

STATISTICAL METHODOLOGY FOR BIOLOGICAL SIGNALS IN THE
PRESENCE OF MEASUREMENT UNCERTAINTY

by

Christopher Robert Barbour

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2018

©COPYRIGHT

by

Christopher Robert Barbour

2018

All Rights Reserved

DEDICATION

I dedicate this dissertation to my family, who have supported me unconditionally throughout my graduate career.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Mark Greenwood and Dr. Bibiana Bielekova, for their support and encouragement throughout my graduate student career. Their guidance has helped me grow as a professional and I can not over-emphasize how grateful I am to have them as mentors, advisors, and friends. I would also like to thank Dr. Mark Owkes for maintaining the LaTeX template used in this dissertation.

This research was supported in part by the Intramural Research Program at the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH) and by the Department of Mathematical Sciences of Montana State University.

TABLE OF CONTENTS

1. INTRODUCTION	1
2. CONSTRUCTED COMPOSITE RESPONSE: A FRAMEWORK FOR CONSTRUCTING TARGETED LATENT VARIABLES.	7
Contribution of Authors and Co-Authors	7
Manuscript Information Page	7
Abstract	9
2.1 Introduction	9
2.2 Proposed Methodology	14
2.2.1 Constructed Composite Response	14
2.2.2 Examination of Exact Solutions	17
2.2.3 Bounded Constructed Composite Response	18
2.2.4 P-value calculation for CCR.	19
2.3 Simulation Study	20
2.3.1 Convergence of CCR algorithm	23
2.3.2 Projection Comparison	23
2.3.3 Type I error estimation	25
2.4 Real Data Example	25
2.5 Conclusions	28
3. EXTENDING CCR METHODOLOGY TO HIGH-DIMENSIONAL DATA TO DEVELOP A SENSITIVE CLINICAL ENDPOINT FOR MULTIPLE SCLEROSIS	57
Contribution of Authors and Co-Authors	57
Manuscript Information Page	57
3.1 Introduction	59
3.2 Penalized Constructed Composite Response	66
3.3 Simulation Study	66
3.4 Real Data Example	70
3.5 Conclusions	71
4. MOLECULAR-BASED DIAGNOSIS OF MULTIPLE SCLEROSIS AND ITS PROGRESSIVE STAGE	82
Contribution of Authors and Co-Authors	82
Manuscript Information Page	83
Abstract	85
4.1 Introduction	85

TABLE OF CONTENTS – CONTINUED

4.2	Methods.....	87
4.2.1	Subjects.....	87
4.2.2	CSF collection and processing	88
4.2.3	SOMAscan.....	88
4.2.4	Assessment of cellular origin of tested biomarkers from a subset of human immune cells and CNS cells	89
4.2.5	Measuring signal-to-noise ratio (SNR) in biomarkers.....	90
4.2.6	Area under the Receiver Operating Characteristics Curve (AUROC).....	91
4.2.7	Statistical learning to develop diagnostic classifiers using a Random Forest methodology.....	92
4.3	Results.....	92
4.3.1	SOMAscan Assay on CSF samples.....	92
4.3.2	Development of diagnostic tests for MS.....	93
4.3.3	Deconvolution of biomarkers' cell of origin.....	96
4.3.4	Biological interpretation of MS versus non-MS di- agnostic test.....	96
4.3.5	Biological interpretation of progressive MS versus RRMS diagnostic test	97
4.3.6	Classifier Properties	98
4.4	Discussion	98
4.5	Acknowledgments:	105
4.6	Author contributions:.....	106
4.7	Potential conflicts of interest:	106
4.8	Supplementary Material	107
4.8.1	Developing Signal-to-noise Metric	107
4.8.2	Simulation Study	109
5.	CONCLUSIONS AND EXTENSIONS	132
	REFERENCES CITED.....	139

LIST OF TABLES

Table	Page
2.1 Case 1: Single variable changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	32
2.2 Case 1: All variables changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	33
2.3 Case 1: Variables changing differently. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	34
2.4 Case 2: Single signal changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	35
2.5 Case 2: Both signals changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	36
2.6 Case 2: Signals changing differently. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	37
2.7 Case 3: Low independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	38
2.8 Case 3: Moderate independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	39

LIST OF TABLES – CONTINUED

Table	Page
2.9 Case 3: High independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.....	40
3.1 Results from simulation study. Mean t -statistics (and standard errors) from each of the considered CCR projections across all cases considered.	81
4.1 Types of $\mathbf{p}$ considered.....	116
4.2 Other simulation parameters.....	116
5.1 CCR Projection coefficients from crabs dataset.	138

LIST OF FIGURES

Figure	Page
2.1 Example path diagram of a multilevel growth curve model with $P=4$ variables measured at $T=3$ unique time points. For identifiability, the loading on the first manifest variable is set to one.	41
2.2 Profile of objective function for two variables with equivalent SNR across varying levels of residual covariance. Optimal solutions become less clearly defined (i.e., flatter surface) as residual covariance of two variables increases.	42
2.3 Simulation results for case 1: one variable changing. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	43
2.4 Simulation results for case 1: all variables changing equally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	44

LIST OF FIGURES – CONTINUED

Figure	Page
2.5 Simulation results for case 1: variables changing unequally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	45
2.6 Simulation results for case 2: one signal changing. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	46
2.7 Simulation results for case 2: both signals changing equally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	47

LIST OF FIGURES – CONTINUED

Figure	Page
2.8 Simulation results for case 2: signal changing unequally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	48
2.9 Simulation results for case 3: low independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	49
2.10 Simulation results for case 3: moderate independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	50

LIST OF FIGURES – CONTINUED

Figure	Page
2.11 Simulation results for case 3: high independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).	51
2.12 Results from simulation study to quantify the Type I error of the CCR projections across a variety of data structures. In each panel, the estimated Type I error rate based on 1000 simulations along with a 95% score interval is displayed.	52
2.13 Example data from Phase II clinical trials of progressive MS. Each patient was measured three times at 6 month intervals leading to the inception of the trial. Each panel represents a different clinical scale measured. The x-axis corresponds to time and the y-axis corresponds to the measurement scale. Lines connect measurements from the same patient.	53
2.14 A beanplot of 500 permutations of validation data test-statistics for CCR solution in addition to individual clinical scales. Black lines represent individual test statistics, red lines show average of test statistics for each scale, yellow areas represent non-parametric density curves of individual test statistics, grey-dashed lines correspond to approximate cut-offs for 5% significance level tests.	54

LIST OF FIGURES – CONTINUED

Figure	Page
2.15 Beanplots of coefficient weights for each scale assigned from 500 permutations of training data for CCR projection (top-left), a bounded CCR (B-CCR) projection (top-right), weights from a SEM (bottom-left), and weights from the first principal component (bottom-right). Black lines represent individual coefficient values, red lines show average of coefficients for each scale, yellow areas represent non-parametric density curves of individual coefficients, grey-dashed lines correspond to 0 (i.e., zero influence).	55
2.16 B-CCR projection in the example data from Phase II clinical trials of progressive MS. Each patient was measured three times at 6 month intervals leading to the inception of the trial. Lines connect measurements from the same patient.	56
3.1 Results from simulation study. a) For each case (rows) and pairwise correlation considered (columns), t -statistics from the independent test dataset for each of the variations of the CCR. b) For each case (rows) and pairwise correlation considered (columns), the mean estimate of correct sparsity, with approximate 95% Wald confidence intervals.	78
3.2 Solution paths from Case 3. With pairwise correlation set to 0 (a) and 0.3 (b), values of c_j are plotted versus $\log(\lambda)$ for unbounded CCR with soft-thresholding (top-left) and hard-thresholding (top-right) and B-CCR with soft-thresholding (bottom-left) and hard-thresholding (bottom-right).	79
3.3 a) Mean cross-validated objective function estimates. Estimates are plotted versus $\log(\lambda)$ for bounded CCR with soft-thresholding. b) Solution paths for motivating example. Values of c_j are plotted versus $\log(\lambda)$ for bounded CCR with soft-thresholding, with optimal $\log(\lambda^*)$ providing 82 non-zero c_j 's.	80

LIST OF FIGURES – CONTINUED

Figure	Page
4.1 Technical (top graphs) and biological (bottom graphs) variance calculation for SOMAmer SL004672. The x-axes correspond to the patient/sample that measurements were produced from. The upper panels show technical replicates (n=88) and the lower panels show biological replicates (n=24). The left panels show the raw measurements (natural-log scale RFU) for SOMAmer SL004672 on the y-axes. The right panels show the identical raw observations with the random intercept effect subtracted to account for subject-to-subject variation. These residuals (after subtracting means of technical or biological replicates) were used to estimate the technical and biological variance, respectively. The horizontal black line is the estimated mean from the technical (top graphs) and biological (bottom graphs) variance models.	117
4.2 Differences in biomarker measurements (log10 RFU) in identical samples analyzed repeatedly (n=88, left), in longitudinal HD samples measured at different time points (n=24, middle), and in patient samples before and after application of immunomodulatory therapy (n=10, right): (i) An average of Spearman rho values calculated across 500 high-signaling SOMAmers with high SNR, and (ii) an average of variabilities (a median of relative percent changes calculated as absolute difference of RFUs for each of the 500 high-signaling SOMAmers between two replicates divided by the average of the two RFUs) for all pairs of replicates in each respective category. Examples of the strongest and the weakest correlations between two samples in each category are visualized on 500 high-signaling SOMAmers	118

LIST OF FIGURES – CONTINUED

Figure	Page
4.3 The SOMAscan assay comprises 1128 SOMAmers (solid black line curves). Calculation of technical and biological SNR reduced the number of SOMAmers considered for further analysis to 500 (dashed red line curves). Using the 500 high-signaling SOMAmers, 124750 biomarker ratios were generated that were subsequently tested for their SNR and their ability to differentiate two diagnostic groups (based on AUC in the modeling cohort), resulting in 5401 high-signaling biomarker ratios for MS versus non-MS diagnostic test, 3626 biomarker ratios for progressive versus RRMS diagnostic test, and 1504 biomarker ratios for SPMS versus PPMS diagnostic test (green dotted line curves). Out-of-bag AUC estimates (bottom graphs) examined from different random forests generated by sequentially adding ratios with the highest variable importance led to a logical cut-off (marked by solid red line) of 22 SOMAmer ratios for MS versus non-MS diagnostic comparison, 21 SOMAmer ratios for RRMS versus progressive MS diagnostic comparison, and 33 SOMAmer ratios for SPMS versus PPMS (blue dash-dot line curves).	119

LIST OF FIGURES – CONTINUED

Figure	Page
4.4 A parallel coordinate plot (PCP) for the 22 most important features that distinguish MS from non-MS. The plot displays individual patients from combined modeling (n=225) and validation cohort (n=85) divided into MS group (RRMS, PPMS, SPMS; thin red lines) and non-MS group (HD, NIND, OIND; thin blue lines). A group average is shown as thick yellow line for PPMS, thick red line for RRMS, thick orange line for SPMS, thick purple line for HD, thick green line for NIND, thick blue line for OIND. The y-axis shows SOMAmer natural log ratios scaled to 0-1 range. SOMAmer ratios were grouped based on the cellular origin and known functions of the individual components into nine groups. Different cell types are shown above the PCP to highlight the cell origin of individual SOMAmers. *ratios in the classifier are inverted.	120
4.5 A parallel coordinate plot (PCP) for the 21 most important variables that distinguish progressive MS from relapsing MS. The plot displays individual MS patients from combined modeling (n=120) and validation cohort (n=47) divided into progressive MS group (PPMS and SPMS; thin blue lines) and relapsing MS group (RRMS; thin red lines). A group average is shown as thick purple line for PPMS, thick blue line for SPMS, and thick red line for RRMS. The y-axis shows SOMAmer natural log ratios scaled to 0-1 range. SOMAmer ratios were grouped based on the cellular origin and known functions of the individual components into nine groups. Different cell types are shown above the PCP to highlight the cell origin of individual SOMAmers. *ratios in the classifier are inverted.	121

LIST OF FIGURES – CONTINUED

Figure	Page
4.6 (A) STARD diagram and (C) confusion matrix for 85 subjects used for validation of the MS molecular diagnostic test and (B) STARD diagram and (D) confusion matrix for 47 subjects used for validation of the progressive MS molecular diagnostic test.....	122
4.7 (A) The MS diagnostic probability (on the y-axis) of 85 subjects from the validation cohort is shown in the graph. Blue circles represent subjects with original non-MS diagnosis (HD, OIND, NIND) and orange circles represent subjects with original MS diagnosis (RRMS, PPMS, SPMS). The red line represents an arbitrary cut-off at 50%. The pink background marks an area between 30% and 70% where the certainty of the molecular classification is weak (contains 22.4% of the validation cohort's subjects). The orange background highlights 70.0% of the validation cohort's MS subjects with highly probable MS molecular diagnosis (>70%) and the blue background labels 86.8% of the validation cohort's non-MS subjects with high probability of non-MS molecular diagnosis (<30%). Grey bars represent a frequency distribution bar chart. Misdiagnosed subjects (pink circles) were evaluated for non-SOMAllogic biomarkers of inflammation: IgG index (B), BCMA (C), sCD27 (D), and CHI3L1 (E) using alternative assays (Komori et al., 2016). The group medians are shown for MS subjects as an orange line and for non-MS subjects as a blue line.....	123

LIST OF FIGURES – CONTINUED

Figure	Page
4.8	Simulation results for the “equal” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g
	124
4.9	Simulation results for the “high-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g
	125

LIST OF FIGURES – CONTINUED

Figure	Page
4.10 Simulation results for the “low-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	126
4.11 Simulation results for the “signal-high” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	127

LIST OF FIGURES – CONTINUED

Figure	Page
4.12 Simulation results for the “signal-low” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	128
4.13 Simulation results for the “signal-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	129

LIST OF FIGURES – CONTINUED

Figure	Page
4.14 Simulation results for the “strong-separation” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	130
4.15 Simulation results for the “weak-separation” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g	131
5.1 Distribution of CCR values across different groupings.....	138

LIST OF ALGORITHMS

Algorithm	Page
1 Bounded CCR Algorithm	31
2 Bounded CCR Algorithm with Soft-thresholding.....	74
3 CCR Algorithm with Soft-thresholding.....	75
4 Bounded CCR Algorithm with Hard-thresholding	76
5 CCR Algorithm with Hard-thresholding	77

ABSTRACT

In recent years, increasing amounts of complex biological data are being collected on patients in many branches of medical research. Many of these signals are being collected with a certain amount of imprecision in the attained measurements. Two such areas in multiple sclerosis (MS) research are clinical scale development and proteomics analysis.

Scales are often constructed from multiple outcome measures to create a combined metric that is a better measure of the true trait of interest than any of the original components. When the interest is in creating a scale that is sensitive to changes over time, developing it using cross-sectional data may not tune the projection to detect changes over time optimally. The proposed methodology, coined the Constructed Composite Response (CCR), was developed to maximize detected longitudinal change. A simulation study, and analysis of a motivating dataset, demonstrated that the CCR methodology performs better at capturing longitudinal change than traditional techniques. Including sparsifying constraints, motivated by penalized regression models, improved the performance of the CCR in high-dimensional data.

In proteomics data, undesirable sources of variation are often present. Examples include temporal fluctuation in control samples and technical variability from multiple assay runs. When developing a molecular classifier of MS, a novel variable screening procedure was implemented to eliminate proteins with high levels of these unwanted sources. A simulation study compared this with traditional screening approaches and findings are discussed. Future extensions and directions of research are also discussed.

CHAPTER ONE

INTRODUCTION

An abundance of data have become available in medical disciplines in recent years. The ability to measure increasingly complex biological signals has led to a wealth of information available to clinicians and researchers when diagnosing and preventing diseases on a patient and/or molecular level, such as individual presence or absence of genes, abundance of cellular proteins, and even individual longitudinal movements and bodily functions. On another technological front, advances in modern computing power have led to methodological developments borrowing from statistics, applied mathematics, and computer science to improve the quality of analyses in such data. The term “big data” has become a buzzword in social and professional settings in many scientific disciplines. While we now have an ability to measure thousands (even millions) of biological signals on patients, measuring some of these signals with perfect accuracy is currently not achievable. Thus, many of the signals being collected in the medical domain today have a certain amount of imprecision in the reported quantities. In other words, there is a certain amount of measurement error inherent in these quantifications. The purpose of this research is to address specific statistical problems in medical research when signals are measured with imprecision, and is motivated by three applications in multiple sclerosis (MS) research. MS is an inflammatory neurological disease that affects the central nervous system. Symptoms of MS include muscle weakness/difficulty with balance and coordination, impaired sensory feeling, blurred/double vision and even blindness.

Traditional measurement error models are a suite of statistical tools that

attempt to correct for known (or assumed) measurement error, adjusting parameter estimates to correct biases induced from said error (Fuller, 1987, Carroll et al., 2006, Buonaccorsi, 2010). In these models, much is assumed to be known in order to perform these corrections, including the type of measurement error present (for example, additive versus multiplicative) as well as an estimate of the measurement error variance, either assumed or estimated. Some popular measurement error correction techniques include method of moments corrections (Buonaccorsi, 2010), the SIMEX correction (Cook and Stefanski, 1994), and Bayesian methods (Carroll et al., 2006). However, in many biological applications, quantifying the amount of measurement error can be either financially costly, impractical, or in some cases impossible, given the application. Quantifying the amount of disease severity in a neurological disorder such as MS is a difficult problem, and gold-standards in such diseases are typically based on measures developed exclusively through expert opinion. In turn, estimating a variance component for measurement error corrections using such outcomes would also be based on said expert opinion, with no verifiable way to assess the quality of the estimate. Conversely, while replicated biological samples can be taken to separate sources of unwanted variation, these are often costly to attain, and the incorporated information often cannot be directly used in more complicated modeling regimes. This research takes the approach of recognizing imprecision in the measured signals, and combining available information in a way to reduce the imprecision in estimating the underlying “true” signal.

The first application is motivated by an interest in developing more sensitive endpoints for progressive MS clinical trials, combining clinical scales currently in practice via a linear projection. Various clinical scales attempt to quantify disease severity in patients, but they have varying degrees of sensitivity to detect disease progression and various levels of imprecision in measuring the underlying disease

severity. For example, the Expanded Disability Status Scale (EDSS, Kurtzke, 1983) is a 0-10 Likert-scale clinical endpoint that is currently utilized by the Food and Drug Administration (FDA) for MS clinical trials. However, once a patient becomes unable to walk (around an EDSS of 7), EDSS is unlikely to change again for several years even though the patient’s disease is still progressing. This unequal change over time is a weakness of the scale and represents measurement error in quantifying the underlying disease severity in a patient.

The developed methodology for this application, coined the Constructed Composite Response (CCR) in Chapter 2, involves determining the coefficients for the linear projection of clinical endpoints that maximize a specified t -statistic from a traditional linear model. While this methodology represents a novel statistical advancement in latent factor construction and clinical endpoint development, it is not the first technique to determine linear combinations that maximize statistical quantities. The generated effect modifier (GEM, Petkova et al., 2017), which determines the linear combination of variables that, when included as a moderator variable, maximizes the detected interaction with a treatment factor on a response variable of interest in clinical trial data, is perhaps the most similar to the CCR methodology. The stark difference between the CCR and the GEM is that the CCR projection is designed to create a response variable in subsequent modeling schemes, not an optimal predictor or moderator.

Other approaches have constructed linear projections to be used to create a response for future analyses. In research motivated by studies in acute kidney injury (Meisner et al., 2017), coefficients were selected to maximize the estimated true positive rate for a selected fixed false positive rate in 2-group problem. Linear Discriminant Analysis (LDA, Fisher, 1936, McLachlan, 2005) determines a linear combination of variables that best linearly separates two groups of observations.

In fact, LDA is a special case of the CCR, and can be recovered if the target is maximizing the t -statistic for a 2-group single factor linear model. However, the CCR methodology can be extended further than LDA, accounting for control variables, while still maximizing detectable change in a model coefficient. Although not discussed in the original CCR manuscript, this methodology can be extended to maximize an F-statistic from a specified contrast vector or matrix. This would allow for linear projections targeting more complicated model components, such as group differences with more than two groups, quadratic change over time, and/or interaction terms if the application suggested these targets.

The second application is motivated by extending the CCR methodology for use in high-dimensional problems (Chapter 3). While the CCR methodology (Chapter 2) shows promise for constructing latent traits targeted to detect changes over time with more sensitivity, it does not allow for exact thresh-holding of unimportant features or automatic variable selection. Regularization is the process of incorporating additional structure or constraints into a problem to achieve a desired result, where the goal of the problem drives the form of the regularization. For example, sparse reconstructions of images can greatly reduce the amount of storage space required, and are often used in image analysis. Regularization is also inherent in Bayesian statistics through the choice of prior distributions on model parameters, particularly in small sample size problems where additional structure is typically incorporated to achieve sensible posterior distributions. A class of regularization techniques used in high-dimensional regression, penalized regression models, have gained popularity in recent years, with entire textbooks devoted to such problems (e.g., Hastie, Tibshirani, and Wainwright, 2015). These techniques involve constraining the solution space, and different constraint regions lead to different properties in the returned solutions. We hypothesized that constraining the projection coefficients for the CCR to lie

in specific constraint regions with desirable properties had potential to strengthen the interpretability and generalizability of the CCR solution, especially in high-dimensional problems. A regularized set of variables for scale construction can also provide considerable cost savings in some applications.

The final application is motivated by an interest in developing molecular-based diagnostic classifiers of MS, as well as its progressive stage. Proteomics data, or data on a large number of cellular proteins, allow researchers to examine disease mechanisms on a molecular level. Identifying biological processes affected by diseases, and whether or not they are present in a particular patient, can allow for therapeutic regimes designed specifically for individualized treatments. Proteomics has led to breakthroughs in modern oncology, notably in breast cancer research (Sorlie et al., 2003), where protein combinations have helped to uncover tumor sub-types with biologically distinguishable features. While this active area of research shows much promise, issues exist in quantifying and using this information. Many sources of variability are inherent in these data, such as natural fluctuation absent of disease processes and technical effects due to different assay runs at different collection centers. Recently, MS research has discovered rich diagnostic information in proteins contained in cerebrospinal fluid (CSF) (e.g., Bielekova et al., 2012; Johanson et al., 2008; Komori et al., 2015, 2016). When developing a CSF-based molecular classifier of MS (Barbour et al., 2017), different screening procedures were used to reduce the number of potential biomarkers into a parsimonious set with optimal predictive abilities. Through this project, a novel form of variable screening, based on the quality of measurement (or in other words, the relative amount of undesirable sources of variability) led to substantial improvements in diagnostic classifiers predictive abilities compared to other forms of variable screening. An appended section to the original manuscript (Barbour et al., 2017) was included, containing a simulation

study examining how this novel screening procedure compares to three traditional approaches: no screening, screening based on marginal relationships (such as univariate correlation measurements or t -tests), and screening based on marginal variance estimates.

The dissertation is organized as follows; Chapter 2 contains the submitted manuscript that motivates the CCR methodology and determines both analytic solutions and an algorithm allowing for coefficients in the projection to be bounded in the non-negative orthant; Chapter 3 extends the CCR methodology to incorporate sparsity inducing penalties on the CCR projection, and concludes with a simulation study comparing different versions of the CCR in low- and high-dimensional situations; Chapter 4 contains the accepted manuscript that develops the molecular classifier of MS and progressive MS, with an additional section containing a simulation study examining the performance of the “data-quality” screening approach implemented in the original paper; Finally, Chapter 5 contains discussion of the findings, future directions for methodological improvement, and conclusions.

CHAPTER TWO

CONSTRUCTED COMPOSITE RESPONSE: A FRAMEWORK FOR
CONSTRUCTING TARGETED LATENT VARIABLES.Contribution of Authors and Co-Authors

Author: Christopher R. Barbour

Contributions: Responsible for the majority of the writing and development of the methodology.

Co-Author: Dr. Mark Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscripts.

Co-Author: Dr. Dominique Zosso

Contributions: Constructed the framework for the developed algorithm and provided feedback on the methodology and drafts of the manuscript.

Co-Author: Dr. Bibiana Bielekova

Contributions: Provided feedback on the scientific application and drafts of the manuscript.

Manuscript Information Page

Authors: Christopher Barbour, Mark Greenwood, Dominique Zosso and Bibiana Bielekova

Journal: *Biometrika*

Status of Manuscript:

☒ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

Abstract

Clinical scales are diagnostic measurements used by clinicians to predict the presence of a disease or level of severity for a disease for individual patients. Scales such as these are often constructed from multiple outcome measures to create a combined metric that better measures the true underlying trait of interest than any of the original outcome measures. These methods typically focus on explaining cross-sectional variation in the outcomes using projections into one or more dimensions to define the combinations of variables. When the interest is in creating a scale that is sensitive to changes over time, developing it using cross-sectional data may not tune the projection to detect changes over time optimally. This research develops methodology for scale creation that is optimized to detect changes over time in longitudinal data. An overview of statistical methods traditionally used in scale development is given. The proposed method, Constructed Composite Response (CCR), is then presented and a simulation study is performed and discussed to examine properties of each projection method across a variety of different data structures. The method applied to a motivating dataset of multiple sclerosis (MS) patients emphasizes its strengths and highlights potential improvements to the CCR. Future extensions are also discussed, such as inclusion of sparse-learning strategies.

2.1 Introduction

There is a long history in statistics and psychometrics of creating scales and performing dimension reduction using weighted combinations of multivariate data (for example, see Spearman, 1904, Anderson and Rubin, 1956, Bollen, 1989). Here we focus on a version of this problem where we are interested in projecting a set of P variables measured on a set of K observations, observed at T points in time in fixed intervals, into an underlying latent factor, or latent scale, assumed to change monotonically over time. This problem is motivated by an interest in developing more sensitive endpoints for progressive multiple sclerosis (MS) clinical trials, combining clinical scales currently in practice. Each of these available scales suffers from some form of measurement error. For example, the Expanded Disability Status Scale (EDSS, Kurtzke, 1983) is a 0-10 Likert-scale clinical endpoint that is currently utilized by the Food and Drug Administration (FDA) for MS clinical trials. However, once a

patient becomes unable to walk (around an EDSS of 7), EDSS is unlikely to change again for several years even though the patient’s disease is still progressing. This unequal change over time is a weakness of the scale and represents measurement error in addressing the underlying disease severity in a patient. Despite its flaws, higher values on EDSS are related to higher levels of disability and it does contain some useful information.

While statistical methods for constructing latent factors and dimensional reduction techniques primarily stem from social science and education research, they have made their way into other scientific domains such as ecology and medicine in recent years. These methods typically focus on explaining cross-sectional variation in the responses using projections into one or more dimensions to define the combinations of variables. In the following, unbolded letters represent scalars, bold lower-case letters represent vectors, and bold upper-case letters represent matrices.

Principal component analysis (PCA, Pearson, 1901, Hotelling, 1933, Hardle and Simar, 2012) is one of the oldest multivariate projection techniques, and takes P variables, $y_1, \dots, y_P$, and creates a new set of orthogonal variables, $y_1^*, \dots, y_L^*$, with $L < P$ that are linear combinations of the original variables. These linear combinations are selected to explain the most variation observed in the original variables. Equivalently, the linear combinations can be formulated as projecting into a lower dimensional space where a measure of the total error is minimized. PCA can be useful in situations where our variables are correlated and can be used to simplify the dimension of our variables’ space. PCA is performed by determining the eigen-decomposition of the real-valued correlation (or covariance) matrix, with coefficient weights for principal components (PCs) being assigned as the eigenvectors associated with decreasing eigenvalues. Scales are extracted by using the PC weights as a linear projection of the original variables.

Similar to PCA, exploratory factor analysis (EFA, Spearman, 1904, Gorsuch, 1983) projects a multivariate set of variables into a smaller set of underlying latent traits that well approximates the covariance structure among the original variables. The traditional EFA model assumes that we have P manifest variables (our observable data $\mathbf{Y}_{n \times P}$ where n is the total sample size and columns representing the observed variables for a single observation) and L unobservable latent traits ($\mathbf{f}_{L \times n}$, assumed to have 0 mean and unit variance without loss of generality and columns representing the latent variables for a single observation) where $P < L$, with the model defined as

$$\mathbf{Y} = \mathbf{\Lambda} \mathbf{f} + \mathbf{u} \text{ and}$$

$$Var(\mathbf{Y}) = \mathbf{\Lambda} \mathbf{\Lambda}^\top + \mathbf{\Psi},$$

where $\mathbf{\Lambda}$ is a $P \times L$ matrix of the factor loadings, $\mathbf{u}$ is a $P \times n$ matrix of measurement errors in the observed variables, and $\mathbf{\Psi}_{P \times P} = Var(\mathbf{u})$. There are multiple ways to estimate these parameters (see Gorsuch, 1983 and Krzanowski, 1988 for more information). A primary criticism of EFA is that these factor loadings are not unique. To see this, consider introducing a $L \times L$ orthogonal matrix $\mathbf{U}$ ($\mathbf{U}^\top \mathbf{U} = \mathbf{U} \mathbf{U}^\top = \mathbf{I}$) into the model:

$$\mathbf{Y} = (\mathbf{\Lambda} \mathbf{U}) (\mathbf{U}^\top \mathbf{f}) + \mathbf{u} \text{ and}$$

$$Var(\mathbf{Y}) = (\mathbf{\Lambda} \mathbf{U}) (\mathbf{\Lambda} \mathbf{U})^\top + \mathbf{\Psi} = \mathbf{\Lambda} \mathbf{\Lambda}^\top + \mathbf{\Psi}.$$

In practice, standard rotations are used to try and get a more interpretable solution, such as the varimax rotation (Kaiser, 1958).

Structural equation models (SEM, Bollen, 1989) are a suite of multivariate statistical models that are frequently used in social and behavioral sciences, although they are gaining popularity in other areas such as ecology (Grace, 2006) and

medicine (Hardcastle et al., 2018). They allow for estimation of links between (unobserved) latent traits and (observed) manifest variables, while allowing for modeling of relationships between latent traits as well as relationships with other observed variables (i.e., regressions). This framework has the benefit over PCA and EFA of the potential to incorporate control variables, such as group level differences. In many clinical settings, we would expect patients to have varying degrees of disease severity at the time of measurement. Failure to account for these could lead to sub-optimal construction of the underlying factor. Confirmatory Factor Analysis (CFA, see Brown, 2006) is a special case of a SEM and is related to EFA, where instead of using the data to uncover potential latent factors as in EFA, we specify the structure of latent factors and can test if our data are consistent with the assumed structure as well as estimating the specified latent factors.

The structure of the SEM for longitudinal data such as these is commonly referred to as a multilevel latent growth curve model (e.g., Bollen and Curran, 2006, Duncan, Duncan, and Strycker, 2006, Little, 2013), where a common latent trait is constructed from the observed variables, and mean patient values (intercepts) and trajectories over time (linear, quadratic, etc.) are estimated. An example path diagram of the assumed SEM model is displayed in Figure 2.1. This path diagram was generated using the `lavaan` (Rosseel, 2012) and `semPlot` (Epskamp and Stuber, 2017) R packages (R Core Team, 2018). This model assumes a) equal loadings for individual observed variables at each time point (also referred to as tau-equivalence, see Brown, 2006), b) equal “error” in measured variables at each time point, c) observation starting values are allowed to vary while the individual slopes are not (same average change over time for all subjects), and d) the longitudinal change is linear. This model is similar to a random-intercept linear mixed model with a common linear time trend for all subjects.

In both EFA and SEM, constructing the latent factor scores (or the predicted “value” of a latent trait) is useful for future analysis. However, there is no single recognized method for calculating these scores (Grice, 2001, DiStefano, Zhu, and Mindrila, 2009). One option is to simply multiply the original (scaled) variables by the estimated factor loadings (called the “weighted sum” method). Other more refined methods exist for preserving specific properties (correlations among loadings and factor scores), which include Bartlett’s method (Bartlett, 1937) and the Regression Method (Thurstone, 1935).

The previously described methods are successful in latent factor construction and dimension reduction when approximating the covariance or correlation matrix of the data is the target. If the interest is in constructing a latent factor that is known (or is assumed) to change monotonically over time, targeting the covariance in the data would not properly allocate coefficient weights to the variables that are changing most over time, unless the level of cross-sectional variation/correlation matched with the level of changes over time. The research in this paper focuses on methodology, coined the Constructed Composite Response (CCR), for scale creation that is optimized to detect changes over time in longitudinal data. Through a simulation study and real-data example, it is shown that the CCR produces scales with equivalent or more sensitivity to detect longitudinal change than previously discussed methods.

The manuscript is organized as follows. In Section 2, we define our proposed methodology, CCR, for unconstrained and constrained solutions. We also develop a permutation based p-value calculation quantifying the strength of the signal in the projection. Section 3 compares our algorithm with traditional scale development techniques via a simulation study. Section 4 applies our method to a motivating dataset of progressive MS patients, where interest lies in constructing a more sensitive and better clinical endpoint to detect changes in disease progression. Section 5

discusses findings and considers future extensions.

2.2 Proposed Methodology

2.2.1 Constructed Composite Response

When interested in creating a scale that is sensitive to changes over time, developing it using cross-sectional data and/or with methods designed for cross-sectional data (e.g., PCA, EFA) may not tune the projection to optimally detect this assumed change. Instead of focusing on creating a projection that optimizes cross-sectional correlation or variance, the proposed method creates projections that maximize a t -statistic for a change over time in a specified linear model. This metric combines the estimated change over time ($\hat{\beta}_{time}$) and its precision ($SE(\hat{\beta}_{time})$). If modeling a clinical scale, larger t -statistics for a time trend correspond to detecting disease progression with more sensitivity, a desirable characteristic in clinical trials. This formulation also allows control variables, such as baseline age or different patient starting values, to be accounted for in the projection.

Let $\mathbf{y}_{jt}$ be a length- P vector containing the P observed response variables for observation (e.g., patient) j ($j = 1, \dots, J$, where J is the total number of observations) at time point t ($t = 1, \dots, T$, where T is the total number of observed time points) and let $\mathbf{s}_p$ be a length- $(J \cdot T)$ vector containing the observed values of response variable p ($p = 1, \dots, P$) for all J subjects and all T time points. Then $n = J \cdot T$ is the total sample size and let $\mathbf{X}$ be a fixed n by q model matrix of full column rank (where q is the number of model parameters to be estimated) containing control variables to be accounted for (e.g., differences among patients) as well as the corresponding variable of interest (such as time). For a linear projection of the response variables at time t , $\mathbf{y}_{jt}$, into a CCR, $z_{jt} = \mathbf{c}^\top \mathbf{y}_{jt}$, we want to determine the row vector $\mathbf{c}^\top$ that maximizes the t -statistic for a linear time change in the assumed

data-generating linear model

$$z_i = \mathbf{X}_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.$$

Let $\hat{\boldsymbol{\beta}}_z$ correspond to the least-squares solution for a fixed projection $\mathbf{c}^\top$, and let $\hat{\boldsymbol{\beta}}_p$ correspond to the least-squares solution for the linear model predicting $\mathbf{s}_p$ from $\mathbf{X}$. Using standard linear model results (e.g., Monahan, 2008), we can show that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_z &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (c_1 \mathbf{s}_1 + c_2 \mathbf{s}_2 + \dots) \\ &= c_1 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{s}_1 + c_2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{s}_2 + \dots \\ &= \sum_{p=1}^P c_p \hat{\boldsymbol{\beta}}_p. \end{aligned} \tag{2.1}$$

Similarly, we can show that the residual vector from modeling $\mathbf{z}$ (call $\mathbf{r}_z$) is a weighted sum of the residual vectors from modeling $\mathbf{s}_p$ ($\mathbf{r}_p$).

$$\begin{aligned} \mathbf{r}_z &= \left(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right) (c_1 \mathbf{s}_1 + c_2 \mathbf{s}_2 + \dots) \\ &= c_1 \left(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right) \mathbf{s}_1 + c_2 \left(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right) \mathbf{s}_2 + \dots \\ &= \sum_{p=1}^P c_p \mathbf{r}_p. \end{aligned} \tag{2.2}$$

Assuming $\hat{\beta}_{time}$ corresponds to the estimated parameter for the assumed change (without loss of generality, assume this is in the first position of the coefficient vector),

simple algebraic results show that the solution to the objective function becomes

$$\begin{aligned}
\hat{\mathbf{c}} &= \operatorname{argmax}_{\mathbf{c}} \frac{|\hat{\beta}_{time}|}{SE(\hat{\beta}_{time})} \\
&= \operatorname{argmax}_{\mathbf{c}} \frac{|\sum_{p=1}^P c_p \hat{\beta}_{time-p}|}{\sqrt{\frac{1}{n-q-1} \sum_{i=1}^n \left(\sum_{p=1}^P c_p r_{ip} \right)^2 (\mathbf{X}^\top \mathbf{X})_{11}^{-1}}} \\
&= \operatorname{argmax}_{\mathbf{c}} \frac{|\sum_{p=1}^P c_p \hat{\beta}_{time-p}|}{\sqrt{\frac{1}{n-q-1} \sum_{i=1}^n \left(\sum_{p=1}^P c_p r_{ip} \right)^2}} \\
&= \operatorname{argmax}_{\mathbf{c}} \frac{|\sum c_p \hat{\beta}_{time-p}|}{\sqrt{\sum_p c_p^2 \hat{\sigma}_p^2 + \sum_p \sum_{p'} c_p c_{p'} \bar{r}_{pp'}}}
\end{aligned} \tag{2.3}$$

where $\hat{\beta}_{time-p}$ is the estimated change over time from modeling $\mathbf{s}_p$, $\hat{\sigma}_p^2$ is the mean-squared error from modeling $\mathbf{s}_p$, r_{ip} is the i 'th residual from modeling $\mathbf{s}_p$, and $\bar{r}_{pp'} = \frac{2}{n-q-1} \sum_i r_{ip} r_{ip'}$. These results show that the coefficient values for each variable ($\hat{c}_p$) are selected to balance the size of the individual coefficient of interest ($\hat{\beta}_{time-p}$) relative to how well the assumed model fits ($\hat{\sigma}_p^2$) and the covariance between pairs of model residuals ($\bar{r}_{pp'}$).

Since both the numerator and denominator of the above equation are always non-negative, the solution maximizing this objective function will also maximize the square of the objective function. Let $\boldsymbol{\gamma} = (\hat{\beta}_{time-1}, \hat{\beta}_{time-2}, \dots, \hat{\beta}_{time-P})^\top$ be a column vector of the coefficients from modeling the individual response variables $\mathbf{s}_1$ through $\mathbf{s}_p$ and let $\mathbf{B} = \boldsymbol{\gamma} \boldsymbol{\gamma}^\top$. Similarly, let $\mathbf{R}$ be a P by n matrix where row p corresponds to the residual vector from modeling $\mathbf{s}_p$, and set $\mathbf{M} = \frac{1}{n-q-1} \mathbf{R} \mathbf{R}^\top$. Note that both $\mathbf{B}$ and $\mathbf{M}$ are symmetric matrices. The solution to the objective function then becomes

$$\hat{\mathbf{c}} = \operatorname{argmax}_{\mathbf{c}} \frac{\mathbf{c}^\top \mathbf{B} \mathbf{c}}{\mathbf{c}^\top \mathbf{M} \mathbf{c}}.$$

If $\mathbf{M}$ is a positive-definite matrix, then an analytic solution exists up to a scaling constant, with the solution being the leading eigenvector of $\mathbf{M}^{-1}\mathbf{B}$ (Hardle and Simar, 2012). An equivalent and less expensive solution is found by computing the Cholesky decomposition of $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$ and reformulating the problem as

$$\begin{aligned}\hat{\mathbf{c}} &= \operatorname{argmax}_{\mathbf{c}} \frac{\mathbf{c}^\top \mathbf{B} \mathbf{c}}{\mathbf{c}^\top \mathbf{M} \mathbf{c}} \\ &= \operatorname{argmax}_{\mathbf{c}} \frac{(\mathbf{L}^\top \mathbf{c})^\top \mathbf{L}^{-1} \mathbf{B} \mathbf{L}^{-\top} (\mathbf{L}^\top \mathbf{c})}{(\mathbf{L}^\top \mathbf{c})^\top (\mathbf{L}^\top \mathbf{c})} \\ &= \operatorname{argmax}_{\mathbf{u}} \frac{\mathbf{u}^\top \mathbf{B}_* \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} \text{ with } \mathbf{u} = \mathbf{L}^\top \mathbf{c} \text{ and } \mathbf{B}_* = \mathbf{L}^{-1} \mathbf{B} \mathbf{L}^{-\top}.\end{aligned}\tag{2.4}$$

The solution, $\hat{\mathbf{c}}$, is obtained by left multiplying the leading eigenvector of $\mathbf{B}_*$, $\hat{\mathbf{u}}$, by $\mathbf{L}^{-\top}$ ($\mathbf{L}^{-\top} = (\mathbf{L}^{-1})^\top$), specifically $\hat{\mathbf{c}} = \mathbf{L}^{-\top} \hat{\mathbf{u}}$.

2.2.2 Examination of Exact Solutions

Considering 2 variables with no covariance in the model (i.e., $\bar{r} = 0$) and constraining the solution to be unit-length, we get the following intuitive result

$$(\hat{c}_1, \hat{c}_2) = \left(\frac{SNR_1}{\sqrt{SNR_1^2 + SNR_2^2}}, \frac{SNR_2}{\sqrt{SNR_1^2 + SNR_2^2}} \right)$$

where $SNR_p = \frac{\hat{\beta}_{time-p}}{\hat{\sigma}_p^2}$. This suggests that coefficient weights are assigned proportional to the magnitude of the individual coefficients of interest relative to how well the assumed model fits the response.

A less intuitive solution is presented when considering 2 variables with covariance in the model residuals (i.e., $\bar{r} \neq 0$). Constraining the solution to be unit-length, as

before, the solution is now

$$(\hat{c}_1, \hat{c}_2) = \left(\frac{SNR_1 - SNR_2 \left(\frac{\bar{r}}{\hat{\sigma}_1^2} \right)}{\alpha}, \frac{SNR_2 - SNR_1 \left(\frac{\bar{r}}{\hat{\sigma}_2^2} \right)}{\alpha} \right),$$

where

$$\alpha = \sqrt{\left(SNR_1 - SNR_2 \left(\frac{\bar{r}}{\hat{\sigma}_1^2} \right) \right)^2 + \left(SNR_2 - SNR_1 \left(\frac{\bar{r}}{\hat{\sigma}_2^2} \right) \right)^2}.$$

To explore this solution further, the two-variable solution is displayed graphically with varying degrees of residual covariance (Figure 2.2). The SNR for both variables were set to 1 ($SNR_1 = \frac{1.5}{1.5}$, $SNR_2 = \frac{1}{1}$) and the solution space was constrained to lie in quadrant 1 ($c_p > 0$ for all p). The value of the objective function was profiled for a grid of coefficient values with the assumed residual covariance taking values of 0 and then 0.1 to 1.5 (by 0.35 units). When no residual covariance is present, the solution for each scale is $\frac{\sqrt{2}}{2}$. As the residual covariance increases, the maximum value of the objective function begins to decrease, and the surface of the profiled objective becomes flatter, indicating a more unstable solution.

2.2.3 Bounded Constructed Composite Response

While the unbounded CCR solution provides a global optimum, it may be of interest in some applications to bound the directions of the individual coefficients (i.e., signs of specific elements of $\mathbf{c}$). For example, if higher values of a certain clinical endpoint are indicative of more disease severity, bounding the coefficient corresponding to that endpoint to be positive may aid in the interpretability of the resulting scale and lead to more scientifically valid solutions by preventing contradictory contributions to the scale (which the non-bounded problem can help identify).

The bounded CCR (B-CCR) problem is formulated as

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}} \mathbf{c}^\top \mathbf{B} \mathbf{c} \quad \text{s.t.} \quad \mathbf{c}^\top \mathbf{M} \mathbf{c} \leq 1, \quad \forall p: c_p > 0, \quad \mathbf{c}^\top \boldsymbol{\gamma} > 0.$$

This constrained quadratic optimization problem was solved using a loose variation of primal-dual hybrid gradients (PDHG) (Chambolle and Pock, 2011). Inspired by successful use of PDHG-type schemes for convex minimization problems in the context of elliptic Partial Differential Equations (Zosso et al., 2017), we derived an extension of the method to the non-convex case here. Roughly speaking, the scheme introduces Karush-Kuhn-Tucker (KKT) multipliers to address the quadratic denominator inequality constraint; further dual variables are introduced to address the quadratic terms in the Lagrangian using Legendre-Fenchel biconjugates; the resulting saddle point problem is solved using a proximal coordinate descent scheme, from which the latter dual variables can be eliminated and only remain as memory terms. In essence, the resulting scheme is first order gradient descent/ascent, but involving memory terms that boost the convergence speed. Although no theoretical guarantees of convergence are proven here (since maximizing a convex function), empirical results from a simulation study (Section 3) suggest this algorithm consistently converges to sensible solutions. The B-CCR algorithm is displayed in the appendix (Algorithm 1).

2.2.4 P-value calculation for CCR.

Due to the nature of the optimization problem, it is possible that the constructed projections achieving large t -statistics may be due to chance. In fact, in development of the CCR, we observed t -statistics as large as 3 or 4 were constructed from simulated data where no true change was actually present. In order to quantify the uncertainty that our projection would reproduce a similar result in datasets from the same

population, the following procedure can be used to calculate a permutation p-value for the projection, testing the null hypothesis of no change over time in the latent factor (for more information on permutation based procedures, see Berry, Johnston, and Mielke, 2014).

1. Perform the (B)CCR on the observed data $(\mathbf{X}, \mathbf{y}_1, \dots, \mathbf{y}_P)$ to get $\hat{\mathbf{c}}_{obs}$. Record the t -statistic from this projection, t_{obs} .
2. Assume (without loss-of-generality) that the coefficient to be maximized occurs in the first column of $\mathbf{X}$. Within each observational grouping (e.g., within each patient), permute the values of column 1 to create permuted data $(\mathbf{X}_{perm}, \mathbf{y}_1, \dots, \mathbf{y}_P)$.
3. Perform the (B)CCR on the permuted data $(\mathbf{X}_{perm}, \mathbf{y}_1, \dots, \mathbf{y}_P)$ to get $\hat{\mathbf{c}}_p$. Record the t -statistic from this projection, t_i .
4. Repeat steps 2-3 n_{perm} times.
5. Record $p_{perm} = \frac{\sum_{i=1}^{n_{perm}} I(|t_i| > |t_{obs}|) + 1}{n_{perm} + 1}$.

A p-value close to 0 would indicate that if a longitudinal change in the underlying latent factor did not exist, we would have a small probability that the (B)CCR projection would produce a t -statistic at least as extreme as the one observed. The number of permutations performed should be driven by the application and how accurately the p-value needs to be estimated.

2.3 Simulation Study

A simulation study was conducted to examine the performance of the various methods across a variety of simulated data. In each case, the cross-sectional (residual)

correlation and the signal-to-noise (change divided by residual variation) as well as the number of unique signals present were varied. A training dataset of $T = 4$ equally spaced time points for each of $J = 100$ subjects was drawn to construct the coefficients for each of the methods. A test dataset of the same size was drawn for assessing the ability of each projection to capture the underlying change over time (i.e., t -statistics from a linear model with a linear time trend and different intercepts across groups). Under each of the cases outlined below, this procedure was repeated 500 times. The data-generating model is

$$\begin{aligned} \mathbf{y}_{jt} &\stackrel{i.i.d}{\sim} MVN(\alpha_j + \boldsymbol{\beta}time_t, \mathbf{D}\mathbf{R}\mathbf{D}) \quad \text{with} \\ \alpha_j &\stackrel{i.i.d}{\sim} N(0, \sigma_\alpha^2), \end{aligned} \tag{2.5}$$

where $\boldsymbol{\beta}$ is a column vector of changes over time in individual scales, $\mathbf{D}$ is a square, diagonal matrix of standard deviations of individual scales, $\mathbf{R}$ is a symmetric matrix of cross-sectional correlations between pairs of scales, and σ_α is the group-to-group (or patient-to-patient) standard deviation in starting values.

Three primary cases will be considered: Case 1 - one signal present (defined as three variables with a non-zero cross-sectional correlation), Case 2 - two signals present, and Case 3 - one signal with three additional variables that are pairwise-independent with every other variable. Cross-sectional correlations were equal across all variables within a signal and took on values of 0.1, 0.4, and 0.7. The diagonal elements of $\mathbf{D}$ were equivalent ($\mathbf{D} = \sigma\mathbf{I}$), and were selected such that the SNR value corresponding to the largest element of $\boldsymbol{\beta}$ took values of 0.05 and 0.1. Starting values (α_j) for each patient were determined by varying σ_α such that the intra-class correlation (ICC) of two observations on the same patient, $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2}$, took values of 0, 0.3, and 0.6. All combinations of these inputs were used for each of the primary cases

considered.

For the single-signal case, three types of longitudinal change were considered: only one variable changing over time ($\beta = (1, 0, 0)^\top$), all variables changing equally over time ($\beta = (1, 1, 1)^\top$), and variables changing unequally over time ($\beta = (2, 1, 0)^\top$). For the two-signal case, three types of longitudinal change were considered: one signal is changing ($\beta_1 = \mathbf{1}^\top$, $\beta_2 = \mathbf{0}^\top$), both signals are changing equally ($\beta_1 = \beta_2 = \mathbf{1}^\top$), and both signals changing with different levels of change ($\beta_1 = \mathbf{3}^\top$, $\beta_2 = \mathbf{1}^\top$). Correlations between the two signals were set to 0. For the single-signal plus 3 independent variables case, three types of longitudinal change were considered: low-level of independent change ($\beta_{signal} = \mathbf{1}^\top$, $\beta_{ind} = \mathbf{1}^\top$), moderate-level of independent change ($\beta_{signal} = \mathbf{1}^\top$, $\beta_{ind} = \mathbf{2}^\top$), and high-level of independent change ($\beta_{signal} = \mathbf{1}^\top$, $\beta_{ind} = \mathbf{3}^\top$).

Under the extreme variance-covariance structures presented (i.e., lowest correlation and lowest ICC, lowest correlation and highest ICC, etc.), the Type I error of the permutation test (Section 2.4) was estimated. To perform this, all elements of β were set to 0, and 1000 training datasets were drawn. For each dataset, the procedure outlined in Section 2.4 was performed using 999 permutations, and the corresponding p -value recorded. The Type I error was estimated as the proportion of p -values less than or equal to 0.05 in the 1000 training datasets.

The following methods were applied: CCR, bounded CCR, SEM, PCA, and EFA. For the CCR method, in addition to the exact solution presented, the derived algorithm minus the non-negativity constraint was implemented to assess the convergence properties. For the SEM, both Bartlett factor scores and standardized factor scores were used to assess the projection's ability to detect change. For EFA, standardized factor scores were used. The resulting EFA projections were nearly identical to those produced from PCA of the correlation matrix, and are not displayed

in the results section.

2.3.1 Convergence of CCR algorithm

Across all the simulated datasets, the Euclidean-norm of the difference between the projection coefficients using the analytic CCR ($\hat{\mathbf{c}}_{CCR}$) and the projection coefficients derived from the B-CCR algorithm (Algorithm 1) without the non-negativity constraint ($\hat{\mathbf{c}}_{A-CCR}$) was calculated. The average observed $\|\hat{\mathbf{c}}_{CCR} - \hat{\mathbf{c}}_{A-CCR}\|$ across all simulations was approximately 0.007. The maximum absolute difference between the validation t -statistics from using the analytic CCR (t_{CCR}) and the validation t -statistics using the algorithm-derived CCR (t_{A-CCR}) was approximately 3.09×10^{-12} . This suggests that Algorithm 1 achieves the global solution in these lower dimensional settings.

2.3.2 Projection Comparison

The results of the simulation study are displayed in Table 1 through Table 9 and Figure 2.3 through Figure 2.11. All sets of projection coefficients have been standardized to unit length for comparability. When the level of cross-sectional correlation is small, or the patient-to-patient variation is small relative to the residual error, the SEM has trouble converging (between 50% to 90% of SEMs for the simulated datasets failed to converge), while the CCR methods are able to converge regardless of the relative correlation structures and underlying changes.

In the single-signal case with one variable changing (Table 1 and Figure 2.3), CCR projections consistently produce larger t -statistics in the validation data compared to SEM and PCA projections. When all variables within the signal are changing equally (Table 2 and Figure 2.4), all methods produce projections with each variable receiving equivalent weights. As the level of cross-sectional correlation increases, the gains in performance above individual variables decrease (for example,

left-column in Figure 2.4a). When variables have different levels of change within the signal (Table 3 and Figure 2.5), the CCR methods improve substantially upon SEM and PCA methods, with bigger gains occurring when cross-sectional correlation is high (for example, right-column in Figure 2.5a). Increasing the ICC does not seem to have a noticeable effect on the t -statistics for the CCR projections (for example, top-row in Figure 2.5a and Figure 2.5b). This would be expected since the CCR method can account for this source of variation. As the level of cross-sectional correlation increases, CCR solutions tend to produce larger t -statistics than B-CCR solutions (bottom-row Figure 2.5a). When high cross-sectional correlation is present, CCR gives negative weight to variables that are not changing, while B-CCR thresholds these coefficients at zero (bottom-row in Figure 2.5b). As noted earlier, the variability in the CCR projection coefficients increases as the correlation between variable pairs increases, while the variability in the projection coefficients decreases for SEM and PCA.

In the two-signal case with one signal changing (Table 4 and Figure 2.6), CCR projections assign small weights to the variables within the signal that are not changing (for example, top-row in Figure 2.6b). When the level of cross-sectional correlation is small, the CCR and B-CCR produce larger t -statistics than the other projection techniques, with these gains becoming more evident with increased patient-to-patient variation (for example, top-row in Figure 2.6a). With high levels of cross-sectional correlation, similar t -statistics are produced from all projection methods (bottom-row of top-portion of Figure 2.6). When both signals are changing equally (Table 5 and Figure 2.7), all projections produce similar t -statistics, also observed in the single signal case. Interestingly, this was also observed in the signals changing at different levels scenario (Table 6 and Figure 2.8), when it could be expected that the CCR projections should perform better.

The performance gains of the CCR methods over SEM and PCA methods are greatest in the single-signal plus three independent variables case (Table 6 through Table 9 and Figure 2.9 through Figure 2.11), with greater improvements as the level of independent change increases. This increased performance also occurs as the level of cross-sectional correlation within the signal increases (for example, left-column in Figure 2.11a).

2.3.3 Type I error estimation

The results from the Type I error estimation simulation are displayed in Figure 2.12. Across the range of correlation structures, both the CCR and B-CCR display estimated Type I error rates close to the nominal Type I error rate, with the 95% Score intervals containing the desired $\alpha = 0.05$ cutoff. This shows that the permutation procedure is well-calibrated. While we did not directly explore the power of the procedure, the CCR produced a small p-value in our application and in various situations where there was true change over time to be detected. The main purpose of the test is to avoid over-optimistically created scales and the simulation results show that it is well-calibrated to at least prevent detection of false signals through the optimization process.

2.4 Real Data Example

We now demonstrate the performance of the CCR and bounded CCR on data from a clinical trial of MS (Kosa et al., 2016). These data consist of a longitudinal cohort of 98 progressive MS patients during a pre-treatment period in the study. Various clinical scales quantify disease progression in patients, but they have varying degrees of sensitivity to detect disease progression and various levels of measurement error. Currently, testing efficacy of treatments in progressive MS, specifically in

Phase II trials, using EDSS requires hundreds of patients being followed for 2-3 years (Kosa et al., 2016). Given the clinical scales currently in practice, we wanted to combine them to create a new scale that is able to better capture smaller changes in disease progression over time. Since each scale has some level of imprecision (or “noise”) in measuring the disease process and measure, to some degree, overlapping biological functions, we would expect that combining them would decrease this level of imprecision in measuring the underlying disease severity. This methodology assumes that disease severity progresses linearly over a short amount of time in untreated patients with MS, an assumption that is supported by clinical observations (Kosa et al., 2016).

An example of the raw data is displayed in Figure 2.13. Each patient was measured 3 times at 6 month intervals leading up to the inception of the trial. The clinical outcomes that were available were: EDSS, Scripps neurological rating scale (SNRS, Sipe et al., 1984), 25 foot walk (25FW, Cutter et al., 1999), nine hole peg test (9HPT, Cutter et al., 1999) separated by dominant hand (DH-9HPT) and non-dominant hand (NDH-9HPT), paced auditory serial addition test (PASAT, Gronwall, 1977, Rao et al., 1991), and symbol digit modality test (SDMT, Smith, 1982). Patients were randomly split into training and validation datasets (approximately 2/3 in the training) and the following methods were performed on the training dataset: CCR, B-CCR, PCA (using all available time points), and an SEM as depicted in Figure 1. The estimated projection coefficients ($\hat{\mathbf{c}}$) for each method based on the training dataset were used to construct latent factors in the independent validation dataset, and the t -statistics for a linear time change (after accounting for patient-to-patient differences with a fixed-effect) in each validation dataset were recorded for each latent factor, as well as the original outcomes. This process was repeated 1,000 times to assess the variability in the constructed projections.

The t -statistics from the independent validation datasets are displayed in Figure 2.14. To the left of the vertical line are the t -statistics from the original clinical scales, and the projection t -statistics are to the right. While EDSS is the current FDA-approved endpoint for progressive MS clinical trials, SNRS achieved the best performance among the original set of clinical scales (achieving an average absolute t -statistic of 3.88). The CCR, B-CCR, and SEM (using Bartlett scores) achieved projections that provided larger t -statistics than SNRS in the majority of runs (larger t -statistics 79.6%, 87.4%, and 73.1% of the time by an average of 0.60, 0.72, and 0.27 units, respectively). Both CCR and B-CCR improved on SEM (using Bartlett scores) in the majority of runs (larger t -statistics 71.5% and 84.5% of the time by an average of 0.34 and 0.46 units, respectively). With the full dataset, the p -value from the CCR projection (using 9,999 permutations of the data) was 0.0001, providing substantial evidence that a real linear time change in the underlying disease process is being captured by the projection and the large t -statistic is not just due to the optimization process.

Figure 2.15 displays the projection coefficients for the CCR, B-CCR, SEM (using Bartlett factor scores), and PCA (using all data) from the different permutations of training data. There is observable agreement in the assigned weights for EDSS, SNRS, log2(25FW), and log2(NDH-9HPT), with the major differences in each projection seen in log2(DH-9HPT), PASAT, and SDMT. For PASAT and SDMT, lower scores are indicative of more disease severity. However, it was observed that the PASAT and SDMT scores were increasing in patients during this pre-treatment period due to a learning effect in taking the test. The CCR projection incorporates this by assigning positive coefficients for these scales. The B-CCR thresholds these coefficients around 0, eliminating the nonsensical solutions. The SEM and PCA assign negative coefficients to these scales, which have negative cross-sectional correlation

with EDSS, even though the direction of change is opposite of what we would expect in the disease course. Figure 2.16 displays the growth curves of the values of the scale obtained from the B-CCR projection using all of the data. These growth curves display smoother trends and increases in the majority of patients over time compared to the individual components displayed in Figure 2.13.

2.5 Conclusions

As noted in the motivation of the method, when variables' level of cross-sectional correlation does not match how the variables are changing over time, the CCR projection methods produce t -statistics that are larger than those from the SEM and PCA projections. The amount of performance gain depends on the strength of this cross-sectional correlation relative to the change and the amount of independent information present. If independent information is present (representing a different dimension of the disease process that is uncorrelated with other dimensions), the CCR has substantial performance gains, with the level of gain increasing as the level of change in these independent dimensions increase. In a motivating dataset from MS clinical trials, the CCR projections produced t -statistics that improved upon the industry standard by a considerable amount. Enforcing the projection coefficients to be bounded in the direction of disease progression allowed for more sensible solutions that track with the disease process with minimal impact on the constructed scales.

While this novel methodology shows promise, there are some aspects that could be improved. The current method does not have a way to automatically perform variable selection, which would have to be done in a *post-hoc* manner in the current implementation. For example, scales with small relative weights in B-CCR could be dropped with little impact on the results. This would be problematic if a large number of signals were to be considered, such as developing a scale from the individual pieces

of a standard neurological exam or from omics-based biomarkers. Extending the CCR methodology to include sparsity-imposing constraints would allow for this extension.

Maximizing a t -statistic allowed for detection of longitudinal change in a single group. However, it may be of interest to develop a clinical endpoint that best separates longitudinal progression rates between MS patients and healthy volunteers. Extending the methodology to maximize an F -statistic for a specified coefficient contrast would allow more complex model components, such as group differences with more than two groups, quadratic change over time, and/or interaction terms, to be targeted.

While SEMs can provide more information on the covariance structure within the data, they are not optimized to construct a projection that detects longitudinal change in the disease process, the goal of a clinical endpoint construction. When the underlying signals being measured are truly measuring the underlying factor (i.e., have some level of correlation and track with the disease process), the SEM performs as well as the CCR methods and is an appropriate choice for latent factor construction. As noted throughout the paper, the objective function of the CCR (for fixed signal-to-noise) decreases as the level of cross-sectional correlation increases. This is likely the reason for the observed lack-of-gain in the two-signal case with unequal changes. As the methodology provides more optimal projections when different dimensions of the disease process are available, we hypothesize that exploiting an *a priori* grouping structure of correlated variables to create approximately independent signals could further improve the CCR methodology.

In the clinical trial example, differences in patient starting values were accounted for using a fixed-effect for patient-to-patient variation. It could be argued that accounting for these differences through a mixed-model formulation would be a more appropriate approach. The t -statistic for a balanced two-level mixed model

for the repeated measures considered here is the same as our linear model result. Extending the CCR for targeting t -statistics in a mixed-model would allow for this type of accounting and would allow extension into multi-level modeling situations where accounting for multiple sources of variation such as different clinic locations or clinicians could be incorporated into the model.

In conclusion, the CCR methodology provides an alternative scale creation technique when an underlying change over time (over another specified domain) is assumed and of interest in the scientific application. A simulation study examining the properties of each projection method across a variety of datasets with varying degrees of longitudinal change, signal strength, and cross-sectional correlation was discussed, and the method applied to a motivating dataset of multiple sclerosis patients, where interest lies in constructing a better clinical endpoint to detect minor changes in disease progression, highlighted strengths and potential improvements to the CCR.

Acknowledgments

The authors thank John Borkowski, Laura Hildreth, Andy Hoegh, and Jing Qin for informative discussions regarding the manuscript. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>) using R and the following packages: `lavaan` (Rosseel, 2012), `semPlot` (Epskamp and Stuber, 2017), `beanplot` (Kampstra, 2008), `ggplot2` (Wickham, 2009), and `psych` (Revelle, 2018). The study was supported in part by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH). *Conflict of Interest:* None declared.

Algorithm 1: Bounded CCR Algorithm

Require: $\mathbf{B}$ and $\mathbf{M}$ matrices, $\boldsymbol{\gamma}$ column-vector.

```

function BOUNDED_CCR( $\mathbf{B}, \mathbf{M}, \boldsymbol{\gamma}, r_{11} = 0.1, r_{12} = 0.1, r_2 = 0.1, r_3 = 5, tol =$ 
 $1 \times 10^{-14}$ )
     $\Delta_a \leftarrow 0$ 
     $\Delta_b \leftarrow 0$ 
     $\mathbf{c}_0 \leftarrow \boldsymbol{\gamma}$ 
     $\mu \leftarrow 0$ 
     $\bar{\mathbf{c}} \leftarrow \mathbf{c}_0$ 
    while  $Error > tol$  do
         $\Delta_a \leftarrow \frac{(\Delta_a + \mathbf{B}\bar{\mathbf{c}})}{1+r_{11}}$  ▷ gradient-accumulators
         $\Delta_b \leftarrow \frac{(\Delta_b + \mu \mathbf{M}\bar{\mathbf{c}})}{1+r_{12}}$ 
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_k + r_{11}r_2\Delta_a - r_{12}r_2\mu\Delta_b$  ▷ gradient descent (primal)
         $\mathbf{c}_{k+1} \leftarrow \max(\mathbf{c}_{k+1}, 0)$  ▷ projection onto first orthant
         $\bar{\mathbf{c}} \leftarrow 2\mathbf{c}_{k+1} - \mathbf{c}_k$  ▷ extra gradient step
         $\mu \leftarrow \mu + \frac{r_3}{2} (\mathbf{c}_{k+1}^\top \mathbf{M} \mathbf{c}_{k+1} - 1)$  ▷ KKT-multiplier update
         $\mu \leftarrow \max(\mu, 0)$  ▷ KKT projection (dual feasibility)
         $Error \leftarrow \|\mathbf{c}_{k+1} - \mathbf{c}_k\|_\infty$ 
    end while
    return  $\mathbf{c}_{k+1}$ 
end function

```

Table 2.1: Case 1: Single variable changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	5.03 (1.08)	4.86 (1.07)	4.94 (1.08)	4.28 (1.72)	4.39 (1.75)	2.50 (1.40)	2.59 (1.16)
		0.30	4.91 (1.01)	4.74 (1.04)	4.82 (1.03)	2.56 (1.08)	2.58 (1.18)	2.51 (1.07)	2.49 (1.08)
		0.60	4.98 (0.98)	4.80 (1.00)	4.88 (0.99)	2.61 (1.07)	2.60 (1.11)	2.60 (1.09)	2.60 (1.07)
	$\rho=0.4$	0.00	5.03 (1.02)	5.58 (1.09)	5.02 (1.02)	2.13 (1.20)	2.31 (1.47)	1.97 (1.13)	1.96 (1.11)
		0.30	5.00 (1.02)	5.47 (1.02)	5.00 (1.02)	2.04 (1.25)	2.04 (1.27)	2.06 (1.22)	2.05 (1.21)
		0.60	5.03 (1.02)	5.58 (1.04)	5.02 (1.02)	1.99 (1.10)	1.94 (1.11)	2.01 (1.10)	2.00 (1.10)
	$\rho=0.7$	0.00	5.08 (1.08)	7.58 (1.11)	5.08 (1.08)	1.71 (1.26)	1.60 (1.30)	1.76 (1.23)	1.72 (1.26)
		0.30	5.00 (1.03)	7.47 (1.08)	5.00 (1.03)	1.71 (1.19)	1.53 (1.23)	1.73 (1.18)	1.72 (1.19)
		0.60	5.10 (1.00)	7.65 (1.06)	5.10 (1.00)	1.72 (1.23)	1.50 (1.28)	1.73 (1.24)	1.73 (1.23)
	$\rho=0.1$	0.00	7.11 (1.13)	7.01 (1.13)	7.05 (1.14)	6.77 (1.39)	6.91 (1.37)	3.46 (1.63)	3.48 (1.29)
		0.30	7.11 (1.03)	7.00 (1.04)	7.05 (1.03)	3.76 (1.11)	3.92 (1.21)	3.62 (1.09)	3.58 (1.08)
		0.60	7.16 (1.03)	7.04 (1.03)	7.10 (1.03)	3.71 (0.98)	3.69 (1.01)	3.72 (0.98)	3.69 (0.98)
SNR=0.1	$\rho=0.4$	0.00	7.09 (1.08)	7.87 (1.05)	7.09 (1.08)	3.78 (1.49)	4.68 (2.30)	2.96 (1.10)	2.93 (1.08)
		0.30	7.01 (1.06)	7.88 (1.11)	7.01 (1.06)	2.89 (1.06)	2.89 (1.23)	2.89 (1.05)	2.85 (1.05)
		0.60	7.12 (1.05)	7.96 (1.08)	7.12 (1.05)	2.97 (1.01)	2.82 (1.09)	2.99 (1.00)	2.98 (1.01)
	$\rho=0.7$	0.00	6.99 (1.02)	10.71 (1.14)	6.99 (1.02)	2.34 (1.12)	2.10 (1.23)	2.41 (1.08)	2.38 (1.08)
		0.30	7.03 (1.04)	10.67 (1.11)	7.03 (1.04)	2.48 (1.02)	2.08 (1.14)	2.54 (1.02)	2.52 (1.02)
		0.60	7.07 (1.08)	10.79 (1.08)	7.07 (1.08)	2.46 (1.13)	1.98 (1.22)	2.49 (1.14)	2.48 (1.14)

Table 2.2: Case 1: All variables changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	5.03 (1.08)	7.86 (1.06)	7.86 (1.06)	7.94 (1.07)	7.89 (1.08)	7.14 (2.04)	7.99 (1.08)
		0.30	4.91 (1.01)	7.68 (1.07)	7.68 (1.07)	7.84 (1.05)	7.80 (1.06)	7.84 (1.05)	7.84 (1.05)
		0.60	4.98 (0.98)	7.77 (1.09)	7.77 (1.09)	7.92 (1.09)	7.90 (1.10)	7.92 (1.09)	7.92 (1.09)
	$\rho=0.4$	0.00	5.03 (1.02)	6.23 (1.00)	6.24 (1.00)	6.41 (1.00)	6.38 (1.00)	6.41 (1.00)	6.41 (1.00)
		0.30	5.00 (1.02)	6.32 (1.06)	6.32 (1.06)	6.51 (1.07)	6.50 (1.07)	6.51 (1.06)	6.51 (1.07)
		0.60	5.03 (1.02)	6.21 (1.00)	6.22 (0.99)	6.40 (0.99)	6.39 (0.99)	6.40 (0.99)	6.40 (0.99)
	$\rho=0.7$	0.00	5.08 (1.08)	5.44 (1.04)	5.50 (1.03)	5.65 (1.03)	5.65 (1.04)	5.65 (1.03)	5.65 (1.03)
		0.30	5.00 (1.03)	5.42 (1.04)	5.47 (1.02)	5.63 (0.99)	5.62 (0.99)	5.63 (0.99)	5.63 (0.99)
		0.60	5.10 (1.00)	5.42 (1.03)	5.48 (1.03)	5.63 (1.03)	5.63 (1.03)	5.63 (1.03)	5.63 (1.03)
	$\rho=0.1$	0.00	7.11 (1.13)	11.13 (1.11)	11.13 (1.11)	11.21 (1.11)	11.16 (1.11)	10.07 (2.46)	11.24 (1.11)
		0.30	7.11 (1.03)	11.09 (1.11)	11.09 (1.11)	11.21 (1.12)	11.17 (1.12)	11.20 (1.13)	11.21 (1.12)
		0.60	7.16 (1.03)	11.11 (1.07)	11.11 (1.07)	11.24 (1.06)	11.21 (1.06)	11.24 (1.06)	11.24 (1.06)
SNR=0.1	$\rho=0.4$	0.00	7.09 (1.08)	9.05 (1.14)	9.05 (1.14)	9.17 (1.14)	9.15 (1.14)	9.17 (1.14)	9.18 (1.14)
		0.30	7.01 (1.06)	8.94 (1.10)	8.94 (1.10)	9.07 (1.09)	9.05 (1.09)	9.07 (1.09)	9.07 (1.09)
		0.60	7.12 (1.05)	9.01 (1.06)	9.01 (1.06)	9.16 (1.06)	9.15 (1.06)	9.16 (1.06)	9.16 (1.06)
	$\rho=0.7$	0.00	6.99 (1.02)	7.68 (1.07)	7.69 (1.07)	7.83 (1.05)	7.82 (1.05)	7.83 (1.05)	7.83 (1.05)
		0.30	7.03 (1.04)	7.73 (1.07)	7.75 (1.06)	7.88 (1.05)	7.88 (1.05)	7.88 (1.05)	7.88 (1.05)
		0.60	7.07 (1.08)	7.72 (1.11)	7.74 (1.10)	7.88 (1.10)	7.88 (1.10)	7.88 (1.10)	7.88 (1.10)

Table 2.3: Case 1: Variables changing differently. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	7.10 (1.10)	7.60 (1.11)	7.60 (1.11)	7.35 (1.09)	7.42 (1.11)	5.03 (1.89)	5.90 (1.15)
		0.30	6.99 (1.03)	7.43 (1.07)	7.43 (1.06)	5.71 (1.04)	5.94 (1.07)	5.45 (1.06)	5.51 (1.04)
		0.60	7.05 (1.00)	7.54 (1.03)	7.53 (1.04)	5.59 (1.07)	5.70 (1.10)	5.56 (1.07)	5.57 (1.07)
	$\rho=0.4$	0.00	7.11 (1.04)	7.79 (1.05)	7.08 (1.04)	5.00 (1.05)	5.66 (1.23)	4.42 (1.01)	4.47 (0.99)
		0.30	7.08 (1.04)	7.76 (1.05)	7.07 (1.05)	4.66 (1.05)	4.92 (1.10)	4.55 (1.05)	4.56 (1.05)
		0.60	7.09 (1.04)	7.78 (1.06)	7.07 (1.06)	4.49 (0.97)	4.60 (1.00)	4.47 (0.96)	4.47 (0.96)
	$\rho=0.7$	0.00	7.15 (1.10)	9.92 (1.16)	7.15 (1.10)	4.02 (1.02)	4.34 (1.06)	3.94 (1.02)	3.95 (1.02)
		0.30	7.08 (1.06)	9.76 (1.13)	7.08 (1.06)	3.96 (0.98)	4.15 (0.99)	3.93 (0.98)	3.93 (0.98)
		0.60	7.18 (1.02)	9.91 (1.09)	7.17 (1.03)	3.97 (1.01)	4.05 (1.03)	3.96 (1.01)	3.96 (1.01)
	$\rho=0.1$	0.00	10.04 (1.17)	10.78 (1.17)	10.75 (1.17)	10.49 (1.17)	10.62 (1.17)	6.96 (2.43)	8.48 (1.26)
		0.30	10.05 (1.06)	10.76 (1.08)	10.73 (1.08)	8.40 (1.13)	9.01 (1.21)	7.70 (1.10)	7.85 (1.08)
		0.60	10.10 (1.07)	10.81 (1.09)	10.78 (1.09)	7.94 (1.02)	8.24 (1.07)	7.83 (1.02)	7.86 (1.02)
SNR=0.1	$\rho=0.4$	0.00	10.02 (1.11)	11.08 (1.10)	10.03 (1.13)	7.81 (1.18)	9.34 (1.25)	6.27 (1.13)	6.40 (1.12)
		0.30	9.94 (1.09)	11.06 (1.14)	9.95 (1.10)	6.54 (1.08)	7.25 (1.21)	6.22 (1.06)	6.26 (1.05)
		0.60	10.06 (1.09)	11.14 (1.06)	10.07 (1.09)	6.44 (1.03)	6.77 (1.07)	6.37 (1.03)	6.38 (1.03)
	$\rho=0.7$	0.00	9.92 (1.07)	13.91 (1.19)	9.92 (1.07)	5.55 (1.03)	6.36 (1.28)	5.31 (1.02)	5.34 (1.02)
		0.30	9.96 (1.08)	13.88 (1.13)	9.96 (1.08)	5.51 (1.03)	5.90 (1.08)	5.42 (1.04)	5.44 (1.03)
		0.60	10.02 (1.12)	14.03 (1.16)	10.02 (1.12)	5.49 (1.07)	5.72 (1.09)	5.46 (1.07)	5.47 (1.07)

Table 2.4: Case 2: Single signal changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	4.99 (1.06)	7.50 (1.09)	7.61 (1.09)	7.66 (1.08)	7.65 (1.09)	4.50 (2.43)	6.67 (1.78)
		0.30	5.06 (1.03)	7.68 (1.11)	7.77 (1.09)	5.77 (1.02)	5.91 (1.06)	5.58 (1.02)	5.68 (1.01)
		0.60	5.03 (1.08)	7.60 (1.10)	7.71 (1.09)	5.65 (1.05)	5.72 (1.08)	5.62 (1.06)	5.64 (1.05)
	$\rho=0.4$	0.00	5.03 (1.03)	6.09 (1.07)	6.24 (1.06)	5.66 (2.35)	5.64 (2.40)	4.26 (1.94)	5.13 (1.69)
		0.30	5.02 (1.02)	6.10 (1.09)	6.23 (1.07)	4.85 (1.04)	5.14 (1.15)	4.54 (1.04)	4.60 (1.03)
		0.60	4.96 (1.02)	6.02 (1.11)	6.17 (1.11)	4.57 (1.09)	4.65 (1.12)	4.54 (1.08)	4.55 (1.09)
	$\rho=0.7$	0.00	4.95 (0.98)	5.12 (1.05)	5.34 (1.02)	2.90 (2.87)	2.85 (2.91)	3.74 (1.47)	4.26 (1.27)
		0.30	5.03 (1.07)	5.12 (1.10)	5.37 (1.08)	4.51 (1.63)	4.52 (2.40)	3.93 (1.01)	3.95 (1.01)
		0.60	5.01 (1.03)	5.13 (1.01)	5.38 (1.02)	4.22 (1.06)	4.47 (2.04)	3.97 (0.98)	3.98 (0.98)
	$\rho=0.1$	0.00	7.09 (1.03)	10.86 (1.12)	10.96 (1.12)	11.02 (1.11)	11.01 (1.11)	6.28 (3.36)	10.45 (1.43)
		0.30	7.15 (1.04)	10.94 (1.12)	11.04 (1.12)	8.35 (1.09)	8.79 (1.16)	7.80 (1.12)	8.07 (1.08)
		0.60	7.06 (1.02)	10.78 (1.10)	10.88 (1.10)	7.78 (1.13)	8.02 (1.16)	7.69 (1.13)	7.74 (1.13)
SNR=0.1	$\rho=0.4$	0.00	7.07 (1.06)	8.78 (1.09)	8.91 (1.10)	9.05 (1.17)	9.07 (1.18)	5.75 (2.64)	8.06 (1.54)
		0.30	7.09 (1.02)	8.80 (1.09)	8.92 (1.08)	7.11 (1.11)	7.79 (1.28)	6.34 (1.05)	6.48 (1.02)
		0.60	7.09 (1.09)	8.83 (1.18)	8.95 (1.20)	6.48 (1.15)	6.76 (1.27)	6.39 (1.14)	6.42 (1.14)
	$\rho=0.7$	0.00	7.10 (1.10)	7.55 (1.10)	7.71 (1.11)	6.95 (2.74)	6.94 (2.79)	5.19 (1.94)	6.32 (1.47)
		0.30	7.08 (1.01)	7.56 (1.06)	7.73 (1.04)	7.02 (1.45)	7.44 (2.08)	5.55 (0.94)	5.60 (0.93)
		0.60	7.08 (1.05)	7.50 (1.10)	7.67 (1.05)	6.08 (1.07)	7.13 (2.01)	5.50 (0.99)	5.51 (0.99)

Table 2.5: Case 2: Both signals changing. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	4.99 (1.06)	10.86 (1.11)	10.86 (1.11)	11.05 (1.09)	10.99 (1.09)	6.49 (3.59)	10.97 (1.11)
		0.30	5.06 (1.03)	10.98 (1.10)	10.98 (1.10)	11.25 (1.06)	11.21 (1.07)	11.21 (1.06)	11.25 (1.06)
		0.60	5.03 (1.08)	10.94 (1.14)	10.94 (1.14)	11.26 (1.13)	11.24 (1.13)	11.26 (1.13)	11.27 (1.13)
	$\rho=0.4$	0.00	5.03 (1.03)	8.77 (1.10)	8.78 (1.09)	7.94 (1.19)	7.44 (1.17)	5.43 (3.61)	8.78 (1.55)
		0.30	5.02 (1.02)	8.84 (1.10)	8.85 (1.10)	9.17 (1.08)	9.09 (1.09)	9.17 (1.09)	9.18 (1.08)
		0.60	4.96 (1.02)	8.80 (1.16)	8.81 (1.16)	9.14 (1.13)	9.12 (1.13)	9.14 (1.13)	9.14 (1.13)
	$\rho=0.7$	0.00	4.95 (0.98)	7.48 (1.07)	7.55 (1.04)	6.11 (1.12)	5.69 (1.06)	4.30 (3.55)	7.42 (1.76)
		0.30	5.03 (1.07)	7.52 (1.06)	7.61 (1.05)	7.55 (1.06)	6.33 (1.06)	7.94 (1.05)	7.94 (1.05)
		0.60	5.01 (1.03)	7.56 (1.01)	7.65 (1.00)	7.91 (1.02)	6.98 (1.05)	7.95 (1.01)	7.95 (1.01)
	$\rho=0.1$	0.00	7.09 (1.03)	15.49 (1.20)	15.49 (1.20)	15.69 (1.20)	15.63 (1.20)	9.16 (4.87)	15.69 (1.20)
		0.30	7.15 (1.04)	15.63 (1.23)	15.63 (1.23)	15.90 (1.21)	15.85 (1.22)	15.85 (1.21)	15.90 (1.21)
		0.60	7.06 (1.02)	15.44 (1.25)	15.44 (1.25)	15.72 (1.25)	15.68 (1.25)	15.71 (1.25)	15.72 (1.25)
SNR=0.1	$\rho=0.4$	0.00	7.07 (1.06)	12.59 (1.17)	12.59 (1.17)	12.40 (1.23)	11.90 (1.36)	7.76 (4.83)	12.86 (1.19)
		0.30	7.09 (1.02)	12.64 (1.15)	12.64 (1.15)	12.94 (1.13)	12.86 (1.14)	12.94 (1.13)	12.95 (1.13)
		0.60	7.09 (1.09)	12.64 (1.26)	12.65 (1.26)	12.95 (1.26)	12.92 (1.26)	12.95 (1.26)	12.95 (1.26)
	$\rho=0.7$	0.00	7.10 (1.10)	10.88 (1.21)	10.91 (1.21)	9.36 (1.19)	8.36 (1.12)	6.27 (4.78)	11.13 (1.33)
		0.30	7.08 (1.01)	10.96 (1.06)	11.00 (1.05)	10.78 (1.06)	9.03 (1.08)	11.26 (1.02)	11.26 (1.02)
		0.60	7.08 (1.05)	10.83 (1.08)	10.87 (1.07)	11.10 (1.07)	9.91 (1.15)	11.15 (1.07)	11.15 (1.07)

Table 2.6: Case 2: Signals changing differently. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	8.66 (1.10)	14.10 (1.15)	14.12 (1.15)	14.28 (1.14)	14.25 (1.15)	8.30 (4.06)	13.97 (1.19)
		0.30	8.73 (1.06)	14.28 (1.16)	14.29 (1.16)	13.25 (1.11)	13.60 (1.15)	12.77 (1.11)	13.03 (1.10)
		0.60	8.70 (1.12)	14.23 (1.20)	14.24 (1.20)	12.98 (1.16)	13.19 (1.20)	12.89 (1.16)	12.94 (1.16)
	$\rho=0.4$	0.00	8.72 (1.09)	11.53 (1.13)	11.57 (1.13)	11.73 (1.13)	11.58 (1.13)	7.47 (3.56)	11.23 (1.21)
		0.30	8.69 (1.05)	11.54 (1.15)	11.57 (1.14)	11.04 (1.12)	11.51 (1.14)	10.43 (1.12)	10.58 (1.11)
		0.60	8.62 (1.06)	11.46 (1.20)	11.50 (1.19)	10.54 (1.15)	10.80 (1.19)	10.46 (1.15)	10.48 (1.15)
	$\rho=0.7$	0.00	8.62 (1.02)	9.87 (1.09)	9.94 (1.07)	9.88 (1.24)	9.68 (1.25)	6.42 (2.92)	9.25 (1.37)
		0.30	8.69 (1.11)	9.88 (1.13)	9.97 (1.13)	10.07 (1.18)	9.93 (1.34)	9.03 (1.07)	9.09 (1.07)
		0.60	8.69 (1.07)	9.92 (1.05)	10.01 (1.05)	9.52 (1.09)	9.79 (1.67)	9.10 (1.03)	9.11 (1.03)
	$\rho=0.1$	0.00	12.28 (1.11)	20.11 (1.31)	20.11 (1.31)	20.22 (1.31)	20.28 (1.30)	11.33 (5.98)	19.72 (1.35)
		0.30	12.35 (1.13)	20.22 (1.31)	20.22 (1.31)	18.92 (1.30)	19.77 (1.34)	17.80 (1.34)	18.41 (1.28)
		0.60	12.25 (1.12)	20.05 (1.31)	20.05 (1.31)	18.11 (1.29)	18.74 (1.34)	17.87 (1.29)	18.00 (1.29)
SNR=0.1	$\rho=0.4$	0.00	12.25 (1.14)	16.37 (1.25)	16.40 (1.25)	16.59 (1.26)	16.37 (1.25)	10.08 (5.11)	15.81 (1.33)
		0.30	12.28 (1.10)	16.42 (1.24)	16.44 (1.24)	15.75 (1.22)	16.56 (1.24)	14.54 (1.22)	14.86 (1.18)
		0.60	12.28 (1.18)	16.43 (1.36)	16.45 (1.37)	14.95 (1.31)	15.71 (1.43)	14.70 (1.30)	14.77 (1.29)
	$\rho=0.7$	0.00	12.29 (1.19)	14.15 (1.24)	14.21 (1.23)	14.29 (1.20)	13.89 (1.18)	8.73 (4.32)	13.23 (1.29)
		0.30	12.28 (1.10)	14.23 (1.16)	14.28 (1.15)	14.31 (1.11)	14.17 (1.15)	12.65 (1.07)	12.78 (1.05)
		0.60	12.26 (1.12)	14.09 (1.16)	14.15 (1.15)	13.48 (1.11)	14.27 (1.19)	12.67 (1.10)	12.70 (1.10)

Table 2.7: Case 3: Low independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	4.99 (1.06)	11.42 (1.11)	11.42 (1.11)	11.47 (1.11)	11.41 (1.12)	5.82 (3.18)	10.82 (1.18)
		0.30	5.06 (1.03)	11.53 (1.10)	11.53 (1.10)	11.73 (1.08)	11.67 (1.09)	11.65 (1.09)	11.70 (1.09)
		0.60	5.03 (1.08)	11.49 (1.14)	11.49 (1.14)	11.76 (1.12)	11.71 (1.12)	11.75 (1.12)	11.75 (1.12)
	$\rho=0.4$	0.00	5.03 (1.03)	10.48 (1.12)	10.49 (1.12)	7.69 (1.16)	7.18 (1.13)	6.28 (1.50)	7.73 (1.19)
		0.30	5.02 (1.02)	10.53 (1.12)	10.53 (1.12)	9.88 (1.08)	9.13 (1.10)	9.80 (1.09)	9.92 (1.07)
		0.60	4.96 (1.02)	10.49 (1.19)	10.50 (1.19)	10.25 (1.13)	9.76 (1.13)	10.22 (1.13)	10.23 (1.13)
	$\rho=0.7$	0.00	4.95 (0.98)	9.94 (1.09)	9.98 (1.08)	6.10 (1.04)	5.73 (1.01)	5.50 (1.22)	6.28 (1.07)
		0.30	5.03 (1.07)	9.99 (1.07)	10.04 (1.06)	7.86 (1.06)	6.28 (1.03)	8.35 (1.08)	8.50 (1.06)
		0.60	5.01 (1.03)	10.04 (1.07)	10.08 (1.07)	8.84 (1.08)	6.64 (1.05)	9.14 (1.08)	9.16 (1.08)
	$\rho=0.1$	0.00	7.09 (1.03)	16.26 (1.22)	16.26 (1.22)	16.34 (1.23)	16.27 (1.23)	8.24 (4.43)	16.01 (1.22)
		0.30	7.15 (1.04)	16.40 (1.25)	16.40 (1.25)	16.58 (1.21)	16.50 (1.22)	16.47 (1.21)	16.54 (1.21)
		0.60	7.06 (1.02)	16.22 (1.26)	16.22 (1.26)	16.42 (1.25)	16.36 (1.25)	16.40 (1.25)	16.41 (1.25)
SNR=0.1	$\rho=0.4$	0.00	7.07 (1.06)	14.97 (1.21)	14.97 (1.21)	12.16 (1.26)	11.14 (1.27)	9.04 (1.73)	12.16 (1.21)
		0.30	7.09 (1.02)	15.01 (1.20)	15.01 (1.20)	14.08 (1.22)	13.10 (1.29)	13.85 (1.25)	14.10 (1.21)
		0.60	7.09 (1.09)	15.02 (1.32)	15.02 (1.32)	14.51 (1.28)	13.80 (1.28)	14.47 (1.28)	14.48 (1.28)
	$\rho=0.7$	0.00	7.10 (1.10)	14.29 (1.25)	14.32 (1.26)	9.19 (1.08)	8.24 (1.04)	7.83 (1.30)	9.70 (1.12)
		0.30	7.08 (1.01)	14.40 (1.15)	14.43 (1.14)	11.29 (1.09)	8.94 (1.04)	11.79 (1.15)	12.17 (1.09)
		0.60	7.08 (1.05)	14.26 (1.14)	14.29 (1.13)	12.44 (1.13)	9.36 (1.09)	12.81 (1.13)	12.86 (1.13)

Table 2.8: Case 3: Moderate independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	7.07 (1.08)	13.15 (1.17)	13.16 (1.17)	13.18 (1.15)	13.20 (1.16)	5.10 (3.31)	11.70 (1.42)
		0.30	7.14 (1.05)	13.29 (1.09)	13.30 (1.09)	12.50 (1.10)	12.58 (1.13)	12.10 (1.13)	12.29 (1.10)
		0.60	7.11 (1.10)	13.25 (1.20)	13.26 (1.20)	12.44 (1.14)	12.45 (1.15)	12.37 (1.14)	12.39 (1.14)
	$\rho=0.4$	0.00	7.12 (1.06)	12.78 (1.16)	12.81 (1.16)	7.27 (2.27)	6.58 (2.46)	4.42 (1.78)	6.66 (1.34)
		0.30	7.10 (1.03)	12.84 (1.17)	12.87 (1.17)	10.21 (1.13)	9.31 (1.29)	9.73 (1.15)	10.04 (1.08)
		0.60	7.03 (1.04)	12.77 (1.23)	12.80 (1.22)	10.72 (1.13)	9.95 (1.19)	10.62 (1.13)	10.65 (1.13)
	$\rho=0.7$	0.00	7.03 (1.00)	12.53 (1.15)	12.58 (1.13)	4.67 (1.06)	4.15 (1.00)	3.83 (1.43)	4.99 (1.13)
		0.30	7.10 (1.09)	12.56 (1.12)	12.63 (1.11)	7.22 (1.08)	4.91 (1.02)	7.98 (1.13)	8.30 (1.07)
		0.60	7.10 (1.05)	12.63 (1.11)	12.69 (1.11)	8.86 (1.10)	5.46 (1.05)	9.37 (1.10)	9.41 (1.09)
	$\rho=0.1$	0.00	10.03 (1.07)	18.74 (1.26)	18.74 (1.26)	18.70 (1.27)	18.81 (1.26)	7.16 (4.62)	17.61 (1.31)
		0.30	10.09 (1.09)	18.86 (1.35)	18.86 (1.35)	17.77 (1.26)	18.11 (1.31)	17.00 (1.30)	17.40 (1.23)
		0.60	10.00 (1.07)	18.66 (1.31)	18.67 (1.31)	17.37 (1.27)	17.54 (1.29)	17.19 (1.27)	17.27 (1.27)
SNR=0.1	$\rho=0.4$	0.00	10.00 (1.10)	18.19 (1.28)	18.21 (1.28)	16.49 (2.06)	16.72 (2.46)	6.46 (2.13)	11.75 (1.36)
		0.30	10.02 (1.06)	18.24 (1.26)	18.25 (1.25)	14.91 (1.29)	14.36 (1.57)	13.62 (1.37)	14.35 (1.24)
		0.60	10.03 (1.14)	18.25 (1.39)	18.26 (1.39)	15.17 (1.30)	14.35 (1.39)	14.93 (1.30)	15.03 (1.29)
	$\rho=0.7$	0.00	10.03 (1.14)	17.92 (1.33)	17.97 (1.32)	7.38 (1.11)	6.09 (1.02)	5.54 (1.45)	8.35 (1.22)
		0.30	10.02 (1.06)	18.03 (1.26)	18.08 (1.26)	10.41 (1.13)	7.04 (1.02)	11.13 (1.26)	11.93 (1.12)
		0.60	10.01 (1.08)	17.89 (1.21)	17.94 (1.20)	12.40 (1.15)	7.70 (1.09)	13.04 (1.16)	13.17 (1.15)

Table 2.9: Case 3: High independent change. Means and standard deviations of the validation test statistics for selected projection techniques and the original variable with the strongest change.

		ICC	Original	CCR	B-CCR	SEM (Loadings)	SEM (Bartlett)	PCA (single)	PCA (all)
SNR=0.05	$\rho=0.1$	0.00	8.66 (1.10)	15.37 (1.21)	15.38 (1.21)	15.40 (1.20)	15.45 (1.20)	5.16 (3.63)	13.74 (1.59)
		0.30	8.73 (1.06)	15.52 (1.11)	15.53 (1.11)	13.78 (1.14)	14.13 (1.18)	12.94 (1.19)	13.35 (1.13)
		0.60	8.70 (1.12)	15.48 (1.25)	15.49 (1.24)	13.52 (1.16)	13.68 (1.20)	13.36 (1.16)	13.41 (1.16)
	$\rho=0.4$	0.00	8.72 (1.09)	15.19 (1.21)	15.23 (1.21)	12.58 (4.00)	12.45 (4.49)	3.57 (2.03)	6.58 (1.61)
		0.30	8.69 (1.05)	15.25 (1.22)	15.29 (1.21)	11.21 (1.20)	10.54 (1.55)	10.15 (1.21)	10.68 (1.10)
		0.60	8.62 (1.06)	15.16 (1.27)	15.21 (1.26)	11.58 (1.15)	10.75 (1.26)	11.38 (1.15)	11.45 (1.14)
	$\rho=0.7$	0.00	8.62 (1.02)	15.01 (1.20)	15.09 (1.19)	4.08 (1.13)	3.46 (1.04)	3.08 (1.56)	4.54 (1.26)
		0.30	8.69 (1.11)	15.04 (1.17)	15.12 (1.16)	7.17 (1.10)	4.36 (1.02)	8.14 (1.19)	8.63 (1.10)
		0.60	8.69 (1.07)	15.11 (1.17)	15.19 (1.17)	9.28 (1.12)	5.04 (1.06)	9.98 (1.12)	10.05 (1.11)
	$\rho=0.1$	0.00	12.28 (1.11)	21.90 (1.32)	21.90 (1.32)	21.78 (1.34)	22.00 (1.32)	7.17 (5.01)	20.35 (1.45)
		0.30	12.35 (1.13)	21.99 (1.45)	22.00 (1.45)	19.73 (1.33)	20.74 (1.42)	18.03 (1.41)	18.90 (1.28)
		0.60	12.25 (1.12)	21.80 (1.39)	21.80 (1.38)	18.86 (1.31)	19.50 (1.37)	18.47 (1.31)	18.66 (1.30)
SNR=0.1	$\rho=0.4$	0.00	12.25 (1.14)	21.57 (1.37)	21.61 (1.37)	20.63 (1.57)	21.25 (1.48)	5.33 (2.39)	12.78 (1.54)
		0.30	12.28 (1.10)	21.62 (1.34)	21.65 (1.34)	16.78 (1.38)	17.38 (1.78)	14.05 (1.48)	15.28 (1.27)
		0.60	12.28 (1.18)	21.63 (1.47)	21.66 (1.47)	16.40 (1.34)	15.91 (1.55)	15.90 (1.33)	16.09 (1.32)
	$\rho=0.7$	0.00	12.29 (1.19)	21.41 (1.40)	21.48 (1.40)	6.66 (1.16)	5.17 (1.02)	4.53 (1.58)	8.13 (1.36)
		0.30	12.28 (1.10)	21.53 (1.37)	21.60 (1.36)	10.28 (1.19)	6.27 (1.02)	11.22 (1.37)	12.41 (1.16)
		0.60	12.26 (1.12)	21.38 (1.30)	21.45 (1.28)	12.89 (1.18)	7.07 (1.10)	13.78 (1.20)	13.99 (1.17)

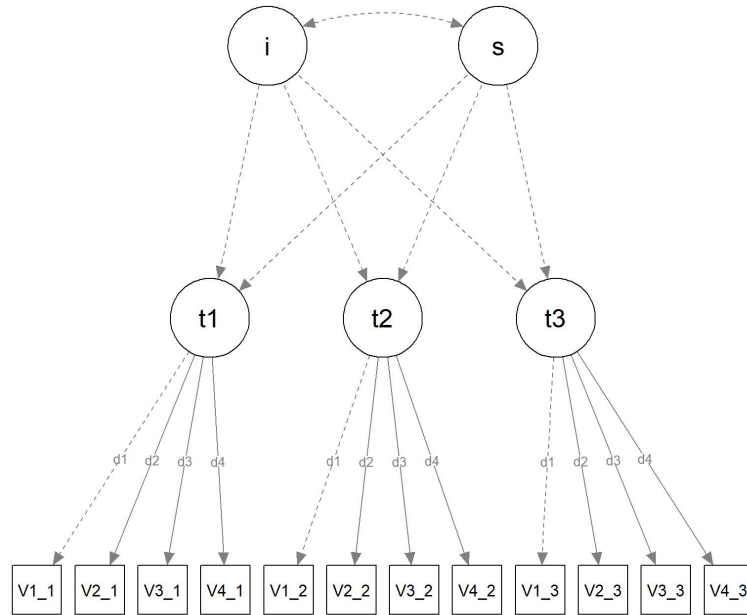


Figure 2.1: Example path diagram of a multilevel growth curve model with $P=4$ variables measured at $T=3$ unique time points. For identifiability, the loading on the first manifest variable is set to one.

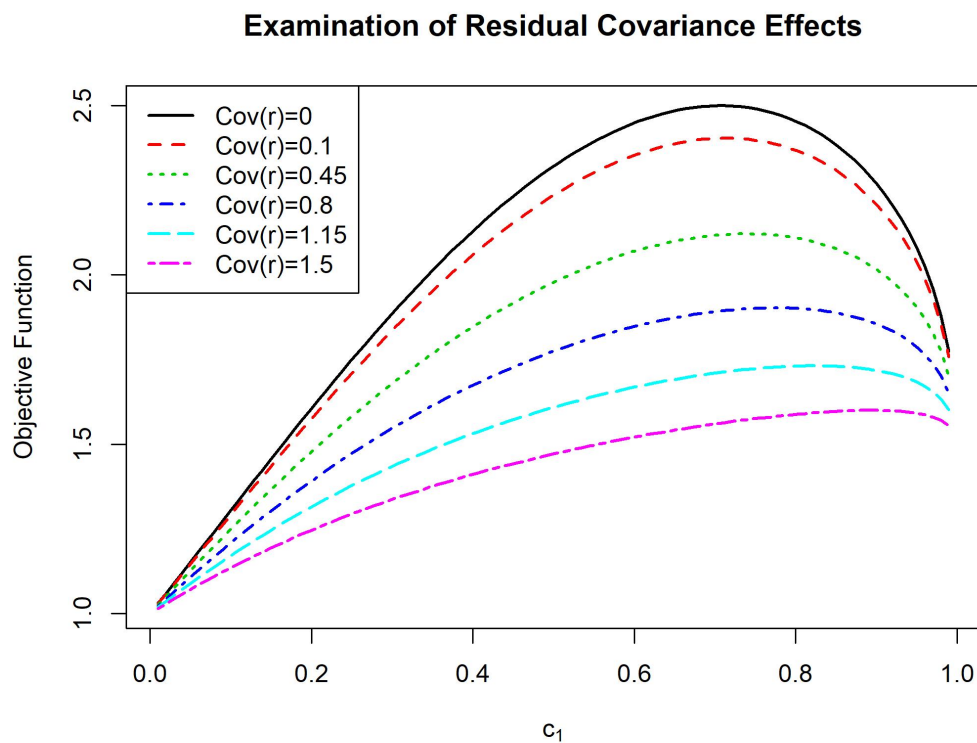


Figure 2.2: Profile of objective function for two variables with equivalent SNR across varying levels of residual covariance. Optimal solutions become less clearly defined (i.e., flatter surface) as residual covariance of two variables increases.

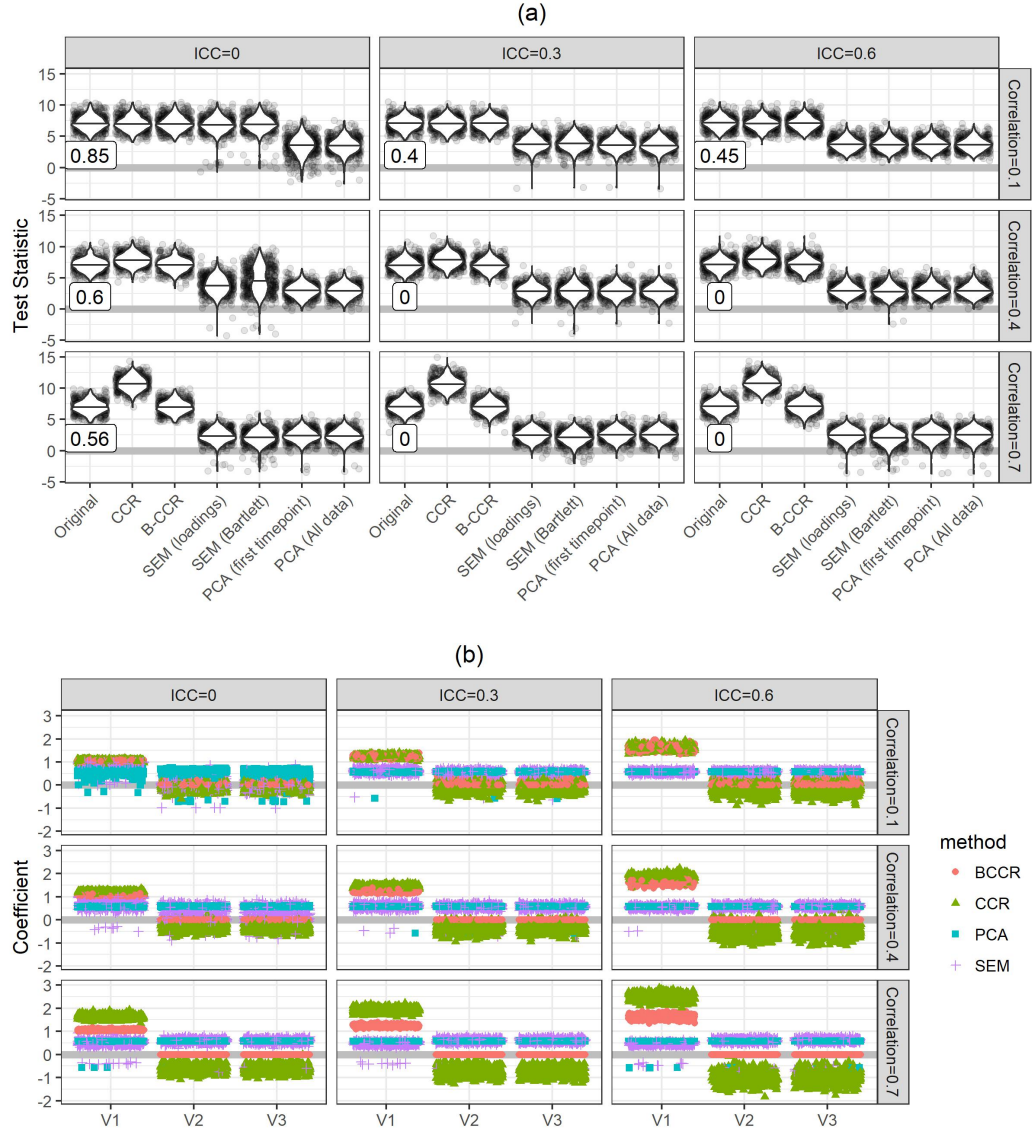


Figure 2.3: Simulation results for case 1: one variable changing. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

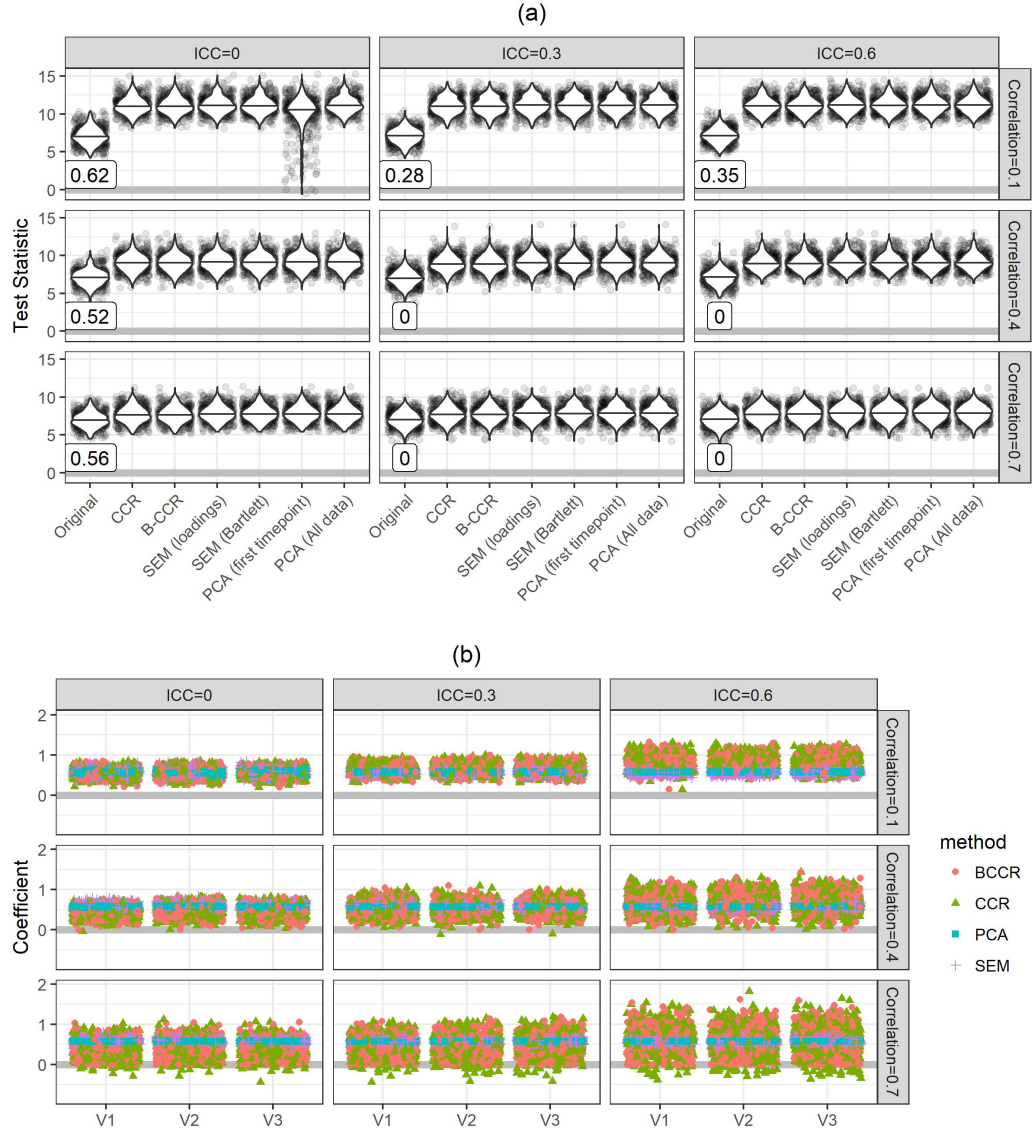


Figure 2.4: Simulation results for case 1: all variables changing equally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

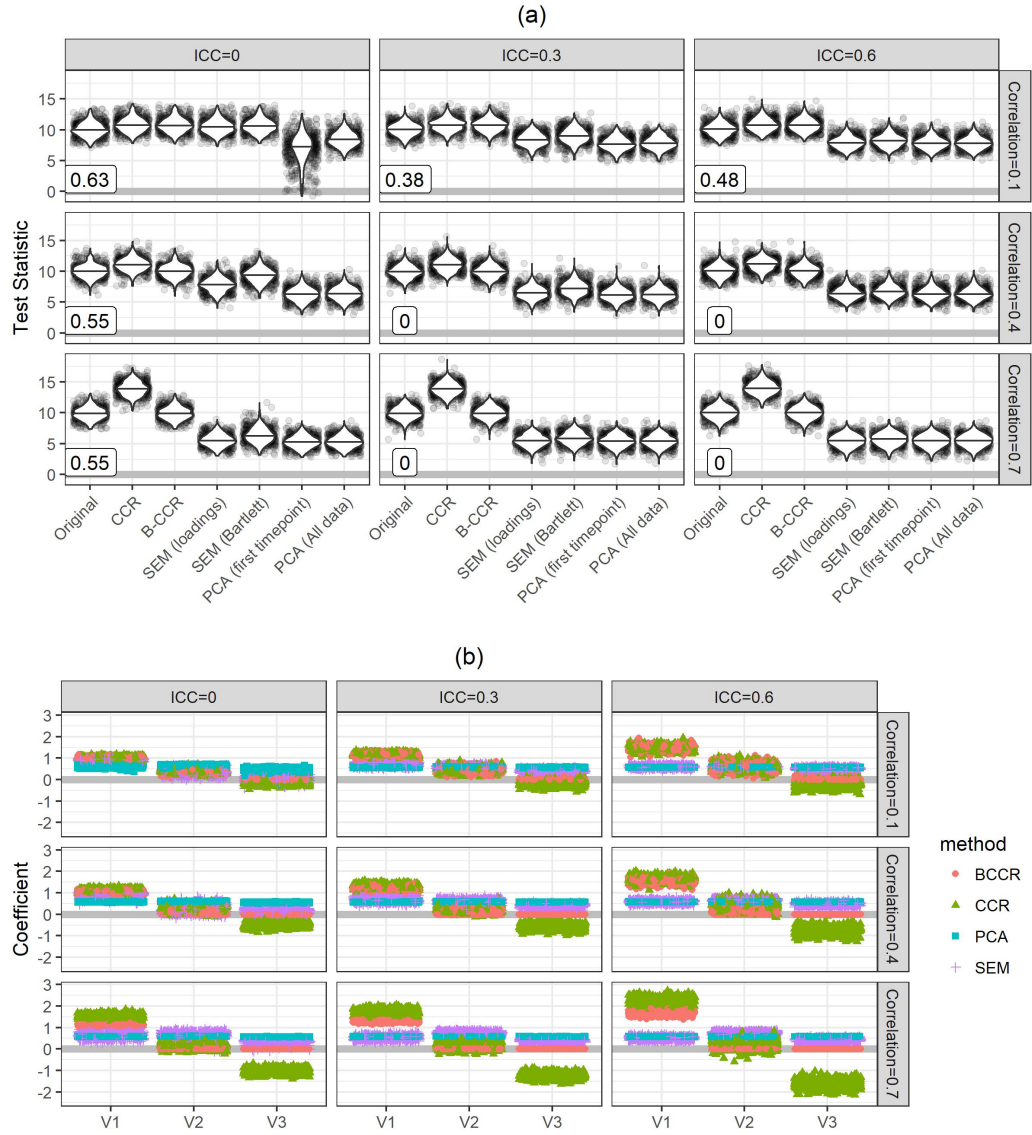


Figure 2.5: Simulation results for case 1: variables changing unequally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

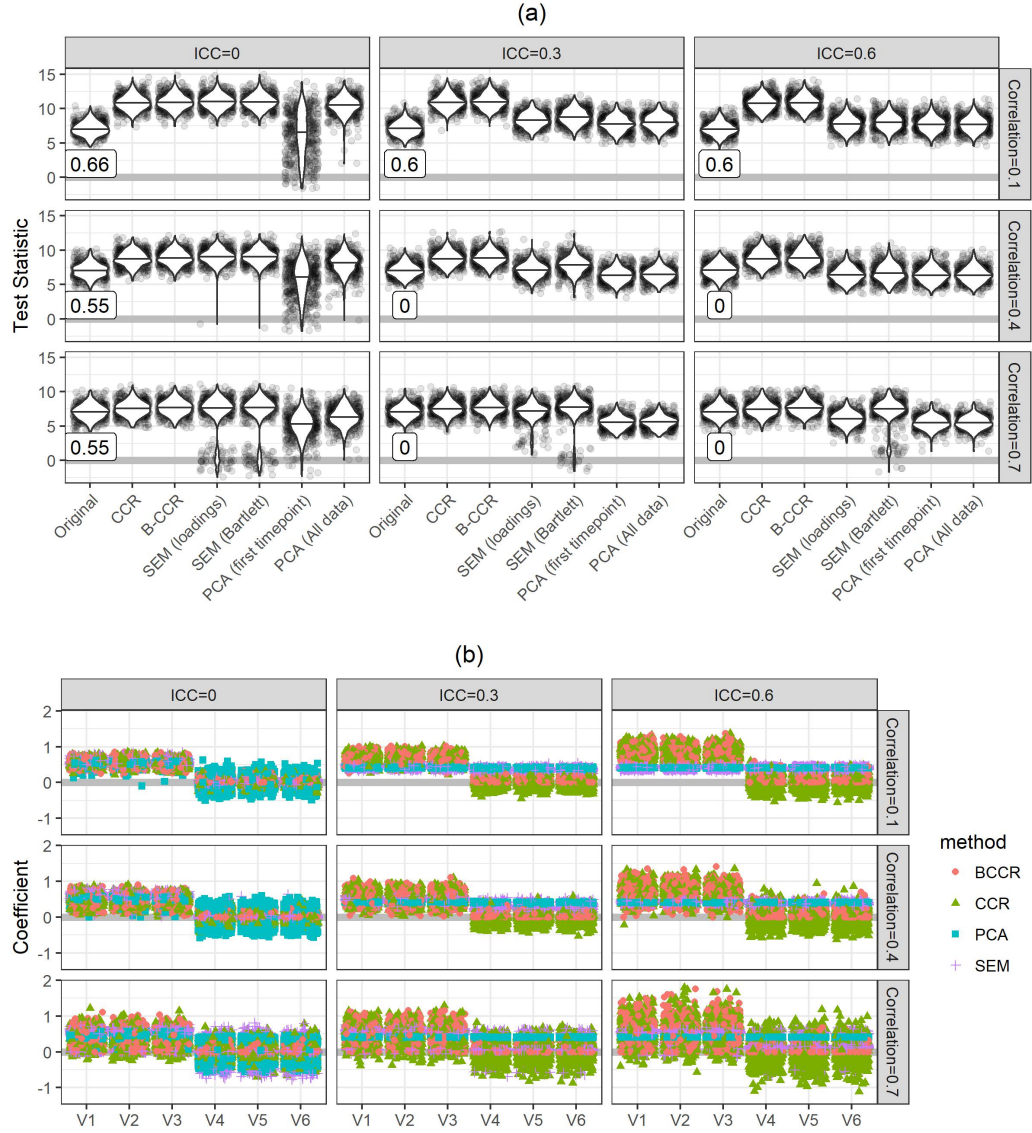


Figure 2.6: Simulation results for case 2: one signal changing. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

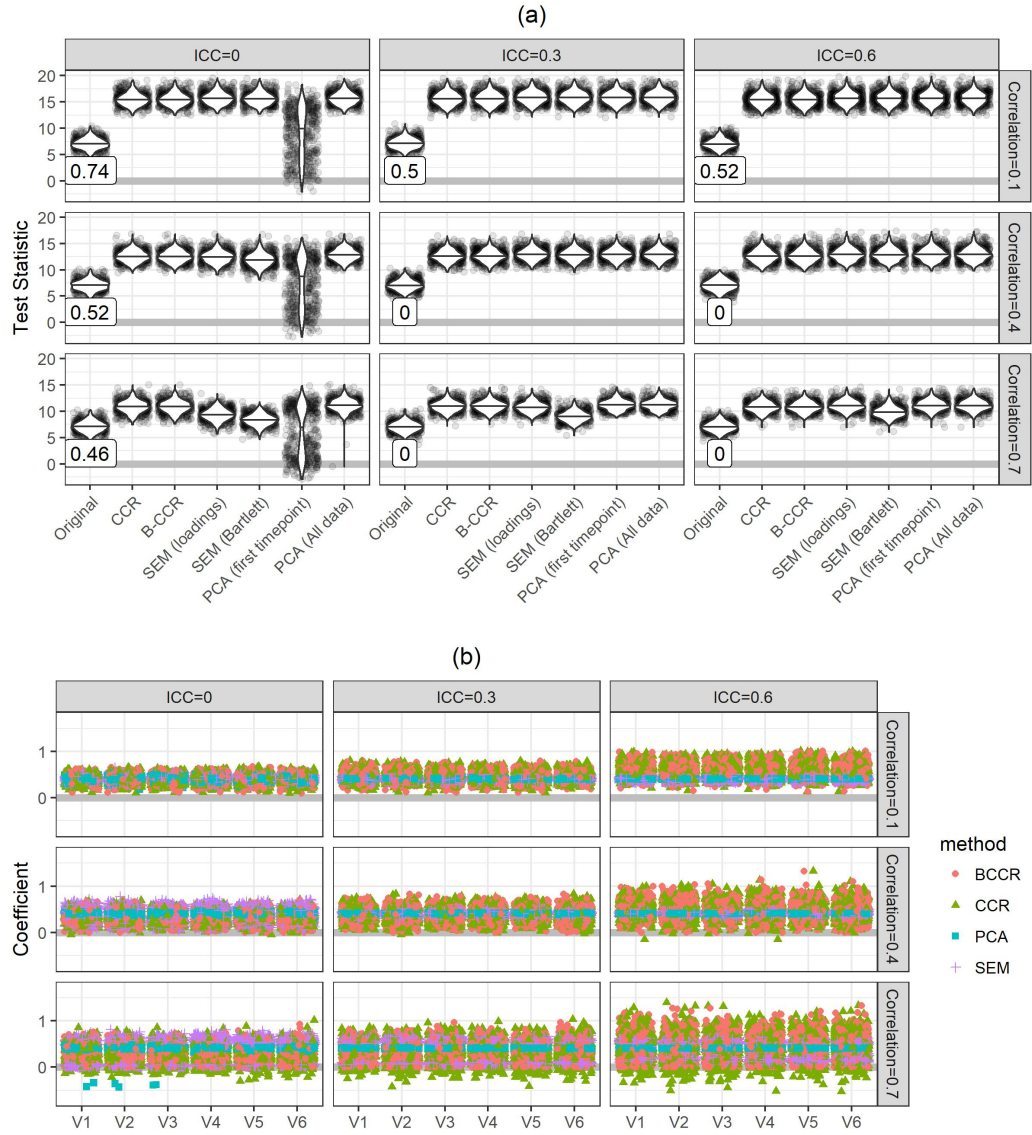


Figure 2.7: Simulation results for case 2: both signals changing equally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

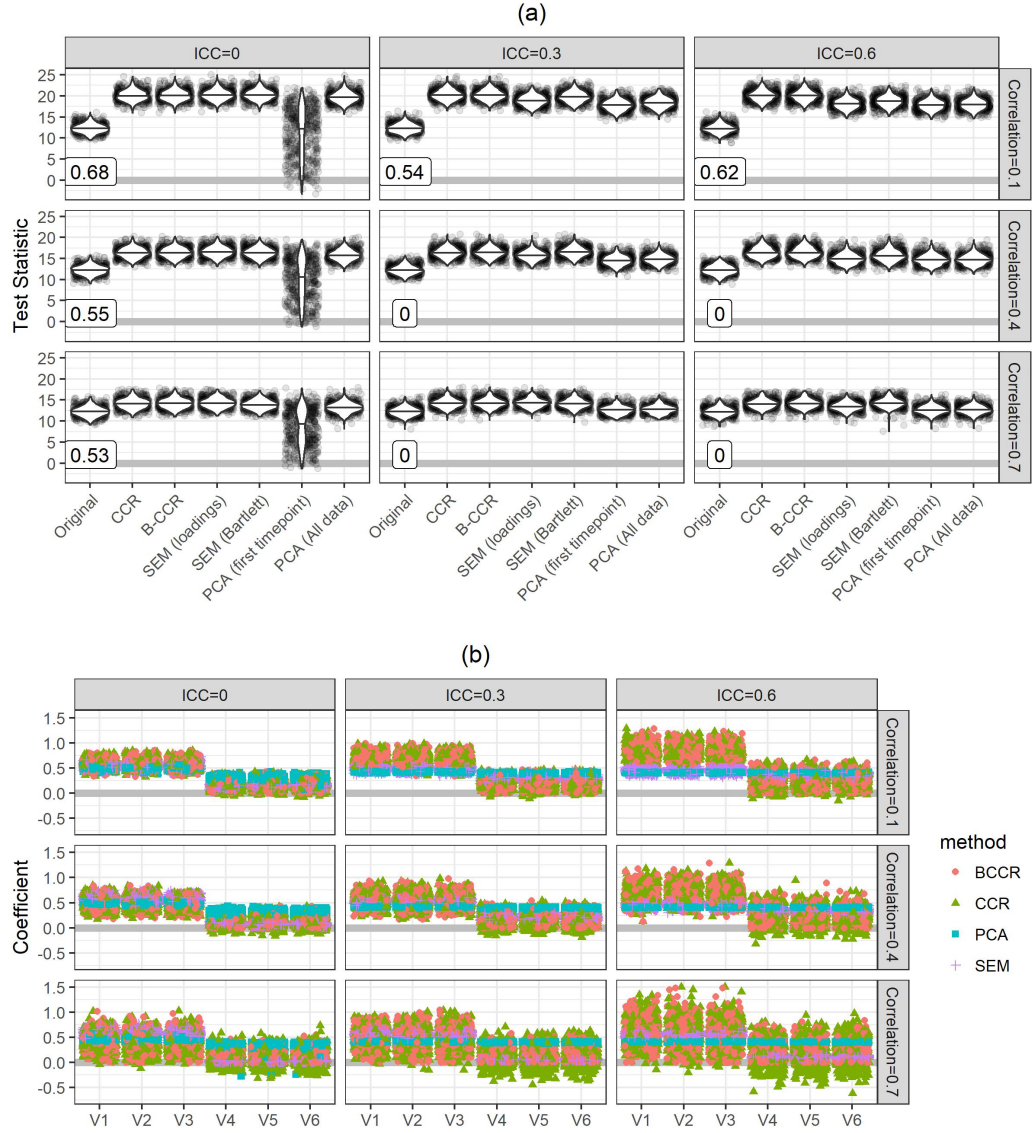


Figure 2.8: Simulation results for case 2: signal changing unequally. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

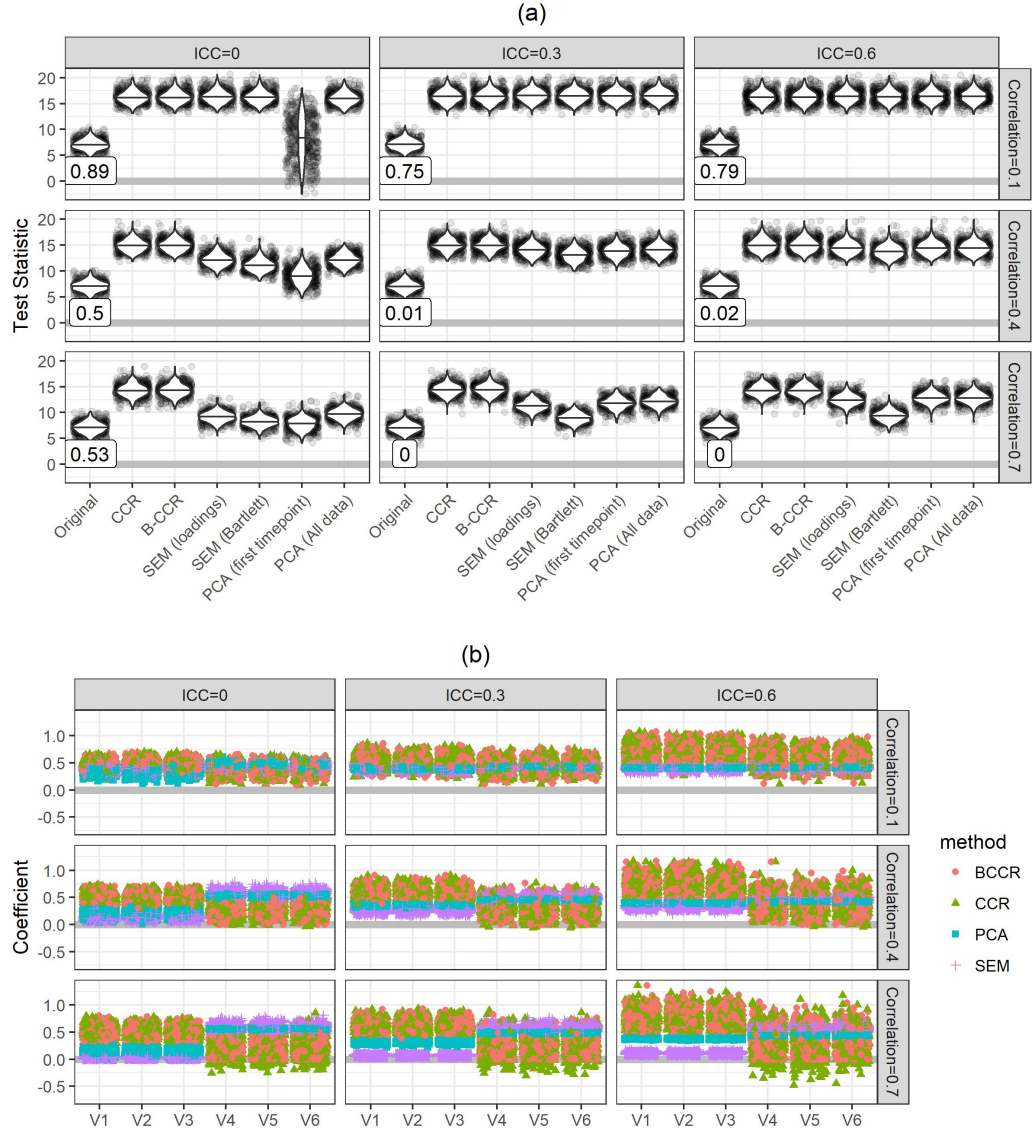


Figure 2.9: Simulation results for case 3: low independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first timepoint), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

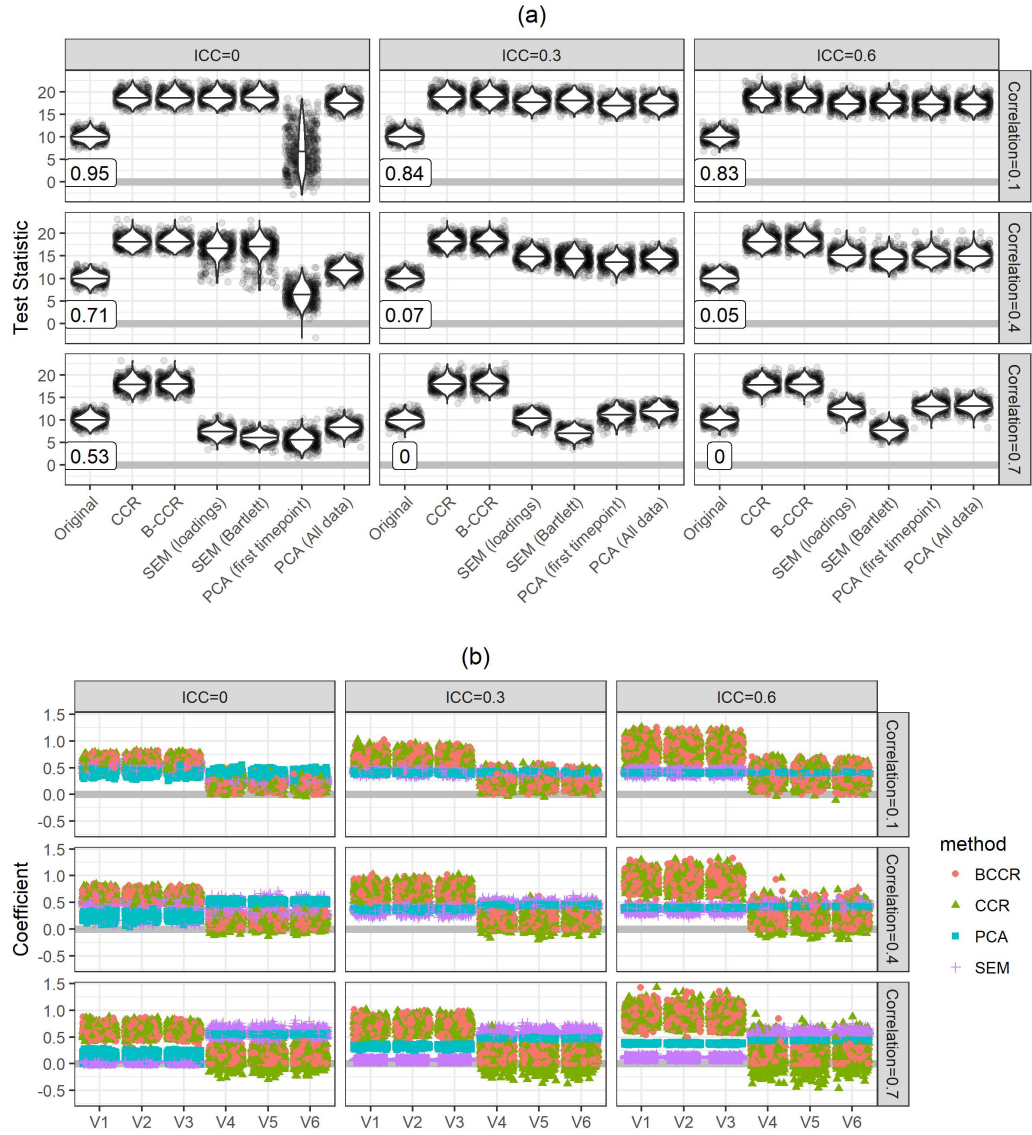


Figure 2.10: Simulation results for case 3: moderate independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

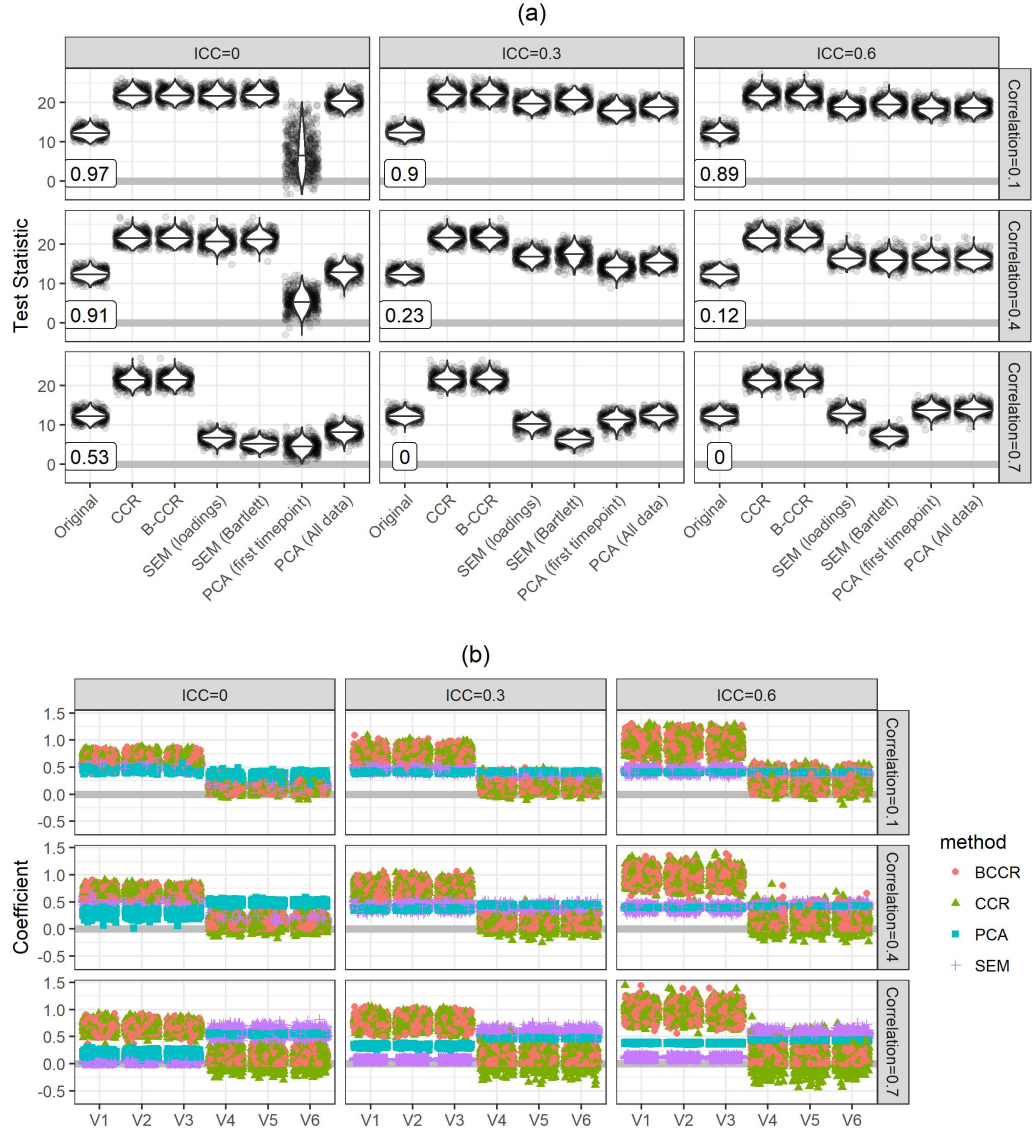


Figure 2.11: Simulation results for case 3: high independent change. (a) Validation test-statistics for the original variable with the largest change and each of the projection methods (CCR, B-CCR, SEM (loadings), SEM (Bartlett scores), PCA (first time point), PCA (All time points)) across different covariance structures, with SNR equal to 1. The proportion of SEM simulations that failed to converge is displayed in the bottom-left corner of each panel. (b) Projection coefficients (CCR, B-CCR, SEM (Bartlett scores), PCA (all time points)) assigned to each of the original variables, with a grey line at 0 (no contribution).

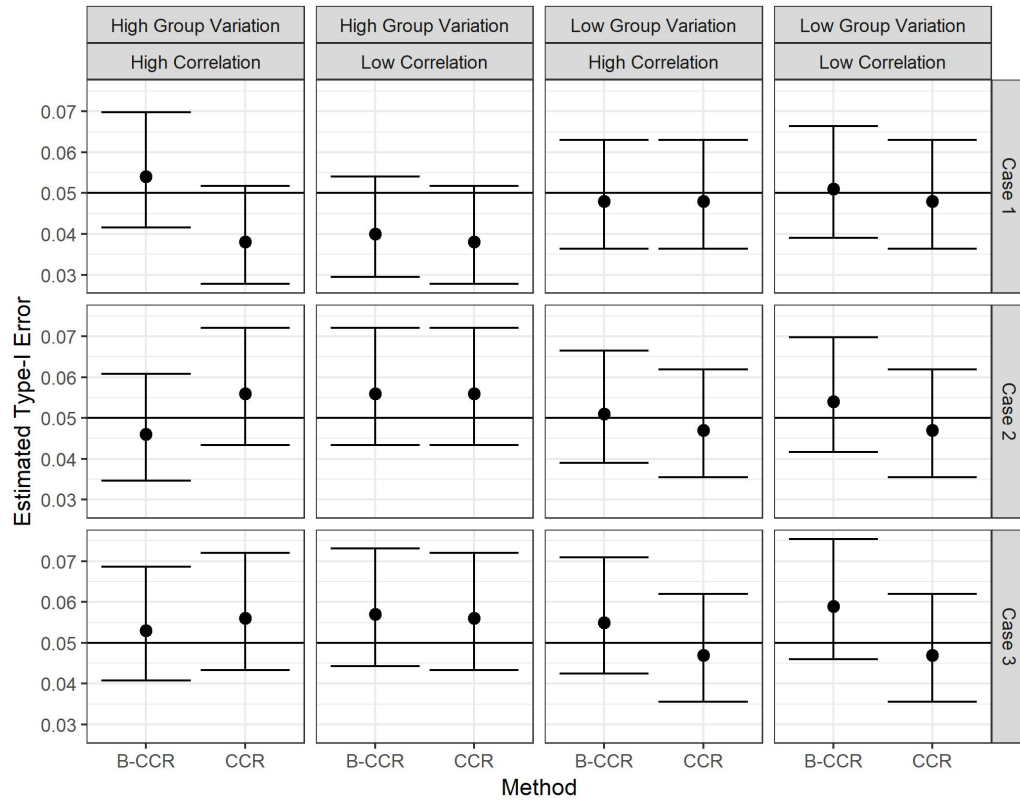


Figure 2.12: Results from simulation study to quantify the Type I error of the CCR projections across a variety of data structures. In each panel, the estimated Type I error rate based on 1000 simulations along with a 95% score interval is displayed.

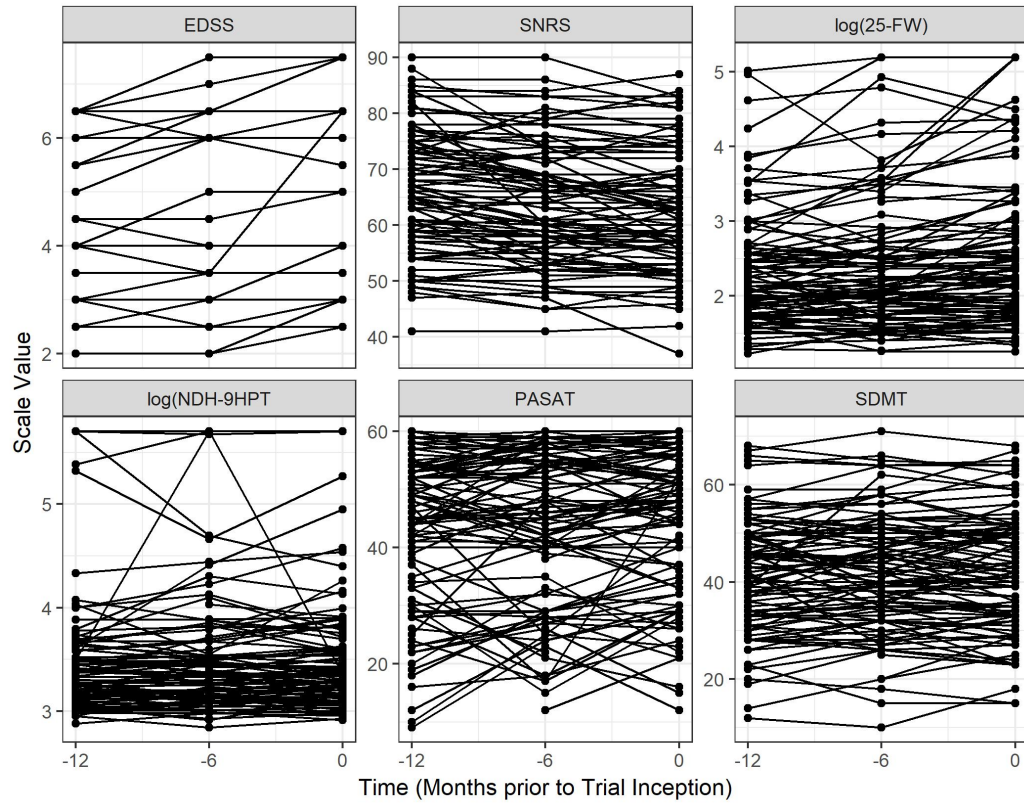


Figure 2.13: Example data from Phase II clinical trials of progressive MS. Each patient was measured three times at 6 month intervals leading to the inception of the trial. Each panel represents a different clinical scale measured. The x-axis corresponds to time and the y-axis corresponds to the measurement scale. Lines connect measurements from the same patient.

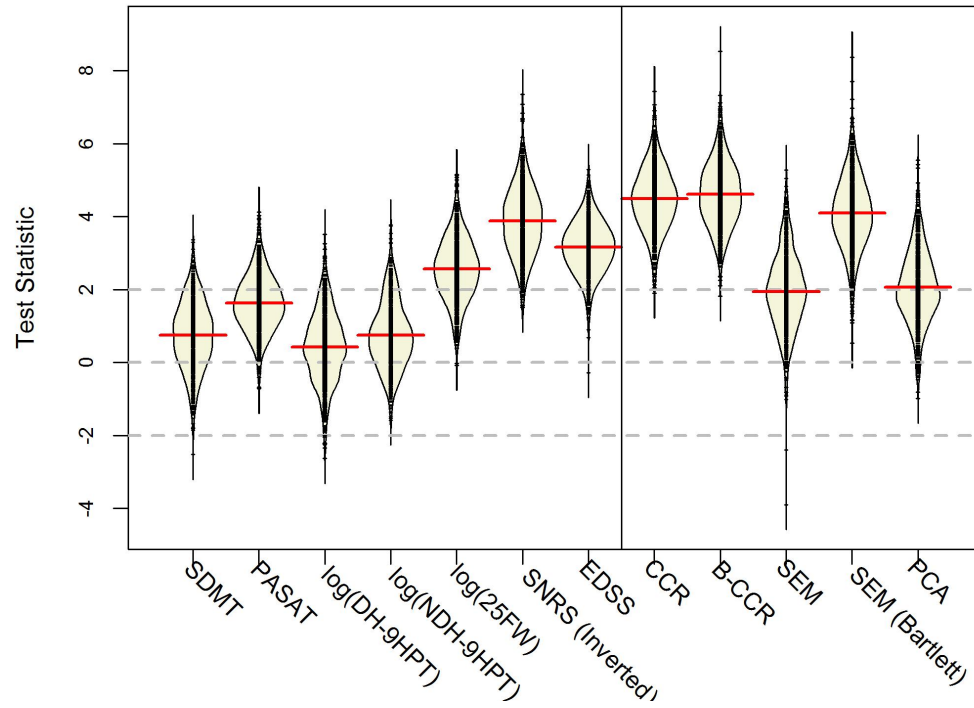


Figure 2.14: A beanplot of 500 permutations of validation data test-statistics for CCR solution in addition to individual clinical scales. Black lines represent individual test statistics, red lines show average of test statistics for each scale, yellow areas represent non-parametric density curves of individual test statistics, grey-dashed lines correspond to approximate cut-offs for 5% significance level tests.

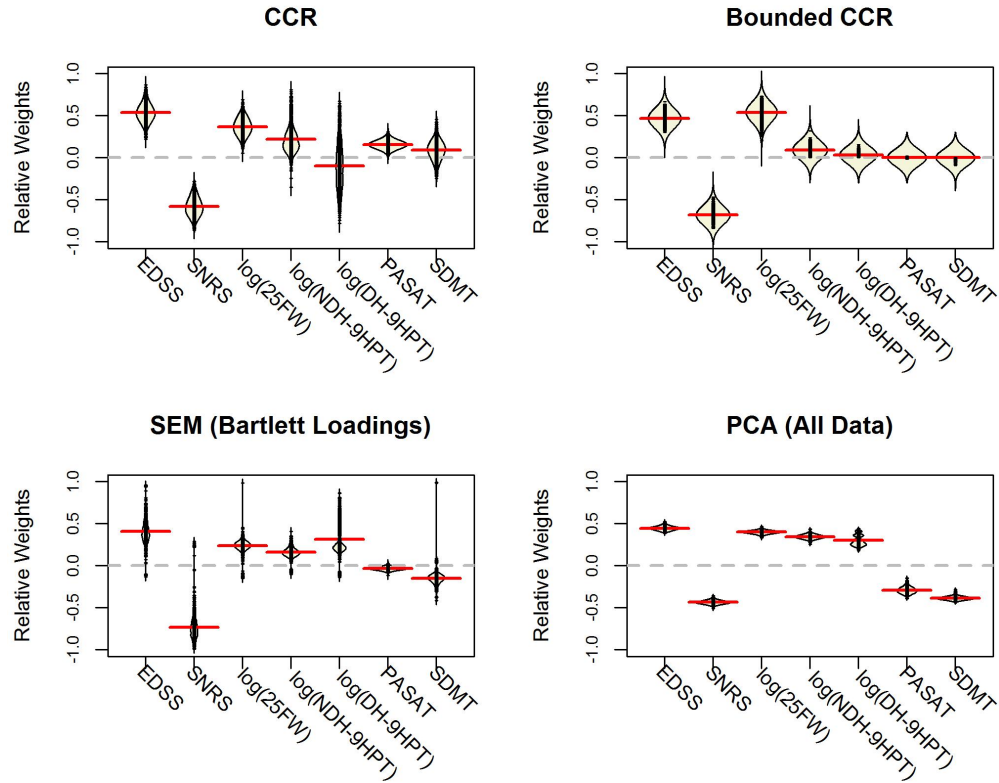


Figure 2.15: Beanplots of coefficient weights for each scale assigned from 500 permutations of training data for CCR projection (top-left), a bounded CCR (B-CCR) projection (top-right), weights from a SEM (bottom-left), and weights from the first principal component (bottom-right). Black lines represent individual coefficient values, red lines show average of coefficients for each scale, yellow areas represent non-parametric density curves of individual coefficients, grey-dashed lines correspond to 0 (i.e., zero influence).

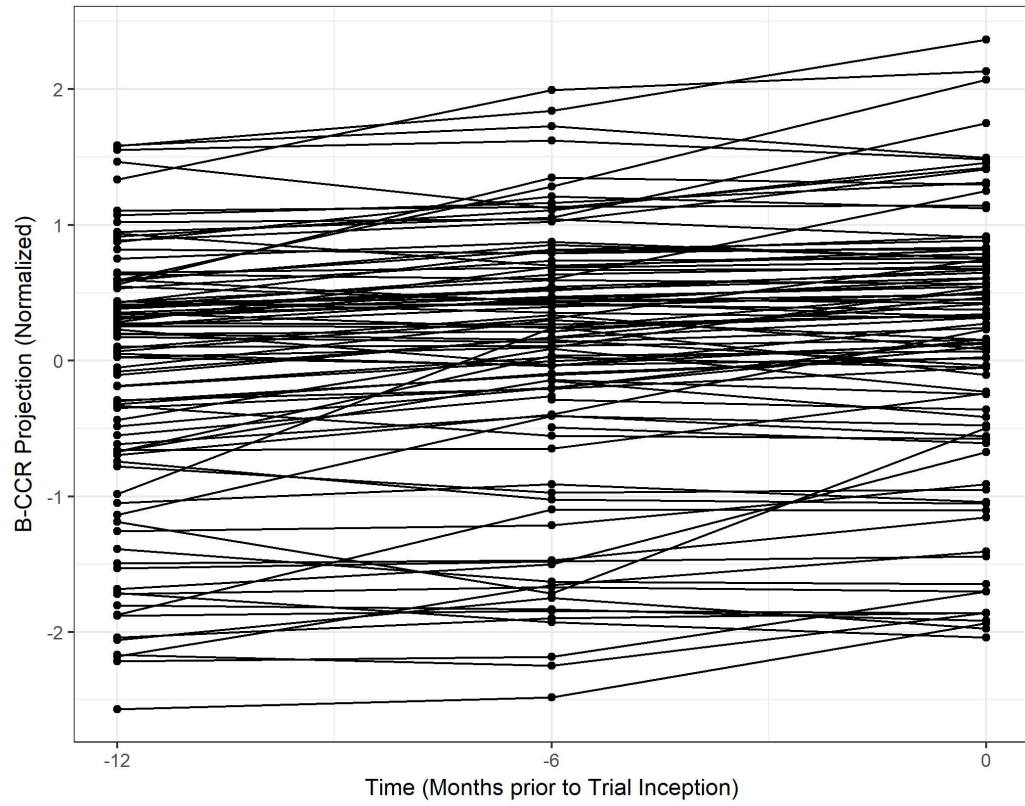


Figure 2.16: B-CCR projection in the example data from Phase II clinical trials of progressive MS. Each patient was measured three times at 6 month intervals leading to the inception of the trial. Lines connect measurements from the same patient.

CHAPTER THREE

EXTENDING CCR METHODOLOGY TO HIGH-DIMENSIONAL DATA TO
DEVELOP A SENSITIVE CLINICAL ENDPOINT FOR MULTIPLE SCLEROSISContribution of Authors and Co-Authors

Author: Christopher R. Barbour

Contributions: Responsible for the majority of the writing and development of the methodology.

Co-Author: Dr. Mark Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscripts.

Co-Author: Dr. Dominique Zosso

Contributions: Constructed the framework for the developed algorithm and provided feedback on the methodology and drafts of the manuscript.

Co-Author: Dr. Bibiana Bielekova

Contributions: Provided feedback on the scientific application and drafts of the manuscript.

Manuscript Information Page

Authors: Christopher Barbour, Mark Greenwood, Dominique Zosso and Bibiana Bielekova

Journal: *Computational and Graphical Statistics*

Status of Manuscript:

☒ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

3.1 Introduction

Clinical scales are diagnostic measurements used by clinicians to determine the presence of a disease or level of impairment/severity for a disease for individual patients. Scales such as these are often constructed from multiple outcome measures to create a combined metric that better measures the true underlying trait of interest than any of the original components. These methods typically focus on explaining cross-sectional variation in the responses using projections into one or more dimensions to define the combinations of variables. When the interest is in creating a scale that is sensitive to changes over time, developing it using cross-sectional data may not tune the projection to detect changes over time optimally. This led to the development of the Constructed Composite Response methodology (Chapter 2). Instead of focusing on creating a projection that optimizes cross-sectional correlation or variance, the proposed method creates projections that maximize a t -statistic for a linear change over time in a specified linear model. This metric combines the estimated change over time ($\hat{\beta}_{time}$) and its precision ($SE(\hat{\beta}_{time})$). While the CCR methodology has shown promise in lower dimensional spaces, it was not tested on high-dimensional problems. This chapter will assess the performance of the CCR in high-dimensional settings, and includes extensions to increase the utility of this method in high-dimensional problems.

Let $\mathbf{y}_{jt}$ be a length- P vector containing the P observed response variables for observation (e.g., patient) j ($j = 1, \dots, J$, where J is the total number of observations) at time point t ($t = 1, \dots, T$, where T is the total number of observed time points) and let $\mathbf{s}_p$ be a length- $(J \cdot T)$ vector containing the observed values of response variable p ($p = 1, \dots, P$) for all J subjects and all T time points. Then $n = J \cdot T$ is the total sample size and let $\mathbf{X}$ be a fixed n by q model matrix of full

column rank (where q is the number of model parameters to be estimated) containing control variables to be accounted for (e.g., differences among patients) as well as the corresponding variable of interest (such as time). For a linear projection of the response variables at time t , $\mathbf{y}_{jt}$, into a CCR, $z_{jt} = \mathbf{c}^\top \mathbf{y}_{jt}$, we want to determine the row vector $\mathbf{c}^\top$ that maximizes the t -statistic for a linear time change in the assumed data-generating linear model

$$z_i = \mathbf{X}_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.$$

Let $\boldsymbol{\gamma} = \left(\hat{\beta}_{time-1}, \hat{\beta}_{time-2}, \dots, \hat{\beta}_{time-P} \right)^\top$ be a column vector of the coefficients from modeling the individual response variables $\mathbf{s}_1$ through $\mathbf{s}_p$ and let $\mathbf{B} = \boldsymbol{\gamma} \boldsymbol{\gamma}^\top$. Similarly, let $\mathbf{R}$ be a P by n matrix where row p corresponds to the residual vector from modeling $\mathbf{s}_p$, and set $\mathbf{M} = \frac{1}{n-q-1} \mathbf{R} \mathbf{R}^\top$. Note that both $\mathbf{B}$ and $\mathbf{M}$ are symmetric matrices. The solution to the objective function then becomes

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}} \frac{\mathbf{c}^\top \mathbf{B} \mathbf{c}}{\mathbf{c}^\top \mathbf{M} \mathbf{c}}.$$

If $\mathbf{M}$ is a positive-definite matrix, then an analytic solution exists up to a scaling constant, with the solution being the leading eigenvector of $\mathbf{M}^{-1} \mathbf{B}$ (Hardle and Simar, 2012). When the level of cross-sectional variation in the observed scales does not match their respective longitudinal change, the CCR methodology consistently produces projections with larger t -statistics compared to more traditional scale development techniques, such as principal component analysis (PCA) and structural equation models (SEM). Bounding the solution space of the CCR allows for more interpretable projection coefficients, as the sign of c_j must match the direction of the expected change, if present (see Chapter 2 for further details).

It is possible, especially in situations where many scales are considered, that

excluding some scales from consideration (resulting in fewer non-zero coefficients) could produce a projection with similar, or even larger t -statistics as a projection considering all available data. This type of high-dimensional data are becoming common in various scientific disciplines, often creating statistical and computational challenges. In order to exclude scales in an objective fashion, some additional constraints need to be imposed on the solution vector, $\hat{\mathbf{c}}$, and different constraints will lead to different properties in the estimated solutions. The properties of many such constraints have been studied in the context of linear regression, an active area for high-dimensional statistical research.

In this setting, we have a response vector, $\mathbf{y}$, of length n and an n by q model matrix $\mathbf{X}$, where n is the total sample size and q is the number of model parameters to be estimated. The goal of linear regression is to determine the coefficient vector $\boldsymbol{\beta}$ that minimizes $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$. As the number of model coefficients, q , approaches the number of observations, n , the least-squares solution to this problem, $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, becomes unstable. When $q \geq n$, least-squares regression cannot be performed, as $\mathbf{X}^\top \mathbf{X}$ becomes singular. In recent years, research has been devoted to extending linear regression to high-dimensional settings where $q \geq n$. These typically focus on minimizing the least-squares criterion, subject to a geometric constraint on the solution vector $\boldsymbol{\beta}$, often referred to as penalized regression. These models are formulated by selecting $\boldsymbol{\beta}$ that minimizes the objective function

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda R(\boldsymbol{\beta})$$

where $R(\boldsymbol{\beta})$ is a regularization function on the model coefficients and λ is a tuning parameter that controls the amount of regularization performed. This is typically performed after all columns of $\mathbf{X}$ (other than the “intercept” column) have been

standardized to mean 0 and unit variance, in order to make the size of the estimated coefficients comparable.

One of the earliest examples of penalized regression was motivated by a near-singular design matrix. Ridge regression (Hoerl and Kennard, 1970) minimizes the least-squares criterion subject to a constraint on the squared l_2 norm of the coefficient vector, formulated as

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

This has a unique closed form for a fixed value of λ , represented by $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$. In effect, the method adds a small value to the “ridge”, or diagonal, of $\mathbf{X}^\top \mathbf{X}$, assisting the matrix inversion when columns of X are highly correlated. The tuning parameter λ controls the amount of regularization to perform. As λ increases, the model coefficients are shrunk towards 0. Although ridge regression creates biased estimators of $\boldsymbol{\beta}$ (Hoerl and Kennard, 1970), it can greatly reduce the variance of the estimator, often worth the trade-off. One of the most influential methods in high-dimensional statistical regression approaches is lasso regression (Tibshirani, 1996), which was motivated by the non-negative garrote (Breiman, 1995). This penalized regression model incorporates an l_1 penalty on the coefficient vector $\boldsymbol{\beta}$,

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_m = \sqrt[m]{\sum_{q=1}^Q |\beta_q|^m}$. Except in the instances of a single covariate or an orthogonal design matrix, an analytic solution to this problem does not exist. Both lasso and ridge regression are special sub-cases of the general class of l_m penalized

regression models (Hastie, Tibshirani, and Wainwright, 2015),

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum |\beta_j|^m,$$

with $m = 1$ corresponding to the lasso criterion and $m = 2$ corresponding to ridge.

Examining the solutions of ridge and lasso regression with an orthogonal design matrix, $\mathbf{X}$, leads to insights into how these penalties operate. Let $\tilde{\beta}_j$ be the least-squares solution for variable j in a linear model. For a fixed value of λ , the ridge regression solution is $\tilde{\beta}_j / (1 + \lambda)$, and the lasso solution is $\text{sign}(\tilde{\beta}_j) \left(|\tilde{\beta}_j| - \lambda \right)_+$, where the t_+ is the soft-thresholding operator, and corresponds to t if t is non-negative and 0 otherwise (Hastie, Tibshirani, and Wainwright, 2015). With increasing values of λ , ridge estimates are shrunk towards 0, but are always retained with a non-zero coefficient. When the number of predictors is large, this can lead to an overly-complicated model that is difficult to interpret. In contrast, the lasso shrinks coefficients towards 0 and automatically sets model coefficients to 0 as λ increases. In effect, this allows for non-contributing variables to have a coefficient of 0, allowing for automatic variable selection with large values of λ .

A stark difference between the ridge and lasso penalties is what happens in the presence of correlated variables; while ridge regression assigns similar coefficient values to highly-correlated variables, lasso regression tends to assign a non-zero coefficient to a single variable in a group of correlated variables. The elastic net regression (Zou and Hastie, 2005, Friedman, Hastie, and Tibshirani, 2010) combines both of these penalties,

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left[(1 - \alpha) \frac{\|\boldsymbol{\beta}\|_2^2}{2} + \alpha \|\boldsymbol{\beta}\|_1 \right],$$

where α is an additional tuning parameter controlling the balance between the ridge

and lasso penalties, in effect allowing for variable selection to be performed while also retaining groups of correlated variables. All three of these methods can be efficiently fit using cyclic coordinate descent in the `glmnet` R package (Friedman, Hastie, and Tibshirani, 2010). The elastic net framework has also been extended to work with Generalized Linear models for count, binomial, and survival response models (Tibshirani and Taylor, 2011).

A notable extension upon lasso regularization involves data where grouping structures exist (either assumed known, or to be estimated from the data). A simple example of such grouping is the inclusion of categorical predictors in a linear model, typically done by fixing one level of the factor as a “baseline” and including indicator variables for the other levels. When performing variable selection, we would likely want to retain or exclude all coefficients for individual factors. Another example of such a grouping arises in gene-expression data, where genes within the same biological pathway form a natural grouping. The group lasso (Yuan and Lin, 2007) allows for *a priori* groupings of predictors to be included in the model formulation, with all coefficients within groups being retained or excluded. Since the original group lasso was proposed, some extensions have been developed to allow for within-group sparsity (Puig, Wiesel, and Hero, 2009), overlap of variables between groups (Jacob and Vert, 2009), and inclusion of interactions between variables (Lim and Hastie, 2014).

Other times, an underlying grouping between variables is expected, but the exact grouping is unknown *a priori*. The octagonal shrinkage and clustering algorithm (OSCAR, Bondell and Reich, 2008), in addition to the l_1 norm, includes pairwise-infinity norms between all predictors that encourages highly correlated variables to be assigned the same $|\hat{\beta}|$, allowing for simultaneous variable selection and forming of variable groups. This was solved via quadratic programming, which becomes impractical to compute as dimensionality increases. The cluster elastic net (Witten,

Shojaie, and Zhang, 2014), which extended upon elastic net to allow for data-driven groupings of variables to be determined, scales better for high-dimensional problems.

Work following the lasso showed that in $Q \gg n$ situations, the amount of shrinkage to eliminate noise variables using the lasso criterion can result in too much shrinkage being performed on the signal variables, and that non-convex functions, if computationally feasible, that perform less shrinking on the signal variables can be attractive in these instances (Meinshausen, 2001, Meinshausen, 2007). Considering the l_m penalized-modeling framework defined above and considering an orthogonal design matrix, $\mathbf{X}$, setting $m = 0$ corresponds to best-subset selection (Hastie, Tibshirani, and Wainwright, 2015), with model coefficients being defined via hard-thresholding, $\tilde{\beta}_j \mathbf{I}(|\tilde{\beta}_j| > \sqrt{2\lambda})$. Both regularization through soft- and hard-thresholding will be applied to the CCR methodology developed in Chapter 2. The lasso type l_1 penalty imposes soft-thresholding on projection coefficients, while the approximate l_0 penalty imposes hard-thresholding. Both operators can impose variable selection, but will differ in convergence properties and generated solutions paths, which will be explored in simulated and real data.

The chapter is organized as follows. In Section 2, we extend the CCR methodology to allow for both l_1 and l_0 regularization in both the CCR and B-CCR algorithms of Chapter 2. Section 3 compares the different CCR formulations via a simulation study. Section 4 applies our method to a motivating dataset of neurological exam data for progressive multiple sclerosis (MS) patients, where interest lies in determining a combination of neurological functions that associates with longitudinal change. Section 5 discusses findings and considers future extensions.

3.2 Penalized Constructed Composite Response

Penalized versions of the CCR algorithm, which include sparsity inducing constraints, can be formulated as

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}} \mathbf{c}^\top \mathbf{B} \mathbf{c} - \lambda \sum |\beta_j|^m \quad s.t. \quad \mathbf{c}^\top \mathbf{M} \mathbf{c} \leq 1, \quad \mathbf{c}^\top \boldsymbol{\gamma} > 0$$

for the algorithm allowing for negative c_j values and

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}} \mathbf{c}^\top \mathbf{B} \mathbf{c} - \lambda \sum |\beta_j|^m \quad s.t. \quad \mathbf{c}^\top \mathbf{M} \mathbf{c} \leq 1, \quad \forall p: c_p > 0, \quad \mathbf{c}^\top \boldsymbol{\gamma} > 0$$

for the algorithm bounding coefficients to be non-negative. In both algorithms, λ represents a regularization parameter determining the amount of sparsity (larger values of λ correspond to more sparsity). Both $m = 0$ (corresponding to hard-thresholding on estimated coefficients) and $m = 1$ (corresponding to soft-thresholding of estimated coefficients) were considered. When the non-negativity constraint is imposed, single-sided thresholding operators are performed. When non-negativity is not imposed, double-sided thresholding operators are performed. The algorithm is solved similarly to the Bounded CCR algorithm in Chapter 2 for a fixed λ . The non-negativity projection is simply replaced by the respective thresholding operator. The updated algorithms are displayed in Algorithms 2 through 5.

3.3 Simulation Study

To explore the performance of the regularized CCR methodology compared to the un-regularized versions, a simulation study was implemented to explore the properties of the algorithm across a range of data structures. The analytic CCR,

bounded-CCR, l_1 -penalized CCR (with and without non-negativity imposed), and l_0 -penalized CCR (with and without non-negativity) were examined. Within each set of simulation parameters, a training dataset was drawn to construct the individual solutions (and solution paths). A separate validation set was used solely for selecting the tuning parameter λ for the penalized CCR methods, where λ was selected that maximized the objective function in the validation dataset. The value of the objective function was assessed on a final test dataset. The proportion of correctly selected and correctly screened variables were computed for each method.

All data were drawn from the data-generating model,

$$\mathbf{y}_{jt} \stackrel{i.i.d}{\sim} MVN(\mu + \boldsymbol{\beta}time_t, \mathbf{D}\mathbf{R}\mathbf{D}), \quad (3.1)$$

where $\boldsymbol{\beta}$ is a column vector of changes over time in individual scales, $\mathbf{D} = \sigma\mathbf{I}$ is a diagonal matrix of standard deviations of individual scales, $\mathbf{R}$ is a symmetric matrix of cross-sectional correlations between pairs of scales, and μ is the average starting value for each observation. Each dataset consisted of $J = 50$ subjects with $T = 6$ equally spaced time points. Four primary cases were considered, three low-dimensional cases, motivated by the simulation study in the elastic net paper (Zou and Hastie, 2005), and one high-dimensional case motivated by the simulation study in the cluster elastic net paper (Witten, Shojaie, and Zhang, 2014).

Case 1: Eight potential response variables were considered, with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and σ set to achieve an SNR ($\frac{\beta}{\sigma^2}$) of 0.1 for the maximum element of $\boldsymbol{\beta}$. Pairwise residual correlations between responses ($\mathbf{R}_{ij}$) were set to $\rho^{|i-j|}$.

Case 2: Same as above, but with $\beta_p = 0.85$ for all p .

Case 3: Forty potential response variables were considered, where

$$\beta = \left(\overbrace{2 \dots, 2}^5, \overbrace{0 \dots, 0}^5, \overbrace{2 \dots, 2}^5, \overbrace{0 \dots, 0}^5, \overbrace{0 \dots, 0}^{20} \right)$$

and the correlation matrix was constructed by

$$\mathbf{I}_4 \otimes \mathbf{R}$$

where $\mathbf{R}$ is a 10×10 symmetric matrix with 1 on the diagonal and ρ on the off-diagonal. This creates 4 pairwise-independent groups of correlated variables, with 2 groups containing some variables changing over time, and 2 groups consisting exclusively of noise variables.

Case 4: The number of potential responses were set to 200, with $\beta_p \sim \text{Unif}(3.9, 4.1)$ for $1 \leq p \leq 10$, $\beta_p \sim \text{Unif}(1.9, 2.1)$ for $26 \leq p \leq 35$, and $\beta_p = 0$ otherwise. The correlation matrix was determined by

$$\mathbf{I}_{25} \otimes \mathbf{R}$$

where $\mathbf{R}$ is a 25×25 symmetric matrix with 1 on the diagonal and ρ on the off-diagonal. This creates 25 pairwise-independent groups of correlated variables, with 2 groups containing some variables changing over time, and 23 groups of only noise variables.

Across all cases, the pairwise correlation parameter ρ took values of 0, 0.15, and 0.3.

The results are displayed in Table 3.1 and Figure 3.1. For the low-dimensional cases, bounding coefficients in the non-negative orthant leads to projections with smaller t -statistics, regardless of the form of regularization used. This was also

observed in the Chapter 2 simulation study, and is observed because the unbounded versions of the algorithm can assign opposing signed coefficients between pairs of correlated variables in order to decrease the MSE from the resulting model. Examining bounded and unbounded algorithms, slightly larger t -statistics are observed for regularized solutions when correlation is low, with approximately equal t -statistics as correlation increases. No noticeable differences in the magnitude of the t -statistics are observed when comparing soft-thresholding versus hard-thresholding regularization.

When examining the correct sparsity, the non-negative constrained methods tend to correctly determine whether variables should be retained. The only instance that the analytic CCR has a high correct-sparsity is the instance when all variables are changing (Figure 3.1b, middle-row). When cross-sectional correlation is low, both bounded and unbounded regularized solutions achieve a high correct sparsity (Figure 3.1b, first-column). In the lowest dimensional case when not all variables are changing over time (Figure 3.1b, first-row), as correlation increases, the unconstrained regularized solutions produce similar projections as the analytic CCR, allowing noise variables to be retained in order to decrease the MSE in the resulting model. For the low-dimensional Case 3 (Figure 3.1b, bottom-row), as correlation increases, the unconstrained regularized solutions have higher correct sparsity estimates compared to the analytic CCR solutions. In the high-dimensional case (Case 4) with many noise variables, performing regularization greatly increases the magnitude of the t -statistics captured by the projection, as many noise variables are eliminated. As the level of cross-sectional correlation increases, this increase in t -statistics becomes less prominent.

To better understand how each penalty is operating, solution paths of $\mathbf{c}^\top$ across a grid of $\log(\lambda)$ were constructed from a single training dataset from Case 3 with ρ of 0 and 0.3, respectively. These solution paths are displayed in Figure 3.2, with signal

variables being displayed as red lines and noise variables being displayed as black lines. As expected, the solution paths for l_1 regularization are much smoother than those for l_0 regularization. When correlation is not present, a portion of the solution path contains the ideal solutions (only retaining signal variables), regardless of the type of regularization and whether or not bounding was present. When correlation is present, unbounded solution paths tend to select many noise variables with non-zero coefficients (more so in l_1 regularization). Bounded solution paths threshold out these noise variables, at the cost of decreased performance. Through the development of the simulation study, it was noted that the selection of tuning parameters for the hard-thresholding operators were more difficult to set, and the soft-thresholding operators from the l_1 regularization converged more regularly.

3.4 Real Data Example

We now demonstrate the performance of the regularized CCR projections on a longitudinal cohort of 919 progressive MS patient-visits (217 total patients) collected between 5/19/2005 and 8/23/2018. MS is a neurological disease that affects the central nervous system. Symptoms of MS include muscle weakness/difficulty with balance and coordination, trouble with sensory feeling, blurred/double vision and even blindness. Various clinical scales quantify disease progression in patients, but they have varying degrees of sensitivity to detect disease progression and various levels of measurement error. For example, the Expanded Disability Status Scale (EDSS, Kurtzke, 1983) is a 0-10 Likert-scale clinical endpoint that is currently utilized by the Food and Drug Administration (FDA) for MS clinical trials. However, once a patient becomes unable to walk (around an EDSS of 7), EDSS is unlikely to change again for several years even though the patient’s disease is still progressing. This unequal change over time is a weakness of the scale and represents measurement error in

addressing the underlying disease severity in a patient. Currently, testing efficacy of treatments in progressive MS, specifically in Phase II trials, using EDSS requires hundreds of patients being followed for 2-3 years (Kosa et al., 2016). Given the features available from a standard neurological examination (Kosa et al., 2018), we want to simultaneously select which features are changing through the disease process and find a linear combination that best associates with this longitudinal change.

The l_0 penalized B-CCR was used to determine a sparse combination of signals that maximize the detected longitudinal change, after accounting for differences in patient starting values. The tuning parameter λ was selected using 5-fold cross-validation, selecting λ that maximized the average objective in the withheld folds. A final projection was constructed using all of the data using the selected λ . A plot displaying the cross-validated objective function across different $\log(\lambda)$, as well as solution paths of $\hat{\mathbf{c}}$ are displayed in Figure 3.3. In the optimal final projection, 82 signals were retained with a non-zero coefficient. Due to ongoing data collection for development of this clinical endpoint, the results of applying the methods to these data will not be discussed further here.

3.5 Conclusions

The inclusion of l_1 regularization in the CCR framework led to automatic variable selection and projections with larger or similarly sized t -statistics as the non-regularized version using fewer signals in a variety of simulated datasets. Through the motivating dataset, it was uncovered that signal relating to muscular strength, brainstem functionality, cerebellar functionality, and bowel and bladder functionality could be effective surrogates for disease progression, which will be further explored as more data become available. Comparison of the regularized CCR methods to regularized versions of common scale development techniques, such as sparse principal

component analysis (Witten, Tibshirani, and Hastie, 2009, Jolliffe, Trendafilov, and Uddin, 2003) will be examined in future research, as we would expect these methods to have similar deficiencies to regular principal component analysis in certain longitudinal settings but would need to test those methods to verify that our expectations are correct.

While this novel methodology shows promise, there are some aspects that could be improved. Maximizing a t -statistic allowed for detection of longitudinal change in a single group. However, it may be of interest to develop a clinical endpoint that best separates longitudinal progression rates between MS patients and healthy volunteers. Extending the methodology to maximize an F -statistic for a specified model contrast would allow more complex model components, such as group differences with more than two groups, quadratic change over time, and/or interaction terms, to be targeted. This extension will be discussed further in Chapter 5.

The current algorithms have a set of tuning parameters that, if not carefully set in advance, can lead to the algorithm alternating between solutions, or completely diverging. This was observed through the development of the simulation study in Chapter 3, and was more of an issue in the high-dimensional Case 4. A theoretical understanding of the tuning parameters could lead to more educated default settings of these parameters, leading to easier implementation of the methodology.

In the neurological exam data, differences in patient starting values were accounted for using a fixed-effect for patient-to-patient differences. It could be argued that accounting for these differences through a mixed-model formulation with random intercepts and possibly slopes included for each subject would be a more appropriate approach. Extending the CCR for targeting t -statistics in a mixed-model would allow for this and would allow extension into multi-level modeling situations where accounting for multiple sources of variation such as different clinic locations, clinicians,

or devices used to obtain data (such as MRI scanners) could be incorporated into the model. This extension may require alternate optimization techniques as the mixed models often do not lead to closed form estimators and the degrees of freedom for the t -statistics can vary based on the observed response patterns.

As noted throughout Chapter 2 and the current chapter, the objective function of the CCR (for fixed signal-to-noise) decreases as the level of cross-sectional correlation between variable pairs increases. As the methodology provides more optimal projections when different uncorrelated dimensions are available, we hypothesize that exploiting an *a priori* grouping structure of correlated variables to create approximately independent signals could further improve the CCR methodology. This is similar to the aggregation of group-level information in the Group Lasso, cluster elastic net, and OSCAR methods discussed earlier and will be explored in future research.

In conclusion, the CCR methodology was extended to incorporate sparsity-inducing constraints, motivated by advances in penalized-regression models. A simulation study that examined the different constraint regions compared to the original CCR across a variety of datasets with varying true sparsity, longitudinal change, dimensionality, and cross-sectional correlation demonstrated the superiority of the penalized CCR, with the l_1 constraint performing the best in terms of ease of convergence, size of targeted t -statistic in data independent of model development, and interpretability of resulting projection coefficients. Applying the penalized CCR to a longitudinal dataset of progressive MS patients identified important neurological functions that associate with longitudinal change in disease and allows for more efficient scale development in high-dimensional situations.

Algorithm 2: Bounded CCR Algorithm with Soft-thresholding

Require: $\mathbf{B}$ and $\mathbf{M}$ matrices, $\boldsymbol{\gamma}$ column-vector, λ .

```

function REGULARIZED_CCR( $\mathbf{B}, \mathbf{M}, \boldsymbol{\gamma}, \lambda, r_{11} = 0.1, r_{12} = 0.1, r_2 = 0.1, r_3 =$ 
5,  $tol = 1 \times 10^{-14}$ )

     $\Delta_a \leftarrow 0$ 
     $\Delta_b \leftarrow 0$ 
     $\mathbf{c}_0 \leftarrow \boldsymbol{\gamma}$ 
     $\mu \leftarrow 0$ 
     $\bar{\mathbf{c}} \leftarrow \mathbf{c}_0$ 

    while  $Error > tol$  do

         $\Delta_a \leftarrow \frac{(\Delta_a + \mathbf{B}\bar{\mathbf{c}})}{1 + r_{11}}$  ▷ gradient-accumulators
         $\Delta_b \leftarrow \frac{(\Delta_b + \mu \mathbf{M}\bar{\mathbf{c}})}{1 + r_{12}}$ 
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_k + r_{11}r_2\Delta_a - r_{12}r_2\mu\Delta_b$  ▷ gradient descent (primal)
         $\mathbf{c}_{k+1} \leftarrow (\mathbf{c}_{k+1} - \lambda r_2)_+$  ▷ single-sided soft-thresholding
         $\bar{\mathbf{c}} \leftarrow 2\mathbf{c}_{k+1} - \mathbf{c}_k$  ▷ extra gradient step
         $\mu \leftarrow \mu + \frac{r_3}{2} (\mathbf{c}_{k+1}^\top \mathbf{M} \mathbf{c}_{k+1} - 1)$  ▷ KKT-multiplier update
         $\mu \leftarrow \max(\mu, 0)$  ▷ KKT-projection (dual feasibility)
         $Error \leftarrow \|\mathbf{c}_{k+1} - \mathbf{c}_k\|_\infty$ 

    end while

    return  $\mathbf{c}_{k+1}$ 

end function

```

Algorithm 3: CCR Algorithm with Soft-thresholding

Require: $\mathbf{B}$ and $\mathbf{M}$ matrices, γ column-vector, λ .

```

function REGULARIZED_CCR( $\mathbf{B}, \mathbf{M}, \gamma, \lambda, r_{11} = 0.1, r_{12} = 0.1, r_2 = 0.1, r_3 =$ 
5,  $tol = 1 \times 10^{-14}$ )

     $\Delta_a \leftarrow 0$ 
     $\Delta_b \leftarrow 0$ 
     $\mathbf{c}_0 \leftarrow \gamma$ 
     $\mu \leftarrow 0$ 
     $\bar{\mathbf{c}} \leftarrow \mathbf{c}_0$ 

    while  $Error > tol$  do

         $\Delta_a \leftarrow \frac{(\Delta_a + \mathbf{B}\bar{\mathbf{c}})}{1 + r_{11}}$  ▷ gradient-accumulators
         $\Delta_b \leftarrow \frac{(\Delta_b + \mu \mathbf{M}\bar{\mathbf{c}})}{1 + r_{12}}$ 
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_k + r_{11}r_2\Delta_a - r_{12}r_2\mu\Delta_b$  ▷ gradient descent (primal)
         $\mathbf{c}_{k+1} \leftarrow \text{sign}(\mathbf{c}_{k+1}) (|\mathbf{c}_{k+1}| - \lambda r_2)_+$  ▷ soft-thresholding
         $\bar{\mathbf{c}} \leftarrow 2\mathbf{c}_{k+1} - \mathbf{c}_k$  ▷ extra gradient step
         $\mu \leftarrow \mu + \frac{r_3}{2} (\mathbf{c}_{k+1}^\top \mathbf{M} \mathbf{c}_{k+1} - 1)$  ▷ KKT-multiplier update
         $\mu \leftarrow \max(\mu, 0)$  ▷ KKT-projection (dual feasibility)
         $Error \leftarrow \|\mathbf{c}_{k+1} - \mathbf{c}_k\|_\infty$ 

    end while

    return  $\mathbf{c}_{k+1}$ 

end function

```

Algorithm 4: Bounded CCR Algorithm with Hard-thresholding

Require: $\mathbf{B}$ and $\mathbf{M}$ matrices, $\boldsymbol{\gamma}$ column-vector, λ .

```

function REGULARIZED_CCR( $\mathbf{B}, \mathbf{M}, \boldsymbol{\gamma}, \lambda, r_{11} = 0.1, r_{12} = 0.1, r_2 = 0.1, r_3 =$ 
5,  $tol = 1 \times 10^{-14}$ )

     $\boldsymbol{\Delta}_a \leftarrow 0$ 
     $\boldsymbol{\Delta}_b \leftarrow 0$ 
     $\mathbf{c}_0 \leftarrow \boldsymbol{\gamma}$ 
     $\mu \leftarrow 0$ 
     $\bar{\mathbf{c}} \leftarrow \mathbf{c}_0$ 

    while  $Error > tol$  do

         $\boldsymbol{\Delta}_a \leftarrow \frac{(\boldsymbol{\Delta}_a + \mathbf{B}\bar{\mathbf{c}})}{1+r_{11}}$  ▷ gradient-accumulators
         $\boldsymbol{\Delta}_b \leftarrow \frac{(\boldsymbol{\Delta}_b + \mu \mathbf{M}\bar{\mathbf{c}})}{1+r_{12}}$ 
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_k + r_{11}r_2\boldsymbol{\Delta}_a - r_{12}r_2\mu\boldsymbol{\Delta}_b$  ▷ gradient descent (primal)
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_{k+1}\mathbf{I}(\mathbf{c}_{k+1} > \sqrt{2\lambda r_2})$  ▷ single-sided hard-thresholding
         $\bar{\mathbf{c}} \leftarrow 2\mathbf{c}_{k+1} - \mathbf{c}_k$  ▷ extra gradient step
         $\mu \leftarrow \mu + \frac{r_3}{2}(\mathbf{c}_{k+1}^\top \mathbf{M}\mathbf{c}_{k+1} - 1)$  ▷ KKT-multiplier update
         $\mu \leftarrow \max(\mu, 0)$  ▷ KKT-projection (dual feasibility)
         $Error \leftarrow \|\mathbf{c}_{k+1} - \mathbf{c}_k\|_\infty$ 

    end while

    return  $\mathbf{c}_{k+1}$ 

end function

```

Algorithm 5: CCR Algorithm with Hard-thresholding

Require: $\mathbf{B}$ and $\mathbf{M}$ matrices, γ column-vector, λ .

```

function REGULARIZED_CCR( $\mathbf{B}, \mathbf{M}, \gamma, \lambda, r_{11} = 0.1, r_{12} = 0.1, r_2 = 0.1, r_3 =$ 
5,  $tol = 1 \times 10^{-14}$ )

     $\Delta_a \leftarrow 0$ 
     $\Delta_b \leftarrow 0$ 
     $\mathbf{c}_0 \leftarrow \gamma$ 
     $\mu \leftarrow 0$ 
     $\bar{\mathbf{c}} \leftarrow \mathbf{c}_0$ 

    while  $Error > tol$  do

         $\Delta_a \leftarrow \frac{(\Delta_a + \mathbf{B}\bar{\mathbf{c}})}{1 + r_{11}}$  ▷ gradient-accumulators
         $\Delta_b \leftarrow \frac{(\Delta_b + \mu \mathbf{M}\bar{\mathbf{c}})}{1 + r_{12}}$ 
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_k + r_{11}r_2\Delta_a - r_{12}r_2\mu\Delta_b$  ▷ gradient descent (primal)
         $\mathbf{c}_{k+1} \leftarrow \mathbf{c}_{k+1} \mathbf{I}(|\mathbf{c}_{k+1}| > \sqrt{2\lambda r_2})$  ▷ hard-thresholding
         $\bar{\mathbf{c}} \leftarrow 2\mathbf{c}_{k+1} - \mathbf{c}_k$  ▷ extra gradient step
         $\mu \leftarrow \mu + \frac{r_3}{2} (\mathbf{c}_{k+1}^\top \mathbf{M} \mathbf{c}_{k+1} - 1)$  ▷ KKT-multiplier update
         $\mu \leftarrow \max(\mu, 0)$  ▷ KKT-projection (dual feasibility)
         $Error \leftarrow \|\mathbf{c}_{k+1} - \mathbf{c}_k\|_\infty$ 

    end while

    return  $\mathbf{c}_{k+1}$ 

end function

```

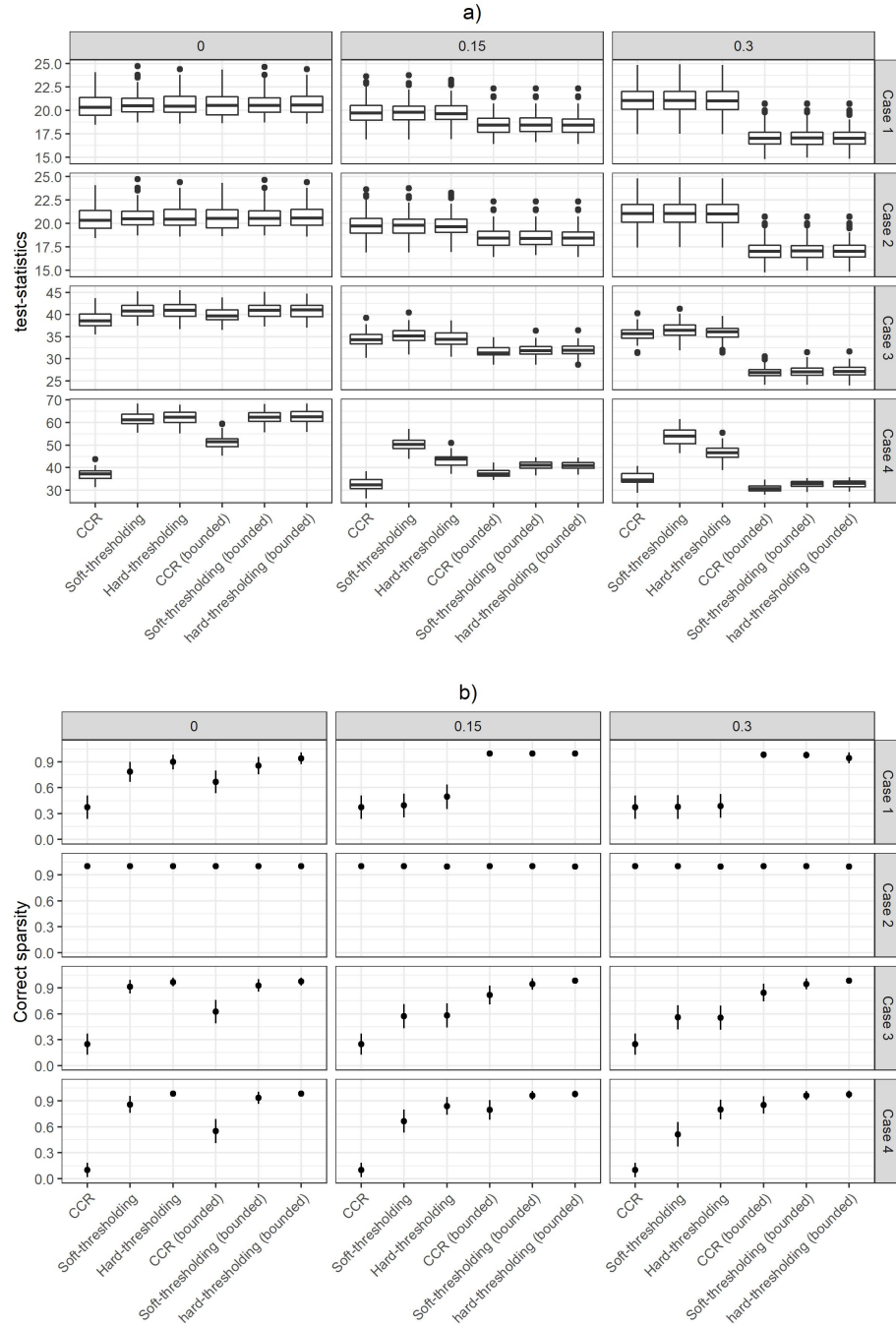


Figure 3.1: Results from simulation study. a) For each case (rows) and pairwise correlation considered (columns), t -statistics from the independent test dataset for each of the variations of the CCR. b) For each case (rows) and pairwise correlation considered (columns), the mean estimate of correct sparsity, with approximate 95% Wald confidence intervals.

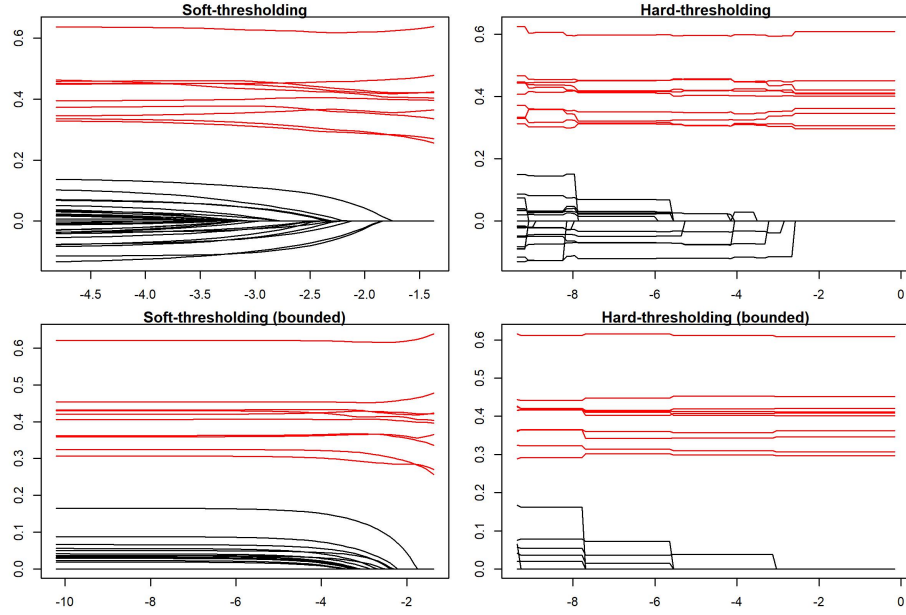
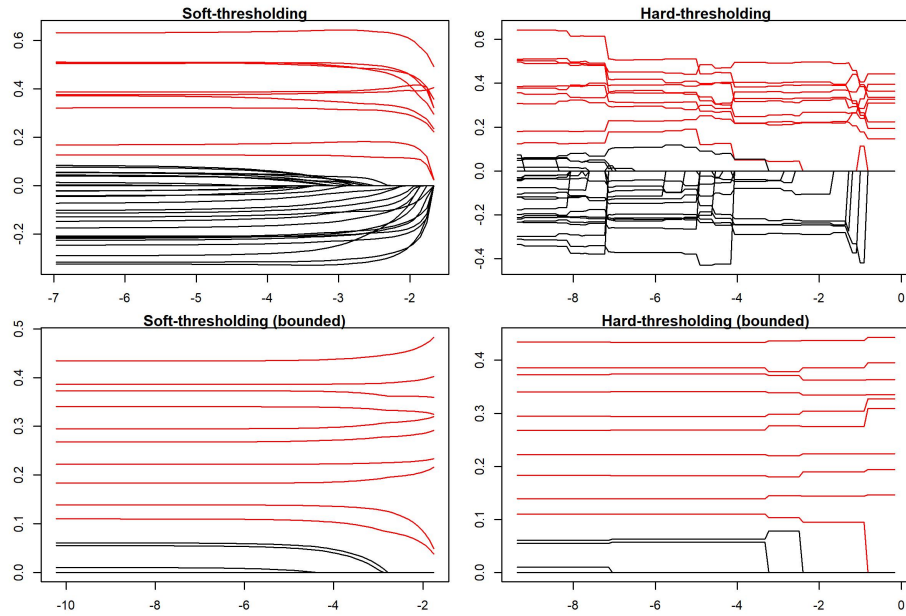
a) $\rho = 0$ b) $\rho = 0.3$ 

Figure 3.2: Solution paths from Case 3. With pairwise correlation set to 0 (a) and 0.3 (b), values of c_j are plotted versus $\log(\lambda)$ for unbounded CCR with soft-thresholding (top-left) and hard-thresholding (top-right) and B-CCR with soft-thresholding (bottom-left) and hard-thresholding (bottom-right).

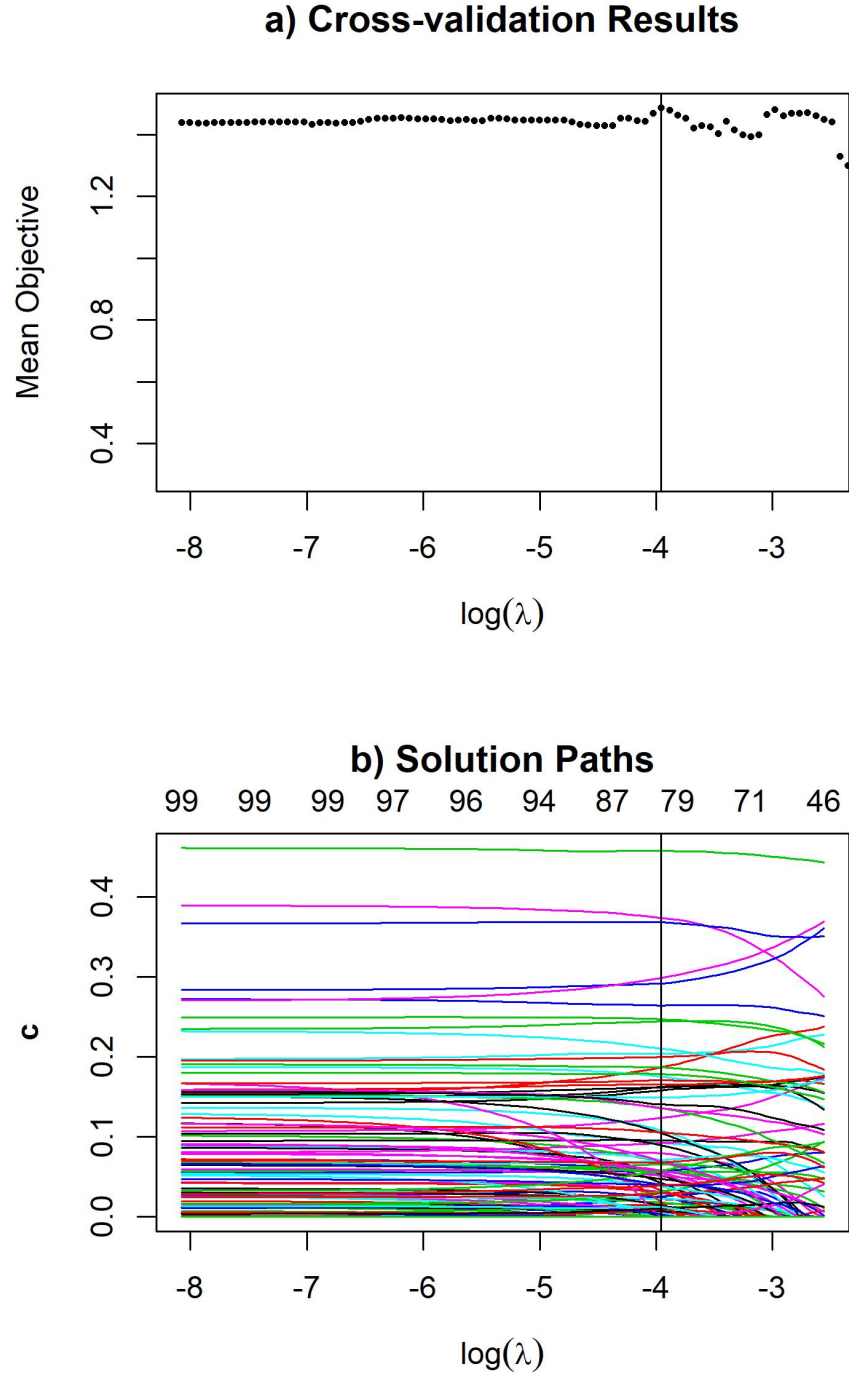


Figure 3.3: a) Mean cross-validated objective function estimates. Estimates are plotted versus $\log(\lambda)$ for bounded CCR with soft-thresholding. b) Solution paths for motivating example. Values of c_j are plotted versus $\log(\lambda)$ for bounded CCR with soft-thresholding, with optimal $\log(\lambda^*)$ providing 82 non-zero c_j 's.

Table 3.1: Results from simulation study. Mean t -statistics (and standard errors) from each of the considered CCR projections across all cases considered.

Case	ρ	CCR	l_1	l_0	B-CCR	B- l_1	B- l_0
Case 1	0.00	20.59 (1.42)	20.75 (1.44)	20.79 (1.45)	20.71 (1.45)	20.76 (1.44)	20.79 (1.44)
	0.15	19.92 (1.47)	19.91 (1.48)	19.87 (1.47)	18.60 (1.39)	18.62 (1.36)	18.58 (1.38)
	0.30	21.10 (1.54)	21.09 (1.55)	21.09 (1.55)	17.23 (1.34)	17.27 (1.31)	17.27 (1.30)
Case 2	0.00	20.59 (1.42)	20.75 (1.44)	20.79 (1.45)	20.71 (1.45)	20.76 (1.44)	20.79 (1.44)
	0.15	19.92 (1.47)	19.91 (1.48)	19.87 (1.47)	18.60 (1.39)	18.62 (1.36)	18.58 (1.38)
	0.30	21.10 (1.54)	21.09 (1.55)	21.09 (1.55)	17.23 (1.34)	17.27 (1.31)	17.27 (1.30)
Case 3	0.00	38.72 (1.65)	40.95 (1.70)	40.99 (1.82)	39.94 (1.53)	40.96 (1.72)	40.99 (1.77)
	0.15	34.50 (1.75)	35.27 (1.84)	34.42 (1.82)	31.59 (1.44)	31.93 (1.55)	32.06 (1.58)
	0.30	35.60 (1.76)	36.37 (1.89)	35.91 (1.79)	26.98 (1.31)	27.20 (1.41)	27.29 (1.42)
Case 4	0.00	36.95 (2.51)	61.61 (3.06)	62.33 (3.07)	51.47 (3.06)	62.34 (2.92)	62.53 (3.05)
	0.15	32.62 (2.65)	50.39 (2.79)	43.15 (2.84)	37.63 (1.75)	40.88 (1.78)	40.98 (1.81)
	0.30	35.00 (2.83)	53.74 (4.01)	46.85 (3.43)	30.86 (1.53)	32.73 (1.48)	32.79 (1.52)

CHAPTER FOUR

MOLECULAR-BASED DIAGNOSIS OF MULTIPLE SCLEROSIS AND ITS
PROGRESSIVE STAGEContribution of Authors and Co-Authors

Author: Christopher R. Barbour

Contributions: Responsible for the majority of the statistical writing and conclusions, methodological developments, and implementation of the additional simulation study included that was not in the original manuscript.

Co-Author: Peter Kosa

Contributions: Responsible for the majority of the biological writing and conclusions and implementation of the study design.

Co-Author: Mika Komori

Contributions: Data acquisition.

Co-Author: Makoto Tanigaw

Contributions: Data acquisition.

Co-Author: Ruturaj Masvekar

Contributions: Data acquisition.

Co-Author: Tianxia Wu

Contributions: Provided analysis of demographic data.

Co-Author: Kory Johnson

Contributions: Provided feedback on statistical analysis and drafts of the manuscripts.

Co-Author: Panagiotis Douvaras

Contributions: Data acquisition.

Co-Author: Valentina Fossati

Contributions: Data acquisition.

Co-Author: Ronald Herbst

Contributions: Data acquisition.

Co-Author: Yue Wang

Contributions: Data acquisition.

Co-Author: Keith Tan

Contributions: Data acquisition.

Co-Author: Mark Greenwood

Contributions: Provided feedback on statistical analysis and drafts of the manuscripts.

Co-Author: Bibiana Bielekova

Contributions: Responsible for conceiving study and writing the manuscript.

Manuscript Information Page

Authors: Christopher Barbour, Peter Kosa, Mika Komori, Makoto Tanigaw, Ruturaj Masvekar, Tianxia Wu, Kory Johnson, Panagiotis Douvaras, Valentina Fossati, Ronald Herbst, Yue Wang, Keith Tan, Mark Greenwood and Bibiana Bielekova

Journal: *Annals of Neurology*. Note that minor edits have been made to the published version (ordering of figures, etc.) for inclusion here. The appendix is new content for the dissertation.

Status of Manuscript:

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☒ Published in a peer-reviewed journal

Publisher: Wiley

Submitted February 4th, 2017

Accepted for publication October 17th, 2017

Volume 82, 795 – 812.

DOI: 10.1002/ana.25083

Abstract

Objective: Biomarkers aid diagnosis, allow inexpensive screening of therapies and guide selection of patient-specific therapeutic regimens in most internal medicine disciplines. In contrast, neurology lacks validated measurements of the physiological status, or dysfunction(s) of cells of the central nervous system (CNS). Accordingly, patients with chronic neurological diseases are often treated with a single disease-modifying therapy without understanding patient-specific drivers of disability.

Therefore, using multiple sclerosis (MS) as an example of a complex polygenic neurological disease, we sought to determine if cerebrospinal fluid (CSF) biomarkers are intra-individually stable, cell type-, disease- and/or process-specific and responsive to therapeutic intervention.

Methods: We used statistical learning in a modeling cohort (n=225) to develop diagnostic classifiers from DNA-aptamer-based measurements of 1128 CSF proteins. An independent validation cohort (n=85) assessed the reliability of derived classifiers. The biological interpretation resulted from in-vitro modeling of primary or stem cell-derived human CNS cells and cell lines.

Results: The classifier that differentiates MS from CNS diseases that mimic MS clinically, pathophysiologically and on imaging, achieved a validated area under receiver-operator characteristic curve (AUROC) of 0.98, while the classifier that differentiates relapsing-remitting from progressive MS achieved a validated AUROC of 0.91. No classifiers could differentiate primary- from secondary-progressive MS better than random guessing. Treatment-induced changes in biomarkers greatly exceeded intra-individual- and technical variabilities of the assay.

Interpretation: CNS biological processes reflected by CSF biomarkers are robust, stable, and disease- or even disease-stage specific. This opens opportunities for broad utilization of CSF biomarkers in drug development and precision medicine for CNS disorders.

4.1 Introduction

Biomarkers play a critical role in diagnostic and therapeutic decisions in many areas of internal medicine. Cell specific analytes (such as liver function tests) provide essential information about functionality in their cells of origin and represent the basis of molecular diagnosis. Molecular dissection of complex disorders allows selection of optimal, individualized therapy. Such “precision” therapy consists of simultaneous

application of (multiple) drugs that collectively target all pathological processes that underlie expression of a disease in particular patient.

In contrast, neurologists lack tools that provide reliable information about the dysfunction of constituent cells of the CNS. This ambiguity leads to 20-40% diagnostic errors (Koga et al., 2015, Rizzo et al., 2016), slow therapeutic progress (Kola and Landis, 2004) and suboptimal clinical outcomes. Complex neurological disorders such as multiple sclerosis (MS) are generally treated by a single disease modifying treatment (DMT), without understanding patient-specific drivers of disability. The multiplicity of mechanisms in neurodegenerative diseases and heterogeneity within patient populations makes successful treatment by a single therapy unlikely. Conversely, proving clinical efficacy of a single therapy is difficult precisely because of limited contribution of the targeted mechanism to the overall disease process.

Thus, reliable quantification of diverse pathogenic processes in the CNS of living subjects is a prerequisite for broad therapeutic progress in neurology. Although cerebrospinal fluid (CSF), an outflow for CNS interstitial fluid (Johanson et al., 2008) is an ideal source for molecular biomarkers, remarkably few CSF biomarkers have reached clinical practice or drug development (Vlassenko et al., 2016). This reality is partly based on a circular argument: CSF examinations are not implemented in clinical trials or clinics because of a lack of validated, commercially-available biomarker measurements, while reliable data on surrogacy of biomarkers to clinical outcomes can be obtained only from clinical trials or wide clinical use.

Consequently, the goal of this proof-of-concept study was to investigate on the example of MS the following hypotheses: 1. A subset of CSF biomarkers are intra-individually stable in the absence of disease process or therapeutic intervention, and such biomarkers can be assembled into clinically useful tests; 2. A subgroup of CSF

biomarkers have restricted cellular origin and can be used to develop clinically-useful classifiers; 3. Healthy and different disease states of the CNS are sufficiently dissimilar on a molecular level that CSF biomarker-based classifiers can differentiate a specific disease from those that have similar clinical phenotype, pathophysiology, or imaging features; 4. CSF biomarker-based classifiers can also quantify evolution of a single disease process, thus differentiating its stages; and 5. Therapy-induced changes in CSF biomarkers can be readily distinguished from intra-individual variability, demonstrating that CSF biomarkers could serve as pharmacodynamic markers in drug development.

4.2 Methods

4.2.1 Subjects

Subjects were prospectively recruited (5/2009-3/2015) as part of a Natural History protocol “Comprehensive Multimodal Analysis of Neurommunological Diseases of the Central Nervous System” (ClinicalTrials.gov Identifier: NCT00794352). The patients’ eligibility criteria included age 18-75 years and presentation with a clinical syndrome consistent with immune-mediated CNS disorder, or neuroimaging consistent with inflammatory or demyelinating CNS disease. The inclusion criteria for healthy donors (HD) were age 18-75 years and vital signs within normal range at the time of the screening visit. The diagnostic workup included a neurological exam, MRI of the brain and laboratory tests (blood, CSF) as described (Komori et al., 2015). Diagnoses of relapsing-remitting MS (RRMS), primary progressive MS (PPMS) and secondary progressive MS (SPMS) were based on 2010 revised McDonald diagnostic criteria (Polman et al., 2011). The remaining subjects were classified as either other inflammatory neurological disorders (OIND; e.g., meningitis/encephalitis, Susac’s Syndrome, CNS vasculitis, Systemic Lupus Erythematosus and genetic immunodef-

ciencies with CNS inflammation) or non-inflammatory neurological disorders (NIND; e.g., epilepsy, vascular/ischemic disorders, leukodystrophy) based on the evidence of intrathecal inflammation as published (Komori et al., 2015, Bielekova et al., 2012). The final clinical diagnostic classification was based on longitudinal follow-up, but reached prior to development of SOMAScan-based diagnostic classifiers. The vast majority of subjects (with few OIND exceptions described elsewhere (Komori et al., 2015)) were not treated by any disease-modifying treatments (DMT) at the time of CSF collection.

Clinical information of the validation cohort (n=85) were not available to developers of the diagnostic classifiers, while the results of the molecular diagnostic tests were not available to clinicians determining diagnoses.

4.2.2 CSF collection and processing

CSF was collected on ice and processed according to a written standard operating procedure. Research CSF aliquots were assigned prospective alpha-numeric codes, and centrifuged (335 g for 10 minutes at 4°C) within 15 minutes of collection. The CSF supernatant was aliquoted and stored in polypropylene tubes at -80°C until use.

4.2.3 SOMAScan

SOMAScan (SomaLogic, Inc., Boulder, CO) is a relative quantification of 1128 proteins (i.e., SOMAScan version available between 6/2012 and 10/2016; (Rohloff et al., 2014)) or 1300 protein (i.e., SOMAScan version available after 10/2016) using single-stranded DNA molecules synthesized from chemically modified nucleotides (SOMAmers; Slow Off-rate Modified DNA Aptamers). Chemical modifications enhance affinity binding to specific proteins. SOMAmers play a dual role of protein affinity-binding reagents and a DNA sequence recognized by complementary DNA

probes. This enables quantification of individual protein concentration using a DNA concentration quantified by hybridization (Gold et al., 2012, Kraemer et al., 2011, Gold et al., 2010). The raw data (relative fluorescent units [RFU]) are normalized and calibrated: hybridization normalization uses a set of twelve hybridization controls and a common pooled calibrator corrects for plate-to-plate variation. SOMAScan focuses on secreted soluble proteins, using a single discovery platform for all research applications.

4.2.4 Assessment of cellular origin of tested biomarkers from a subset of human immune cells and CNS cells

Fresh peripheral blood mononuclear cells (PBMC) of two HD were obtained from Ficoll gradient-treated lymphocytopheresis samples. Monocytes, B-cells, CD4+ T cells, CD8+ T cells, natural killer cells, innate lymphoid cells, and myeloid dendritic cells were Fluorescence-activated cell sorted (FACS). Each purified cell subtype was cultured (1×10^6 cells/ml) in serum-free X-VIVO 15 medium (Lonza, Walkersville, MD) with or without $10 \mu\text{g/mL}$ phorbol 12-myristate 13-acetate (PMA) and $1 \mu\text{M}$ Ionomycin. Supernatants were collected after 48 hours and frozen until use.

Isolated primary human CNS cells or cell lines: human neurons (ScienCell, Carlsbad, CA), human astrocytes (ScienCell), human brain endothelial cell line (HCMEC/D3; provided by Pierre-Olivier Couraud, PhD, INSERM, France (Weksler et al., 2005)), human microglia cell line (CHME5; provided by Nazira El-Hage, PhD, Florida International University, USA), and human choroid plexus epithelial cells (hCPEpiC; ScienCell) were plated (10^5 cells/ml; 10 ml/flask). Cells were treated with PBMC culture media (control) and an inflammatory mediators (supernatant from lipopolysaccharide- and CD3/CD28 beads-stimulated human PBMCs; 50% v/v). Oligodendrocytes were differentiated from the NIH approved human embryonic stem

cell line RUES1 using published protocol (Douvaras and Fossati, 2015). Cell-culture supernatants were collected after 24-hour incubation and frozen until use.

4.2.5 Measuring signal-to-noise ratio (SNR) in biomarkers

Differences in biomarker measurements in identical samples analyzed blindly at different times (n=29; 88 samples) quantified the technical variability. Similarly, differences in longitudinal samples of HDs (n=11; 24 samples) collected ~ 1 year apart quantified the biological variation. From each technical or biological replicate, we calculated the relative percent change for each SOMAmer as the difference in the repeated measures divided by the average of the replicates.

To constrain the number of biomarkers used for statistical learning, we calculated signal-to-noise ratios (SNR) using R statistical software (R Core Team, 2018) as exemplified in Figure 4.1. Briefly, we estimated the residual variance for each log-transformed biomarker from a linear mixed model (Pinheiro et al., 2016), after accounting for subject-to-subject variation using the random intercept adjustment. The residual variance measured when identical samples were analyzed repeatedly, represents “technical” variation (i.e., variation caused by differences in assay runs). Analogously, the variance measured in multiple different CSF samples derived from individual HDs represents “biological” variation (i.e., intra-individual variation in healthy state). A third type of variance, which we call “clinical”, reflects how individual biomarkers vary in the presence of different disease states (e.g., variation across all disease states). This variance was estimated by using one observation for each biomarker from each of the n=225 subjects in the training dataset. The SNR was calculated as the clinical variance divided by the sum of all (the clinical, biological, and technical) variances. Thus, SNR reflects the proportion of the total variability that is attributable to variation among subjects from different diagnostic

categories. In other words, biomarkers with high SNR (i.e., values close to 1) are biomarkers with large differences across subjects in the training cohort, but with low variation in longitudinal sampling of healthy people that can be detected with minimal variations across different assay runs. These markers have a high diagnostic potential. In contrast, biomarkers with SNR close to 0 represent proteins that are either not detectable in the CSF, or offer low diagnostic value because the physiological (intra-individual) variation or assay noise are comparable to the difference between diagnostic categories we desire to measure. An analogous SNR procedure restricted the number of biomarker ratios used for statistical learning.

4.2.6 Area under the Receiver Operating Characteristics Curve (AUROC)

The AUROC (R statistical software using the roc function in the pROC package (Robin et al., 2011)) quantified the ability of biomarkers, biomarker ratios, and diagnostic classifiers to differentiate diagnostic categories. Higher AUROC values imply a larger potential for separating diagnostic groups. The AUROCs were calculated for the 124,750 ratios formed from the top 500 SNR SOMAmers in the modeling cohort (n=225) and these were used to restrict the number of biomarker ratios used for statistical learning. Graphical exploration of the distributions of the SNR and AUROCs were used to determine cutoffs for the best ratios on the two criteria in each situation. For differentiating MS (RRMS, PPMS, SPMS) from non-MS subjects (OIND, NIND, HD), cutoffs of AUROC > 0.65 and SNR > 0.75 were selected. Similarly, cutoffs of AUROC > 0.7 and SNR > 0.7 were used in differentiating progressive (PPMS, SPMS) MS from RRMS. Finally, cutoffs of AUROC > 0.65 and SNR > 0.65 were used for differentiating SPMS from PPMS. Ratios meeting these cutoffs were used in statistical learning to assemble diagnostic classifiers.

The performance of resulting classifiers was assessed by AUROC, along with its 95% bootstrapped confidence interval (CI), in the independent validation cohort.

4.2.7 Statistical learning to develop diagnostic classifiers using a Random Forest methodology

Random forests (Hastie, Tibshirani, and Friedman, 2009) were built using the randomForest R package (Liaw and Wiener, 2002) by sequentially estimating multiple classification trees (between 500-1500; selected based on the stability of the out-of-bag [OOB] error) using bootstrapped training cohort samples (n=225) with a random subset of predictors for each node. The trees in the “forest” are averaged together, providing more reliable predictions than are possible using a single classification tree. Biomarkers and biomarker ratios were natural log-transformed prior to classifier construction. Variable importance measures (average decrease in accuracy from permuting each predictor across all trees) (Breiman, 2001, Friedman, 2001) evaluated the contribution of individual biomarkers to the classifier.

4.3 Results

4.3.1 SOMAscan Assay on CSF samples.

Using SOMAscan we analyzed CSF samples in blinded fashion in two independent cohorts of subjects: 1. A training cohort (n=225) consisted of untreated subjects from six diagnostic groups (RRMS, n=40; PPMS, n=40; SPMS, n=40; NIND, n=39; OIND, n=41; and HD, n=25). 2. An independent validation cohort (n=85 untreated subjects) consisted of 14 subjects per non-MS diagnostic categories (OIND and NIND), 16 subjects per RRMS and PPMS group, 15 SPMS subjects and 10 HD.

In addition, to assess the technical and biological variability of the SOMAscan, we analyzed 88 CSF samples representing technical replicates (identical CSF aliquots analyzed at different time-points) and 24 longitudinal CSF samples collected from 11 HD over one-year span, serving as biological replicates. The average technical and biological relative percent change (Figure 4.2) were 11.9% and 12.9%, respectively. To determine the effect of immunomodulatory DMT on SOMAmers, we analyzed 10 longitudinal CSF samples from five OIND patients collected before and after therapy with high dose methylprednisolone. The resulting average relative percent change of 52.1% exceeded the highest technical and biological relative percent change. Therefore, we concluded that the SOMAscan reliably measures CSF biomarkers and can detect an effect of DMT in subjects over time.

4.3.2 Development of diagnostic tests for MS

Considering the complex biological mechanisms underlying MS, a single biomarker cannot reliably differentiate MS from all other CNS diseases. Machine learning strategies (Hastie, Tibshirani, and Friedman, 2009), such as random forests (Liaw and Wiener, 2002), combine multiple biomarkers using a statistical algorithm that enhances sensitivity and specificity (Bielekova et al., 2014), resulting in clinically-useful classifiers.

We employed random forests trained using the modeling (n=225) cohort to develop three classifiers: 1. One that differentiates MS from all other diagnostic groups, 2. One that differentiates RRMS from progressive MS (i.e., PPMS+SPMS) and, 3. One that differentiates PPMS from SPMS. Because the validity of the classifier must be tested in an independent cohort not used for the model construction (Ioannidis, 2011), we assessed the performance of the diagnostic tests in the independent (n=85) cohort by predicting AUROCs and their CIs (Figure 4.3).

Classifiers that used all 1128 SOMAmers achieved validated AUROC=0.91 (CI: 0.84 – 0.97) for the MS versus non-MS test, AUROC=0.73 (CI: 0.57-0.90) for the RRMS versus progressive MS test and AUROC 0.64 (CI: 0.44-0.84) for SPMS versus PPMS test (Figure 4.3).

We expected that not all 1,128 proteins measured by SOMAscan will be detectable in the CSF. We filtered out noise stemming from the poorly-detectable biomarkers by restricting the number of SOMAmers to the 500 with the highest SNR (see Methods for details). We reasoned that the most useful biomarkers will vary greatly among subjects from different diagnostic categories, whereas they will have stable physiological levels (i.e., they will have low variance in biological replicates of HDs) and can be measured with high precision (i.e., low variance in technical replicates). This reduced set of 500 biomarkers improved performance of the classifiers: AUROC for the RRMS versus progressive MS classifier increased to 0.80 (CI: 0.65-0.95); the performance of the diagnostic test for MS marginally improved (AUROC=0.92, CI: 0.86-0.98), and remained unchanged for SPMS versus PPMS test (Figure 4.3).

While biomarkers are secreted by diverse cells under physiological or pathological states, many biomarkers are biologically related; e.g., they physically interact or belong to the same network. However, random forests consider biomarkers only sequentially within any given tree. For related biomarkers, such as receptor-ligand pairs, the pathogenic process may depend more on their stoichiometry than on their respective concentrations. Therefore, we hypothesized that considering related biomarkers simultaneously, for example as ratios, will add discriminatory value. Mathematically, this corresponds to broadening the biomarker-based random forests from partitioning the predictor space based on absolute concentrations of individual markers to considering predictors built from the relative proportion of biomarkers.

Consequently, we used the 500 high-SNR SOMAmers to generate 124,750 biomarker ratios. To build random forests only from ratios with highest clinical utility, we combined AUROC with SNR values in the modeling (n=225) cohort, selecting logical cut-offs described in Methods that allowed for enough diversity to capture the biological processes while limiting the dimension of the search space. This led to 5,401 retained ratios for MS versus non-MS, 3,626 retained ratios for progressive versus relapsing MS, and 1,504 retained ratios for SPMS versus PPMS. Using these sets of ratios strongly enhanced the performance of random forest models distinguishing MS from non-MS and RRMS from progressive MS (validated AUROC=0.95; CI: 0.91-0.99 and AUROC=0.88; CI: 0.76-1.00, respectively). However, the performance of SPMS versus PPMS diagnostic test remained low (AUROC=0.45; CI: 0.24-0.67).

A clinical test has to fulfill technical requirements of reproducible measurement across different laboratories, often achieved by using standard curves; this makes measuring several hundreds of proteins prohibitive. Therefore, we sought to identify the smallest number of biomarkers that can be assembled into the random forest classifiers without a significant loss of accuracy. To achieve this, we examined the out-of-bag (OOB) AUROC estimates from random forests generated from the modeling (n=225) cohort by sequentially adding ratios with the highest variable importance. The point of inflection where the OOB AUROC appeared to stabilize was selected to achieve models with high predictive ability at a lower complexity (Figure 4.3). Interestingly, the reduction of the number of ratios further improved the performance of the classifiers in the validation cohort: the 22 most important ratios in the MS vs non-MS classifier led to validated AUROC=0.98 (CI: 0.94-1.00) and the 21 most important ratios distinguished RRMS from progressive MS with AUROC=0.91 (CI: 0.80-1.00) (Figure 4.3). The 33 most important ratios distinguishing SPMS from PPMS led to a classifier with performance comparable to random guessing

(AUROC=0.58; CI: 0.37-0.79) and therefore this model was abandoned from further analyses.

In all three classifiers, variable importance measures were dominated by ratios in models that also included single SOMAmers, validating our mathematical and biological foundation of the ratio-based hypothesis.

4.3.3 Deconvolution of biomarkers' cell of origin

To investigate whether statistical learning preferentially selected biomarkers with restricted cellular origin into clinically-useful tests, we used in-vitro modeling on human primary immune and CNS cells (see Methods), complemented with data from the public domain, such as the RNA sequencing database of human CNS cells (Zhang et al., 2014, Hawrylycz et al., 2012) and The Human Protein Atlas (Uhlen et al., 2015) to assess cellular origin of the biomarkers in the optimized random forest classifiers (Figure 4.4 and Figure 4.5). We analyzed cell cultured media in resting state and upon activation: CNS cells/cell lines were exposed to supernatants from LPS- and CD3/CD28 beads-stimulated human PBMCs to mimic inflammatory conditions, while immune cells were activated by PMA/Ionomycin to achieve robust activation of all immune cells. Supernatants were analyzed by SOMAscan assay at time 0 and after 24-hour incubation/stimulation. Results are shown as the stimulation index - a ratio of RFUs at 24 hour and time 0 for each condition for 48 SOMAmers that form the two diagnostic classifiers. These results are the basis for the biological interpretation of the diagnostic classifiers.

4.3.4 Biological interpretation of MS versus non-MS diagnostic test

The 22 most important ratios of the MS diagnostic test were dominated by immune cell-specific biomarkers (Figure 4.4). Twenty-one of these contain plasma cell-specific biomarkers TNFRSF17 (BCMA) or IGG. The main insight from the

MS diagnostic classifier is that in MS, activation of humoral immunity represented by plasma cells and plasmablasts is out of proportion to activation of all other cellular components of innate or adaptive immunity. Combining cellular origin and known biological functions, the biomarkers constituting the MS diagnostic test can be divided into 9 subgroups, where plasma cell activation/levels is compared to: overall intrathecal inflammation (Figure 4.4; group 1a; PDCD1LG2, SLAMF6, CD48, CSF3, CXCL13, TNFRSF4) and amount of activation of myeloid lineage (Figure 4.4; groups 1b and 1c; PLA2G7, CCL7, TLR4 LY96, PRTN3). MS patients also show higher plasma cell activation/levels in comparison to vascular injury and/or ongoing CNS stress (Figure 4.4; group 1d, 2a, and 2c; FLT4, CDKN1B, TNFRSF6B, DSG2, CRK, PGK1, MAPK14, F9, DCTPP1). A ratio of IgG and IgM (Figure 4.4; group 2b) points to differences in immunoglobulin subtypes between MS and non-MS subjects. Higher plasma cell immunoglobulin secretion in comparison to levels of CNS injury and higher plasma cell activation in comparison to astrocyte activation (Figure 4.4; group 2d; TNC) was also observed in MS. Lastly, MS subjects have increased epithelial stress compared to activation of neutrophils (Figure 4.4; group 3; MMP7, PRTN3).

4.3.5 Biological interpretation of progressive MS versus RRMS diagnostic test

The 21 top ratios distinguish RRMS from progressive MS (Figure 4.5). Seven were ratios of EDA2R or EDAR with markers released predominantly by CNS cells, especially neurons and oligodendrocytes: STX1A, EPHA5, JAM3, NTRK3, RGMA, BOC, and UNC5 (Figure 4.5; group 1a); all of these ratios demonstrate relative loss of CNS-specific markers in progressive MS. Moreover, three ratios show proportional loss of neuronal and oligodendroglial markers in relation to ICOSLG, expressed on activated antigen-presenting cells, in progressive MS (Figure 4.5; group 1b; EPHA5, JAM3, TYRO3). Similarly, the ratio INHBA/JAM3 measures relative

loss of an oligodendroglial marker in comparison to a marker secreted predominantly by myeloid cells during wound healing and tissue remodeling (Figure 4.5; group 1c). Progressive MS patients also show increased epithelial stress in comparison to overall intrathecal inflammation (Figure 4.5; group 1d; SELL, EDAR). Another group of biomarker ratios points out enhanced alternative pathways of complement activation in comparison to overall intrathecal inflammation (Figure 4.5; groups 2a and 2b; CFD, GZMA, SELL, SERPING1, IL22). Finally, a group of ratios relates to dysregulation of LTA/LTB and IL22 pathways, which play an important role in the formation of tertiary lymphoid follicles in progressive MS (Figure 4.5; groups 2c, 2d, 2e, and 3; LILRB2, SELL, LTA LTB, IL10, PRTN3, ETHE1, GP6, CLEC1B) and also in platelet aggregation (Figure 4.5; group 3).

4.3.6 Classifier Properties

The clinical properties of the validated models are summarized in Figure 4.6. When using 50% cut-off to convert continuous probabilities into dichotomous classifiers, the MS molecular diagnostic test shows 87.2% sensitivity (CI: 77.7%-96.8%) and 94.7% specificity (CI: 87.6%-100.0%) with a diagnostic odds ratio of 123.0. The progressive MS classifier differentiates RRMS from progressive MS with 93.5% sensitivity (CI: 84.9%-100.0%) and 81.3% specificity (CI: 62.1%-100.0%), reaching a diagnostic odds ratio of 62.8.

4.4 Discussion

Exploring recent advances in proteomics, we asked whether CSF biomarkers can reliably measure intrathecal processes and thus facilitate diagnosis, drug development, and clinical management of patients with complex CNS diseases. We hypothesized and confirmed that intra-individually stable CSF biomarkers with restricted cellular

origin are over-represented in clinically-useful classifiers. Collected data provide proof-of-principle evidence that molecular diagnosis of polygenic CNS diseases is feasible with current technologies.

In contrast to internal medicine disciplines that utilize molecular biomarkers (Morgan et al., 2012), contemporary diagnostic process and therapeutic decisions for polygenic neurological diseases are based on clinical findings and structural imaging, both of which lack molecular specificity. This may contribute to high misclassification rates in neurodegenerative diseases against pathology (Koga et al., 2015, Rizzo et al., 2016). Indeed, finding that one out of the first three “MS” subjects who succumbed to natalizumab-induced progressive multifocal leukoencephalopathy demonstrated no pathological evidence of MS (Kleinschmidt-DeMasters and Tyler, 2005) suggests that a >20% misdiagnosis rate may also be applicable to MS, despite advances provided by imaging. While we observed approximately 10% discrepancy between clinical- and biomarker-based MS diagnosis (Figure 4.7A), the absence of pathological evidence prevents determining which classification is correct. The majority of SOMAscan-misclassified “MS” patients lacked defining biological features of MS: intrathecal activation of plasma cells and adaptive immunity, validated by alternative assays (Figure 4.7B). Therefore, patients misclassified by molecular classifiers either exhibited a non-inflammatory form of MS, observed by pathologists at frequencies analogous to our MS misclassification rate (Lucchinetti et al., 2000), or had alternative conditions. Regardless of what we call such ailments, these patients lack targets of immunomodulatory DMTs and are unlikely to reap their benefit. Thus, providing therapeutically-relevant information represents the first advantage of molecular diagnosis. The second advantage is reporting diagnostic probabilities as a continuous variable that captures the strength of biological evidence, in comparison to a dichotomous clinical diagnosis. Indeterminate results close to 50% probability

(which represented 88% of subjects with discrepant clinical and molecular diagnosis; Figure 4.7A) should be repeated, ideally after cessation of the acute process that prompted diagnostic testing.

Diagnostic specificity is of high clinical importance, because false positive results expose subjects to potential harms of unnecessary therapies. Because the MS diagnostic classifier is antigen-nonspecific, dysregulated immunity targeting non-MS antigens may be misclassified as MS if it elicits qualitatively similar intrathecal inflammation. Indeed, we observed that two OIND patients (one with CTLA-4 haploinsufficiency and another with chronic aseptic meningitis) were misclassified as MS. It is plausible that analogously to mutations shared among different cancers, conditions with pathogenic mechanisms similar to MS may respond to MS treatments. This is the third advantage of molecular taxonomy, as it could promote CNS therapeutics from disease-specific monotherapies to process-specific therapies, where treatments are shared among pathophysiologically-related conditions and rationally assembled into patient-specific polypharmacy regimens.

Our study has following limitations: 1. SOMAscan represents a selection of proteins that are not specifically targeted to the CNS. This drawback, however, also proves that molecular signatures of distinct diseases are sufficiently robust that sampling $\sim 1\%$ of the relevant proteome can reliably differentiate among them. Our observation that unbiased statistical learning selected virtually all available SOMAmers with restricted CNS cellular origin suggests that deliberate broadening of the sampled proteome to more CNS-relevant biomarkers has a potential to improve classifications, and expand understanding of disease mechanisms. 2. SOMAscan is a discovery platform, routinely optimized and expanded, and therefore lacking standards of clinical applications. We dealt with this problem by embedding many technical replicates that allowed normalization between different assay runs. However,

even after normalizing and focusing on biomarkers with high SNR, the inter-assay variability (measured by technical replicates) decreased the performance of the classifiers. Furthermore, during the submission and review of this article, Somalogic updated SOMAscan from an assay that quantifies 1128 proteins to an assay that uses different buffers, dilutions, and quantifies 1300 proteins. We subsequently used the original validation cohort to test whether random forests constructed from the selected set of biomarker ratios depicted in Figures 4.4 and 4.5 can still reliably differentiate MS from other diseases and RRMS from progressive MS using the new version of SOMAscan. We observed OOB AUROC of 0.89 (CI: 0.82-0.97) for MS versus non-MS classifier and OOB AUROC of 0.84 (CI: 0.67-1.00) for RRMS versus progressive MS classifier. This shows that our results are not assay-dependent, but reflect true biological processes. Nevertheless, biomarker-based precision neurology cannot be achieved without the biotech industry, which needs to develop fully quantitative, CSF-targeted assays that conform to technical requirements of clinical tests. To facilitate this, we considered “assay economy”, as an optimum between assay cost (dependent on the number of proteins that need absolute and relative quantification) and accuracy. Biomarker ratios simplify assay commercialization, by limiting the need to run standard curves for every analyte and providing internal normalization that avoids false positive results caused by, e.g., high protein levels. However, absolute quantification of at least the dominant biomarker partners, such as BCMA and EDA2R will likely be necessary for quality assurance for clinical applications.

The creative use of SNR screened out biomarkers of low clinical value; this improved the efficacy of the random forest algorithm, which searched for optimal biomarkers in the lower dimensional search space. Although such use of external data and domain-expert knowledge is encouraged in statistical learning as it typically improves performance (as it did in our case), we acknowledge that this methodology

leads to some arbitrariness in the selection of markers and thus it may not work in other settings or for other cohorts. To assure adequate representation of all MS subtypes and both inflammatory and non-inflammatory MS mimics in the classifier construction, we designed this study to have approximately equal representation from all patient groups. This may not be representative of the rates in the real population of patients, where e.g., RRMS/SPMS patients are much more frequently encountered than PPMS subjects. Because SNR is dependent on the composition of the training cohort, we acknowledge that different compositions of the training cohort may lead to selection of different biomarkers, potentially better or worse for separating certain disease states. This behavior is inherent to any statistical learning process and, therefore, sampling and population structure must be considered carefully in the study design. We view our population selection as appropriate for the stated goals, because in addition to healthy donors, controls included subjects with varied inflammatory and non-inflammatory CNS diseases, who presented for the diagnostic work-up of MS or related neuroimmunological diseases and who must be differentiated from all three MS subtypes.

Supporting the notion that CSF biomarkers can expand understanding of CNS diseases, the following knowledge was gained from the current study: The essential difference between MS and its mimics is selective expansion/activation of B cell/plasma-cell lineages, out of proportion to the activation of other immune cells and to the resultant injury/stress of CNS-resident cells. An ancillary pathway that helps to diagnose MS is linked to a marker of tissue remodeling and repair, MMP7. These features are shared by all MS subtypes, indicating that PPMS is not a pathophysiologically-distinct “non-inflammatory” entity (Stys et al., 2012), but rather an equivalent disease stage to SPMS. This conclusion is supported by the observed inability to validate a molecular classifier that differentiates PPMS from

SPMS with accuracy higher than random guessing and by therapeutic response of PPMS to immunomodulation by ocrelizumab (Montalban et al., 2016).

The dominance of plasma cell biomarkers in the molecular classifier poses a question of its value against current CSF tests such as IgG index and oligoclonal bands (OCB). We have included IgG index and MS classifier prediction rates to Figure 4.7C to demonstrate the superiority of the classifier. Similar data were obtained for OCB; in the cohort of patients with available OCB data the sensitivity (93.9%; CI: 90.3% - 97.6%) of OCB test was comparable to the sensitivity of molecular classifier (96.4%; CI: 93.5% - 99.2%), the specificity of the OCB test (80.0%; CI: 72.8% - 87.2%) was highly outperformed by the specificity of the molecular classifier (98.3%; CI: 96.0% - 100.0%).

Statistical learning also enhanced our understanding of progressive MS by demonstrating that PPMS and SPMS are biologically indistinguishable. These data argue for merging PPMS and SPMS cohorts in future drug development and clinical considerations. Features that differentiate progressive MS from RRMS are greater CNS tissue destruction, including more wide-spread endothelial/epithelial cell stress and reactive gliosis with increased permeability of CNS barriers and greater activation of innate immunity. In addition to proportional loss of oligodendroglial and neuronal biomarkers that likely reflect injury or loss of their cells of origin, there are also immunological differences between RRMS and progressive MS. These relate to innate immunity (complement, myeloid lineage, and antigen presentation), and to pathways involved in the formation of tertiary lymphoid follicles, such as lymphotoxin complex and IL-22. This is consistent with the pathological evidence of tertiary lymphoid follicles in progressive MS (Magliozzi et al., 2007) and with a recent report that the level of compartmentalization of immune responses to the CNS can differentiate RRMS from two progressive MS subtypes (Komori et al., 2015).

Finally, it is intriguing that 4 out of 21 ratios that differentiate progressive MS from RRMS (i.e., containing SERPING1 and CFD, which is essential in alternative complement activation by cleaving C3) are linked to “neurotoxic reactive astrocytes”, recently shown to mediate neuronal death in MS and other neurodegenerative diseases (Liddel et al., 2017).

One may ask to what degree the identified MS progression-specific processes reflect aging. Re-analyzing probabilities of progressive MS in patients younger and older than 45 years demonstrated that the molecular classifier correctly differentiates RRMS from progressive MS irrespective of age. Thus, the biological interpretation of MS classifiers offers the following unifying hypothesis for future longitudinal studies: while aberrant activation of B/plasma cell lineage is essential for development of MS, the complex response of CNS tissue, exemplified by microglial activation, toxic astrogliosis and endothelial/epithelial stress, determines the extent and irreversibility of demyelination and neuronal death, which underlie progressive accumulation of disability in MS.

Although longitudinal data represented only a small part of the current study, they were instrumental for selecting high SNR biomarkers, which improved the accuracy of the molecular classifiers. They also demonstrated the ability of CSF biomarkers to measure broad biological effects of applied therapies in the intrathecal compartment. Expanding CSF biomarker studies to longitudinal cohorts could identify molecular signatures that forecast therapeutic efficacy, as well as biological synergisms among different treatments. Longitudinal cohorts are also required to determine the extent and stability of pathogenic heterogeneity (Lucchinetti et al., 2000). Implementation of CSF biomarkers to Phase I/II trials can guide dose and patient selection, and eliminate unpromising agents without accruing excessive costs and sequestering large numbers of available patients (Komori et al., 2016). Such

biomarker-supported trials thus offer the promise to propel CSF-biomarker-based precision-medicine into neurology practice. While the presented results make these prospects realistic, they cannot be achieved without broader, visionary investment of efforts and resources to exploit the full potential of CSF biomarkers in neurology.

4.5 Acknowledgments:

We thank Dr. Pierre-Olivier Courad (Institute Cochin, INSERM, France) for providing the HCMEC/D3 cell line. We thank Elena Romm for processing of CSF samples. We thank clinicians Peter Williamson, Anil Panackal, Alison Wichman, Jamie Cherup, Irene Cortese, Joan Ohayon, Kaylan Fenton, Camilo Toro, Dennis Landis, Adeline Vanderver, Elisabeth Wells, Carlos Pardo, Lauren Krupp and research nurse Jenifer Dwyer for expert patient care, regulatory nurse Rosalind Hayden for help with regulatory paperwork and schedulers Anne Mayfield and Kewounie Pumphrey for patient scheduling. We thank Dragan Maric for help with cell sorting. Finally, we thank the patients, their caregivers and healthy volunteers, without whom this work could not be possible. The study was supported by the intramural research program of the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH) and the Material Transfer Agreement (MTA) between NINDS and Medimmune, LLC (A member of the AstraZeneca Group) that partially funded the SOMAscan assays. P.K. received fellowship support from the Myelin Research Foundation and M.K. received post-doctoral fellowship support from the Japan Society of the Promotion of Science.

4.6 Author contributions:

Study concept and design: B.B.; data acquisition and analysis: C.B., P.K., M.K., M.T., R.M., T.W., K.J., P.D., V.F., R.H., Y.W., K.T., M.G. and B.B.; drafting the manuscript and figures: C.B., P.K. and B.B.

4.7 Potential conflicts of interest:

BB, PK, MK, CB, and MG are co-inventors of US patent application number 62/038,530: Biomarkers for Diagnosis and Management of Neuro-immunological Diseases, which pertains to the results of this paper. BB, PK, and MK have assigned their patent rights to the US Department of Health and Human Services.

4.8 Supplementary Material

4.8.1 Developing Signal-to-noise Metric

To calculate the required variance components, the following procedure was performed for each of the biomarkers:

For biological variability, the data set used consists of the biological replicates that came from a single run (8 samples from the original validation cohort, 2 samples for each of 4 healthy donors). Then using natural $\log(\text{marker})$ as a response, a linear mixed model (Pinheiro et al., 2016) was fit that contained a constant mean, μ , and a random effect for subject j , $Patient_j$. There is also a residual error, ϵ_{ij} , associated with observations i within each subject j :

$$\log(\text{marker}_{ij}) = \mu + Patient_j + \epsilon_{ij},$$

$$Patient_j \stackrel{iid}{\sim} N(0, \sigma_{patient}^2),$$

$$\epsilon_{ij} \stackrel{iid}{\sim} (0, \sigma_{bio}^2), \text{ and}$$

$$\epsilon_{ij} \perp Patient_j.$$

The estimated residual variance from this model ($\hat{\sigma}_{bio}^2$) represents the estimated variability in the responses within each subject (controlling for the subject-to-subject variability), and we used this as an estimate of the biological variation. Note that all $Patient_j$ and ϵ_{ij} are assumed to be independent in this model.

For technical variability, the data set used consists of technical replicates after patients in our testing data set were removed (66 measurements, representing 31 unique samples). Then using $\log(\text{marker})$ as a response, a linear mixed model was fit that used a constant mean, μ , and a random effect for sample j , $Sample_j$. The

residual error, ϵ_{ij} , is then associated with observations i within each sample j :

$$\log(\text{marker}_{ij}) = \mu + \text{Sample}_j + \epsilon_{ij},$$

$$\text{Sample}_j \stackrel{iid}{\sim} N(0, \sigma_{\text{sample}}^2),$$

$$\epsilon_{ij} \stackrel{iid}{\sim} (0, \sigma_{\text{tech}}^2), \text{ and}$$

$$\epsilon_{ij} \perp \text{Sample}_j.$$

The estimated residual variance from this model ($\hat{\sigma}_{\text{tech}}^2$) represents the estimated variability in the responses within each sample (controlling for the sample-to-sample variability), and this was used as an estimate of the technical variation. Note that all Sample_j and ϵ_{ij} are assumed to be independent in this model.

For clinical variability (or variability in the marker across our represented population), the data set used consists of the 225 patients that are used to train the classifier. These patients have varying disease states, representing healthy individuals, patients with varying degrees of MS, and patients with other inflammatory and non-inflammatory neurological diseases. Using $\log(\text{marker})$ as a response, a linear model was fit that contained a constant mean, μ ,

$$\log(\text{marker}_{ij}) = \mu + \epsilon_{ij} \text{ with}$$

$$\epsilon_{ij} \stackrel{iid}{\sim} (0, \sigma_{\text{clinical}}^2).$$

The estimated residual variance from this model ($\hat{\sigma}_{\text{clinical}}^2$) represents the variability in the measurements in our patients, and we used this as an estimate of the clinical variation.

In order to construct a single metric that quantifies a marker's utility for

diagnostic purposes, we constructed a single SNR measure:

$$SNR = \frac{\hat{\sigma}_{clinical}^2}{\hat{\sigma}_{clinical}^2 + \hat{\sigma}_{tech}^2 + \hat{\sigma}_{bio}^2},$$

where values close to 1 indicate a marker with a high potential for clinical utility. We explored the SNR of the log(biomarkers) and made decisions based on retaining a proportion of the top (larger) SNR biomarkers for constructing classifiers.

4.8.2 Simulation Study

The signal-to-noise metrics were used to reduce the dimensionality of the predictor space, leading to performance gains. These calculated metrics represent a form of screening based on data quality, or the level of contamination from unwanted sources of variability. Other forms of variable screening commonly used in medical studies include screening based on marginal relationships with a response (such as an area under the curve, correlation coefficient, or t -test), screening based on raw variance calculations, or no screening.

Screening based on univariate test-statistics is useful when the relationship between the predictors and the response are marginal and additive (i.e., no interactions between pairs of variables). In these instances, the test-statistic accurately captures the relationship between the predictors and the response (i.e., if a Pearson correlation is used, there is not a nonlinear relationship). This is an unrealistic expectation in gene-expression data, where interactions between pairs of biomarkers in the same pathway and complicated nonlinear relationships with the response and between pathways are likely to occur.

While no-screening allows all information present to be used in the model fitting, it can lead to difficulties when interpreting results. This is particularly difficult in gene-expression studies and other high-dimensional settings, when visual examination

of all the predictors' relationship with the response is unfeasible.

Screening based on variance calculations is common when analyzing medical data (e.g., Witten and Tibshirani, 2010, Tan et al., 2014), and is based on the assumption that biomarkers that are associated with disease diagnosis (or severity) will have a larger spread of values when examined with a group of controls. This is similar to the clinical variance that was reported in this manuscript. While this assumption is reasonable in many instances, it ignores other sources of variation that are not related to the response of interest. In this manuscript, we identified two such sources of error, namely the technical variation from the same sample on different assay runs and the natural biological variation in healthy individuals. Our SNR metric represented a form of data quality, where higher values represent variables where the sources of unwanted variability were small in comparison to the observed variability. If this degree of contamination is large relative to the measurable signal, then screening based on data quality is likely to perform better than screening on raw variance alone.

To explore the properties of each of these screening procedures, a simulation study was performed. The simulation study considered that four variable types were present; variables that were associated with the response variable (signal) and were measured with high precision (high-quality), variables that were associated with the response variable and were measured with low precision (low-quality), variables that were not associated with the response (noise) and were measured with high precision (high-quality), and variables that were not associated with the response (noise) and were measured with low precision (low-quality). The proportion of these different variable types is represented by

$$\mathbf{p} = [p_{\text{signal,high-quality}}, p_{\text{signal,low-quality}}, p_{\text{noise,high-quality}}, p_{\text{noise,low-quality}}].$$

Let $\mathbf{D} = \text{diag} \left(\overbrace{a, a, \dots, a}^{n_{\text{signal}}}, \overbrace{b, b, \dots, b}^{n_{\text{noise}}} \right)$ be a diagonal matrix of the variable standard deviations, where $a = 0.1$ and $b = 0.05$. This is done because biomarkers that are associated with a disease state of interest tend to have more variability than those that do not. Let $p = n_{\text{signal}} + n_{\text{noise}}$ be the total number of predictor variables and $\mathbf{R}_{p \times p} = \mathbf{I}_{20 \times 20} \otimes \mathbf{C}$ be the correlations between pairs of variables, where $\mathbf{C}$ is a $\frac{p}{20} \times \frac{p}{20}$ matrix with 1 on the diagonal and ρ in the off-diagonal and $\otimes$ is the Kronecker product. This creates 20 clusters of variables that share a common correlation ρ , with independence among clusters. Then $\mathbf{\Sigma} = \mathbf{D}\mathbf{R}\mathbf{D}$ is the variance-covariance matrix of our predictors. For each observation i ($i = 1, \dots, n$ where n is the total sample size), the variables, $\mathbf{X}_i$, were simulated from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$. The first n_{signal} entries, $\mathbf{X}_i^s$, represent the signal variables and the last n_{noise} entries, $\mathbf{X}_i^n$, represent the noise variables.

The relationship between the signal variables and the response is

$$y_i = f(\mathbf{X}_i^s) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

where

$$f(\mathbf{x}) = \sum_{j=1}^{N_{sig}-1} 100 (x_{j+1} - x_j^2)^2 + (1 - x_j)^2$$

is the generalized Rosenbrock equation (Rosenbrock, 1960). The measure of data quality was determined for each variable j by

$$p_j \stackrel{iid}{\sim} Unif(0, 0.1) I(j = low) + Unif(0.9, 1) I(j = high)$$

and the level of contamination for each variable within each patient was determined

by

$$\gamma_{ij} \stackrel{iid}{\sim} N(0, (1 - p_j) \sigma_g^2).$$

The “observed” data, $\mathbf{O}$, to be used to construct a predictive model was constructed by

$$O_{ij} = X_{ij} + \gamma_{ij}.$$

Therefore, variables that were lower-quality are observed to have an extra source of variation. The contaminated variables were used to predict the response y .

Using this data-generating model, the error standard deviation, σ , was fixed to 0.1, the total sample size, n , was fixed at 200, and the proportion of variables retained after screening (proportion kept) was set at 0.25. The proportions of different data types ($\mathbf{p}$), the number of variables (p), the pairwise correlation within clusters (ρ), and the data contamination standard deviation (σ_γ) were varied according to Table 4.1 and Table 4.2. For every $\mathbf{p}$ in Table 4.1, all combinations of parameters in Table 4.2 were implemented.

Under each parameter scenario, a training dataset was drawn from the generating model. Using the observed data ($\mathbf{O}$), variables were screened using four methods; screening based on variance of $\mathbf{O}_j$, screening based on univariate Spearman correlations of $(y, \mathbf{O}_j)$, screening based on p_j , and no-screening (all variables retained). The number of each variable type were recorded, and a random forest model using standard tuning was constructed (Wright and Ziegler, 2017). An independent validation dataset was then drawn from the data-generating model, and predictions ($\hat{y}$) were made from each of the 4 models. The root mean-squared error of prediction, $\sqrt{\frac{1}{n} (y - \hat{y})^2}$, was recorded. This was performed 500 times for each parameter scenario.

The results are displayed in Figure 4.8 through Figure 4.15. Across all instances

of $\mathbf{p}$, screening based on univariate Spearman correlations had the worst performance of the methods. Given the complexity of the underlying function, this is to be expected. In general, increasing the variance of the data contamination worsens the performance of the raw variance screening and improves the performance of the data quality screening compared to other methods. All methods tend to perform better when correlation is present, especially in higher dimensions. When no correlation is present, the performance over the mean-only model decreases as dimensionality increased. When correlation is present, information from signal variables can be used interchangeably, leading to information gains with fewer variables.

When all variable types have equal representation (Figure 4.8), screening based on raw variance performs better with low contamination levels, while screening based on data-quality performs better with higher contamination levels. When contamination is low, variance screening is able to pull out a large portion of the low-quality signal variables in addition to some high-quality signal variables, which quality screening does not. With high contamination levels, variance screening starts pulling out only low-quality signal variables and some low-quality noise variables. Data-quality screening consistently pulls out half noise variables and half signaling variables, both of high-quality.

When variables are equally split between high-quality signal variables and high-quality noise variables (Figure 4.9), screening based on variance estimates has better performance than screening based on data-quality, with performance gains becoming less substantial as contamination severity increases. As all variables are of high-quality, those variables with larger variance estimates will correspond to the signal variables. Screening based on variance consistently pulls out high-quality signal variables, while screening based on data-quality selects half noise and half signal variables.

When variables are equally split between low-quality signal variables and low-quality noise variables (Figure 4.10), screening based on variance estimates performs slightly better than screening based on data-quality when amount of contamination is low, and approximately equal with high levels of contamination. As all variables are of poor quality, those with a higher estimated variance are likely to have more signal. However, as contamination levels increase, the signal present from variables retained in variance screening becomes less apparent. Screening based on data quality pulls out half signal and half noise variables, both of low-quality. Variance screening grabs all low-quality signal variables when contamination levels are low, but begins to pull out low-quality noise variables as contamination levels increase.

When all variables are high-quality signal variables (Figure 4.11), screening based on data-quality and screening based on variance estimates perform similarly, with slight gains for screening based on data-quality when contamination levels are high. When all variables are high-quality, those that are slightly contaminated will become less reliable when contamination levels are high. Similarly, when all variables are low-quality signal variables (Figure 4.12), both screening based on data-quality and screening based on variance estimates perform poorly, as all variables contain large levels of contamination.

When the variables are split between high-quality signal variables and low-quality signal variables (Figure 4.13), screening based on data-quality performs better than screening based on variance estimates, with increased performance as amount of contamination increases. Screening based on data-quality leads to selecting the signal variables that are not being contaminated. In contrast, screening based on variance estimates leads to selecting primarily low-quality signal variables.

When variables are split between high-quality signal variables and low-quality noise variables (Figure 4.14), screening based on data-quality performs much better

than screening based on variance estimates. This is what we would expect, as screening based on data-quality selects the high-quality signal variables, while screening based on variance estimates primarily selects the low-quality noise variables. In contrast, when variables are split between low-quality signal variables and high-quality noise variables, screening based on variance components performs better than screening based on data-quality. While screening based on variance estimates selects the low-quality signal, screening based on data-quality selects the high-quality noise variables.

This simulation study highlights the importance of considering variable quality when performing initial variable screening. Overall, if the level of contamination is high, screening based on data quality tends to produce variable sets that predict better than other methods. If the level of contamination is low, and signaling variables have higher variability in a diverse population of patients (as commonly assumed in biological studies), screening based on variance estimates tends to perform better. Across all scenarios, screening based on univariate relationships with the response performed poorly, highlighting the importance of considering interactions and nonlinear relationships in complicated problems such as these. Future extensions will consider additional data-generating models (changing $f(\boldsymbol{x})$) and different modeling schemes.

Table 4.1: Types of $\mathbf{p}$ considered.

$\mathbf{p}$	Name
[0.25, 0.25, 0.25, 0.25]	All equal
[0.50, 0.00, 0.50, 0.00]	High-quality split
[0.00, 0.50, 0.00, 0.50]	Low-quality split
[1.00, 0.00, 0.00, 0.00]	High-quality signal
[0.00, 1.00, 0.00, 0.00]	Low-quality signal
[0.50, 0.50, 0.00, 0.00]	High-signal split
[0.50, 0.00, 0.00, 0.50]	Strong separation
[0.00, 0.50, 0.50, 0.00]	Weak separation

Table 4.2: Other simulation parameters.

Parameter	Values
p	{100, 200, 400}
ρ	{0.00, 0.30, 0.60}
σ_γ	{0.05, 0.10, 0.20}

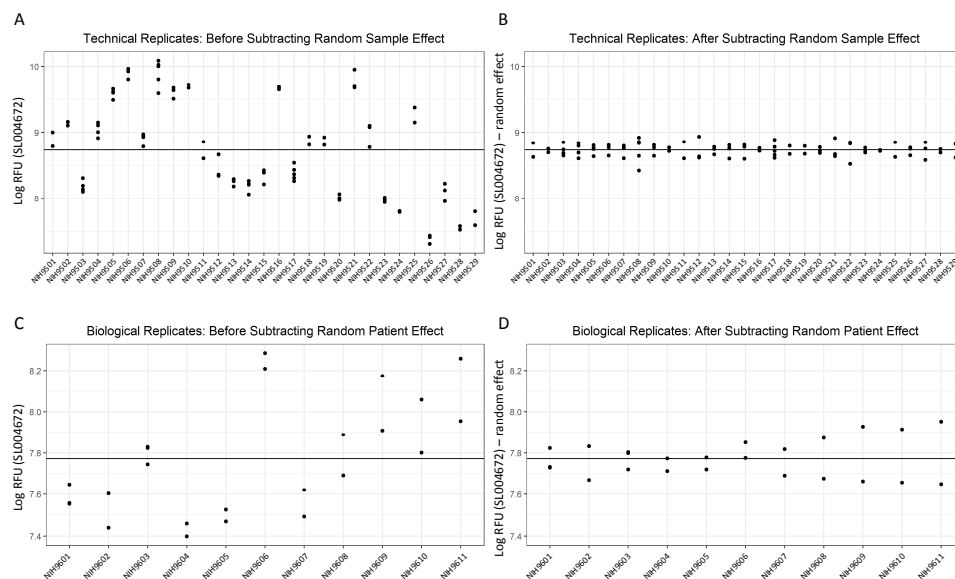


Figure 4.1: Technical (top graphs) and biological (bottom graphs) variance calculation for SOMAmer SL004672. The x-axes correspond to the patient/sample that measurements were produced from. The upper panels show technical replicates (n=88) and the lower panels show biological replicates (n=24). The left panels show the raw measurements (natural-log scale RFU) for SOMAmer SL004672 on the y-axes. The right panels show the identical raw observations with the random intercept effect subtracted to account for subject-to-subject variation. These residuals (after subtracting means of technical or biological replicates) were used to estimate the technical and biological variance, respectively. The horizontal black line is the estimated mean from the technical (top graphs) and biological (bottom graphs) variance models.

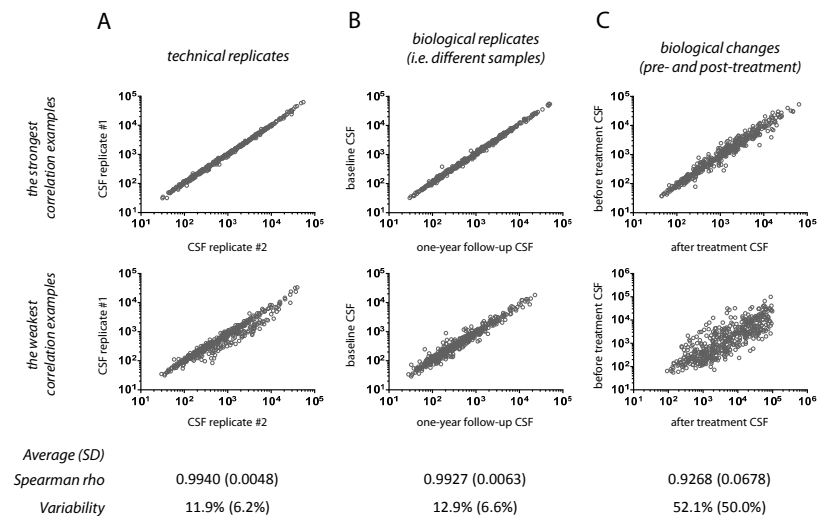


Figure 4.2: Differences in biomarker measurements (log10 RFU) in identical samples analyzed repeatedly (n=88, left), in longitudinal HD samples measured at different time points (n=24, middle), and in patient samples before and after application of immunomodulatory therapy (n=10, right): (i) An average of Spearman rho values calculated across 500 high-signaling SOMAmers with high SNR, and (ii) an average of variabilities (a median of relative percent changes calculated as absolute difference of RFUs for each of the 500 high-signaling SOMAmers between two replicates divided by the average of the two RFUs) for all pairs of replicates in each respective category. Examples of the strongest and the weakest correlations between two samples in each category are visualized on 500 high-signaling SOMAmers

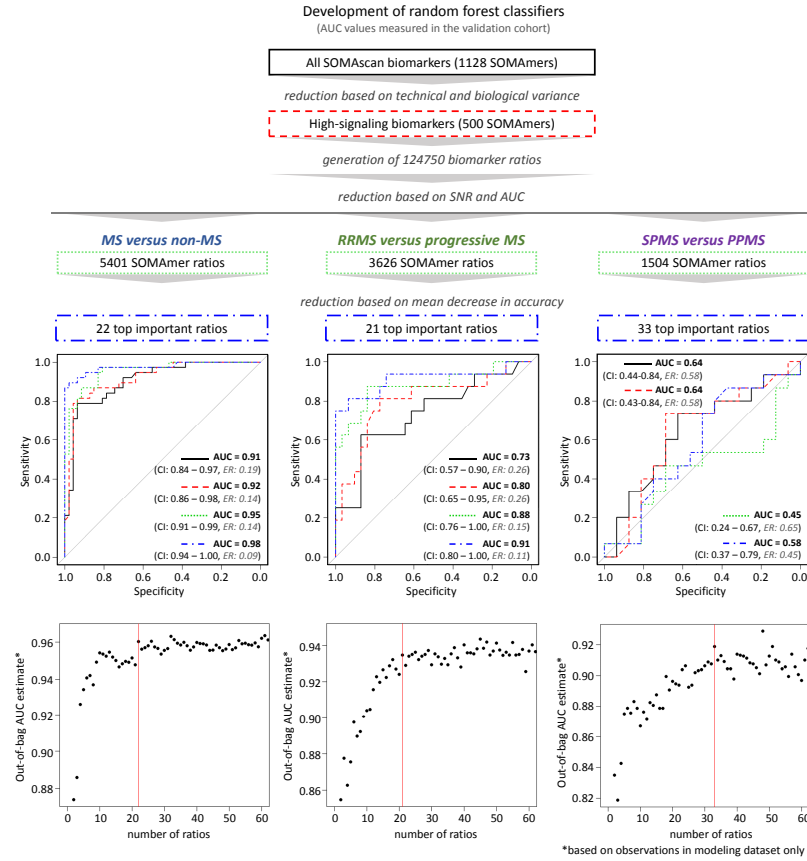


Figure 4.3: The SOMAscan assay comprises 1128 SOMAmers (solid black line curves). Calculation of technical and biological SNR reduced the number of SOMAmers considered for further analysis to 500 (dashed red line curves). Using the 500 high-signaling SOMAmers, 124750 biomarker ratios were generated that were subsequently tested for their SNR and their ability to differentiate two diagnostic groups (based on AUC in the modeling cohort), resulting in 5401 high-signaling biomarker ratios for MS versus non-MS diagnostic test, 3626 biomarker ratios for progressive versus RRMS diagnostic test, and 1504 biomarker ratios for SPMS versus PPMS diagnostic test (green dotted line curves). Out-of-bag AUC estimates (bottom graphs) examined from different random forests generated by sequentially adding ratios with the highest variable importance led to a logical cut-off (marked by solid red line) of 22 SOMAmer ratios for MS versus non-MS diagnostic comparison, 21 SOMAmer ratios for RRMS versus progressive MS diagnostic comparison, and 33 SOMAmer ratios for SPMS versus PPMS (blue dash-dot line curves).

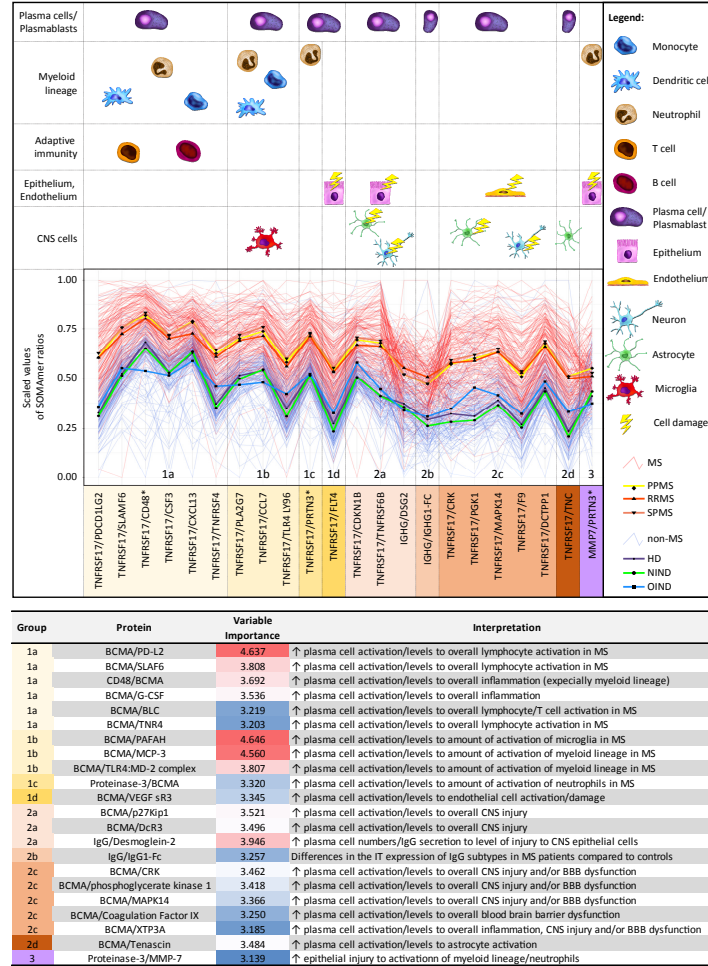


Figure 4.4: A parallel coordinate plot (PCP) for the 22 most important features that distinguish MS from non-MS. The plot displays individual patients from combined modeling (n=225) and validation cohort (n=85) divided into MS group (RRMS, PPMS, SPMS; thin red lines) and non-MS group (HD, NIND, OIND; thin blue lines). A group average is shown as thick yellow line for PPMS, thick red line for RRMS, thick orange line for SPMS, thick purple line for HD, thick green line for NIND, thick blue line for OIND. The y-axis shows SOMAmer natural log ratios scaled to 0-1 range. SOMAmer ratios were grouped based on the cellular origin and known functions of the individual components into nine groups. Different cell types are shown above the PCP to highlight the cell origin of individual SOMAmers. *ratios in the classifier are inverted.

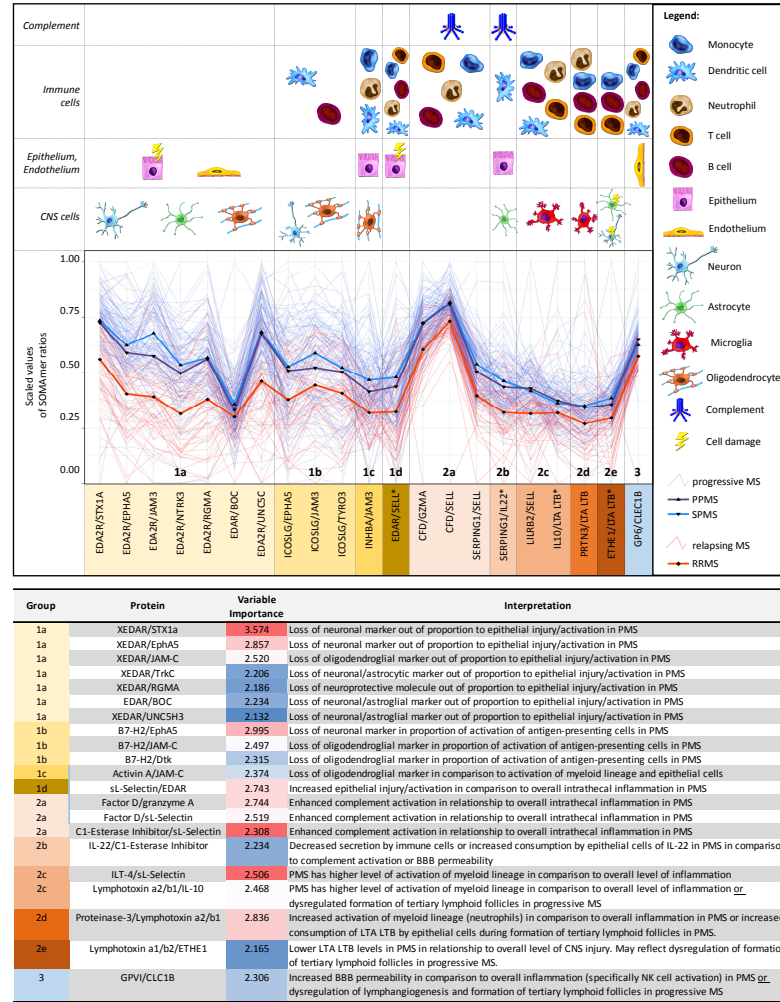


Figure 4.5: A parallel coordinate plot (PCP) for the 21 most important variables that distinguish progressive MS from relapsing MS. The plot displays individual MS patients from combined modeling (n=120) and validation cohort (n=47) divided into progressive MS group (PPMS and SPMS; thin blue lines) and relapsing MS group (RRMS; thin red lines). A group average is shown as thick purple line for PPMS, thick blue line for SPMS, and thick red line for RRMS. The y-axis shows SOMAmer natural log ratios scaled to 0-1 range. SOMAmer ratios were grouped based on the cellular origin and known functions of the individual components into nine groups. Different cell types are shown above the PCP to highlight the cell origin of individual SOMAmers. *ratios in the classifier are inverted.

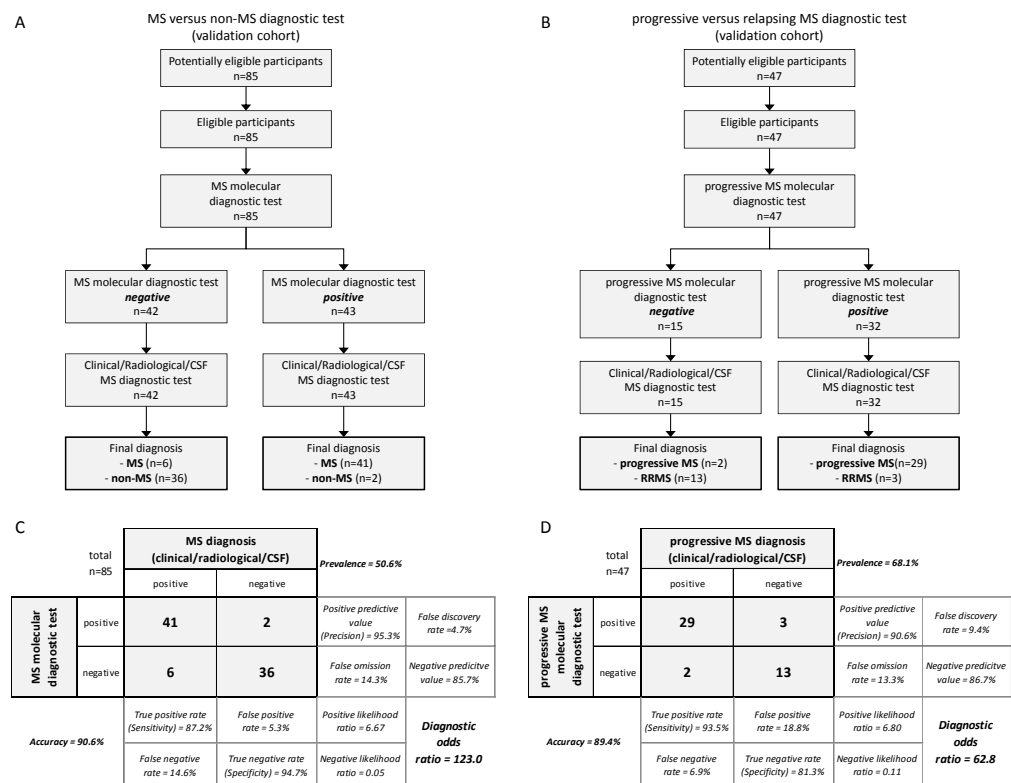


Figure 4.6: (A) STARD diagram and (C) confusion matrix for 85 subjects used for validation of the MS molecular diagnostic test and (B) STARD diagram and (D) confusion matrix for 47 subjects used for validation of the progressive MS molecular diagnostic test.

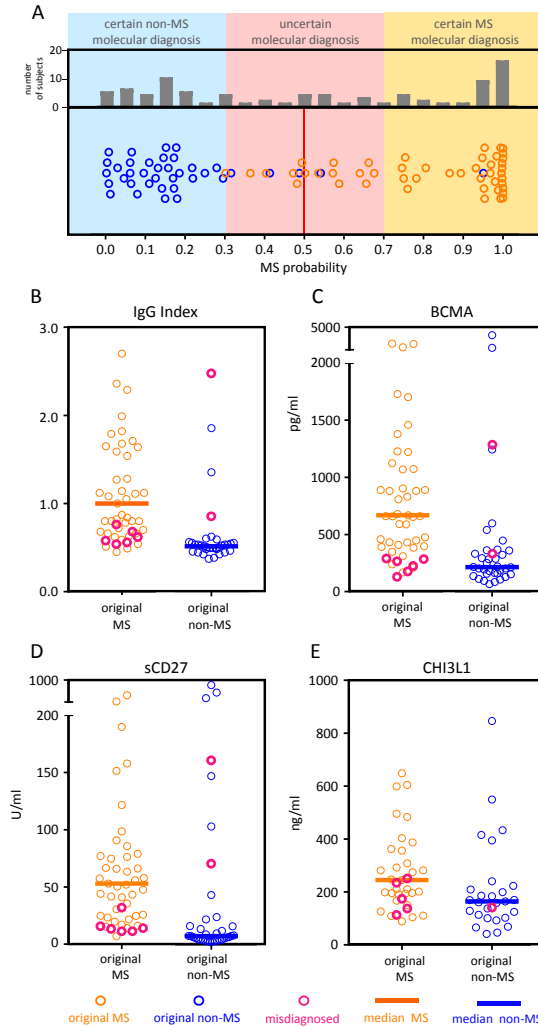


Figure 4.7: (A) The MS diagnostic probability (on the y-axis) of 85 subjects from the validation cohort is shown in the graph. Blue circles represent subjects with original non-MS diagnosis (HD, OIND, NIND) and orange circles represent subjects with original MS diagnosis (RRMS, PPMS, SPMS). The red line represents an arbitrary cut-off at 50%. The pink background marks an area between 30% and 70% where the certainty of the molecular classification is weak (contains 22.4% of the validation cohort's subjects). The orange background highlights 70.0% of the validation cohort's MS subjects with highly probable MS molecular diagnosis (>70%) and the blue background labels 86.8% of the validation cohort's non-MS subjects with high probability of non-MS molecular diagnosis (<30%). Grey bars represent a frequency distribution bar chart. Misdiagnosed subjects (pink circles) were evaluated for non-SOMALogic biomarkers of inflammation: IgG index (B), BCMA (C), sCD27 (D), and CHI3L1 (E) using alternative assays (Komori et al., 2016). The group medians are shown for MS subjects as an orange line and for non-MS subjects as a blue line.

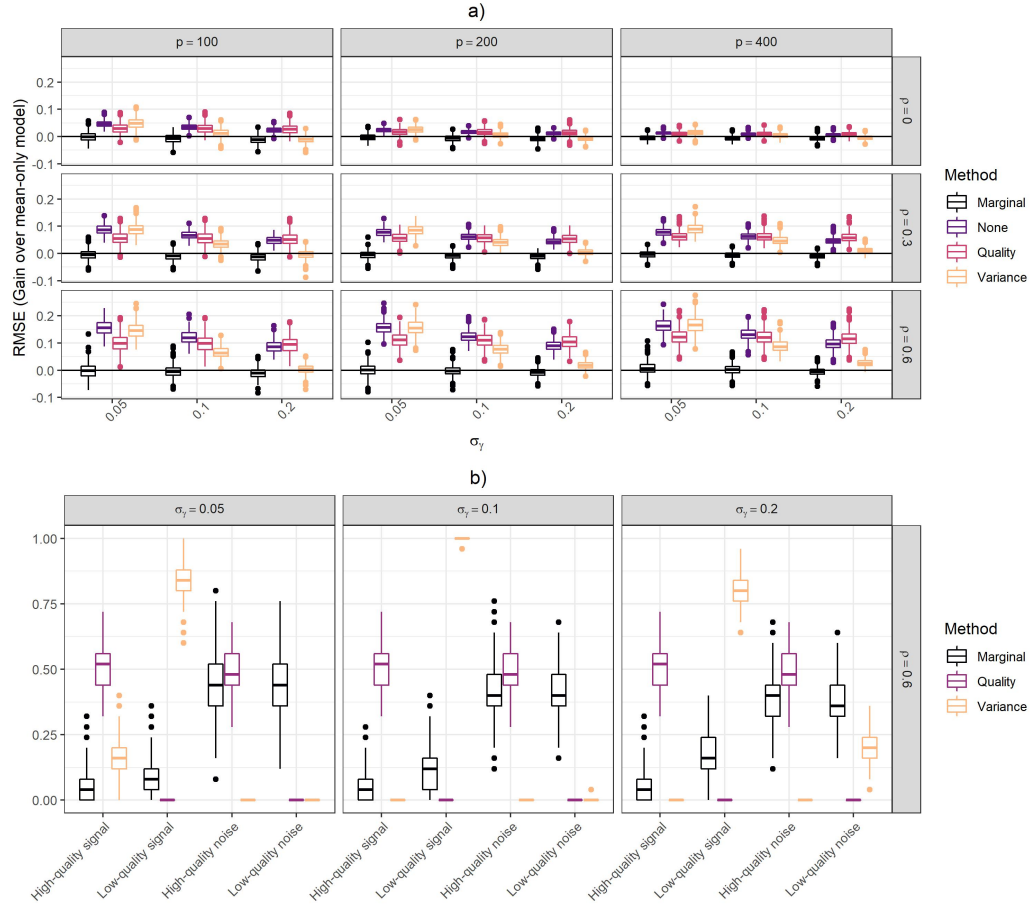


Figure 4.8: Simulation results for the “equal” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

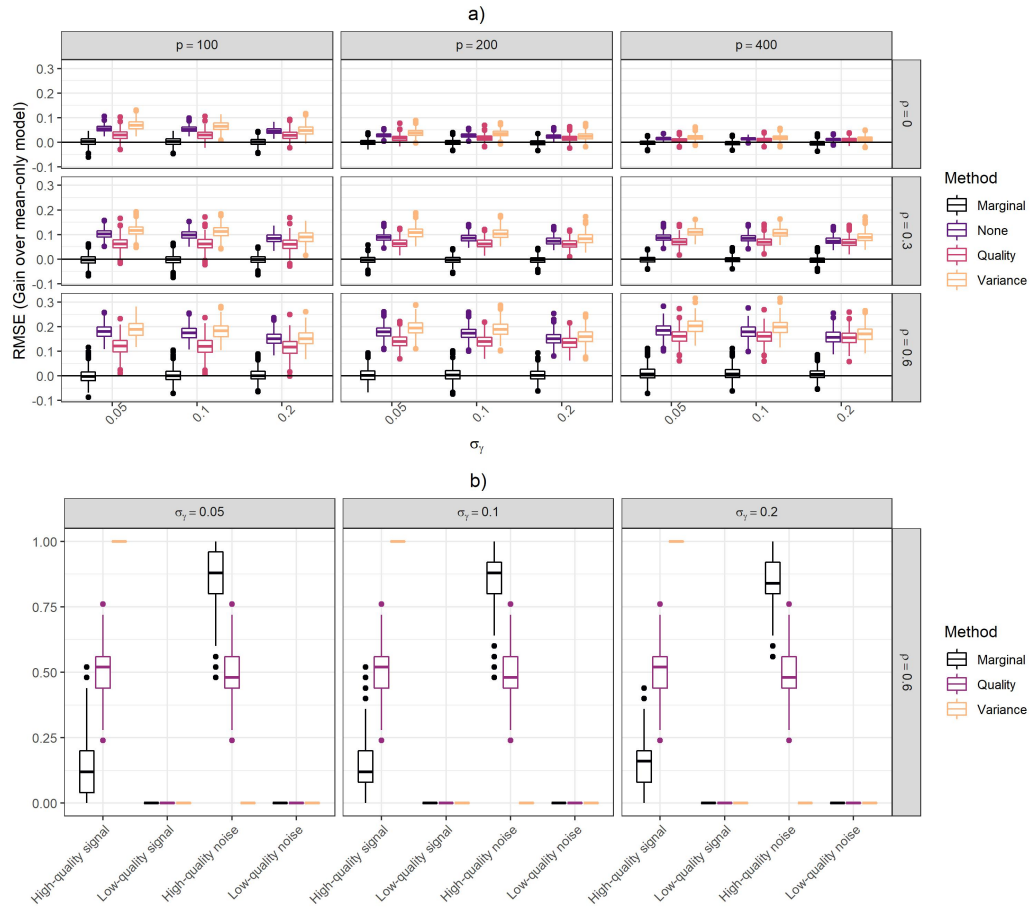


Figure 4.9: Simulation results for the “high-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

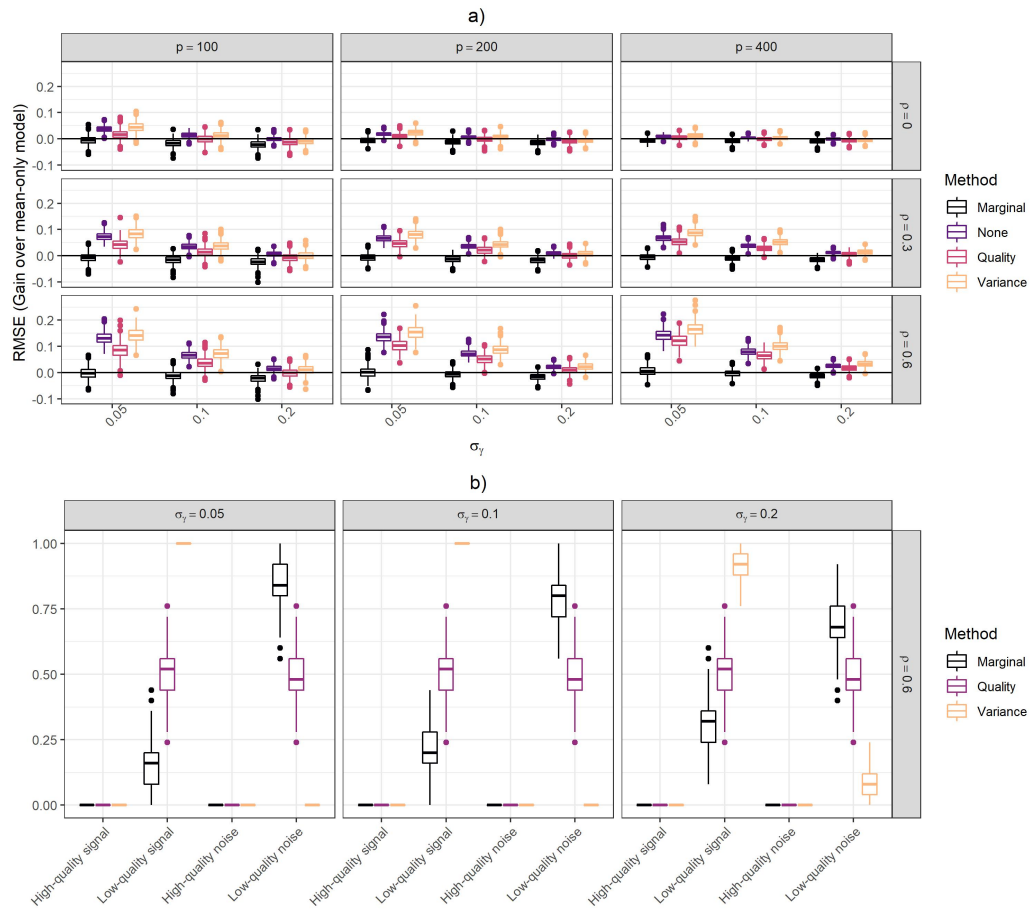


Figure 4.10: Simulation results for the “low-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

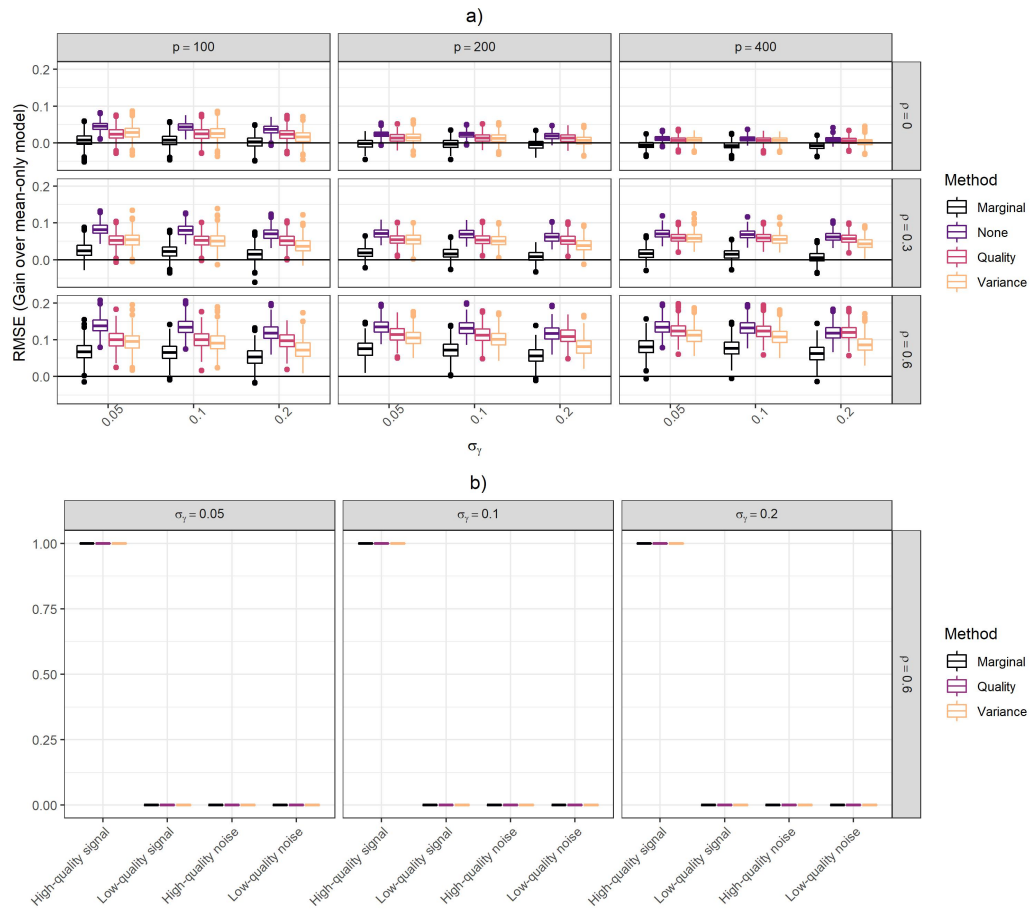


Figure 4.11: Simulation results for the “signal-high” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

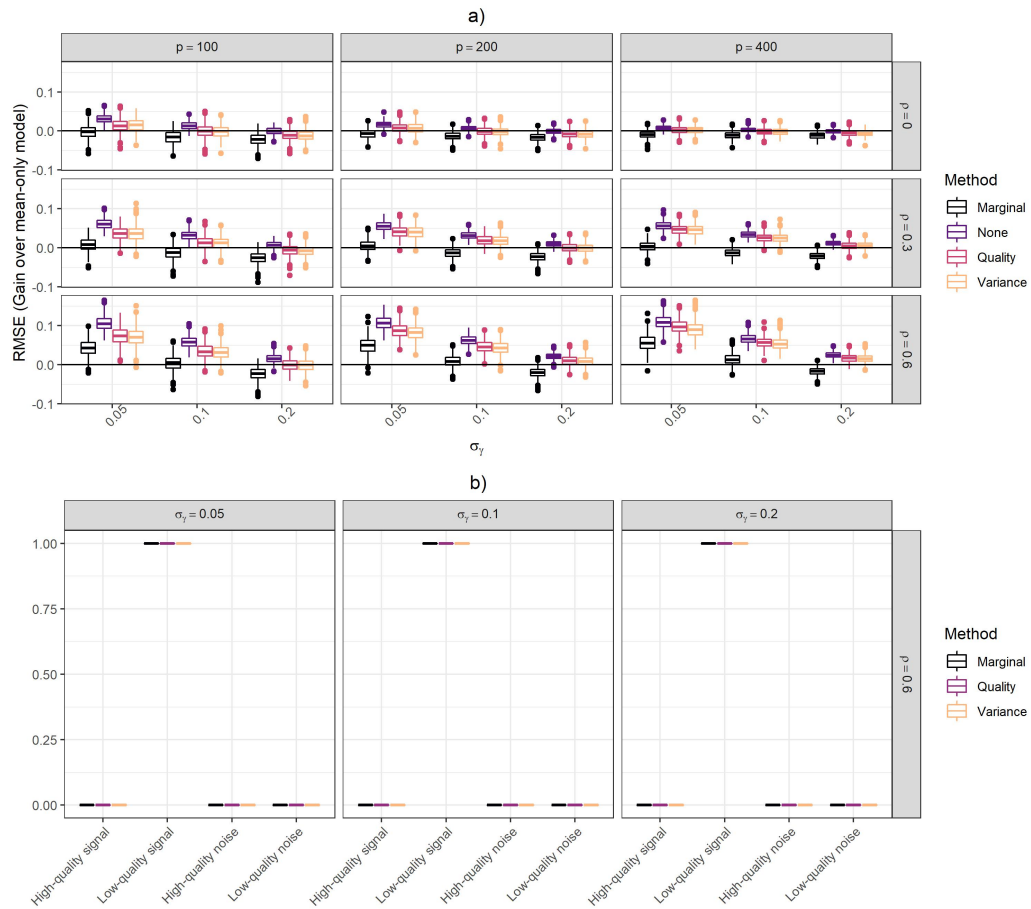


Figure 4.12: Simulation results for the “signal-low” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

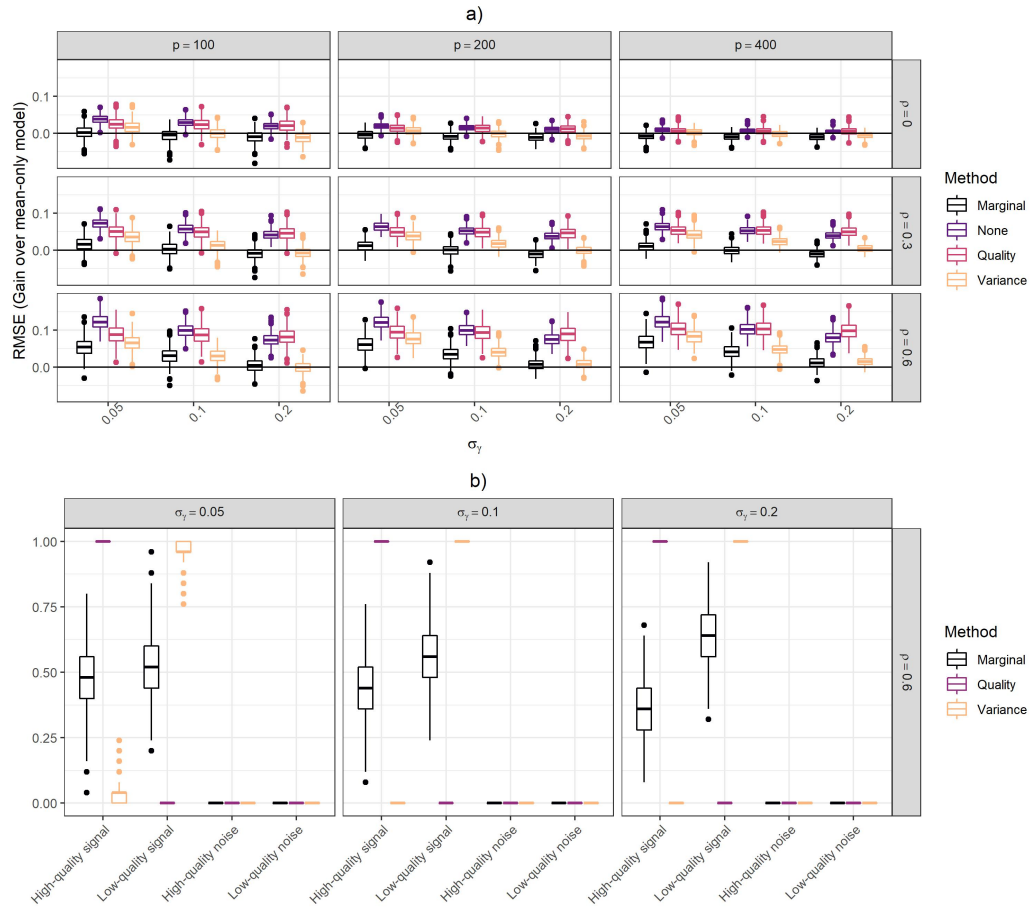


Figure 4.13: Simulation results for the “signal-split” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

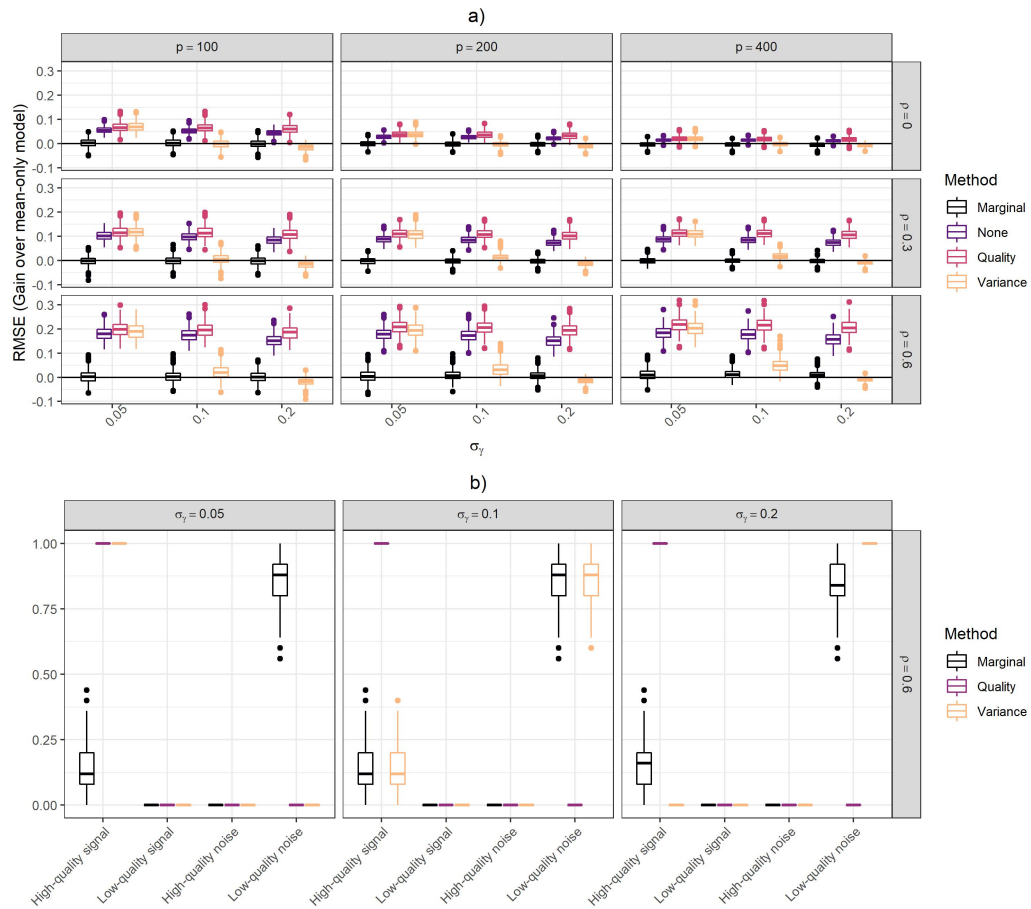


Figure 4.14: Simulation results for the “strong-separation” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

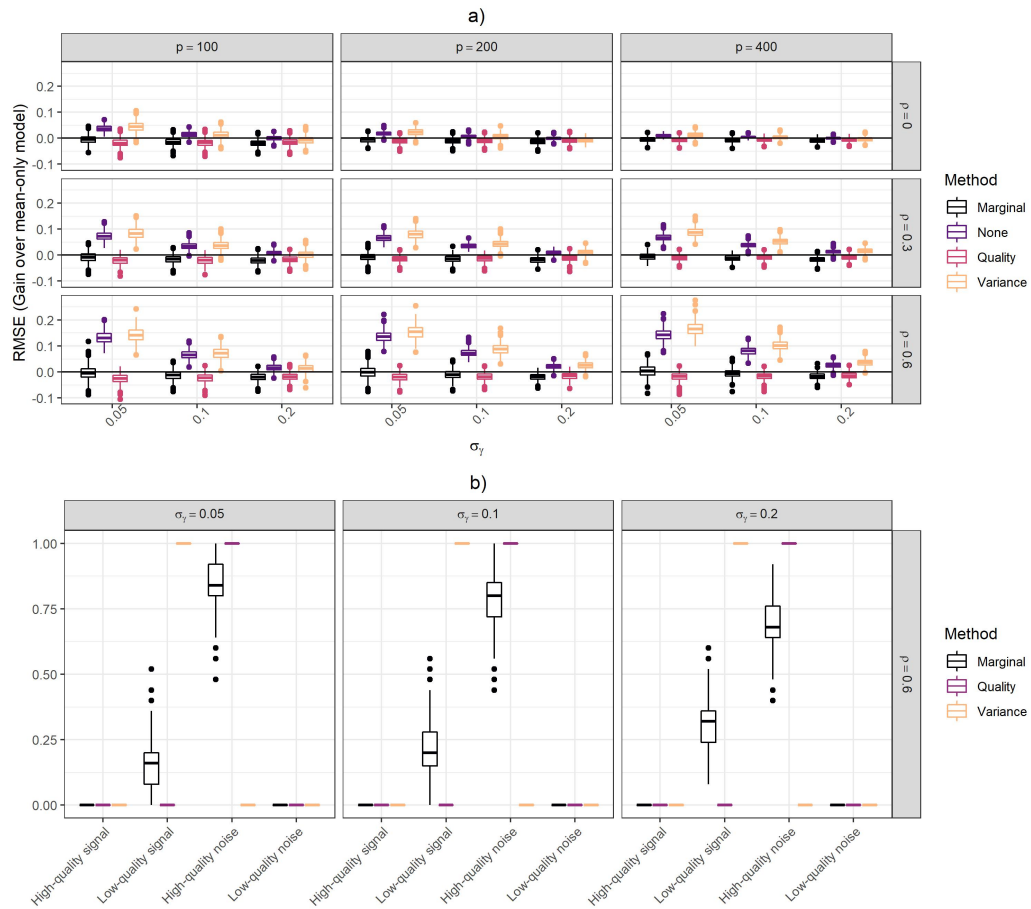


Figure 4.15: Simulation results for the “weak-separation” proportion case. a) The root-mean-squared error, expressed as a percentage gain compared to predictions from a mean-only model is displayed on the y-axis, increasing values of σ_g are displayed on the x-axis, and colors represent the various screening techniques. Columns represent increasing dimensionality and rows represent increasing pairwise correlation. b) For the $p = 100$ and $\rho = 0.6$ case, the proportion of the different types of variables selected is displayed on the y-axis, the variable type is displayed on the x-axis, with colors representing the different screening approaches. Columns represent increasing values of σ_g .

CHAPTER FIVE

CONCLUSIONS AND EXTENSIONS

As technology and scientific understanding continues advancing, the amount of biological signals we are able to collect will continue to grow. With this growth, statistical methods tailored towards specific problems of interest and data forms, with varying quality of measurement precision, will become increasingly important. The goal of this dissertation was to develop novel techniques specific to various research problems in the medical field, specifically neurological disorders and Multiple Sclerosis (MS) research relating to clinical endpoint development and proteomics analysis.

In complex neurological diseases, developing a sensitive outcome to accurately measure disease severity (or progression) is difficult, and data-driven approaches have not been extensively considered in the past. Through the motivation to use clinical scales currently utilized in MS research to create a linear projection that maximizes sensitivity to longitudinal change, the Constructed Composite Response (CCR, Chapter 2) methodology was developed. Through a simulation study and a motivating dataset of progressive MS patients, it was shown that this methodology was able to construct projections that were able to better detect longitudinal change across a variety of data structures compared to traditional methods. Chapter 3 extended this methodology to include sparsity-imposing constraints on the projection coefficients, motivated by advances in penalized regression models. The l_1 (or Lasso) constraint was found to perform the best in terms of ease of convergence, size of targeted t -statistic in data independent of model development, and interpretability of the resulting projection coefficients. This extended the CCR methodology to be more effective in high-dimensional problems and was used to uncover neurological

functions that associate with longitudinal change.

As discussed in Chapters 2 and 3, extending the CCR methodology to maximize F -statistics from a specified contrast matrix has the potential to target more complicated model components and could be used to design scales that target more sophisticated biological questions. For a specified contrast matrix, this can be solved by applying said contrast to the model coefficient vector for each potential response, $\hat{\beta}_p$, and constructing a new $\mathbf{B}$ matrix similar to that in Chapter 2. The solution (up to a sign) can be solved using the analytic approach presented in Chapter 2. An example of this was performed on the `crabs` dataset (Campbell and Mahon, 1974), accessed in the `MASS` R package (Venables and Ripley, 2002). The dataset contains 5 morphological measurements (frontal lobe size, rear width, carapace length, carapace width, and body depth, all measured in millimeters) on 200 crabs of the species *Leptograpsus variegatus* collected in Western Australia, 50 each of blue males, blue females, orange males, and orange females. The original authors were able to differentiate color sub-types and sex within color sub-types through two weighted contrasts.

A CCR projection was constructed using a linear combination of these 5 measurements to maximize the overall F -statistic from a linear model with the 4 groupings as a single factor. All of these measurements were standardized to mean 0 and unit variance prior to the CCR construction. The projection coefficients are displayed in Table 5.1, and the distributions of CCR values, z_i , across the sex/color combinations are displayed in Figure 5.1. A clear separation between blue and orange colors is evident, regardless of sex. Within colors, the differences between sex are less prominent, with males having slightly higher values compared to females within the respective colors. The original manuscript claimed that a weighted contrast of carapace width versus body depth and front lip separated crab color, and a weighted contrast of rear width versus body depth separated sex within color. The CCR

projection was dominated by the weighted contrast of carapace width versus body depth and front lip length and much less by the contrast of rear width versus carapace length. If instead, the CCR was used to target the difference between sex within color after accounting for differences in color, it is likely that the weighted contrast of rear width versus carapace length would be more prominent. When targeting an F -statistic, careful scientific reasoning will be needed to determine the “target” of the projection, and more research is needed on the practical considerations.

Two future extensions of the CCR methodology will be the incorporation of *a priori* variable groupings and targeting model coefficients from more complicated linear models that do not contain closed form solutions. The former could potentially be solved by introducing a Group Lasso constraint on the projection coefficients, where all coefficients within a grouping were either retained or excluded depending on the level of regularization. A simpler, yet intuitive approach would be to first apply the CCR (or B-CCR) to the individual groupings. If there was a high level of dependence within the groupings, but approximate pairwise independence between the groupings, we would hope to see approximate independence between the CCR values from these projections. Using these CCR values as a new set of response variables in a final projection may lead to increased performance over inclusion of all variables in a single projection.

Although the ridge penalty (and by extension the elastic-net penalty) were discussed in Chapter 3, how these penalties affect the behavior of the penalized CCR were not explored in this research. Inclusion of the ridge penalty has been shown in some instances to increase performance in the presence of correlated variables. It could be hypothesized that including it in the CCR objective may lead to performance gains in the presence of residual correlation. However, in other instances, a ridge penalty is simply used as an identifiability constraint, as is the case in the cluster

elastic-net. In the CCR formulation, an identifiability constraint is already used to handle the denominator matrix, $\mathbf{M}$, as it led to a more computationally feasible solution. Instead, a ridge-like identifiability constraint could have been used instead. Whether including a ridge penalty along with the current identifiability constraint has potential to improve the CCR algorithm is unclear, and how this extension will affect the CCR solution will be explored in future research.

Assuming an unbounded solution space, exact analytic CCR solutions exist when a standard linear model is assumed. While certain scenarios lend themselves to this modeling formulation, many data structures would require unreasonable assumptions to be made in order to use this framework. For example, when observing longitudinal data, we may expect that residual values closer together in time will have some level of dependence that damps over time. Or, we could expect that the variability in the observed values of the constructed signal could become more variable as we observe them over time. As previously stated, incorporating crossed or nested random effects would allow us to account for various batch effects due to observations collected within the same patient, clinician, or machine. All of these scenarios imply that the residual variance covariance matrix changes from $\sigma^2 \mathbf{I}$ to $\sigma^2 \mathbf{V}$, and $\mathbf{V}$ needs to be estimated unless assumed known. These models are typically fit by calculating a set of regression coefficients, $\hat{\boldsymbol{\beta}}$, assuming independence among observations, using the residuals from this model to estimate the variance-covariance matrix of a particular form, $\hat{\mathbf{V}}$, re-estimating $\hat{\boldsymbol{\beta}}$ using the estimate of $\hat{\mathbf{V}}$, re-estimating $\hat{\mathbf{V}}$ using the residuals from the updated model, and iterating until convergence (Pinheiro and Bates, 2000). We hypothesize the CCR methodology could be applied for these models by estimating $\hat{\mathbf{c}}^\top$ assuming independence, using the CCR values z_i to estimate $\hat{\mathbf{V}}$ from the specified model, and iterating until convergence. The merit of this hypothesis will be examined in future studies.

In the applications that motivated the CCR methodology, determining which individual scales should be included in the projection was the primary interest. In other applications, we may have some *a priori* knowledge that a subset of scales should be retained with probability 1, and the interest would instead be in determining which additional features could be included to further strengthen the captured signal. For example, it is commonly observed that lower ambulatory functions such as walking degrade as the disease progresses, and interest could lie in examining which additional neurological systems improve the ability to measure these changes in disease severity. This extension could be solved by introducing different thresholding operators to the individual projection coefficients, c_j . The practical implications of such an approach are unknown and not common in penalized regression models, and will need to be extensively studied in future research.

While a p-value calculation was presented for the CCR methodology in Chapter 2, creating an estimate of the standard error of the projection coefficients, $SE(\hat{\mathbf{c}})$, was not explored. An estimate of this standard error could be used to construct confidence regions of the estimated coefficients, $\hat{\mathbf{c}}$, to examine uncertainty in the projected observations, or for directly incorporating into the modeling framework as an estimate of measurement error (e.g., Weeding, 2016). The standard errors of this non-standard confidence region could potentially be estimated through a bootstrap procedure (Efron and Tibshirani, 1993, Hastie, Tibshirani, and Wainwright, 2015) or through a Bayesian framework (Weeding, 2016, Gelman et al., 2013, Park and Casella, 2008).

Chapter 4 addressed a problem of variable screening in proteomics data, where many sources of unwanted variation are often assumed to be present, and in some instances can be estimated. A novel variable screening metric was developed based on data-quality, where data-quality ranks individual variables based on the relative size

of this undesired additional variation. Through a simulation study comparing this approach to traditional screening approaches, it was shown that depending on the level of this unwanted variation relative to the variability attributable to the signal of interest, screening based on data quality or screening based on marginal variance estimates tended to perform the best. While these results show promise, they are based on a single data-generating function (e.g., the Rosenbrock function). Future research will extend the simulation study to examine other linear and non-linear data-generating processes for both regression and classification problems.

Table 5.1: CCR Projection coefficients from crabs dataset.

Frontal lobe	Rear width	Carapace length	Carapace width	Body depth
5.10	-0.45	4.19	-12.37	4.49

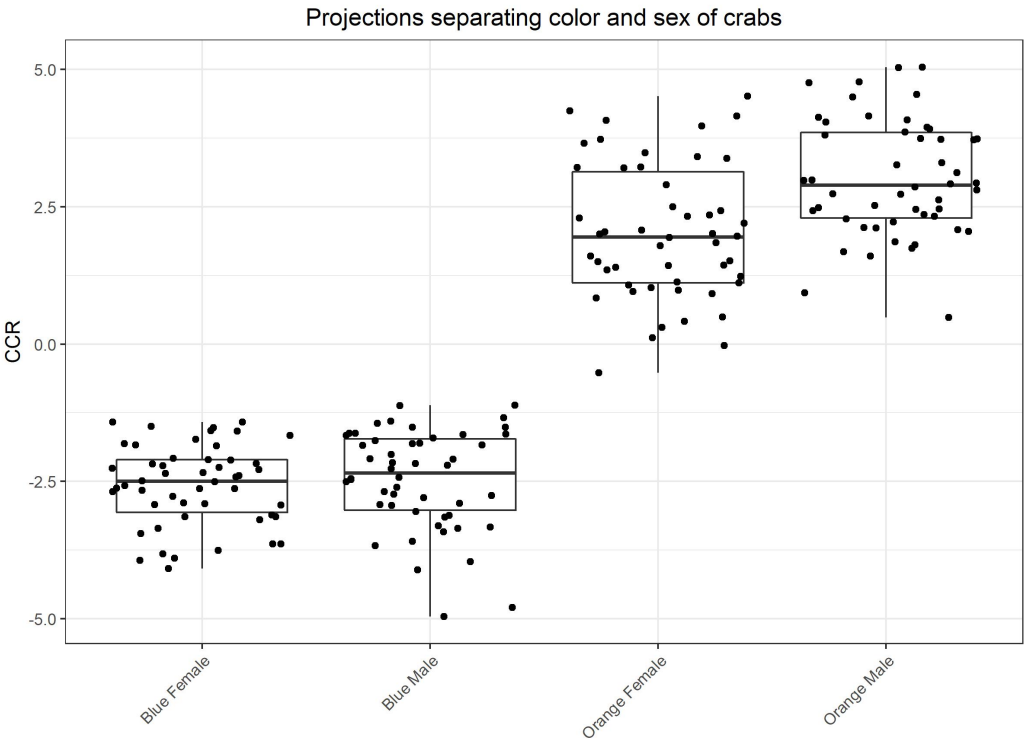


Figure 5.1: Distribution of CCR values across different groupings.

REFERENCES CITED

- [1] T. Anderson and H. Rubin. “Statistical Inference in Factor Analysis”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (1956), pp. 111–150.
- [2] C. Barbour et al. *Molecular-Based Diagnosis of Multiple Sclerosis and Its Progressive Stage*. 2017.
- [3] M. Bartlett. “The statistical conception of mental scores”. In: *British Journal of Psychology* 28 (1937), pp. 97–104.
- [4] K. Berry, J. Johnston, and P. Mielke. *A Chronicle of Permutation Statistical Methods. 1920 - 2000, and beyond*. Springer, 2014. ISBN: 978-3-319-02743-2.
- [5] B. Bielekova et al. “Cerebrospinal Fluid IL-12p40, CXCL13 and IL-8 as a Combinatorial Biomarker of Active Intrathecal Inflammation”. In: *PLoS ONE* 7.11 (2012).
- [6] B. Bielekova et al. “How implementation of systems biology into clinical trials accelerates understanding of diseases”. In: *Front Neurol* 5 (2014), p. 102.
- [7] K. Bollen. *Structural models with latent variables*. New York: Wiley, 1989.
- [8] K. Bollen and P. Curran. *Latent Curve Models: A Structural Equation Perspective*. New York: Wiley, 2006.
- [9] H. Bondell and B. Reich. “Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR”. In: *Biometrics* 64.1 (2008), pp. 115–123.
- [10] L. Breiman. “Better subset selection using the nonnegative garrote”. In: *Technometrics* 37 (1995), pp. 738–754.

- [11] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.
- [12] T. Brown. *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press, 2006.
- [13] J.P. Buonaccorsi. *Measurement error: models, methods, and applications*. CRC Press, 2010.
- [14] N.A. Campbell and R.J. Mahon. “A Multivariate Study of Variation in Two Species of Rock Crab of the Genus *Leptograpsus*”. In: *Australian Journal of Zoology* 22 (1974), pp. 417–425.
- [15] R.J. Carroll et al. *Measurement error in nonlinear models: a modern perspective*. CRC Press, 2006.
- [16] A. Chambolle and T. Pock. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging”. In: *Journal of Mathematical Imaging and Vision* 40 (2011), pp. 120–145.
- [17] J.R. Cook and L.A. Stefanski. “Simulation-extrapolation estimation in parametric measurement error models”. In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1314–1328.
- [18] G. Cutter et al. “Development of a multiple sclerosis functional composite as a clinical trial outcome measure.” In: *Brain* 122.5 (1999), pp. 871–882.
- [19] C. DiStefano, M. Zhu, and D. Mindrila. “Understanding and Using Factor Scores: Considerations for the Applied Researcher”. In: *Practical Assessment, Research and Evaluation* 14.20 (2009).
- [20] P. Douvaras and V. Fossati. “Generation and isolation of oligodendrocyte progenitor cells from human pluripotent stem cells”. In: *Nat Protocols* 10.8 (2015), pp. 1143–1154.

- [21] T. Duncan, S. Duncan, and L. Strycker. *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*. 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 2006.
- [22] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, NY, 1993. ISBN: 978-0-412-04231-7.
- [23] S. Epskamp and S. Stuber. *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*. R package version 1.1. 2017. URL: <https://CRAN.R-project.org/package=semPlot>.
- [24] R.A. Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Annals of Eugenics* 7.2 (1936), pp. 179–188.
- [25] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: <http://www.jstatsoft.org/v33/i01/>.
- [26] J.H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29 (2001), pp. 1189–1232.
- [27] W. Fuller. *Measurement Error Models*. Wiley, 1987. ISBN: 978-0-470-09571-3.
- [28] A. Gelman et al. *Bayesian Data Analysis*. 3rd. Chapman and Hall/CRC Press, 2013.
- [29] L. Gold et al. “Aptamer-based multiplexed proteomic technology for biomarker discovery”. In: *PLoS ONE* 5.12 (2010).
- [30] L. Gold et al. “Advances in human proteomics at high scale with the SOMAscan proteomics platform”. In: *New Biotechnology* 29.5 (2012), pp. 543–549.

- [31] R. Gorsuch. *Factor Analysis*. 2nd ed. Mahwah NJ: Lawrence Erlbaum Associates, 1983.
- [32] J. Grace. *Structural Equation Modeling and Natural Systems*. Cambridge University Press, 2006.
- [33] J. Grice. “Computing and Evaluating Factor Scores”. In: *Psychological Methods* 6.4 (2001), pp. 430–450.
- [34] D. Gronwall. “Paced auditory serial addition task: A measure of recovery from concussion”. In: *Perceptual Motor Skills* 44 (1977), pp. 367–373.
- [35] S. Hardcastle et al. “Knowledge, attitudes, and practice of oncologists and oncology health care providers in promoting physical activity to cancer survivors: an international survey.” In: *Support Care Cancer* (2018).
- [36] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. 3rd ed. Berlin: Springer-Verlag, 2012.
- [37] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- [38] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Taylor and Francis, 2015.
- [39] M.J. Hawrylycz et al. “An anatomically comprehensive atlas of the adult human brain transcriptome”. In: *Nature* 489.7416 (2012), pp. 391–399.
- [40] A. Hoerl and R. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 12 (1970), pp. 55–67.
- [41] H. Hotelling. “Analysis of a complex of statistical variables into principal components”. In: *The Journal of Educational Psychology* 24 (1933), pp. 417–441.

- [42] J.P. Ioannidis. “A roadmap for successful applications of clinical proteomics”. In: *Proteomics Clin Appl* 5.5–6 (2011), pp. 241–247.
- [43] G. Jacob L. Obozinski and J.P. Vert. “Group Lasso with overlap and graph lasso”. In: *Proceedings of the 26th International Conference on Machine Learning* (2009).
- [44] C.E. Johanson et al. “Multiplicity of cerebrospinal fluid functions: New challenges in health and disease”. In: *Cerebrospinal fluid research* 5 (2008), p. 10.
- [45] I.T. Jolliffe, N.T. Trendafilov, and M. Uddin. “A modified principal component technique based on the Lasso”. In: *Journal of Computational and Graphical Statistics* 12 (2003), pp. 531–547.
- [46] H. Kaiser. “The varimax criterion for analytic rotation in factor analysis”. In: *Psychometrika* 23.3 (1958).
- [47] P. Kampstra. “Beanplot: A Boxplot Alternative for Visual Comparison of Distributions”. In: *Journal of Statistical Software, Code Snippets* 28.1 (2008), pp. 1–9. URL: <http://www.jstatsoft.org/v28/c01/>.
- [48] B.K. Kleinschmidt-DeMasters and K.L. Tyler. “Progressive multifocal leukoencephalopathy complicating treatment with natalizumab and interferon beta-1a for multiple sclerosis”. In: *N Engl J Med.* 353.4 (2005), pp. 369–374.
- [49] S. Koga et al. “When DLB, PD, and PSP masquerade as MSA: an autopsy study of 134 patients.” In: *Neurology* 85.5 (2015), pp. 404–412.
- [50] I. Kola and J. Landis. “Can the pharmaceutical industry reduce attrition rates?” In: *Nat Rev Drug Discov* 3.8 (2004), pp. 711–5.

- [51] M. Komori et al. “Cerebrospinal fluid markers reveal intrathecal inflammation in progressive multiple sclerosis”. In: *Annals of Neurology* 78.1 (2015), pp. 3–20.
- [52] M. Komori et al. “Insufficient disease inhibition by intrathecal rituximab in progressive multiple sclerosis.” In: *Annals of clinical and translational neurology* 3.3 (2016), pp. 166–179.
- [53] P. Kosa et al. “Development of a sensitive outcome for economical drug screening for progressive multiple sclerosis”. In: *Frontiers in Neurology. Multiple Sclerosis* 7.131 (2016).
- [54] P. Kosa et al. “NeurEx: digitalized neurological examination offers a novel high-resolution disability scale”. In: *Annals of Clinical and Translational Neurology* (2018). DOI: 10.1002/acn3.640.
- [55] S. Kraemer et al. “From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay”. In: *PLoS ONE* 6.10 (2011).
- [56] W. Krzanowski. *Principles of Multivariate Analysis: A User’s Perspective*. New York: Oxford University Press, 1988.
- [57] J.F. Kurtzke. “Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)”. In: *Neurology* 33.11 (1983), pp. 1444–1452.
- [58] A. Liaw and M. Wiener. “Classification and Regression by randomForest”. In: *R News* 2.3 (2002), pp. 18–22. URL: <http://CRAN.R-project.org/doc/Rnews/>.
- [59] S.A. Liddel et al. “Neurotoxic reactive astrocytes are induced by activated microglia.” In: *Nature* (2017).

- [60] M. Lim and T. Hastie. “Learning interactions via hierarchical group-Lasso regularization”. In: *Journal of Computational and Graphical Statistics* 24.3 (2014), pp. 627–654.
- [61] T. Little. *Longitudinal Structural Equation Modeling*. New York: The Guilford Press, 2013.
- [62] C. Lucchinetti et al. “Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination”. In: *Ann Neurol.* 47.6 (2000), pp. 707–717.
- [63] R. Magliozzi et al. “Meningeal B-cell follicles in secondary progressive multiple sclerosis associate with early onset of disease and severe cortical pathology.” In: *Brain* 130.4 (2007), pp. 1089–1104.
- [64] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, Inc, 2005.
- [65] N. Meinshausen. “Variable selection via penalized likelihood”. In: *Journal of the American Statistical Association* 86 (2001), pp. 1348–1360.
- [66] N. Meinshausen. “Relaxed Lasso”. In: *Computational Statistics and Data Analysis* 52 (2007), pp. 374–393.
- [67] A. Meisner et al. “Combining Biomarkers by Maximizing the True Positive Rate for a Fixed False Positive Rate”. In: *UW Biostatistics Working Paper Series* Working Paper 420 (2017). URL: <https://biostats.bepress.com/uwbiostat/paper420>.
- [68] J. Monahan. *A Primer on Linear Models*. Boca Raton, FL: Chapman and Hall, 2008.

- [69] X. Montalban et al. “Efficacy and Safety of Ocrelizumab in Primary Progressive Multiple Sclerosis: Results of the Phase III Double-Blind, Placebo-Controlled ORATORIO Study (S49.001)”. In: *Neurology* 86 (2016).
- [70] P. Morgan et al. “Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival”. In: *Drug discovery today* 17.9–10 (2012), pp. 419–424.
- [71] T. Park and G. Casella. “The Bayesian Lasso”. In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686.
- [72] K. Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. In: *Philosophical Magazine* 2.11 (1901), pp. 559–572.
- [73] E. Petkova et al. “Generated effect modifiers (GEM’s) in randomized clinical trials”. In: *Biostatistics* 18.1 (2017), pp. 105–118.
- [74] J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000. ISBN: 978-1-441-90317-4.
- [75] J. Pinheiro et al. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128. 2016. URL: <http://CRAN.R-project.org/package=nlme>.
- [76] C.H. Polman et al. “Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria”. In: *Annals of neurology* 69.2 (2011), pp. 292–302.
- [77] A. Puig, A. Wiesel, and A. Hero. “A multidimensional shrinkage thresholding operator”. In: IEEE. 2009, pp. 113–116.
- [78] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: <https://www.R-project.org/>.

- [79] S. Rao et al. “Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction.” In: *Neurology* 41 (1991), pp. 685–691.
- [80] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.8.4. Northwestern University. Evanston, Illinois, 2018. URL: <https://CRAN.R-project.org/package=psych>.
- [81] G. Rizzo et al. “Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis”. In: *Neurology* 6 (2016), pp. 566–576.
- [82] X. Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12 (2011), p. 77.
- [83] J.C. Rohloff et al. “Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents”. In: *Molecular therapy Nucleic acids* 3 (2014).
- [84] H.H. Rosenbrock. “An automatic method for finding the greatest or least value of a function”. In: *The Computer Journal* 3.3 (1960), pp. 175–184.
- [85] Y. Rosseel. “lavaan: An R Package for Structural Equation Modeling”. In: *Journal of Statistical Software* 48.2 (2012), pp. 1–36. URL: <http://www.jstatsoft.org/v48/i02/>.
- [86] J. Sipe et al. “A neurological rating scale (NRS) for use in multiple sclerosis”. In: *Neurology* 34.10 (1984), pp. 1368–1372.
- [87] A. Smith. *Symbol digits modalities test*. Western Psychological Services. Los Angeles, CA, 1982.
- [88] R. Sorlie et al. “Repeated observation of breast tumor subtypes in independent gene expression data sets”. In: *PNAS* 100.14 (2003), pp. 8418–8423.

- [89] C. Spearman. “General intelligence objectively determine and measured”. In: *The American Journal of Psychology* 15 (1904), pp. 201–293.
- [90] P.K. Stys et al. “Will the real multiple sclerosis please stand up?” In: *Nat Rev Neurosci* 13.7 (2012), pp. 507–514.
- [91] K.T. Tan et al. “Learning Graphical Models With Hubs”. In: *Journal of Machine Learning Research* 15 (2014), pp. 3297–3331.
- [92] L. Thurstone. *The vectors of mind*. Chicago: University of Chicago Press, 1935.
- [93] R. Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society, Series B* 58 (1996), pp. 267–288.
- [94] Ry. Tibshirani and J. Taylor. “The solution path of the generalized Lasso”. In: *Annals of Statistics* 39.3 (2011), pp. 1335–1371.
- [95] M. Uhlen et al. “Proteomics. Tissue-based map of the human proteome.” In: *Science* 347.6220 (2015).
- [96] W.N. Venables and Ripley. B.D. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: <http://www.stats.ox.ac.uk/pub/MASS4>.
- [97] A.G. Vlassenko et al. “Imaging and cerebrospinal fluid biomarkers in early preclinical alzheimer disease”. In: *Ann Neurol* 80.3 (2016), pp. 379–387.
- [98] J. Weeding. “Bayesian Measurement Error Modeling with Application to the Area Under the Curve Summary Measure”. PhD thesis. Montana State University, 2016.
- [99] B.B. Weksler et al. “Blood-brain barrier-specific properties of a human adult brain endothelial cell line”. In: *Faseb J.* 19.13 (2005), pp. 1872–1874.

- [100] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: <http://ggplot2.org>.
- [101] D. Witten, A. Shojaie, and F. Zhang. “The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping”. In: *Technometrics* 56.1 (2014), pp. 112–122.
- [102] D. Witten and R. Tibshirani. “A framework for feature selection in clustering”. In: *Journal of the American Statistical Association* 105.490 (2010), pp. 713–726.
- [103] D. Witten, R. Tibshirani, and T Hastie. “A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis”. In: *Biometrika* 10 (2009), pp. 515 –534.
- [104] M.N. Wright and A. Ziegler. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R”. In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.
- [105] M. Yuan and Y. Lin. “Model selection and estimation in regression with grouped variables”. In: *Journal of the Royal Statistical Society Series B* 68.1 (2007), pp. 49–57.
- [106] Y. Zhang et al. “An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex”. In: *J Neurosci* 34.36 (2014), pp. 11929 –11947.
- [107] D. Zosso et al. “An Efficient Primal-Dual Method for the Obstacle Problem”. In: *Journal of Scientific Computing* 73 (2017), pp. 1–22.

- [108] H. Zou and T. Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society Series B* 67.2 (2005), pp. 301–320.