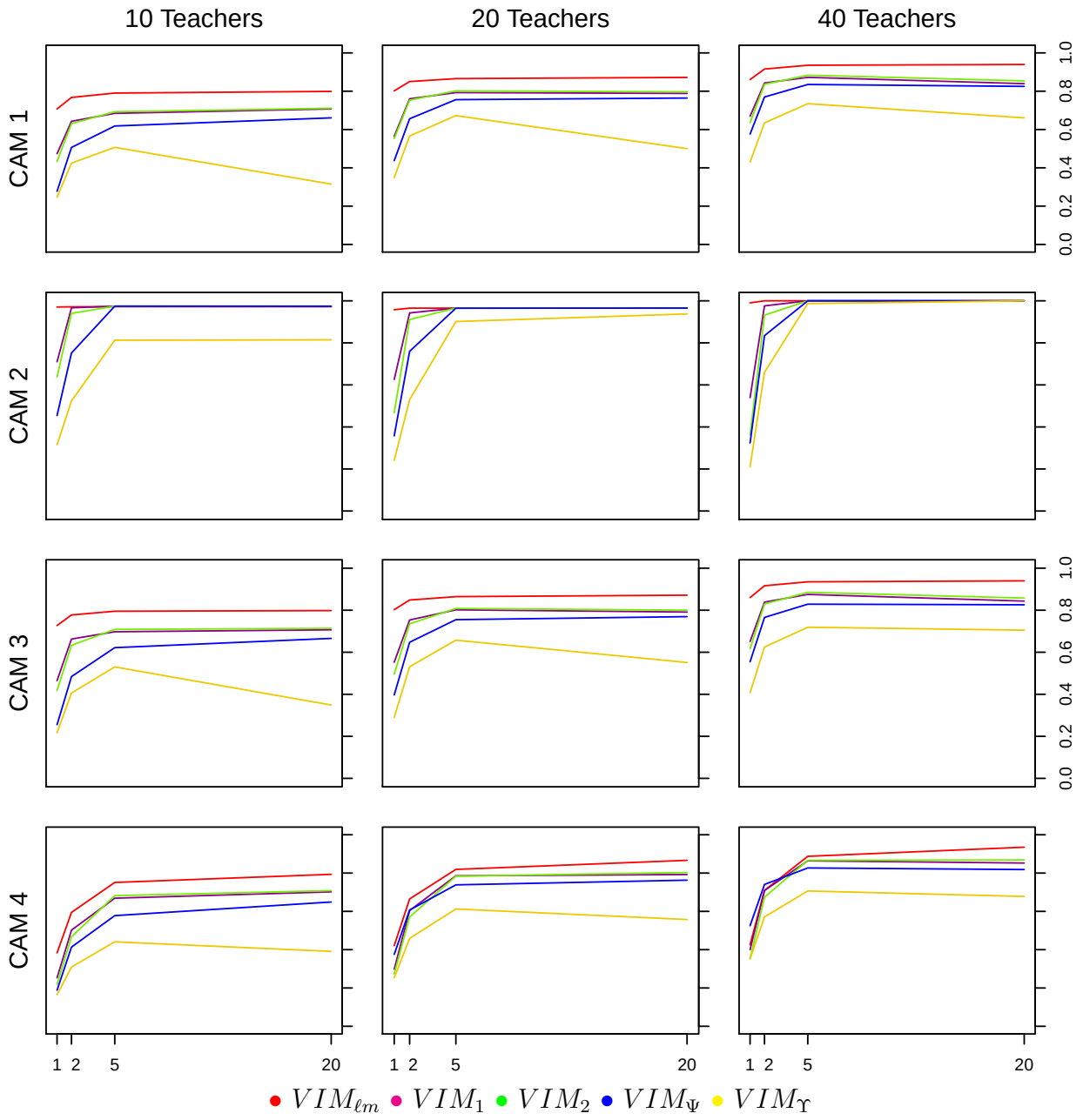

Figure 3.8: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teachers in group 1 to group 2 ratio is 30/18.

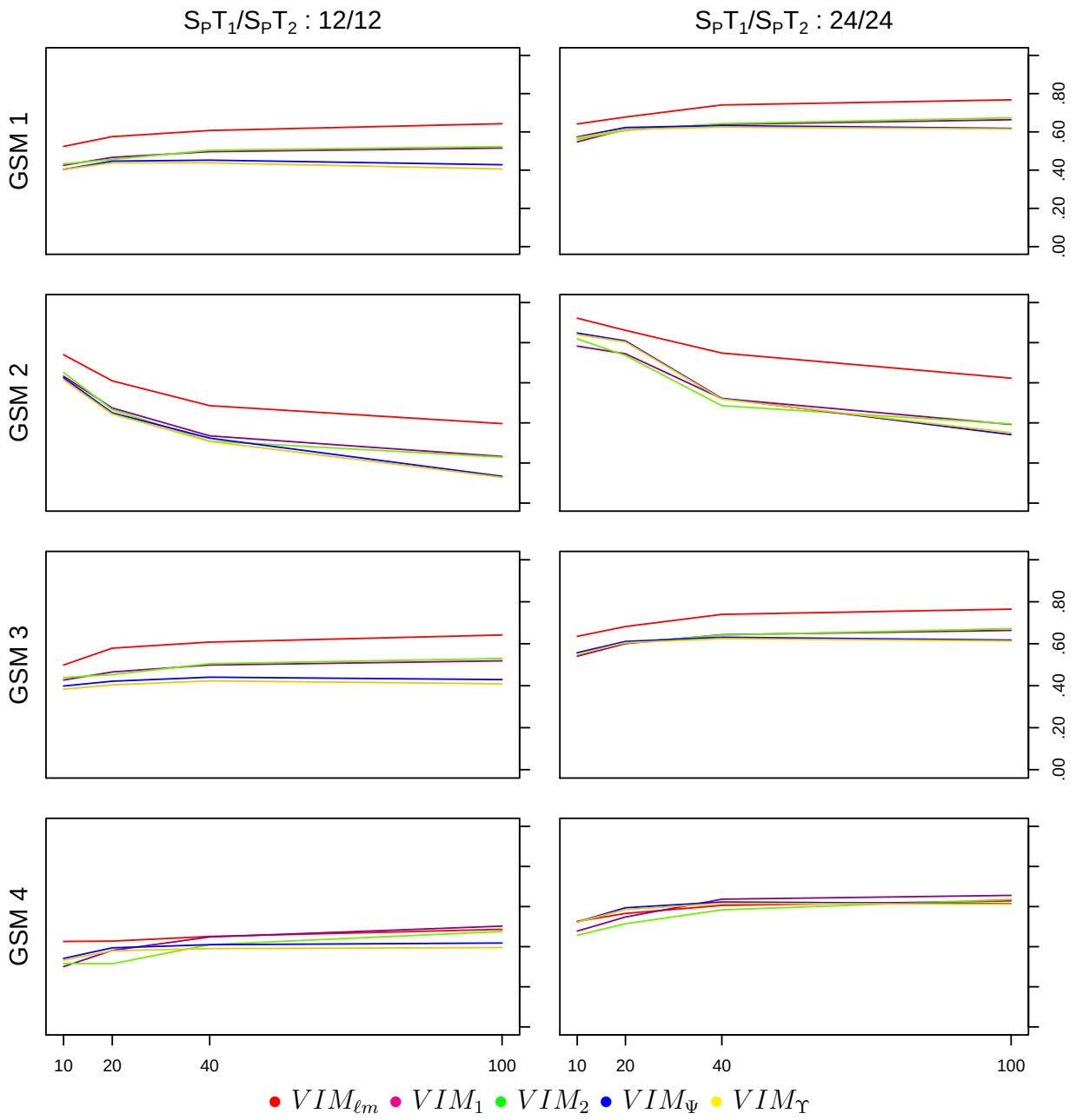

Figure 3.9: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different GSM models, the number of students per teacher is 12 (left column) or 24 (right column), and $\sigma^2_\tau/\sigma^2 = 2$.

students per teacher (right column graphs) are considered and the ratio of teacher variance over student variance is 2.

For the GSM baseline model, GSM_1, when the number of students per teacher was 12, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any

number of teachers. When the number of students per teacher increased, VIM_1 tended to improve in performance, while VIM_Ψ did not. $VIM_{\ell m}$ still significantly outperformed the rest in each case. A similar trend was obtained when the number of students per teacher was 24. For instance, when the number of teachers was 40 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, 0.61 and 0.74 for $VIM_{\ell m}$, 0.50 and 0.64 for VIM_1, and 0.45 and 0.63 for VIM_Ψ. The 95% confidence intervals of the difference between the correlations of $VIM_{\ell m}$ and VIM_Ψ were (.133, .178) and (.092, .124), respectively. When the number of teachers was 100 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, .64 and .77 ($VIM_{\ell m}$), .52 and .66 (VIM_1), and .43 and .62 (VIM_Ψ), and the 95% confidence intervals for the difference between the correlations of $VIM_{\ell m}$ and VIM_Ψ were (.199, .231) and (.140, .159), respectively. When the number of students per teacher was 24, all the VIMs based on random forest improved and reduced slightly the gap towards $VIM_{\ell m}$ results in comparison to the case with 12 students per teacher.

For the GSM *good teacher - bad teacher model*, GSM_2, when the number of teachers increased, the mean association measure values were smaller for all VIMs. When the number of students per teacher was 12, $VIM_{\ell m}$ significantly outperformed the random forest measures for any number of teachers. Among the data mining VIM measures, VIM_1 and VIM_2 performed better than the proposed measures. In the case of 24 students, all measures yielded high mean association. As an illustration, when the number of teachers was 100, the 95% confidence interval for the difference between $VIM_{\ell m}$ and VIM_Ψ correlations was (0.226, 0.337).

For the GSM *simple interactions model*, GSM_3, the results were similar to the GSM_1 results. For both scenarios, 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. However, as the number of students per teacher increased from 12 to 24, the random forest VIMs tended to improve in performance and the difference in performance between

$VIM_{\ell m}$ and the random forest VIMs was reduced. Specifically, when the number of teachers was 40 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, 0.61 and 0.74 for $VIM_{\ell m}$, 0.50 and 0.64 for VIM_1, and 0.44 and 0.63 for VIM_Ψ. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference were (.143, .191) and (.093, .125), respectively. When the number of teachers was 100 and the number of students per teacher was 12 and 24, the mean correlations were, respectively, .64 and .76 ($VIM_{\ell m}$), .52 and .66 (VIM_1), and .43 and .62 (VIM_Ψ). The 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference were (.197, .228) and (.138, .157), respectively.

For the GSM *complex interactions* model, GSM_4, Figure 3.9 shows that, when the number of teachers was 10 or 20, VIM_Ψ and VIM_Υ significantly outperformed the other random forest measures and were not significantly worse than $VIM_{\ell m}$. $VIM_{\ell m}$ significantly outperformed the proposed measures with 10 teachers and 12 students per teacher. In this scenario, the sample mean correlations were .43 for $VIM_{\ell m}$ and .34 for VIM_Ψ and the 95% confidence intervals of $VIM_{\ell m}$ and VIM_Ψ correlations difference was (.013, .156). For the cases with 10 teachers and 24 students per teacher, 20 teachers and 12 or 24 students per teacher, and 40 teachers and 24 students per teacher, VIM_Ψ was not significantly worse than $VIM_{\ell m}$. Specifically, with 10 teachers and 24 students per teacher, the sample mean correlations were 0.52 for VIM_Ψ and 0.53 for $VIM_{\ell m}$ and the 95% confidence interval for the Spearman's correlation difference of VIM_Ψ and $VIM_{\ell m}$ was (-.052, .044). With 20 teachers and 12 or 24 students per teacher the mean correlations were, respectively, 0.39 or 0.59 for VIM_Ψ and 0.43 or 0.57 for $VIM_{\ell m}$, respectively. The corresponding 95% confidence intervals for the correlation difference of VIM_Ψ and $VIM_{\ell m}$ were (-.081, .015) and (-.005, .062), respectively. With 40 teachers and 24 students per teacher the mean correlations were 0.62 for VIM_Ψ and 0.61 for $VIM_{\ell m}$ and the 95% confidence interval for the correlation difference of VIM_Ψ and $VIM_{\ell m}$ was (-.004, .037). When the number of teachers was 40 or 100, VIM_Ψ and VIM_Υ

did not outperform significantly the other two random forest measures, VIM_1 and VIM_2. An increase of the number of students per teacher from 12 to 24 produced slightly higher correlations for every VIM in the study.

Figure 3.10 plots the mean correlation/association for the five measures across the number of teachers (x-axis) when 12 students per teacher (left column graphs) or 24 students per teacher (right column graphs) are considered and the ratio of teacher variance over student variance is 20. Results for conditions where teacher variance over student variance is 5 are similar and are included in the Appendix A.

For GSM_1, in scenarios with 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. When the number of students per teacher increased from 12 to 24, little change was observed in the VIMs based on random forest. An illustration of this is provided by the case where the number of teachers was 40 and the number of students per teacher was 12 or 24. There, mean correlations were found to be 0.932 or 0.937 for $VIM_{\ell m}$, 0.847 or 0.849 for VIM_1, and 0.836 or 0.834 for VIM_Ψ. The 95% confidence intervals for the difference of $VIM_{\ell m}$ and VIM_Ψ correlations were (.071, .123) or (.077, .130), respectively.

For GSM_2, for every combination of number of teacher and students per teacher, $VIM_{\ell m}$ was not significantly better or worse than any of the random forest VIMs. Again, this occured because all measures adequately identified the *good* and *bad* teacher effects and placed them at the top of the ranking. This transpired in almost every replicate, thereby producing association measures approaching 1 when the number of teachers was 10 or 20, and equal to 1 when the number of teachers was 40 or 100.

For GSM_3, the results were similar to the CAM_1 results. For both scenarios, 12 and 24 students per teacher, $VIM_{\ell m}$ yielded significantly higher correlations than the remaining measures for any number of teachers. As the number of students per teacher increased from 12 to 24, the random forest VIMs tended to improve slightly in

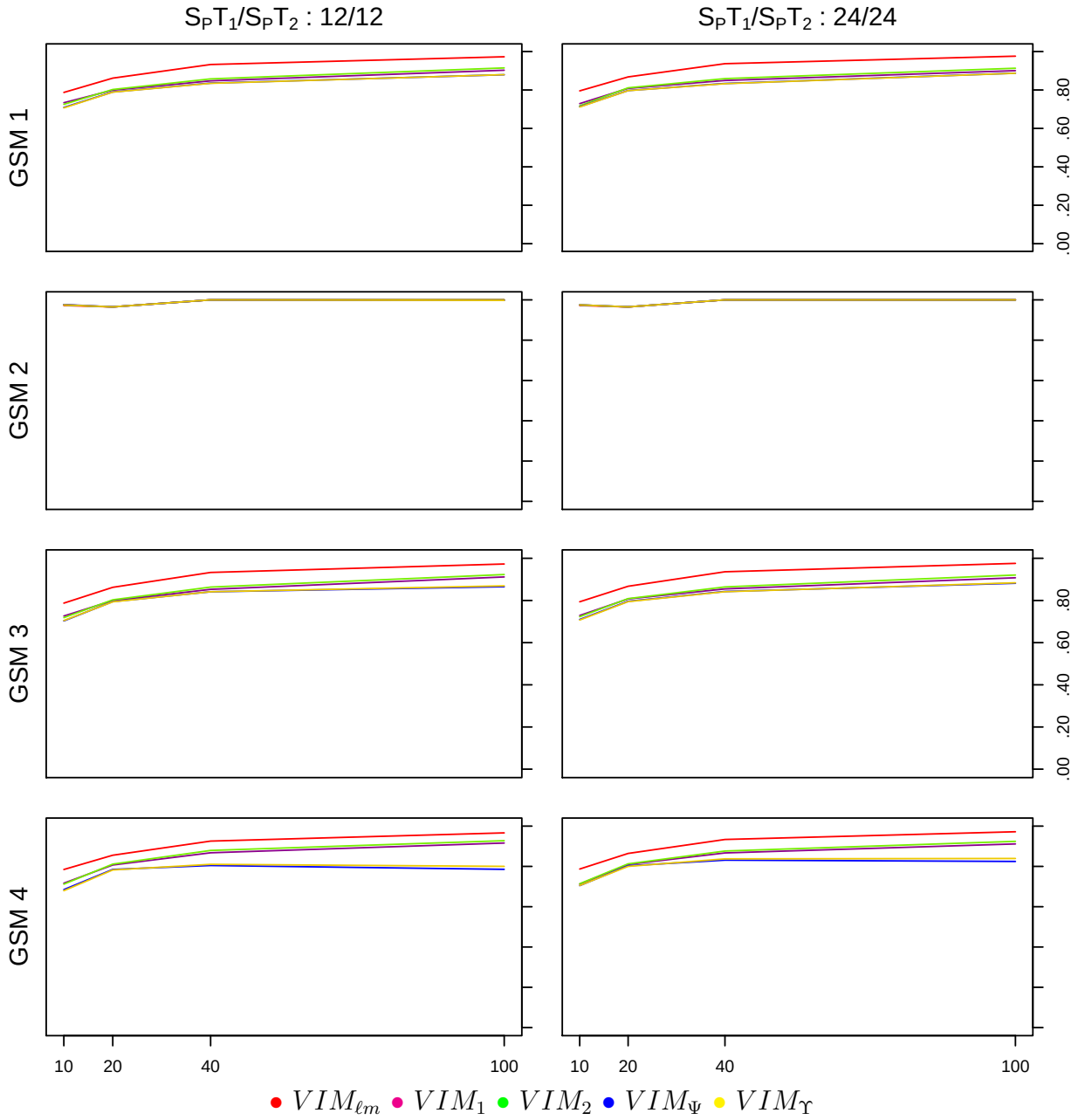

Figure 3.10: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different GSM models, the number of students per teacher is 12 (left column) or 24 (right column), and $\sigma_\tau^2/\sigma^2 = 20$.

performance. In the specific case where the number of teachers was 40 and the number of students per teacher was 12 or 24, the mean correlations were, respectively, 0.933 or 0.936 for $VIM_{\ell m}$ and 0.841 or 0.843 for VIM_{Ψ}. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_{Ψ} correlations difference were (.065, .118) and (.067, .120), respectively.

The GSM_4 results for a teacher/student variance ratio of 2 in Figure 3.9 came out quite differently than for the case where the ratio was 20 in Figure 3.10. In this instance, VIM_{Ψ} and VIM_{Υ} were significantly outperformed by $VIM_{\ell m}$ in all the cases. One instance of this was for when the number of teachers was 20 and the number of students per teacher was 24. In that case the 95% confidence interval for the Spearman's correlation difference of $VIM_{\ell m}$ and VIM_{Ψ} was (.030, .092). An increase in the number of teacher or the number of students per teacher produced higher correlations in most cases for every VIM that was studied; the only exception was for the proposed measures when the number of teachers increased from 40 to 100 teachers.

Figure 3.11 plots the mean correlation/association for the five measures across the ratios of the number of students per teacher in group 1 to those in group 2 (x-axis) when teacher variances range from 1 to 20 (columns) with 40 teachers. We should note that in conditions with fewer teachers (10 and 20, respectively), results tended to follow similar patterns, although correlations across conditions for all measures tended to be lower. Associated graphs for 10 and 20 teachers are provided in Appendix A.

For GSM_1, $VIM_{\ell m}$ outperformed significantly the random forest VIMs for all values of teacher variance. As seen in Figure 3.11, as teacher variance increased, the gap between $VIM_{\ell m}$ and the random forest measures decreased. The lowest correlations were found in conditions with group 1 to group 2 ratio equal to 36 to 12, for all measures. For example, when the teacher variance was 20 and the group 1 to group 2 ratio was 30 to 18, the sample mean correlations were .94 for $VIM_{\ell m}$ and .81 for VIM_{Ψ} and the 95% confidence intervals for the difference of $VIM_{\ell m}$ and VIM_{Ψ} correlations was (.097, .155).

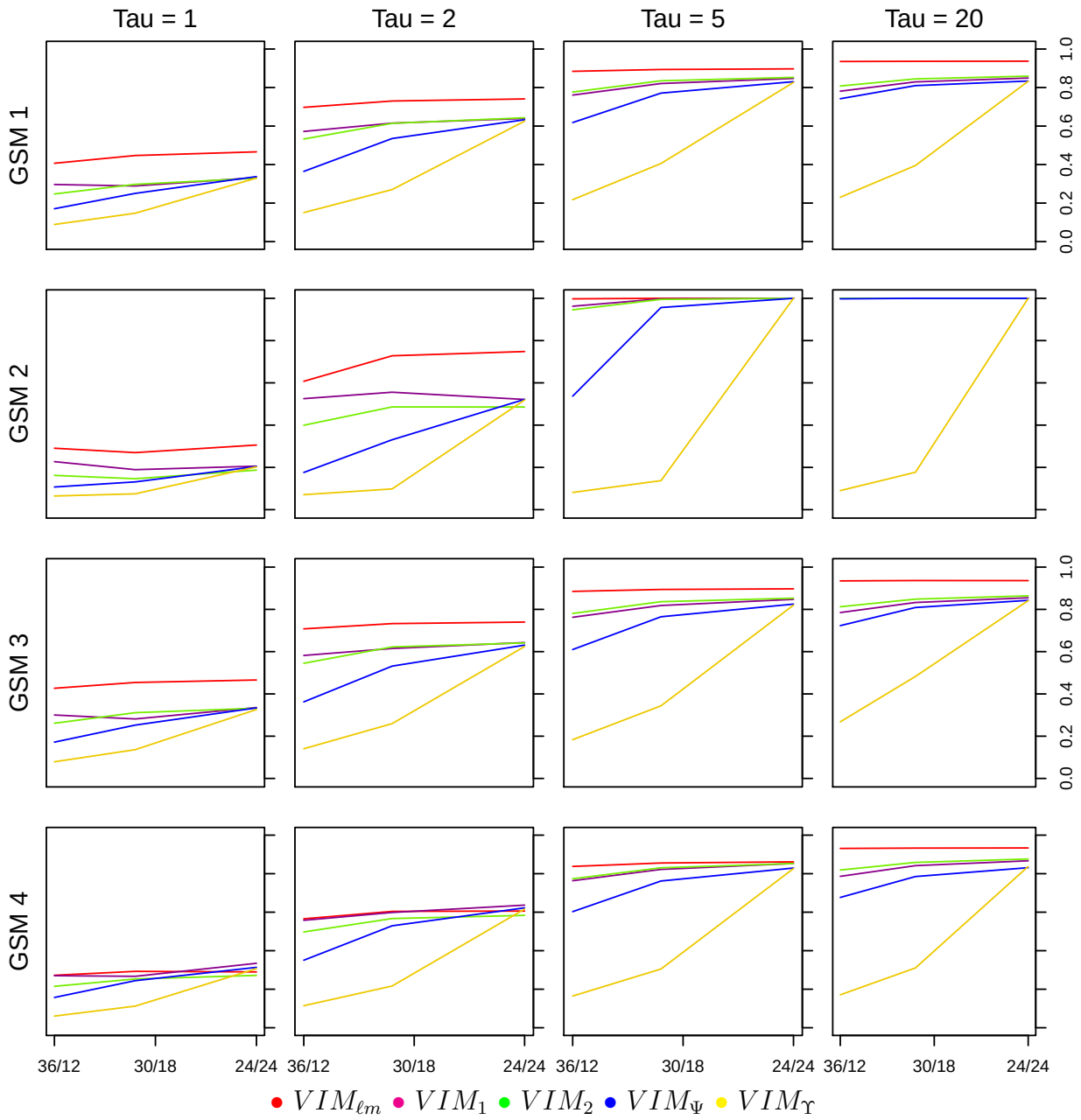

Figure 3.11: Mean correlation/association between the VIMs and the absolute value of true teacher effects when SpT_1/SpT_2 varies for different GSM models and different $\sigma^2_\tau/\sigma^2 = 2$ when the number of teachers is 40.

In GSM_2, when teacher variances were 1 or 2, $VIM_{\ell m}$ statistically outperformed the random forest methods in every case. When the teacher variance was 1 and the number of students per teacher in group 1 to group 2 ratios were 30 to 18 or 36 to 12, the mean association measures were, respectively, 0.27 or 0.29 for $VIM_{\ell m}$ and 0.13 or 0.11 for VIM_{Ψ}. The corresponding 95% confidence intervals of $VIM_{\ell m}$ and VIM_{Ψ} association difference were (.094, .183) and (.138, .229), respectively. In contrast, when the teacher variances were 5 or 20, the mean association in most cases was 1 or close to 1 and $VIM_{\ell m}$ was not significantly better than any random forest VIMs.

We see from Figure 3.11 that the results for simple interactions are similar to those for the baseline model. Beyond that, we note that regardless of the group 1 to group 2 ratio, $VIM_{\ell m}$ outperformed significantly all random forest methods. The performance gap between $VIM_{\ell m}$ and the random forest measures narrowed as the teacher variance increased and the group 1 to group 2 ratio became balanced.

For GSM_4, when teacher variance was 1 or 2, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$ in conditions where the number of students per teacher in group 1 to group 2 ratio was 24 to 24 (balanced case). Specifically, when the teacher variance was 1 in the balanced case, the mean correlations were 0.31 for VIM_{Ψ} and 0.29 for $VIM_{\ell m}$ and the 95% confidence interval for the correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ was (-.009, .056). When the teacher variance was 2 in the balanced case, the mean correlations were 0.62 for VIM_{Ψ} and 0.61 for $VIM_{\ell m}$ and the 95% confidence interval for the correlation difference of VIM_{Ψ} and $VIM_{\ell m}$ was (-.004, .037). When the teacher variance was 5 or 20, $VIM_{\ell m}$ signifcantly outperformed VIM_{Ψ} for any combination of the number of students per teacher in group 1 to group 2 ratio.

Figure 3.12 shows the mean correlations for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers

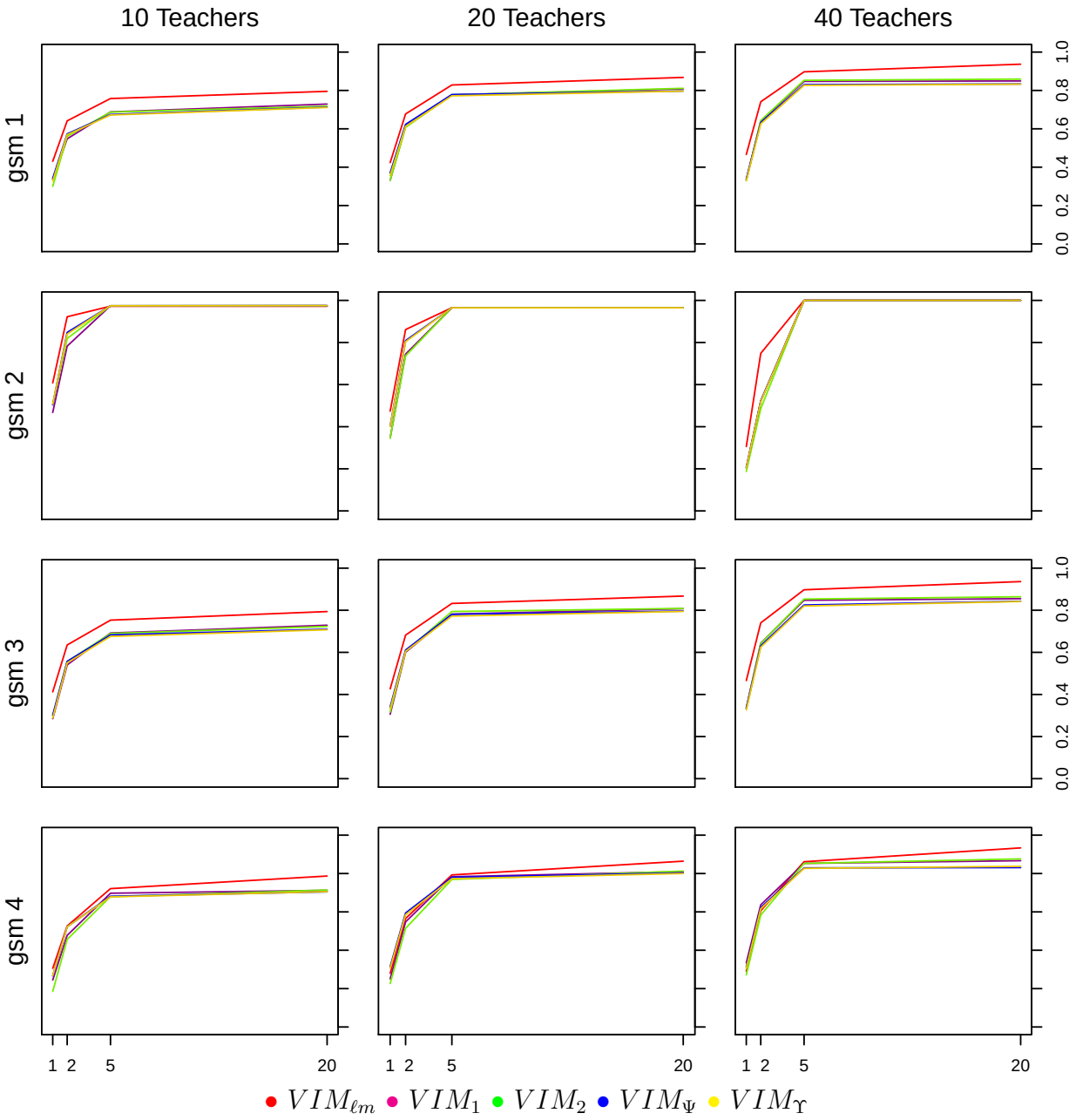

Figure 3.12: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different GSM models and different number of teachers when the group 1 to group 2 ratio is 24/24.

(columns) when the group 1 to group 2 ratio is $\frac{24}{24}$ (balanced case). Due to similarity of results, associated graphs for $\frac{12}{12}$ and $\frac{36}{36}$ have been placed in Appendix A.

For GSM_1, $VIM_{\ell m}$ yielded significantly higher correlations than random forest VIMs in every case studied. As the teacher variance increased, the random forest

measures performed better. Overall, the highest correlations for all measures tended to be found in conditions with 40 teachers and teacher variance of 20. The results for cases with 100 teachers (not shown in the figure) were slightly better than those with 40 teachers. The results for GSM_3 were similar to those described for GSM_1. For GSM_2, when the teacher variance was 5 or 20, VIM_Ψ was not significantly worse than $VIM_{\ell m}$ for each case studied. On the other hand, when the teacher variance was 1 or 2, $VIM_{\ell m}$ outperformed significantly VIM_Ψ in each case studied. When, for example, the teacher variance was 1 and the number of teachers 40, the 95% confidence interval for the difference in correlation of $VIM_{\ell m}$ and VIM_Ψ was (.060, .143). For GSM_4, Figure 3.12 shows that for those cases with teacher variance equal to 1 or 2, VIM_Ψ was not significantly worse than $VIM_{\ell m}$. This was also the case when the number of teachers was 20 and teacher variance was 5. For the remaining cases with teacher variance 5 as well as all cases with teacher variance 20, $VIM_{\ell m}$ was significantly better than VIM_Ψ. All the measures' performances tended to improve as the number of teachers increased although did not carry over to the random forest measures when we considered an additional case with 100 teachers.

Figure 3.13 summarizes the mean correlations/associations for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers (columns). The number of students per teacher in group 1 to group 2 ratio is unbalanced at $\frac{36}{12}$.

For GSM_1 and GSM_3, $VIM_{\ell m}$ significantly outperformed VIM_Ψ in every case. For GSM_2, when the teacher variance was 1, 2, or 5, $VIM_{\ell m}$ significantly outperformed VIM_Ψ for any number of teachers. On the other hand, when the teacher variance was 20, VIM_Ψ was not significantly worse than $VIM_{\ell m}$. For example, with 20 teachers and teacher variance 20, the mean association for $VIM_{\ell m}$ was .965 and for VIM_Ψ was .966. The 95% confidence interval for $VIM_{\ell m}$ and VIM_Ψ association difference was (-.003, .001). For GSM_4, $VIM_{\ell m}$ outperformed significantly VIM_Ψ in all cases. For example, with 40 teachers and teacher variance 1, the mean correlations for $VIM_{\ell m}$ was .272 and

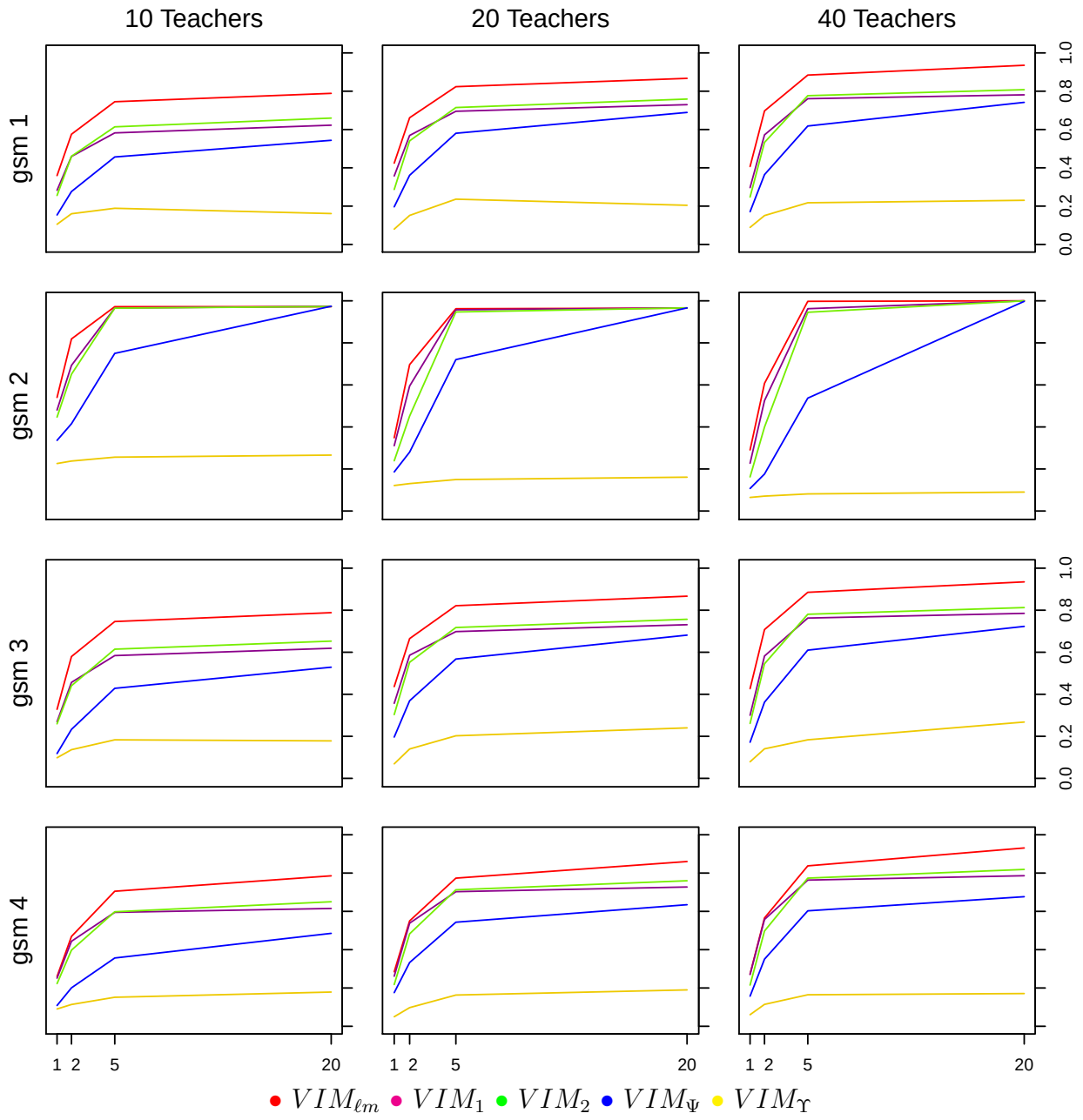

Figure 3.13: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different GSM models and different number of teachers when the number of students per teacher in group 1 to group 2 ratio is 36/12.

for VIM_{Ψ} was .157. The 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} correlation difference was (-.080, .150).

Figure 3.13 also demonstrates that in this unbalanced case, for all four GSM specifications, the *covariate proportion* VIM, VIM_{Υ}, produced consistently low association measures. As we explained for the CAM results, this is an artifact of the way the measure is constructed.

Figure 3.14 graphically depicts the mean correlations/associations for the five measures across the various values of teacher variance over student variance (x-axis) with 10, 20 or 40 teachers (columns) when the number of students per teacher is unbalanced at $\frac{30}{18}$. For GSM_1 and GSM_3, $VIM_{\ell m}$ significantly outperformed VIM_{Ψ} in every case. For GSM_2, when the teacher variance was 1 or 2, $VIM_{\ell m}$ significantly outperformed VIM_{Ψ} for any number of teachers. On the other hand, when teacher variance was 5 and number of teachers 10 or 20, or when teacher variance was 20 for any number of teachers, VIM_{Ψ} was not significantly worse than $VIM_{\ell m}$. For example, with 10 teachers and teacher variance 5, the mean association for $VIM_{\ell m}$ was .972 and for VIM_{Ψ} was .970. The 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} association difference was (-.005, .011). For GSM_4, $VIM_{\ell m}$ significantly outperformed VIM_{Ψ} in nearly all cases. Only when the number of teachers was 20 and the teacher variance was 1 was VIM_{Ψ} found not to be significantly worse than $VIM_{\ell m}$. In this case, the mean correlations for $VIM_{\ell m}$ was .312 and for VIM_{Ψ} was .267. The 95% confidence interval for $VIM_{\ell m}$ and VIM_{Ψ} correlation difference was (-.004, .101).

As was true for the CAM, in all GSM scenarios, the mean correlation/association differences between the random forest methods and $VIM_{\ell m}$ that were observed when the group 1 to group 2 ratio was 36 to 12, were larger than in the case where this ratio was 30 to 18. This transpires for the same reason it did in the analogous CAM setting.

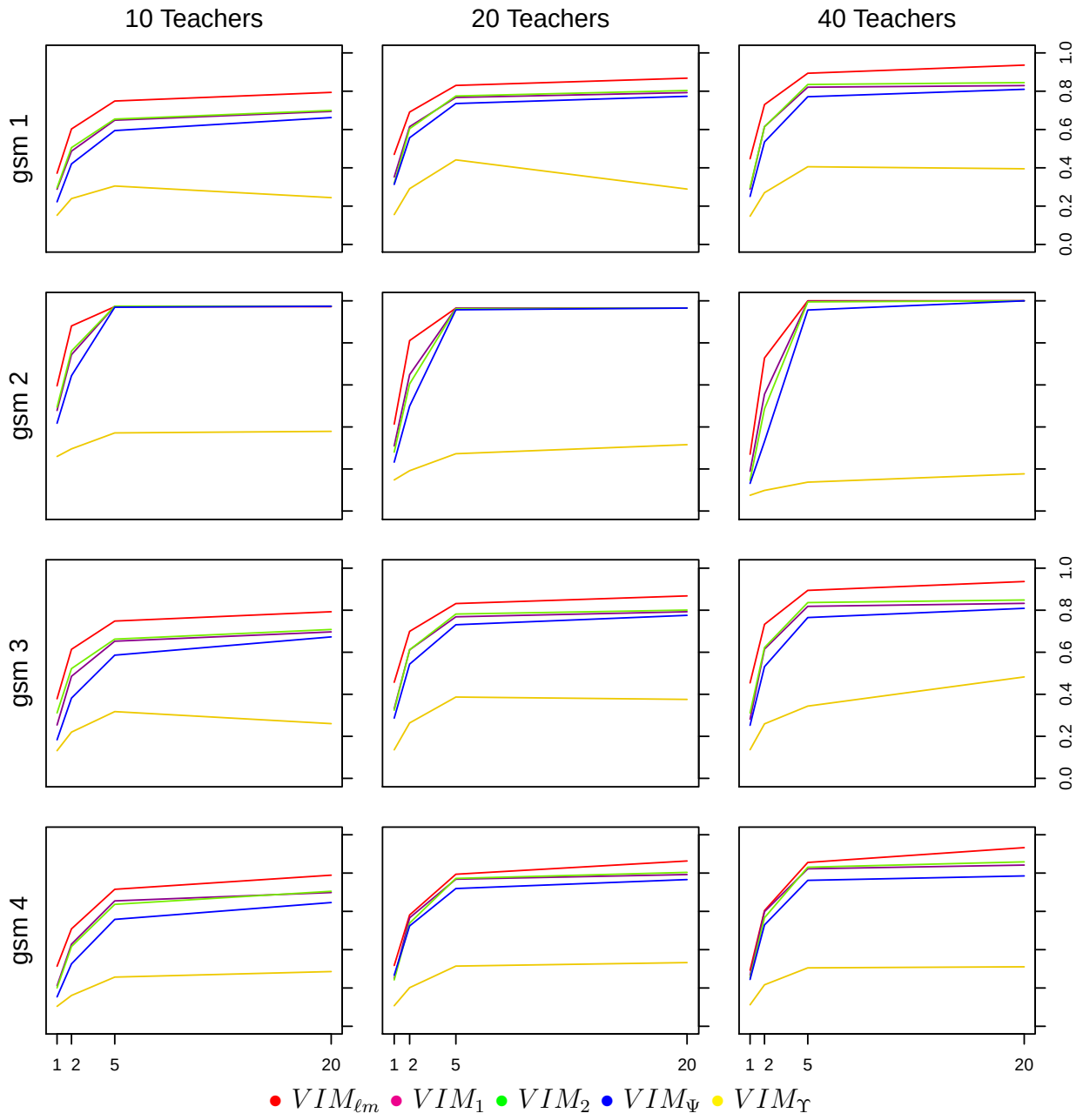

Figure 3.14: Mean correlation/association between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teacher in group 1 to group 2 ratio is 30/18.

3.3 Discussion

Based on the configuration of the terminal nodes in a random forest, we have proposed two new measures to rank the input variables based on their influence in prediction in the context of VAMs; namely, the node-proportion and the covariate-proportion VIMs. For each simulation setting we compared the average across 100 replicates of the correlation/association between the ranks produced by the absolute value of the true (teacher) random effects and each of the different VIMs. The rank produced by $VIM_{\ell m}$ was constructed assuming always a generic linear mixed model formulation, given by (2.6) or (2.10). However, for certain simulation settings, we specified the model allowing modifications and/or extensions to these linear model formulations. The purpose of this exercise was to determine, given a model misspecification, how VIM_{Ψ} and VIM_{Υ} performed in comparison to $VIM_{\ell m}$. In particular, the simulation considered models that included third-order interactions among three covariates, two of them associated with fixed effects and one associated with random effects. We called this formulation the *complex interaction* model, and simulation results based on this formulation were represented by CAM_4 or GSM_4.

In terms of the CAM models, the simulation results with CAM_4 provide the central justification for the relevance of our proposed measures. These results are summarized in the figures in Section 3.2; particularly, in Figures 3.3 and 3.4. To better display the cases where VIM_{Ψ} performance was at least as good as $VIM_{\ell m}$, Table 3.1 presents all the factor combinations in the simulation study that correspond to CAM_4 where VIM_{Ψ} is not significantly worse than $VIM_{\ell m}$. In this and subsequent tables we use N_{teach} to represent the number of teachers for each particular simulation setting. In addition, recall that SpT_{ℓ} represent the number of students per teacher in group ℓ for $\ell = 1, 2$ and $r(\Psi)$ and $r(\ell m)$ are the correlation/association averages defined in (3.9). The paired t-test statistic with the associated lower and upper limits ($CI_{\ell\ell}$ and $CI_{u\ell}$) for a 95% confidence interval and the p-values are also presented in the table.

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
10	12	12	1	0.26	0.19	-0.07	-1.89	-0.154	0.004	6.20E-02
10	24	24	1	0.38	0.37	-0.01	-0.21	-0.077	0.062	8.31E-01
10	36	36	1	0.40	0.44	0.04	1.19	-0.029	0.113	2.38E-01
10	24	24	2	0.61	0.59	-0.02	-0.63	-0.062	0.032	5.27E-01
10	36	36	2	0.64	0.64	-0.00	-0.17	-0.045	0.038	8.68E-01
20	12	12	1	0.31	0.30	-0.01	-0.42	-0.066	0.043	6.75E-01
20	24	24	1	0.41	0.51	0.10	4.01	0.049	0.146	1.19E-04
20	36	36	1	0.49	0.58	0.10	5.05	0.058	0.133	2.00E-06
20	30	18	1	0.42	0.38	-0.05	-1.63	-0.102	0.010	1.07E-01
20	12	12	2	0.57	0.56	-0.01	-0.55	-0.046	0.026	5.82E-01
20	24	24	2	0.65	0.72	0.07	4.68	0.041	0.103	9.28E-06
20	36	36	2	0.72	0.76	0.04	3.14	0.013	0.060	2.25E-03
20	24	24	5	0.82	0.80	-0.02	-1.66	-0.038	0.003	9.94E-02
40	12	12	1	0.29	0.41	0.12	7.36	0.091	0.158	5.39E-11
40	24	24	1	0.42	0.62	0.20	14.95	0.176	0.230	3.92E-27
40	36	36	1	0.51	0.71	0.20	16.26	0.173	0.222	1.06E-29
40	36	12	1	0.39	0.36	-0.03	-1.42	-0.061	0.010	1.58E-01
40	30	18	1	0.43	0.53	0.10	5.78	0.065	0.134	8.57E-08
40	12	12	2	0.59	0.70	0.11	8.68	0.082	0.130	8.07E-14
40	24	24	2	0.71	0.81	0.10	12.04	0.085	0.119	4.15E-21
40	36	36	2	0.76	0.85	0.09	12.48	0.073	0.100	4.95E-22
40	30	18	2	0.71	0.74	0.03	3.80	0.015	0.047	2.47E-04
40	12	12	5	0.84	0.83	-0.01	-1.82	-0.029	0.001	7.24E-02
100	12	12	1	0.30	0.53	0.23	20.21	0.206	0.251	6.47E-37
100	24	24	1	0.44	0.71	0.27	27.56	0.249	0.287	3.08E-48
100	36	12	1	0.42	0.49	0.07	7.81	0.055	0.092	6.16E-12
100	30	18	1	0.44	0.64	0.19	22.26	0.175	0.209	2.58E-40
100	12	12	2	0.61	0.78	0.17	23.84	0.160	0.189	8.32E-43
100	24	24	2	0.74	0.88	0.14	23.43	0.126	0.150	3.52E-42
100	36	12	2	0.71	0.71	0.00	0.59	-0.008	0.015	5.55E-01
100	30	18	2	0.74	0.83	0.10	18.30	0.087	0.108	1.59E-33
100	12	12	5	0.88	0.88	0.00	0.36	-0.007	0.011	7.19E-01

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
100	24	24	5	0.93	0.92	-0.00	-0.95	-0.014	0.005	3.42E-01

Table 3.1: All the factor combinations for the CAM *complex interaction* scenarios, CAM_4, where VIM_Ψ was not significantly worse than $VIM_{\ell m}$. Remaining factor combinations, the mean correlations for VIM_Ψ and $VIM_{\ell m}$ and the inferential study results of paired samples t-tests are shown.

Table 3.1 shows that for any number of teachers, when the teacher variance was 1 or 2 and the number of students in group 1 to group 2 ratio was balanced (12 to 12, 24 to 24, or 36 to 36), the node-proportion is not significantly worse than the other measures, including $VIM_{\ell m}$. In the cases where the number of teachers was at least 20 and the number of students per teacher was at least 24, the proposed measures outperformed significantly all the other measures. This happens because the covariate-proportion and node-proportion better capture the complex structure of the model. More precisely, the random forest captures the complex structure of the model and these measures reflect more accurately this information. Observe as well that the proposed measures better reflect the teacher effects when the number of teachers increases or the number of student per teacher increases.

While not unexpected, the results were completely different when simulations were obtained using the baseline model, CAM_1, or the simple interaction model, CAM_3. In each of those situations and for any factor combination, VIM_Ψ measure was significantly worse than $VIM_{\ell m}$.

Table 3.2 presents similar results to those in Table 3.1 for the *good teacher - bad teacher model, CAM_2*. Recall that in this model, there is no teacher variance, rather σ_τ^2 represents a multiplier to obtain the magnitude of the good teacher effect ($1.5 * \sigma_\tau^2$) and the bad teacher effect ($-1 * \sigma_\tau^2$). Table 3.2 shows that VIM_Ψ was not significantly worse than $VIM_{\ell m}$ when σ_τ^2 was 5 or 20, or when the number of students per teacher was 36 for all teachers and σ_τ^2 was 2.

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
10	36	36	2	0.97	0.97	0.00	0.28	-0.006	0.008	7.78E-01
10	12	12	5	0.97	0.97	0.00	1.22	-0.002	0.008	2.25E-01
10	24	24	5	0.97	0.97	0.00	0.78	-0.003	0.007	4.36E-01
10	36	36	5	0.98	0.97	0.00	1.26	-0.002	0.008	2.09E-01
10	30	18	5	0.97	0.97	0.00	1.04	-0.002	0.007	2.99E-01
10	12	12	20	0.98	0.97	0.00	1.19	-0.003	0.012	2.38E-01
10	24	24	20	0.97	0.97	0.00	1.18	-0.002	0.007	2.41E-01
10	36	36	20	0.98	0.97	0.00	1.44	-0.001	0.009	1.54E-01
10	36	12	20	0.97	0.97	0.00	1.34	-0.001	0.005	1.84E-01
10	30	18	20	0.97	0.97	0.00	1.59	-0.001	0.005	1.16E-01
20	36	36	2	0.96	0.97	-0.00	-0.91	-0.007	0.003	3.65E-01
20	12	12	5	0.96	0.96	-0.00	-0.60	-0.001	0.000	5.53E-01
20	24	24	5	0.96	0.97	-0.00	-0.74	-0.001	0.001	4.60E-01
20	36	36	5	0.97	0.97	0.00	0.37	-0.001	0.001	7.09E-01
20	30	18	5	0.96	0.97	-0.00	-0.60	-0.001	0.001	5.51E-01
20	12	12	20	0.97	0.97	0.00	0.16	-0.001	0.002	8.74E-01
20	24	24	20	0.97	0.97	0.00	0.08	-0.001	0.001	9.36E-01
20	36	36	20	0.97	0.97	0.00	0.03	-0.001	0.001	9.76E-01
20	36	12	20	0.97	0.97	0.00	1.41	-0.001	0.003	1.62E-01
20	30	18	20	0.97	0.97	0.00	0.61	-0.001	0.002	5.41E-01
40	36	36	2	1.00	1.00	0.00				
40	12	12	5	1.00	1.00	0.00				
40	24	24	5	1.00	1.00	0.00				
40	36	36	5	1.00	1.00	0.00				
40	30	18	5	1.00	1.00	0.00				
40	12	12	20	1.00	1.00	0.00				
40	24	24	20	1.00	1.00	0.00				
40	36	36	20	1.00	1.00	0.00				
40	36	12	20	1.00	1.00	0.00				
40	30	18	20	1.00	1.00	0.00				
100	24	24	2	0.99	1.00	-0.01	-1.75	-0.016	0.001	8.32E-02
100	12	12	5	1.00	1.00	-0.00	-1.00	-0.007	0.002	3.20E-01

N_{teach}	SpT_1	SpT_2	σ^2_τ	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
100	24	24	5	1.00	1.00	0.00				
100	36	12	5	1.00	1.00	0.00				
100	30	18	5	1.00	1.00	0.00				
100	12	12	20	1.00	1.00	0.00				
100	24	24	20	1.00	1.00	0.00				
100	36	12	20	1.00	1.00	0.00				
100	30	18	20	1.00	1.00	0.00				

Table 3.2: Factor combinations for the *good teacher - bad teacher model, CAM_2,* where VIM_Ψ was at least not significantly worse than $VIM_{\ell m}$. Remaining factor combinations, the mean association for VIM_Ψ and $VIM_{\ell m}$ and the simulation study results are shown.

As was the case for CAM, the central results for the GSM portion of our simulation study are realized when analyzing selected factor combinations for GSM_4. A summary of those results is presented in the figures in Section 3.2; particularly, in Figure 3.9. Table 3.3 presents all the factor combinations in the simulation study that correspond to GSM_4 where VIM_Ψ is not significantly worse than $VIM_{\ell m}$.

N_{teach}	SpT_1	SpT_2	σ^2_τ	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
10	24	24	1	0.27	0.31	-0.04	-1.11	-0.109	0.031	2.72E-01
10	36	36	1	0.33	0.31	0.02	0.54	-0.048	0.083	5.93E-01
10	24	24	2	0.52	0.53	-0.00	-0.15	-0.052	0.044	8.81E-01
10	36	36	2	0.58	0.55	0.03	1.55	-0.009	0.075	1.24E-01
20	12	12	1	0.16	0.17	-0.02	-0.52	-0.076	0.044	6.01E-01
20	24	24	1	0.31	0.28	0.03	1.50	-0.011	0.081	1.38E-01
20	36	36	1	0.42	0.38	0.04	2.00	0.000	0.086	4.78E-02
20	30	18	1	0.27	0.32	-0.05	-1.84	-0.106	0.004	6.89E-02
20	12	12	2	0.39	0.43	-0.03	-1.37	-0.082	0.015	1.73E-01
20	24	24	2	0.59	0.57	0.03	1.66	-0.005	0.062	9.94E-02
20	36	36	2	0.68	0.66	0.02	1.41	-0.008	0.045	1.62E-01
20	24	24	5	0.78	0.79	-0.01	-1.00	-0.038	0.013	3.22E-01
20	36	36	5	0.80	0.82	-0.02	-1.81	-0.052	0.002	7.37E-02
40	12	12	1	0.15	0.17	-0.02	-0.99	-0.056	0.019	3.27E-01
40	24	24	1	0.31	0.29	0.02	1.44	-0.009	0.056	1.53E-01

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
40	36	36	1	0.44	0.39	0.05	2.79	0.013	0.080	6.27E-03
40	24	24	2	0.62	0.61	0.02	1.59	-0.004	0.037	1.14E-01
40	36	36	2	0.72	0.68	0.04	4.34	0.024	0.064	3.41E-05

Table 3.3: All the factor combinations for GSM_4, where VIM_Ψ was not significantly worse than $VIM_{\ell m}$. Remaining factor combinations, the mean correlations for VIM_Ψ and $VIM_{\ell m}$ and the inferential study results of paired samples t-tests are shown.

In Table 3.3, we can observe that for 10, 20, or 40 teachers, when the teacher variance is 1 or 2 and the number of students per teacher in group 1 to group 2 ratio is balanced (12 to 12, 24 to 24, or 36 to 36), the two proposed VIMs are not significantly worse than the other measures, including $VIM_{\ell m}$. In the cases where the number of teachers was 40 and the number of students per teacher was 36, the proposed measures actually outperformed (significantly) $VIM_{\ell m}$. As with the CAM scenarios, this happens because the random forest captures the complex structure of the model and these measures reflect more accurately this information. Observe again that the proposed measures better reflect the teacher effects when the number of teachers increases or the number of students per teacher increases.

We found that VIM_Ψ was significantly worse than $VIM_{\ell m}$ in any instances for the baseline model, GSM_1, or simple interaction model, GSM_3. Table 3.4 presents similar results to Table 3.3 for the *good teacher - bad teacher model*. Table 3.4 shows that VIM_Ψ was not significantly worse than $VIM_{\ell m}$ when σ_τ^2 was 5 or 20, or when the number of students per teacher was 12 for all teachers and σ_τ^2 was 1.

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
10	12	12	1	0.43	0.44	-0.00	-0.02	-0.042	0.042	9.87E-01
10	24	24	5	0.97	0.97	0.00	0.78	-0.003	0.007	4.36E-01
10	36	36	5	0.98	0.97	0.00	1.26	-0.002	0.008	2.09E-01
10	30	18	5	0.97	0.97	-0.00	-0.63	-0.011	0.006	5.32E-01
10	12	12	20	0.98	0.97	0.00	1.19	-0.003	0.012	2.38E-01

N_{teach}	SpT_1	SpT_2	σ_τ^2	$r(\ell m)$	$r(\Psi)$	$r(\Psi)$-$r(\ell m)$	t	$CI_{\ell\ell}$	$CI_{u\ell}$	p
10	24	24	20	0.97	0.97	0.00	1.18	-0.002	0.007	2.41E-01
10	36	36	20	0.98	0.97	0.00	1.44	-0.001	0.009	1.54E-01
10	36	12	20	0.97	0.97	0.00	1.34	-0.001	0.005	1.84E-01
10	30	18	20	0.97	0.97	0.00	1.59	-0.001	0.005	1.16E-01
20	24	24	5	0.96	0.97	-0.00	-0.74	-0.001	0.001	4.60E-01
20	36	36	5	0.97	0.97	0.00	0.37	-0.001	0.001	7.09E-01
20	30	18	5	0.96	0.97	-0.01	-1.80	-0.016	0.001	7.54E-02
20	12	12	20	0.97	0.97	0.00	0.16	-0.001	0.002	8.74E-01
20	24	24	20	0.97	0.97	0.00	0.08	-0.001	0.001	9.36E-01
20	36	36	20	0.97	0.97	0.00	0.03	-0.001	0.001	9.76E-01
20	36	12	20	0.97	0.97	0.00	1.41	-0.001	0.003	1.62E-01
20	30	18	20	0.97	0.97	0.00	0.61	-0.001	0.002	5.41E-01
40	24	24	5	1.00	1.00	0.00				
40	36	36	5	1.00	1.00	0.00				
40	12	12	20	1.00	1.00	0.00				
40	24	24	20	1.00	1.00	0.00				
40	36	36	20	1.00	1.00	0.00				
40	36	12	20	1.00	1.00	-0.00	-1.00	-0.007	0.002	3.20E-01
40	30	18	20	1.00	1.00	0.00				
100	12	12	20	1.00	1.00	-0.00	-1.00	-0.007	0.002	3.20E-01
100	24	24	20	1.00	1.00	0.00				
100	30	18	20	1.00	1.00	0.00				

Table 3.4: Factor combinations for the GSM_2, where VIM_Ψ was at least not significantly worse than $VIM_{\ell m}$. Remaining factor combinations, the mean correlations for VIM_Ψ and $VIM_{\ell m}$ and the t-statistics are shown.

CHAPTER 4

INTERACTIONS

One of the advantages of the random forest method is its ability to perform well when the data exhibit interactions among input variables. In the literature, "performance" has generally been assessed in terms of predictive ability. The results in the previous chapter represent an attempt to extend the notion of performance to include the ability to detect important variables. The present chapter proceeds in a similar vein except now the goal is the explicit identification of the variables that interact using a random forest approach.

In what follows we propose methods, based on random forest, that identify and/or measure interactions among input variables. We begin by developing a new statistic that can be used to identify variable interactions based on unique patterns observed in the structure of the trees under limited modelling specification. We consider the specific case of data from a linear model and explore the properties of our measure in that context. While still under the linear model specification, we restrict our statistic in a way that not only identifies but also estimates the interaction effects. To conclude, the results of a simulation study are presented that provide evidence of both the reach and limitations of our methodology.

Subsequent mathematical developments require a refinement of the notation laid out in Chapter 2. Figure 4.1 provides a simple illustration that we will use to explain the new ideas that are involved.

As in Chapter 2 a tree is a representation of partitions corresponding to some data set, $\mathcal{L}_N$. A typical internal node in a tree is denoted as η. When necessary, we write η_k for the k-th internal node or even η_k^t for the k-th internal node in the t-th tree. The root node η_1 corresponds to the entire set of observations. Any other node is a subset of $\mathcal{L}_N$ given by partitions that are determined by subsets corresponding to categories obtained

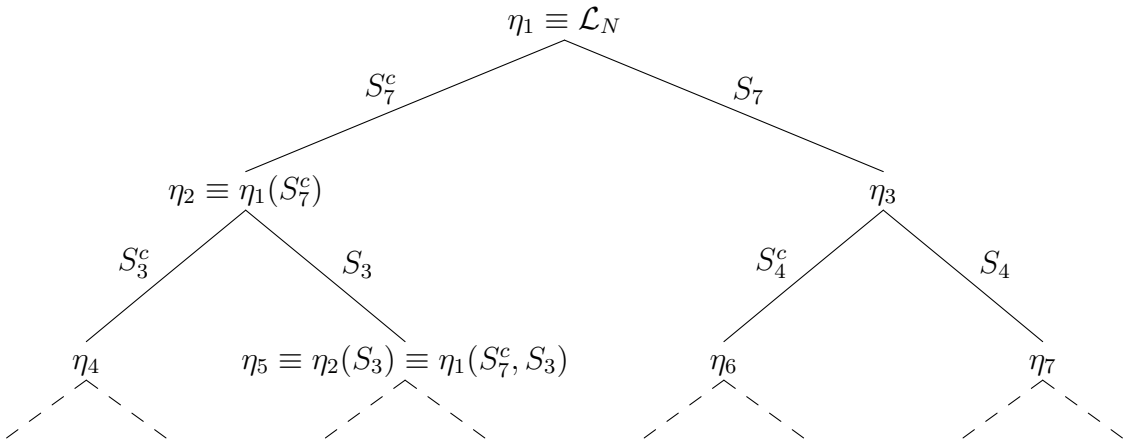

Figure 4.1: A graphical representation of a tree showing 7 non-terminal nodes and implicitely 3 splitting variables, X_7, X_3, and X_4. Their corresponding subsets of categories are shown explicitely. The root node is equal to the entire set of observations used to grow the tree, $\mathcal{L}_N$. For nodes η_2 and η_5, alternative representations using the parent nodes are also shown.

by splitting the range of a predictor variable. We will use $\mathcal{T}(\eta)$ to indicate the subtree that arises from viewing a particular η as a root node.

To describe the splitting idea let C_p be the set of categories of X_p for $p = 1, \ldots, P$. Then, a split on X_p can be represented as a set $S_p \subset C_p$ with observations being assigned to the right child node if its value of X_p is in S_p and to the left child node if it is in S_p^c. For example, in Figure 4.1, the splitting variable in the root node is X_7 and $S_7, S_7^c \subset C_7$ are the subsets of values for X_7 that are used to partition the data into the two observation subsets that represent the nodes η_2 and η_3.

A branch is the unique path or history of partitions that produces a subset of observations (i.e., a node) in the tree. We denote the branch corresponding to node η by $H(\eta)$. This branch can be described via an ordered list of the form

$$H(\eta) = (S_{1p_1}, S_{2p_2}, \ldots, S_{Lp_L}). \tag{4.1}$$

Here L is the number of nodes on the branch and $\{p_1, \ldots, p_L\}$ are all variable indices in $\{1, \ldots, P\}$ with $S_{\ell p_\ell}$ the subset of values for X_{p_ℓ} that was used to create the ℓ-th

partition. Note that a variable index can appear more than once. For example, if $p_i = p_j$ and $i < j$ then $S_{jp_j}, S^c_{jp_j} \subset S_{ip_i}$. When each variable appears at most once in a branch we alternatively express the branch in the simpler notation

$$H(\eta) = (S_{p_1}, S_{p_2}, \ldots, S_{p_L}). \tag{4.2}$$

In particular, in the next Section we will focus on binary variables which, by their dichotomous nature, can only be used in a single split. As an illustration, the branch corresponding to node η_6 in Figure 4.1 is $H(\eta_6) = (S_7, S^c_4)$.

If no confusion arises, we will treat the list $H(\eta)$ as being synonymous with the set that contains the subsets in the list and we write $S_p \in H(\eta)$ only if two conditions are satisfied: X_p splits at least one of the nodes in $H(\eta)$ and after that node, the resulting partition keeps only the observations having realized values of X_p in S_p. Because of that, if $S_p \in H(\eta)$ then $S^c_p \notin H(\eta)$; on the other hand, when the variable X_p has not been used as a splitting variable in $H(\eta)$, $S_p \notin H(\eta)$ and $S^c_p \notin H(\eta)$. For example, in Figure 4.1, $S^c_4 \in H(\eta_6)$, $S_7 \notin H(\eta_4)$, and $S_2, S^c_2 \notin H(\eta_6)$.

We will use $\phi(\eta)$ to represent the function that indicates the variable used to create η's child nodes; for example in Figure 4.1, $\phi(\eta_3) = X_4$.

Somewhat more generally, we will denote by $\eta(S_p)$ the subset of η given by those observations with realized values of X_p in S_p, for any arbitrary partition that need not be one in the tree. However, if $\phi(\eta) = X_p$ and S_p is the corresponding subset of X_p used to split the observations, there is a node $\tilde{\eta}$ with $\tilde{\eta} = \eta(S_p)$; i.e., $\eta(S_p)$ is a child node of η. In the instance where no variable has been used for more than one split, we can represent this by writing

$$H(\tilde{\eta}) \equiv H(S_p; \eta) \equiv H(\eta) \cup S_p \equiv (S_{p_1}, S_{p_2}, \ldots, S_{p_L}, S_p).$$

On the other hand, when $\eta(S_p)$ is not a child node of η, the resulting subset of observations is $\eta(S_p) = \eta$ if $S_p \in H(\eta)$, $\eta(S_p) = \emptyset$ if $S^c_p \in H(\eta)$, or any subset of

observations in between when $S_p, S_p^c \notin H(\eta)$. Similarly, in the case of more than one partition, $\eta(S_{p_1}, \ldots, S_{p_M}) \subseteq \eta$ is the subset of the elements in η that satisfy the conditions for the predictor variable values determined by $S_{p_1}, \ldots, S_{p_M}$ where the S_j may or may not be partitions that were used to create η. Figure 4.1 illustrates this idea with alternative node representations for nodes η_2 and η_5.

Finally, we will use $|\eta|$ to represent the number of observations in a node η and

$$\bar{y}(\eta) = \frac{1}{|\eta|} \sum_{i:(\mathbf{x}_i, y_i) \in \eta} y_i$$

to denote its mean output value. These representations extend to any partition of η when this partition exists. For example, if $\eta(S_p, S_q)$ is not empty we have

$$\bar{y}(S_p, S_q; \eta) = \frac{1}{|\eta(S_p, S_q)|} \sum_{i:(\mathbf{x}_i, y_i) \in \eta(S_p, S_q)} y_i.$$

On the other hand, when $\eta(S_p, S_q)$ is the empty set, we take $\bar{y}(S_p, S_q; \eta)$ to simply be undefined.

4.1 A New Approach to Interaction Identification

We now introduce a new approach to identifying interaction effects. The method uses the random forest estimator $\hat{F}_\alpha$ in (2.17) for the regression function F in (2.1). Given a data set $\mathcal{L}_N = \{\mathbf{x}_i, y_i\}_{i=1}^N$ consisting of N realizations of the random vector $(\mathbf{X}, Y)$, the random forest procedure returns an estimator that is an average of regression tree estimators $F_\alpha\big(\cdot; \hat{\Xi}(\mathcal{B}_t)\big)$ for $t = 1, \ldots, T$, corresponding to trees created from bootstrap samples $\mathcal{B}_1, \ldots, \mathcal{B}_T$ of $\mathcal{L}_N$. We will use the resulting trees to construct a method to identify interactions among input variables.

Our basic premise is that, since random forest performs well when data exhibit interaction effects, it is expected that the tree structure will reflect the presence of interactions. We illustrate this idea in a simple setting. Assume momentarily that for each tree in the random forest the root nodes have all used the same splitting variable X_p and partition the data with the same subset of categories, S_p and S_p^c, as in Figure 4.2. If there

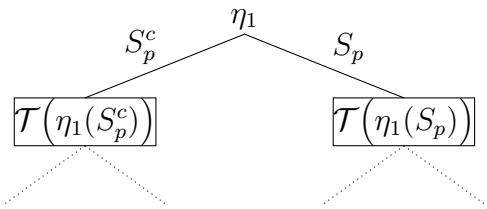

Figure 4.2: A tree having the root node partitioned by variable X_p and generating two subtrees.

are no interaction effects, there are no *a priori* influences to make the subtrees $\mathcal{T}\Big(\eta_1^t(S_p)\Big)$ and $\mathcal{T}\Big(\eta_1^t(S_p^c)\Big)$ consistently different with η_1^t being the root node for the t-th tree. We are not implying that the structure of both subtrees will look alike or even similar in any given tree. This is highly unlikely for two reasons. First, different (random) groups of potential splitting variables are evaluated at each node; therefore, the variable chosen to split a node in one subtree might not even be considered for splitting in the corresponding node in the other subtree. Second, even if the same group of variables is considered on each equivalent node in both subtrees, the unaccounted differences in the bootstrap samples may generate different subtrees. What we are suggesting is that, under the assumption of no interaction effects, if we were to grow T trees, each time with a different bootstrap sample, the collective or ensemble structure of the subtrees $\{\mathcal{T}\Big(\eta_1^t(S_p)\Big)\}_{t=1}^T$ should be similar to that of the subtrees $\{\mathcal{T}\Big(\eta_1^t(S_p^c)\Big)\}_{t=1}^T$, for T sufficiently large.

To analyze the structure for particular variables X_p and X_q in a specific tree we might use the distance between the nodes that used these variables to create splits. This distance can be measured in various ways. For example, it can simply be the number of internal nodes between the nodes where the two splits occur. For now, however, we will leave our choice of distance measure unspecified and use the word "distance" in a generic sense. Notwithstanding the specific choice, when the values of this measure are accumulated across bootstrap samples we will then obtain an approximation for the distribution of the distance. If the collection of tree structures is similar, the distribution of distances between X_p and X_q in $\{\mathcal{T}\Big(\eta_1^t(S_p)\Big)\}_{t=1}^T$ should be similar to the analogous

distribution in $\{\mathcal{T}\Big(\eta_1^t(S_p^c)\Big)\}_{t=1}^T$. Clearly, if X_p and X_q do not interact, we would expect the two distribution to coincide in an asymptotic sense.

On the contrary, if X_q interacts with X_p, we would expect the distribution of distances between X_p and X_q in each collection of subtrees to differ. Moreover, the specific differences in the distribution of distances between collections of subsets might also give an indication of the strength of the interaction effect. This may happen because, if T is large enough, the proportion of times X_q is part of the subset of variables considered for splits is similar at each node in both collections of subtrees. If the interaction effect is strong enough, X_q would be chosen more often in the first few nodes of one collection of subtrees than in the other. Hence, if shorter distances between X_p and X_q in $\{\mathcal{T}\Big(\eta_{1_t}(S_p)\Big)\}_{t=1}^T$ differ from the corresponding shorter distances in $\{\mathcal{T}\Big(\eta_{1_t}(S_p^c)\Big)\}_{t=1}^T$, stronger interaction effects might be present. Our proposed measure takes into account this characteristic.

We fixed $\phi(\eta_1) = X_p$ in the previous discussion for clarity in exposition. This is not necessary and for large enough T, we could perform the same analysis for any locations of X_p and X_q in the tree, provided the comparisons are made for the same specific subsets S_p and S_p^c of categories for X_p. That is, for any node η, if $\phi(\eta) = X_p$, then we apply the same ideas to the subtrees given by $\eta(S_p)$ and $\eta(S_p^c)$. Note that in the case where X_p is binary, S_p and S_p^c will always coincide. Hence, the problem is to determine if the distribution of the distances between X_p and X_q in subtrees given by

$$\{\mathcal{T}\Big(\eta_k^t(S_p)\Big) : \forall k \text{ s.t. } \phi(\eta_k^t) = X_p, t = 1, \ldots, T\} \tag{4.3}$$

is different than that in

$$\{\mathcal{T}\Big(\eta_k^t(S_p^c)\Big) : \forall k \text{ s.t. } \phi(\eta_k^t) = X_p, t = 1, \ldots, T\}. \tag{4.4}$$

In this general situation, we also need to consider the relative location of both variables in the trees. When both variables are closer to the root node, this would also suggest the presence of stronger interactions.

The actual comparison of distance distributions can be based on summary measures. For example, if there are no interactions, we would expect the sum of distances between X_p and X_q in (4.3) to be very similar to the corresponding sum in (4.4). This is precisely what we do to obtain our proposed measure. The *distribution-based* interaction identification measure between X_p and X_q is given by

$$\begin{aligned}\Gamma(p,q) =& \frac{1}{\omega_1}\sum_{t=1}^{T}\sum_{k=1}^{D_t}\Big(I(\phi(\eta_k^t)=X_q)I(S_p\in H(\eta_k^t))\gamma(p,q,\eta_k^t)\Big)\\ &-\frac{1}{\omega_2}\sum_{t=1}^{T}\sum_{k=1}^{D_t}\Big(I(\phi(\eta_k^t)=X_q)I(S_p^c\in H(\eta_k^t))\gamma(p,q,\eta_k^t)\Big)\end{aligned} \tag{4.5}$$

with D_t the total number of interior nodes in the t-th tree. In each sum, the first indicator function in (4.5) allows the measure to keep only those nodes with corresponding splitting variable X_q. The second indicator function keeps those nodes where S_p is part of η_k^t's branch in the first sum or those nodes where S_p^c is part of η_k^t's branch in the second sum. Effectively, it separates the expressions corresponding to sums of those nodes in (4.3) from those in (4.4). The values for ω_1 and ω_2 provide us with the flexibility of being able to weight each sum differently.

The function $\gamma(p,q,\eta_k^t)$ in (4.5) assigns a weight to each selected node based on different criteria. Following the discussion above, we will use this function to consider weights based on the distance between X_p and X_q as well as the relative location of the node in the tree, i.e. the distance from this node to the root node. The function $\gamma(p,q,\eta_k^t)$ need not be restricted only to distance measures. It can represent any additional information related to variables X_p, X_q, and the node η_k^t. For example, in the following section we will propose a second measure derived from (4.5) that includes the outcome means of η_k^t's child nodes. We will postpone giving choices for $\gamma(\cdot,\cdot,\cdot)$ until the next section. For now, it suffices to observe that any formulation that includes distances should assign at least as much weight to shorter distances or nodes closer to the root node as to other distances and nodes. When such is the case, we expect $\Gamma(p,q)$ to be close to zero if there is no interaction between X_p and X_q and nonzero otherwise.

Finally, notice that this analysis has accounted for cases where X_p splits a node before X_q. To account for situations where X_q splits a node before X_p we need to analyze the results obtained with $\Gamma(q,p)$ simultaneously with $\Gamma(p,q)$.

At this point it becomes expedient to analyze how the splitting variable is chosen in a particular node η for some given tree. Recall that at any given node, the random forest algorithm selects the splitting variable from a random subset of variables. There are then different criteria that are employed to choose an "optimal" splitting variable. A common criterion, and the one used in this analysis, is to select the variable that produces the largest reduction in error sum of squares between a node outcome and its potential child nodes' outcomes. Let $\mathcal{P} \subseteq \{1, 2, \ldots, P\}$ be the index subset of potential splitting variables for node η. The chosen splitting variable, X_p with $p \in \mathcal{P}$, is then determined by

$$p = \arg\max_{q \in \mathcal{P}} \Bigg\{ \sum_{i \in \eta} \Big(y_i - \bar{y}(\eta)\Big)^2 - \Big[\sum_{i \in \eta(S_q)} \Big((y_i - \bar{y}(S_q;\eta)\Big)^2 + \sum_{i \in \eta(S_q^c)} \Big(y_i - \bar{y}(S_q^c;\eta)\Big)^2 \Big] \Bigg\}. \tag{4.6}$$

This allows us to make comparisons between the outcome means of potential splitting variables. The following theorem provide certain necessary conditions for a variable to be used in a split.

Theorem 1. Let $\mathcal{P}$ and η be as defined above and for any $q \in \mathcal{P}$, let S_q be the subset such that $\bar{y}(S_q;\eta) \geq \bar{y}(\eta)$. If the splitting variable X_p with $p \in \mathcal{P}$ is obtained using (4.6) and for any fixed but arbitrary $q \in \mathcal{P}$, $\frac{|\eta(S_p)|}{|\eta(S_p^c)|} \leq \frac{|\eta(S_q)|}{|\eta(S_q^c)|}$, then

$$\bar{y}(S_p;\eta) \geq \bar{y}(S_q;\eta). \tag{4.7}$$

Conversely, if $\frac{|\eta(S_p)|}{|\eta(S_p^c)|} \geq \frac{|\eta(S_q)|}{|\eta(S_q^c)|}$,

$$\bar{y}(S_p^c;\eta) \leq \bar{y}(S_q^c;\eta). \tag{4.8}$$

In particular, if $\frac{|\eta(S_p)|}{|\eta(S_p^c)|} = \frac{|\eta(S_q)|}{|\eta(S_q^c)|}$, $\bar{y}(S_p^c;\eta) \leq \bar{y}(S_q^c;\eta) \leq \bar{y}(S_q;\eta) \leq \bar{y}(S_p;\eta)$.

Theorem 1 has the implication that, when $\bar{y}(S_p;\eta) > \bar{y}(\eta)$ at least one of $\bar{y}(S_p;\eta) \geq \bar{y}(S_q^*;\eta)$ or $\bar{y}(S_p^c;\eta) \leq \bar{y}(S_q^*;\eta)$ holds for any $q \in \mathcal{P}$ with S_q^* being either S_q or

S_q^c. It allows us to use outcome means of potential splitting variables as an alternative to expressions such as (4.6) to determine the best splitting variable and the best splitting point. The theorem is used in the next section to further analyze the properties of our proposed interaction measure.

Proof of Theorem 1. It is convenient to express (4.6) as

$$p = \arg\max_{q\in\mathcal{P}} \Big\{ |\eta(S_q)|\big(\bar{y}(S_q;\eta)\big)^2 + |\eta(S_q^c)|\big(\bar{y}(S_q^c;\eta)\big)^2 - |\eta|\big(\bar{y}(\eta)\big)^2 \Big\}. \tag{4.9}$$

Note that $|\eta|$ and $\bar{y}(\eta)$ do not depend on the choice of q, since they are the number of observations and outcome in the node where we evaluate alternative splitting variables. Therefore, (4.9) shows that the chosen variable X_p produces the largest sum of the child nodes outcome mean squares, weighted by their respective node sizes.

Now, (4.9) can also be written as

$$p = \arg\max_{r\in\mathcal{P}} \Big\{ |\eta| \frac{|\eta(S_q)|}{|\eta(S_q^c)|}\big(\bar{y}(S_q;\eta) - \bar{y}(\eta)\big)^2 \Big\}. \tag{4.10}$$

Since X_p is the chosen variable this implies that

$$\frac{|\eta(S_p)|}{|\eta(S_p^c)|}\big(\bar{y}(S_p;\eta) - \bar{y}(\eta)\big)^2 \geq \frac{|\eta(S_q)|}{|\eta(S_q^c)|}\big(\bar{y}(S_q;\eta) - \bar{y}(\eta)\big)^2, \qquad \forall q \in \mathcal{P}. \tag{4.11}$$

Let

$$R_q = \left(\frac{|\eta(S_q)|}{|\eta(S_q^c)|}\right)^{\frac{1}{2}}, \qquad \forall q \in \mathcal{P}.$$

Then, when (4.11) holds,

$$|\bar{y}(S_p;\eta) - \bar{y}(\eta)| \geq \frac{R_q}{R_p}|\bar{y}(S_q;\eta) - \bar{y}(\eta)|, \qquad \forall q \in \mathcal{P}. \tag{4.12}$$

But, by assumption, $\bar{y}(S_q;\eta) \geq \bar{y}(\eta)$ and, hence,

$$\bar{y}(S_p;\eta) \geq \bar{y}(S_q;\eta) + \left(\frac{R_q}{R_p} - 1\right)(\bar{y}(S_q;\eta) - \bar{y}(\eta)), \qquad \forall q \in \mathcal{P}, \tag{4.13}$$

giving (4.7) when $R_p \leq R_q$.

To show (4.8), observe that we can rewrite (4.10) in terms of S_p^c and S_q^c as

$$p = \arg\max_{r \in \mathcal{P}} \left\{ |\eta| \frac{|\eta(S_q^c)|}{|\eta(S_q)|} \big(\bar{y}(S_q^c; \eta) - \bar{y}(\eta) \big)^2 \right\}. \tag{4.14}$$

Thus, if $\bar{y}(S_q; \eta) \geq \bar{y}(\eta)$, it is also the case that $\bar{y}(S_q^c; \eta) \leq \bar{y}(\eta)$. An analogous argument for $\bar{y}(S_q^c; \eta)$ leads to

$$\bar{y}(S_p^c; \eta) \leq \bar{y}(S_q^c; \eta) + \left(\frac{R_p}{R_q} - 1 \right) (\bar{y}(S_q^c; \eta) - \bar{y}(\eta)), \qquad \forall q \in \mathcal{P} \tag{4.15}$$

and (4.10) holds when $R_p \geq R_q$. Finally, when $R_p = R_q$ the result holds by combining (4.13) and (4.15). ■

The conditions on $\bar{y}(S_p; \eta)$ and $\bar{y}(S_p^c; \eta)$ are necessary for p to be chosen as the splitting variable but not sufficient. Inequalities (4.13) and (4.15) provide necessary and sufficient conditions that we can use to extract some additional information about the outcome mean corresponding to the optimal splitting variable. For example, (4.13) could be written as

$$\bar{y}(S_p; \eta) \geq \frac{R_q}{R_p} \bar{y}(S_q; \eta) + \left(1 - \frac{R_q}{R_p} \right) \bar{y}(\eta), \qquad \forall q \in \mathcal{P}. \tag{4.16}$$

Now if $\frac{|\eta(S_p)|}{|\eta(S_p^c)|} > \frac{|\eta(S_q)|}{|\eta(S_q^c)|}$ so that $R_p > R_q$, then $\bar{y}(S_p; \eta)$ is greater than the weighted average of $\bar{y}(S_q; \eta)$ and $\bar{y}(\eta)$, with weight $\frac{R_q}{R_p}$. When $\frac{R_q}{R_p}$ is close to 1, it is still possible to have $\bar{y}(S_p; \eta) \geq \bar{y}(S_q; \eta)$. On the other hand, if $R_p \gg R_q$, the value of $\bar{y}(S_q; \eta)$ has less influence on the optimization criterion because its weight in (4.16) is very small. What this means is that when the proportion of observations in the node $\eta(S_p)$ with respect to $\eta(S_p^c)$ is much larger than the corresponding proportion given by q, the variable X_p could be chosen even when $\bar{y}(S_q; \eta) \geq \bar{y}(S_p; \eta)$.

Analysis with a Linear Model Specification

In this section we specialize to an explicit form for F in (2.1); namely, a linear model with interactions. This allows us to further explore our measure's ability to idenfify interactions in a setting where the variable interactions take a specific, common form.

The model we will study has a continuous output variables and binary input variables. We restrict our analysis to the case of binary variables for two reasons. First, we are interested in understanding how the structure of the tree works in simple settings. The random forest algorithm handles two optimization problems at each stage of the iterative process; it chooses the best splitting variable among those randomly selected and determines the best split point for this variable. When dealing with binary variables, the split point is predetermined and the selection process is based only on one optimization problem. It is simpler to track the effect of one optimization problem and relate it to the structure of the tree. Second, the results obtained using variables with two categories are useful in a variety of settings; for example, in situations like those studied in Chapter 3.

We represent the two categories with $C_p = \{1, 0\}$, $S_p = \{1\}$, $S_p^c = \{0\}$ for $p = 1, \ldots, P$. A variable can only appear in a branch once and the length of a branch in the tree can be at most equal to the number of input variables in the data set. For example, let $P = 4$ and $H(S_3^c; \eta) = (S_1, S_4^c, S_2, S_3^c)$ in which case $\eta(S_3^c)$ is a terminal node. Variable X_1 splits the root node with the first subset of observations (right child node) being determined by those observations with $X_1 = 1$. Variable X_4 splits this subset of observations keeping those with $X_4 = 0$, and so on. The leaf of this branch is composed of observations with covariate vectors of the form $\mathbf{x} = (1, 1, 0, 0)$.

The model is now defined explicitely by

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} \beta_{pq} X_p X_q + \epsilon \tag{4.17}$$

for binary variables $X_1, \ldots, X_p$. As before, we use y_i and x_{ip} to denote the observed values of Y and X_p with associated random errors $\epsilon_1, \ldots, \epsilon_N$ that are independent copies of the random variable ϵ that is assumed to be $N(0, \sigma^2)$.

For any arbitrary but fixed tree that is constructed using data from model (4.17), we can now express the outcome mean in node η by means of the coefficients of the

model; namely,

$$\begin{aligned}
\bar{y}(\eta) &= \frac{1}{|\eta|} \sum_{i:(y_i,\mathbf{x}_i)\in\eta} y_i \\
&= \frac{1}{|\eta|} \sum_{i:(y_i,\mathbf{x}_i)\in\eta} \left(\beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} + \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} \beta_{pq} x_{ip} x_{iq} + \epsilon_i \right) \\
&= \frac{1}{|\eta|} \left(\sum_{i:(y_i,\mathbf{x}_i)\in\eta} \beta_0 + \sum_{p=1}^{P} \beta_p \sum_{i:(y_i,\mathbf{x}_i)\in\eta} x_{ip} + \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} \beta_{pq} \sum_{i:(y_i,\mathbf{x}_i)\in\eta} x_{ip} x_{iq} + \sum_{i:(y_i,\mathbf{x}_i)\in\eta} \epsilon_i \right) \\
&= \beta_0 + \sum_{p=1}^{P} \frac{|\eta(S_p)|}{|\eta|} \beta_p + \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} \frac{|\eta(S_p, S_q)|}{|\eta|} \beta_{pq} + \bar{\epsilon}(\eta), \qquad (4.18)
\end{aligned}$$

where $\bar{\epsilon}(\eta) = \frac{1}{|\eta|} \sum_{i:(y_i,\mathbf{x}_i)\in\eta} \epsilon_i$. Recall that if $S_q \in H(\eta)$, then $|\eta(S_q)| = |\eta|$ and $|\eta(S_q^c)| = 0$. Using this fact, we can rewrite the last equality in (4.18) as

$$\bar{y}(\eta) = \Delta_1(H(\eta)) + \Delta_2(H(\eta)) \qquad (4.19)$$

with

$$\Delta_1(H(\eta)) = \beta_0 + \sum_{p:S_p\in H(\eta)} \beta_p + \sum_{\substack{p,q:S_p,S_q\in H(\eta) \\ p<q}} \sum \beta_{pq}$$

and

$$\begin{aligned}
\Delta_2(H(\eta)) = &\sum_{p:S_p\notin H(\eta)} \frac{|\eta(S_p)|}{|\eta|} \beta_p + \sum_{\substack{p:S_p\in H(\eta) \\ q:S_q\notin H(\eta)}} \sum \frac{|\eta(S_q)|}{|\eta|} \beta_{pq} + \\
&+ \sum_{\substack{p,q:S_p,S_q\notin H(\eta) \\ p<q}} \sum \frac{|\eta(S_p, S_q)|}{|\eta|} \beta_{pq} + \bar{\epsilon}(\eta).
\end{aligned}$$

In (4.19) we have disaggregated each sum in the last equality in (4.18) based on its characteristics: the sums in $\Delta_1(H(\eta))$ correspond to those variables used for partitions in $H(\eta)$ while the sums in $\Delta_2(H(\eta))$, that we also refer to as off-sums, correspond to those with at least one variable not used for partitions in $H(\eta)$.

At this point it will be useful to consider an example of (4.19) for that simple case of $P = 4$ covariates. Figure 4.3 provides a graphical representation of the situation. Here, the nodes where all the covariates are included are easier to interpret. For example,

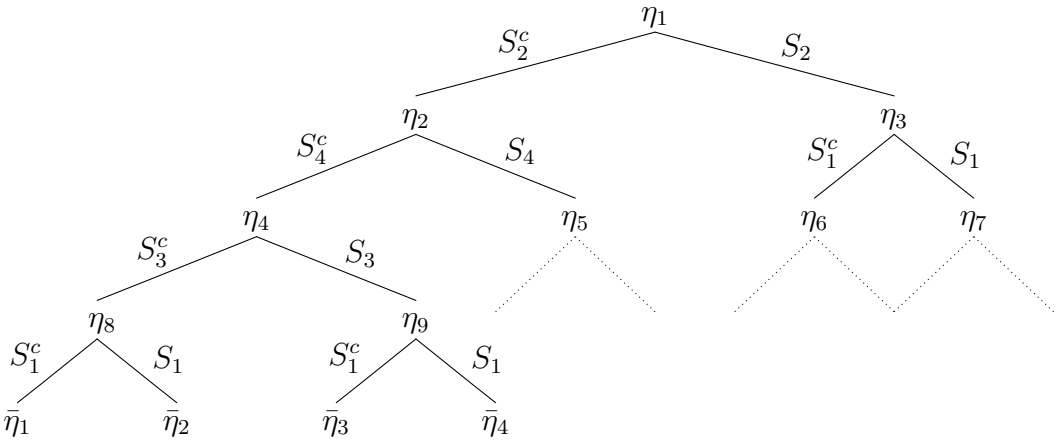

Figure 4.3: Partial outcome of a regression tree with four covariates. The set of observations at the terminal node $\bar{\eta}_1$ is obtained following the branch $H(\bar{\eta}_1) = (S_2^c, S_4^c, S_3^c, S_1^c)$. The resulting set of observations is a subset of $\mathcal{L}_N$ having the value $(0, 0, 0, 0)$ for $\mathbf{X}$.

$H(\bar{\eta}_1) = (S_2^c, S_4^c, S_3^c, S_1^c)$ fully describes the branch of the tree corresponding to the first terminal node. Observe that $H(\bar{\eta}_1) \equiv H(S_1; \eta_8)$.

Because all the input variables have been used in $H(\bar{\eta}_1)$ all the sums in $\Delta_2(\bar{y}(\bar{\eta}_1))$ are equal to zero. In addition, observe that

$$\{p : S_p^c \in H(S_1^c; \eta_1)\} = \{1, 2, 3, 4\} \text{ and } \{p : S_p \in H(\bar{\eta}_1)\} = \{p : S_p^c \notin H(S_1^c; \eta_1)\} = \emptyset,$$

i.e., the subsets of indices used for the sums in $\Delta_1(\bar{y}(\bar{\eta}_1))$ are empty sets because the observations included in $\bar{\eta}_1$ have all realized values for $X_p = 0$ for $p = 1, \ldots, 4$. Therefore, (4.19) reduces to

$$\bar{y}(\bar{\eta}_1) = \beta_0 + \bar{\epsilon}(\bar{\eta}_1).$$

Similarly, the $\bar{\eta}_4$ branch is given by $H(\bar{\eta}_4) = (S_2^c, S_4^c, S_3, S_1)$ and

$$\{p : S_p \in H(\bar{\eta}_4)\} = \{1, 3\}, \{p : S_p^c \in H(\bar{\eta}_4)\} = \{2, 4\}, \{p : S_p \notin H(\bar{\eta}_4)\} = \emptyset.$$

The outcome mean given by (4.19) in this case reduces to

$$\bar{y}(\bar{\eta}_4) = \beta_0 + \beta_1 + \beta_3 + \beta_{13} + \bar{\epsilon}(\bar{\eta}_4).$$

For branches that do not account for all the variables, the representation is more intricate. For example, for the η_7 branch, $H(\eta_7) = (S_2, S_1)$, we have

$$\{p : S_p \in H(\eta_7)\} = \{1, 2\}, \{p : S_p^c \in H(\eta_7)\} = \emptyset, \{p : S_p \notin H(\eta_7)\} = \{3, 4\}$$

By (4.19), the outcome mean is

$$\begin{aligned}
\bar{y}(\eta_7) &= \beta_0 + \beta_1 + \beta_2 + \beta_{12} + \frac{|\eta_7(S_3)|}{|\eta_7|}\beta_3 + \frac{|\eta_7(S_4)|}{|\eta_7|}\beta_4 + \\
&\quad + \frac{|\eta_7(S_3)|}{|\eta_7|}\beta_{13} + \frac{|\eta_7(S_3)|}{|\eta_7|}\beta_{23} + \frac{|\eta_7(S_4)|}{|\eta_7|}\beta_{14} \\
&\quad + \frac{|\eta_7(S_4)|}{|\eta_7|}\beta_{24} + \frac{|\eta_7(S_3, S_4)|}{|\eta_7|}\beta_{34} + \bar{\epsilon}(\eta_7) \\
&= \beta_0 + \beta_1 + \beta_2 + \beta_{12} + \frac{|\eta_7(S_3)|}{|\eta_7|}(\beta_3 + \beta_{13} + \beta_{23}) \\
&\quad + \frac{|\eta_7(S_4)|}{|\eta_7|}(\beta_4 + \beta_{14} + \beta_{24}) + \frac{|\eta_7(S_3, S_4)|}{|\eta_7|}\beta_{34} + \bar{\epsilon}(\eta_7). \qquad (4.20)
\end{aligned}$$

Let us now return to the interpretation of (4.19). Assume that a new observation is assigned to a terminal node $\bar{\eta}$ of a particular tree. Then $\bar{y}(\bar{\eta})$ is the tree's predicted value for the new observation's outcome. While the true parameter values for each β_p or β_{pq} for $p, q = 1, \ldots, N$ in (4.19) are, of course, unknown, the functional form for $\bar{y}(\bar{\eta})$ is exactly what we would expect from model (4.17); namely, the expected value for responses in $\bar{\eta}$. Now, as seen from our example, if $H(\bar{\eta})$ were to include every single input variable, the sums in $\Delta_2(H(\bar{\eta}))$ would be zero and the estimation would then be off from the true mean value only by the magnitude of $\bar{\epsilon}(\eta)$.

In contrast, if the number of partitions in $H(\bar{\eta})$ is not equal to P, the terms included in $\Delta_2(H(\bar{\eta}))$ would be nonzero and this would have two potential effects in the prediction. To see why this is so, let us work with a particular predictor X_p and assume that $S_p, S_p^c \notin H(\bar{\eta})$. If the new observation's value for X_p is 1, we would like for β_p to be accounted for in the prediction. However, it is possible that $|\bar{\eta}(S_p)| < |\bar{\eta}|$ and only a fraction of β_p would appear in the prediction. On the other hand, if the observed value is in 0, we would prefer for β_p not to be considered in the prediction and in this instance, it

is possible that $|\bar{\eta}(S_p)| > 0$. So a fraction of the value β_p would still be employed in the prediction. Both scenarios correspond to cases where the realizations of X_p for some of the observations in $\bar{\eta}$ are different than the observed value of X_p in the new observation.

The random forest predicted value at a given value $\mathbf{X} = \mathbf{x}$ is

$$F_\alpha(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \bar{y}(\bar{\eta}_k^t) \tag{4.21}$$

with η_k^t corresponding to the k-th terminal node in the t-th tree such that if $S_p \in H(\bar{\eta}_k^t)$ then $x_p = 1$ and if $S_p^c \in H(\bar{\eta}_k^t)$ then $x_p = 0$ for $p = 1, \ldots, P$. We know from our discussions in Chapter 2 that (4.21) provides accurate predictions even when the trees that appear in the average are restricted to a limited size and the sums in $\Delta_2(H(\bar{\eta}_k^t))$ for each tree are nonzero. Intuitively, we can provide some reasons why this result is possible. First, for a variable to split a node, it needs to be preselected and then chosen over the other preselected variables. As long as the trees are not too small, the most important effects would be included in $\Delta_1(H(\bar{\eta}_k^t))$ as soon as the corresponding variables are preselected in a node. Important variables mostly appear in $\Delta_2(H(\bar{\eta}_k^t))$ if they have not been preselected in any node. Second, $\Delta_2(H(\bar{\eta}_k^t))$ is a sum of fractions of effects that should have been considered entirely and fractions of effects that should not be considered at all. It is possible that the latter contributes to reduce the missing part in the former. Finally, observe that the larger the number of observations considered to obtain $\bar{\epsilon}(\bar{\eta}_k^t)$, the lower its variability. Therefore, there is a tradeoff between the magnitude of $\Delta_2(H(\bar{\eta}_k^t))$ and $\bar{\epsilon}(\bar{\eta}_k^t)$. If a tree is built allowing a large number of nodes, the number of coefficients only partially considered in any prediction is small; however, the prediction is made using only a few observations, affecting $\bar{\epsilon}(\bar{\eta}_k^t)$. In any event, if the random forest solution is adequate one must conclude that the average of $\Delta_2(H(\bar{\eta}_k^t))$ across trees does not adversely influence prediction in a substantial way.

The Distribution Based Measure in the Linear Model Specification

One motivation for the distribution-based interaction measure was that because the random forest prediction performs well when interactions among variables are present, the interactions should influence the structure of the constitutive trees in very specific ways that make it possible to identify interactions via the analysis of the trees' structure. We are now interested in providing evidence of the plausibility of this premise and to do so we will show that the structure of the trees in random forest provides evidence of interactions explicitely when the model specification is given by (4.17). It is clear in this situation that the existence and strength of the interaction between two variables, X_p and X_q, is entirely determined by β_{pq}. We will show that once X_p has been chosen as a splitting variable in a node and, as a consequence, has generated two subtrees from that node, the difference between choosing X_q as a splitting variable in one subtree or the other can be tied directly to the value of β_{pq}.

Figure 4.4 describes how variables X_p and X_q might relate to each other in a particular tree. The first node shown in the figure is η^o. The splitting variable in this case is X_p and the path from the root node to η^o is given by $H(\eta^o)$. This node partitions the overall tree into two subtrees, $\mathcal{T}\Big(\eta^o(S_p^c)\Big)$ and $\mathcal{T}\Big(\eta^o(S_p)\Big)$. In both subtrees, nodes are then shown where the other variable of interest, X_q, is selected as a splitting variable: at η' for the left subtree and at η for the right subtree. The broken lines represent different paths thereby suggesting potential asymmetries.

Thus, in terms of Figure (4.4) our immediate task is to find an expression where variable X_q is chosen among all preselected variables in η, and compare it with the corresponding representation when X_q is chosen among others in η'. To proceed in that direction observe that when a tree is formed, e.g., the one in Figure 4.4, subsets of variables are randomly selected and compared to determine the splitting variable in nodes η and η'. Let these subsets be described by index sets $\mathcal{P}$ and $\mathcal{P}'$, respectively. In the case

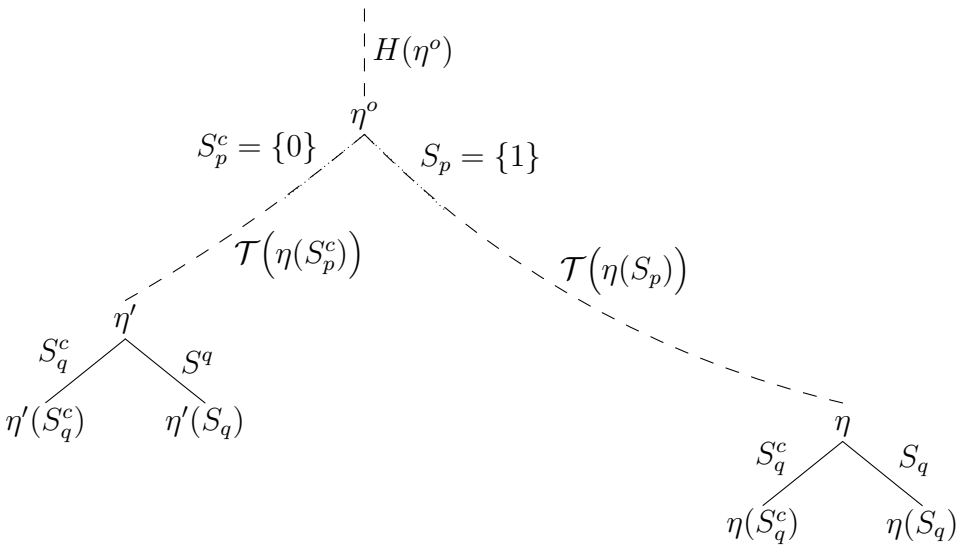

Figure 4.4: A section of a tree showing two variables and their relationship. The first node shown, η^o, has splitting variable X_p and the history of η^o's branch is given by $H(\eta^o)$. X_m partitions the data in two subtrees represented by $\mathcal{T}\Big(\eta^o(S_p^c)\Big)$ and $\mathcal{T}\Big(\eta^o(S_p)\Big)$. One node in each subtree indicated by η' and η represent points where X_q is the splitting variable.

of node η, let $q \in \mathcal{P}$ such that X_q is the chosen splitting variable in η with $\bar{y}(S_q;\eta) > \bar{y}(\eta)$ and assume that

$$\frac{|\eta(S_q)|}{|\eta(S_q^c)|} \approx \frac{|\eta(S_r)|}{|\eta(S_r^c)|},$$

for r in $\mathcal{P}$. Then, according to Theorem 1,

$$\bar{y}(S_q^c;\eta) \leq \bar{y}(S_r^*;\eta) \leq \bar{y}(S_q;\eta) \tag{4.22}$$

holds, where S_r^* is either S_r or S_r^c. Similarly, for node η', let $q \in \mathcal{P}'$ such that X_q is also the chosen splitting variable in η' with $\bar{y}(S_q;\eta') > \bar{y}(\eta')$ and

$$\frac{|\eta'(S_q)|}{|\eta'(S_q^c)|} \approx \frac{|\eta'(S_r)|}{|\eta'(S_r^c)|},$$

for r in $\mathcal{P}'$. Then

$$\bar{y}(S_q^c;\eta') \leq \bar{y}(S_r^*;\eta') \leq \bar{y}(S_q;\eta'). \tag{4.23}$$

In order to compare (4.22) and (4.23) we assume that η and η' are symmetrical with respect to X_p, $\mathcal{P} = \mathcal{P}'$ and the ratios of number of observations for symmetrical

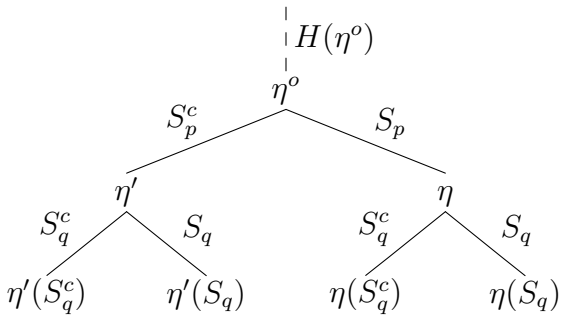

Figure 4.5: A symmetrical part of a tree consisting of two branches, both with the initial branch $H(\eta^o)$. In node η^o the splitting variable is X_p and the variable splitting both child nodes is X_q.

groups are similar, so that

$$\frac{|\eta(S_\ell, S_q)|}{|\eta(S_q)|} \approx \frac{|\eta'(S_\ell, S_q)|}{|\eta'(S_q)|},$$

for $\ell \in \mathcal{P}$. We say that η and η' are symmetrical when the only difference between the branches of these nodes is $S_p \in H(\eta)$ and $S_p^c \in H(\eta')$ or alternatively

$$\{\ell : S_\ell \in H(\eta) \backslash H(\eta')\} = \{\ell : S_\ell^c \in H(\eta') \backslash H(\eta)\} = \{p\}.$$

A simple example of symmetry is shown in Figure 4.5, when X_p splits the parent node and X_q splits both child nodes. Under these assumptions, comparison of (4.22) and (4.23) is tantamount to consideration of

$$[\bar{y}(S_q; \eta) - \bar{y}(S_r^*; \eta)] - [\bar{y}(S_q; \eta') - \bar{y}(S_r^*; \eta')]. \tag{4.24}$$

To show that (4.24) is a function of β_{pq} we express each $\bar{y}(\cdot;\cdot)$ as in (4.19) and start by comparing $\bar{y}(S_q; \eta)$ and $\bar{y}(S_r^*; \eta)$. Note that $H(S_q; \eta)$ represents an existent path in the tree while $H(S_r^*; \eta)$ is a hypothetical path that could have existed if X_r was chosen over X_q to split η. Rearranging and expanding those terms associated with X_q and X_r we

obtain

$$\begin{aligned}\bar{y}(S_q;\eta) =&\beta_0 + \sum_{\ell:S_\ell\in H(\eta)}\beta_\ell + \beta_q + \frac{|\eta(S_r,S_q)|}{|\eta(S_q)|}\beta_r + \\ &+ \sum_{\substack{\ell,m:S_\ell,S_m\in H(\eta)\\ \ell<m}}\sum\beta_{\ell m} + \sum_{\ell:S_\ell\in H(\eta)}\beta_{\ell q} + \frac{|\eta(S_r,S_q)|}{|\eta(S_q)|}\beta_{rq} + \sum_{\ell:S_\ell\in H(\eta)}\frac{|\eta(S_r,S_q)|}{|\eta(S_q)|}\beta_{\ell r} \\ &+ \sum_{\ell:S_\ell\notin H(r;\eta)\cup\{S_r\}}\left[\frac{|\eta(S_\ell,S_q)|}{|\eta(S_q)|}\beta_{\ell q} + \frac{|\eta(S_\ell,S_r,S_q)|}{|\eta(S_q)|}\beta_{\ell r}\right] \\ &+ \Delta\Big(H(S_q;\eta)\cup\{S_r\}\Big), \qquad \forall r\in\mathcal{P}, \end{aligned} \tag{4.25}$$

where

$$\begin{aligned}\Delta\Big(H(S_q;\eta)\cup\{S_r\}\Big) = &\sum_{\ell:S_\ell\notin H(S_q;\eta)\cup\{S_r\}}\frac{|\eta(S_\ell,S_q)|}{|\eta(S_q)|}\beta_\ell + \sum_{\substack{S_\ell\in H(\eta)\\ S_m\notin H(S_q;\eta)\cup\{S_r\}}}\sum\frac{|\eta(S_m,S_q)|}{|\eta(S_q)|}\beta_{\ell m} + \\ &+ \sum_{\substack{S_\ell,S_m\notin H(S_q;\eta)\cup\{S_r\}\\ \ell<m}}\sum\frac{|\eta(S_\ell,S_m,S_q)|}{|\eta(S_q)|}\beta_{\ell m} + \bar{\epsilon}(S_q;\eta).\end{aligned}$$

A similar expression holds for $\bar{y}(S_r^*;\eta)$. However, if $S_r^* = S_r^c$, the terms containing β_r and $\beta_{r\ell}$ are zero. We account for that with an indicator function and multiply such terms by $I(S_r^* = S_r)$ that is "1" only when $S_r^* = S_r$. In addition, since $S_p \in H(\eta)$, we want to explicitely display the terms associated with X_p. After cancelling common terms and rearranging some expressions, we obtain

$$\begin{aligned}\bar{y}(S_q;\eta) - \bar{y}(S_r^*;\eta) \;=& \left(1 - I(S_r^* = S_r)\frac{|\eta(S_q,S_r^*)|}{|\eta(S_r^*)|}\right)\left(\beta_q + \beta_{pq} + \sum_{\ell:S_\ell\in H(\eta)\setminus\{S_p\}}\beta_{\ell q}\right) \\ &- \left(I(S_r^* = S_r) - \frac{|\eta(S_r,S_q)|}{|\eta(S_q)|}\right)\left(\beta_r + \beta_{rp} + \sum_{\ell:S_\ell\in H(\eta)\setminus\{S_p\}}\beta_{\ell r}\right) \\ &+ \sum_{S_\ell\notin H(S_q;\eta)\cup\{S_r\}}\left(\frac{|\eta(S_\ell,S_q)|}{|\eta(S_q)|} - I(S_r^* = S_r)\frac{|\eta(S_\ell,S_q,S_r^*)|}{|\eta(S_r^*)|}\right)\beta_{\ell q} \\ &+ \sum_{S_\ell\notin H(S_q;\eta)\cup\{S_r\}}\left(I(S_r^* = S_r)\frac{|\eta(S_\ell,S_r^*)|}{|\eta(S_r^*)|} - \frac{|\eta(S_\ell,S_r,S_q)|}{|\eta(S_q)|}\right)\beta_{\ell r} \\ &+ \Delta(H(S_q;\eta)\cup\{S_r\}) - \Delta(H(S_r^*;\eta)\cup\{S_q\}),\ \forall r\in\mathcal{P}, \end{aligned} \tag{4.26}$$

where the set of subindices $\{r : S_r \in H(\eta)\setminus\{S_p\}\}$ accounts for each subindex in $\{r : S_r \in H(\eta)\}$ except p. To simplify the analysis we subsequently assume that the

remainder term

$$\Delta(H(S_q;\eta)\cup\{S_r\}) - \Delta(H(S_r^*;\eta)\cup\{S_q\})$$

is negligible and can be ignored.

The analogous representation of the differences in outcome means in node η' is similar. However, recall that $S_p^c \in H(\eta')$. That is, the variable X_p partitions some node in branch $H(\eta')$ keeping the observations with $X_p \in S_p^c$. With this in mind we obtain

$$\begin{aligned}
\bar{y}(S_q;\eta')-\bar{y}(S_r^*;\eta') \;&= \left(1 - I(S_r^* = S_r)\frac{|\eta'(S_q,S_r^*)|}{|\eta'(S_r^*)|}\right)\left(\beta_q + \sum_{\ell:S_\ell\in H(\eta')\setminus\{S_p\}}\beta_{\ell q}\right)\\
&\quad - \left(I(S_r^* = S_r) - \frac{|\eta'(S_r,S_q)|}{|\eta'(S_q)|}\right)\left(\beta_r + \sum_{\ell:S_\ell\in H(\eta')\setminus\{S_p\}}\beta_{\ell r}\right)\\
&\quad + \sum_{\ell:S_\ell\notin H(S_q;\eta')\cup\{S_r\}}\left(\frac{|\eta'(S_\ell,S_q)|}{|\eta'(S_q)|} - I(S_r^* = S_r)\frac{|\eta'(S_\ell,S_q,S_r^*)|}{|\eta'(S_r^*)|}\right)\beta_{\ell q}\\
&\quad + \sum_{\ell:S_\ell\notin H(S_q;\eta')\cup\{S_r\}}\left(I(S_r^* = S_r)\frac{|\eta'(S_\ell,S_r^*)|}{|\eta'(S_r^*)|} - \frac{|\eta'(S_\ell,S_r,S_q)|}{|\eta'(S_q)|}\right)\beta_{\ell r}\\
&\quad + \Delta(H(S_q;\eta')\cup\{S_r\}) - \Delta(H(S_r^*;\eta')\cup\{S_q\}),\ \forall r\in\mathcal{P}'. \qquad (4.27)
\end{aligned}$$

As we did in (4.26) we will assume that the remainder term

$$\Delta(H(S_q;\eta')\cup\{S_r\}) - \Delta(H(S_r^*;\eta')\cup\{S_q\})$$

can be ignored.

The difference between (4.26) and (4.27) provides the representation for (4.24) and is given by

$$\begin{aligned}
&[\bar{y}(S_q;\eta) - \bar{y}(S_r^*;\eta)] - [\bar{y}(S_q;\eta') - \bar{y}(S_r^*;\eta')] =\\
&\quad = \left(1 - I(S_r^* = S_r)\frac{|\eta(S_q,S_r^*)|}{|\eta(S_r^*)|}\right)\beta_{pq} - \left(I(S_r^* = S_r) - \frac{|\eta(S_r,S_q)|}{|\eta(S_q)|}\right)\beta_{rp}, \qquad (4.28)
\end{aligned}$$

for all $r\in\mathcal{P}$. Clearly (4.24) is a function of β_{pq}, as we wanted to show. Observe that (4.24) is also a function of β_{rp}. However, a different β_{rp} is considered for each choice of $r\in\mathcal{P}$. Hence, (4.24) will be consistently positive or negative only due to β_{pq}.

The Mean-Based Interaction Measure

The outcome mean representations given by (4.28) allow us to discern the influence of the interaction effect between X_p and X_q. This result serves as motivation to propose a second measure that is based on linear combinations of outcome means at nodes that are affected by X_p and X_q.

We call our new measure the *mean-based* interaction measure for X_p and X_q. It is a weighted sum of the difference of differences of the outcome means, where the weights take into consideration the existence of symmetrical sections and the location of X_q (and therefore X_p) from the root node. The mean-based interaction measure is defined as

$$\begin{aligned}\Omega(p,q) = &\frac{1}{\omega_1}\sum_{t=1}^{T}\sum_{k=1}^{D_t} I(\phi(\eta_k^t) = X_q) I(S_p \in H(\eta_k^t))\frac{\delta(p,q,\eta_k^t)}{\mu(\eta_k^t)}(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)) \\ &- \frac{1}{\omega_2}\sum_{t=1}^{T}\sum_{k=1}^{D_t} I(\phi(\eta_k^t) = X_q) I(S_p^c \in H(\eta_k^t))\frac{\delta(p,q,\eta_k^t)}{\mu(\eta_k^t)}(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)), \qquad (4.29)\end{aligned}$$

where

$$\omega_1 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big) I\Big(S_p \in H(\eta_k^t)\Big)\frac{\delta\Big(p,q,\eta_k^t\Big)}{\mu(\eta_k^t)},$$

and

$$\omega_2 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big) I\Big(S_p^c \in H(\eta_k^t)\Big)\frac{\delta\Big(p,q,\eta_k^t\Big)}{\mu(\eta_k^t)}.$$

The value of ω_1 is the weight of all the nodes that split with $X_p = 1$ first and then X_q, while ω_2 is the corresponding value of nodes that split with $X_p = 0$ first and then X_q. The value of $\delta(p,q,\eta_k^t)$ for each node η_k^t and each tree t, is a weight that represents the distance between X_p and X_q. Similarly, $\mu(\eta_k^t)$ represents the weight of node η_k^t and depends on its relative location. The value of $\delta(p,q,\eta_k^t)$ and $\mu(\eta_k^t)$ could also depend on the existence of symmetry of $H(\eta_k^t)$ with any other branch in tree t or any other tree.

An estimator for β_{pq} is

$$\tilde{\beta}(p,q) = \frac{1}{2}\Big(\Omega(p,q) + \Omega(q,p)\Big), \qquad (4.30)$$

i.e., the value of (4.30) is the average of $\Omega(p,q)$ and $\Omega(q,p)$, where $\Omega(p,q)$ provides an estimate for β_{pq} when X_p splits a node in a branch before X_q and $\Omega(q,p)$ provides that estimate when X_q splits a node in a branch before X_p.

It is noteworthy that (4.29) is a special case of (4.5) with

$$\gamma(p,q,\eta_k^t) = \frac{\delta\big(p,q,\eta_k^t\big)}{\mu(\eta_k^t)}\Big(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)\Big). \tag{4.31}$$

Although the distribution-based measure in (4.5) provides only an interaction identification measure in the general case, the connection between (4.5) and (4.30) suggest the possibility that, given a model specification, an estimation measure could be obtained from (4.5).

Note that we have not explicitely defined $\delta(\cdot,\cdot,\cdot)$ and $\mu(\cdot)$. As an illustration, if we let $\delta(\cdot,\cdot,\cdot) = 1$ and $d(\cdot) = 1$, (4.29) could be expressed as

$$\begin{aligned}\Omega(p,q) = &\frac{1}{\omega_1}\sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p \in H(\eta_k^t)\Big)\Big(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)\Big)\\ &- \frac{1}{\omega_2}\sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p^c \in H(\eta_k^t)\Big)\Big(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)\Big).\end{aligned} \tag{4.32}$$

Observe also that in general, when all the trees are added together, we would expect to have a comparable number of branches with $S_m \in H(\eta_k^t)$ as with $S_m^c \in H(\eta_k^t)$, and therefore $\omega_1 \approx \omega_2$. If this is the case, (4.32) is just the unweighted sum of difference of differences of outcome means.

We now show why (4.30) is an estimate of β_{pq}. We start the analysis with a symmetrical case such as the one presented in Figure 4.5: namely, X_p splits node η^o and X_q splits both child nodes, η and η'. We rearrange the terms in (4.19) and expand those terms associated with X_q to obtain

$$\bar{y}(S_q;\eta) = \beta_0 + \sum_{\ell:S_\ell \in H(\eta)} \beta_\ell + \beta_q + \sum_{\substack{\ell,m:S_\ell,S_m \in H(\eta)\\ \ell<m}}\sum \beta_{\ell m} + \sum_{\ell:S_\ell \in H(\eta)} \beta_{\ell q} + \Delta_2\big(H(S_q;\eta)\big), \tag{4.33}$$

with $\Delta_2(\cdot)$ as described in (4.19). Using a similar expression for $\bar{y}(S_q^c;\eta)$ and expanding the terms associated with X_p, the difference of $\bar{y}(S_q;\eta)$ and $\bar{y}(S_q^c;\eta)$ is given by

$$\bar{y}(S_q;\eta) - \bar{y}(S_q^c;\eta) = \beta_q + \sum_{\ell: S_\ell \in H(\eta)\backslash\{S_p\}} \beta_{\ell q} + \beta_{pq} + \Delta_2\Big(H(S_q;\eta)\Big) - \Delta_2\Big(H(S_q^c;\eta)\Big). \tag{4.34}$$

Similarly, for node η', we have

$$\bar{y}(S_q;\eta') - \bar{y}(S_q^c;\eta') = \beta_q + \sum_{\ell: S_\ell \in H(\eta')\backslash\{S_p\}} \beta_{\ell q} + \Delta_2\Big(H(S_q;\eta')\Big) - \Delta_2\Big(H(S_q^c;\eta')\Big). \tag{4.35}$$

Recall that $S_p^c \in H(\eta')$. In the symmetric case $H(\eta)\backslash\{S_p\} = H(\eta')\backslash\{S_p\}$ and the difference of (4.34) and (4.35) is given by

$$\begin{aligned}\Big[\bar{y}(S_q;\eta) - \bar{y}(S_q^c;\eta)\Big] - \Big[\bar{y}(S_q;\eta') - \bar{y}(S_q^c;\eta')\Big] = \\ = \beta_{pq} + \Delta_2\Big(H(S_q;\eta)\Big) - \Delta_2\Big(H(S_q^c;\eta)\Big) - \Delta_2\Big(H(S_q;\eta')\Big) + \Delta_2\Big(H(S_q^c;\eta')\Big).\end{aligned} \tag{4.36}$$

We can understand (4.36) as indicating that $\tilde{\beta}(p,q)$ is an estimator of β_{pq} with the difference of differences of the $\Delta_2(\cdot)$ terms as its bias. When considered in the context of a collection of trees, the difference of differences of the average of those $\Delta_2(\cdot)$ terms will determine the quality of our β_{pq} estimator in (4.30).

Our intuitive random forest argument suggest that the average of the off-sums in each $\Delta_2(\cdot)$ have a small influence in the outcome prediction in (4.19). However, the random forest prediction of an observation is always determined using one and only one terminal node from each tree: namely, the one that corresponds to the tree's prediction of that observation in (4.19). In contrast, the subset of nodes considered to find the random forest interaction estimation between two variables is comprised of all the nodes in each tree whose branches include the two variables from which the interaction is to be estimated. For binary variables there are four different partitions of realized values among

those two variables that can be realized; therefore, four different type of nodes can be part of this subset. When a node of any of these four types is part of this subset, none of its subsequent nodes (child nodes) in the rest of the tree are considered. For any given tree, any of the four partitions could be represented by none, one, or several nodes.

In other words, the effect of the off-sums inside the $\Delta_2(\cdot)$ terms in (4.36) is different than the analogous effect in (4.19), in the context of random forest. Hence, the average of each $\Delta_2(\cdot)$ term in (4.19) might not be as small as the average of $\Delta_2(\cdot)$ term in (4.36). Nevertheless, there are additional arguments that support the contention that the effect of the difference of difference of $\Delta_2(\cdot)$ terms in (4.36) is small in the context of random forest. In the symmetrical case, we have

$$\{\ell : S_\ell \notin H(S_q;\eta)\} = \{\ell : S_\ell \notin H(S_q^c;\eta)\} = \{\ell : S_\ell \notin H(S_q;\eta')\} = \{\ell : S_\ell \notin H(S_q^c;\eta')\}$$

which has the consequence that

$$\begin{aligned}
&\Delta_2\Big(H(S_q;\eta)\Big) - \Delta_2\Big(H(S_q^c;\eta)\Big) - \Delta_2\Big(H(S_q;\eta')\Big) + \Delta_2\Big(H(S_q^c;\eta')\Big) = \\
&\quad = \sum_{\ell:S_\ell\notin H(S_q;\eta)\}} \left(\frac{|\eta(S_\ell,S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_\ell,S_q^c)|}{|\eta(S_q^c)|} - \frac{|\eta'(S_\ell,S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_\ell,S_q^c)|}{|\eta'(S_q^c)|}\right)\beta_\ell \\
&\qquad + \sum_{\substack{S_\ell\in H(S_q;\eta)\\ S_m\notin H(S_q;\eta)}}\sum \left(\frac{|\eta(S_m,S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_m,S_q^c)|}{|\eta(S_q^c)|} - \frac{|\eta'(S_m,S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_m,S_q^c)|}{|\eta'(S_q^c)|}\right)\beta_{\ell m} \\
&\qquad + \sum_{\substack{S_\ell,S_m\notin H(S_q;\eta)\\ \ell<m}}\sum \left(\frac{|\eta(S_\ell,S_m,S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_\ell,S_m,S_q^c)|}{|\eta(S_q^c)|}\right. \\
&\qquad\qquad \left. - \frac{|\eta'(S_\ell,S_m,S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_\ell,S_m,S_q^c)|}{|\eta'(S_q^c)|}\right)\beta_{\ell m} \\
&\qquad + \bar{\epsilon}(S_q;\eta) - \bar{\epsilon}(S_q^c;\eta) - \bar{\epsilon}(S_q;\eta') + \bar{\epsilon}(S_q^c;\eta'). \qquad (4.37)
\end{aligned}$$

The sums involve coefficients of at least one variable not included in $H(S_q;\eta)$ or $H(S_q;\eta')$. First consider the influence of a variable X_ℓ for some $\ell : S_\ell \notin H(S_q;\eta)$. In that case, the coefficient multiplying β_ℓ in (4.37) is

$$\frac{|\eta(S_\ell,S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_\ell,S_q^c)|}{|\eta(S_q^c)|} - \frac{|\eta'(S_\ell,S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_\ell,S_q^c)|}{|\eta'(S_q^c)|}. \qquad (4.38)$$

For any given ℓ such that $S_\ell \notin H(S_q;\eta)$, the first term in the coefficient gives the proportion of observations in $\eta(S_q)$ with $X_\ell = 1$ while the second term provides the proportion of observations in $\eta(S_q^c)$ with $X_\ell = 1$. The only difference between the first and second terms is that the former has observations with $X_q = 1$, while the latter with $X_q = 0$. So, one would expect these proportions to be similar, unless there is a strong relationship between X_q and X_ℓ. And even then, the third and fourth terms in (4.38) are the proportion of observations in $\eta'(S_q)$ and $\eta'(S_q^c)$, respectively, with $X_\ell = 1$. If a strong relationship exists between X_q and X_ℓ, it should also be reflected in these two terms, making the difference of differences in (4.38) small.

Similarly, if X_p and X_ℓ are related, the number of observations in the first two terms in (4.38) might be very different from the number of observations in the last two terms. However, the difference of the first and second terms would be small, as would the difference between the third and fourth term.

An analogous argument can be made for the other two off-sums in (4.37). Notice however that the second summation is given by

$$\begin{aligned}
&\sum_{\substack{S_\ell \in H(S_q;\eta) \\ S_m \notin H(S_q;\eta)}} \sum \left(\frac{|\eta(S_m, S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_m, S_q^c)|}{|\eta(S_q^c)|} - \frac{|\eta'(S_m, S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_m, S_q^c)|}{|\eta'(S_q^c)|} \right) \beta_{\ell m} = \\
&= \sum_{\substack{S_\ell \in H(\eta) \\ S_m \notin H(S_q;\eta)}} \sum \left(\frac{|\eta(S_m, S_q)|}{|\eta(S_q)|} - \frac{|\eta(S_m, S_q^c)|}{|\eta(S_q^c)|} - \frac{|\eta'(S_m, S_q)|}{|\eta'(S_q)|} + \frac{|\eta'(S_m, S_q^c)|}{|\eta'(S_q^c)|} \right) \beta_{\ell m} \\
&\quad + \sum_{\ell : S_\ell \notin H(S_q;\eta)\}} \left(\frac{|\eta(S_\ell, S_q)|}{|\eta(S_q)|} - \frac{|\eta'(S_\ell, S_q)|}{|\eta'(S_q)|} \right) \beta_{\ell q}. \qquad (4.39)
\end{aligned}$$

For the factor in the last summation, the difference of only two ratios of observations are considered. This difference also contributes to make the effect of the last summation in (4.39) small. However, if there is a strong relationship between X_p and X_ℓ for some $\ell : S_\ell \notin H(S_q;\eta)$ this off-sum could have a bigger effect in (4.36), because there are no third and fourth terms to regulate this difference.

Finally, as we observed previously, the value of (4.37) seems to be the result of a trade off in the influence of the off-sums and the expression with error terms. Fewer elements in the set $\{\ell : S_\ell \notin H(S_q;\eta)\}$ imply fewer extra terms considered in (4.37); but, fewer observations are used to calculate the outcome means leading to more variability in the error terms expression. On the other hand, more elements in the set $\{\ell : S_\ell \notin H(S_q;\eta)\}$ imply more extra terms in (4.37) with a larger number of observations and less variability in the error terms expression.

Our analysis so far has assumed the symmetrical situation described in Figure 4.5. We have done this, because the random forest solution, on average, approaches the symmetrical case. However, we deem it important to also analyze additional characteristics in the asymmetrical scenario. If we can characterize the additional effects of asymmetries, we could try to adjust our measure to account for these effects.

In the asymmetrical case, the number of variables that partition the data and do not overlap in $H(\eta)$ and $H(\eta')$ is larger than $\{S_p\}$. The expression for the difference of differences of the corresponding outcome means is given by

$$\begin{aligned}\Big[\bar{y}(S_q;\eta) - \bar{y}(S_q^c;\eta)\Big] - \Big[\bar{y}(S_q;\eta') - \bar{y}(S_q^c;\eta')\Big] = \\ = \beta_{pq} + \sum_{p\in H(\eta)\setminus[H(\eta')\cup\{m\}]} \left(1 - \frac{|\eta'(p,r)|}{|\eta'(r)|}\right)\beta_{\ell q} \\ - \sum_{p\in H(\eta')\setminus H(\eta)} \left(1 - \frac{|\eta(p,r)|}{|\eta(r)|}\right)\beta_{\ell q} + \Delta_5, \end{aligned} \tag{4.40}$$

where Δ_5 has a larger number of different summation terms in comparison to (4.36) because fewer variables used in partitions for $H(\eta)$ and $H(\eta')$ overlap and a larger number of different summation terms do not cancel. Nevertheless, each term in each summation in Δ_5 is a linear combination of two ratios of observations (a difference) or a linear combination of four ratios of observations (a difference of differences). Hence the arguments presented for (4.37) and (4.39) are still valid here. The explicit expression in

(4.40) as well as the parallel expressions of (4.19), (4.33), and (4.36) for the asymmetrical case can be found in Appendix C.

In (4.40) we still are able to separate the effect of β_{pq}; however, we have additional terms that have a larger impact, those corresponding to the interactions of X_q with those other variables that do not overlap in $H(\eta)$ and $H(\eta')$. Observe that, the closer we are to the symmetrical case, the smaller influence these additional terms have. More specifically, the closer the variables X_p and X_q are from each other in node η (or η'), the smaller the number of terms in the first summation (or second summation) in (4.40). Similarly, in Δ_5, the closer X_p and X_q are to each other, the smaller will be the number of summations with terms composed by only two ratios (a difference), and the larger will be the number of summations with terms composed of four ratios (a difference of differences). In both situations we obtain a smaller net effect, but the former could be more sensitive to strong relationships.

4.2 Simulation Study

We now present the results of a simulation study that was conducted to evaluate the performance of the interaction measures that were introduced in the previous sections of this chapter. For that purpose we now consider several formulations for the mean-based interaction measure in (4.30) and the distribution-based interaction identification measure in (4.5).

The mean-based interaction measure involves the functions $\delta(\cdot,\cdot,\cdot)$ and $\mu(\cdot)$ that measure distance between variables and distance from the root node, respectively. Here we consider some specific choices for these weight functions and corresponding interaction measures that will be used in the simulation.

Our first choice for the mean-based measure weights is to use $\delta(\cdot,\cdot|\cdot) = 1$ and $\mu(\cdot) = 1$. With this choice, (4.29) becomes

$$\Omega_1(p,q) = \frac{1}{\omega_1}\sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p \in H(\eta_k^t)\Big)\Big(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)\Big)$$
$$- \frac{1}{\omega_2}\sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p^c \in H(\eta_k^t)\Big)\Big(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t)\Big), \qquad (4.41)$$

with

$$\omega_1 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p \in H(\eta_k^t)\Big)$$

and

$$\omega_2 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p^c \in H(\eta_k^t)\Big).$$

From $\Omega_1(p,q)$ we obtain the estimator

$$\tilde{\beta}_1(p,q) = \frac{1}{2}\Big(\Omega_1(p,q) + \Omega_1(q,p)\Big). \qquad (4.42)$$

The second choice again has $\delta(\cdot,\cdot,\cdot) = 1$ but now $\mu(\eta) = |H(\eta)|$, the number of nodes from η to the root node. This produces the measure

$$\Omega_2(p,q) = \frac{1}{\omega_1}\sum_{t=1}^{T}\sum_{k=1}^{D_t} I(\phi(\eta_k^t) = X_q)I(S_p \in H(\eta_k^t))\frac{(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t))}{|H(\eta_k^t)|}$$
$$- \frac{1}{\omega_2}\sum_{t=1}^{T}\sum_{k=1}^{D_t} I(\phi(\eta_k^t) = X_q)I(S_p^c \in H(\eta_k^t))\frac{(\bar{y}(S_q;\eta_k^t) - \bar{y}(S_q^c;\eta_k^t))}{|H(\eta_k^t)|}, \qquad (4.43)$$

$$\omega_1 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I(\phi(\eta_k^t) = X_q)I(S_p \in H(\eta_k^t))\frac{1}{|H(\eta_k^t)|},$$

and

$$\omega_2 = \sum_{t=1}^{T}\sum_{k=1}^{D_t} I\Big(\phi(\eta_k^t) = X_q\Big)I\Big(S_p^c \in H(\eta_k^t)\Big)\frac{1}{|H(\eta_k^t)|}.$$

with

$$\tilde{\beta}_2(p,q) = \frac{1}{2}\Big(\Omega_2(p,q) + \Omega_2(q,p)\Big). \qquad (4.44)$$

the estimator of β_{pq}. Yet another option is provided by taking $\mu(\eta) = |H(\eta)|$ as for Ω_2 but with

$$\delta(p, q, \eta_k^t) = \begin{cases} \frac{1}{|H(\eta_k^t)|}, & \text{if } \exists j : H(\eta_k^t) \backslash H(\eta_j^t) = \{S_p\} \text{ or } \{S_p^c\}, \\ 1, & \text{otherwise;} \end{cases} \tag{4.45}$$

i.e., $\delta(p, q, \eta) = |H(\eta)|^{-1}$ only if there is another node symmetric to η in the same tree. This produces the measure $\Omega_3(p, q)$ and corresponding estimator $\tilde{\beta}_3(p, q)$.

Our last choice for the mean-based measure weights uses $\mu(\eta) = |H(\eta)|$ and

$$\delta(p, q, \eta_k^t) = \begin{cases} \frac{1}{|H(\eta_k^t)|}, & \text{if } \exists j, r : H(\eta_k^t) \backslash H(\eta_j^r) = \{S_p\} \text{ or } \{S_p^c\}, j = 1, \ldots, D_r, r = 1, \ldots, T, \\ 1, & \text{otherwise.} \end{cases} \tag{4.46}$$

This weight considers a weaker condition of symmetry wherein $\delta(\cdot, \cdot, \cdot)$ takes into account whether the branches are symmetric not only in the same tree but anywhere in the collection of trees in the random forest. The resulting measure will be denoted by $\Omega_4(p, q)$ with $\tilde{\beta}_4(p, q)$ the associated estimator of β_{pq}.

Intuitively, our first mean-based measure representation could be understood as the unweighted mean difference of differences, while the weights in the other measure representations are given by the branch length (distance) for the second mean-based measure, and a combination of the branch length and symmetry for the third and fourth mean-based measures.

The distribution-based interaction measure involves the functions $\gamma(p, q, \eta), \omega_1$, and ω_2. We let

$$\gamma(p, q, \eta) = \frac{\delta(p, q, \eta)}{\mu(\eta)}$$

and $\omega_1 = \omega_2 = 1$ where both $\delta(\cdot, \cdot, \cdot)$ and $\mu(\cdot)$ will represent distance measures as in the case of the mean-based measure. For the distribution-based measure, our first option is to

use $\delta(\cdot,\cdot|\cdot)=1$ and $\mu(\cdot)=1$. Relation (4.5) simplifies to

$$\Gamma_1(p,q)=\sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t)=X_q\Big)I\Big(S_p\in H(\eta_k^t)\Big) - \sum_{t=1}^{T}\sum_{k_t=1}^{D_t} I\Big(\phi(\eta_k^t)=X_q\Big)I\Big(S_p^c\in H(\eta_k^t)\Big). \tag{4.47}$$

The second measure denoted by $\Gamma_2(p,q)$ uses

$$\delta(p,q,\eta)=\frac{1}{|H(\eta)\backslash H(\eta^o)|},$$

where $H(\eta^o)\subset H(\eta)$ and $\phi(\eta^o)=X_p$ with $\mu(\eta)=|H(\eta)|$. Finally, the third measure $\Gamma_3(p,q)$ employs $\mu(\eta)=|\eta|$ and $\delta(p,q,\eta)=|\eta^o|$.

Observe that all three distribution-based measures are obtained as the difference of two sums. Both sums account for all the nodes whose splitting variable is X_q. The branches of the nodes in the first sum include $X_p\in S_p$ while the branches in the second sum include $X_p\in S_p^c$. The $\Gamma_1(p,q)$ measure is the difference of two unweighted sums while the weights in $\Gamma_2(p,q)$ are the branch lengths and the distance from the node with splitting variable X_q to the node with splitting variable X_p. The weights in $\Gamma_3(p,q)$ are the number of observations in node η and the ratio between the number of observations in η^o and η where η^o is the node with splitting variable X_p.

Data Structure and Design

The data for the simulation study is generated from model (4.17) with $P=10,20$, or 40 and $N=400$. The binary variables are independently Bernoulli distributed random variables with success probability $\mathrm{P}(X_p\in S_p=\{1\})$ for variable X_p. The error terms were simulated as a random sample from the $N(0,1)$ distribution.

For the regression coefficients in (4.17) we considered three cases. The first two are simple constants: either $\beta_p=1$ for all $p=1,\ldots,P$, or $\beta_p=5$ for all $p=1,\ldots,P$. The third option has $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_P)$ obtained as a random sample with replacement from $(-5,-4,\ldots,4,5)$.

We considered interactions between X_1 and X_2, X_1 and X_5, and X_6 and X_7, respectively. Five levels were produced by taking. β_{pq} =0, 1, 5, 10, or 20, for (p, q) or (q, p) = (1,2), (1,5), and (6,7).

Two levels were used for the success probability. Either $\mathrm{P}(X_p = 1) = 0.5$ or $\mathrm{P}(X_p = 1) = 0.75$ for all $p = 1, \ldots, P$.

The tuning parameters for the random forest method are the number of (bootstrap) trees to be grown, the number of variables that are used at each node to determine a split and the number of terminal nodes for each tree. In this regard, the number of trees was taken to be either 500, 1000, or 2000, either 4, 5, or 7 potential splitting variables were used for the nodes in a tree, and the number of terminal nodes in each tree was set at 8, 16, 32, 64, or 128.

Procedures and Results

The full factorial design would yield a total of 4050 combinations. However, based on partial results and arguments that we present below, we have chosen unique factor levels for the number of trees to be grown, the number of random variables used at each node to determine the splitting variable, and the number of final nodes obtained at each tree. With these simplifications the experiment reduces to a factorial design yielding a total of 90 combinations, each one used to generate 100 replicates. Each interaction measure is then calculated. The results are presented in two parts. First, we discuss the selection process and rationale for both the mean-based and distribution-based interaction measures, as well as the arguments for the selection of parameters in the random forest algorithm. Second, the outcome of the simulation is summarized.

Selecting The mean-based Interaction Measure. As a preliminary step we analyzed the results for each replicate in our simulation and each factor combination to determine which of the four mean-based measures had the most desirable properties. To illustrate the idea, consider the information in Table 4.1 concerning the first mean-based measure.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**4.36**	-0.32	-0.37	**5.84**	0.88	-0.03	-0.10	0.79	-0.39
X2	**4.86**		0.54	1.04	0.10	0.19	-0.82	-0.81	0.11	0.32
X3	0.27	0.17		-0.59	-0.74	0.82	-0.14	0.31	0.11	0.24
X4	-0.62	1.17	0.68		0.37	0.29	-0.48	0.81	-0.26	0.10
X5	**4.18**	-0.95	1.28	0.07		-2.29	0.27	0.07	0.47	-0.50
X6	-0.30	-0.16	0.51	-0.27	-1.27		**4.34**	-0.08	-1.00	-1.19
X7	-0.33	-0.70	-0.06	-0.61	0.14	**5.97**		0.63	0.58	0.32
X8	-0.19	-0.10	-0.11	0.61	0.18	0.34	0.37		-0.33	0.58
X9	0.85	0.24	0.07	-0.26	-0.12	-0.76	0.44	-0.28		0.25
X10	-0.57	0.30	0.86	-0.16	-1.24	-1.81	-0.02	0.17	0.36	

Table 4.1: Values of $\Omega_1(p,q)$ for $p,q = 1,\ldots,10, p \neq q$, when the true values are $\beta_{pq} = 5$ for (p,q) = (1,2), (2,1), (1,5), (5,1), (6,7), and (7,6) and $\beta_{pq} = 0$ otherwise. The coordinates represent the estimated interaction between X_p and X_q. The row indicates which of the two variables appears first in the branch.

The results in the table are for data generated using model (4.17) with 10 variables, success probability $\mathrm{P}(X_p = 1) = .5$ for all $p = 1,\ldots,10$,

$$\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{10}) = (3,1,0,2,-4,-4,2,-2,-4,3)$$

obtained as a random sample with replacement from the vector $(-5,-4,\ldots,4,5)$, and $\beta_{pq} = 5$ for (p,q) or (q,p) = (1,2), (1,5), and (6,7). Using this data set, the random forest solution is obtained based on 1000 trees. In each tree, 4 covariates are randomly selected at every node to determine the splitting variable, and each tree could grow up to 32 terminal nodes. Once the trees are grown, the values of $\Omega_1(p,q)$ and $\Omega_1(q,p)$ were obtained. They are shown in Table 4.1. Recall that in (4.17) we assume that $\beta_{pq} = \beta_{qp}$ because we are interested in a unique interaction effect between X_p and X_q. The values presented in Table 4.1 correspond to (4.41) or (4.29). The interaction estimates, $\tilde{\beta}_1(p,q)$, are given by (4.30) as the average of $\Omega_1(p,q)$ and $\Omega_1(q,p)$. They are shown in Table 4.2. The values shown in boldface correspond to the estimates of β_{pq} for $(p,q) = (1,2),(1,5)$, and $(6,7)$ that estimate the true interaction coefficient $\beta_{pq} = 5$.

The corresponding results for $\tilde{\beta}_i(p,q), i = 2,3$, and 4 are presented in Tables 4.3, 4.4, and 4.5, respectively. From this we see that the estimates obtained with all four

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**4.61**	-0.02	-0.49	**5.01**	0.29	-0.18	-0.14	0.82	-0.48
X2			0.36	1.10	-0.43	0.01	-0.76	-0.46	0.17	0.31
X3				0.05	0.27	0.67	-0.10	0.10	0.09	0.55
X4					0.22	0.01	-0.54	0.71	-0.26	-0.03
X5						-1.78	0.20	0.12	0.17	-0.87
X6							**5.16**	0.13	-0.88	-1.50
X7								0.50	0.51	0.15
X8									-0.30	0.38
X9										0.30
X10										

Table 4.2: Values of $\tilde{\beta}_1(p,q)$ for $p = 1,\ldots,9, q = (p+1),\ldots,10$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. The coordinates represent the estimated interaction.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**4.63**	-0.06	-0.52	**5.08**	0.25	-0.22	-0.19	0.87	-0.46
X2			0.35	1.10	-0.42	0.04	-0.76	-0.52	0.17	0.31
X3				0.06	0.31	0.68	-0.11	0.09	0.08	0.54
X4					0.22	0.03	-0.57	0.71	-0.27	-0.02
X5						-1.79	0.22	0.14	0.20	-0.83
X6							**5.16**	0.14	-0.91	-1.54
X7								0.52	0.57	0.16
X8									-0.35	0.41
X9										0.28
X10										

Table 4.3: Values of $\tilde{\beta}_2(p,q)$ for $p = 1,\ldots,9, q = (p+1),\ldots,10$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. The coordinates represent the estimated interaction.

measures are very similar. However, these results correspond to a single replicate given a unique combination of factor levels. Figure 4.6 shows the boxplots for all 100 replicates for $\tilde{\beta}_1(p,q)$.

The boxplots for all 100 replicates for $\tilde{\beta}_i$, $i = 2, 3$, and 4 are similar to those presented in Figure 4.6. It is difficult to identify by simple inspection any relevant difference between all four mean-based interaction representations. We therefore introduce a notion of efficiency of our estimators. We compare the sum across all 100 replicates and

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**4.61**	-0.08	-0.46	**5.02**	0.29	-0.29	-0.11	0.86	-0.43
X2			0.53	1.11	-0.23	0.05	-0.75	-0.53	0.14	0.37
X3				0.31	0.50	0.58	-0.02	0.18	-0.33	0.23
X4					0.20	-0.00	-0.59	0.82	-0.25	-0.06
X5						-1.74	0.19	0.04	0.32	-0.78
X6							**5.06**	0.13	-0.92	-1.42
X7								0.55	0.51	0.16
X8									-0.23	0.43
X9										0.28
X10										

Table 4.4: Values of $\tilde{\beta}_3(p,q)$ for $p = 1,\ldots,9, q = (p+1),\ldots,10$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. The coordinates represent the estimated interaction.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**4.60**	0.13	-0.42	**5.03**	0.17	-0.21	-0.19	0.84	-0.47
X2			0.50	1.00	-0.37	-0.24	-0.83	-0.38	0.14	0.41
X3				0.60	0.65	1.32	-0.01	0.21	-0.19	0.29
X4					0.29	0.28	-0.61	0.72	-0.35	-0.04
X5						-1.99	0.15	0.11	0.23	-0.75
X6							**5.10**	0.06	-1.07	-1.31
X7								0.53	0.45	0.14
X8									-0.23	0.38
X9										0.34
X10										

Table 4.5: Values of $\tilde{\beta}_4(p,q)$ for $p = 1,\ldots,9, q = (p+1),\ldots,10$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. The coordinates represent the estimated interaction.

across all (p,q) for $p,q = 1,\ldots,P$, $p \neq q$ of the squared errors between the values obtained with $\tilde{\beta}_i(p,q)$ for $i = 1,\ldots,4$ and the true values β_{pq}, for each measure. For this particular combination of factor levels, the sums are 47.92, 45.98, 47.25, and 50.60, for the first, second, third, and fourth mean-based measures, respectively. The double sum for the second mean-based measure is slightly smaller than for the other three measures.

When this same calculation is carried out for each combination of factor levels in the study, we found that in 78.15% of the cases the second mean-based measure had the

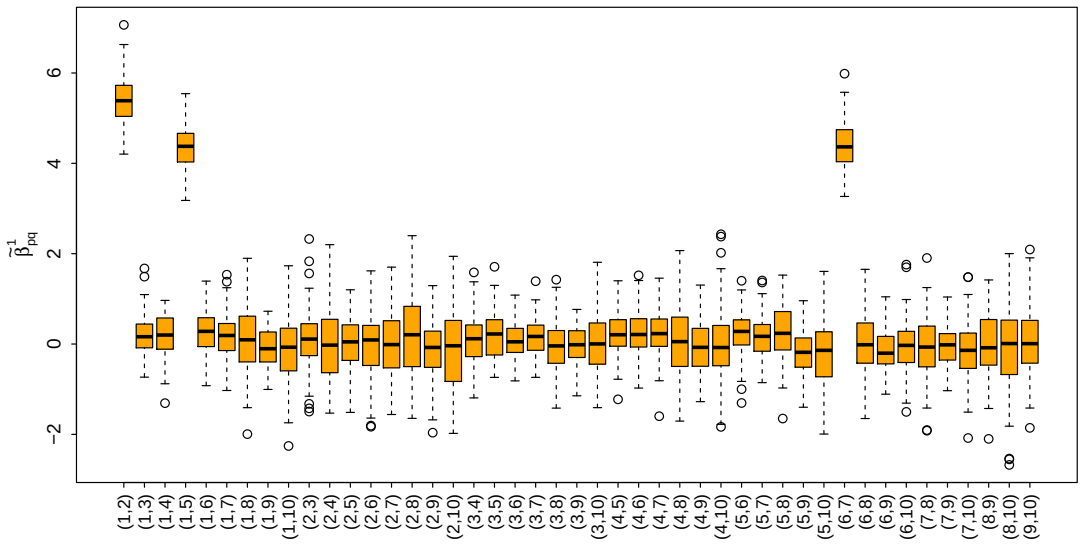

Figure 4.6: Boxplots of $\tilde{\beta}_1(p,q)$ for $p = 1,\ldots,9, q = (p+1),\ldots,10$. The true interaction values are $\beta_{pq} = 5$ for $(p,q) = (1,2),(1,5)$, and $(6,7)$ and $\beta_{pq} = 0$ otherwise, $P = 10$, β_p is sampled from $(-5,\ldots,5)$, and $P(X_p = 1) = .5$ for all $p = 1,\ldots,10$.

smallest sum of squared errors. Similarly, in 15.19%, 6.30%, and 0.04% of the cases, the third, first, and fourth measures had the smallest sum of squared errors. Based on these considerations, the second mean-based interaction measure seems more effective and accordingly, we will focus subsequent discussions only in its direction.

Recall that the weights used in $\tilde{\beta}_2(p,q)$ were based exclusively on the branch length, i.e., the number of nodes between the second variable node and the root node. The weights considered in $\tilde{\beta}_3(p,q)$ and $\tilde{\beta}_4(p,q)$ also accounted for the existence of symmetric branches. Based on the simulation study results, it seems that accounting for symmetry with the weights given by (4.45) and (4.46) does not improve the interaction estimation. In what follows, we refer to $\tilde{\beta}_2(p,q)$ simply as $\tilde{\beta}(p,q)$.

Selecting The Distribution-Based Interaction Measure. The procedure we used to select from among our three distribution-based interaction measures is similar to what we

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**1069**	-43	-236	**-1307**	-75	122	88	5	309
X2	**46**		-11	313	-10	-123	-15	3	-7	-223
X3	15	13		12	-4	-16	-7	-19	8	-11
X4	-9	36	57		-20	-51	-44	-8	2	20
X5	**11**	-46	-27	37		109	-29	-16	-13	12
X6	-6	7	4	-74	25		**95**	-33	-33	40
X7	-64	42	91	16	130	**-710**		7	184	241
X8	32	-43	4	35	11	29	11		-9	-34
X9	43	7	11	42	-89	144	-22	76		-164
X10	11	-77	22	0	21	119	10	-37	-29	

Table 4.6: Values of $\Gamma_1(p,q)$ for $p,q = 1,\ldots,10, p \neq q$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. Numbers of large magnitude, positive or negative, provide evidence of interaction.

used for the mean-based measure. The distribution-based measure, rather than producing interaction estimates, are only designed to detect the presence of variable interaction.

Using the data for the same specific replication as before we produced the values of $\Gamma_1(p,q)$ for this case. These values are shown in Table 4.6. Analogous results for $\Gamma_2(p,q)$ and $\Gamma_3(p,q)$ are presented in Tables 4.7 and 4.8, respectively.

It is interesting to observe that the results are very different than the ones for the mean-based measures. In this respect, we are not surprised to see markedly different values for the two different measures. What might seem unexpected, however, is the lack of symmetry in each matrix of results of $\Gamma_j(\cdot,\cdot)$ for $j = 1,2,3$: i.e., the large difference between $\Gamma_j(p,q)$ and $\Gamma_j(q,p)$ for $j = 1,2,3$ particularly for those values that correspond to $\beta_{pq} \neq 0$. In addition, some of the results corresponding to pairs of variables with positive interaction are negative (coordinates (1,5) and (7,6) in Table 4.6).

Small values in the matrix that produce asymmetries or negative values are the result of the presence of error terms or off sums such as the ones in (4.37). These are actually consistent with the meaning of our measure and provide additional information that is relevant. Recall that both coordinates (p,q) and (q,p) measure the number of

branches with paths containing both variables X_p and X_q, where the row coordinate indicates the variable that appears first in the branch. Thus, large asymmetries indicate the presence of interactions, particularly when the interaction effect, β_{pq}, is much larger or acts in the opposite direction than the second direct effect. In addition, the sign of the value in the measure is not directly related to the sign of the interaction value corresponding to those coordinates. Rather, it indicates that there are more branches splitting with $X_p = 0$ somewhere in their path and ending in X_q than branches splitting with $X_p = 1$ in their path and ending in X_q. This behavior is present when the sign of β_{pq} is opposite to the sign of β_q. In our example, $\beta_{15} = 5$ and $\beta_5 = -4$.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**236.3**	-3.1	-33.5	**-244.5**	-2.4	8.9	36.0	-46.1	44.8
X2	**7.1**		-1.9	45.3	-2.2	-11.8	-15.2	10.7	-5.9	-28.1
X3	2.8	2.7		2.5	-1.2	-2.2	-1.1	-3.7	0.6	-1.6
X4	-3.1	8.8	9.8		-2.2	-7.7	-11.3	-1.8	0.3	4.9
X5	**3.8**	-8.3	-1.1	2.8		18.9	-1.7	-2.7	-9.0	1.4
X6	-4.3	0.2	1.7	-13.9	5.7		**18.5**	-8.3	-7.2	9.8
X7	8.4	7.2	11.4	2.2	17.6	**-109.4**		5.6	13.1	40.6
X8	-1.4	-7.4	1.8	6.6	-4.1	2.2	5.5		3.9	-3.2
X9	0.3	-7.5	0.9	9.9	-19.5	22.1	3.0	15.5		-22.3
X10	0.5	-5.0	3.0	-2.8	4.4	22.1	6.2	-13.5	-10.8	

Table 4.7: Values of $\Gamma_2(p, q)$ for $p, q = 1, \ldots, 10, p \neq q$, when the true values are $\beta_{pq} = 5$ for (p, q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. Numbers of large magnitude, positive or negative, provide evidence of interaction.

Next, we select the distribution-based measure that performs the best. Unlike with the mean-based measures, where all the weighted mean difference of differences are adjusted to preserve the magnitude of the outcome mean, the distribution-based measures are entirely defined by weights. We purposely do not adjust the weights prior to comparison with all the pairs on variables in each random forest. As a consequence, it is not possible to directly determine which distribution-based measure is the most adequate. We do not have a reference to determine which numbers are more adequate than others for the values (p, q) where $\beta_{pq} \neq 0$. Therefore, we compare our measures only considering

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1		**72.03**	2.88	4.06	**-34.33**	5.42	20.35	18.81	5.21	23.21
X2	**5.16**		0.34	9.69	0.85	0.09	-0.16	5.08	2.86	-0.36
X3	0.26	0.47		0.20	0.02	-0.29	0.05	-0.28	0.05	-0.07
X4	-1.35	-0.45	0.50		-0.17	-1.89	-3.14	-1.27	-1.73	-0.28
X5	**1.05**	-1.20	-0.09	-0.12		2.87	0.65	-0.35	-2.07	-0.31
X6	-1.99	-1.01	-0.48	-2.65	-0.66		**1.60**	-2.42	-2.34	0.73
X7	12.10	11.04	3.04	3.56	3.60	**-13.42**		5.60	8.14	16.14
X8	-0.81	-1.64	0.32	0.76	-0.46	0.39	1.42		1.24	-0.45
X9	-1.59	-4.11	-0.21	2.11	-2.72	2.96	-0.00	2.83		-2.88
X10	0.29	0.58	0.24	0.01	1.32	3.59	1.52	-2.85	-1.65	

Table 4.8: Values of $\Gamma_3(p,q)$ for $p,q = 1,\ldots,10, p \neq q$, when the true values are $\beta_{pq} = 5$ for (p,q) =(1,2), (1,5), and (6,7) and $\beta_{pq} = 0$ otherwise. Numbers of large magnitude, positive or negative, provide evidence of interaction.

the pairs (p,q) with $\beta_{pq} = 0$. We standardize the measures and obtain the sum of squared errors. The sums corresponding to Tables 4.6 to 4.8 equal 0.45, 0.57, and 1.85, for the first, second, third distribution-based measures, respectively.

When this comparison is obtained for each combination of factor levels in the study, there is a weak preference for $\Gamma_2(p,q)$ over $\Gamma_1(p,q)$, while $\Gamma_3(p,q)$ is almost never chosen as the preferred option. As a result we decided to use only $\Gamma_2(p,q)$ in our subsequent empirical investigation. In what follows, we refer to $\Gamma_2(p,q)$ simply as $\Gamma(p,q)$.

Results for Different Combinations of Random Forest Conditions.

In terms of the results from our experiments, we found that the outcomes for 500, 1000 or 2000 trees were similar. When comparing measure performance between 500 and 1000 trees, in some instances more accuracy was achieved with 1000 trees when considering larger numbers of variables. By contrast, almost no improvement was observed when comparing results between 1000 and 2000 trees. Thus, in what follows, all the random forest results were obtained using 1000 trees.

Subsets of $4, 5$ and 7 potential splitting variables were considered based on the rule of thumb that the subsets of potential splitting variables for regression trees should be

approximately equal to the square root of the total number of variables. When studying our proposed interactions measures, slightly smaller numbers of subsets seem to work just as well and using subsets with $4, 5$ or 7 variables led to similar conclusions about the interaction measures. In what follows, we present an example where we compare results using subsets of 4 and 7 potential splitting variables. All the remaining results of the study use subsets of 4 potential splitting variables.

We first considered 8, 16, 32, 64, and 128 terminal nodes in each tree. The number of terminal nodes is one method to control for the tree size and therefore the complexity of each tree. While there are alternative methods that are employed to regulate tree size, using the number of terminal nodes is a natural option for our setting since it allows us to think in terms of the length of branches.

In principle, it is more convenient to account for a large number of terminal nodes in each tree, as it allows us to consider more branches with any pair of variables. Even when a random forest produces several trees, some of the variables might not be used in the first few nodes if the direct effect, β_p, and interaction effect, β_{pq}, are small relative to other variables effect. However, the number of potential splitting variables also determines how often variables with small effects could appear in the first few nodes. Hence, when using a small subset of splitting variables, it is possible to obtain accurate results even for trees with few terminal nodes, especially with multiple replicates.

A Specific Example. Before we present the results for the entire simulation study it will be useful to examine the application of our interaction measures in the context of a single replicate data set that was used in our experiments. This will allow us to see some of the effects of different choices for the random forest tuning parameters as well as what might be important to analyze when all the replicates are taken into account.

The data we will examine was obtained using 40 variables, success probability $\mathrm{P}(X_p = 1) = .5$, $\beta_p = 1$ for all $p = 1, \ldots, 40$, $\beta_{pq} = 20$ for (p, q) or $(q, p) =$ (1,2), (1,5), and

(6,7), and 1000 trees to obtain the random forest solution. We first describe the results for when the number of preselected variables at each node was either 4 or 7 while 8 terminal nodes were used in each tree.

We could try to analyze the results using tables or matrices such as Tables 4.1 to 4.8. However, a matrix of 40 rows and 40 columns has 1600 potential numbers to consider, and understanding the patterns of the interaction effects for all pairs of variables becomes challenging. It is more convenient to try to visualize these patterns and Figure 4.7 provides one possible graphical representation of the values in the matrices for $\Omega(p, q)$ and $\Gamma(p, q)$. The individual values for each coordinate are represented by colors based on a graded scale that has the largest numbers represented by bright yellow and the smallest by bright blue. This figure corresponds to a random forest solution with 4 preselected variables compared at each node and 8 terminal nodes.

Notice that in Figure 4.7, some of the cells in the mean-based interaction results (left matrix) are white in color which indicates that they are empty. This occurs because with 8 terminal nodes for each tree, only selected branches are generated and some combinations of variables do not appear in any of those branches.

Since the mean-based measure produces the estimated interaction effects for each pair of variables we would expect Figure 4.7 to consist of six bright yellow cells, each corresponding to those coordinates with $\beta_{pq} \neq 0$, while the rest of the cells with colors around zero. While this is certainly the case for those cells that correspond to nonzero interactions several other cells are also bright yellow or blue. These results are not unexpected and occur because not enough branches are provided by the trees to effectively identify and estimate the interactions. The mean-based measure is a weighted average of difference of differences of outcome means. To effectively estimate interactions two different types of branches are necessary: one connecting the first and second variables with the first variable equal to one, and the other branch connecting the first

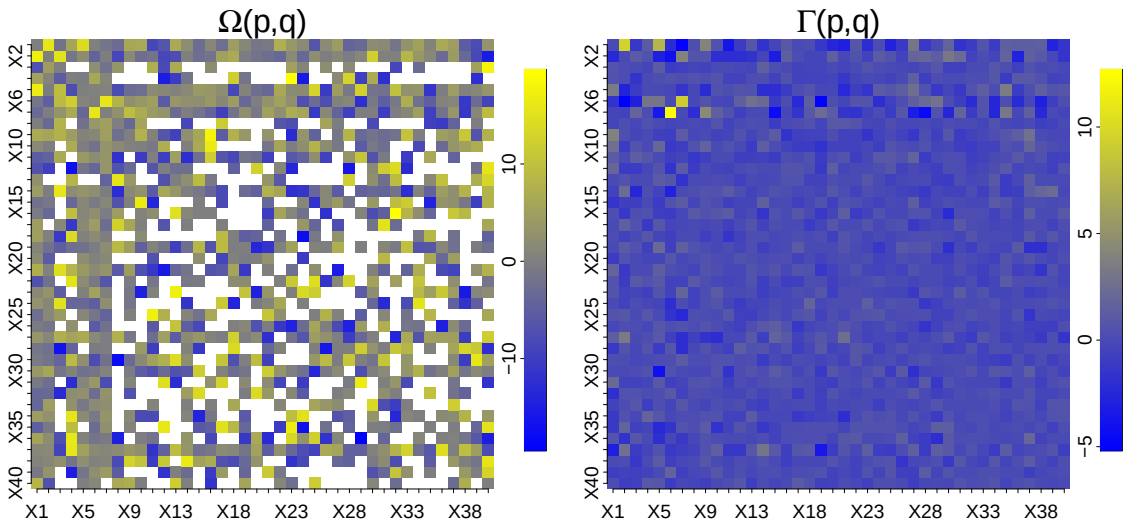

Figure 4.7: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ for $p,q = 1,\ldots,40$, for one replicate obtained from a random forest output with 1000 trees, 8 terminal nodes in each tree, and a subset of 4 potential splitting variables. The data set was generated using 40 variables, $\beta_p = 1$ for all $p = 1,\ldots,40$, $\beta_{pq} = 20$, for $(p,q) = (1,2),(1,5)$, and $(6,7)$, and $P(X_p = 1) = .5$. White cells in the left figure correspond to variable combinations that did not arise in the tree and are viewed as being empty.

and second variables with the first variable equal to zero. When the trees in a random forest do not produce any of these two branches, the resulting cell is empty. When only one branch is produced, the results are based only on one of the differences producing poor estimation results such as the ones presented in the left matrix of Figure 4.7.

In contrast to the mean-based estimator, the graphical representation of the matrix for the distribution-based measure does not contain any empty cells. Although the same branches are used here as in the mean-based case, the corresponding values for non-existent branches is zero. More importantly, the figure detects the presence of the specific interactions that are present in the data, at least in one of the corresponding coordinates for each interaction.

Figure 4.8 shows the analogous results for a case where the subset of potential splitting variables used is 7. The results are similar to those presented in Figure 4.7. In the matrix of $\Omega(p,q)$ values, the number of empty cells is larger than in Figure 4.7. Since

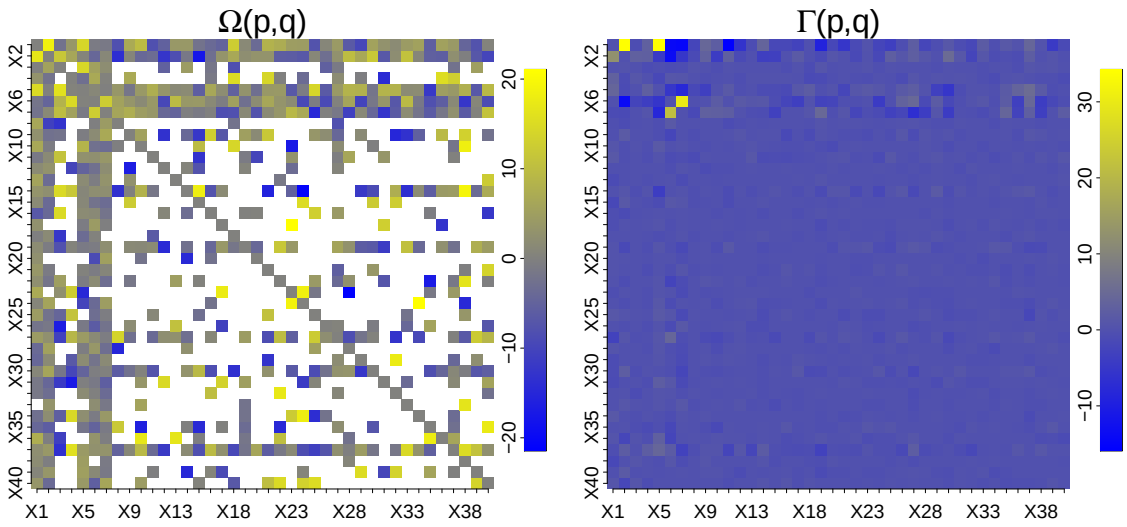

Figure 4.8: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ for one replicate obtained from a random forest output with 1000 trees, 8 terminal nodes in each tree, and a subset of 7 potential splitting variables. The data set was generated using 40 variables, $\beta_p = 1$ for all $p = 1, \ldots, 40$, $\beta_{pq} = 20$, for $(p,q) = (1,2), (1,5)$, and $(6,7)$, and $P(X_p = 1) = .5$.

more variables are used at each node to select the splitting variable, cases with a small direct effect are selected less often. The matrix of $\Gamma(p,q)$ values is similar to that in Figure 4.7. As expected, coordinates (1,2), (1,5), and (6,7) are bright yellow, and (2,1) and (7,6) are pale yellow. Although most other coordinates contain values that are around zero, there are few cells that show larger negative values such as in coordinates (1,6), (1,7) and (1,11). This can be attributed to a sort of interaction *spread* effect that occurs when an interaction effect is very large relative to the direct effects or other interaction effects as we now explain.

If an interaction exists between X_p and X_q, the number of branches with $X_p \in S_p$ and X_q is different than the number of branches with $X_p \in S_p^c$ and X_q. If the interaction effect is large relative to other interaction effects and direct effects, the proportion of these branches with respect to the total number of branches in the random forest is also high. Therefore, the difference in the number of branches will *spread* to other variables that do not interact with X_p but appear in those branches, thereby producing a difference in the number of branches containing X_p and those variables.

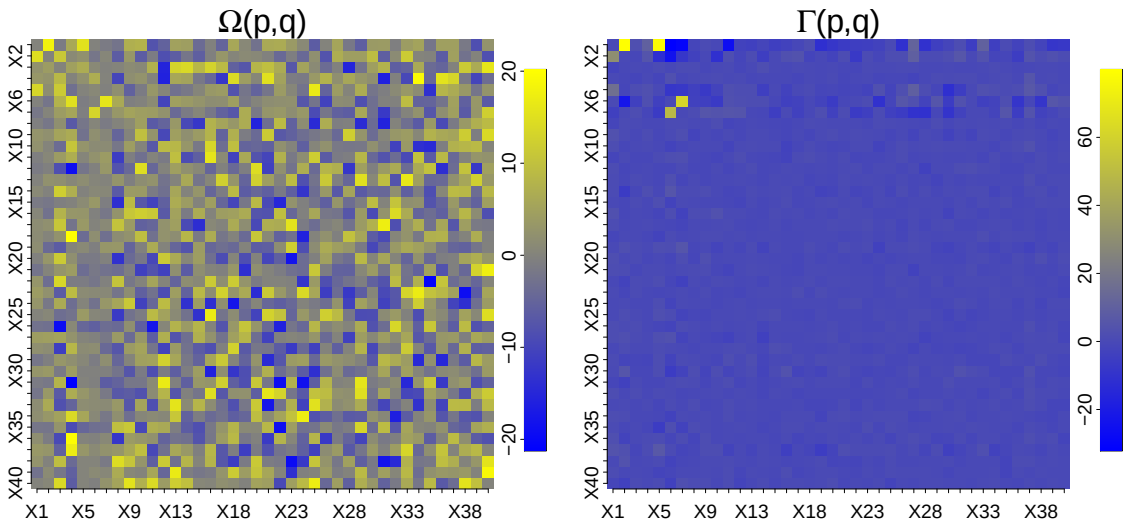

Figure 4.9: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ for one replicate obtained from a random forest output with 1000 trees, 32 terminal nodes in each tree, and a subset of 7 potential splitting variables. The data set was generated using 40 variables, $\beta_p = 1$ for all $p = 1, \ldots, 40$, $\beta_{pq} = 20$, for $(p,q) = (1,2), (1,5)$, and $(6,7)$, and $P(X_p = 1) = .5$.

On the other hand, the *spread* effect has no direct influence in the mean-based measure. The mean-based estimates for coordinates (1,6), (1,7) and (1,11) are all close to zero. We observed the same pattern in a number of other specific replicates that we examined individually. It is noteworthy to point out that the two measures seemed to complement each other in the sense that only potential interactions that were correctly predicted simultaneously by both measures were precisely those with true interaction effects.

Figure 4.9 depicts the corresponding mean-based and distribution-based interaction results when the subset of potential splitting variables is 7 and the number of terminal nodes is 32. Most of what we see in this case is similar to what was found in Figure 4.7 and 4.8. However, notice that when 32 terminal nodes are used, the matrix with mean-based values no longer contains empty cells. The values for coordinates (2,1) and (5,1) in the right hand figure are also more visible than before.

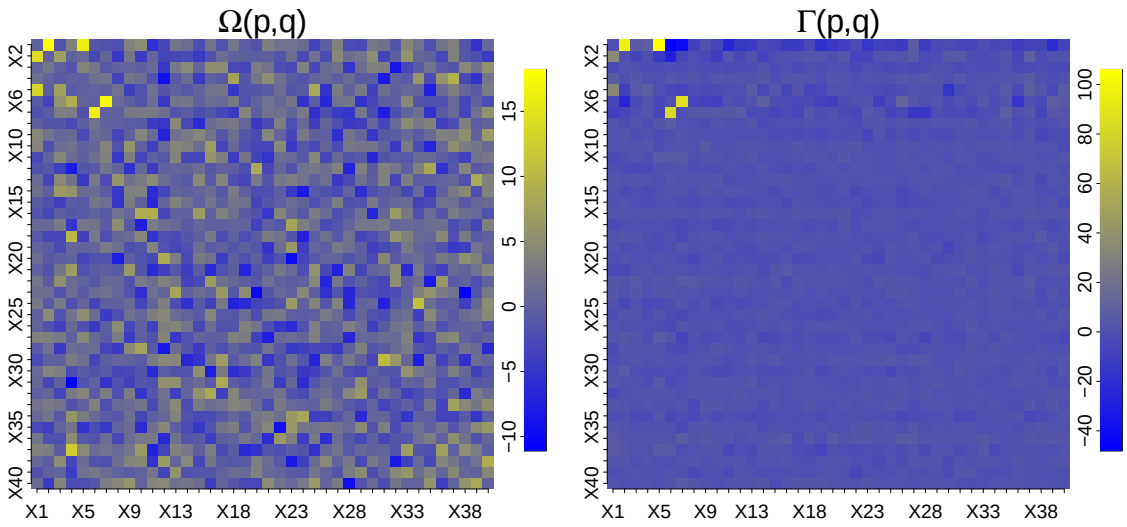

Figure 4.10: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ for one replicate obtained from a random forest output with 1000 trees, 128 terminal nodes in each tree, and a subset of 7 potential splitting variables. The data set was generated using 40 variables, $\beta_p = 1$ for all $p = 1, \ldots, 40$, $\beta_{pq} = 20$, for $(p,q) = (1,2), (1,5)$, and $(6,7)$, and $P(X_p = 1) = .5$.

Figure 4.10 shows the corresponding mean-based and distribution-based interaction results when the subset of potential splitting variables is 7 and the number of terminal nodes is 128. Although the conclusions obtained are similar to those for previous figures, it is worth noting that since the matrix of mean-based values now takes into account many more branches, fewer coordinates are producing false positives and the matrix is approaching the ideal representation.

At least in terms of this replication with these specific factor level combinations, the mean-based and distribution-based interaction measures seem to adequately determine both the presence of interactions and the estimated interaction effects. When considering both interaction measures simultaneously, even a small subset of potential splitting variables and a small number of terminal nodes produced potentially satisfactory results. However, when using a small number of terminal nodes, the results using only the mean-based estimates would be inadequate and similarly, the distribution-based measure by itself incorrectly identified interactions due to *spread* effects.

Results for Data Generated With Different Combinations of Factor Levels. In order to present results of the comprehensive study in a manageable manner, we examined first the patterns of the interaction measure results within each studied factor. Then, based on these patterns, we discuss in detail selected cases of interest. Additional results are presented in Appendix B. For what follows, we use the chosen mean-based and distribution-based interaction measures, random forest with 1000 trees, subsets of 4 potential splitting variables, and 8 terminal nodes.

For each factor combination, we obtain the average of 100 replicates for both the mean-based and distribution-based measures. To show variability, we report the range of the $\tilde{\beta}(p, q)$ averages for coordinates p, q with $\beta_{pq} \neq 0$ and the range for coordinates p, q with $\beta_{pq} = 0$. We view the range as being somewhat more informative in our setting.

We first analyze the impact that the magnitude of the true interaction effects has on our proposed measures. Figure 4.11 shows matrices corresponding to the mean of 100 replicates for the mean-based (left column) and distribution-based (right column) measures, when the interaction effects are $\beta_{pq} = 0, 1, 5$, and 20 for $(p, q) = (1, 2), (1, 5)$, and $(6, 7)$. In this instance, $P = 10$, $P(X_p = 1) = .5$, and $\beta_p = 1$ for $p = 1, \ldots, 10$. The measures perform well for these combinations of factor levels. In terms of the mean-based measure, when $\beta_{pq} = 0$, the matrices present no particular patterns with all the coordinate values around zero ranging from -0.05 to 0.03. When $\beta_{pq} \neq 0$, all three interaction effects are clearly identified, with the estimated mean-based interaction effects ranging from 0.70 to 0.82 when $\beta_{pq} = 1$, from 4.32 to 4.55 when $\beta_{pq} = 5$, and from 17.85 to 19.36 when $\beta_{pq} = 20$. Hence, in these situations, the mean-based estimator reflects the true values of the interaction effects but appears to be biased toward smaller values. The range of estimated values for coordinates without interaction effects is -0.03 to 0.18, -0.49 to 0.48, and -1.43 to 1.55, respectively. In terms of the distribution-based measure, the ranges for the identification values are -3.69 to 2.97 when $\beta_{pq} = 0$ for all p, q. When $\beta_{pq} \neq 0$, at least one relevant cell, (p, q) or (q, p), has values that range from 37.02 to

68.23 when $\beta_{pq} = 1$, 56.82 to 107.02 when $\beta_{pq} = 5$, and 57.63 to 106.75 when $\beta_{pq} = 20$. The range of values for coordinates without interaction effects is -30.08 to 1.12, -59.23 to 1.59, and -58.60 to 2.00, respectively. The large negative values correspond to the *spread* effect for most variables paired with X_1 as indicated by the light blue color of the corresponding cells. This effect is particularly strong for X_6 and X_7.

Figure 4.12 presents the output for data generated with 20 variables, but otherwise the same combination of factor levels as in Figure 4.11. The results are very similar. The measures perform well for these combinations of factor levels. For the mean-based measure, when $\beta_{pq} = 0$, the matrices present no particular patterns and all the cell values range from -0.15 to 0.11. When $\beta_{pq} \neq 0$, the interaction effects range from 0.66 to 0.74 when $\beta_{pq} = 1$, from 4.43 to 4.68 when $\beta_{pq} = 5$, and from 19.09 to 19.72 when $\beta_{pq} = 20$. Again, the estimated interaction effects are in the neighborhood of the true interaction effects, but appear to be biased toward smaller values. The ranges of estimated values for cells without interaction effects are -0.17 to 0.20, -0.58 to 0.58, and -1.83 to 2.24, respectively. For the distribution-based measure, the ranges for all cell values are -1.07 to 1.18 when $\beta_{pq} = 0$, while in the case of $\beta_{pq} \neq 0$, at least one relevant cell, (p, q) or (q, p), have values that range from 11.14 to 17.05 when $\beta_{pq} = 1$, 24.19 to 33.62 when $\beta_{pq} = 5$, and 25.42 to 34.71 when $\beta_{pq} = 20$. The range of values for cells without interaction effects is -2.71 to 0.72, -13.21 to 0.98, and -13.40 to 0.90, respectively. There is also a *spread* effect for most variables (particularly X_6 and X_7) paired with X_1 represented by the large negative values.

Figure 4.13 summarizes the output for data generated with 40 variables, but otherwise the same combinations of factor levels as found in Figures 4.11 and 4.12. The results are again similar. In terms of the mean-based measure, when $\beta_{pq} = 0$, the matrices present no particular patterns and all the cell values range from -0.27 to 0.35. When $\beta_{pq} \neq 0$, the estimated interaction effects range from 0.60 to 0.79 when $\beta_{pq} = 1$, from 4.14 to 4.63 when $\beta_{pq} = 5$, and from 19.19 to 20.02 when $\beta_{pq} = 20$. The bias toward smaller

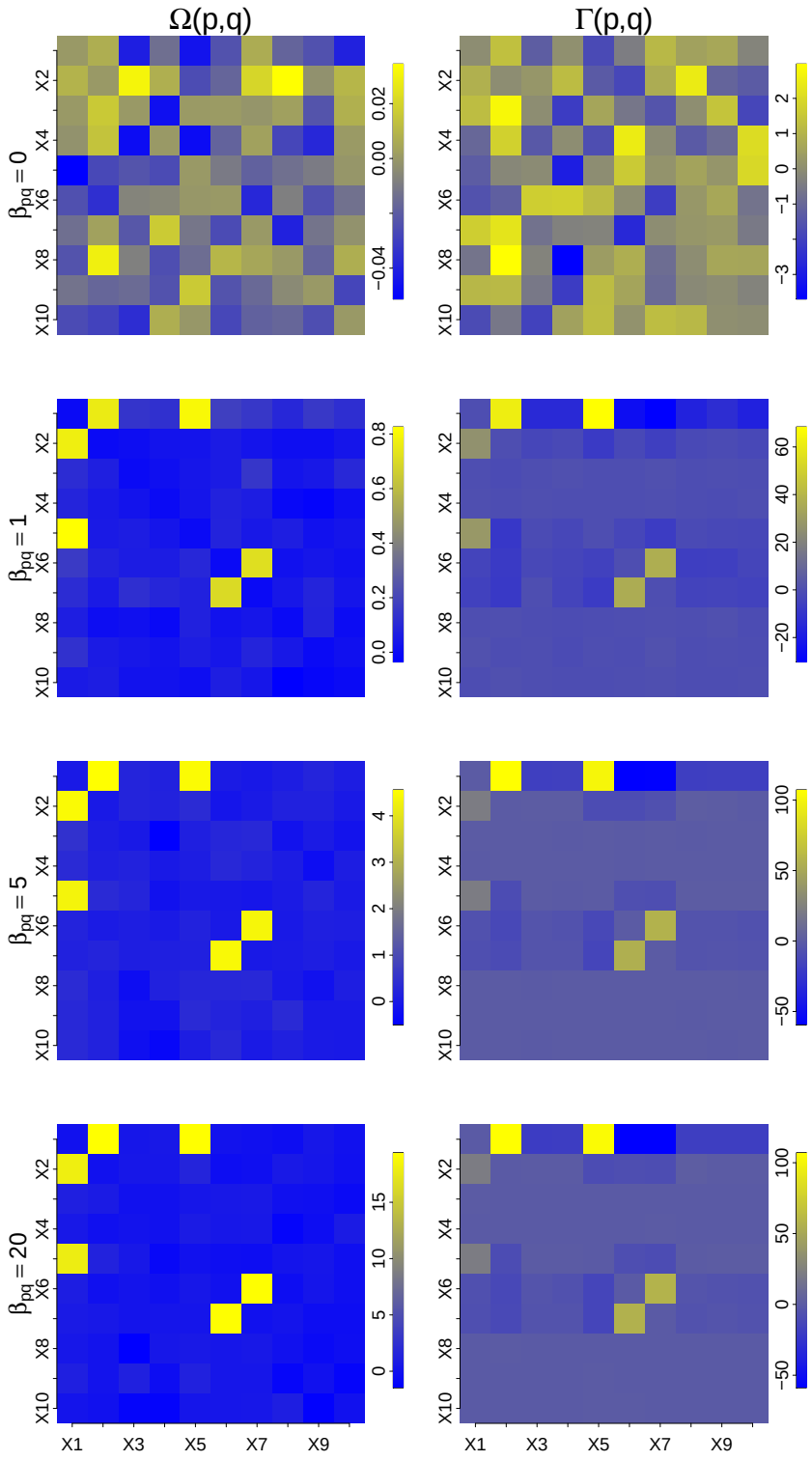

Figure 4.11: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 10$, $\beta_p = 1$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .5$.

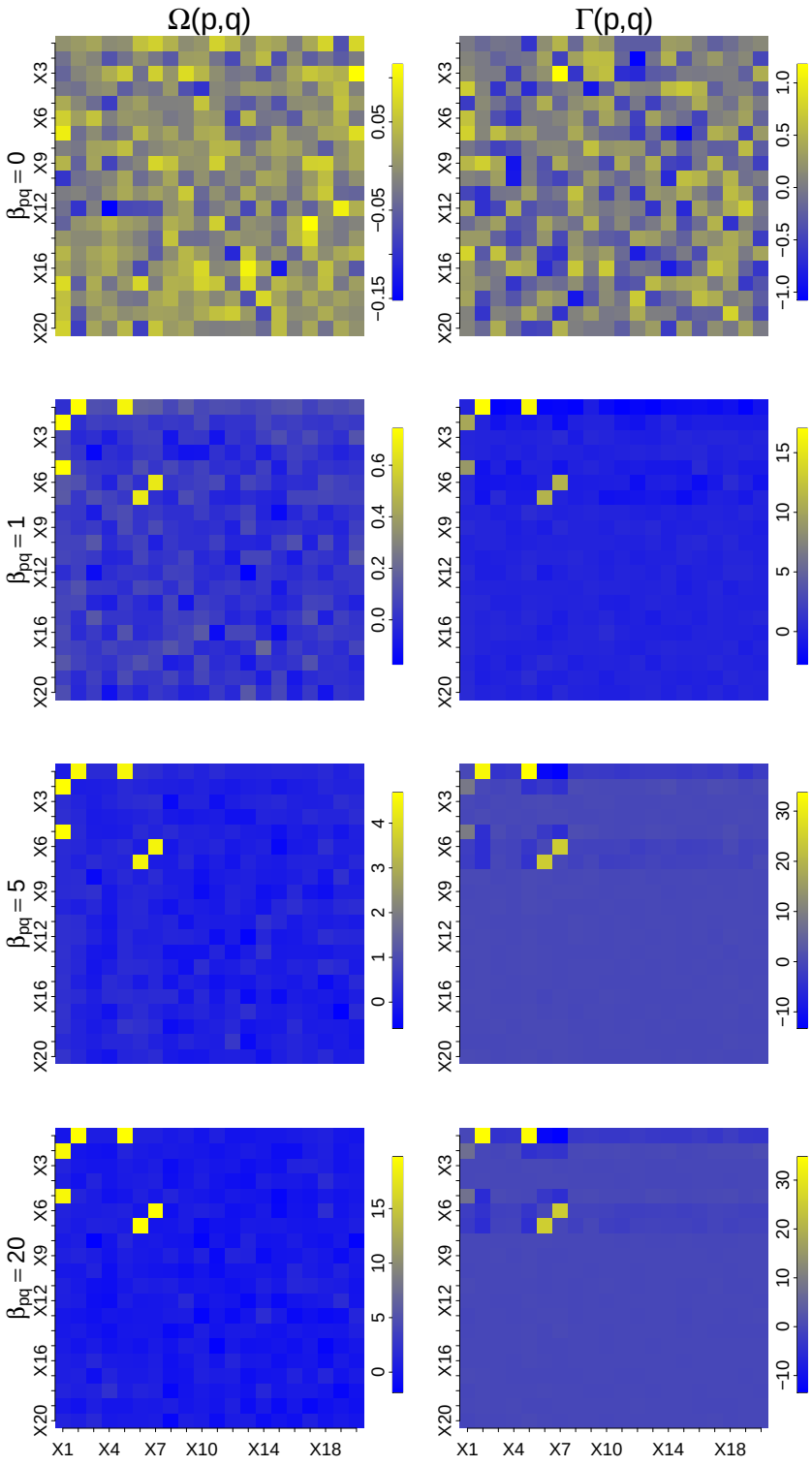

Figure 4.12: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, $\beta_p = 1$ for all $p = 1, \ldots, 20$, and $P(X_p = 1) = .5$.

values for the mean-based estimator continues for these factor combinations. The range of estimated values for coordinates without interaction effects is -0.42 to 0.41, -1.05 to 0.99, and -4.17 to 3.97, respectively. In terms of the distribution-based measure, the ranges of values when $\beta_{pq} = 0$, for all p, q, are -0.38 to 0.41. When $\beta_{pq} \neq 0$, at least one relevant cell, (p, q) or (q, p), has values that range from 2.01 to 3.83 when $\beta_{pq} = 1$, from 7.53 to 9.55 when $\beta_{pq} = 5$, and from 8.23 to 9.76 when $\beta_{pq} = 20$. The range of values for cells without interaction effects was -0.76 to 0.61, -3.06 to 0.47, and -2.71 to 0.48, respectively.

Figure 4.14 corresponds to the case with 40 variables and $\beta_p = 5$ for $p = 1, \ldots, 40$. Otherwise the data is generated using the same combinations of factors as the data presented in Figures 4.11 to 4.13. When $\beta_{pq} = 0$, the matrices present no particular patterns and all the mean-based cell values range from -1.42 to 1.54. When $\beta_{pq} = 1$, both measures failed to determine the presence of the interaction effects. The values in the mean-based matrix range from -1.66 to 1.52. When $\beta_{pq} = 5$, all three interaction effects are identified; however, the mean-based measure underestimates the true value with estimates ranging from 2.86 to 3.69, while the range of estimated values for coordinates without interaction effects is from -2.66 to 2.07. When $\beta_{pq} = 20$, the estimates of all three interactions measures are underestimated ranging from 16.51 to 18.87 for cells with interactions and from -4.72 to 5.03 for cells with no interactions. In terms of the distribution-based measure, when $\beta_{pq} \neq 0$, at least one relevant cell, (p, q) or (q, p), has values that range from 2.17 to 3.97 when $\beta_{pq} = 5$, and 7.30 to 9.91 when $\beta_{pq} = 20$. The corresponding range of values for coordinates without interaction effects is -0.68 to 0.38, and -2.43 to 0.50, respectively. The *spread* effect was less noticable than before.

Figure 4.15 presents the output corresponding to data generated with 40 variables when β_p is sampled from the vector $(-5, \ldots, 5)$ for $p = 1, \ldots, 40$. Otherwise, the data is generated using the same combinations of factor levels as before. The results are very similar to those found in Figure 4.14. When $\beta_{pq} = 0$ or 1, the matrices presented no particular patterns and neither measure could detect any interaction effects. In terms of

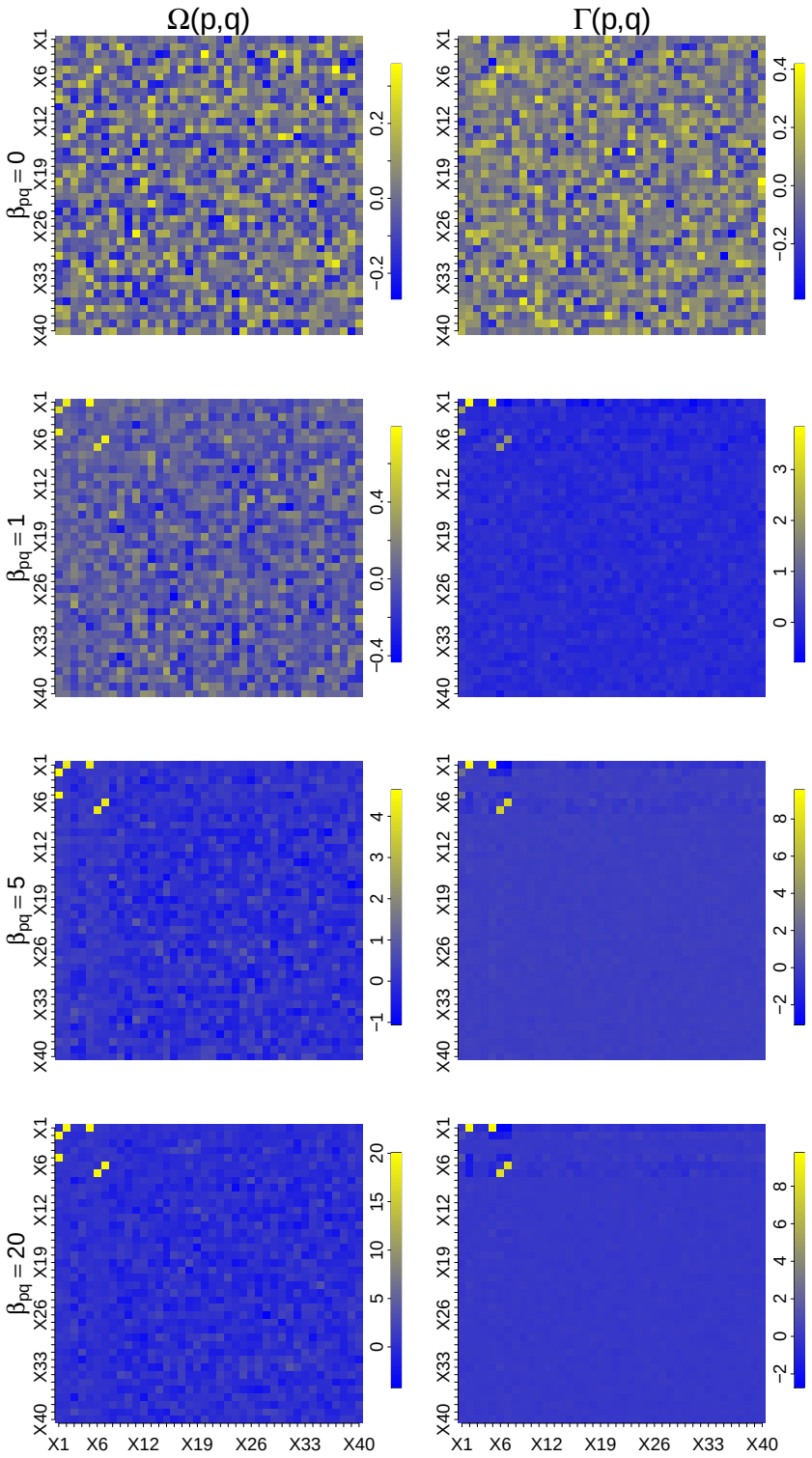

Figure 4.13: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, $\beta_p = 1$ for all $p = 1, \ldots, 40$, and $P(X_p = 1) = .5$.

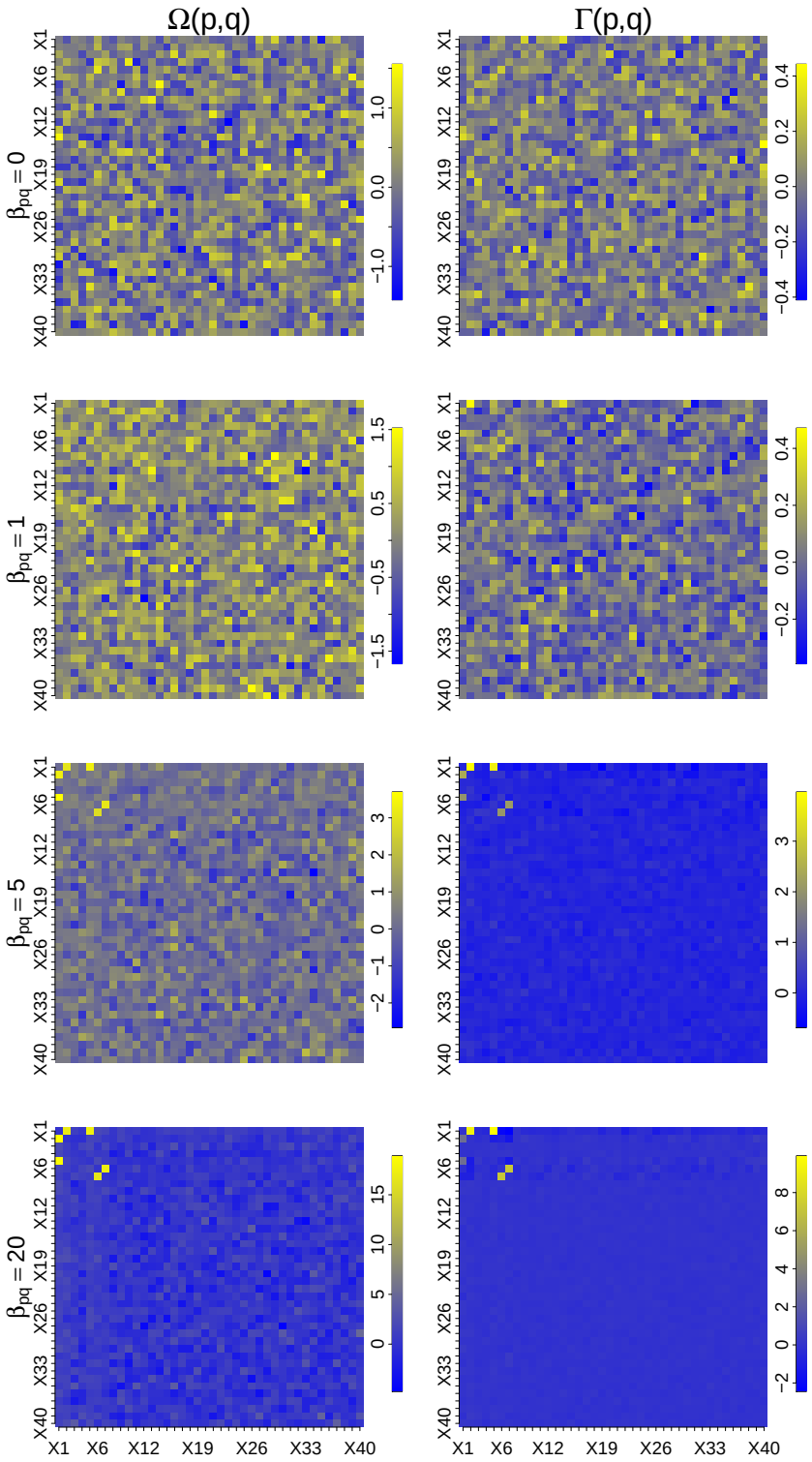

Figure 4.14: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, $\beta_p = 5$ for all $p = 1, \ldots, 40$, and $P(X_p = 1) = .5$.

the mean-based measure, the cell values ranged from -1.87 to 1.86 when $\beta_{pq} = 0$ and from -1.86 to 1.61 when $\beta_{pq} = 1$. When $\beta_{pq} = 5$ or 20, all three interaction effects are identified. When $\beta_{pq} = 5$, the estimates range from 4.25 to 7.26 for cells with interactions and from -1.71 to 1.89 for cells without interactions. When $\beta_{pq} = 20$, the estimates range from 18.10 to 22.32 for cells with interactions and from -4.46 to 4.37 for cells without interactions. For the distribution-based measure, the ranges for all cell values are -0.31 to 0.51 when $\beta_{pq} = 0$, and -0.31 to 0.46 when $\beta_{pq} = 1$. In the case of $\beta_{pq} = 5$ or 20, at least one relevant cell, (p, q) or (q, p), have values that range from .64 to 1.12 when $\beta_{pq} = 5$ and 6.92 to 11.59 when $\beta_{pq} = 20$. The range of values for cells without interaction effects is -0.32 to 0.46, and -1.24 to 0.79, respectively. The *spread* effect was less noticable than before.

Results for different combinations of interaction coefficients and success probabilities. Figure 4.16 shows the means across 100 replicates of the mean-based interaction measures for three combinations of interaction effects and two success probabilities. The data is generated with 20 variables, β_p sampled from the vector $(-5, \ldots, 5)$ for $p = 1, \ldots, 20$. The interaction effects are β_{pq}=0, 5, and 20, for $(p, q) = (1, 2), (1, 5)$, and $(6, 7)$, and the success probabilities are $P(X_p = 1) = 0.5$ and 0.75 for $p = 1, \ldots, 20$. All interaction effects are correctly identified for both success probabilities when the interaction effects are 5 and 20. When $P(X_p = 1) = 0.5$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 5$, the mean-based estimates range from 4.02 to 6.73 for cells with interactions and from -1.48 to 1.72 for cells without interactions. When $P(X_p = 1) = 0.5$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 20$, the mean-based estimates range from 19.19 to 22.39 for cells with interactions and from -2.63 to 2.28 for cells without interactions. When $P(X_p = 1) = 0.75$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 5$, the mean-based estimates range from 3.66 to 7.33 for cells with interactions and from -2.00 to 1.73 for cells without interactions. Finally, when $P(X_p = 1) = 0.75$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 20$, the mean-based estimates range from 18.51 to 23.41 for cells with interactions and from -4.74 to 3.02 for cells without interactions.

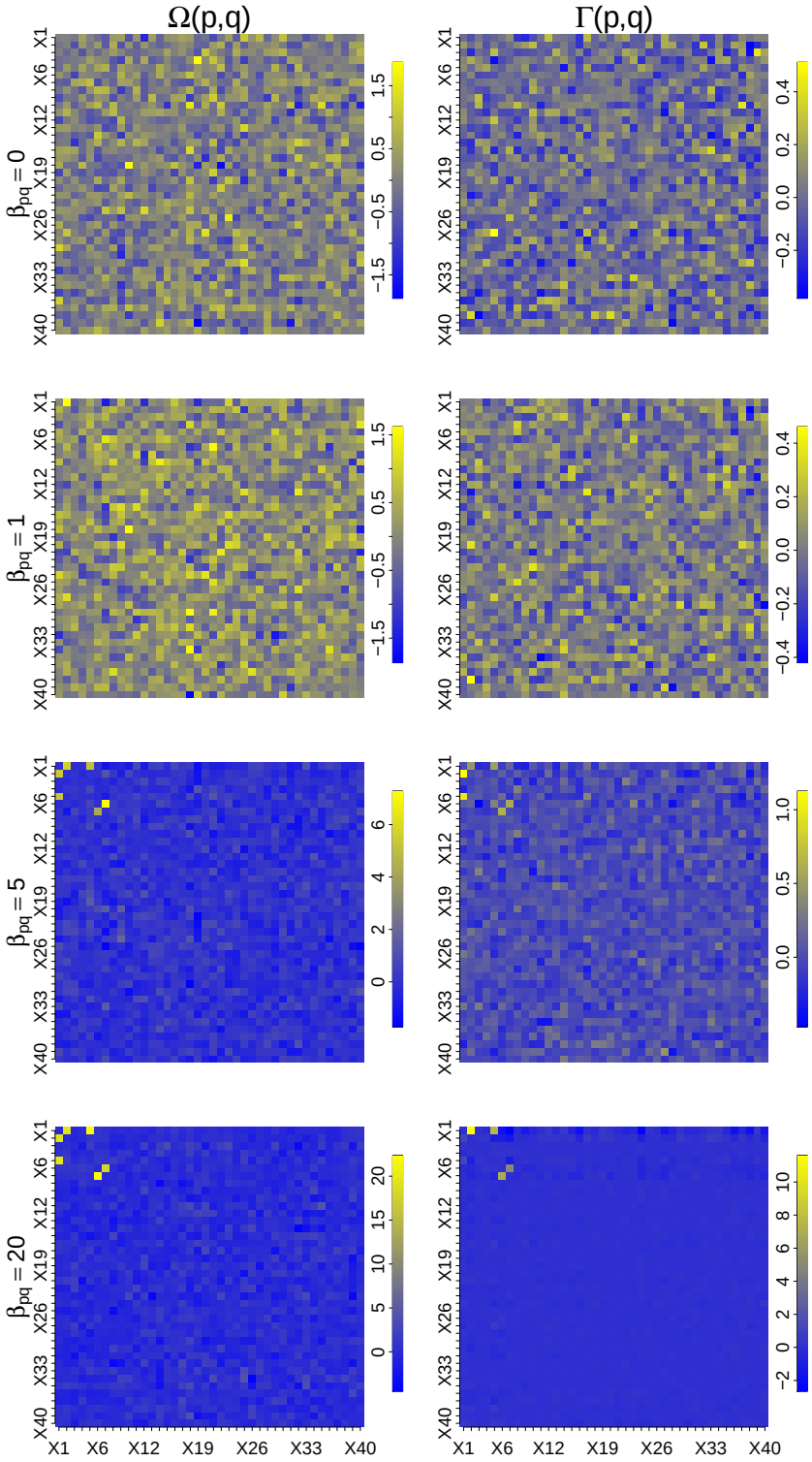

Figure 4.15: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, β_p is sampled from $(-5,\ldots,5)$ for all $p = 1,\ldots,40$, and $P(X_p = 1) = .5$.

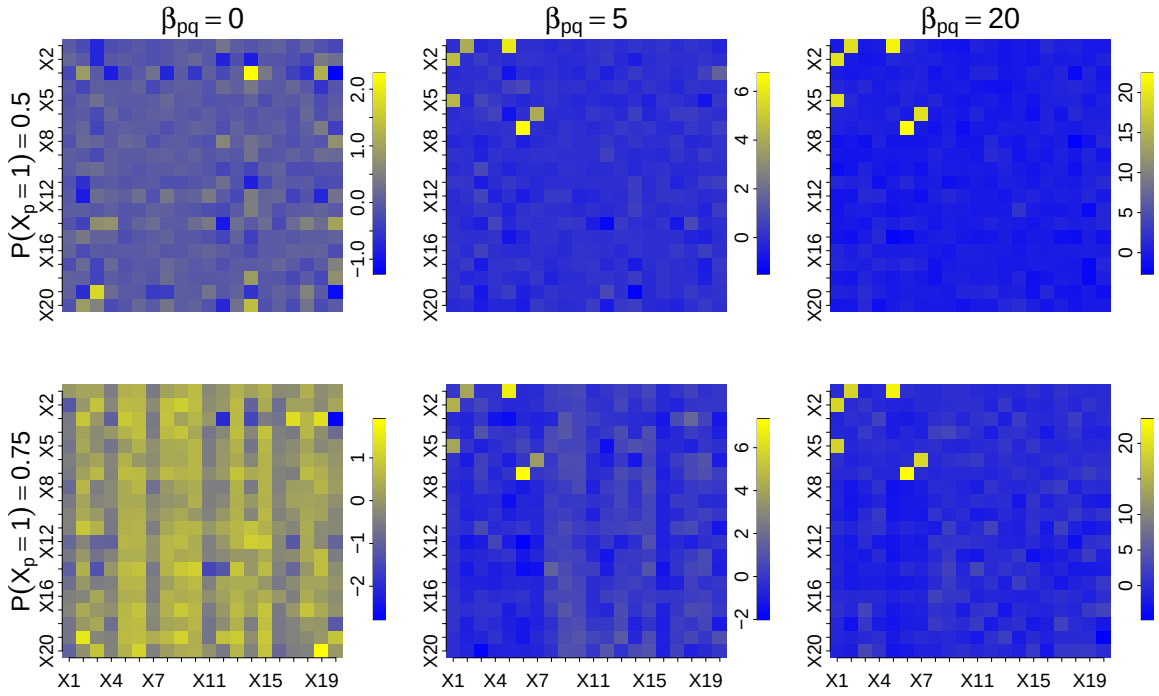

Figure 4.16: Estimated interaction effects for mean-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 20 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 20$.

Figure 4.17 contains analogous results to those in Figure 4.16 for the distribution-based measure. Observe that all interaction effects are identified in at least one of the corresponding cells, (p, q) or (q, p), for both success probabilities when the interaction effects are 5 and 20. These values could be positive or negative; therefore, we report absolute values in the following description of ranges. When $P(X_p = 1) = 0.5$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 5$, the distribution-based absolute values range from 11.85 to 38.89 for at least one of the corresponding cells with interactions and from 2.31 to 4.50 for cells without interactions. When $P(X_p = 1) = 0.5$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 20$, the distribution-based absolute values range from 21.58 to 41.45 for at least one of the corresponding cells with interactions and from 1.21 to 10.27 for cells without interactions. When $P(X_p = 1) = 0.75$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 5$, the distribution-based absolute values range from 12.13 to 35.04 for cells with interactions and from 2.47 to 6.67 for cells

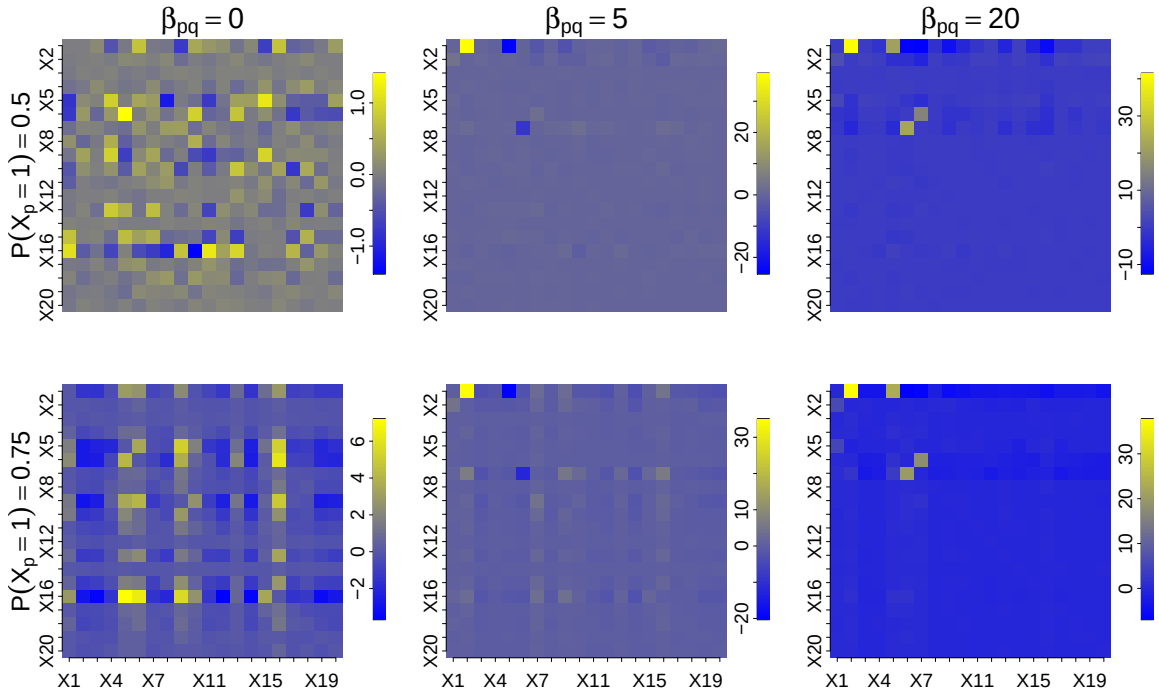

Figure 4.17: Estimated interaction effects for distribution-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 20 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 20$.

without interactions. Finally, when $P(X_p = 1) = 0.75$ for all p and $\beta_{12} = \beta_{15} = \beta_{67} = 20$, the distribution-based absolute values range from 20.09 to 37.76 for at least one of the corresponding cells with interactions and from 1.59 to 6.90 for cells without interactions.

4.3 Discussion

Based on the structure of the trees in random forest, we have proposed two new measures to identify and estimate interaction effects: the distribution-based and the mean-based measures, respectively. Four versions of the mean-based and three of the distribution-based measure were formulated and one representation for each measure was selected.

The selected distribution-based and mean-based interaction measures were able to identify and estimate the interaction effects in most of the scenarios we studied. When looking at the interaction measure for specific data sets, we found the mean-based

interaction estimates to be sensitive to the number of terminal nodes obtained for each tree in random forest. When a small number of terminal nodes was considered, several pairs of variables could be identified as having interactions when in reality no interaction effect existed between them. Increasing the number of terminal nodes improved the accuracy of the mean-based estimates. On the other hand, the distribution-based identification measure did not demonstrate sensitivity to the number of terminal nodes.

The proposed interaction measures were capable of identifying and estimating interactions even when the interaction effects were as small as the variance of the error terms in the model and when these effects were about the same size of the variables direct effect. However, all the mean-based interaction estimates were biased toward smaller values than the true interaction effects. Only when the interaction effect was as small as the error term variance and the direct effects were considerably larger did the proposed measures fail to detect the presence of interactions.

The binary variables in the study were generated from a Bernoulli distribution. In this respect, our interaction measures were able to identify and estimate interactions for different success probabilities that were used to create the data.

In some scenarios, the distribution-based measure incorrectly identified interactions between two variables. We called this unintended result a *spread* interaction effect. This occured when an interaction effect was large relative to other interaction effects or direct effects. The mean-based interaction measure was not sensitive to the *spread* effect.

Overall, it appears there may be some merit to using both measures in tandem for estimation of interaction effects. The distribution-based measure would be employed to identify the interactions and the mean-based measure could then be used for the corresponding point estimator. By doing so, it may be possible to obtain interaction estimates even when *spread* effects are present and the number of terminal nodes is small.

The interaction estimates obtained with the mean-based measure produced estimates that were biased toward smaller values; a problem that will be investigated in future research. Table 4.9 presents the average bias for those pairs of variables that interact, Bias($\beta_{pq} \neq 0$), and those that do not interact, Bias($\beta_{pq} = 0$), for all the factor combinations with $\beta_p = 1$ for $p = 1, \ldots, P$. The results when $\beta_p = 5$ or sampled from $(-5, \ldots, 5)$ for $p = 1, \ldots, P$, are presented in Appendix B.

β_p	Num. Variables (P)	$Prob(X_p = 1)$	β_{pq}	$\tilde{\beta}_{pq}$	Bias($\beta_{pq} \neq 0$)	Bias($\beta_{pq} = 0$)
1	10	0.50	0	-0.02	-0.02	-0.01
1	20	0.50	0	0.01	0.01	-0.00
1	40	0.50	0	0.03	0.03	0.00
1	10	0.75	0	-0.06	-0.06	-0.10
1	20	0.75	0	-0.05	-0.05	-0.08
1	40	0.75	0	-0.12	-0.12	-0.09
1	10	0.50	1	0.76	-0.24	0.06
1	20	0.50	1	0.71	-0.29	0.02
1	40	0.50	1	0.69	-0.31	0.01
1	10	0.75	1	0.63	-0.37	-0.05
1	20	0.75	1	0.61	-0.39	-0.07
1	40	0.75	1	0.59	-0.41	-0.07
1	10	0.50	5	4.43	-0.57	0.11
1	20	0.50	5	4.54	-0.46	0.09
1	40	0.50	5	4.34	-0.66	0.04
1	10	0.75	5	4.53	-0.47	0.05
1	20	0.75	5	4.44	-0.56	0.01
1	40	0.75	5	4.41	-0.59	0.03
1	10	0.50	10	9.19	-0.81	0.13
1	20	0.50	10	9.50	-0.50	0.12
1	40	0.50	10	9.29	-0.71	0.06
1	10	0.75	10	9.40	-0.60	0.10
1	20	0.75	10	9.36	-0.64	0.01
1	40	0.75	10	9.43	-0.57	0.01

β_p	Num. Variables (P)	$Prob(X_p = 1)$	β_{pq}	$\tilde{\beta}_{pq}$	Bias($\beta_{pq} \neq 0$)	Bias($\beta_{pq} = 0$)
1	10	0.50	20	18.84	-1.16	0.17
1	20	0.50	20	19.49	-0.51	0.20
1	40	0.50	20	19.54	-0.46	0.11
1	10	0.75	20	19.17	-0.83	0.09
1	20	0.75	20	19.23	-0.77	-0.13
1	40	0.75	20	19.13	-0.87	-0.15

Table 4.9: Average interaction estimation bias for those pairs of covariates that interact, Bias($\beta_{pq} \neq 0$), and those that do not interact, Bias($\beta_{pq} = 0$), when $P = 10, 20$, or 40, $\beta_p = 1$, $Prob(X_p = 1) = 0.5$ or 0.75 for $p = 1, \ldots, P$, and β_{pq} varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, or $(6, 7)$.

CHAPTER 5

CONCLUSION

Summary of Methods and Results

The original motivation for the topics studied in this dissertation was to expand the scope and usefulness of VAMs in education. As a result, we proposed new methods to determine characteristics of the underlying models based on the random forest procedure. We focused on two such characteristics: the variable importance measures and interactions. The novelty of the proposed methods is that they were constructed by taking into account not only the final outcome values, as is traditionally done, but also characteristics of the structure of the random forest, i.e., patterns found in the constitutive trees.

The contributions of this work are contained in two central chapters of this dissertation: Chapters 3 and 4. In both chapters, we present the formulation and development of measures and evaluate their potential usefulness via simulation studies. A brief summary of the key contributions from each chapter is provided next, followed by a discussion of the limitations of the current study. We conclude with a discussion of future research directions.

In Chapter 3, we proposed two novel VIMs. These measures were determined by the final configuration of the terminal nodes on each tree. The first VIM we proposed, the node-proportion, was formulated as follows. The importance measure for a particular covariate was obtained by the average of relative importance of that covariate on each terminal node in each tree. For a given terminal node in a given tree, a covariate's relative importance was measured as the proportion of observations affected by this covariate in that terminal node versus observations affected by it in the entire tree. The second new VIM, the covariate-proportion, was constructed similarly to the first with the exception of the relative importance formulation. Here, the covariate's relative importance was assessed by the proportion of observations affected by this covariate in

that terminal node versus the total number of observations in that terminal node. In order to examine the usefulness of the proposed VIMs, using a simulation study, under a variety of conditions, we produced a ranking of random effects from data generated based on the covariate adjustment and gain score models. We also obtained the corresponding rankings based on existent VIMs and EBLUPs. We compared whether the VIM rankings were more accurate than the EBLUP rankings based on the Spearman's correlation with the true random effects rankings. These comparisons were made when the linear mixed model was correctly specified and when it was misspecified. The correctly specified models showed that the EBLUP rankings were more accurate, although the VIM rankings were often nearly as accurate. On the other hand, VIM rankings were sometimes more accurate than the EBLUP rankings when the model was misspecified, particularly when third-order interaction effects were present that were not included in the assumed model. In these situations, the proposed VIM rankings outperformed both the EBLUP and traditional VIM rankings.

The main contribution of the proposed VIMs to the VAM literature is that these measures can be used as a complementary tool to determine if the assumptions about the underlying model are adequate when obtaining the EBLUPs. If the EBLUP rankings are similar to those produced by the VIMs, then we might conclude that the underlying model used to obtain the EBLUPs is adequate; otherwise, important effects might be unnaccounted for in the model specification.

In Chapter 4, we studied interaction effects. We proposed two measures to identify and/or estimate second-order interaction effects: the distribution-based and the mean-based interaction measures. The rationale for the proposed measures relies on the assumption that, independently of the nature of the unknown model specification, certain patterns in the structure of the resulting trees in a random forest provide information about the existence of variable interactions. The proposed measures were constructed precisely with the goal of capturing those patterns. Specifically, given any two variables,

the distribution-based interaction measure was built taking into account the frequency of their appearance in different nodes in each tree as well as their relative position with respect to each other and to the root node.

In order to study this further, we restricted our analysis to a linear model as presented in (4.17). The mean-based measure was constructed as a special case of the distribution-based measure also taking into account the response values obtained in each tree. This additional consideration allowed the mean-measure not only to identify the interaction effects, as was the case with the distribution-based measure, but also to estimate the interaction effects.

As with our new VIMs, the interaction measures we devised were evaluated in a simulation study under a number of conditions. The results suggested that the distribution-based measure identified the interaction effects in most of the scenarios that were studied and the mean-based measure produced estimates of the true interaction effects that approached the true values but were biased toward smaller values. Furthermore, the interaction measures were affected by the random forest characteristics, largely by the constitutive trees' size and the number of variables used to select the splitting variable at each node. Larger branches in trees and a relatively small number of variables used to select the splitting variable yielded more accurate results for the proposed interaction measures. Finally, the study found that by using both measures simultaneously, the distribution-based measure to identify and the mean-based measure to estimate the interaction effects, we could obtain useful interaction estimates even when *spread* effects are present and the number of terminal nodes is small.

The main contribution of the proposed interaction measures is that they could be used as a self-standing mechanism to determine interaction effects or as a complementary tool to improve traditional methods of statistical modeling. For example, we could use

these measures to identify potential interaction effects and include those effects in the model specification, that in turn could be used to estimate the interaction effects.

Limitation of the Study

The current study has several limitations. While the parameters chosen in the simulation studies were made to correspond to a realistic scenario, the generalizability of the conclusions is limited to the particular choices. Specifically, in Chapter 3, we used the covariate adjustment and gain score models to obtain comparisons between the proposed VIMs and the EBLUPs. Although these models allowed us to draw conclusions about comparisons between the traditional approach and the data mining approach, additional considerations are needed when studying generalized linear mixed models, as the one described in McCaffrey et al. (2004): namely, the extent by which multimembership random effect structure of these models could be captured by regression trees in the random forest procedure. In addition, alternative scenarios should be studied in order to examine the usefulness of the proposed VIMs, including additional nonlinearities in the model specification and correlation among covariates.

Based on the study design in Chapter 4, we considered a linear model that used binary variables or categorical variables with two categories. Hence, the proposed interaction measures were only evaluated within this framework. It is unknown how well the measures will perform when the variables are continuous or categorical with several categories. Variables with several categories could potentially appear multiple times in the same branch of a tree and/or could appear before or after another covariate for which the interaction measure is studied. Furthermore, as mentioned previously, the proposed measures were constructed based on a linear model; i.e., trees in random forest were built based on data generated from a linear model, and patterns on those trees were considered to produce the proposed measures. The mean-based interaction measure, in particular, was devised using unique attributes of the linear model and should be used with caution to estimate interactions if the underlying model is unknown or assumed nonlinear. Even

when the underlying model is linear, this measure produces an estimator that is biased toward smaller values. The distribution-based interaction measure, on the other hand, might still be adequate beyond linear model specifications; however, additional work is needed to investigate this possibility.

Future Research

Given these limitations and the novelty of the proposed methods, we believe that future research on the use of data mining methods to gain insights into the structure of an underlying model is warranted. Some of this work should address the limitations of the current work rather directly, for example, expanding on the study design choices to include scenarios that were not studied here. Other work may require additional considerations and modifications to the proposed measures or the data mining methods to allow for more generalizable results. Finally, as suggested below, future research may depend on the results of those additional investigations and conclusions obtained.

With respect to the proposed VIMs, the immediate future work will focus on proposing alternative random forest formulations that take into account the multimembership structure of the value-added models, in particular for the complete persistence model described by Mariano et al. (2010). To consider this model, we should study not only the presence or absence of a variable in each observation, but also a range of possible values that this variable could adopt. Hence, the proposed VIMs need to consider variables with different numbers of categories. Consequently, selection bias problems for the random forest method need to be addressed.

With respect to the proposed interaction measures, the first task for future research is to determine how to effectively correct the bias in the mean-based measure. In addition, a larger range of values for the true interaction effects relative to the direct effects and the variance of the error terms need to be considered. In particular, the proposed interaction measures were able to identify interaction effects that in magnitude

were as small as the direct effects and the error term variances; however, interaction effects as small as the error term variance and five times smaller than the direct effects were not identified. Additional ranges of comparisons need to be studied to determine the threshold for interaction identification. Furthermore, extensions for the interaction measures that take into account different variable types, e.g., continuous or with different number of categories, are needed. The following discussion indicates some specific avenues of inquiry that might be followed in relation to this latter problem.

To account for a variable with several categories, we can express each variable as a set of binary (dummy) variables and analyze the accuracy of the proposed interaction measures in this modified data set. An alternative approach comes from the random forest construction. Regardless of the original variable type, the chosen splitting variable could also be considered a binary variable, albeit this classification is a result of an optimization process that locates an optimal split point. The proposed measures might still yield meaningful results provided they take into account the additional information produced by those splits.

Once we start considering variables with several categories, we need to determine if variables with different numbers of categories produce selection bias in the interaction measures. As mentioned previously, the interaction measures are based on the assumption that because random forest produces an accurate prediction model, the constitutive trees contain information about the underlying model specification. However, if random forest does not produce adequate results, the tree structure may no longer be useful. Future research is needed to better understand whether the tree structure may still reflect interactions adequately, even in the presence of variable selection bias. There is a reason to hope that such might be the case because selection bias affects the tree structure in ways different than interactions. For instance, a covariate with a larger number of categories will be chosen more often than covariates with fewer categories. However, the

proposed interaction measures are not directly affected by the frequency with which variable is chosen in the tree.

The problem of selection bias can also be addressed from a different perspective. In the last few years alternative methods have been developed to solve the limitations of random forests in terms of variable selection bias and correlation. Two such methods are GUIDE (as in Loh (2002)) and Unbiased Recursive Partitioning (due to Hothorn et al. (2006)). Hence, a potential area of research is to study if the proposed interaction measures are still adequate when used on trees based on these alternative methods. Notice that we have used unbiased recursive partitioning (Hothorn et al., 2006) in Chapter 3, but the potential selection bias was restricted to a unique continuous variable (pre-scores) that was not part of the VIM rankings. In that case, we found that the results obtained by the traditional random forest algorithm were preferable.

Model specifications that are truly nonlinear present another potential area for future research. The simulations studies presented here are restricted to model specifications where the nonlinearity is expressed only through the introduction of interactions.

An important addition to this research would be developing a statistical framework that would allow us to formalize the inferential aspects of our methodology. A starting point would be to develop tests of significance for the interaction measures and VIMs. In terms of interaction measures, we could start by attempting to determine if there is a relationship between the accuracy of the interaction measure estimate and the prediction estimate for the corresponding variables. For example, an interval estimation upper bound could be determined if the off-sums in (4.24) and (4.36) could be expressed in terms of the off-sums in (4.19). Hence, the interaction estimates could be bounded by the accuracy of the prediction estimates of the random forest solution.

Taken all together, should our avenues of current and future study produce results that validate the type of measures proposed in this dissertation, it may be feasible to expand the underlying premise to produce a general new methodology. This methodology would center on developing statistical learning and/or data mining techniques that take into account not only the final outcome, but also the resulting estimator structures that correspond to the methods that are used.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. journal of educational data mining. *Journal of Educational Data Mining*, 1:3–17.

Beck, J. E. and Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *Proceedings of the 9th International Conference on Intelligent Tutoring System*, pages 353–362.

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Jounal of the Royal Statistical Society*, 8:65–75.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forest and other averaging classifiers. *Journal of Machine Learning Research*, 9:2039–2057.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forest. *Machine Learning*, 45:5–32.

Breiman, L. (2004). Consistency for a simple model of random forests. Technical report, University of California at Berkeley, Statistics Department.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Briggs, D. and Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value- added analysis underlying the effectiveness rankings of los angeles unified school district teachers by the los angeles times. National Education Policy Center.

Bryk, A. S. and Weisberg, H. I. (1976). Value added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1:127–155.

Chong, H. Y., DiGangi, S., Jannasch-Pennell, A., and Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8:307–325.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley, Hoboken, NJ.

Deng, H. (2011). *System Complexity Reduction via Feature Selection*. PhD thesis, Arizona State University.

Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Operation Research Quarterly*, 24:465–467.

Doyle, P. and Fenwick, I. (1975). The pitfalls of AID. *Journal of Marketing Research*, 12:408–413.

Friedman, J. H. and Popescu, B. E. (2003). Importance sampled learning ensembles. Technical report, Stanford University, Department of Statistics.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2:916–954.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31:2071–2344.

Goldhaber, D. and Hansen, M. (2010). Assesing the potential of using value-added estimates of teacher job performance for making tenure decisions. Working Paper 31. National Center for Analysis of Longitudinal Data in Education Research.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, 60:280–288.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 14:351–388.

Heald, G. I. (1972). The application of the automatic interaction detector (A.I.D.) programme and multiple regression techniques to the assessment of store performance and site selection. *Operational Research Quarterly*, 23:445–457.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Karl, A., Yang, Y., and Lohr, S. (2011). Exploring missing data in value-added models in education. *Proceedings of the Social Statistics Section, American Statistical Association*, pages 2449–2460.

Kass, G. (1975). Significance testing in automatic interaction detection (a.i.d.). *Journal of the Royal Statistical Society*, 24:178–189.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society*, 29:119–127.

Kelly, C. and Okada, K. (2012). Variable interaction measures with random forest classifiers. *International Symposium on Biomedical Imaging*, pages 154–157.

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, 3:333–362.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2:18–22.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386.

Loh, W.-Y. (2012). Variable selection for classification and regression in large p, small n problems. *Probability Approximations and Beyond*, 205:133–157.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7:815–840.

Mariano, L. T., McCaffrey, D. F., and Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 35:253–279.

McCaffrey, D. F. and Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, 5:773–797.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29:67–101.

Mendez, G., Buskirk, T. D., Lohr, S., and Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, 97:57–70.

Messenger, R. and Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67:768–772.

Morgan, J. N. and Messenger, R. C. (1973). *THAID, a sequential analysis program for the analysis of nominal scale dependent variables*. Survey Research Center, Institute for Social Research, University of Michigan, MI.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of suvey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

Patterson, J. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaumann Publishers, San Mateo, CA.

Robinson, G. K. (1991). The BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6:15–34.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4:537–571.

Rothstein, J. (2010). Teacher quality in education production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125:175–214.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). *The Tennessee Value-Added Assessment System: A Quantitative, outcomes-based approach to Educational Assessment*, pages 137–162. Corwin Press, Thousand Oaks, CA.

Scott, A. and Knott, M. (1976). An approximate test for use with AID. *Jounal of the Royal Statistical Society*, 25:103–106.

Searle, S. (1971). Topics in variance component estimation. *Biometrics*, 27:1–76.

Sonquist, J. A., Baker, E. L., and Morgan, J. N. (1971). *Searching for Structure (Alias - AID - III) An approach to analysis of substantial bodies of micro-data and documentation for a computer program*. Institute for Social Research, University of Michigan, Michigan.

Sonquist, J. A. and Morgan, J. N. (1964). *The detection of interaction effects; a report on a computer program for the selection of optimal combinations of explanatory variables*. Suvey Research Center, Institute for Social Research, University of Michigan, Michigan.

Stewart, B. E. (2006). Value-added modeling: The challenge of measuring educational outcomes. *Carnegie Corporation of New York*.

Strasser, H. and Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 8:220–250.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14:323–348.

Therneau, T. and Atkinson, E. (1997). An introduction to recursive partitioning using the rpart routines. Division of Biostatistics 61, Mayo Clinic.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106:1626–1636.

Tuv, E., Borisov, A., Runger, G., and Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, 10:1341–1366.

Williams, R. H. and Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, 20:59–69.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., and Biernacka, J. M. (2012). SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics*, 13:164.

APPENDIX A
ADDITIONAL CHARTS FOR VARIABLE IMPORTANCE MEASURES

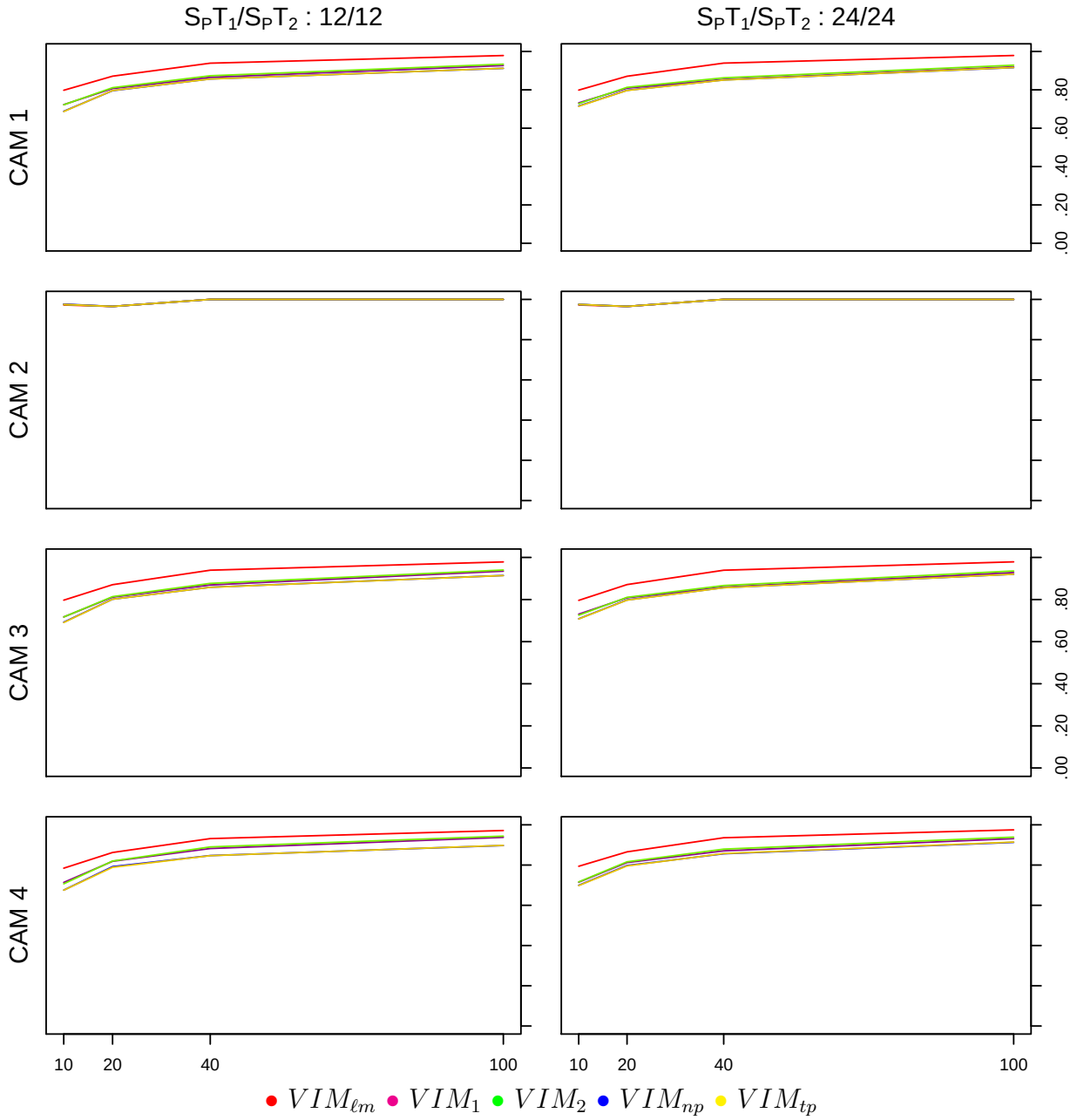

Figure A.1: Mean correlation between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different CAM models and different student per teacher ratios. $\sigma_\tau^2/\sigma^2 = 20$.

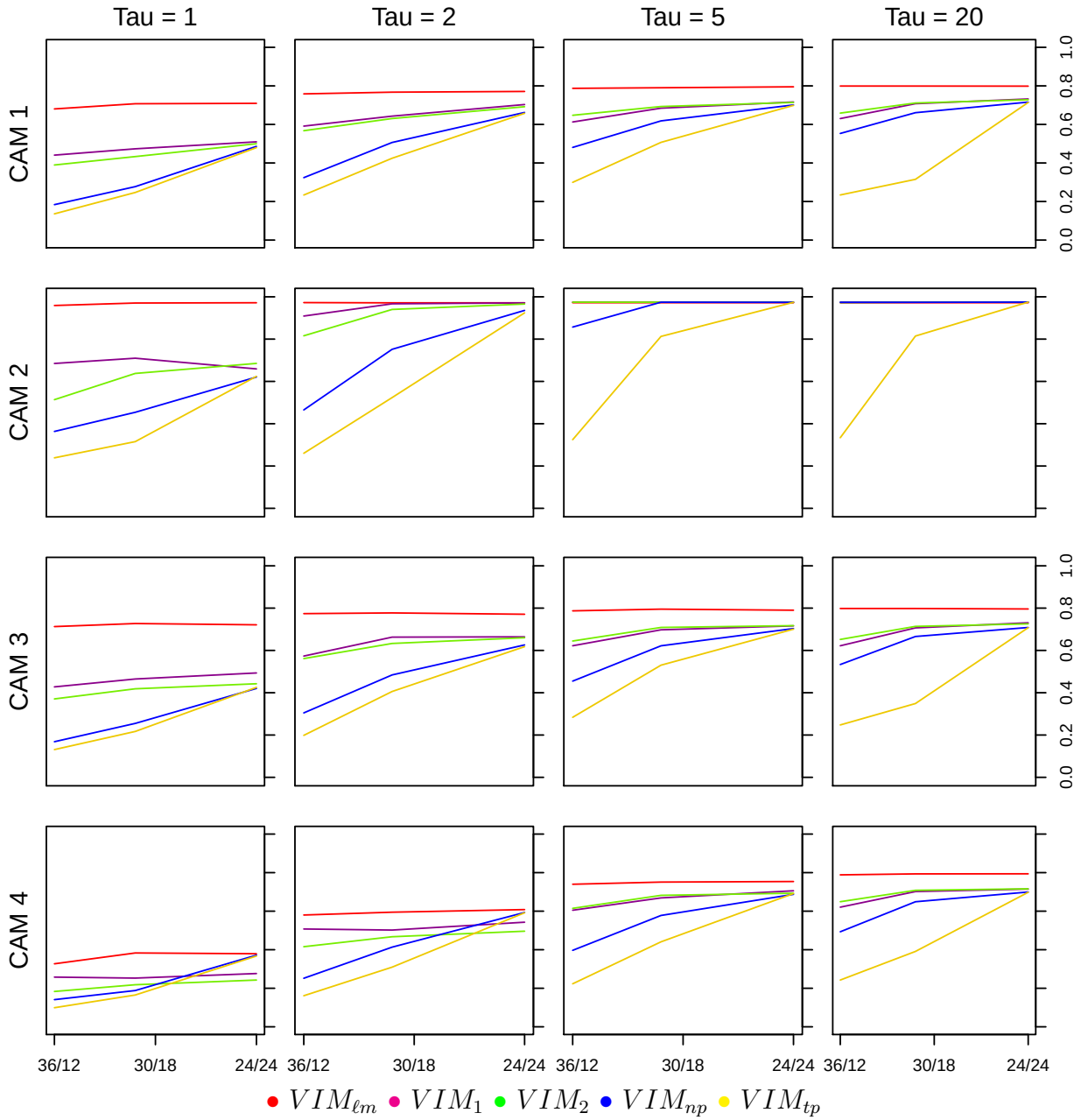

Figure A.2: Mean correlation between the VIMs and the absolute value of true teacher effects when the ratio of the number of students per teacher, SpT_1/SpT_2, varies for different CAM models and different $\sigma^2_\tau/\sigma^2 = 2$ when the number of teachers is 10.

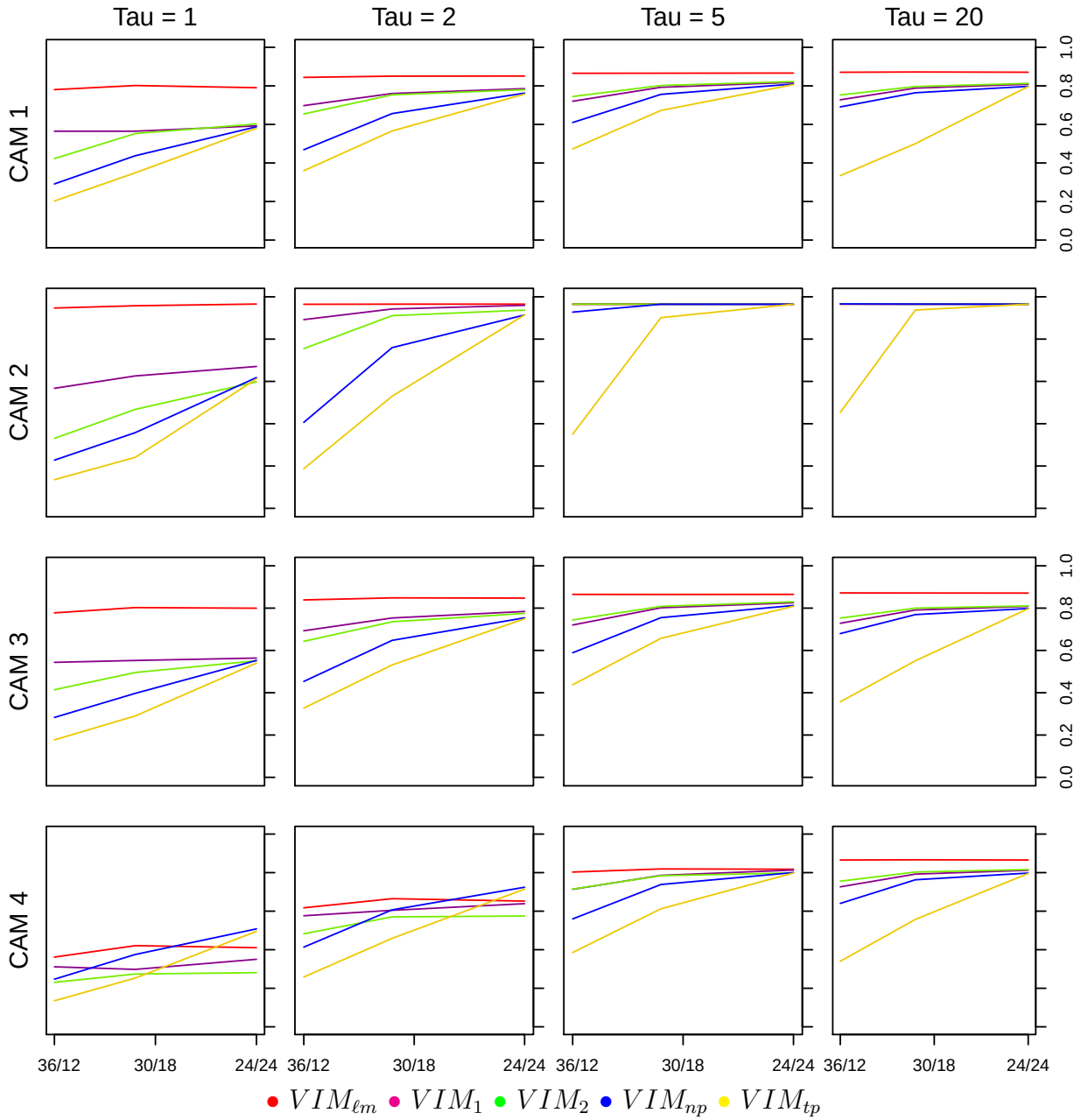

Figure A.3: Mean correlation between the VIMs and the absolute value of true teacher effects when the ratio of the number of students per teacher, SpT_1/SpT_2, varies for different CAM models and different $\sigma^2_\tau/\sigma^2 = 2$ when the number of teachers is 20.

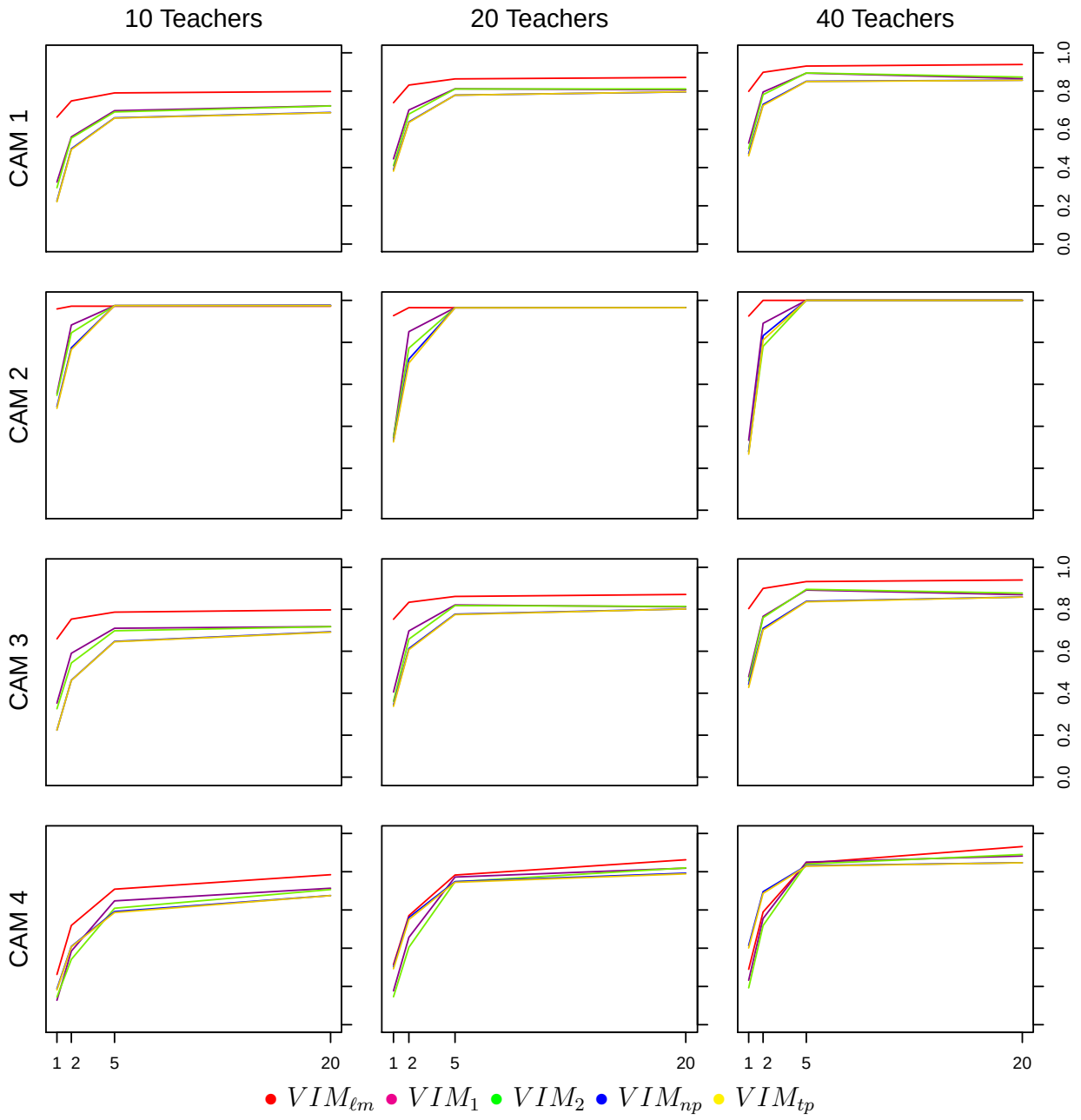

Figure A.4: Mean correlation between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teacher ratio is 12/12.

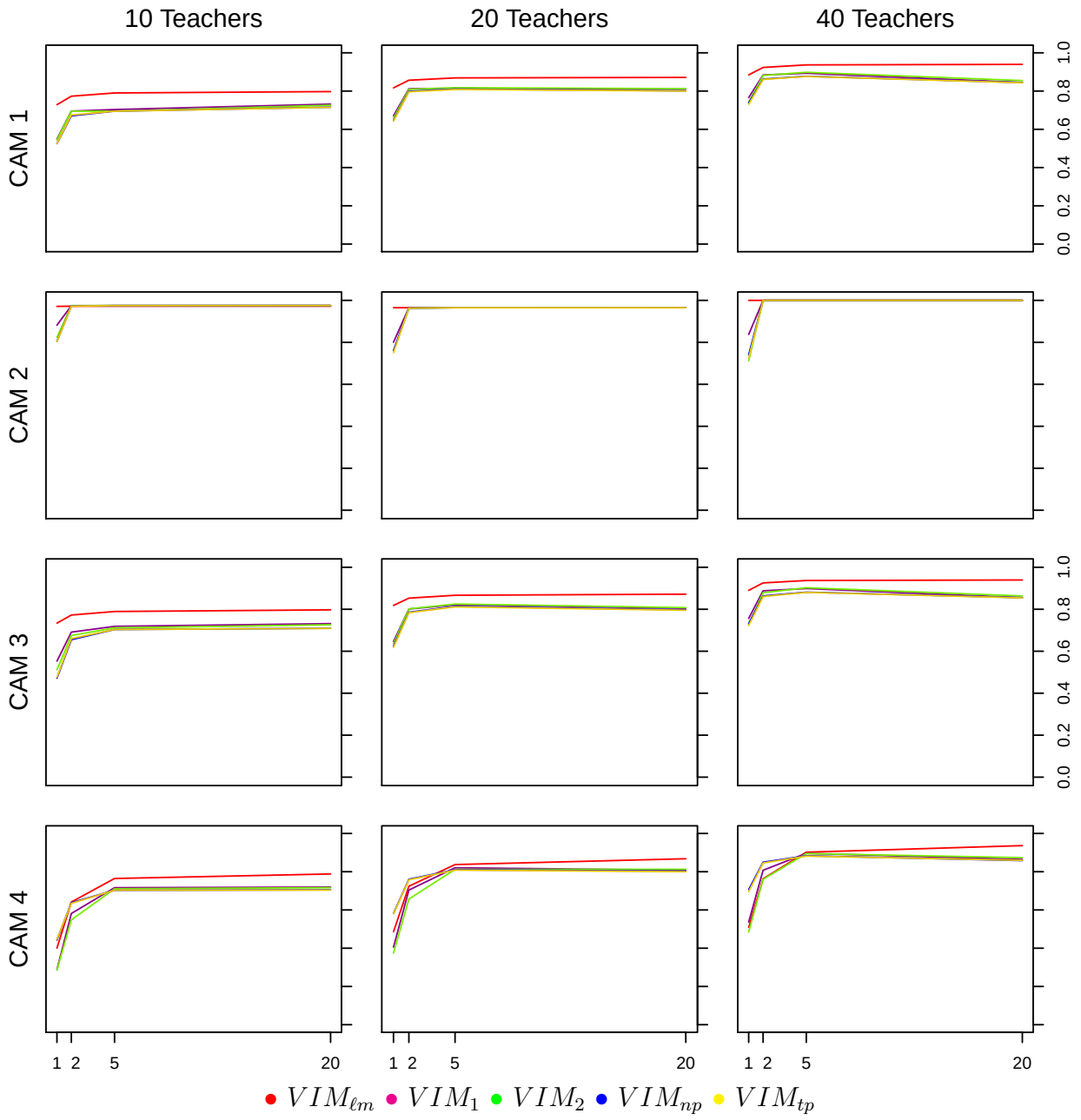

Figure A.5: Mean correlation between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different CAM models and different number of teachers when the number of students per teacher ratio is 36/36.

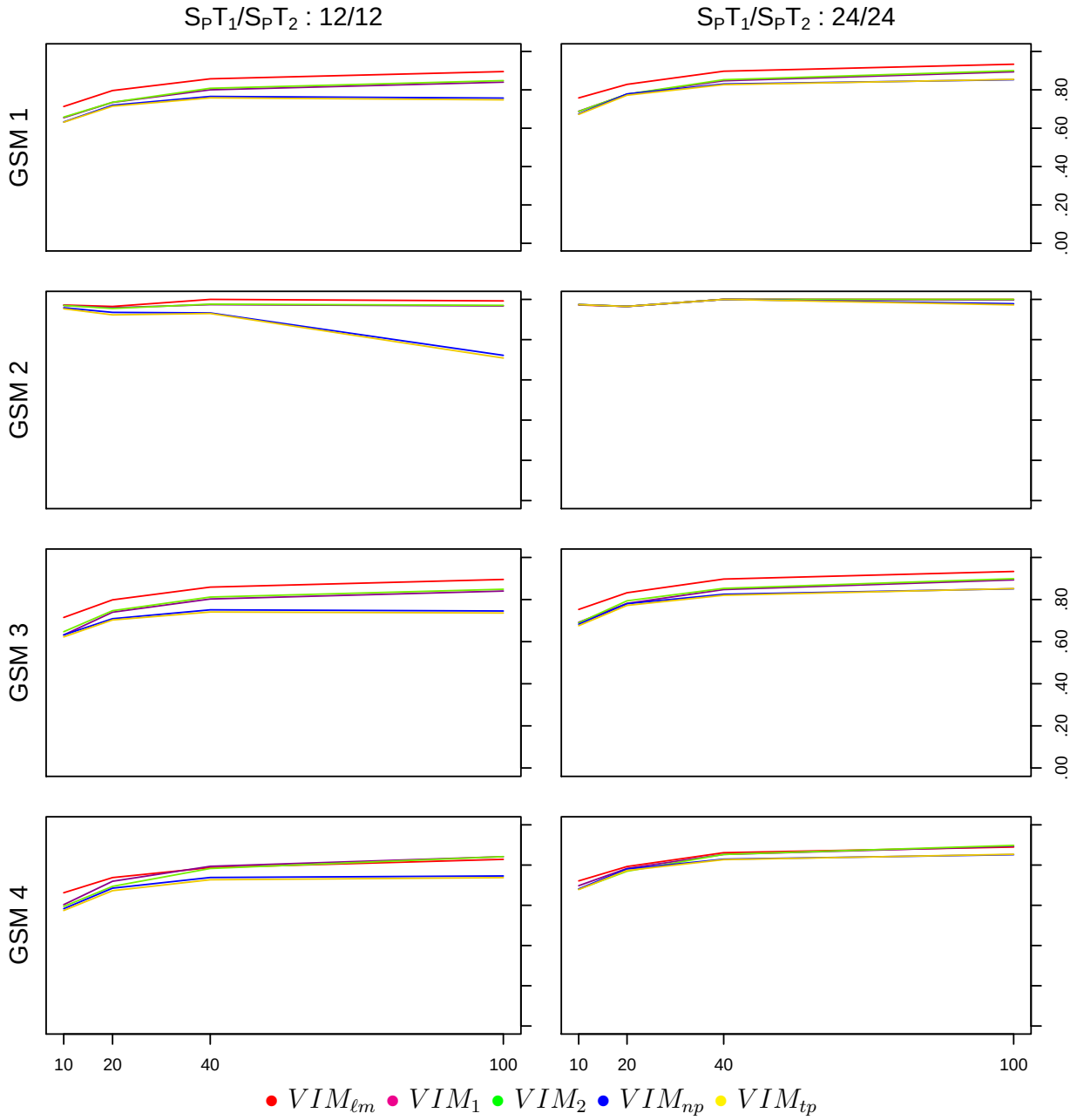

Figure A.6: Mean correlation between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different GSM models and different student per teacher ratios. $\sigma_\tau^2/\sigma^2 = 5$.

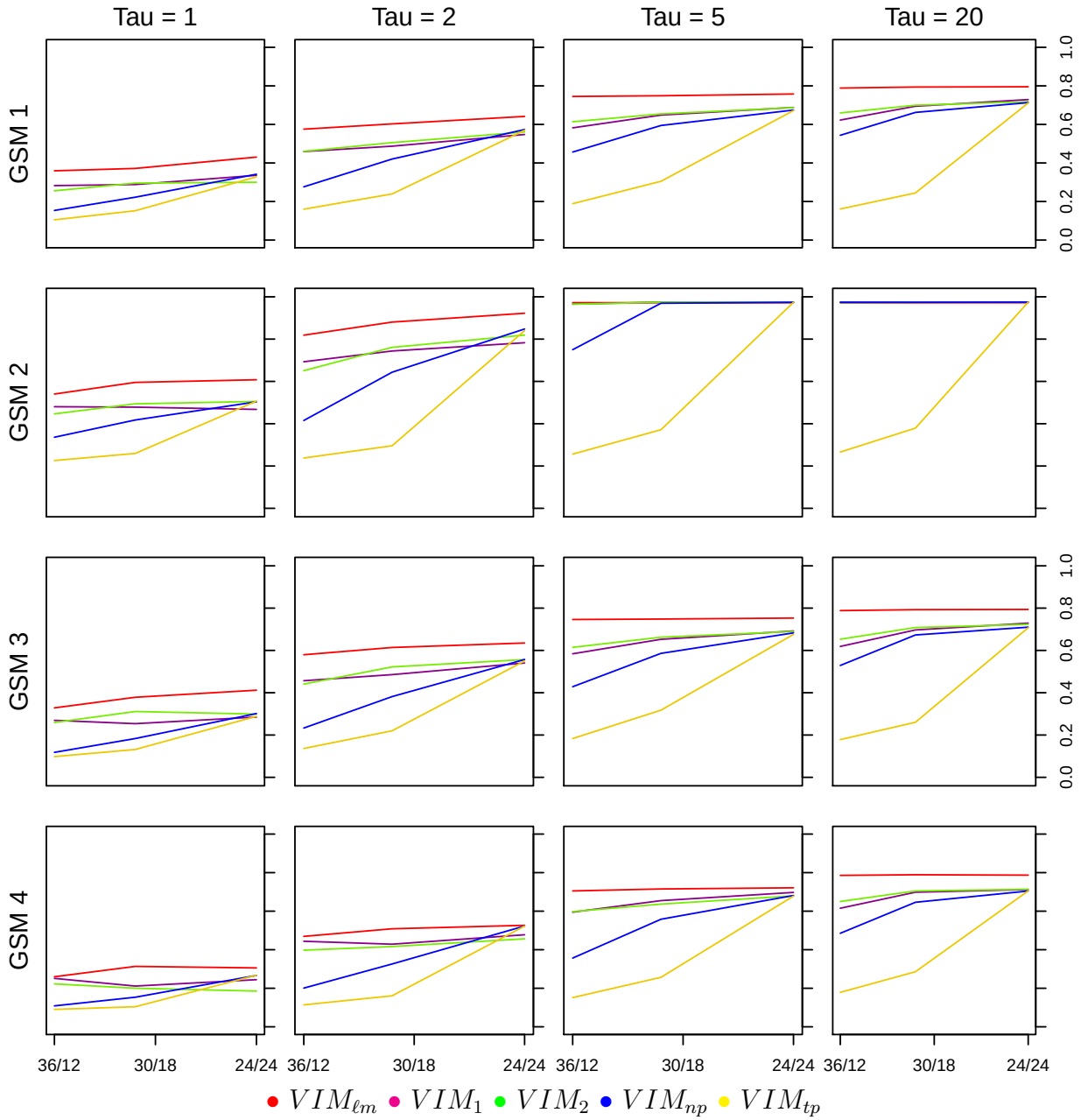

Figure A.7: Mean correlation between the VIMs and the absolute value of true teacher effects when the ratio of the number of students per teacher, SpT_1/SpT_2, varies for different GSM models and different $\sigma^2_\tau/\sigma^2 = 2$ when the number of teachers is 10.

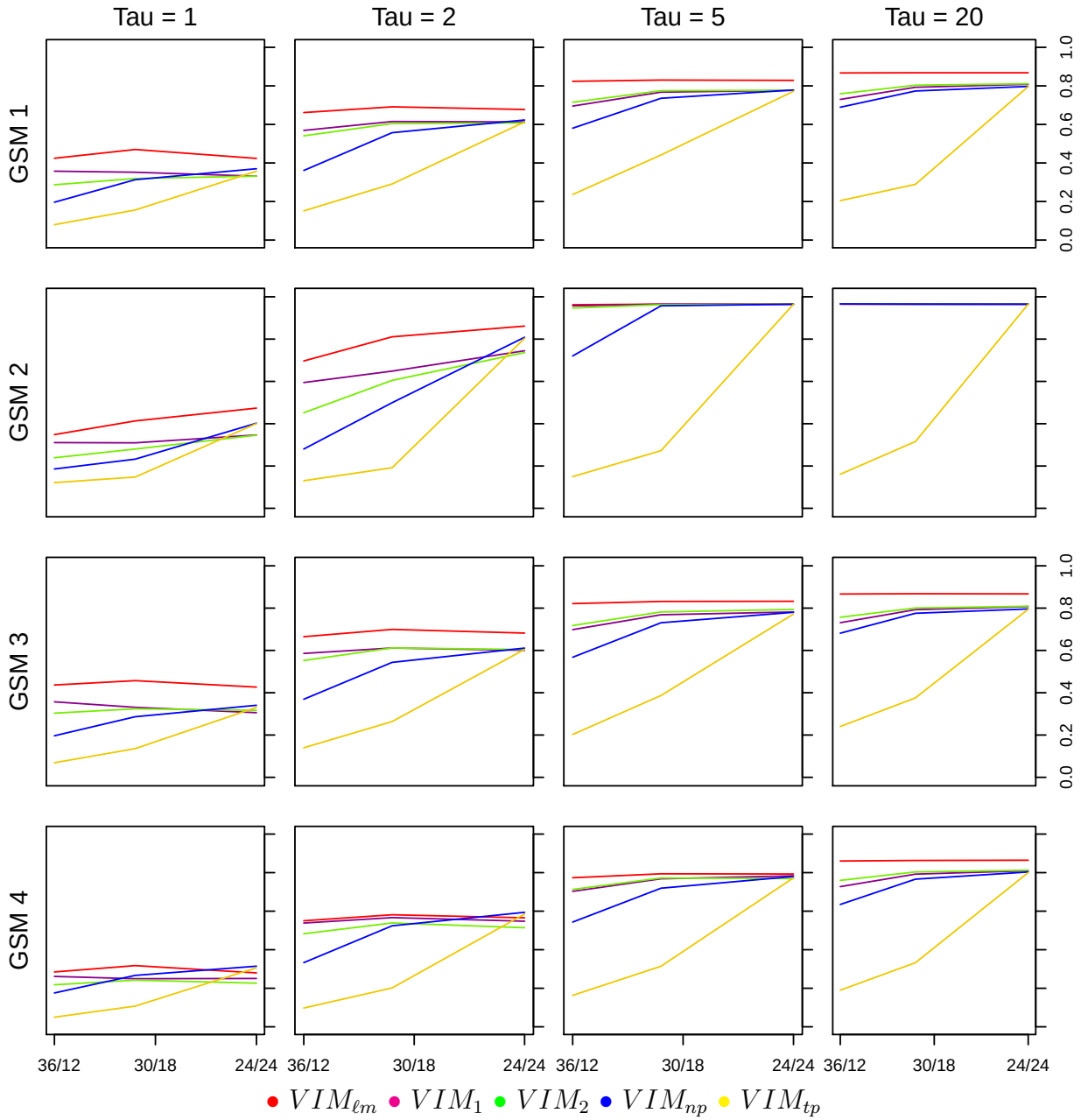

Figure A.8: Mean correlation between the VIMs and the absolute value of true teacher effects when the ratio of the number of students per teacher, SpT_1/SpT_2, varies for different GSM models and different $\sigma^2_\tau/\sigma^2 = 2$ when the number of teachers is 20.

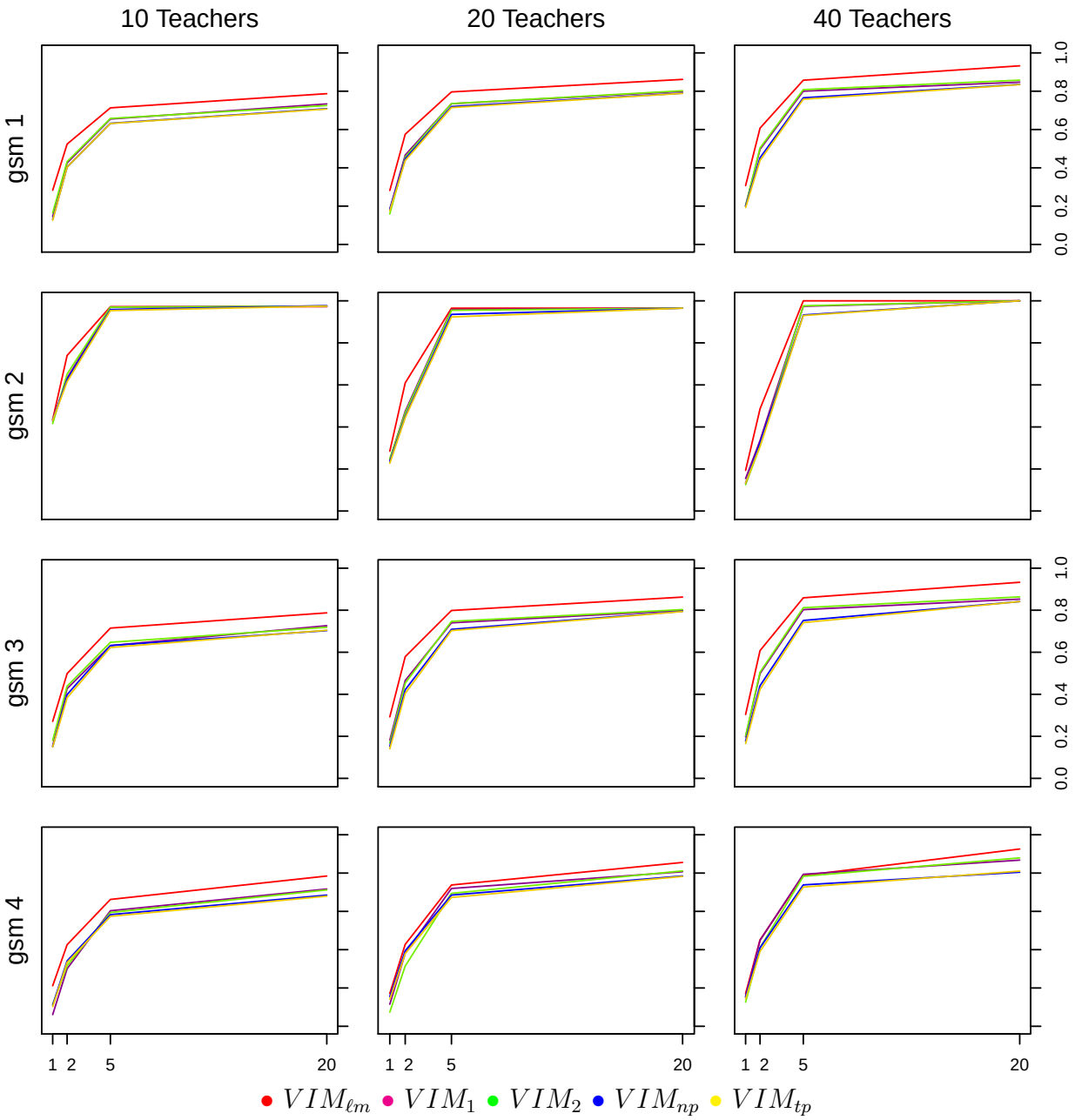

Figure A.9: Mean correlation between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different GSM models and different number of teachers when the number of students per teacher ratio is 12/12.

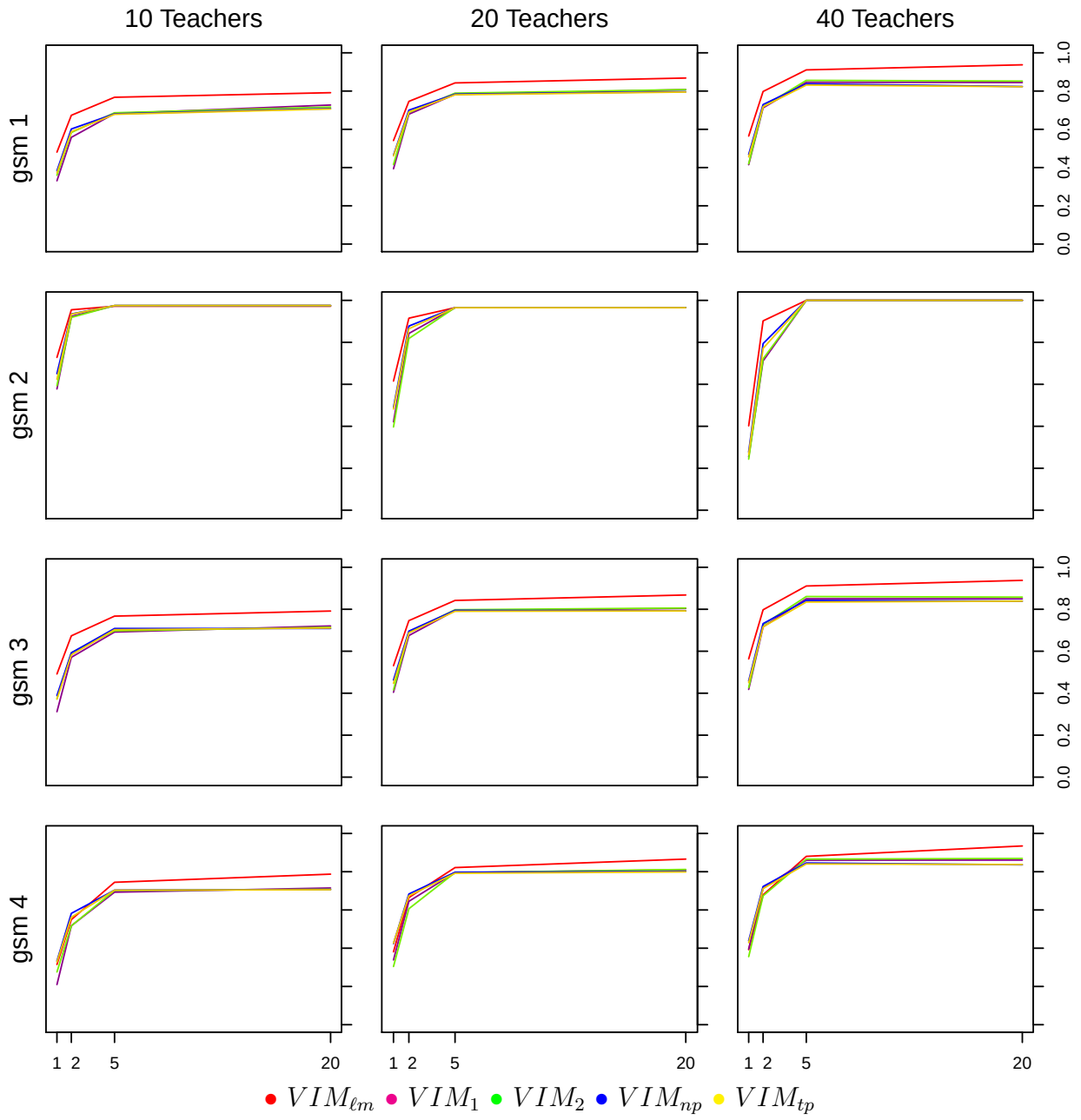

Figure A.10: Mean correlation between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different GSM models and different number of teachers when the number of students per teacher ratio is 36/36.

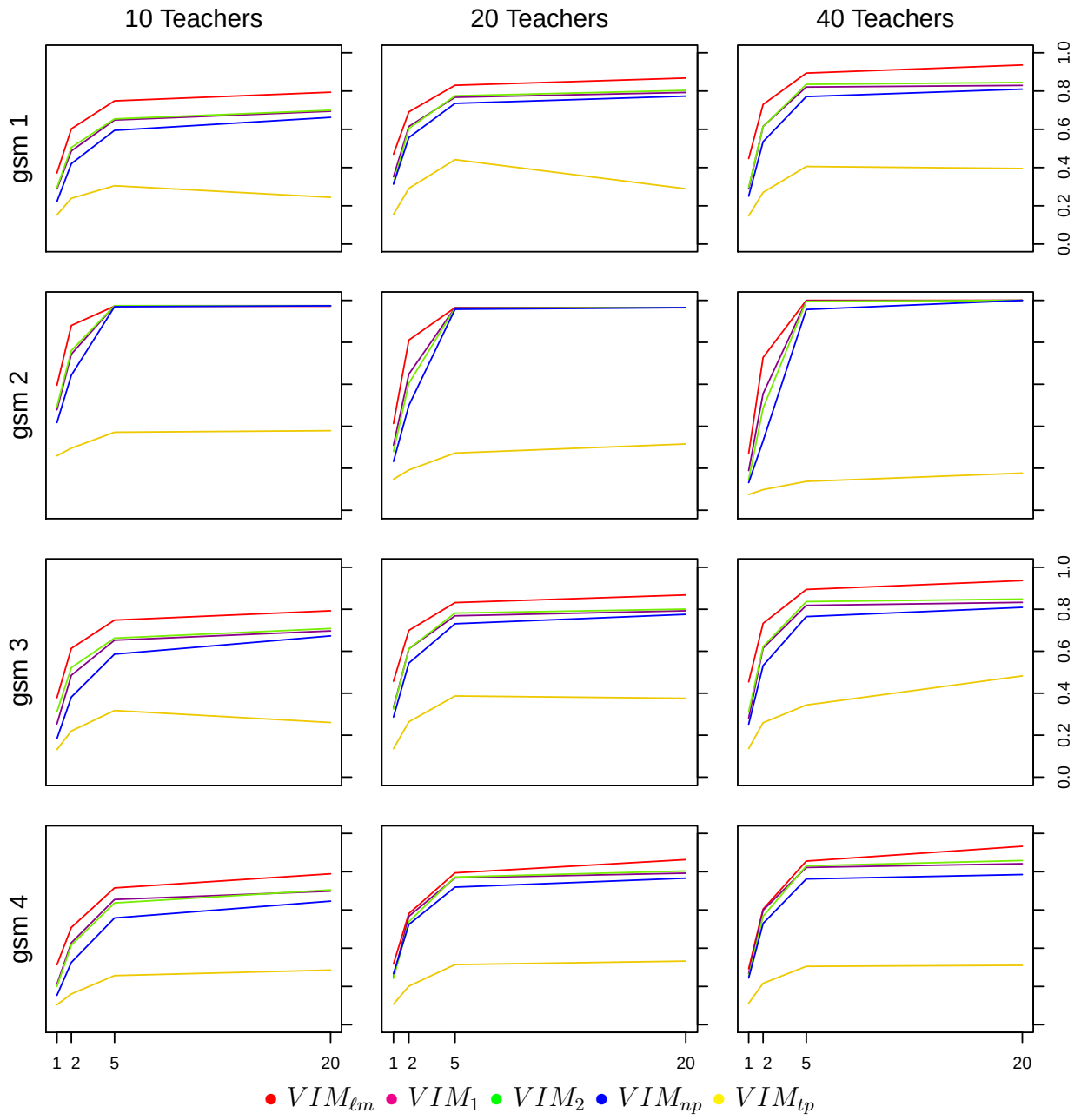

Figure A.11: Mean correlation between the VIMs and the absolute value of true teacher effects when the teacher variance over student variance σ_τ^2/σ^2 varies for different GSM models and different number of teachers when the number of students per teacher ratio is 30/18.

APPENDIX B

ADDITIONAL CHARTS AND TABLES FOR INTERACTION MEASURES

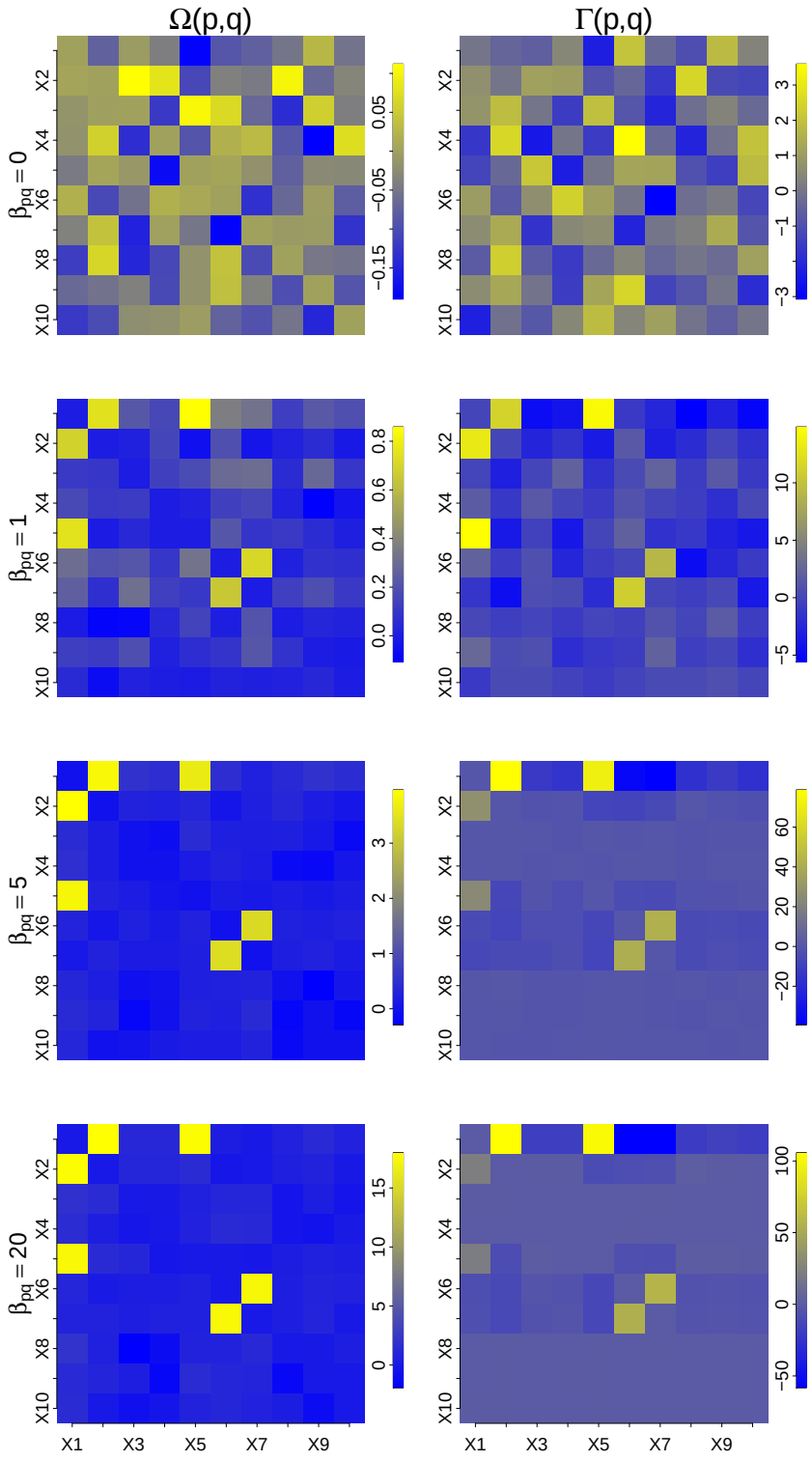

Figure B.1: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 10$, $\beta_p = 5$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .5$.

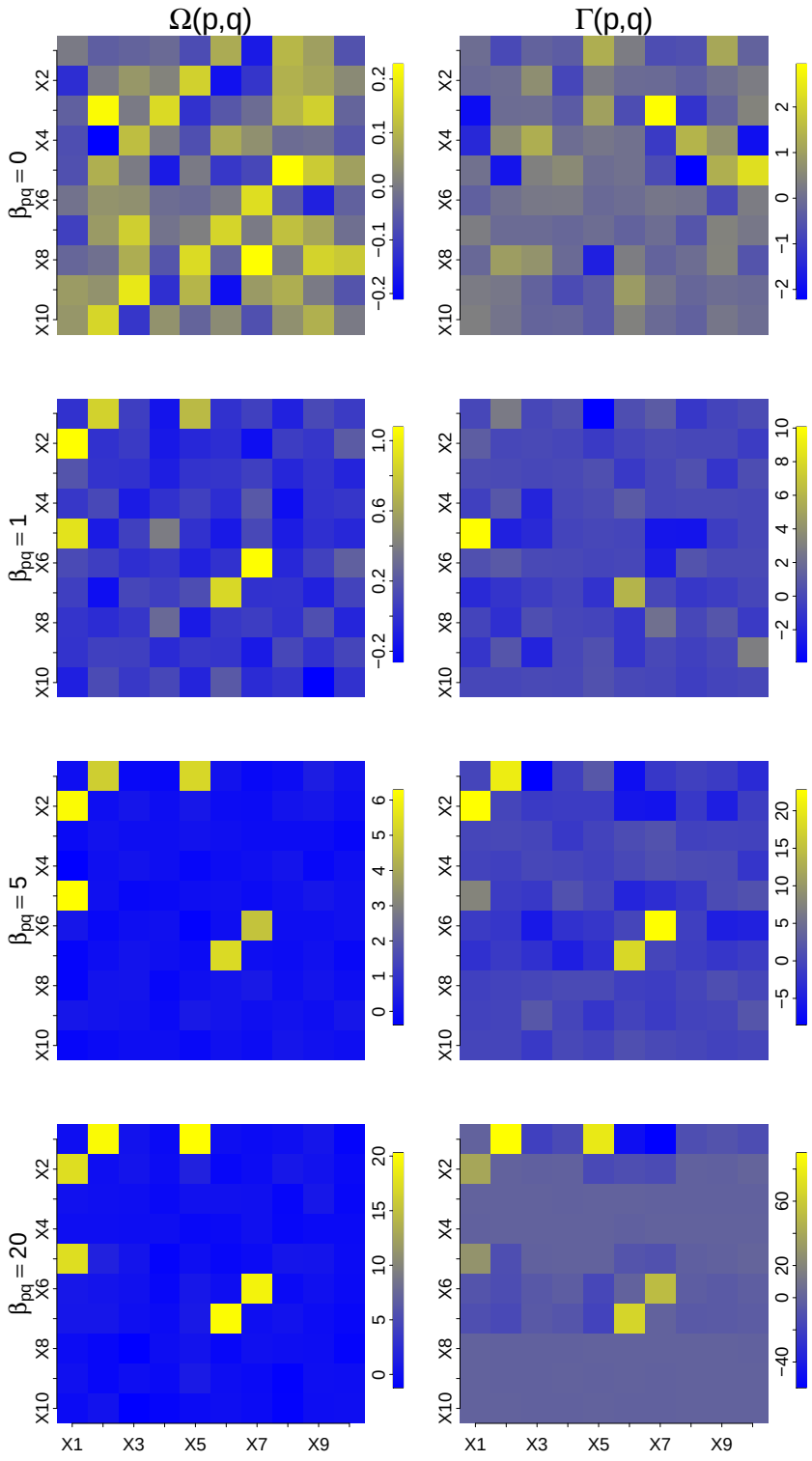

Figure B.2: The values of $\Omega(p, q)$ and $\Gamma(p, q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. $P = 10$, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .5$.

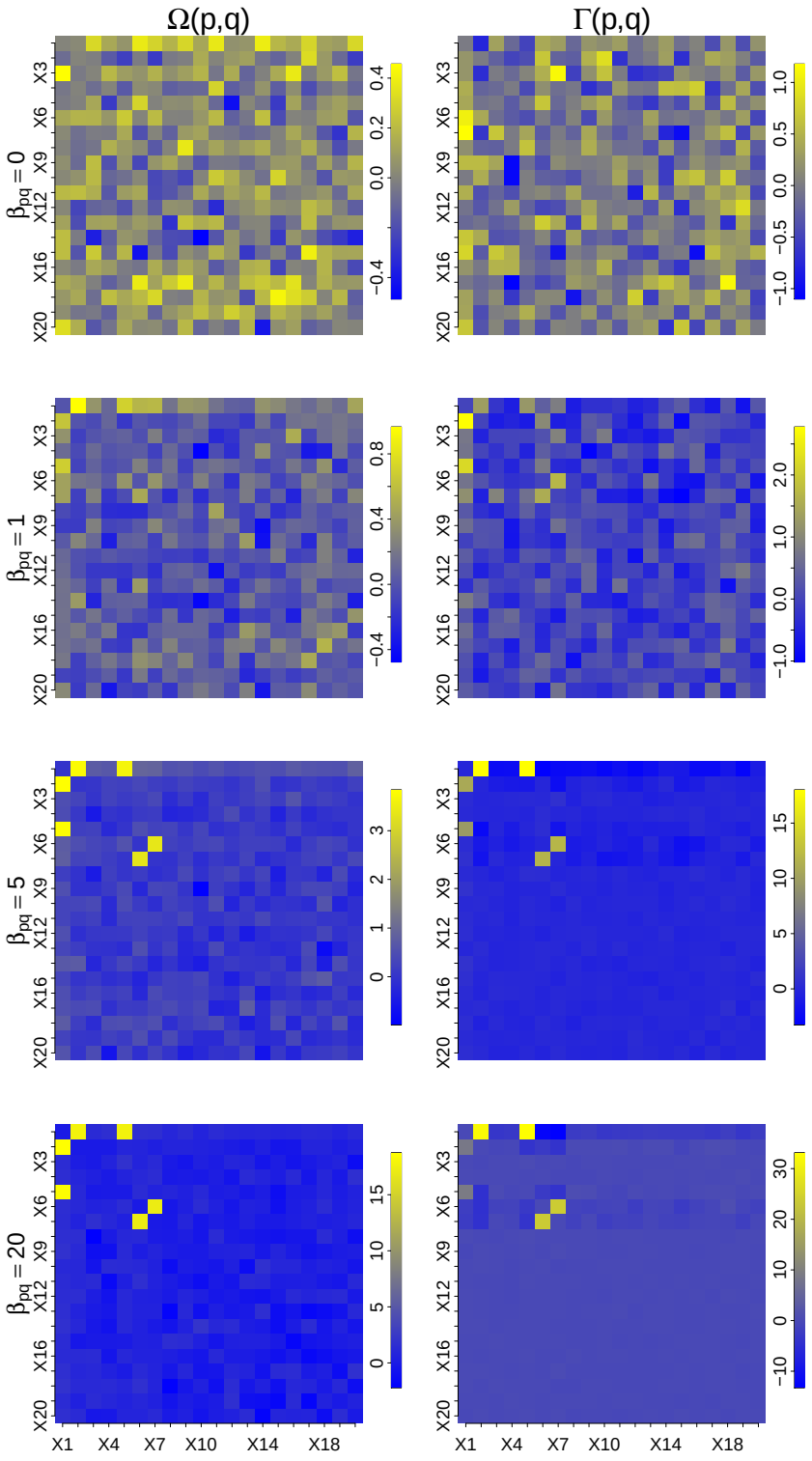

Figure B.3: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, $\beta_p = 5$ for all $p = 1, \ldots, 20$, and $P(X_p = 1) = .5$.

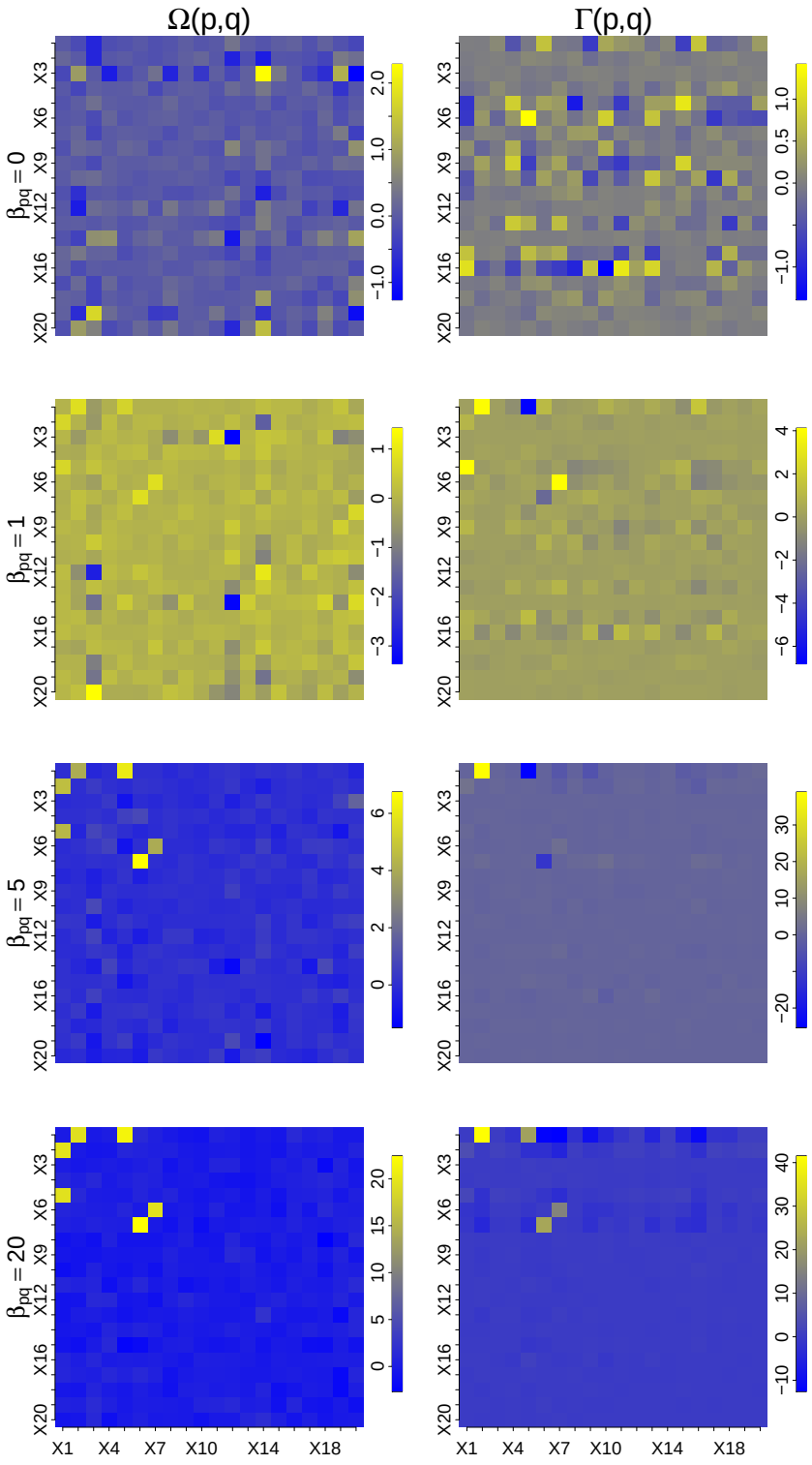

Figure B.4: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, β_p is sampled from $(-5,\ldots,5)$ for all $p = 1,\ldots,20$, and $P(X_p = 1) = .5$.

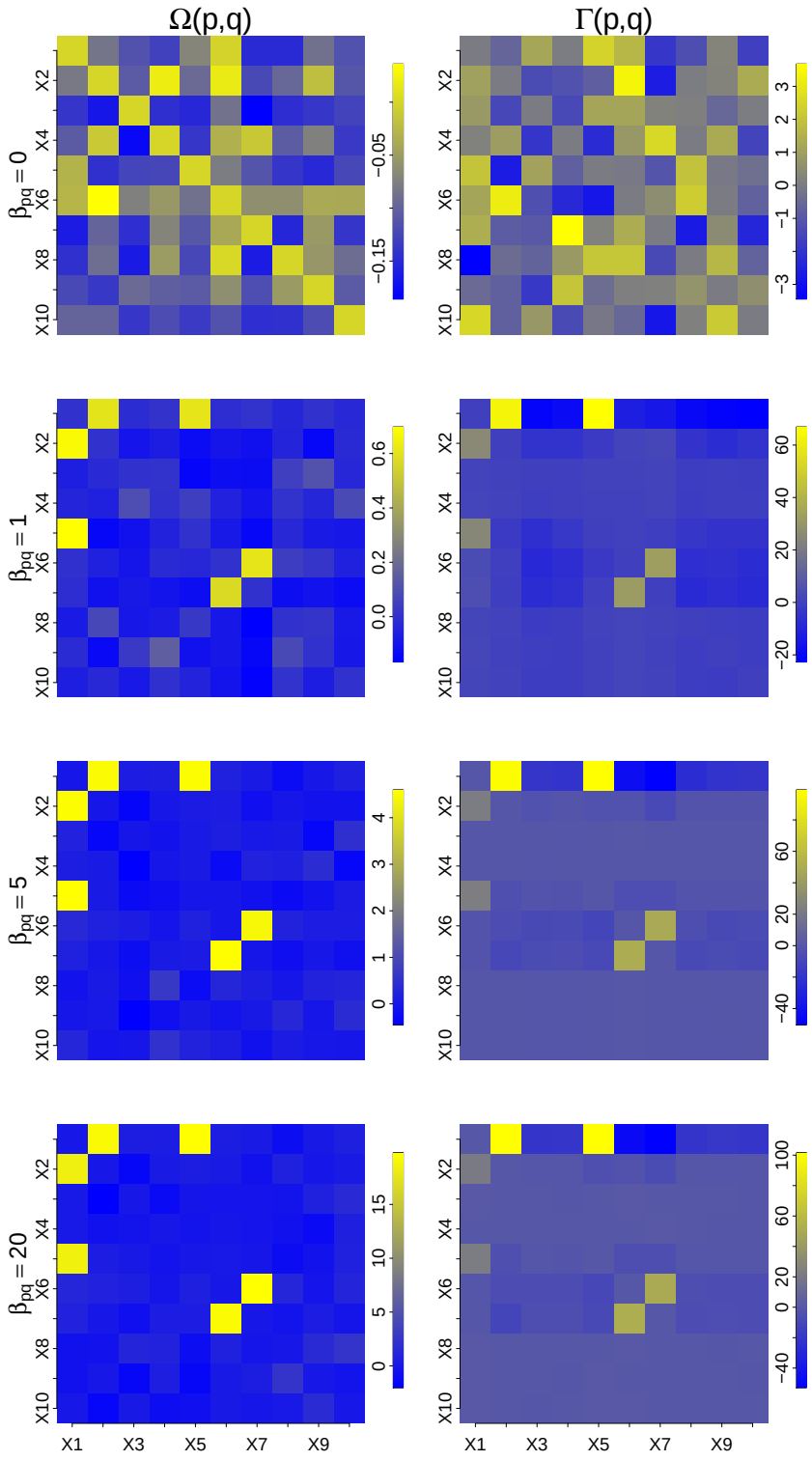

Figure B.5: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 10$, $\beta_p = 1$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .75$.

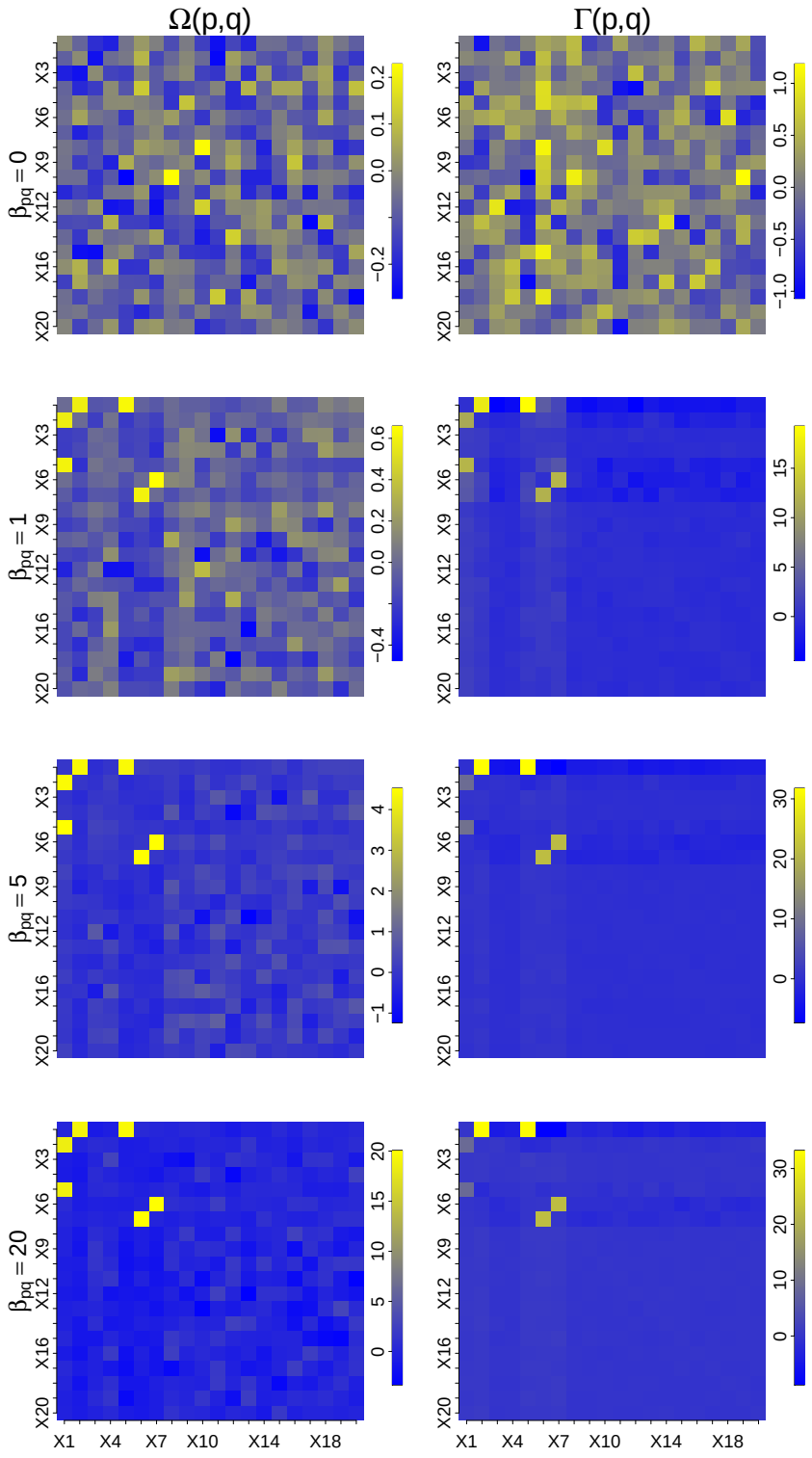

Figure B.6: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, $\beta_p = 1$ for all $p = 1, \ldots, 20$, and $P(X_p = 1) = .75$.

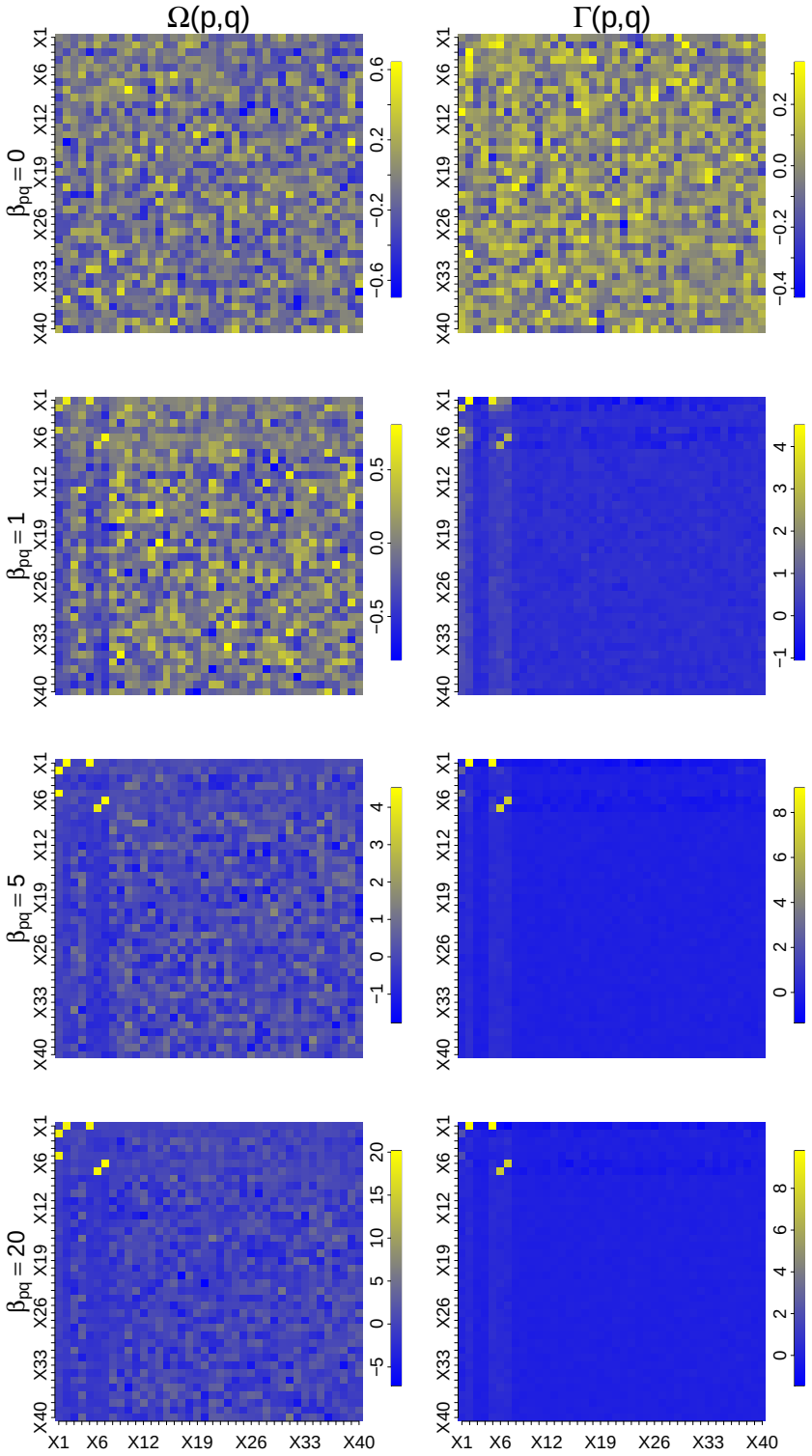

Figure B.7: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, $\beta_p = 1$ for all $p = 1, \ldots, 40$, and $P(X_p = 1) = .75$.

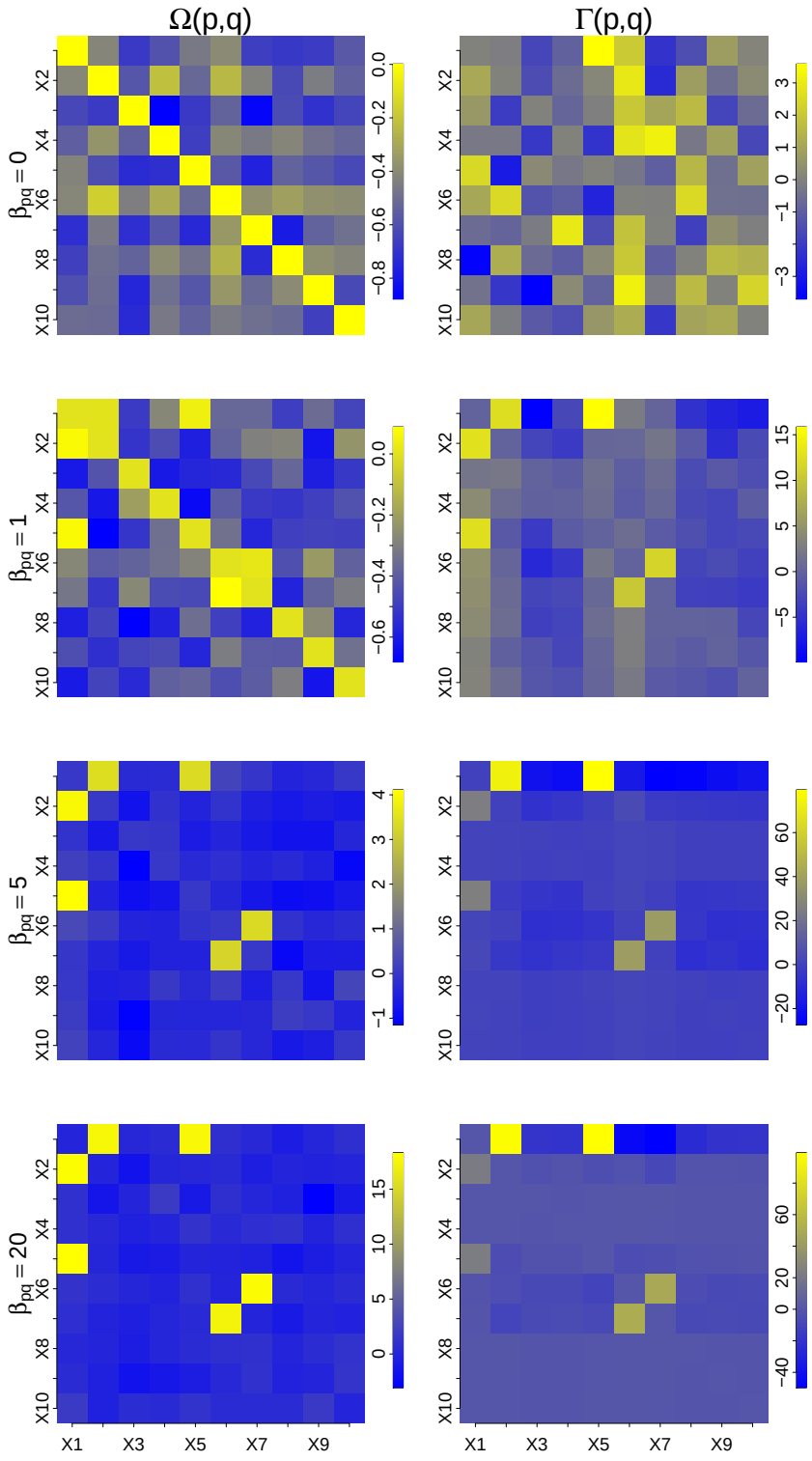

Figure B.8: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 10$, $\beta_p = 5$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .75$.

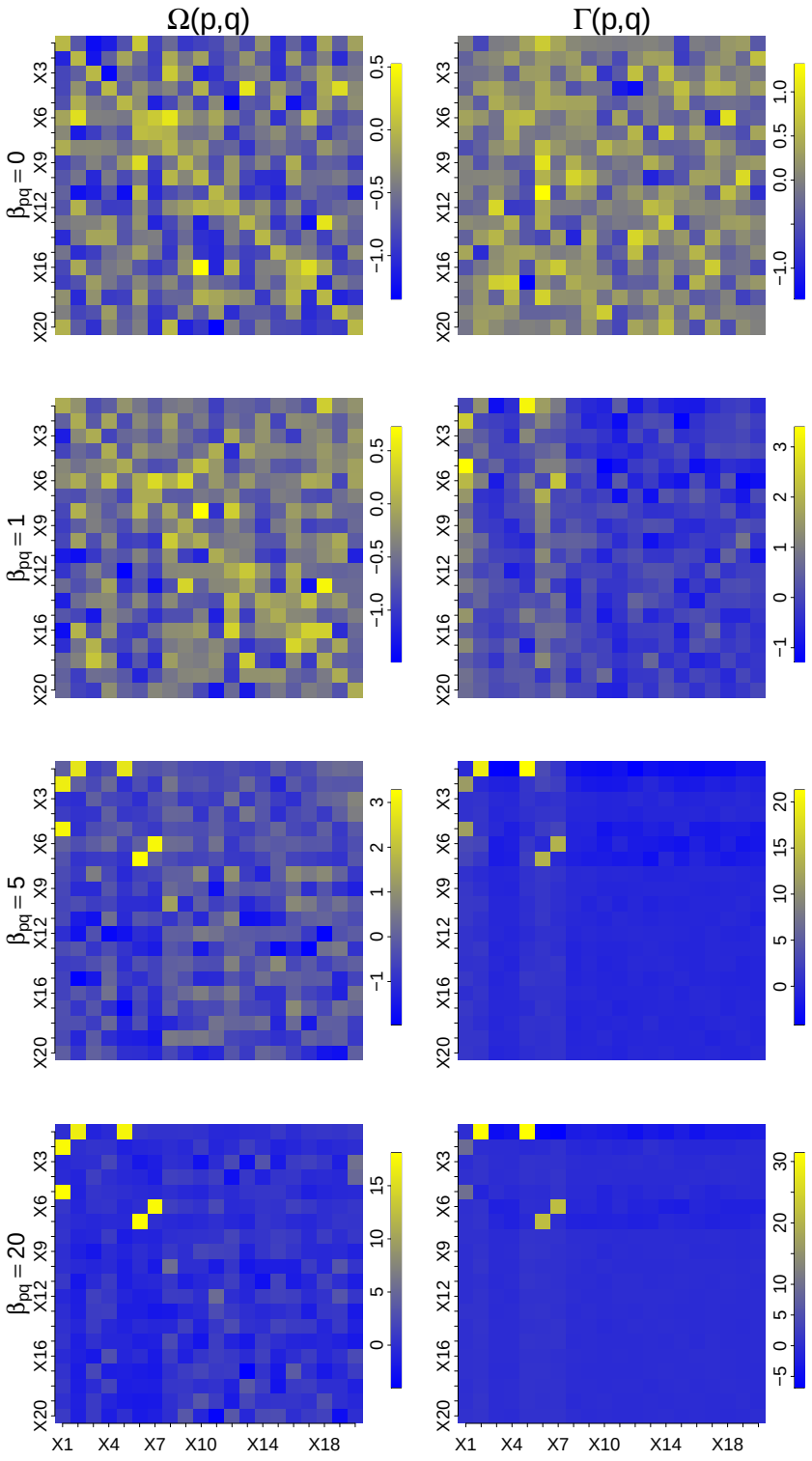

Figure B.9: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, $\beta_p = 5$ for all $p = 1, \ldots, 20$, and $P(X_p = 1) = .75$.

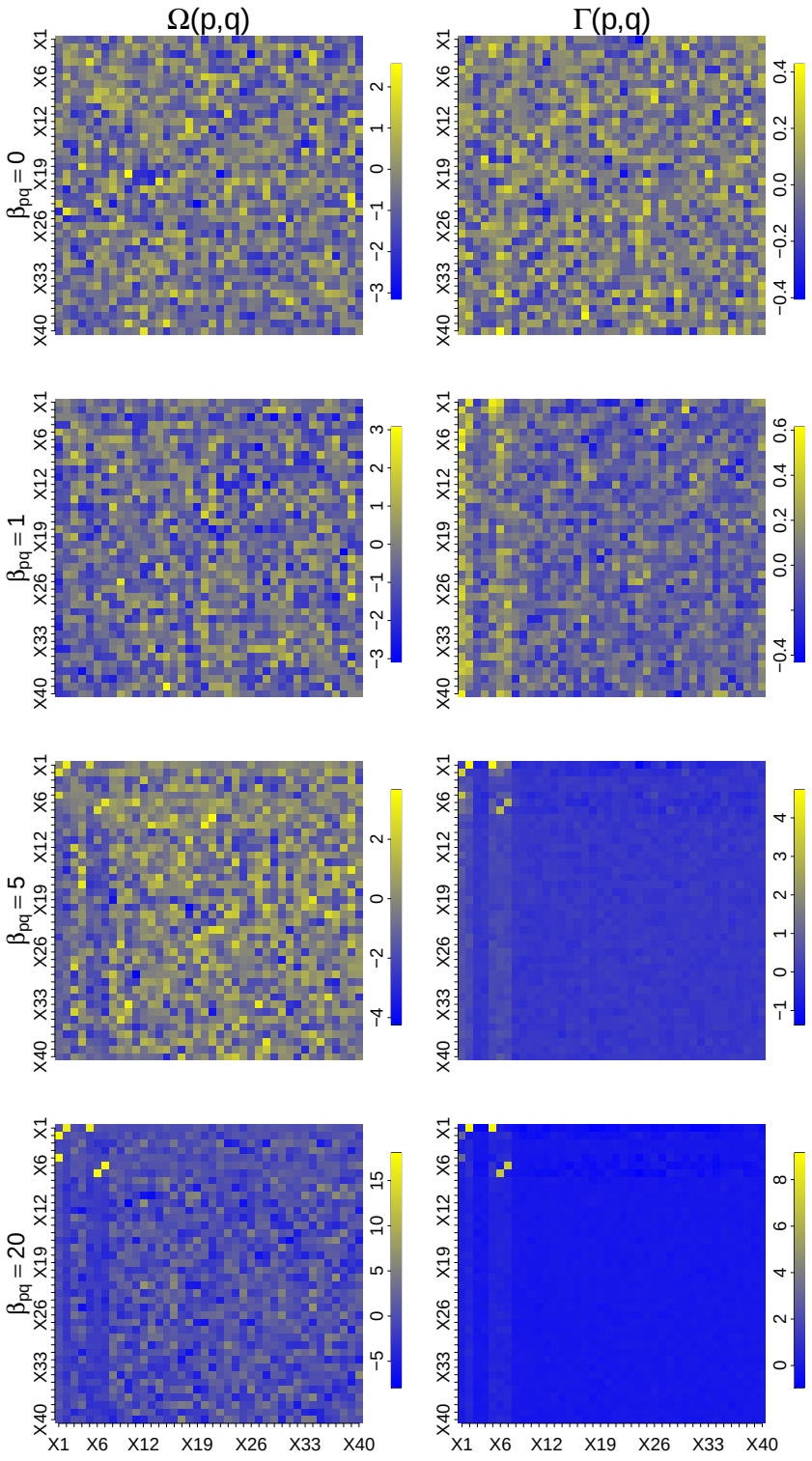

Figure B.10: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, $\beta_p = 5$ for all $p = 1, \ldots, 40$, and $P(X_p = 1) = .75$.

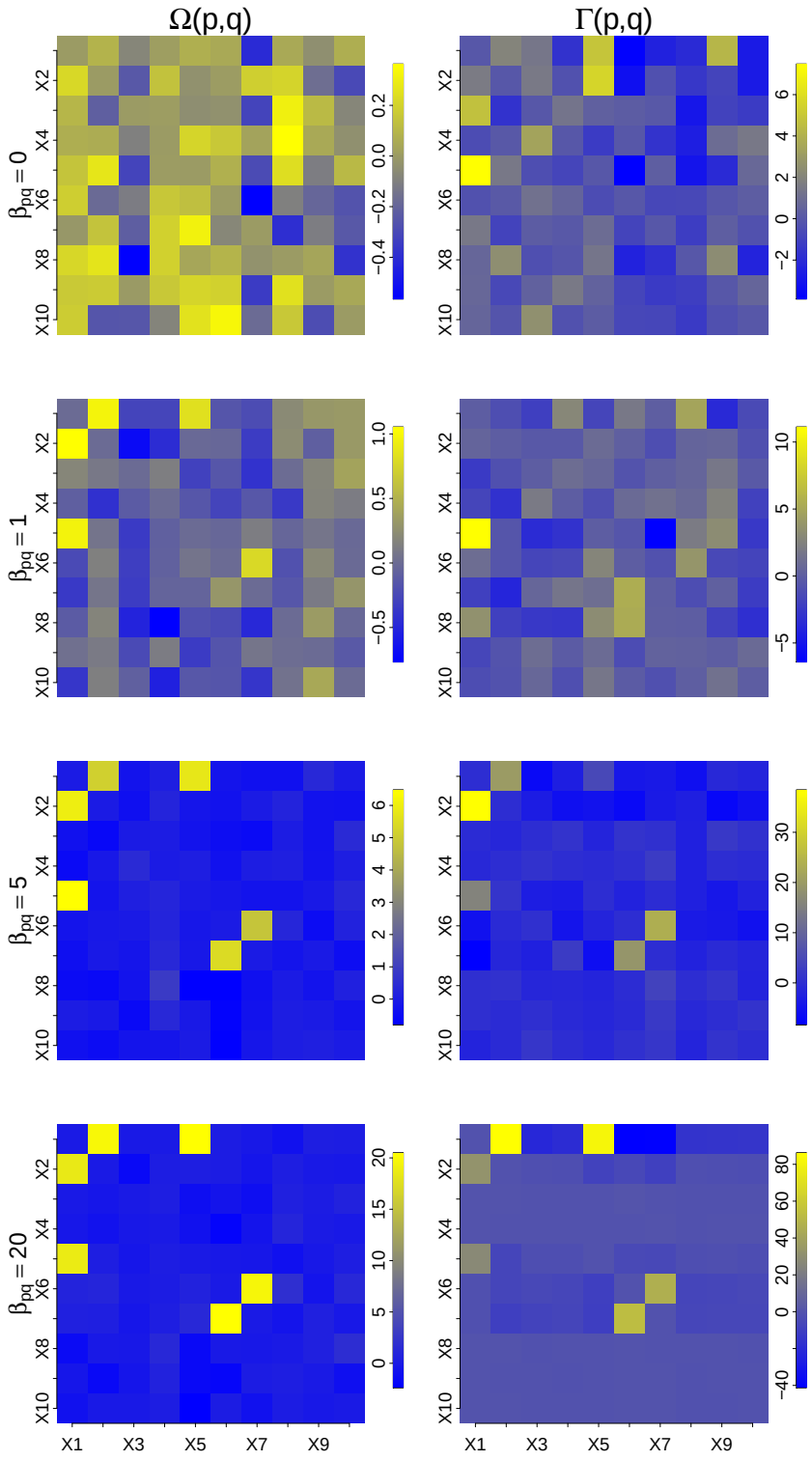

Figure B.11: The values of $\Omega(p, q)$ and $\Gamma(p, q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. $P = 10$, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 10$, and $P(X_p = 1) = .5$.

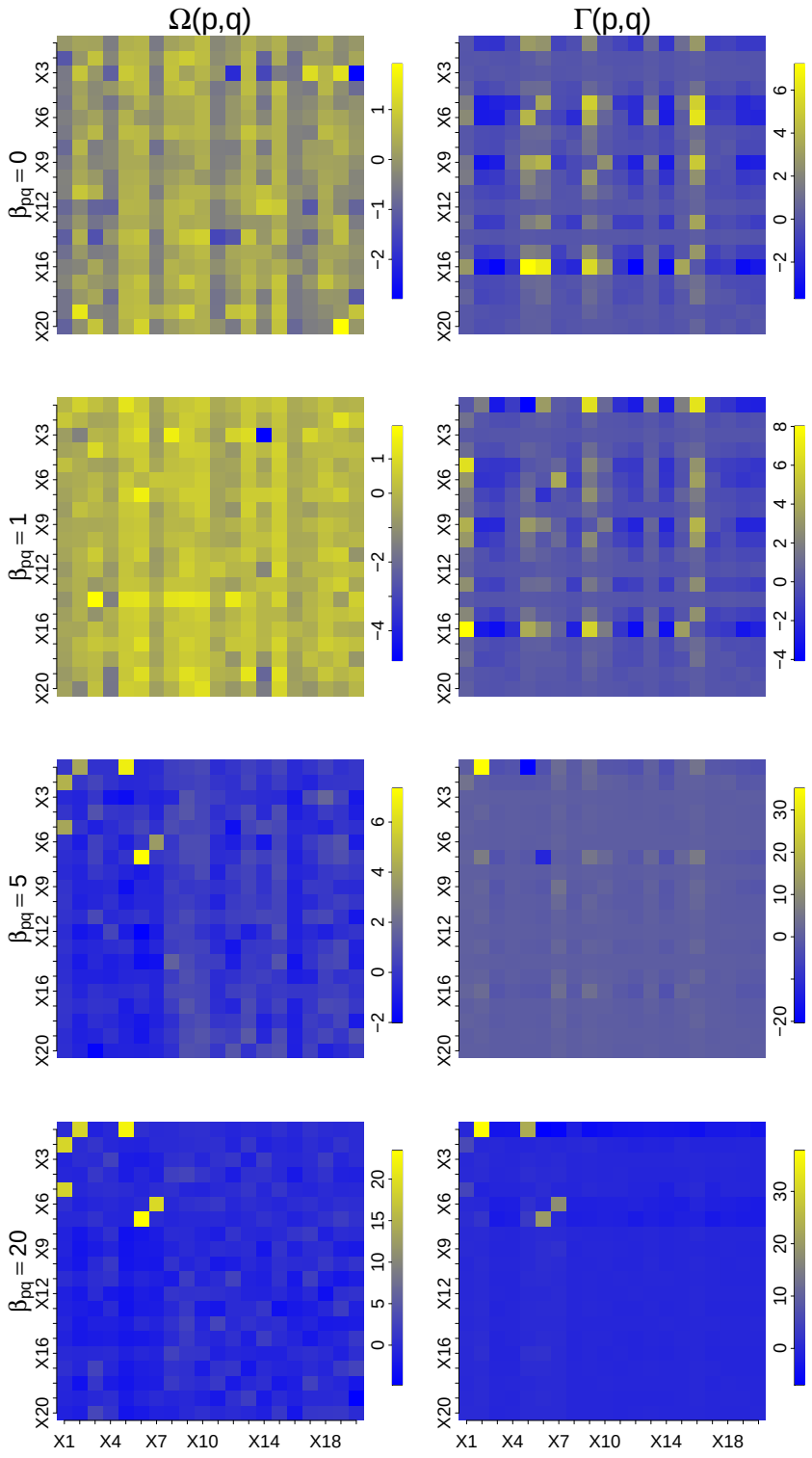

Figure B.12: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 20$, β_p is sampled from $(-5,\ldots,5)$ for all $p = 1,\ldots,20$, and $P(X_p = 1) = .5$.

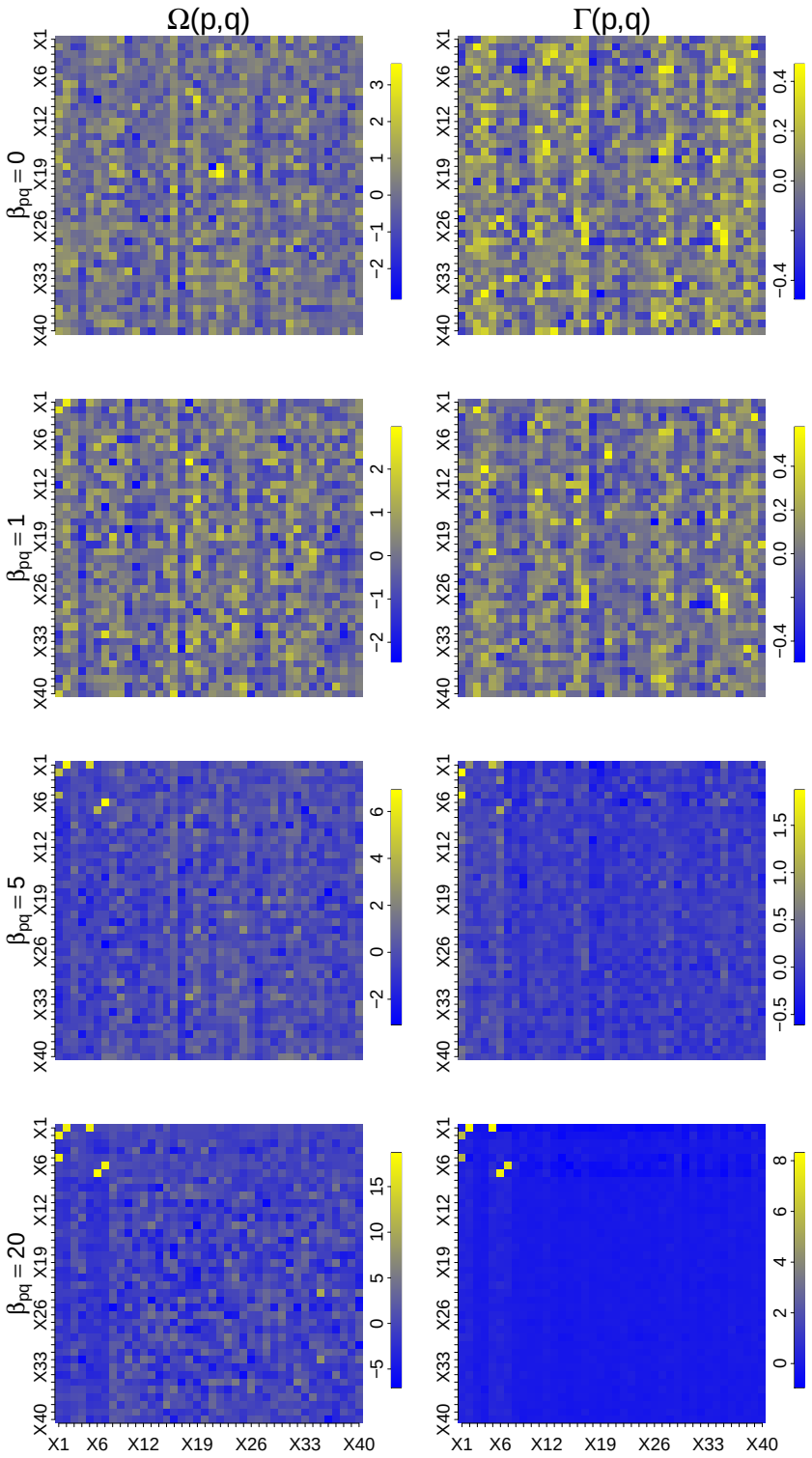

Figure B.13: The values of $\Omega(p,q)$ and $\Gamma(p,q)$ when the true interaction values, β_{pq}, varies from 0 to 20, for (p,q) or $(q,p) = (1,2), (1,5)$, and $(6,7)$. $P = 40$, β_p is sampled from $(-5,\ldots,5)$ for all $p = 1,\ldots,40$, and $P(X_p = 1) = .5$.

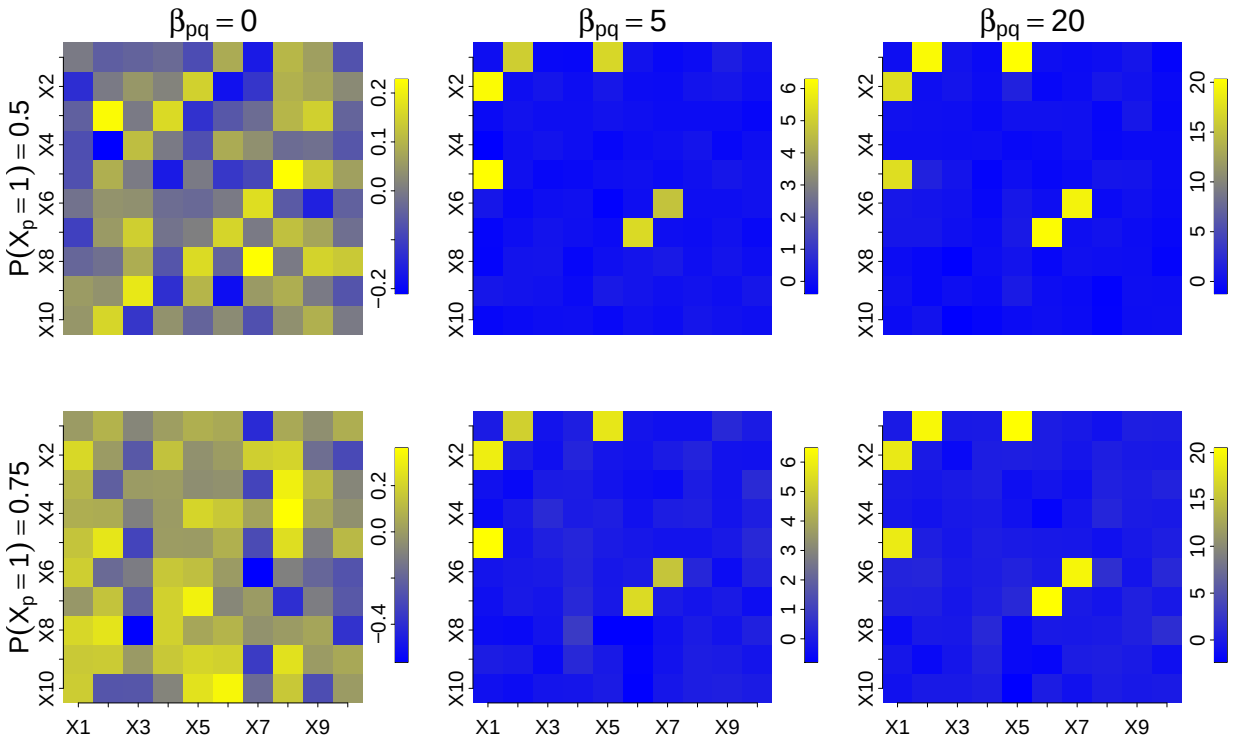

Figure B.14: Estimated interaction effects for mean-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 10 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 10$.

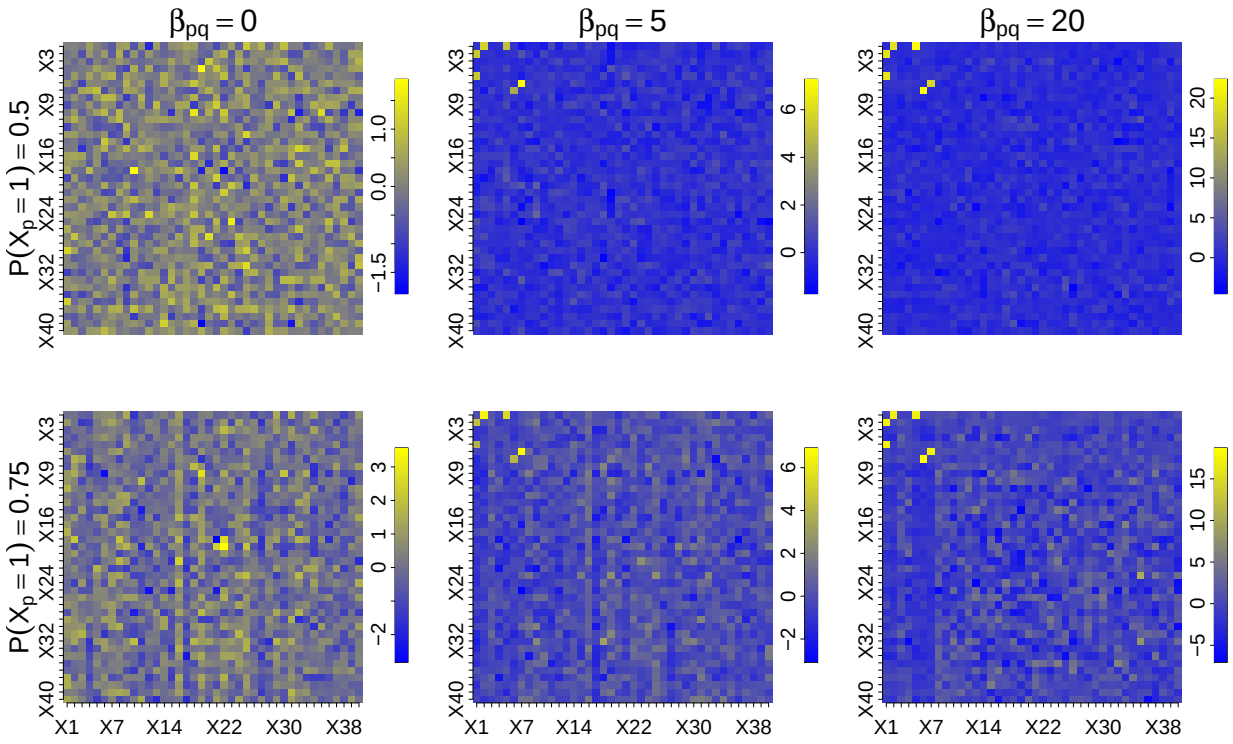

Figure B.15: Estimated interaction effects for mean-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 40 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 40$.

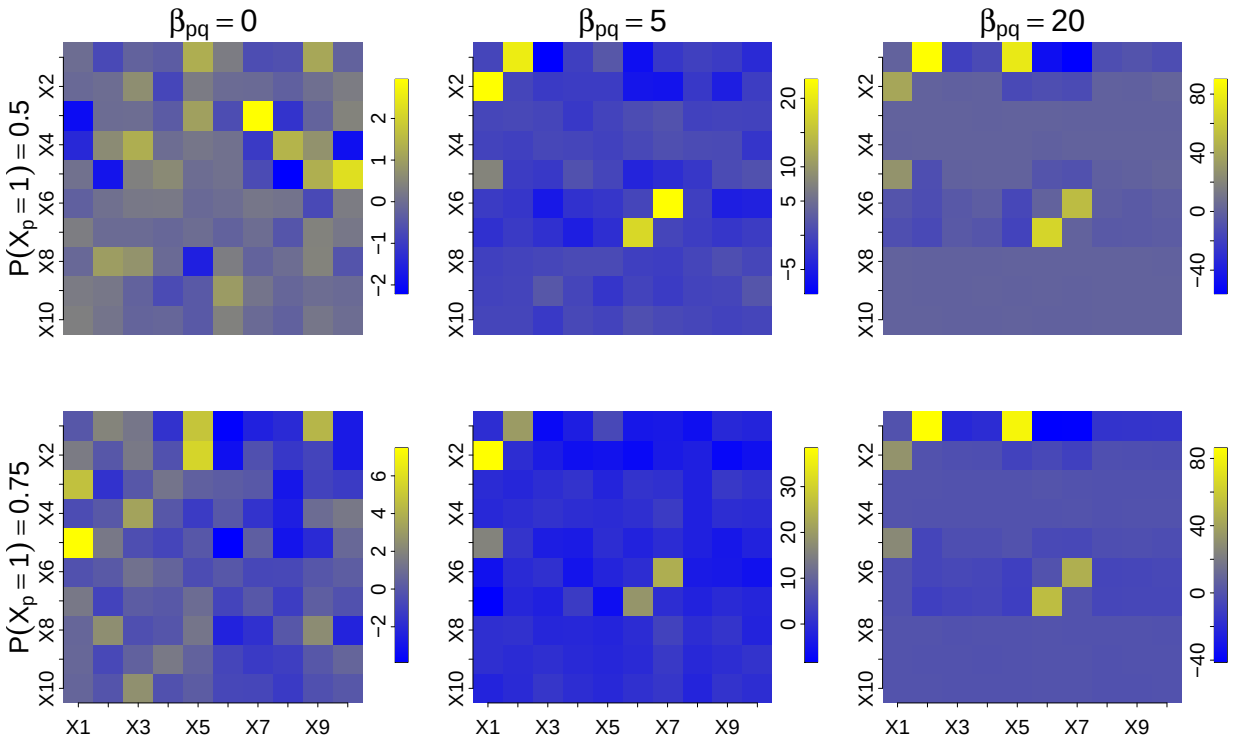

Figure B.16: Estimated interaction effects for distribution-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 10 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 10$.

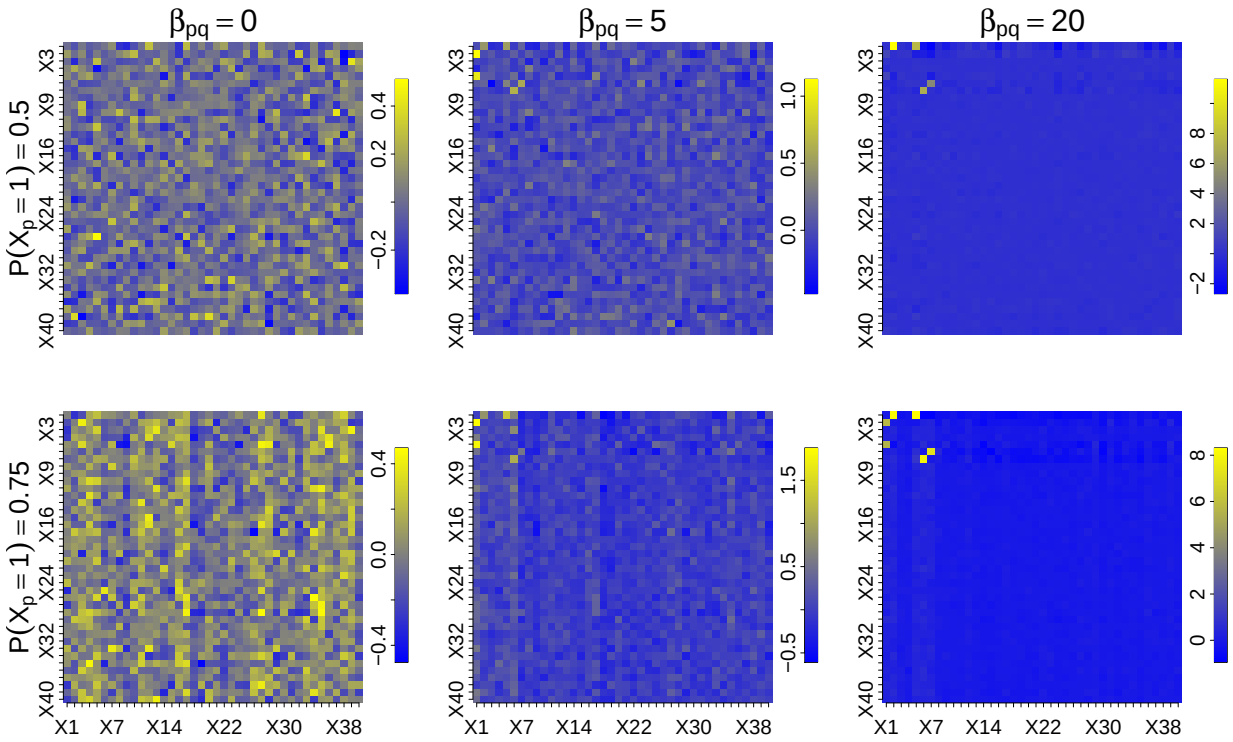

Figure B.17: Estimated interaction effects for distribution-based interaction measure when X_p success probability, $P(X_p = 1)$, is either .5 or .75 for all $p = 1, \ldots, 20$, and the true interaction effect, β_{pq}, varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, and $(6, 7)$. 40 variables are considered, β_p is sampled from $(-5, \ldots, 5)$ for all $p = 1, \ldots, 40$.

β_p	Num. Variables (P)	$Prob(X_p = 1)$	β_{pq}	$\hat{\beta}_{pq}$	Bias($\beta_{pq} \neq 0$)	Bias($\beta_{pq} = 0$)
5	10	0.50	0	-0.10	-0.10	-0.04
5	20	0.50	0	-0.00	-0.00	0.03
5	40	0.50	0	-0.05	-0.05	0.03
5	10	0.75	0	-0.41	-0.41	-0.54
5	20	0.75	0	-0.52	-0.52	-0.58
5	40	0.75	0	-0.80	-0.80	-0.53
5	10	0.50	1	0.73	-0.27	0.10
5	20	0.50	1	0.64	-0.36	0.04
5	40	0.50	1	0.50	-0.50	0.03
5	10	0.75	1	0.05	-0.95	-0.45
5	20	0.75	1	-0.17	-1.17	-0.53
5	40	0.75	1	0.39	-0.61	-0.49
5	10	0.50	5	3.66	-1.34	0.19
5	20	0.50	5	3.62	-1.38	0.13
5	40	0.50	5	3.22	-1.78	0.08
5	10	0.75	5	3.56	-1.44	-0.38
5	20	0.75	5	2.98	-2.02	-0.46
5	40	0.75	5	2.68	-2.32	-0.46
5	10	0.50	10	8.20	-1.80	0.33
5	20	0.50	10	8.09	-1.91	0.27
5	40	0.50	10	7.57	-2.43	0.11
5	10	0.75	10	8.21	-1.79	-0.12
5	20	0.75	10	7.75	-2.25	-0.23
5	40	0.75	10	7.43	-2.57	-0.29
5	10	0.50	20	17.56	-2.44	0.52
5	20	0.50	20	17.82	-2.18	0.42
5	40	0.50	20	17.61	-2.39	0.23
5	10	0.75	20	17.80	-2.20	0.18
5	20	0.75	20	17.49	-2.51	0.06
5	40	0.75	20	17.20	-2.80	-0.06

Table B.1: Average interaction estimation bias for those pairs of covariates that interact, Bias($\beta_{pq} \neq 0$), and those that do not interact, Bias($\beta_{pq} = 0$), when $P = 10, 20$, or 40, $\beta_p = 5$, $Prob(X_p = 1) = 0.5$ or 0.75 for $p = 1, \ldots, P$, and β_{pq} varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, or $(6, 7)$.

β_p	Num. Variables (P)	$Prob(X_p = 1)$	β_{pq}	$\hat{\beta}_{pq}$	Bias($\beta_{pq} \neq 0$)	Bias($\beta_{pq} = 0$)
(-5:5)	10	0.50	0	-0.00	-0.00	0.02
(-5:5)	20	0.50	0	-0.03	-0.03	0.00
(-5:5)	40	0.50	0	-0.24	-0.24	-0.01
(-5:5)	10	0.75	0	-0.02	-0.02	-0.00
(-5:5)	20	0.75	0	-0.12	-0.12	0.07
(-5:5)	40	0.75	0	0.31	0.31	0.01
(-5:5)	10	0.50	1	0.92	-0.08	0.02
(-5:5)	20	0.50	1	0.71	-0.29	-0.05
(-5:5)	40	0.50	1	0.87	-0.13	0.02
(-5:5)	10	0.75	1	0.82	-0.18	-0.08
(-5:5)	20	0.75	1	0.76	-0.24	0.03
(-5:5)	40	0.75	1	1.73	0.73	-0.03
(-5:5)	10	0.50	5	5.45	0.45	0.01
(-5:5)	20	0.50	5	4.99	-0.01	0.01
(-5:5)	40	0.50	5	5.48	0.48	0.00
(-5:5)	10	0.75	5	5.59	0.59	-0.12
(-5:5)	20	0.75	5	5.06	0.06	-0.12
(-5:5)	40	0.75	5	5.21	0.21	-0.12
(-5:5)	10	0.50	10	9.80	-0.20	0.00
(-5:5)	20	0.50	10	9.92	-0.08	0.04
(-5:5)	40	0.50	10	10.07	0.07	-0.01
(-5:5)	10	0.75	10	9.98	-0.02	-0.12
(-5:5)	20	0.75	10	10.04	0.04	-0.18
(-5:5)	40	0.75	10	10.30	0.30	-0.16
(-5:5)	10	0.50	20	19.02	-0.98	-0.02
(-5:5)	20	0.50	20	20.20	0.20	0.06
(-5:5)	40	0.50	20	20.10	0.10	0.06
(-5:5)	10	0.75	20	19.58	-0.42	-0.09
(-5:5)	20	0.75	20	20.14	0.14	-0.36
(-5:5)	40	0.75	20	17.59	-2.41	-0.29

Table B.2: Average interaction estimation bias for those pairs of covariates that interact, Bias($\beta_{pq} \neq 0$), and those that do not interact, Bias($\beta_{pq} = 0$), when $P = 10, 20$, or 40, β_p is sampled from $(-5, \ldots, 5)$, $Prob(X_p = 1) = 0.5$ or 0.75 for all $p = 1, \ldots, P$, and β_{pq} varies from 0 to 20 for (p, q) or $(q, p) = (1, 2), (1, 5)$, or $(6, 7)$.

APPENDIX C
THE ASYMMETRICAL CASE

In what follows, we group summation terms based on the partitions in η and $\tilde{\eta}$. As an illustration,

$$\begin{aligned}\{\ell, m : S_\ell, S_m \in H(\eta) \cup H(\tilde{\eta})\} &= \{\ell, m : S_\ell, S_m \in H(\eta)\backslash H(\tilde{\eta})\} \\ &\cup \{\ell, m : S_\ell, S_m \in H(\tilde{\eta})\backslash H(\eta)\} \cup \{\ell, m : S_\ell, S_m \in H(\eta) \cap H(\tilde{\eta})\} \\ &\cup \{\ell, m : S_\ell \in H(\eta)\backslash H(\tilde{\eta}) \text{ and } S_m \in H(\tilde{\eta})\backslash H(\eta)\} \\ &\cup \{\ell, m : S_\ell \in H(\eta)\backslash H(\tilde{\eta}) \text{ and } S_m \in H(\eta) \cap H(\tilde{\eta})\} \\ &\cup \{\ell, m : S_\ell \in H(\eta) \cap H(\tilde{\eta}) \text{ and } S_m \in H(\tilde{\eta})\backslash H(\eta)\} \end{aligned} \tag{C.1}$$

We rewrite $\bar{y}(q;\eta)$ as $\bar{y}(q;\eta) = \Delta_3(H(q;\eta)) + \Delta_4(H(q;\eta))$ where

$$\begin{aligned}\Delta_3(H(q;\eta)) &= \beta_0 + \beta_q + \beta_p + \beta_{pq} \\ &+ \sum_{\ell \in H(\eta)\backslash[H(\tilde{\eta})\cup\{p\}]} \beta_\ell + \sum_{\ell \in H(\tilde{\eta})\cap H(\eta)} \beta_\ell + \sum_{\ell \in H(\tilde{\eta})\backslash H(\eta)} \frac{|\eta(\ell,q)|}{|\eta(q)|}\beta_\ell \\ &+ \sum_{\ell \in H(\eta)\cap H(\tilde{\eta})} \beta_{\ell q} + \sum_{\ell \in H(\eta)\backslash[H(\tilde{\eta})\cup\{p\}]} \beta_{\ell q} + \sum_{\ell \in H(\tilde{\eta})\backslash H(\eta)} \frac{|\eta(\ell,q)|}{|\eta(q)|}\beta_{\ell q} \\ &+ \sum_{\substack{\ell,m \in H(\eta)\cap H(\tilde{\eta}) \\ \ell<m}} \beta_{\ell m} + \sum_{\substack{\ell,m \in H(\eta)\backslash H(\tilde{\eta}) \\ \ell<m}} \beta_{\ell m} + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \in H(\eta)\backslash H(\tilde{\eta})}} \beta_{\ell m} \\ &+ \sum_{\substack{\ell \in H(\eta)\backslash H(\tilde{\eta}) \\ m \in H(\tilde{\eta})\backslash H(\eta)}} \frac{|\eta(m,q)|}{|\eta(q)|}\beta_{\ell m} + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \in H(\tilde{\eta})\backslash H(\eta)}} \frac{|\eta(m,q)|}{|\eta(q)|}\beta_{\ell m} \\ &+ \sum_{\substack{\ell,m \in H(\tilde{\eta})\backslash H(\eta) \\ \ell<m}} \frac{|\eta(\ell,m,q)|}{|\eta(q)|}\beta_{\ell m} \end{aligned} \tag{C.2}$$

and

$$\begin{aligned}\Delta_4(H(q;\eta)) &= \sum_{\ell \notin H(q;\eta)\cup H(q;\tilde{\eta})} \frac{|\eta(\ell,q)|}{|\eta(q)|}\beta_\ell \\ &+ \sum_{\ell \notin H(q;\eta)\cup H(q;\tilde{\eta})} \frac{|\eta(\ell,q)|}{|\eta(q)|}\beta_{\ell q} + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}} \frac{|\eta(m,q)|}{|\eta(q)|}\beta_{\ell m} \\ &+ \sum_{\substack{\ell \in H(\eta)\backslash H(\tilde{\eta}) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}} \frac{|\eta(m,q)|}{|\eta(q)|}\beta_{\ell m} + \sum_{\substack{\ell \in H(\tilde{\eta})\backslash H(\eta) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}} \frac{|\eta(\ell,m,q)|}{|\eta(q)|}\beta_{\ell m} \\ &+ \sum_{\substack{\ell,m \notin H(q;\eta)\cup H(q;\tilde{\eta}) \\ \ell<m}} \frac{|\eta(\ell,m,q)|}{|\eta(q)|}\beta_{\ell m} + \bar{\epsilon}(q;\eta). \end{aligned} \tag{C.3}$$

The difference of differences for $\Delta_3(\cdot)$ and $\Delta_4(\cdot)$ are given by

$$
\begin{aligned}
&\Delta_3\Big(H(q;\eta)\Big) - \Delta_3\Big(H(q^c;\eta)\Big) - \Delta_3\Big(H(q;\tilde{\eta})\Big) + \Delta_3\Big(H(q^c;\tilde{\eta})\Big) = \quad \beta_{pq} + \\
&\quad + \sum_{\ell \in H(\eta)\setminus[H(\tilde{\eta})\cup\{p\}]} \left(1 - \frac{|\tilde{\eta}(\ell,q)|}{|\tilde{\eta}(q)|}\right)\beta_{\ell q} - \sum_{\ell \in H(\tilde{\eta})\setminus H(\eta)} \left(1 - \frac{|\eta(\ell,q)|}{|\eta(q)|}\right)\beta_{\ell q} \\
&\quad + \sum_{\ell \in H(\eta)\setminus H(\tilde{\eta})} \left(-\frac{|\tilde{\eta}(\ell,q)|}{|\tilde{\eta}(q)|} + \frac{|\tilde{\eta}(\ell,q^c)|}{|\tilde{\eta}(q^c)|}\right)\beta_{\ell} \\
&\quad + \sum_{\ell \in H(\tilde{\eta})\setminus H(\eta)} \left(\frac{|\eta(\ell,q)|}{|\eta(q)|} - \frac{|\eta(\ell,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell} \\
&\quad + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \in H(\eta)\setminus H(\tilde{\eta})}}\sum \left(-\frac{|\tilde{\eta}(m,q)|}{|\tilde{\eta}(q)|} + \frac{|\tilde{\eta}(m,q^c)|}{|\tilde{\eta}(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \in H(\tilde{\eta})\setminus H(\eta)}}\sum \left(\frac{|\eta(m,q)|}{|\eta(q)|} - \frac{|\eta(m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell \in H(\eta)\setminus H(\tilde{\eta}) \\ m \in H(\tilde{\eta})\setminus H(\eta)}}\sum \left(\frac{|\eta(m,q)|}{|\eta(q)|} - \frac{|\eta(m,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(m,q)|}{|\tilde{\eta}(q)|} + \frac{|\tilde{\eta}(m,q^c)|}{|\tilde{\eta}(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell,m \in H(\eta)\setminus H(\tilde{\eta}) \\ \ell<m}}\sum \left(\frac{-|\tilde{\eta}(\ell,m,q)|}{|\tilde{\eta}(q)|} + \frac{|\tilde{\eta}(\ell,m,q^c)|}{|\tilde{\eta}(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell,m \in H(\tilde{\eta})\setminus H(\eta) \\ \ell<m}}\sum \left(\frac{|\eta(\ell,m,q)|}{|\eta(q)|} - \frac{|\eta(\ell,m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m}, \qquad \text{(C.4)} \\
&\Delta_4\Big(H(q;\eta)\Big) - \Delta_4\Big(H(q^c;\eta)\Big) - \Delta_4\Big(H(q;\tilde{\eta})\Big) + \Delta_4\Big(H(q^c;\tilde{\eta})\Big) = \\
&\quad = \sum_{\ell \notin H(q;\eta)\cup H(q;\tilde{\eta})} \left(\frac{|\eta(\ell,q)|}{|\eta(q)|} - \frac{|\eta(\ell,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(\ell,q)|}{|\tilde{\eta}(q)|} + \frac{|\eta(\ell,q^c)|}{|\eta(q^c)|}\right)(\beta_{\ell} + \beta_{\ell q}) \\
&\quad + \sum_{\substack{\ell \in H(\eta)\cap H(\tilde{\eta}) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}}\sum \left(\frac{|\eta(m,q)|}{|\eta(q)|} - \frac{|\eta(m,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(m,q)|}{|\tilde{\eta}(q)|} + \frac{|\eta(m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell \in H(\eta)\setminus H(\tilde{\eta}) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}}\sum \left(\frac{|\eta(m,q)|}{|\eta(q)|} - \frac{|\eta(m,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(m,q)|}{|\tilde{\eta}(q)|} + \frac{|\eta(m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell \in H(\tilde{\eta})\setminus H(\eta) \\ m \notin H(q;\eta)\cup H(q;\tilde{\eta})}}\sum \left(\frac{|\eta(\ell,m,q)|}{|\eta(q)|} - \frac{|\eta(\ell,m,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(\ell,m,q)|}{|\tilde{\eta}(q)|} + \frac{|\eta(\ell,m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m} \\
&\quad + \sum_{\substack{\ell,m \notin H(q;\eta)\cup H(q;\tilde{\eta}) \\ \ell<m}}\sum \left(\frac{|\eta(\ell,m,q)|}{|\eta(q)|} - \frac{|\eta(\ell,m,q^c)|}{|\eta(q^c)|} - \frac{|\tilde{\eta}(\ell,m,q)|}{|\tilde{\eta}(q)|} + \frac{|\eta(\ell,m,q^c)|}{|\eta(q^c)|}\right)\beta_{\ell m} \\
&\quad + \bar{\epsilon}(q;\eta) - \bar{\epsilon}(q^c;\eta) - \bar{\epsilon}(q;\tilde{\eta}) + \bar{\epsilon}(q^c;\tilde{\eta}). \qquad \text{(C.5)}
\end{aligned}
$$