QDA.SEM: Combining QDA and SEM to Improve Predictive Accuracy in Classifications: A Simulation Study

Laane, Florian

Universiteit Leiden

Faculteit der Sociale Wetenschappen

QDA.SEM: Combining QDA and SEM to Improve Predictive Accuracy in Classifications

A Simulation Study

F.R.M. Laane

Master's Thesis Methodology and Statistics in Psychology

Methodology and Statistics Unit, Institute of Psychology
Faculty of Social and Behavioral Sciences, Leiden University

Date: 15 March 2025

Student number: s3966046
Supervisor: Dr. J.D. Karch

Abstract

In recent years, the focus in statistics has expanded beyond explanation to include prediction as a central goal. De Rooij et al. (2023) developed a prediction method that can use reflective Structural Equation Models to make out-of-sample predictions on the indicator level. However, their method does not support classification. Building on their work, this thesis introduces QDA.SEM, a new classification method combining Structural Equation Modeling (SEM) with Quadratic Discriminant Analysis (QDA). By incorporating prior theoretical knowledge through SEM, class-specific model-implied covariance matrices are estimated and imputed in the probability density functions used in the QDA framework to calculate posterior probabilities. The estimates of the model-implied covariance matrices separate measurement variance from structural variance, this introduces bias but reduces variance, effectively aiming to improve predictive accuracy. A simulation study compares predictive performance of QDA.SEM with QDA under varying conditions and different forms of model misspecification. Results show that QDA.SEM improves predictive accuracy in conditions that benefit most from his trade-off, such as small sample sizes and a high indicator-to-latent-variable ratio, even when model fit on the training data is poor.

Table of Contents

Introduction

Traditionally, the main focus in statistics has been *explanation*, aiming to understand associations between independent and dependent variables (Breiman, 2001). This approach begins with a theory on how variables relate, followed by statistical modeling to estimate the true parameters of these (causal) mechanisms. Statistical inference, such as hypothesis tests and confidence intervals, is then used to evaluate the theory (De Rooij & Weeda, 2020). In this framework, the parameter estimates of the statistical models are the main focus, as they represent the true underlying mechanisms of nature.

Though the explanation approach to data has been scientifically valuable, the emphasis on explaining the causes of behavior through mechanistic models rarely display the ability to accurately predict future behavior (Yarkoni & Westfall, 2017). As a result, many published findings in the social sciences fail to replicate when the same experiments and analyses are conducted independently (Open Science collaboration, 2015). The difficulty in replicating these findings is frequently linked to questionable research practices, such as 'p-hacking' (John et al., 2012; Simmons et al, 2011).

As an alternative, *prediction* focuses on how accurate the model's predictions are, rather than the specific values of its parameters. (Breiman, 2001; Yarkoni & Westfall, 2017). Predictive modeling is considered effective to the extent that the model correctly predicts out-of-sample cases. To assess this effectiveness, datasets are commonly split into a training set for model estimation and a test set for performance evaluation. The statistical model is a mathematical function estimated on the trainig data, that can generate predicted values for the dependent variables of cases in the test data, based on values of their independent variables. In supervised classification settings, the misclassification rate (MCR) is a commonly used method to assess predictive performance by measuring the percentage of incorrectly classified cases (James et al., 2013). However, other metrics, such as sensitivity, specificity, and AUC-ROC, may also be used depending on the context and specific goals of the analysis (James et al., 2013). When data is scarce, k-fold cross-validation can be used to accurately evaluate predictive performance by cyclically splitting up the data into training and test sets (De Rooij & Weeda, 2020).

An important problem in predictive modeling is so-called overfitting, which occurs when a model captures not only the signal pattern in the training data but also (sampling) noise. This

results in a model that performs well on training data but poorly generalizes to new, unseen data (Ghojogh & Crowley, 2019). Overfitting is best explained through the bias-variance trade-off (Hastie et al., 2009). Complex models estimate more parameters. This allows them to closely mirror the patterns in the training data, minimizing bias. However, the fitting is sensitive to small sampling variations, leading to volatile parameter estimations from sample to sample. In contrast, simpler models may not capture all the patterns in training data. This leads to higher bias but provides more stable estimates, reducing variance (James et al., 2013). Besides complexity, the misrepresentation of the true functional form could also introduce bias. For example, when a linear model is used to approximate a non-linear relationship, this can lead to systematic bias. When two models can capture the functional relationships equally well, the simpler model often predicts better (Ghojogh & Crowley, 2019). This issue underscores a key difference between explanation and prediction: while explanatory models focus on minimizing bias, predictive models try to achieve accurate prediction by balancing bias and variance (Shmueli, 2010).

Despite the procedural differences between explanation and prediction, both methods can complement each other. Shmueli and Koppius (2011, p. 554) argue that predictive modeling can contribute to "generating new theory, developing new measures, comparing competing theories, improving existing theories, assessing the relevance of theories, and assessing the predictability of empirical phenomena". A stronger focus on prediction can ultimately enhance our ability to understand causal mechanisms by identifying patterns that traditional explanatory models might overlook. Moreover, in many areas of applied psychology, the primary goal is often to achieve accurate predictions rather than to explain underlying mechanisms (Yarkoni & Westfall, 2017). This perspective suggests that integrating predictive approaches into research in the social sciences not only strengthens empirical findings but also refines theoretical insights over time.

Most statistical prediction methods are data-driven, relying purely on patterns in the training data. This thesis introduces a new classification method, QDA.SEM, which combines the traditional machine learning approach Quadratic Discriminant Analysis (QDA) with Structural Equation Modeling (SEM). The aim is to incorporate prior theoretical knowledge into a data-driven framework to improve predictive accuracy.

Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is a commonly used method for classifying data (Hastie et al., 2009; James et al., 2013). It calculates the posterior probability of class membership using Bayes' theorem. Mean vectors and covariance matrices are first derived from the training data and then inserted into multivariate normal density functions. QDA assigns each observation to the class with the highest posterior probability, taking prior probabilities into account (Wu et al., 1996; Tharwat, 2016; Ghojogh & Crowley, 2019).

Structural Equation Modeling (SEM)

Structural Equation Modeling (SEM) is a statistical technique commonly used in social sciences to analyze relationships among observed and latent variables. Latent variables are unobservable constructs, inferred from multiple observed indicators that contain measurement error. SEM models how latent variables influence these observed variables, explaining the covariation among them. In other words, the covariation in the indicator variables is due to their dependence on one or more latent variables (Beaujean 2014).

SEM allows to model the relationship between indicators and latent variables, while taking measurement error into account (Schreiber et al., 2006). It does so by separating unique variance (measurement error or noise) from shared variance, which is assumed to reflect the true relationship between variables. This approach leads to more reliable and valid representations of constructs than relying on single observed measures, which may be biased or inconsistent. For instance, a latent factor "intelligence" aggregates information from various IQ test items, providing a more accurate measurement than individual test items (Bollen, 1989; Schreiber et al., 2006).

Prediction with SEM

In recent years, there has been growing interest in modeling latent variables to estimate relationships between observed variables and latent constructs for predictive purposes (Hair et al., 2011; Hair et al., 2012a; Hair et al., 2012b; Henseler et al., 2009; Ringle et al., 2012). De Rooij et al. (2020) demonstrated that it is possible to achieve out-of-sample predictions using a reflective SEM by developing a prediction rule that can predict the values of the indicators (y_{ir}) of exogenous latent variables (Y) based on the indicators (x_{ir}) of endogenous latent variables (X). where i refers to the individual cases (i = 1, …, I) and r denotes the specific indicators (r = 1, …, R) for each

latent variable. This method addresses the challenge of making predictions at the item-level using the SEM framework, previously thought of as being not feasible (Shmueli, 2016; Sarstedt et al., 2016).

The prediction rule invented by De Rooij et al, (2020) uses SEM to make assumptions about the data that add restrictions to the covariance matrix. This approach introduces bias but reduces variance. Under specific conditions that benefit most from this bias-variance tradeoff - such as small to medium sample size and minimal specification error – this new method has been shown to outperform multiple state-of-the-art methods, underscoring the potential usefulness of SEM in the field of prediction.

Though the method by De Rooij et al., (2020) has proved useful in predicting exogenous constructs at the item level, it cannot be used for classification. As of yet, no SEM-based prediction rule exists for categorical data. To fill this gap, this thesis proposes a SEM-based prediction that combines SEM with QDA, named QDA.SEM.

QDA.SEM

Traditionally, QDA uses the class priors, sample covariance matrices, and sample means to determine decision boundaries. QDA.SEM replaces the sample covariance matrices with model-implied covariance matrices derived from SEM. By leveraging SEM, this approach aims to produce covariance matrices that might introduce bias but reduce sampling variance.

Unlike standard QDA, which directly estimates covariance matrices from raw training data, QDA.SEM employs model-implied covariance matrices based on an analysis model defined by the researcher. This framework can account for the relationships among variables specified in the SEM, allowing the separation of measurement error from the latent constructs. The main advantage of QDA.SEM lies in its ability to reduce the influence of sampling variance through this model-based approach. Because every SEM is a simplification of reality it is, by definition, wrong. However, as long as the predefined analysis model sufficiently approximates the true data-generating process, the reduction in variance should not come at the cost of too much bias. This balance can lead to improved predictive accuracy.

A way to conceptualize the difference between QDA.SEM and standard QDA is by considering the application of a saturated analysis model. In a saturated model, no constraints are imposed on the relationships among variables, resulting in a model-implied covariance matrix that

is identical to the sample covariance matrix. In this scenario, QDA.SEM reduces to traditional QDA and predictions will be the same. However, when a less saturated model is employed, QDA.SEM incorporates structural assumptions about the covariance matrix by fixing parameters to zero. This introduces bias but reduces variance.

In conclusion, by separating measurement error from the latent variables, SEM offers a more accurate representation of the underlying constructs, improving the reliability of the estimated covariance matrices and, consequently, the predictions in QDA.SEM. This thesis performs a simulation study to compare the two methods under varying conditions of sample sizes, number of predictors in the model, and varying forms of model misspecification.

Methods

QDA

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are commonly used methods for classifying data (Hastie et al., 2009; James et al., 2013). LDA constructs a linear decision boundary based on training data, assuming that each class follows a Gaussian distribution with a common covariance matrix but different means (James et al., 2013).

QDA extends on LDA by dropping the common covariance assumption, allowing each class to have its own covariance matrix. This adds flexibility and makes QDA useful for situations in which the classes differ in their variance-covariance structure, as it can produce quadratic decision boundaries (Wu et al., 1996; Tharwat, 2016; Ghojogh & Crowley, 2019).

QDA classifies observations by modeling the distribution of the predictor variables $x_1 \dots x_\mathrm{p}$ separately for each response class *k* (i.e., for each possible value of *Y*) using multivariate normal distributions. Bayes' theorem is then applied to convert these class-specific distributions into conditional probabilities, $\Pr(Y = k \mid X = x)$, providing the basis for making classification decisions (James et al., 2013)

$$\Pr(Y = k \mid X = x) = \frac{f_k(x)\pi_\mathrm{k}}{\sum_{\mathrm{l}=1}^{K} f_l(x)\pi_\mathrm{l}}. \quad (1)$$

To compute posterior probabilities using Bayes' theorem, QDA relies on two key components: the prior probabilities π_k and the class-specific density functions $f_k(x)$. The prior probability π_k represents the probability that a randomly chosen observation comes from the *k*-th class. The class-specific density function is defined by $f_k(x) \equiv \Pr(X \mid Y = k)$. QDA assumes that the predictor variables $x_1 \dots x_\mathrm{p}$ for each class k follow a multivariate normal distribution with class-specific mean vector $\boldsymbol{\mu_\mathbf{k}}$ and covariance matrix $\boldsymbol{\Sigma_\mathbf{k}}$

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma_\mathbf{k}}|^{1/2}} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu_\mathbf{k}})^T \boldsymbol{\Sigma_\mathbf{k}^{-1}} (x - \boldsymbol{\mu_\mathbf{k}})\right). \ [1]$$

[1] Note that π implies the mathematical constant $\pi \approx 3.14$ in this formula, not to be confused with the prior probabilities π_k in formula (1).

The class specific densities $f_k(x)$ depend only on the class specific mean vectors $\boldsymbol{\mu_k}$, and the class specific covariance matrices $\boldsymbol{\Sigma_k}$. QDA estimates these values from the training data.

To classify an observation into one of K distinct categories, the qualitative response variable Y is assumed to take on K possible, unordered values. For the binary classification case (K = 2), QDA assigns an observation to class $k = 0$ if $\Pr(Y = 1 \mid X = x) < .5$ and to class $k = 1$ otherwise (Wu et al., 1996; Tharwat, 2016; Ghojogh & Crowley, 2019).

SEM

In the social sciences, Structural Equation Modeling (SEM) is a popular approach for examining how latent variables and observed variables relate to each other. It combines aspects of multiple regression, factor analysis and path analysis. In *Figure 1,* an example of an SEM is displayed. The rectangles in the model denote the observed variables, the ellipses are the latent variables. The single-headed arrows indicate linear regression relationships between the variables in the model, where one variable is predicted by another. The double-headed arrows represent covariances between variables that are exogenous, i.e. do not have single-headed arrows pointed at them. For endogenous variables, that do have single-headed arrows pointed at them, the double-headed arrows indicate residual covariances, i.e. the covariances that remain unexplained by the directed relationships. The double-headed arrows that connect a variable to itself indicate the variance of that variable when it is exogenous, and the residual variance when it is endogenous (Karch, 2025).

The framework of SEM can be represented through matrix algebra in the Reticular Action Model (RAM) notation. The RAM notation computes model-implied covariance matrices through the following computation

$$\boldsymbol{\Sigma_{\text{model}}} = \mathbf{F}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}(\mathbf{I} - \mathbf{A})^{-\mathrm{T}}\mathbf{F}^{\mathrm{T}},$$

where **I** is an identity matrix. The **A** matrix represents the directed relationships between all variables in the model, both observed and latent. The parameters in **A** are structured such that element $\mathbf{A}_{\mathrm{ij}}$ represents the effect of variable j on variable i, the single-headed arrows in *Figure 1*. **S** is a matrix containing the (residual) covariances between all variables in the model, latent and observed. As well as the (residual) variances of the variables, on the diagonal. Note that in *Figure*

1, no covariances are estimated, so the off-diagonal elements in this example are constrained to zero in **S**. Lastly, the **F** matrix selects observed variables from the full variable set, it makes sure that only observed - not latent - variables are in the final model-implied covariance matrices. The rows in **F** correspond to the observed variables, the columns to all variables in the model, both observed and latent. The matrix consists of ones to indicate the positions of observed variables in the model, all other values are zero. If the **A** and **S** matrices are arranged such that the observed variables come first, followed by the latent variables, then the **F** matrix looks like an identity of matrix of size $p\ x\ p$ where p are the observed variables, followed by a $p\ x\ l$ matrix containing zeroes, where l is the number of latent variables (J. D. Karch, personal communication, February 27, 2025).

Maximum Likelihood estimation is applied to iteratively estimate the values of **A** and **S** that minimize the statistical distance between the model-implied covariance matrix $\mathbf{\Sigma_{model}}$ and the observed covariance matrix $\mathbf{\Sigma}$ (Bollen, 1989). In a multigroup setting, this process is repeated separately for all classes, resulting in a $\mathbf{\Sigma_{model}}$ for each class. **A** and **S** provide a clear separation of structural variance and measurement variance, allowing for a more accurate representation of the structural information contained in the noisy indicators. Moreover, SEM allows to constrain parameter values to zero, effectively ensuring that spurious results caused by sampling variance are not modeled into the $\mathbf{\Sigma_{model}}$ (J. D. Karch, personal communication, February 27, 2025).

Summarizing, SEM derives class-specific model-implied covariance matrices $\mathbf{\Sigma_{k\ model}}$ based on the constraints of the model to minimize the statistical distance to the class-specific sample covariance matrices $\mathbf{\Sigma_k}$. Therefore, the quality of an SEM depends heavily on the analysis model used to describe the theoretical relations between latent and observed variables. The challenge in this step is the theoretically unlimited number of possible model configurations that could be considered. Since the true data-generating process is generally unknown, one needs to rely on prior knowledge, theoretical frameworks, and insights from the literature to define a model (Schreiber et al., 2006). Fit indices like the Root Mean Square Error of Approximation (RMSEA) measure how well the specified analysis model aligns with the observed data (Hu & Bentler, 1999).

QDA.SEM

The QDA.SEM function extends traditional QDA by replacing class-specific covariance matrices in the probability density functions $f_k(x)$ with model-implied covariance matrices derived via SEM. These model-implied covariance matrices are obtained during the SEM model fitting. In this study, the model-implied covariance matrices are estimated using the *lavaan* package in *R* (Rosseel, 2012).

When a saturated model is used as an analysis model, the SEM has 0 *df*. In this specific case, the model-implied covariance matrix is identical to the sample covariance matrix. Consequently, QDA.SEM reduces to QDA. This illustrates the key principle of QDA.SEM. From the formula to estimate $\mathbf{\Sigma_{model}}$ follows that a simpler analysis model incorporates more structural assumptions about the data and estimates less parameters in matrices **A** and **S**, in other words, more parameters are constrained to zero. This introduces bias but reduces variance.

This bias-variance trade-off underscores the value of QDA.SEM. Traditional QDA does not impose any assumptions about the structure of the covariance matrix. In contrast QDA.SEM incorporates prior knowledge, by grouping variables into latent constructs and specifying relationships among observed variables. To the extent that the specified SEM sufficiently approximates the population, these constraints should enhance the quality of the covariance matrix estimates and, consequently, improve predictive performance.

Simulation design

This thesis employs a Monte Carlo simulation study to compare the predictive performance of the newly proposed QDA.SEM method against the standard QDA method. The applicability of SEM, and thus QDA.SEM, heavily depends on the quality of the predefined analysis model. The true data-generating process is in practice unknown, and a model is by definition a simplification of reality. Thus, the analysis model used is subject to misspecification. This simulation examines three scenarios:

1. The true data-generating model matches the analysis model.
2. The analysis model contains local misspecification, as compared to the true data-generating model. The term local misspecification is used to refer to unaccounted correlations between error terms of indicators.

3. The analysis model contains global misspecification, as compared to the true data-generating model. By global misspecification, an incorrect number or configuration of latent variables is meant.

These scenarios are chosen to assess the robustness of QDA.SEM to different forms of model misspecification. Performance is evaluated using the misclassification rate (MCR). Specifically, across 1000 repetitions, a 'win' is recorded each time the QDA.SEM function achieves a lower MCR than standard QDA. The percentage of wins indicates how often the QDA.SEM function predicts more accurately than QDA. Simulation runs where the MCR of both methods is exactly the same are also reported as 'ties'.

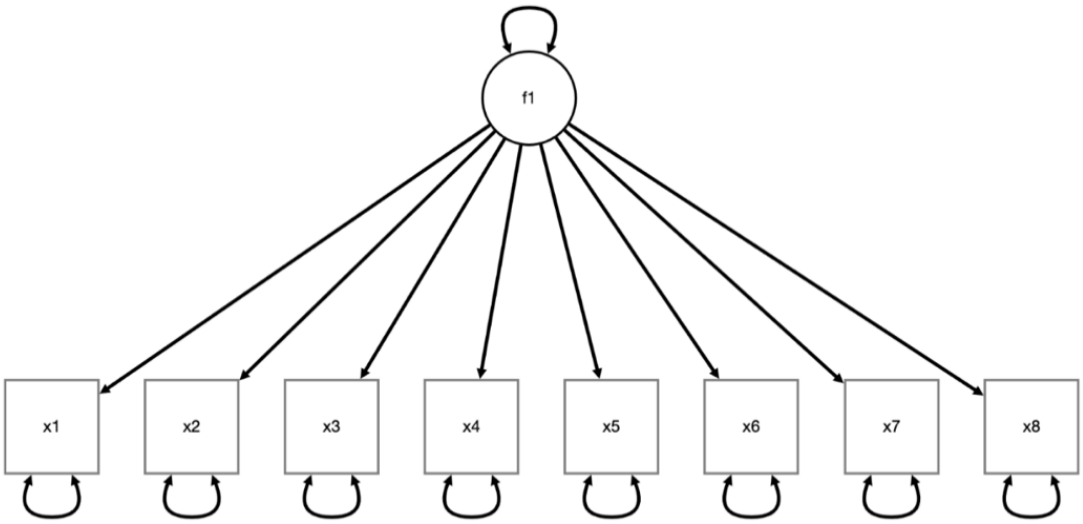
Class membership is equally distributed (50/50) in both the training and test sets, with the test set size fixed at n = 1000. This study is restricted to scenarios where all manifest predictor variables are continuous and follow a multivariate normal distribution. All simulations systematically vary the following conditions:

- The training set sample sizes: 200, 500, 1000, 10000
- The number of indicators: 8, 12, 16

In all conditions, the analysis model follows a simple one-factor structure, where a single latent variable loads onto all manifest variables. No correlated error variances are assumed among the manifest variables. An example of the one-factor model used with eight predictors is shown in *Figure 1*.

Figure 1

Example of the 1 Factor Model Used as Analysis Model with 8 Predictors



1. **True model equals analysis model**

   This scenario represents the most favorable conditions for QDA.SEM. The true data-generating process follows a one-factor model that is identical to the analysis model. Essentially, this method reduces bias but does not introduce variance. This makes it the most advantageous scenario and serves as a sanity check, as such a perfect match between the true model and the analysis model is unlikely in real-world applications. Following the reasoning of the bias-variance trade-off, the QDA.SEM approach is expected to have less variance, by the restrictions added, but no bias, because the assumptions correspond to the true data-generating process. Thus, the expectation is that QDA.SEM clearly outperforms QDA in this condition.

2. **Local misspecification**

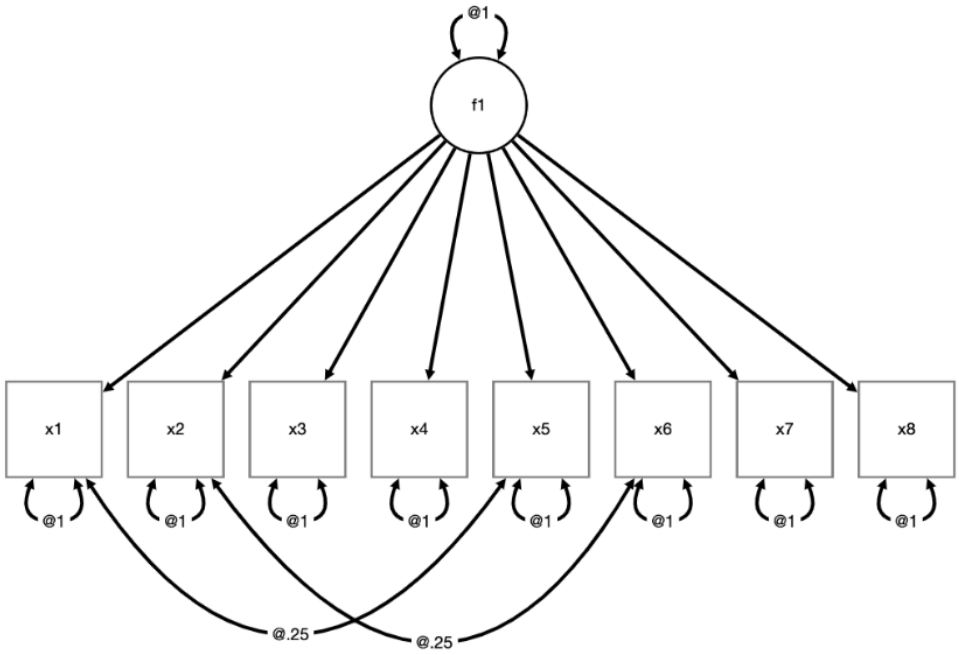
   The true data-generating model is a one-factor model with local misspecification, where error variances of 50% of the manifest variables are correlated. For example, in the case of eight indicators, four manifest variables have correlated error variances, forming two pairs. The strength of the correlated errors is set at .25, ensuring a range of fit quality from good to adequate based on the guidelines of Hu and Bentler (1999). *Figure 2* illustrates the true data-generating model under local misspecification.

3. **Global misspecification**

The true data-generating model is a two-factor model, where 50% of the manifest variables load onto one latent factor, and the other 50% load onto a second latent factor. The correlation between the two latent factors varies across four levels: 0, .3, .5, and .7. See *Figure 3.*

Figure 2

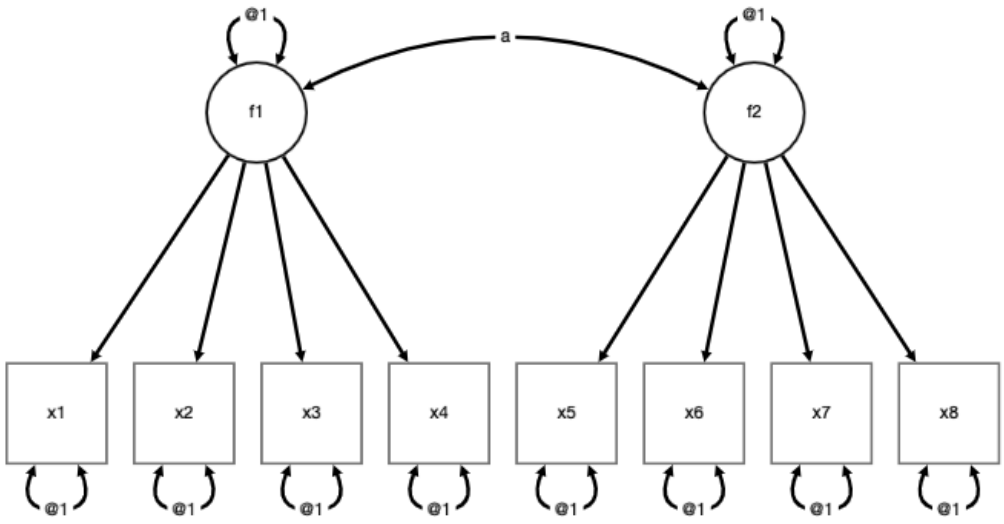
Illustration of the True Data Generating Model in the Local Misspecification Condition with 8 Predictors



Notes. 50 percent of the predictors have correlating error variances of .25

Figure 3

Example of the True Data Generating Model in the Global Misspecification Condition with 8 Predictors



Notes. The correlation between the latent variables varies with parameter a taking on values 0, .3, .5, and .7

In all simulations, the *simulateData()* function of *Lavaan* is used in *R* (Rosseel, 2012). The variances of all latent and manifest variables are set equal to 1. Class differences are simulated by assigning class-specific factor loadings to the manifest variables. For each class, factor loadings are defined as four pairs of values that were randomly drawn between 0.2 and 0.8. The first value of each pair is assigned to class k = 0, and the second value to class k = 1. This pattern is applied cyclically across all manifest variables, ensuring that every four consecutive indicators share the same loadings. For example, the 5th (and, if applicable, the 9th and 13th) indicator has the same loading as the 1st. The 6th indicator has the same loading as the 2nd, and so on, see *Table 1*. The reasoning behind the cyclically assigning of the loadings is that the differences in predictive accuracy across the simulations with 8, 12, and 16 predictors only depends on the number of predictors and not on a difference in loadings.

Table 1

Randomly Drawn Values for Factor Loadings to Simulate Class Differences

Manifest variable	1, 5, 9, 13	2, 6, 10, 14	3, 7, 11, 15	4, 8, 12, 16
Class k = 0	0.8	0.4	0.4	0.3
Class k = 1	0.8	0.7	0.3	0.7

In the single-latent-variable simulations (scenario 1 and 2), the mean of the latent variable is set to 0.5 for class k = 0 and to 1.5 for class k = 1. In the two-latent-variable simulations (scenario 3), the first latent variable (LV1) followed the same class-specific mean structure, with means set to 0.5 and 1.5 for classes 0 and 1, respectively. The second latent variable (LV2) has the same mean value of 0 for both classes.

In each simulated run, the RMSEA value is recorded to assess the average model fit of the analysis model on the training data. RMSEA measures how well the model approximates the observed data, with lower values indicating better fit. The RMSEA fit measure is decided upon because it is not influenced by sample size, enabling a reliable comparison across the different simulation conditions. According to the guidelines proposed by Hu and Bentler (1999), an RMSEA value of less than 0.06 indicates a good fit, values between 0.06 and 0.08 suggest an adequate fit, and values greater than 0.1 are considered indicative of a poor fit. The primary focus of this study is on scenarios where the analysis model demonstrates at least an adequate fit to the training data. This approach is based on the practical consideration that a model with a poor fit should not be used as a predictive tool.

Results

Simulation Study 1.

Analysis Model = True Data Generating Model

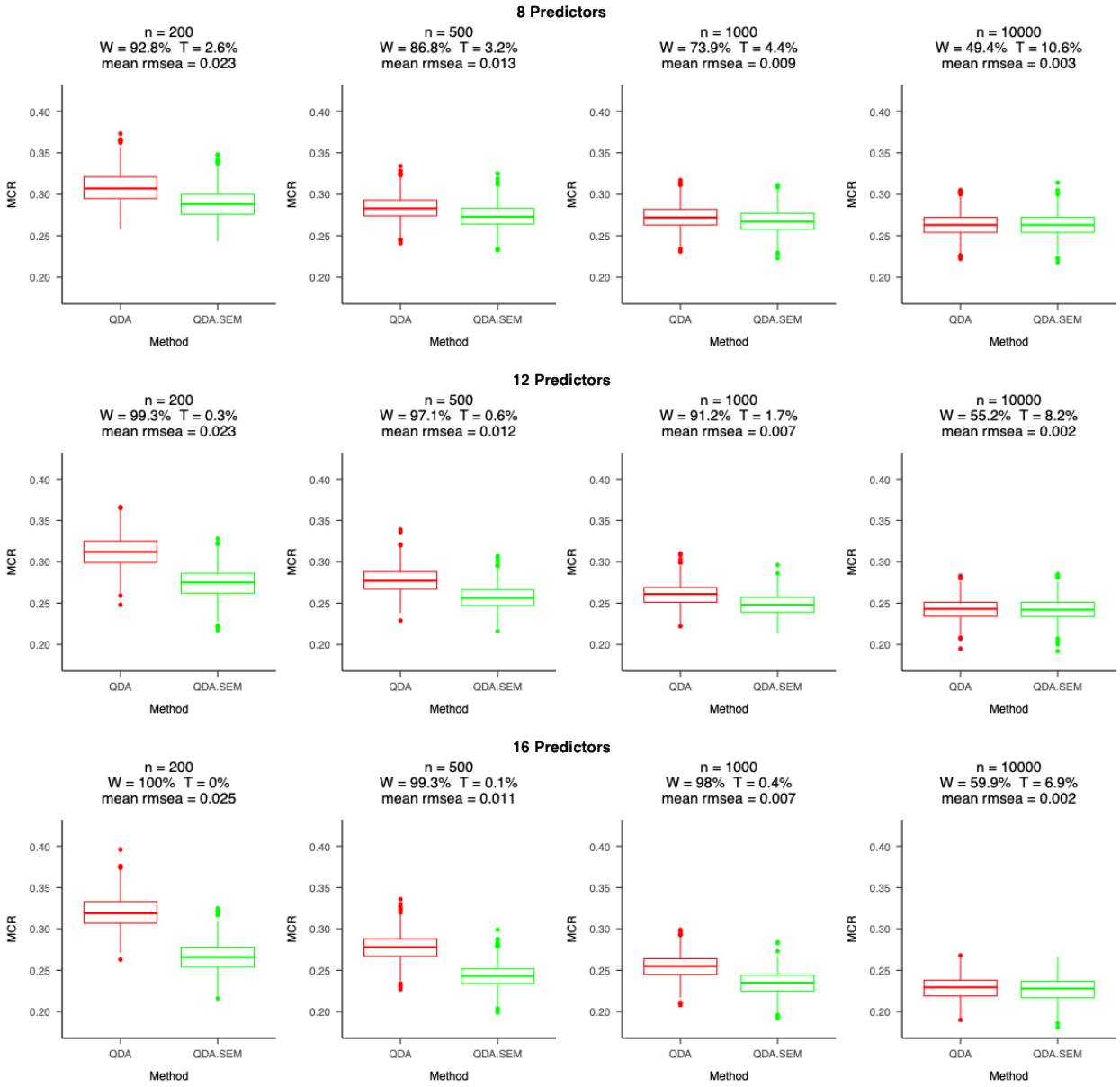
The results of the first simulation study are displayed in *Figure 4.* With the lowest mean Root Mean Square Error of Approximation (RMSEA) values of 0.002 to a maximum mean RMSEA 0.025, the RMSEA scores of the models are particularly low in these simulations, proving a near perfect fit. As expected through the reduction in variance without the introduction of bias, the QDA.SEM function consistently achieves a lower misclassification rate (MCR) than traditional Quadratic Discriminant Analysis (QDA) and thus generates more wins when the analysis model matches the data-generating model. Specifically, QDA.SEM outperforms QDA by recording 91% - 100% wins when $n \leq 1000$ and the number of predictors is 12 or 16. With 8 predictors and $n \leq 1000$, the improvement over QDA is still present but slightly smaller, 74% - 93% of wins. Effectively demonstrating its advantage when the model is correctly specified.

Thus, the advantage of QDA.SEM is most pronounced in smaller sample sizes and increases with more predictors. This pattern provides evidence that Structural Equation Modeling (SEM) is more effective at separating shared variance from measurement variance when there are more indicators measuring the same latent construct (Marsh et al., 1989; Gagne & Hancock, 2016). The increased number of predictors increases the precision of the estimated covariance structure, allowing QDA.SEM to produce a more stable representation of the underlying relationships.

As sample size increases to $n = 10000$, the methods converge in their prediction accuracy, resulting in 50% - 60% wins for QDA.SEM and 7% - 10% ties. This pattern can be explained through the bias-variance trade-off: QDA.SEM reduces sampling variance, aiming to make the model-implied covariance matrix a more accurate and stable representation of the population covariance matrix. By the law of large numbers, when sample sizes are very big (i.e. $n \leq 10000$), the raw sample covariance matrix already approaches the population covariance matrix, and the effect of sampling variance reduces, hence the advantage of QDA.SEM. Essentially, in large sample sizes the effect of sampling variance is already negligible, so a further reduction in variance does not lead to better predictions.

Figure 4

Results Simulation Study 1



Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is the same as the true data generating model (simulation study 1).

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

Simulation Study 2.

Local Misspecification

The second simulation study assesses the robustness of QDA.SEM to misspecification by introducing local misspecification in the data-generating model. The results displayed in *Figure* 5 show similar patterns as those observed in the first simulation study: QDA.SEM outperforms traditional QDA in smaller sample sizes and when more predictors measure the same latent construct. Though in the largest sample sizes, the traditional QDA predicts more accurately than QDA.SEM.

The average RMSEA scores indicate that the 8-predictor condition falls slightly outside the threshold for good fit of ≤ 0.06 (Hu & Bentler, 1999) with a mean RMSEA ≈ 0.065. Whereas the 12-predictor (mean RMSEA ≈ 0.055) and 16-predictor (minimum mean RMSEA = 0.047, maximum = 0.054) conditions do meet the good fit criterion. Despite this misspecification, QDA.SEM maintains strong positive performance, demonstrating its robustness to minor model misspecifications.

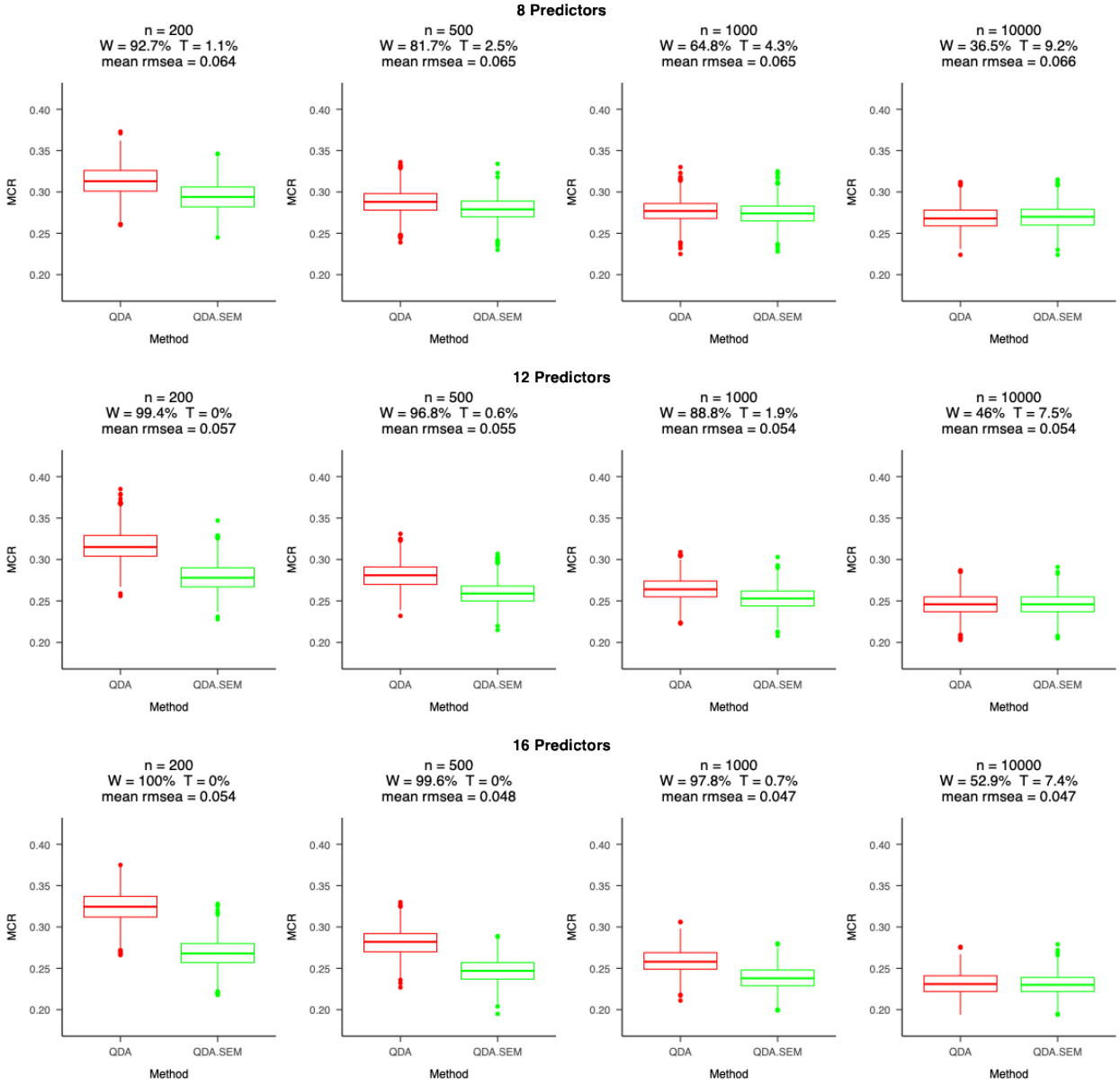
QDA.SEM records more wins than QDA across all conditions when $n \leq 1000$. In the 16-predictor condition, QDA.SEM outperforms QDA in over 97% of the simulations. Even in the 8-predictor condition, QDA.SEM remains superior, achieving a 92.7% win rate when $n = 200$, when sample size increases, the number of wins decreases to 81.7 % at $n = 500$ and 64.8% at $n = 1000$.

As sample size becomes very large, $n = 10000$, the bias introduced by the model misspecification starts to outweigh the reduction in variance. Especially with 8 predictors, the number of wins drops to 36.5%, with 9.2% ties. In the 12- and 16-predictor conditions, the performances of both QDA and QDA.SEM are very close.

In summary, QDA.SEM predicts more accurately than QDA despite the analysis model containing local misspecification, provided the number of predictors is high, and sample size is small to moderate.

Figure 5

Results Simulation Study 2.



Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is wrongly specified through local misspecification (simulation study 2).

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

Simulation Study 3.

Global Misspecification

Simulation Study 3 examines whether QDA.SEM predicts more accurately than QDA under global misspecification of the analysis model. *Figure 6* contains the results under the condition where the two latent variables in the true data generating process have a correlation of .3. [2]

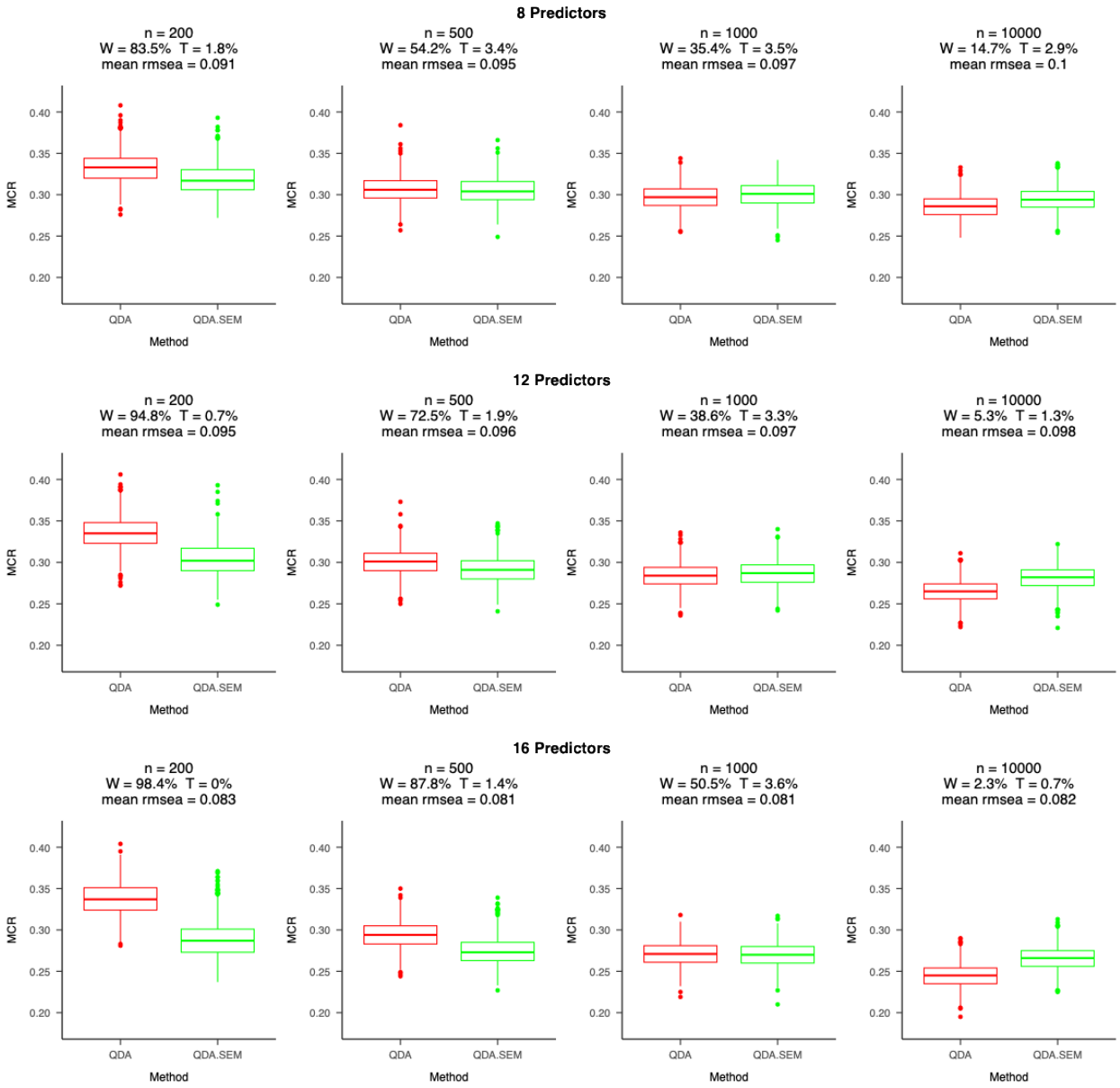
The mean RMSEA values indicate an adequate (RMSEA $\leq$ 0.08) to poor (RMSEA > 0.1) fit for all conditions (Hu & Bentler, 1999). With the lowest mean RMSEA being 0.081 and the highest being 0.1. In spite of the poor fit, QDA.SEM wins more simulations than QDA when sample size is $n \leq 500$, even winning 98.4% of the simulation runs with 16 predictors and $n = 200$. At $n = 1000$ QDA.SEM records slightly more wins than QDA in the 16-predictor condition, whereas in the 8- and 12-predictor condition, QDA more often predicts better. This pattern underscores the interaction between sample size and the number of predictors, as was found in simulation study 1 and 2. This pattern is caused by the bias-variance trade-off.

In the simulations with $n = 10000$, QDA.SEM wins less than 15% of the simulations, providing evidence that fitting a wrongly specified SEM to a large sample size does not lead to improved predictions by QDA.SEM as compared to QDA. Essentially, the reduction of variance becomes minimal, not outweighing the introduction of bias. It is noteworthy that in the $n = 10000$ condition, the pattern of number of predictors is reversed. Under global misspecification, and with a big sample size, having more predictors *reduces* the predictive accuracy of QDA.SEM as compared to QDA.

[2] Appendix 1.1 summarizes the simulations for the latent variables with a correlation parameter of .5 and .7. In those conditions, the simulations show a highly similar pattern compared to the simulations where the true data generating model have a correlation parameter of .3, yet QDA.SEM outperforms the QDA function in a greater number of simulated runs, as expected due to the increased correlation size, and hence better model fit.

Figure 6

Results Simulation Study 3



Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is wrongly specified through global misspecification (simulation study 3). The true data generating model consists of 2 latent constructs that have a correlation of .3.

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

The simulations with the most severe form of misspecification – where the true data-generating model consists of two uncorrelated latent variables – is illustrated in *Figure 7*. In this scenario, the mean RMSEA indicate an adequate model fit (RMSEA ≤ 0.08) for the 16-predictor condition, with mean RMSEA ≈ 0.095, and a poor fit (RMSEA > 0.1) for the 8- and 12-predictor conditions (Hu & Bentler, 1999), ranging from a minimum mean RMSEA of 0.103 to a maximum of 0.116.

Despite the significant model misspecification, QDA.SEM maintains a lower MCR than QDA in all conditions where $n = 200$: 92.5% wins with 16 predictors, 84.7% with 12 predictors and 70.2% with 8 predictors. With 8-predictors and $n = 500$, the performance of QDA.SEM and QDA is nearly identical with 50% wins and 3.4% ties. In the 12-predictor condition, QDA.SEM records 49.4% wins and 3.1% ties when $n = 1000$. In the 16-predictor condition; when $n = 200$ QDA.SEM wins in 92.5% of the simulations, though its advantage decreases to 74.6% wins when sample size increases to $n = 500$, and to 62% with $n = 1000$. Generally, as sample size increases, QDA gradually outperforms QDA.SEM.

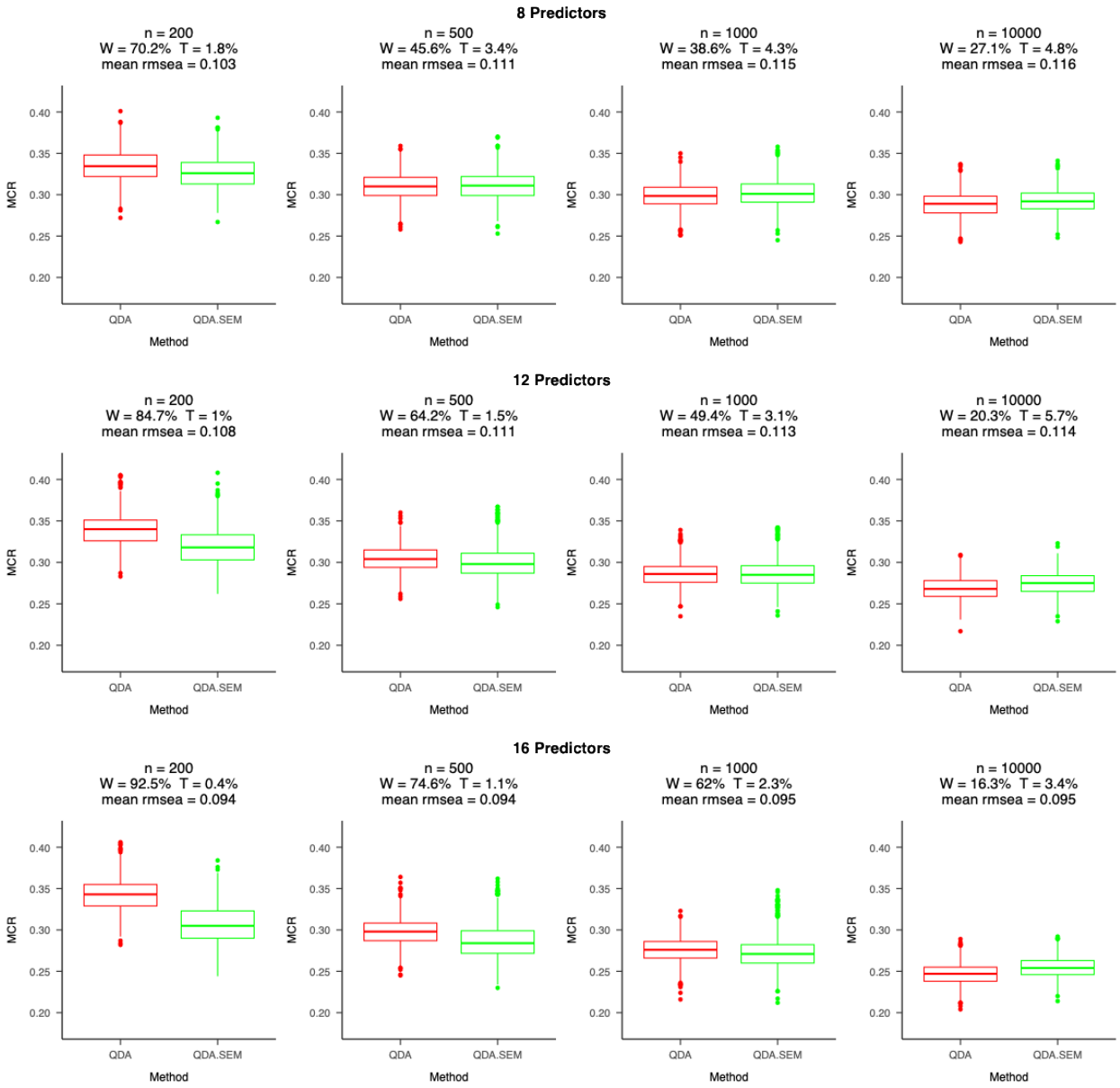
As shown in the simulations where the latent variables had a .3 correlation, the relationship between the number of predictors and classification performance has reversed in very large samples. At $n = 10000$, QDA.SEM achieves a higher win rate in the 8-predictor condition (27.1%) than in the 12- predictor condition (20.3%) 16-predictor condition (16.3%).

Surprisingly the difference between QDA and QDA.SEM is smaller with $n = 10000$ in the .3 correlation condition than in the uncorrelated condition, despite the model fit being better. QDA.SEM only wins 2.3% of the simulations with 16-predictors versus 16.3% in the uncorrelated condition, a similar pattern is observed in the 8- and 12- predictor conditions. A possible reason is that two uncorrelated latent variables function as two independent one-factor models. In this case, a one-factor analysis model might capture key patterns that go unnoticed when the latent variables have a weak correlation.

In summary, even under significant global misspecification, QDA.SEM consistently achieves a lower MCR than QDA when sample size is small to moderate, with its advantage increasing as the number of predictors rises. However, in very large samples, QDA outperforms QDA.SEM, and the relationship between predictor count and classification performance is reversed.

Figure 7

Results Simulation Study 3.



Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is wrongly specified through global misspecification (simulation study 3). The true data generating model consists of 2 latent constructs that are uncorrelated with each other.

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

Discussion

In line with the findings of De Rooij et al. (2020), who demonstrated that Structural Equation Modeling (SEM) can be used to predict the values of indicators for dependent latent variables, this thesis shows that SEM can be leveraged to enhance predictive performance in classification settings. In the current study, an extension of Quadratic Discriminant Analysis (QDA), QDA.SEM, was investigated, where the sample covariance matrix is not based on the raw data but is substituted with a model-implied covariance matrix estimated using SEM.

The implementation of QDA.SEM requires specifying an analysis model that reflects the underlying data structure, based on literature and prior knowledge. Three simulation studies were conducted, using different data-generating models, varying sample sizes, and different indicator-to-latent-variable ratios, to investigate whether and under what conditions QDA.SEM leads to more accurate predictions than QDA.

The first simulation study investigated predictive performance in cases where the analysis model was correctly specified, i.e. matched the true data-generating process. This was the most favorable simulation study and essentially served as a sanity check, as the restrictions of the SEM will ensure less variance, but no bias. It is unrealistic, however, to have such an ideal scenario in real-world applications. QDA.SEM outperformed QDA in all tested conditions. The advantage of QDA.SEM is most pronounced for smaller sample sizes and larger numbers of indicators reflecting the latent variable. The performances converge when sample sizes become very large (n = 10000), because the effect of sampling variance diminishes (Yarkoni & Westfall, 2017).

In the second and third simulation studies the robustness of QDA.SEM to local or global misspecification was tested. Even under the most severe model misspecification conditions tested (Root Mean Square Error of Approximation scores > 0.09), QDA.SEM predicted more accurately than QDA in small to moderate sample sizes (i.e. n = 200 and n = 500) [3], with the advantage increasing when the ratio of indicators to latent variables is larger. Both patterns can be explained by QDA.SEM's ability to reduce sampling variance. When more indicators measure the same latent constructs, SEM is better at differentiating measurement variance from structural variance

[3] With the true data generating model being a 2-factor model with uncorrelated latent variables, 8 predictors and n = 500, QDA.SEM recorded 45.6% wins and 3.4% ties, performing slightly worse than QDA.

(Marsh et al., 1989; Gagne & Hancock, 2006), furthermore, smaller samples benefit more from a reduction in variance. In larger sample sizes, however, a wrongly specified model results worse predictions by QDA.SEM compared to QDA, and a higher ratio of indicators to latent variables further decreases the predictive performance.

Interestingly, when the true data generating process consists of two latent variables that are completely uncorrelated, under $n = 10000$, the predictive performance of QDA.SEM and QDA is more similar than when the latent variables have a correlation of .3. This is surprising because a worse-fitting analysis model would typically be expected to reduce the performance of QDA.SEM. One possible explanation is that two uncorrelated latent variables are essentially two separate 1-factor models. In that case, an analysis model consisting of a 1-factor model might pick up on some important patterns that are overlooked when the latent variables are weakly correlated. Since this pattern only emerges in a scenario where QDA.SEM performed worse than QDA, it is not highly relevant for the application of QDA.SEM in real-world settings. I leave it to future research to further investigate this.

An essential aspect of QDA.SEM is the alignment between the analysis model and the true data-generating process. However, only a limited number of scenarios have been explored in this study. The simulations tested three different data-generating conditions, meaning the results do not account for all possible mismatches. Future research could investigate a broader range of data-generating models, as well as variations in analysis models. The current study examined either local or global misspecification but not, for example, a combination of both.

It is common practice in SEM modeling to use an iterative approach. One starts with an initial model based on literature and prior knowledge, then uses fit indices to assess how well the model represents the data. If the model does not fit well, modification indices are used to suggest potential adjustments, such as adding or removing constraints (Beaujean, 2014). This cycle of testing and refining helps researchers develop a model that both fits the data and aligns with theoretical expectations. The current study did not discuss model selection, nor did it examine modification indices prior to making predictions. QDA.SEM could be combined with K-fold cross-validation to arrive at an optimal analysis model, though caution is advised not to rely too much on a data-driven approach and to maintain a strong theoretical foundation.

This study relies on simulated data. While real-world data may contain different, more complex, characteristics not fully accounted for in the data simulation (Burton et al., 2006; Morris

et al., 2019), the findings provide valuable insights that can be useful in real-world applications. The simulated data used in this study were continuous and normally distributed, aligning with the normality assumption in SEM (Bollen, 1989). Although, real-world data is rarely perfectly normally distributed, many datasets approximate normality sufficiently for the results to remain informative. QDA.SEM's robustness to violations of the normality assumption have not been tested. Furthermore the study fixed the class-specific factor loadings. In practice, these parameter values can vary, which might influence predictive performance in ways not tested for by the simulations.

Acknowledging these limitations, this thesis demonstrates that under a specific set of conditions, SEM can be used to improve predictive performance in a classification setting, even under significant model misspecifications. QDA.SEM can be a valuable addition to the statistician's toolbox in social sciences, particularly in scenario's where sample sizes are small to moderate, and when the indicator to latent variable ratio is high. In social sciences, sample sizes of $n \leq 200$ are frequently encountered, as is the use of SEM, emphasizing the potential usefulness of QDA.SEM.

References

Beaujean, A. A. (2014). *Latent variable modeling using R*: A step-by-step guide. Routledge.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*(24), 4279–4292.

De Rooij, M., Karch, J. D., Fokkema, M., Bakk, Z., Pratiwi, B. C., & Kelderman, H. (2023). SEM-based out-of-sample predictions. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(1), 132-148.

De Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248-263.

Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. *Journal of Business Research*, *69*(10), 4565-4582.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate behavioral research, 41(1), 65-83.

Ghojogh, B., & Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. *arXiv preprint arXiv:1906.02590*

Hair Jr, J. F., Matthews, L. M., Matthews, R. L., & Sarstedt, M. (2017). PLS-SEM or CB-SEM: updated guidelines on which method to use. *International Journal of Multivariate Data Analysis*, *1*(2), 107-123.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, *19*(2), 139-152.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012a). An assessment of the use of partial least squares structural equation modeling in marketing research. Journal of the academy of marketing science, 40, 414-433.

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012b). The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long range planning*, *45*(5-6), 320-340.

Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (pp. 277-319). Emerald Group Publishing Limited.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incen- tives for truth telling. *Psychological Science, 23*, 524–532. doi:10.1177/0956797611430953

Karch, J. D. (2025). lavaangui: A Web-Based Graphical Interface for Specifying Lavaan Models by Drawing Path Diagrams. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-12.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivariate behavioral research, 33(2), 181-220.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102.

Open Science Collaboration. (2015). Estimating the repro- ducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-SEM in" MIS Quarterly". *MIS quarterly*, iii-xiv.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software*, *48*, 1-36.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detecting misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 561-582.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies!. *Journal of business research*, *69*(10), 3998-4010.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323–338.

Shmueli, G. (2010). To explain or to predict?.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS quarterly*, 553-572.

Shmueli, G., Ray, S., Estrada, J. M. V., & hatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research, 69*, 4552–4564

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analy- sis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145-180

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., & Erni, F. (1996). Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta*, *329*(3), 257-265.

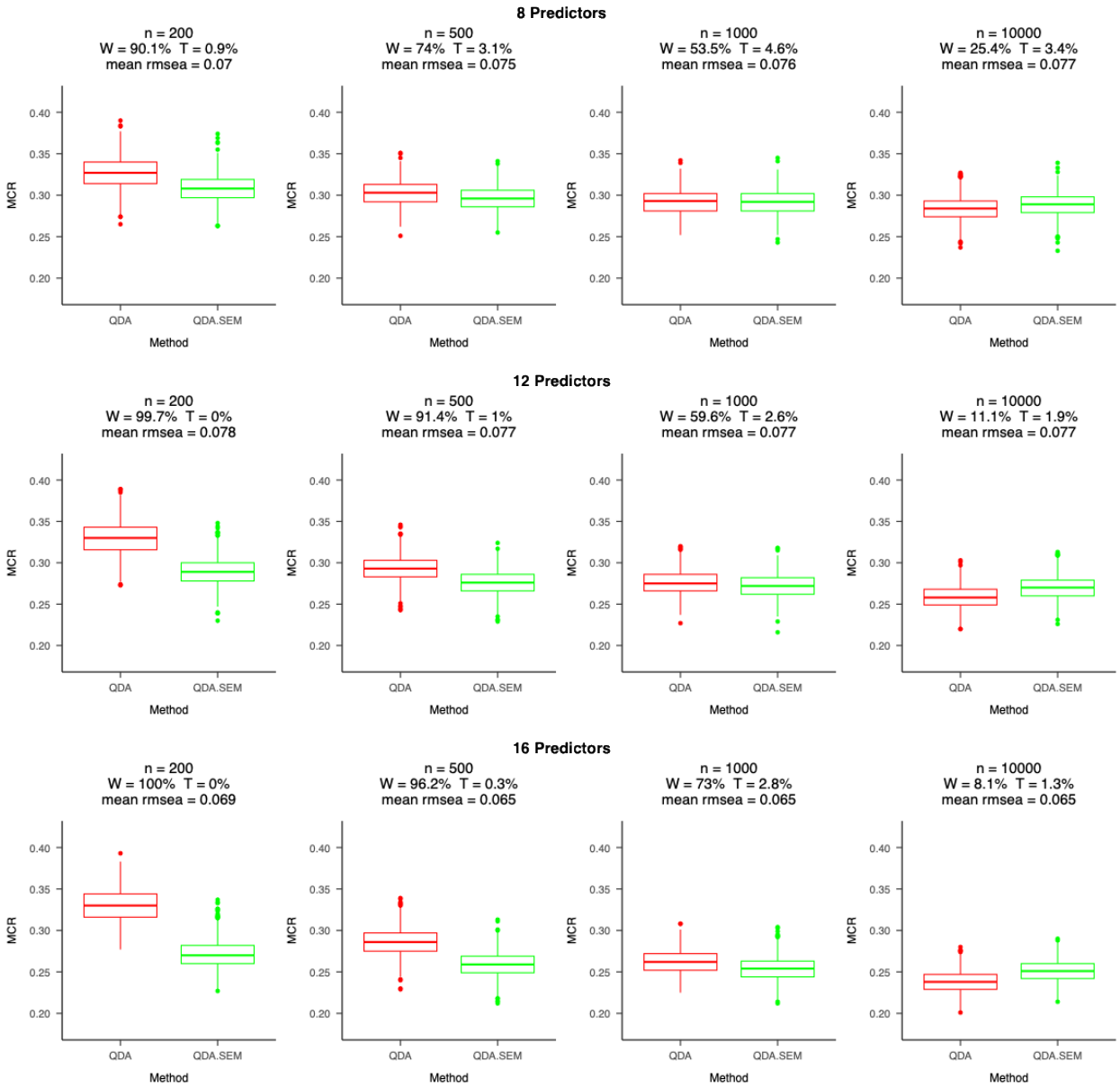
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Appendices

Appendix 1: Figures Extra Results Simulation Study 3

Figure 8

Results Simulation Study 3.



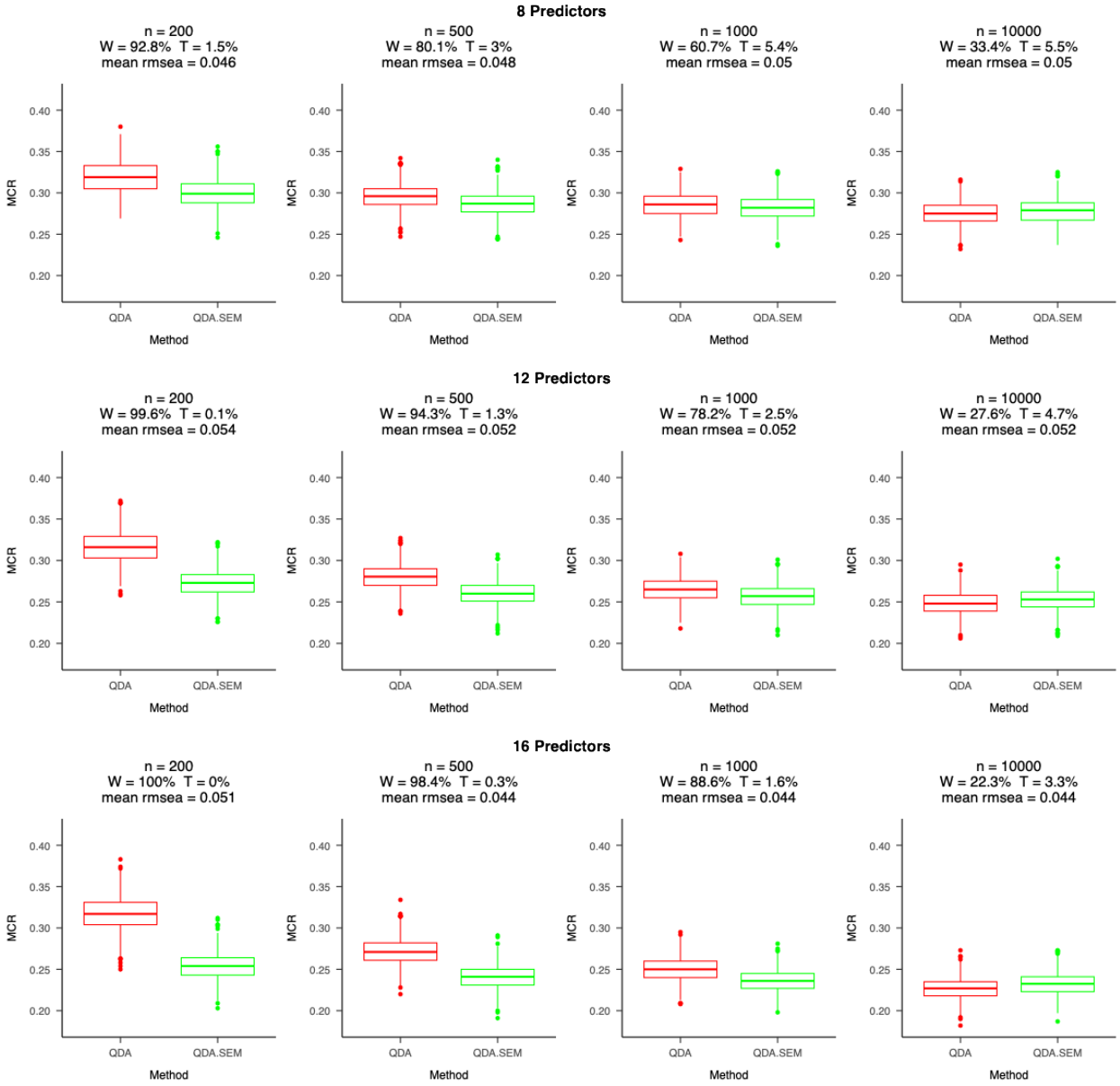
Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is wrongly specified through global misspecification (simulation study 3). The true data generating model consists of 2 latent constructs that have a correlation of .5.

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

Figure 9

Results Simulation Study 3.



Notes. Misclassification Rate (MCR) of Quadratic Discriminant Analysis (QDA) versus QDA.SEM when the analysis model is wrongly specified through global misspecification (simulation study 3). The true data generating model consists of 2 latent constructs that have a correlation of .7.

W indicates the percentage of simulations QDA.SEM achieved a lower MCR than QDA.

T indicates how often the MCR was exactly the same.

Appendix 2: *R*-Code

See attached *R*-files.