TECHNISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

Sufficient Dimension Reduction for Longitudinal Data

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Statistik und Wirtschaftsmathematik

unter der Anleitung von

Univ.Prof. Efstathia Bura PhD

eingereicht von

Roman Parzer
Matrikelnummer 01604879

ausgeführt am Institut für Stochastik und Wirtschaftsmathematik
der Fakultät für Mathematik und Geoinformation
an der Technischen Universität Wien

Wien, am 07.09.2021

(Unterschrift Verfasser) (Unterschrift Betreuerin)

Kurzfassung

In dieser Arbeit wird ein neuer Ansatz zur suffizienten Dimensionsreduktion für longitudinal gemessene Prädiktoren und eine reellwertige Zielvariable untersucht. Die meisten bestehenden Reduktionsverfahren, die die longitudinale Struktur ignorieren, können zu einem Informationsverlust führen, da die Zeitinformation nicht genutzt wird. Wir stellen zunächst eine der ersten linearen Reduktionsmethoden, den Sliced Inverse Regression (SIR) Algorithmus, und seine Anpassung für Longitudinaldaten (LSIR) vor. Letztere wird als Benchmark für die Prognoseleistungen der Reduktionen dienen. Anschließend definieren wir das Structured Time-Dependent Inverse Regression (STIR) Modell, das den bedingten Mittelwert der Prädiktoren in Abhängigkeit von der Zielvariable mit Hilfe von Funktionen der Zeit und der Zielvariable modelliert. Für dieses Modell leiten wir Kleinste-Quadrate und Maximum-Likelihood-Schätzer der Parameter her und finden eine Reduktion der Marker, wobei der Zeiteffekt im Modell berücksichtigt wird. Die Vorteile des Modells liegen darin, dass verschiedene Zeitpunkte für jedes Individuum modelliert werden können und dass eine einfachere Interpretation der Reduktion ermöglicht wird, da nur die Marker reduziert werden. Wir untersuchen die Genauigkeit der Parameterschätzer und die Prognosefähigkeit der Reduktionen für dieses Modell in einer umfangreichen Simulationsstudie. In den meisten Simulationsszenarien und für einen realen Datensatz ist STIR konkurrenzfähig mit dem longitudinal angepassten SIR-Algorithmus und mit Standardregressionsmethoden, die die vektorisierten, nicht reduzierten Prädiktoren verwenden. In bestimmten Simulationsszenarien für eine binäre Zielvariable, bei denen das erste und zweite Moment der Prädiktoren von der Zielvariable abhängen, übertrifft STIR die anderen Methoden in der Vorhersageleistung deutlich.

Abstract

This thesis explores a new sufficient dimension reduction approach for longitudinally measured predictors and a real response. Most existing reduction techniques ignore the longitudinal structure and can lead to a loss of information, since the time information is not used. We first introduce one of the first linear reduction methods, the Sliced Inverse Regression (SIR) algorithm, and its adaptation for longitudinal data (LSIR), which serves as benchmark for the predictive performance of the reductions. Then, we define the Structured Time-Dependent Inverse Regression (STIR) model, which models the conditional mean of the predictors given the response using functions of time and the response. For this model we derive least squares and maximum likelihood based parameter estimates, and find a reduction of the markers that accounts for the time effect. Advantages of the model are that different time points for individuals can be modeled and the reduction of markers only allows an easier interpretation of the reduction. We assess the estimation accuracy of the parameter estimates and the predictive ability of the reductions for this model in extensive simulation studies. Throughout most of the simulation settings and on a real data set, STIR is competitive with the longitudinally adapted SIR algorithm and to standard regression methods using the vectorized unreduced predictors. In certain simulation settings for a binary response, where the first and second moment of the predictors relate with the response, STIR excels in predictive performance against the other methods.

Acknowledgement

I would like to thank Prof. E. Bura for giving me the opportunity to be part of this STIR research project and for supervising this thesis, as well as R. Pfeiffer and M. Song, the other co-authors of the project.

Furthermore, I thank L. Neubauer and L. Riess, two of my fellow students, for their companionship and collaboration throughout my entire study.

Finally, I thank my parents for providing me with more than enough support to pursue this master degree with full focus.

The thesis was supported in part by Prof. Bura's TU Wien start-up fund.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 7. September 2021

Roman Parzer

Contents

1. Introduction

In many studies aiming to find strong indicators of diseases, multiple biomarkers are measured, because there is no single biomarker with high classification accuracy. Often, these measurements are also taken repeatedly over time, e.g. in cohort studies during follow up. In practice, the sample size of such studies is often rather modest, which limits the use of statistical methods that estimate many parameters or rely on asymptotic properties. Dimension reduction techniques, such as Sliced Inverse Regression (SIR) by Li [15], aim to combine the relevant information of multiple markers into a lower dimensional score, that contains sufficient information for the regression of the outcome on the markers.

However, in the case of longitudinally measured predictors, or matrix-valued predictors in general, the data structure can contain additional useful information and accommodating it is beneficial in modeling.

[17] proposed and studied first moment based dimension reductions by assuming that the first and second moment of the predictors can be separated into a Kronecker product of time and marker specific components, reducing the complexity of the first-moment sufficient dimension reduction (FMSDR) space. They also proposed the LSIR algorithm, an extension of SIR for longitudinally measured predictors. In simulations and a real data set, the resulting reduction yielded better predictive performance than the SIR algorithm applied to the vectorized predictors.

In another work, [18], assumed a Kronecker structure only for the first moment without requiring a specific structure for the covariance. Under a linear model framework, they proposed and studied computationally efficient least squares based estimates of sufficient reductions.

This thesis explores a new sufficient dimension reduction approach for longitudinally measured predictors and a real response based on modeling their first moment using known functions of time and the response. The method is called Structured Time-Dependent Inverse Regression (STIR). Using a least squares based estimator of coefficients, the derived reduction reduces the markers, while accounting for the effect of time.

In Chapter 2, the theoretical setting for dimension reduction is defined and, after a short overview of multivariate linear regression, the first linear reduction methods, the SIR algorithm, and its adaptation to longitudinal data are described. LSIR serves as benchmark for the predictive performance of the reductions later in simulations. We define the STIR model in Chapter 3, and derive least-squares and maximum likelihood based parameter estimates for a sufficient reduction. We also provide a hypothesis test for variable selection in that model. Advantages of this model are that different time points for individuals can be modeled and the reduction of markers allows an easier interpretation of the reduction.

We assess the estimation accuracy of the parameter estimates and the predictive ability of the reductions for this model in extensive simulation studies in Chapter 4 and on a real data set from a study on brain cancer in Chapter 5. Throughout most of the simulation settings and on the real data set, STIR is competitive to the longitudinally adapted SIR-algorithm and to standard regression methods using the vectorized unreduced predictors. In certain simulation settings for a binary response, where the first and second moment of the predictors relate with the response, STIR excels in predictive performance, beating the other methods by far.

The STIR method was developed by Song, Bura, Parzer and Pfeiffer in [22]. This thesis serves as an extension of that paper, providing more detailed theoretical derivations of estimators and their properties, and deriving a maximum likelihood (ML) based parameter estimate in the model. Using this ML estimator to calculate the reductions did not yield an advantage in predictive performance over the more simple least squares based estimator. The derivation of an alternative representation of the variance-covariance matrix of the estimator was also practically relevant to allow a more efficient calculation. Additional important contributions of this thesis are the efficient computer implementation of the STIR method and carrying out extensive simulations. The source code was implemented in `R` [19] and can be downloaded from `https://github.com/RomanParzer/STIR-Functions-and-Simulations/releases`.

2. Theory on Sufficient Dimension Reduction

In this section we will introduce some general theory of sufficient dimension reduction (SDR) and then, after a short overview of multivariate linear regression, present one of the first linear reduction methods, the SIR algorithm, and its adaptation to longitudinal data.

We start by introducing the notation in this thesis. For matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{k\times l}$ we let $\mathbf{A}' \in \mathbb{R}^{n\times m}$ denote the transpose of $\mathbf{A}$, $\text{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$ the vectorized version of $\mathbf{A}$, where the columns of $\mathbf{A}$ are stacked to form $\text{vec}(\mathbf{A})' = (A_{11}, A_{21}, \dots, A_{m1}, A_{12}, \dots, A_{mn})$, and $\mathbf{A} \otimes \mathbf{B}$ the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ given by the block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} A_{11}\mathbf{B} & \cdots & A_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ A_{m1}\mathbf{B} & \cdots & A_{mn}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{mk\times nl}.$$

See Lemma A.2 in the Appendix for some properties of this Kronecker product, which will be used often in this work. If not mentioned otherwise, $\mathbf{A}_{i\cdot} \in \mathbb{R}^n$ will be the i-th row of $\mathbf{A}$ (as column vector) and $\mathbf{A}_{\cdot j} \in \mathbb{R}^m$ will be the j-th column of $\mathbf{A}$.

Suppose we have two random quantities $\mathbf{X} \in \mathbb{R}^p$, the predictors, and $Y \in \mathbb{R}$, the response. In regressing the response on the predictors involves deducing the information contained in $\mathbf{X}$ relevant to Y.

Definition 2.1 (Reduction)**.** *For two (jointly) random quantities $\mathbf{X} \in \mathbb{R}^p, Y \in \mathbb{R}$ a reduction function, or simply reduction, is a function $R : \mathbb{R}^p \to \mathbb{R}^d$ with $d \ll p$, such that*

$$F(Y|\mathbf{X}) = F(Y|R(\mathbf{X})), \tag{2.1}$$

where $F(Y \mid \cdot)$ is the conditional cumulative distribution function of the response given the predictors. If R is linear, that is $R(\mathbf{X}) = \boldsymbol{\eta}'\mathbf{X}$ for some full-rank $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$, R is called a linear reduction.

We aim to find such a reduction because most regression methods work better with fewer explanatory variables that still contain all the relevant information. For a reduction R we have that $R(\mathbf{X})$ is *sufficient* for the regression of Y on $\mathbf{X}$; that is, it contains all the information for the effect that $\mathbf{X}$ has on Y.

It is easy to see[1] that R satisfies (2.1) if and only if

$$F(\mathbf{X}|Y, R(\mathbf{X})) = F(\mathbf{X}|R(\mathbf{X})). \tag{2.2}$$

[1] see Lemma A.1 in the Appendix with a proof for the case of a joint continuous distribution.

So we can also find the sufficient reduction R by finding a sufficient statistic for Y after assuming a parametric model, treating Y as a parameter. This approach is called *model-based* dimension reduction.

Later in Chapter 3 we will build a parametric model for the inverse regression of $\mathbf{X}$ on Y and try to find a reduction of $\mathbf{X}$ from this model. For this inverse regression we will use an ordinary least squares (OLS) estimator for parameter estimation. Since $\mathbf{X}$ is the multivariate response in inverse regression, we introduce the Multivariate Linear Regression Model in the next section.

2.1. Multivariate Linear Regression

This section is based on [11]. It is an extension of the classical linear regression model, where a multivariate response $\mathbf{Y} = (Y_1, \ldots, Y_m)' \in \mathbb{R}^m$ is regressed on p predictor variables $\mathbf{X} = (X_1, \ldots, X_p)' \in \mathbb{R}^p$. In scalar form the model with n observations can be written as follows.

Definition 2.2 (Multivariate Linear Regression Model)**.** *Assume*

$$y_{ik} = \sum_{j=1}^{p} x_{ij} b_{jk} + e_{ik}, \quad i = 1 \ldots, n, k = 1, \ldots, m, \tag{2.3}$$

where

- $y_{ik} \in \mathbb{R}$ *is the* k*-th response for observation* i*,*
- $x_{ij} \in \mathbb{R}$ *is the* i*-th observation of the* j*-th predictor,*
- $b_{jk} \in \mathbb{R}$ *is the* j*-th predictor's regression coefficient for the* k*-th response, and*
- $e_{ik} \in \mathbb{R}$ *is the* i*-th error term for the* k*-th response.*

In addition to (2.3) *we assume that*

1. y_{ik} *and* x_{ij} *are observed (known) random variables,*
2. *the predictor matrix* $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$ *has full rank* p *(a.s.),*
3. $\mathbf{b}_k = (b_{1k}, \ldots, b_{pk})' \in \mathbb{R}^p$ *are unknown, non-random constants, and*
4. $(e_{i1}, \ldots, e_{im}) \overset{iid}{\sim} N(\mathbf{0}_m, \boldsymbol{\Delta})$ *are unobserved Gaussian error vectors independent of the predictors* x_{ij}*, with* $\boldsymbol{\Delta} \in \mathbb{R}^{m \times m}$ *positive definite and unknown.*

In the last statement, $\mathbf{0}_m$ denotes the zero vector of dimension m. The meaning of this statement is that within the same observation the errors for modeling the m responses are correlated via $\boldsymbol{\Delta}$, but they are independent across observations. The same is true for the

responses y_{ik} when considering the conditional model, where all the predictors x_{ij} are given. The assumption on the rank of the predictor matrix ensures that there is no redundant information in the explanatory variables. For a model with intercept, we can set $x_{i1} = 1$ for all $i = 1, \dots, n$ and interpret b_{1k} as the intercept for the k-th response.

Equivalently, model (2.2) can be written in matrix form as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{2.4}$$

with

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_m \end{pmatrix} \in \mathbb{R}^{p \times m},$$
$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}, \quad \mathbf{E} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}.$$

The distributional assumption on the errors translates to

$$\operatorname{vec}(\mathbf{E}) \sim N(\mathbf{0}_{nm}, \boldsymbol{\Delta} \otimes \mathbf{I}_n),$$
$$\operatorname{vec}(\mathbf{Y}) | \mathbf{X} \sim N((\mathbf{I}_m \otimes \mathbf{X}) \operatorname{vec}(\mathbf{B}), \boldsymbol{\Delta} \otimes \mathbf{I}_n),$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.

The OLS problem in (2.4) is to find an estimate of $\mathbf{B}$ that minimizes the function

$$S(\mathbf{B}) = \|\mathbf{Y} - \mathbf{XB}\|_F^2 = \operatorname{tr}(\mathbf{Y}'\mathbf{Y}) - 2\operatorname{tr}(\mathbf{Y}'\mathbf{XB}) + \operatorname{tr}(\mathbf{B}'\mathbf{X}'\mathbf{XB}), \tag{2.5}$$

with $\|\mathbf{A}\|_F$ denoting the Frobenius norm and $\operatorname{tr}(\mathbf{A})$ the trace of a matrix $\mathbf{A}$, and using that $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}'\mathbf{A})$. For calculating the derivative we use the facts on matrix derivatives in Lemma A.3 in the Appendix. By the first order condition

$$\frac{\partial S(\mathbf{B})}{\partial \mathbf{B}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{XB} \stackrel{!}{=} 0,$$

the OLS estimator is given by

$$\widehat{\mathbf{B}} = \operatorname{argmin}_{\mathbf{B} \in \mathbb{R}^{p \times k}} S(\mathbf{B}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.6}$$

If we let $\widehat{\mathbf{b}}_k$ and $\mathbf{y}_k$ denote the k-th columns of $\widehat{\mathbf{B}}$ and $\mathbf{Y}$, we can express $\widehat{\mathbf{b}}_k$ as

$$\widehat{\mathbf{b}}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_k, \quad k = 1, \dots, m.$$

Thus, the OLS estimate can be obtained by fitting m univariate linear regressions with the same predictors in parallel, but with the m regression errors being correlated via $\boldsymbol{\Delta}$.

Next we want to take a look at the ML estimator in this model. Let $\mathbf{y}_{i\cdot} \in \mathbb{R}^m$ and $\mathbf{x}_{i\cdot} \in \mathbb{R}^p$ denote the i-th row of $\mathbf{Y}$ and $\mathbf{X}$. Under assumption 4 in model (2.2), we have

$$\mathbf{y}_{i\cdot} | \mathbf{x}_{i\cdot} \sim N(\mathbf{B}'\mathbf{x}_{i\cdot}, \boldsymbol{\Delta}),$$

independent over $i = 1, \ldots, n$. Therefore, the log-likelihood is given by

$$\ell(\mathbf{B}|\mathbf{Y}, \mathbf{X}) = -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{i.} - B'\mathbf{x}_{i.})'\boldsymbol{\Delta}^{-1}(\mathbf{y}_{i.} - B'\mathbf{x}_{i.}) + c,$$

where c is a constant not depending on $\mathbf{B}$. Noting that

$$\begin{aligned}(\mathbf{y}_{i.} - \mathbf{B}'\mathbf{x}_{i.})'\boldsymbol{\Delta}^{-1}(\mathbf{y}_{i.} - \mathbf{B}'\mathbf{x}_{i.}) &= \operatorname{tr}(\boldsymbol{\Delta}^{-1}(\mathbf{y}_{i.} - \mathbf{B}'\mathbf{x}_{i.})(\mathbf{y}_{i.} - \mathbf{B}'\mathbf{x}_{i.})') \\ &= -\operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{y}_{i.}\mathbf{x}_{i.}'\mathbf{B}) - \operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{B}'\mathbf{x}_{i.}\mathbf{y}_{i.}') + \operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{B}'\mathbf{x}_{i.}\mathbf{x}_{i.}'\mathbf{B}) + c,\end{aligned}$$

and again by Lemma A.3 in the Appendix,

$$\begin{aligned}\frac{\partial \ell(\mathbf{B}|\mathbf{Y}, \mathbf{X})}{\partial \mathbf{B}} &= -\frac{1}{2}\sum_{i=1}^{n}\Big(-2\mathbf{x}_{i.}\mathbf{y}_{i.}'\boldsymbol{\Delta}^{-1} + 2\mathbf{x}_{i.}\mathbf{x}_{i.}'\mathbf{B}\boldsymbol{\Delta}^{-1}\Big) \\ &= (\sum_{i=1}^{n}\mathbf{x}_{i.}\mathbf{y}_{i.}')\boldsymbol{\Delta}^{-1} - (\sum_{i=1}^{n}\mathbf{x}_{i.}\mathbf{x}_{i.}')\mathbf{B}\boldsymbol{\Delta}^{-1} \\ &= \mathbf{X}'\mathbf{Y}\boldsymbol{\Delta}^{-1} - \mathbf{X}'\mathbf{X}\boldsymbol{\Delta}^{-1},\end{aligned}$$

so by the first order condition, the ML estimator agrees with the OLS estimator $\widehat{\mathbf{B}}$.

This result can also be obtained, as in [14, Section 6.2], by vectorizing the transposed version of (2.4). The resulting model has the structure of a (univariate) generalized linear model. After unvectorizing, the generalized OLS estimator in that model, which is known to agree with the ML estimator, agrees with our $\widehat{\mathbf{B}}$ in (2.6).

In model (2.2), we have

$$\begin{aligned}\mathbb{E}[\widehat{\mathbf{B}}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{B}, \\ \operatorname{Cov}(\operatorname{vec}(\widehat{\mathbf{B}})|\mathbf{X}) &= \operatorname{Cov}((\mathbf{I}_m \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\operatorname{vec}(\mathbf{Y})|\mathbf{X}) \\ &= (\mathbf{I}_m \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\underbrace{\operatorname{Cov}(\operatorname{vec}(\mathbf{Y})|\mathbf{X})}_{=\boldsymbol{\Delta}\otimes\mathbf{I}_n}(\mathbf{I}_m \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= \boldsymbol{\Delta} \otimes (\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Thus,

$$\operatorname{vec}(\widehat{\mathbf{B}})|\mathbf{X} \sim N(\operatorname{vec}(\mathbf{B}), \boldsymbol{\Delta} \otimes (\mathbf{X}'\mathbf{X})^{-1}).$$

Let $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}$ be the fitted values, and

$$\begin{aligned}\mathbf{U} &= \mathbf{Y} - \widehat{\mathbf{Y}} \\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{E}\end{aligned}$$

be the residuals. Then,

$$\begin{aligned}\mathbb{E}[(\mathbf{U}'\mathbf{U})_{jl}] &= \mathbb{E}[(\mathbf{E}'(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{E})_{jl}] = \mathbb{E}[\mathrm{tr}(\mathbf{e}'_{\cdot j}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{e}_{\cdot l})] \\ &= \mathrm{tr}((\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\underbrace{\mathbb{E}[\mathbf{e}_{\cdot l}\mathbf{e}'_{\cdot j}]}_{=\boldsymbol{\Delta}_{jl}\cdot\mathbf{I}_n}) \\ &= \mathrm{tr}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\cdot\boldsymbol{\Delta}_{jl} = (n-p)\boldsymbol{\Delta}_{jl}, \quad j,l = 1,\dots,m,\end{aligned}$$

where $\mathbf{e}_{\cdot j}$ denotes the j-th column of $\mathbf{E}$. Therefore, the unknown $\boldsymbol{\Delta}$ can be estimated by the unbiased estimator

$$\widehat{\boldsymbol{\Delta}} = \frac{1}{n-p}\mathbf{U}'\mathbf{U}.$$

This section serves as a warm-up for our STIR model in Chapter 3, where we use similar methods and estimators. When we only assume the error has the same mean and covariance structure, but it is not normally distributed, then the OLS estimator stays the same with the same mean and covariance structure, but it is also not normal. Without assuming a specific distribution, we can not derive a ML estimator. The estimator of the error covariance, $\widehat{\boldsymbol{\Delta}}$, remains unbiased for $\boldsymbol{\Delta}$.

2.2. Linear Sufficient Dimension Reductions

In the coming section we introduce some basic results for linear sufficient dimension reductions.

We go back to the formulation of Definition 2.1 with two (jointly) random quantities $\mathbf{X} \in \mathbb{R}^p, Y \in \mathbb{R}$, where the (marginal) covariance matrix $\boldsymbol{\Sigma}_x$ of $\mathbf{X}$ is full rank, and want to find a linear reduction $R: \mathbb{R}^p \to \mathbb{R}^d: \mathbf{X} \mapsto \boldsymbol{\eta}'\mathbf{X}$ for some full-rank $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ satisfying

$$F(Y|\mathbf{X}) = F(Y|\boldsymbol{\eta}'\mathbf{X}), \tag{2.7}$$

where, again, $F(Y \mid \cdot)$ is the conditional cumulative distribution function of the response given the predictors.

It is easy to see that there is no unique $\boldsymbol{\eta}$ satisfying (2.7). For example when scaling the columns of $\boldsymbol{\eta}$ by non-zero constants or for a change of basis $\tilde{\boldsymbol{\eta}} = \mathbf{A}\boldsymbol{\eta}$ for a full rank $\mathbf{A} \in \mathbb{R}^{d\times d}$, the conditional distribution $Y|\tilde{\boldsymbol{\eta}}'\mathbf{X}$ does not change at all.

In [15], Li defines the span of the columns $\mathrm{span}(\boldsymbol{\eta})$ as an *effective dimension reduction space*, since any basis of $\mathrm{span}(\boldsymbol{\eta})$ is a sufficient dimension reduction. However, also the span is not unique, since adding additional columns to $\boldsymbol{\eta}$ preserves the sufficient-reduction property (2.7).

In [5, Section 6.4, pp 108–112], Cook found that under mild conditions the *central dimension reduction space*, defined as intersection of all effective dimension reduction spaces

$$\mathcal{S}_{Y|\mathcal{X}} = \bigcap_{\boldsymbol{\eta} \text{ satisfying (2.7)}} \mathrm{span}(\boldsymbol{\eta}),$$

is also sufficient. Moreover, it exists and is unique if the support of $\mathbf{X}$ is convex.

From now on we assume that $\mathcal{S}_{Y|\mathcal{X}}$ exists and let $\boldsymbol{\eta}$ be a basis of that space, i.e. $\operatorname{span}(\boldsymbol{\eta}) = \mathcal{S}_{Y|\mathcal{X}}$. The first SDR approaches tried to estimate the reduction space by using *core* matrices $\boldsymbol{\Omega}$ calculated from the moments of the conditional distribution $\mathbf{X}|Y$ with $\operatorname{span}(\boldsymbol{\Omega}) \subseteq \operatorname{span}(\boldsymbol{\eta})$ [18]. These approaches are called *moment-based* SDR and one such approach is explained in the remainder of this section.

Li derived the following important result using the first moment of $\mathbf{X}|Y$ in [15, Theorem 3.1], which we state without proof.

Theorem 2.3 (Li, 1991)**.** *Suppose that*

$$Y = f(\boldsymbol{\eta}_1'\mathbf{X}, \dots, \boldsymbol{\eta}_d'\mathbf{X}, \varepsilon), \tag{2.8}$$

where ε is independent of $\mathbf{X}$ and $f : \mathbb{R}^{d+1} \to \mathbb{R}$ is an unknown link function. If for any $\mathbf{b} \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}[\mathbf{b}'\mathbf{X}|\boldsymbol{\eta}_1'\mathbf{X}, \dots, \boldsymbol{\eta}_d'\mathbf{X}]$ is linear in $\boldsymbol{\eta}_1'\mathbf{X}, \dots, \boldsymbol{\eta}_d'\mathbf{X}$, then for all y in the support of Y

$$\boldsymbol{\Sigma}_x^{-1}(\mathbb{E}[\mathbf{X}|Y = y] - \mathbb{E}[\mathbf{X}]) \in \operatorname{span}(\boldsymbol{\eta}), \tag{2.9}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_d)$.

In [25] it is shown that the condition (2.8) is equivalent to (2.7), i.e. to the general linear regression problem. The second condition, called *linear design condition*, is a property of the marginal distribution of $\mathbf{X}$. It is satisfied for an elliptically symmetric distribution such as the normal distribution [15].

From this theorem, we can define the first moment sufficient dimension reduction space as

$$\mathcal{S}_{\text{FMSDR}} = \boldsymbol{\Sigma}_x^{-1} \operatorname{span}(\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}]). \tag{2.10}$$

To recover all of $\mathcal{S}_{Y|\mathcal{X}}$, meaning $\mathcal{S}_{\text{FMSDR}} = \mathcal{S}_{Y|\mathcal{X}}$, at least one of $\operatorname{Var}(\operatorname{Var}[\mathbf{b}'\mathbf{X}|Y])$ or $\operatorname{Var}(\mathbb{E}[\mathbf{b}'\mathbf{X}|Y])$ must be positive for every non-zero $\mathbf{b} \in \mathcal{S}_{Y|\mathcal{X}}$ [20, Section 2]. This holds, for example, when the predictors $\mathbf{X}|Y$ have a conditional multivariate normal distribution where the covariance does not depend on Y.

Theorem 2.3 tells us that the centered inverse regression curve falls into the effective dimension reduction space scaled by $\boldsymbol{\Sigma}_x$. However, the inverse regression curve is difficult to estimate in general.

Following the arguments in [15, Section 3], let us take a look at the standardized predictors $\mathbf{Z} = \boldsymbol{\Sigma}_x^{-1/2}(\mathbf{X} - \mathbb{E}[\mathbf{X}])$. If we let $\boldsymbol{\gamma}_k = \boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\eta}_k, k = 1, \dots, d$, then it is easy to see that the conditions of Theorem 2.3 are equivalent to

$$\begin{aligned} & Y = f(\boldsymbol{\gamma}_1'\mathbf{Z}, \dots, \boldsymbol{\gamma}_d'\mathbf{Z}, \varepsilon), \quad \mathbf{Z}, \varepsilon \text{ independent}, \\ \forall \mathbf{b} \in \mathbb{R}^p : \quad & \mathbb{E}[\mathbf{b}'\mathbf{Z}|\boldsymbol{\gamma}_1'\mathbf{Z}, \dots, \boldsymbol{\gamma}_d'\mathbf{Z}] \text{ is linear in } \boldsymbol{\gamma}_1'\mathbf{Z}, \dots, \boldsymbol{\gamma}_d'\mathbf{Z}. \end{aligned}$$

For $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$, applying Theorem 2.3 to $\mathbf{Z}$ gives

$$\mathbb{E}[\mathbf{Z}|y] \in \operatorname{span}(\boldsymbol{\gamma}) \text{ for all } y \text{ in the support of } Y.$$

If we take any $\mathbf{u} \in \operatorname{span}(\boldsymbol{\gamma})^{\perp}$ orthogonal to $\operatorname{span}(\boldsymbol{\gamma})$, we therefore have $\mathbb{E}[\mathbf{Z}|Y]'\mathbf{u} = 0$ and

$$\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y])\mathbf{u} = \mathbb{E}[\mathbb{E}[\mathbf{Z}|Y]\underbrace{\mathbb{E}[\mathbf{Z}|Y]'\mathbf{u}}_{=0}] = 0,$$

which implies

$$\operatorname{span}(\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y])) \subseteq \operatorname{span}(\mathbb{E}[\mathbf{Z}|Y]) \subseteq \operatorname{span}(\boldsymbol{\gamma}). \tag{2.11}$$

In [8], which was made publicly available in [9], Eaton gives the following proposition.

Proposition 2.4 ([9, Prop. 2.7])**.** *Let $\widetilde{\mathbf{X}}$ be a random vector in a linear space V with an inner product and suppose that $\operatorname{Cov}(\widetilde{\mathbf{X}}) = \boldsymbol{\Sigma}_x$ exists. Then*

$$\mathbb{P}(\widetilde{\mathbf{X}} \in \mathbb{E}[\widetilde{\mathbf{X}}] + \operatorname{span}(\boldsymbol{\Sigma}_x)) = 1. \tag{2.12}$$

Applying this proposition to the random variable $\mathbb{E}[\mathbf{Z}|Y]$ obtains that the span of $\mathbb{E}[\mathbf{Z}|Y]$ lies in $\operatorname{span}(\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y]))$ with probability 1. Together with (2.11) yields that $\operatorname{span}(\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y])) = \operatorname{span}(\mathbb{E}[\mathbf{Z}|Y])$ (with probability 1).

This way, we get the following corollary for the original $\mathbf{X}$ variables by scaling back.

Corollary 2.5. *Under the conditions of Theorem 2.3, we have*

$$\mathcal{S}_{\text{FMSDR}} = \boldsymbol{\Sigma}_x^{-1/2} \operatorname{span}(\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y])) \subseteq \operatorname{span}(\boldsymbol{\eta}). \tag{2.13}$$

This corollary is the foundation for the Sliced Inverse Regression (SIR) algorithm given in the next section.

2.3. SIR Algorithm

In this section we will introduce the SIR algorithm and give an illustrative example for its application.

Starting from Corollary 2.5, Li proposes the following algorithm to find a linear reduction by estimating $\operatorname{Cov}(\mathbb{E}[\mathbf{Z}|Y])$ [15].

Algorithm 2.6 (SIR)**.** *Suppose we have data $(\mathbf{x}_i, y_i), i = 1, \ldots n$. For a chosen H, the number of slices for Y, and k, the estimated dimension of the reduction space:*

1. *Standardize the predictors to get $\mathbf{z}_i = \widehat{\boldsymbol{\Sigma}}_x^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\widehat{\boldsymbol{\Sigma}}_x$ and $\bar{\mathbf{x}}$ denote the empirical covariance matrix and mean of $\mathbf{X}$, respectively.*

2. *Divide the range of* Y *into* H *slices* $S_1, \dots, S_H$ *and let* $\hat{p}_h = (1/n) \sum_{i=1}^{n} \mathbb{1}(y_i \in S_h)$ *be the portion of* y_i*s in the* h*-th slice, where* $\mathbb{1}$ *is the indicator function.*
3. *Within each slice, compute the sample mean* $\hat{\mathbf{m}}_h$ *of the* $\mathbf{z}_i$*s:* $\hat{\mathbf{m}}_h = (1/n\hat{p}_h) \sum_{y_i \in S_h} \mathbf{z}_i$.
4. *Let* $\widehat{\boldsymbol{\Delta}} = \sum_{h=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h'$ *be the estimated covariance of* $\mathbb{E}[\mathbf{Z}|Y]$ *and* $\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_k$ *be the eigenvectors corresponding to the* k *largest eigenvalues of* $\widehat{\boldsymbol{\Delta}}$.
5. *Output* $\hat{\boldsymbol{\eta}}_j = \widehat{\boldsymbol{\Sigma}}_x^{-1/2} \hat{\boldsymbol{\gamma}}_j$ *for* $j = 1, \dots, k$.

Steps 2 and 3 aim to (roughly) estimate the inverse regression curve $\mathbb{E}[\mathbf{Z}|Y]$. Only the main orientation of the estimated curve is needed, so [15] advocates to use the sliced Y due to its simplicity over more complex non-parametric regression methods such as smoothing splines or nearest neighbour. The k eigenvectors in step 4 build the most important subspace to describe the inverse regression curve $\mathbb{E}[\mathbf{Z}|Y]$. After scaling back in step 5, we can estimate the effective dimension reduction space by $\text{span}(\hat{\boldsymbol{\eta}}_j, j = 1, \dots, k)$.

In general, when we want to fit a regression model between predictors $\mathbf{X} \in \mathbb{R}^p$ and a response $Y \in \mathbb{R}$, we can now first use SIR to reduce the predictors and then fit some regression model

$$Y = g(\widehat{\boldsymbol{\eta}}'\mathbf{X}, \varepsilon),$$

where $\widehat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_1, \dots, \hat{\boldsymbol{\eta}}_k)$. With this model we can also predict responses for new $\mathbf{X}$-observations by plugging in the reduced values in the regression model. This reduced regression model can outperform a model using the unreduced predictors, because for large p finding the right model can be challenging. We illustrate this procedure on an easy example (adapted from [2]).

Example 2.7. *We generate* $n = 200$ *observations from*

$$\begin{aligned} &\mathbf{X} \sim N(0_{10}, \mathbf{I}_{10}), \\ &Y = (X_1 + X_2 + 1)^3 + N(0, 0.5) \quad \textit{(unknown)}. \end{aligned}$$

In order to find a suitable model for the regression of Y *on* $\mathbf{X}$ *we first have a look at the marginal plots of* Y *against the* X_j*s. The one for* X_1 *is shown in the left plot in Figure 2.1. There seems to be a relation, but the exact functional form can not be determined. For* X_2 *we obtain a similar picture, while there is no visible relation for the remaining variables.*

Using `R`*, we fit a linear model with all ten variables, where we also include the interaction term between* X_1 *and* X_2*, as well as the squares and cubes of those two variables. This model leads to a coefficient of determination of* $R^2 = 0.9297$*. Applying model selection techniques, such as stepwise regression using some information criterion or LASSO, could improve the predictive performance, but the* R^2 *would not increase. Also, there is an unknown effect on the validity of inference after all this data processing.*

In comparison, we can apply the `R` *function* `dr()` *in the* `dr`*-package [23] with* `method="sir"` *on the original predictors. A built-in test estimates the reduction dimension to be* 1*. The*

right plot shows a scatter plot of that one-dimensional reduced predictor against Y. We see a cubic relation and fitting a linear model using this reduced predictor, as well as the square and cube of it, yields a coefficient of determination of $R^2 = 0.9891$.

The algorithm identified that this is a one-dimensional problem and the complexity of modeling Y was drastically reduced by the reduction.

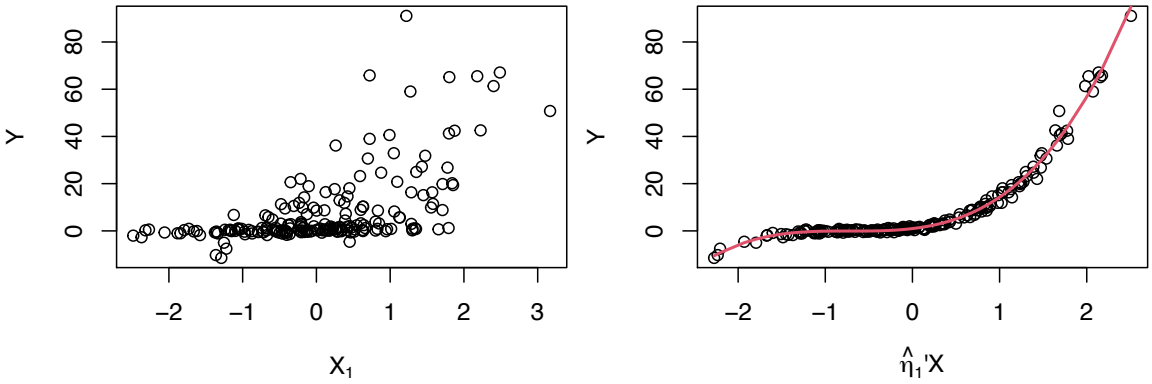

Figure 2.1.: Plots illustrating Example 2.7. Left: Scatterplot of Y against the first component of $\mathbf{X}$. Right: Scatterplot of Y against the one-dimensional reduction found by SIR. The red line gives the fitted values for a linear model using up to cubic terms of that reduction.

In [15, Section 5], Li also gives an argument for root n consistency of the SDR directions obtained by the SIR algorithm. One problem of this algorithm is that if the standardized inverse regression curve falls into a proper subspace of the span of the $\boldsymbol{\gamma}_k$s, then it cannot recover all directions of the effective dimension reduction space [15, Remark 4.5]. An example for this is

$$Y = g(\boldsymbol{\eta}_1'\mathbf{X}) + \varepsilon,$$

with some symmetric function g and with $\boldsymbol{\eta}_1'\mathbf{X}$ symmetric around 0. Then the inverse regression curve is 0 and the algorithm fails to provide a valid estimate for $\boldsymbol{\eta}_1$.

To overcome the inability of SIR to detect certain types of regression relations, Cook and Weisberg proposed the following theorem [7].

Theorem 2.8. *Additionally to the assumptions of Theorem 2.3, assume that* $\mathrm{Var}(\mathbf{X}|\boldsymbol{\eta}'\mathbf{X})$ *is a non-random matrix (constant variance condition). Then*

$$\boldsymbol{\Sigma}_x^{-1}\,\mathrm{span}(\boldsymbol{\Sigma}_x - \mathrm{Var}(\mathbf{X}|Y)) \subseteq \mathrm{span}(\boldsymbol{\eta}). \tag{2.14}$$

This theorem leads to the Sliced Average Variance Estimation (SAVE) algorithm, which also uses the second moment of $\mathbf{X}|Y$.

There are some limitations with all the methods presented in this section. They might not be exhaustive, and they only aim to find linear reductions $R(\mathbf{X}) = \boldsymbol{\eta}'\mathbf{X}$ and miss other functional relations. For example, Bura and Forzani showed in [3] that for $\mathbf{X}|Y \sim N_p(\boldsymbol{\mu}_y, c_y\boldsymbol{\Delta})$, where $c_y \in \mathbb{R}$ is a scaling constant, the minimal sufficient reduction is given by

$$R(\mathbf{X}) = (\boldsymbol{\alpha}'(\mathbf{X} - \mathbb{E}[\mathbf{X}]), (\mathbf{X} - \mathbb{E}[\mathbf{X}])'\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \mathbb{E}[\mathbf{X}])), \tag{2.15}$$

with $\operatorname{span}(\boldsymbol{\alpha}) = \boldsymbol{\Sigma}_x^{-1}\operatorname{span}(\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}])$. So the minimal reduction has a non-linear component. This result also generalizes to elliptically contoured distributions. Here, SIR and other moment-based methods are not exhaustive.

Starting from equation (2.2), we can assume a model and find a reduction by finding a sufficient statistic for Y, treating Y as a parameter. The advantages of this model-based approaches are that after assuming a model for $\mathbf{X}|Y$, we do not need any further assumptions. One can use existing theory for sufficient statistics to obtain exhaustive and (possibly) non-linear reductions, such as Fisher's factorization theorem or exponential families.

2.4. LSIR Algorithm

In this section we go back to a first moment based approach to find a linear reduction R, but now with longitudinally measured predictors, where for each individual we observe the p markers over several time points. We summarize the work by Pfeiffer et al. [17], and introduce their LSIR (Longitudinal Sliced Inverse Regression) algorithm, which extends the SIR algorithm 2.6 to longitudinally measured predictors. In [17], the predictors are modeled as $p \times T$ matrices. To be consistent with the modeling in Chapter 3, we state the results in the transposed version.

At the population level we now have still a single real response $Y \in \mathbb{R}$, but the predictors are $T \times p$ random matrices, $\mathbf{X} \in \mathbb{R}^{T\times p}$, where T is a fixed number of different time points and the entry X_{tj} is the measurement for the j-th marker at the t-th time point. The aim is to estimate the first moment sufficient dimension reduction space $\mathcal{S}_{\text{FMSDR}}$ of the vectorized predictors

$$\mathcal{S}_{\text{FMSDR}} = \boldsymbol{\Sigma}_x^{-1}\operatorname{span}(\operatorname{vec}(\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}])), \tag{2.16}$$

where $\boldsymbol{\Sigma}_x = \operatorname{Cov}(\operatorname{vec}(\mathbf{X}))$.

The main assumption is the following structure of the conditional mean $\mathbf{X}|Y$. We assume

$$\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}] = \boldsymbol{\psi}\mathbf{G}_y\boldsymbol{\phi}' \tag{2.17}$$

$$\iff \operatorname{vec}(\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}]) = (\boldsymbol{\phi} \otimes \boldsymbol{\psi})\operatorname{vec}(\mathbf{G}_y), \tag{2.18}$$

for some $r \times m$ matrix $\mathbf{G}_y$, which depends on Y with $\mathbb{E}[\mathbf{G}_y] = 0$ and $\det(\operatorname{Cov}(\operatorname{vec}(\mathbf{G}_y))) > 0$. The matrix $\boldsymbol{\psi} \in \mathbb{R}^{T\times r}$ captures the mean structure over the time points, and $\boldsymbol{\phi} \in \mathbb{R}^{p\times m}$ captures the mean structure over the markers regardless of time.

In the practically important binary case, i.e. when Y only takes the values 0 or 1, we have $r = m = 1$, implying that both $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are vectors. Otherwise the covariance matrix of $\text{vec}(\mathbf{G}_y)$ cannot be full rank. Then the condition (2.17) is also satisfied, if we let one of $\boldsymbol{\phi}$ or $\boldsymbol{\psi}$ depend on Y. Assume for example that $\text{vec}(\mathbb{E}[\mathbf{X}|Y]) = \boldsymbol{\phi}_y \otimes \boldsymbol{\psi}$, which would mean that $\mathbb{E}[X_{tj}|Y] = \boldsymbol{\phi}_j(Y) \cdot \boldsymbol{\psi}_t$ for $j = 1, \dots, p, t = 1, \dots, T$. So, the conditional mean of each marker only depends on time through a multiplicative scalar, which is the same for all markers in both groups given by Y. If we let $p_y = \mathbb{P}(Y = y)$ for $y = 0, 1$ and using that $\mathbb{E}[\mathbf{X}] = p_0(\boldsymbol{\phi}_0 \otimes \boldsymbol{\psi}) + (1 - p_0)(\boldsymbol{\phi}_1 \otimes \boldsymbol{\psi})$, we obtain

$$\text{vec}(\mathbb{E}[\mathbf{X}|Y = y] - \mathbb{E}[\mathbf{X}]) = (1 - p_y)(\boldsymbol{\phi}_0 - \boldsymbol{\phi}_1) \otimes \boldsymbol{\psi},$$

which shows that condition (2.17) is actually satisfied with $\mathbf{G}_y = (1 - p_y)$ and $\widetilde{\boldsymbol{\phi}} = \boldsymbol{\phi}_0 - \boldsymbol{\phi}_1$ for this example in the binary case.

Let us now move back to the general setting. The second assumption is that the covariance structure can be decomposed as a Kronecker product into a marker and a time component, i.e.

$$\text{Cov}(\text{vec}(\mathbf{X})) = \boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_p \otimes \boldsymbol{\Sigma}_T, \tag{2.19}$$

with $\boldsymbol{\Sigma}_p \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Sigma}_T \in \mathbb{R}^{T \times T}$ both positive definite. This structure implies that

$$\text{Cov}(X_{si}, X_{tj}) = \sigma^p_{ij}\sigma^T_{st}, \quad i, j = 1, \dots, p, \quad s, t = 1, \dots, T,$$

meaning that the correlation of a fixed marker observed at different time points only depends on these time points and not the marker. Similarly, the correlation of two markers observed at the same time point only depends on the two markers but not on the time point. If we let $\mathbf{X}_{\cdot j}$ and $\mathbf{X}_{t\cdot}$ denote the j-th column and the t-th row (as column vector) of $\mathbf{X}$, then (2.19) implies

$$\text{Cov}(\mathbf{X}_{\cdot j}) = \sigma^p_{jj}\boldsymbol{\Sigma}_T$$
$$\text{Cov}(\mathbf{X}_{t\cdot}) = \sigma^T_{tt}\boldsymbol{\Sigma}_p.$$

So, the covariance structure between the T time points depends on the marker only through a multiplicative constant, and the same holds true for the dependence of the covariance structure between the p markers on time.

Assumption (2.19) might be reasonable if the data come from a prospective cohort study, but there are settings in which it is not, e.g. when the longitudinal data arise from a retrospective case-control study [17]. A slightly less restrictive assumption is that

$$\boldsymbol{\Delta} := \mathbb{E}[\text{Cov}(\text{vec}(\mathbf{X})|Y)] = \boldsymbol{\Delta}_p \otimes \boldsymbol{\Delta}_T, \tag{2.20}$$

with $\boldsymbol{\Delta}_p \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Delta}_T \in \mathbb{R}^{T \times T}$ again both positive definite. An easy example for when this condition is satisfied is $\text{Cov}(\text{vec}(\mathbf{X})|Y) = \boldsymbol{\Delta}_p(Y) \otimes \boldsymbol{\Delta}_T(Y)$, where only one of $\boldsymbol{\Delta}_p$ or $\boldsymbol{\Delta}_T$ depends on Y.

By the law of total variance, we have

$$\text{Cov}(\text{vec}(\mathbf{X})) = \mathbb{E}[\text{Cov}(\text{vec}(\mathbf{X})|Y)] + \text{Cov}(\mathbb{E}[\text{vec}(\mathbf{X})|Y]).$$

The mean assumption (2.17) implies that

$$\mathrm{Cov}(\mathbb{E}[\mathrm{vec}(\mathbf{X})|Y]) = (\boldsymbol{\phi} \otimes \boldsymbol{\psi})\,\mathrm{Cov}(\mathrm{vec}(\mathbf{G}_y))(\boldsymbol{\phi} \otimes \boldsymbol{\psi})',$$

which does not have a Kronecker structure in general unless $\mathrm{Cov}(\mathrm{vec}(\mathbf{G}_y))$ itself is decomposed as a Kronecker product with symmetric matrices of dimensions m and r. This gives an idea for assumption (2.20) being less restrictive than (2.19).

The following two theorems are taken from [17].

Theorem 2.9 ([17, Theorem 1])**.** *Suppose that the mean assumption* (2.17) *holds and the vectorized predictors have a Kronecker product covariance structure as in* (2.19)*. Then*

$$\mathcal{S}_{\mathrm{FMSDR}} = \mathrm{span}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\phi} \otimes \boldsymbol{\Sigma}_T^{-1}\boldsymbol{\psi}). \tag{2.21}$$

Theorem 2.10 ([17, Theorem 2])**.** *Suppose that the mean assumption* (2.17) *holds and the conditional vectorized predictors have an expected Kronecker product covariance structure as in* (2.20)*. Then*

$$\mathcal{S}_{\mathrm{FMSDR}} = \mathrm{span}(\boldsymbol{\Delta}_p^{-1}\boldsymbol{\phi} \otimes \boldsymbol{\Delta}_T^{-1}\boldsymbol{\psi}).$$

The main benefit of the two theorems is that under these assumptions, the first moment reduction space also has a Kronecker structure and there are a lot fewer parameters to estimate. The covariance matrix alone in (2.16) has $Tp(Tp+1)/2$ entries, compared to $T(T+1)/2 + p(p+1)/2$ for the two covariance matrices in Theorem 2.9 or 2.10.

Depending on which Theorem's assumptions are satisfied, we can use $\boldsymbol{\Sigma}_x$ or $\boldsymbol{\Delta}$ respectively to scale the predictors. It can be shown by direct calculation, as in the proof of Corollary 3.4 in [6], that when using $\widehat{\boldsymbol{\Delta}}_p$ and $\widehat{\boldsymbol{\Delta}}_T$ instead of $\widehat{\boldsymbol{\Sigma}}_p$ and $\widehat{\boldsymbol{\Sigma}}_T$ in the algorithm given later, the estimated subspace does not change [17].

In the following algorithm we will therefore rather use $\boldsymbol{\Sigma}_x$ than $\boldsymbol{\Delta}$ to scale the data, since it is easier to estimate. Similar to the SIR algorithm 2.6, we will use the standardized variable

$$\mathrm{vec}(\mathbf{Z}) = \boldsymbol{\Sigma}_x^{-1/2}\,\mathrm{vec}(\mathbf{X} - \mathbb{E}[\mathbf{X}]), \tag{2.22}$$

for numeric stability. In the derivation of Corollary 2.5 we saw that

$$\mathrm{span}(\mathrm{Cov}(\mathrm{vec}(\mathbb{E}[\mathbf{Z}|Y]))) = \mathrm{span}(\mathrm{vec}(\mathbb{E}[\mathbf{Z}|Y])) \quad \text{(a.s.)},$$

which implies

$$\mathcal{S}_{\mathrm{FMSDR}} = \boldsymbol{\Sigma}_x^{-1/2}\,\mathrm{span}(\mathrm{vec}(\mathbb{E}[\mathbf{Z}|Y])) = \boldsymbol{\Sigma}_x^{-1/2}\,\mathrm{span}(\mathrm{Cov}(\mathrm{vec}(\mathbb{E}[\mathbf{Z}|Y]))). \tag{2.23}$$

From Theorems 2.9 and 2.10 we know that this span has a Kronecker structure (under the corresponding assumptions). So the adapted SIR algorithm aims to estimate $\boldsymbol{\Sigma}$ and $\mathrm{Cov}(\mathrm{vec}(\mathbb{E}[\mathbf{Z}|Y]))$ as a Kronecker structure and find the subspace this way.

Algorithm 2.11 (LSIR,[17]). *Given a data sample* $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$ *with* $\mathbf{X}_i \in \mathbb{R}^{T\times p}$ *and* $Y_i \in \mathbb{R}$, H, *the number of slices for* Y, *and estimated ranks* r, m *of the time and marker component:*

1. *Scale the predictors by*

$$\operatorname{vec}(\mathbf{Z}_i) = \widehat{\mathbf{\Sigma}}_x^{-1/2} \operatorname{vec}(\mathbf{X}_i - \bar{\mathbf{X}}),$$

where $\bar{\mathbf{X}}$ *is the mean over all* $\mathbf{X}_i$ *and* $\widehat{\mathbf{\Sigma}}_x = \widehat{\mathbf{\Sigma}}_p \otimes \widehat{\mathbf{\Sigma}}_T$ *is found by the following estimate. Let* $\mathbf{x}^i_{\cdot j}$ *be the* j*-th column of* $\mathbf{X}_i$, $\mathbf{x}^i_{t\cdot}$ *the* t*-th row of* $\mathbf{X}_i$ *(as column vector), and* $\bar{\mathbf{x}}_{\cdot j}, \bar{\mathbf{x}}_{t\cdot}$ *the corresponding means over all observations. Then*

$$\widehat{\mathbf{\Sigma}}_{pt} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^i_{t\cdot} - \bar{\mathbf{x}}_{t\cdot})(\mathbf{x}^i_{t\cdot} - \bar{\mathbf{x}}_{t\cdot})', \quad \widehat{\mathbf{\Sigma}}_p = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{\Sigma}}_{pt} \in \mathbb{R}^{p\times p},$$
$$\widehat{\mathbf{\Sigma}}_{Tj} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^i_{\cdot j} - \bar{\mathbf{x}}_{\cdot j})(\mathbf{x}^i_{\cdot j} - \bar{\mathbf{x}}_{\cdot j})', \quad \widehat{\mathbf{\Sigma}}_T = \frac{1}{p}\sum_{j=1}^{p}\widehat{\mathbf{\Sigma}}_{Tj} \in \mathbb{R}^{T\times T},$$

and set $\widehat{\mathbf{\Sigma}}_x = \widehat{\mathbf{\Sigma}}_p \otimes \widehat{\mathbf{\Sigma}}_T$.

2. *Divide the range of* Y *into* H *slices to estimate*

$$\widehat{\operatorname{Cov}}(\operatorname{vec}(\mathbb{E}[\mathbf{Z}|Y])) = \widehat{\mathbf{\Omega}}_p \otimes \widehat{\mathbf{\Omega}}_T,$$

where $\widehat{\mathbf{\Omega}}_p$ *and* $\widehat{\mathbf{\Omega}}_T$ *are found be the following estimate. Let* p_h *be the proportions of observation in slice* h, $\bar{\mathbf{z}}^{(h)}_{\cdot j}$ *the mean over observations within slice* h *of the* j*-th column of the standardized predictors and* $\bar{\mathbf{z}}^{(h)}_{t\cdot}$ *the mean over observations within slice* h *of the* t*-th row of the standardized predictors (as column/vector) for* $h = 1, \ldots, H$. *Then,*

$$\widehat{\mathbf{\Omega}}_{pt} = \sum_{h=1}^{H} p_h \bar{\mathbf{z}}^{(h)}_{t\cdot}\bar{\mathbf{z}}^{(h)\prime}_{t\cdot}, \quad \widehat{\mathbf{\Omega}}_p = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{\Omega}}_{pt} \in \mathbb{R}^{p\times p},$$
$$\widehat{\mathbf{\Omega}}_{Tj} = \sum_{h=1}^{H} p_h \bar{\mathbf{z}}^{(h)}_{\cdot j}\bar{\mathbf{z}}^{(h)\prime}_{\cdot j}, \quad \widehat{\mathbf{\Omega}}_T = \frac{1}{p}\sum_{j=1}^{p}\widehat{\mathbf{\Omega}}_{Tj} \in \mathbb{R}^{T\times T}.$$

3. *Compute the first* r *left singular vectors* $\widehat{\mathbf{U}}_T = (\widehat{\mathbf{U}}_{T1}, \ldots, \widehat{\mathbf{U}}_{Tr})$ *from the singular value decomposition of* $\widehat{\mathbf{\Omega}}_T$ *and the first* m *left singular vectors* $\widehat{\mathbf{U}}_p = (\widehat{\mathbf{U}}_{p1}, \ldots, \widehat{\mathbf{U}}_{pm})$ *from the singular value decomposition of* $\widehat{\mathbf{\Omega}}_p$.

4. *Output* $\widehat{\boldsymbol{\phi}} = \widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{U}}_p \in \mathbb{R}^{p\times m}$ *and* $\widehat{\boldsymbol{\psi}} = \widehat{\mathbf{\Sigma}}_T^{-1/2}\widehat{\mathbf{U}}_T \in \mathbb{R}^{T\times r}$ *to form the estimate*

$$\widehat{\mathcal{S}}_{\text{FMSDR}} = \operatorname{span}(\widehat{\boldsymbol{\phi}} \otimes \widehat{\boldsymbol{\psi}}).$$

One difference to the SIR Algorithm 2.6 is the use of the left singular vectors instead of eigenvectors. Since $\mathbf{\Omega}_T$ and $\mathbf{\Omega}_p$ are symmetric, they do agree up to the sign and their span

is the same. Note that $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\psi}}$ are not (pointwise) estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ in (2.17), but their spans are estimates for $\text{span}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\phi})$ and $\text{span}(\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\psi})$.

The Kronecker structures of $\boldsymbol{\Sigma}_x$ and $\text{Cov}(\text{vec}(\mathbb{E}[\mathbf{Z}|Y]))$ are estimated by first calculating the empirical covariance across one dimension (time or markers) for every level of the other dimension and then averaging over these levels of the other dimension. When finding a Kronecker structure of a given matrix $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2$, where the dimensions of $\mathbf{A}_1$ and $\mathbf{A}_2$ are fixed, $\mathbf{A}_1$ and $\mathbf{A}_2$ are only unique up to scale. However, for standardizing in the first step we use the full matrix, which is unique, and the span of the output is not affected by the scaling.

The true dimensions r and m are unknown. They can be estimated separately as the ranks of $\widehat{\boldsymbol{\Omega}}_T$ and $\widehat{\boldsymbol{\Omega}}_p$ by using the weighted chi-square test for dimension in [4], and since $\text{rank}(\widehat{\boldsymbol{\phi}} \otimes \widehat{\boldsymbol{\psi}}) = \text{rank}(\widehat{\boldsymbol{\phi}}) \cdot \text{rank}(\widehat{\boldsymbol{\psi}})$, the dimension of the estimated FMSDR space is then the product of these two estimated ranks.

From the algorithm, we can obtain the estimated reduction as

$$\widehat{R}(\mathbf{X}) = \widehat{\boldsymbol{\psi}}' \mathbf{X} \widehat{\boldsymbol{\phi}} \in \mathbb{R}^{r \times m}, \tag{2.24}$$

or in vectorized form as

$$\text{vec}(\widehat{R}(\mathbf{X})) = (\widehat{\boldsymbol{\phi}} \otimes \widehat{\boldsymbol{\psi}})' \,\text{vec}(\mathbf{X}) \in \mathbb{R}^{rm}. \tag{2.25}$$

For categorical Y, it is natural to use the categories as slices. It can also be applied for continuous Y, but one has to choose how to divide the range of Y into H slices. In the binary case (where $r = m = 1$), a scalar score of the predictors is obtained by projecting vectorized new observations on the estimated span via $\text{vec}(R(\mathbf{X})) = (\widehat{\boldsymbol{\phi}} \otimes \widehat{\boldsymbol{\psi}})' \,\text{vec}(\mathbf{X}) \in \mathbb{R}^{rm} = \mathbb{R}$. With this score, a AUC value for the prediction of the new responses can be obtained. This procedure will be explained in more detail later in Chapter 4.

Simulations for a binary case scenario in [17] show that LSIR reached an approximately 5% higher AUC value than the standard SIR, which ignores the longitudinal structure of the predictors. This difference leads to a significantly better discriminatory power [17]. The method also proved to be robust to violations of the Kronecker structure assumptions on the covariance. On a practically relevant data set, LSIR also outperformed SIR by 3% in terms of AUC.

This method will serve as a competitor against the method we will introduce now in Chapter 3 for both the simulations in Chapter 4 and the data analysis in Chapter 5.

3. STIR

We will now introduce the structured time-dependent inverse regression (STIR) model from [22]. Here we again consider longitudinal data and we model the first moment of the matrix-valued predictors by known functions of time and output. We estimate the coefficients by a least squares procedure. Using the resulting estimate, the method reduces only the markers to capture their joint effect on the outcome, while accounting for the effect of time.

3.1. Model

Here we consider a single real response $Y \in \mathbb{R}$ and a set of p markers $\mathbf{X} \in \mathbb{R}^p$, which are observed repeatedly over T time points for each individual. We assume that the number of marker-observations T and the set of (ordered) time points $\mathcal{T} = (t_1, \ldots, t_T)$ are the same for all individuals, although we will allow different time points $\mathcal{T}_i$ for each individual $i = 1, \ldots, n$ later on. For each individual i we therefore observe a response Y_i and a $T \times p$ predictor matrix

$$\mathbf{X}_i = \begin{pmatrix} X_{11}^i & \cdots & X_{1p}^i \\ \vdots & \ddots & \vdots \\ X_{T1}^i & \cdots & X_{Tp}^i \end{pmatrix} \in \mathbb{R}^{T \times p}, \tag{3.1}$$

where the entry X_{sj}^i is the observation of the j-th marker at the s-th time point t_s for individual i. We assume that the observations arise from the following model on the population level.

Definition 3.1 (STIR Model)**.** *Let $Y \sim F_Y$ be a real response and*

$$\mathbf{X} = (\mathbf{S}' \otimes \mathbf{G}(Y))\mathbf{B} + \mathbf{E} \in \mathbb{R}^{T \times p}, \tag{3.2}$$

where $\mathbf{E}$ is independent of Y with $\mathbb{E}[\mathbf{E}] = 0$ and $\mathrm{Cov}(\mathrm{vec}(\mathbf{E})) = \boldsymbol{\Delta} \in \mathbb{R}^{Tp \times Tp}$ positive definite and the elements of the mean structure are

$$\mathbf{S} = \begin{pmatrix} s_1(t_1) & \cdots & s_1(t_T) \\ \vdots & \ddots & \vdots \\ s_d(t_1) & \cdots & s_d(t_T) \end{pmatrix} \in \mathbb{R}^{d \times T}, \quad \mathbf{G}(Y) = (g_1(Y), \ldots, g_H(Y)) \in \mathbb{R}^{1 \times H}, \tag{3.3}$$

with $s_1, \ldots, s_d$ being pre-specified and centered functions of time and $g_1, \ldots, g_H$ centered functions of the response Y.

The conditional mean of $\mathbf{X}|Y$ is given by

$$\mathbb{E}[\mathbf{X}|Y] = (\mathbf{S}' \otimes \mathbf{G}(Y))\mathbf{B},$$

that is, it consists of a Kronecker product of time effect covered by $\mathbf{S}$ and effect of response covered by $\mathbf{G}$, multiplied by unknown coefficients $\mathbf{B}$. The $\mathbf{B}$ matrix can be written as

$$\mathbf{B} = \begin{pmatrix} \beta_{11}^1 & \cdots & \beta_{1p}^1 \\ \vdots & \ddots & \vdots \\ \beta_{11}^H & \cdots & \beta_{1p}^H \\ \vdots & \ddots & \vdots \\ \beta_{d1}^1 & \cdots & \beta_{dp}^1 \\ \vdots & \ddots & \vdots \\ \beta_{d1}^H & \cdots & \beta_{dp}^H \end{pmatrix} \in \mathbb{R}^{dH \times p},$$

where the entry β_{lj}^h is the coefficient for the l-th function of time and h-th function of Y of the j-th marker. The parameters d, H and functions $s_l, l \leq d$ and $g_h, h \leq H$ should be chosen such that $\mathbf{S}$ has rank d (implying $d \leq T$) and

$$\mathbf{G}(\mathbf{Y}) = \begin{pmatrix} \mathbf{G}(y_1) \\ \vdots \\ \mathbf{G}(y_n) \end{pmatrix} \in \mathbb{R}^{n \times H} \tag{3.4}$$

has rank H for observations of the response $y_1, \ldots, y_n$. Typical choices for the time functions are polynomials or Fourier basis elements. For continuous Y, we can also choose these for the g_hs. If Y is categorical with values in $\{0, 1, \ldots, H\}$, a natural choice is $g_h(Y) = \mathbb{1}(Y = h) - \mathbb{P}(Y = h)$.

To give a better understanding of the model and equation (3.2), the expectation of the j-th marker at the s-th time point for given $Y = y$ can be written as

$$\mathbb{E}[X_{sj}] = s_1(t_s)\left(\beta_{1j}^1 g_1(y) + \cdots + \beta_{1j}^H g_H(y)\right) + \cdots + s_d(t_s)\left(\beta_{dj}^1 g_1(y) + \ldots + \beta_{dj}^H g_H(y)\right).$$

The formulation from Definition 3.1 implies that $\mathbf{X}$ has mean 0. This does not result in a loss of generality, since for a general $\widetilde{\mathbf{X}}$ we can use the centered $\mathbf{X} = \widetilde{\mathbf{X}} - \mathbb{E}[\widetilde{\mathbf{X}}]$ in our model.

3.2. Reduction

Next, we give the setting for finding a reduction in that model. In general, we want to find a linear reduction on the vectorized predictors, as in LSIR, such that

$$F(Y|\operatorname{vec}(\mathbf{X})) = F(Y|\boldsymbol{\eta}'\operatorname{vec}(\mathbf{X})), \boldsymbol{\eta} \in \mathbb{R}^{Tp \times a}, \quad a \ll Tp. \tag{3.5}$$

Our method aims to reduce the markers while accounting for the time component, so we will have $a = Tk$ for some $k \ll p$.

Here we consider the time points as given. One could also model the time effect as random. Then, there would be two scenarios. If the distribution of the response depends on the time points $\mathcal{T}$, i.e. the time points capture relevant information for the outcome, we would need to find a reduction as a function of both $\mathbf{X}$ and $\mathcal{T}$. This scenario can occur e.g. in a health study where sicker individuals seek healthcare (where their predictors are observed) more frequently or earlier. However, modeling the distribution of $\mathcal{T}|Y$ can be quite challenging and is not considered here any further. The other scenario, where the distribution of the response Y is independent of the time points $\mathcal{T}$, is equivalent to our setting when conditioning on the time points.

As described in Chapter 2, finding a reduction satisfying (3.5) is equivalent to finding $\boldsymbol{\eta} \in \mathbb{R}^{Tp \times a}$ with

$$F(\operatorname{vec}(\mathbf{X})|Y, \boldsymbol{\eta}' \operatorname{vec}(\mathbf{X})) = F(\operatorname{vec}(\mathbf{X})|\boldsymbol{\eta}' \operatorname{vec}(\mathbf{X})).$$

This can be accomplished by finding a sufficient statistic for Y, treated as a parameter, in the inverse regression model $\mathbf{X}|Y$. Since under our model (3.2), the response Y relates to $\mathbf{X}$ only through the conditional mean, the linear design condition in Theorem 2.3 is not required. For categorical Y and normal errors $\mathbf{E}$, the marginal distribution of our $\mathbf{X}$ would be a mixture of multivariate normal distributions with different means and would not be elliptically symmetric.

The following theorem states how to find a sufficient reduction in our STIR model from Definition 3.1 [22, Theorem 1].

Theorem 3.2. *Under the* STIR *model* (3.2) *with* $\boldsymbol{\Sigma}_x = \operatorname{Cov}(\operatorname{vec}(\mathbf{X}))) \in \mathbb{R}^{Tp \times Tp}$ *and* $k = \operatorname{rank}(\mathbf{B}') = \operatorname{rank}(\mathbf{B})$*, a sufficient reduction for the regression of* Y *on* $\mathbf{X}$ *reducing the rows of* $\mathbf{X}$ *is given by*

$$R(\mathbf{X}) = \operatorname{unvec}(\boldsymbol{\Sigma}_x^{-1} \operatorname{vec}(\mathbf{X}))\boldsymbol{\alpha}_p, \tag{3.6}$$

where $\boldsymbol{\alpha}_p \in \mathbb{R}^{p \times k}$ *such that* $\operatorname{span}(\boldsymbol{\alpha}_p) = \operatorname{span}(\mathbf{B}')$.

Proof. Under (3.2), the rows of $\mathbb{E}[\mathbf{X}|Y] = (\mathbf{S}' \otimes \mathbf{G}(Y))\mathbf{B}$ are linear combinations of the rows of $\mathbf{B}$, so the idea is to find a base $\boldsymbol{\alpha}_p$ of $\operatorname{span}(\mathbf{B}')$ for the reduction. Since

$$R(\mathbf{X}) = \operatorname{unvec}(\boldsymbol{\Sigma}_x^{-1} \operatorname{vec}(\mathbf{X}))\boldsymbol{\alpha}_p \iff \operatorname{vec}(R(\mathbf{X})) = (\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{\alpha}_p \otimes \mathbf{I}_T))' \operatorname{vec}(\mathbf{X}),$$

we need to show that

$$\operatorname{span}(\operatorname{vec}(\mathbb{E}[\mathbf{X}|Y])) \subseteq \operatorname{span}(\mathbf{B}' \otimes \mathbf{I}_T).$$

Then the span of the supposed reduction covers the first moment sufficient dimension reduction space $\mathcal{S}_{\text{FMSDR}} = \boldsymbol{\Sigma}_x^{-1} \operatorname{span}(\operatorname{vec}(\mathbb{E}[\mathbf{X}|Y]))$.

For any $m \in \mathbb{N}$ and $i = 1, \ldots, m$ let $c_i \in \mathbb{R}$ be coefficients and y_i realizations of Y. Using the properties of the Kronecker product in A.3, we get

$$\sum_{i=1}^{m} c_i \operatorname{vec}((\mathbf{S}' \otimes \mathbf{G}(y_i))\mathbf{B}) = (\mathbf{I}_p \otimes \mathbf{S}' \otimes \underbrace{\sum_{i=1}^{m} c_i \mathbf{G}(y_i)}_{:=\mathbf{C}}) \operatorname{vec}(\mathbf{B}) = \operatorname{vec}((\mathbf{S}' \otimes \mathbf{C})\mathbf{B})$$
$$= (\mathbf{B}' \otimes \mathbf{I}_T) \operatorname{vec}(\mathbf{S}' \otimes \mathbf{C}) \in \operatorname{span}(\mathbf{B}' \otimes \mathbf{I}_T) \subseteq \mathbb{R}^{Tp},$$

which completes the proof. □

In general we have

$$\operatorname{span}(\operatorname{vec}(\mathbb{E}[\mathbf{X}|Y])) \subsetneq \operatorname{span}(\mathbf{B}' \otimes \mathbf{I}_T),$$

so our reduction might cover more than the first moment and is not minimal.

So far, we considered the error distributions to be independent of Y. In the simulations we will let the error covariance depend on Y, such that $\operatorname{vec}(\mathbf{E})|Y \sim N(0, \mathbf{\Delta}_Y)$. More precisely, we will let $\mathbf{\Delta}_Y = c_y \mathbf{\Delta}$ for a scaling constant $c_y \in \mathbb{R}$. As mentioned earlier in Chapter 2, Bura and Forzani showed in [3] that for $\mathbf{X}|Y \sim N_p(\mu_y, c_y\mathbf{\Delta})$, the minimal sufficient reduction is given by

$$R(\mathbf{X}) = (\boldsymbol{\alpha}'(\mathbf{X} - \mathbb{E}[\mathbf{X}]), (\mathbf{X} - \mathbb{E}[\mathbf{X}])'\mathbf{\Sigma}_x^{-1}(\mathbf{X} - \mathbb{E}[\mathbf{X}])), \tag{3.7}$$

with $\operatorname{span}(\boldsymbol{\alpha}) = \mathbf{\Sigma}_x^{-1} \operatorname{span}(\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}])$. Therefore, including second order terms of our reduction (3.6) for fitting a model to Y can possibly increase the predictive performance.

3.3. Estimation of the Coefficients

To estimate the reduction (3.6), we need to find the coefficients $\mathbf{B}$ in model (3.2). We can find an approximation of the row span of $\mathbf{B}$ from this estimate. In this section we will derive an OLS based estimator and an ML estimator assuming a normal error distribution, both for equal and unequal time points for the individuals. The predictive performance for the different resulting reductions will be compared in simulations in Chapter 4.

As a general case we will allow now different time points for each individual $\mathcal{T}_i = (t_1^i, \ldots, t_T^i)$ and let $\mathbf{S}_i$ be the time matrix for those time points, meaning

$$\mathbf{S}_i = \begin{pmatrix} s_1(t_1^i) & \cdots & s_1(t_T^i) \\ \vdots & \ddots & \vdots \\ s_d(t_1^i) & \cdots & s_d(t_T^i) \end{pmatrix} \in \mathbb{R}^{d \times T}.$$

However, we always assume a fixed number of time points T. Then the STIR model in Definition 3.1 at the observation level is

$$\mathbf{X}_i = (\mathbf{S}_i{}' \otimes \mathbf{G}(y_i))\mathbf{B} + \mathbf{E}_i, \quad i = 1, \ldots, n, \tag{3.8}$$

or in vectorized version

$$\operatorname{vec}(\mathbf{X}_i) = (\mathbf{I}_p \otimes \mathbf{S}_i' \otimes \mathbf{G}(y_i)) \operatorname{vec}(\mathbf{B}) + \operatorname{vec}(\mathbf{E}_i), \quad i = 1, \ldots, n. \tag{3.9}$$

By stacking the rows of the $\mathbf{X}_i$s in a convenient way, we will obtain a version of the multivariate linear regression model from Definition 2.2, where we distinguish between the case of same or different time points. Note that we are looking at an inverse regression problem here, so $\mathbf{X}$ plays the role of the response of the multivariate linear regression model. Afterwards we will assume a normal distribution for the errors $\mathbf{E}_i$ and derive the ML estimator for $\mathbf{B}$ in this model.

3.3.1. OLS Estimates

Same Time Points

We first look at the case where $\mathbf{S}_i = \mathbf{S}$ for all $i = 1, \ldots, n$. Let $\mathbf{X}_{s.}^i \in \mathbb{R}^{1\times p}$ denote the s-th row of $\mathbf{X}_i$ (here as row). Then by stacking these rows with varying the observation index i before the time index, we can write

$$\mathbb{X}_{\text{same}} = \begin{pmatrix} \mathbf{X}_{1.}^1 \\ \vdots \\ \mathbf{X}_{1.}^n \\ \vdots \\ \mathbf{X}_{T.}^1 \\ \vdots \\ \mathbf{X}_{T.}^n \end{pmatrix} = (\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y}))\mathbf{B} + \mathbb{E}_{\text{same}} \in \mathbb{R}^{nT\times p}, \tag{3.10}$$

with $\operatorname{Cov}(\operatorname{vec}(\mathbb{E}_{\text{same}})) = \operatorname{Cov}(\operatorname{vec}(\mathbb{X}_{\text{same}})|\mathbf{Y}) = \boldsymbol{\Delta} \otimes \mathbf{I}_n$ and $\mathbf{G}(\mathbf{Y}) \in \mathbb{R}^{n\times H}$ with rows $\mathbf{G}(y_i)$. The OLS estimator of $\mathbf{B}$ in equation (3.10) derived in Section 2.1 is given by

$$\widehat{\mathbf{B}}_{\text{same}} = \left((\mathbf{S}\mathbf{S}')^{-1}\mathbf{S} \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}\mathbf{G}(\mathbf{Y})'\right) \mathbb{X}_{\text{same}} \in \mathbb{R}^{dH\times p}, \tag{3.11}$$

with the $dHp \times dHp$ covariance matrix

$$\operatorname{Cov}(\operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y}) = (\mathbf{I}_p \otimes \mathbf{W})(\boldsymbol{\Delta} \otimes \mathbf{I}_n)(\mathbf{I}_p \otimes \mathbf{W}'), \tag{3.12}$$

where $\mathbf{W} = \left((\mathbf{S}\mathbf{S}')^{-1}\mathbf{S} \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}\mathbf{G}(\mathbf{Y})'\right)$. The formulation in (3.10) does not exactly satisfy the assumptions for the multivariate linear regression model, since $\mathbb{X}_{\text{same}}$ in (3.10) would need to have error covariance structure $\widetilde{\boldsymbol{\Delta}}_p \otimes \mathbf{I}_{nT}$ for some $\widetilde{\boldsymbol{\Delta}}_p \in \mathbb{R}^{p\times p}$. However, we can still use this unbiased estimator. It has a different covariance structure and, when assuming a normal error distribution, does not agree with the ML estimator.

Different Time Points

Now we consider the more general case of different time points and time matrices $\mathbf{S}_i$. By stacking the n equations for the observations of (3.8), we get

$$\mathbb{X}_{\text{diff}} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{S}_1' \otimes \mathbf{G}(y_1) \\ \vdots \\ \mathbf{S}_n' \otimes \mathbf{G}(y_n) \end{pmatrix}}_{:=\mathbf{Z} \in \mathbb{R}^{nT \times dH}} \mathbf{B} + \mathbb{E}_{\text{diff}} = \mathbf{Z}\mathbf{B} + \mathbb{E}_{\text{diff}} \in \mathbb{R}^{nT \times p}, \tag{3.13}$$

with

$$\text{Cov}(\text{vec}(\mathbb{E}_{\text{diff}})) = \text{Cov}(\text{vec}(\mathbb{X}_{\text{diff}})|\mathbf{Y}) = \begin{pmatrix} \mathbf{I}_n \otimes \mathbf{\Delta}_{11} & \cdots & \mathbf{I}_n \otimes \mathbf{\Delta}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_n \otimes \mathbf{\Delta}_{p1} & \cdots & \mathbf{I}_n \otimes \mathbf{\Delta}_{pp} \end{pmatrix}, \tag{3.14}$$

where $\mathbf{\Delta}_{jk} = \mathbb{E}[\mathbf{E}_{\cdot j}\mathbf{E}_{\cdot k}'] \in \mathbb{R}^{T \times T}$ for $j, k = 1, \ldots, p$ denotes the (j,k)-th sub-matrix of $\mathbf{\Delta} \in \mathbb{R}^{Tp \times Tp}$. The different order of the row stacking leads to a different covariance structure compared to the equal time setting. If we assume $\mathbf{\Delta} = \mathbf{\Delta}_p \otimes \mathbf{\Delta}_T$, then the covariance matrix in (3.14) simplifies to

$$\text{Cov}(\text{vec}(\mathbb{E}_{\text{diff}})) = \mathbf{\Delta}_p \otimes \mathbf{I}_n \otimes \mathbf{\Delta}_T.$$

Similar to the case of equal time points, the OLS estimator of $\mathbf{B}$ in (3.13) is

$$\widehat{\mathbf{B}}_{\text{diff}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbb{X}_{\text{diff}}, \tag{3.15}$$

with the $dHp \times dHp$ covariance matrix

$$\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{\text{diff}})|\mathbf{Y}) = (\mathbf{I}_p \otimes (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \begin{pmatrix} \mathbf{I}_n \otimes \mathbf{\Delta}_{11} & \cdots & \mathbf{I}_n \otimes \mathbf{\Delta}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_n \otimes \mathbf{\Delta}_{p1} & \cdots & \mathbf{I}_n \otimes \mathbf{\Delta}_{pp} \end{pmatrix} (\mathbf{I}_p \otimes \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}).$$

This structure implies that for $j, k = 1, \ldots, p$ the (j,k)-th sub-matrix of dimension $dH \times dH$ is given by

$$\begin{aligned} \text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{\text{diff}})|\mathbf{Y})_{jk} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}_n \otimes \mathbf{\Delta}_{jk})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = \\ &= (\mathbf{Z}'\mathbf{Z})^{-1} \left(\sum_{i=1}^{n} (\mathbf{S}_i \mathbf{\Delta}_{jk} \mathbf{S}_i' \otimes \mathbf{G}(y_i)'\mathbf{G}(y_i)) \right) (\mathbf{Z}'\mathbf{Z})^{-1} \end{aligned} \tag{3.16}$$

The following proposition tells us that this estimator is indeed a generalization of the estimator $\widehat{\mathbf{B}}_{\text{same}}$ in (3.11) for equal time points.

Proposition 3.3. *If we have equal time points for all individuals and therefore $\mathbf{S}_i = \mathbf{S}$ for all $i = 1, \ldots, n$, then the estimator $\widehat{\mathbf{B}}_{same}$ in (3.11) agrees with $\widehat{\mathbf{B}}_{diff}$ in (3.15).*

Proof. $\widehat{\mathbf{B}}_{\text{same}}$ can also be written as

$$\widehat{\mathbf{B}}_{\text{same}} = \left((\mathbf{S}\mathbf{S}') \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))\right)^{-1} (\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y}))'\mathbb{X}_{\text{same}}.$$

For $\mathbf{S}_i = \mathbf{S}$,

$$\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} (\mathbf{S}_i \otimes \mathbf{G}(y_i)')(\mathbf{S}_i' \otimes \mathbf{G}(y_i)) = \mathbf{S}\mathbf{S}' \otimes \left(\sum_{i=1}^{n} \mathbf{G}(y_i)'\mathbf{G}(y_i)\right) = \mathbf{S}\mathbf{S}' \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y})),$$

so what is left to show is that

$$(\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y}))'\mathbb{X}_{\text{same}} = \mathbf{Z}'\mathbb{X}_{\text{diff}} \in \mathbb{R}^{dH \times p}.$$

Let $l \in \{1, \dots, d\}, h \in \{1, \dots, H\}$ and $j \in \{1, \dots, p\}$. Then, element-wise comparison gives

$$\begin{aligned}
(\mathbf{Z}'\mathbb{X}_{\text{diff}})_{h+H(l-1),j} &= \sum_{i=1}^{n} ((\mathbf{S}_i)_{l\cdot} \otimes g_h(y_i))\mathbf{X}^i_{\cdot j} = \sum_{i=1}^{n}\sum_{s=1}^{T} ((\mathbf{S}_i)_{ls} \otimes g_h(y_i))X^i_{sj} = \\
&= \sum_{s=1}^{T} ((\mathbf{S})_{ls} \otimes \mathbf{G}(\mathbf{Y})_{\cdot h}) \begin{pmatrix} X^1_{sj} \\ \vdots \\ X^n_{sj} \end{pmatrix} = ((\mathbf{S})_{l\cdot} \otimes (\mathbf{G}(\mathbf{Y})')_{h\cdot})(\mathbb{X}_{\text{same}})_{\cdot j} = \\
&= ((\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y}))'\mathbb{X}_{\text{same}})_{h+H(l-1),j},
\end{aligned}$$

which proves the claim. □

Even though $\widehat{\mathbf{B}}_{\text{same}}$ and $\widehat{\mathbf{B}}_{\text{diff}}$ do theoretically agree for equal time points, using (3.11) to calculate the estimator is computationally more efficient and stable than (3.15).

If we apply the formula (3.16) in the case of equal time $\mathbf{S}_i = \mathbf{S}$, we obtain the following alternative representation of $\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y})$.

Corollary 3.4. *The (j,k)-th sub-matrix of* $\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{same})|\mathbf{Y}) \in \mathbb{R}^{dHp \times dHp}$ *in* (3.12) *for* $j, k = 1, \dots, p$ *is given by*

$$\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{same})|\mathbf{Y})_{jk} = \left((\mathbf{S}\mathbf{S}')^{-1}\mathbf{S}\boldsymbol{\Delta}_{jk}\mathbf{S}'(\mathbf{S}\mathbf{S}')^{-1}\right) \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1} \in \mathbb{R}^{dH \times dH}. \quad (3.17)$$

Proof. For $\mathbf{S}_i = \mathbf{S}$, the previous proposition tells us that we can also use (3.16) to calculate $\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y})$. In that case we have

$$\begin{aligned}
\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} (\mathbf{S}_i\mathbf{S}_i' \otimes \mathbf{G}(y_i)'\mathbf{G}(y_i)) &\overset{\mathbf{S}_i = \mathbf{S}}{=} (\mathbf{S}\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y})) \\
\implies (\mathbf{Z}'\mathbf{Z})^{-1} = (\mathbf{S}\mathbf{S}')^{-1} &\otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}.
\end{aligned}$$

Then, directly applying (3.16) gives

$$\begin{aligned}
\text{Cov}(\text{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y})_{jk} &= (\mathbf{Z}'\mathbf{Z})^{-1} \left(\sum_{i=1}^{n} (\mathbf{S}\boldsymbol{\Delta}_{jk}\mathbf{S}' \otimes \mathbf{G}(y_i)'\mathbf{G}(y_i))\right) (\mathbf{Z}'\mathbf{Z})^{-1} \\
= \left((\mathbf{S}\mathbf{S}')^{-1} \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}\right) &\left(\mathbf{S}\boldsymbol{\Delta}_{jk}\mathbf{S}' \otimes \mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y})\right) \left((\mathbf{S}\mathbf{S}')^{-1} \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}\right) \\
&= \left((\mathbf{S}\mathbf{S}')^{-1}\mathbf{S}\boldsymbol{\Delta}_{jk}\mathbf{S}'(\mathbf{S}\mathbf{S}')^{-1}\right) \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1} \in \mathbb{R}^{dH \times dH}.
\end{aligned}$$

□

For practical use, this makes a huge difference in computation times of $\mathrm{Cov}(\mathrm{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y})$. For data generated from scenario STIR-1 in Table 4.4, the computing times of direct implementation of equation (3.12) were around $0.11, 3.35$ and 19 seconds for $n = 100, 500, 1000$, while they were around 0.002 seconds for all n when using Corollary 3.4.

3.3.2. ML Estimate

Now we assume a normal distribution of the errors $\mathbf{E}_i$ in (3.8). Then the vectorized predictors in (3.9) satisfy

$$\mathrm{vec}(\mathbf{X}_i)|(Y_i = y_i) \overset{indep.}{\sim} N(\mathbf{M}_i \,\mathrm{vec}(\mathbf{B}), \boldsymbol{\Delta}), i = 1, \ldots, n,$$

where $\mathbf{M}_i = (\mathbf{I}_p \otimes \mathbf{S}_i{}' \otimes \mathbf{G}(y_i)) \in \mathbb{R}^{Tp \times dHp}$. The log-likelihood of $\mathbf{B}$ is then given by

$$\begin{aligned}\ell(\mathbf{B}|\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_n) &= -\frac{1}{2}\sum_{i=1}^{n}(\mathrm{vec}(\mathbf{X}_i) - \mathbf{M}_i \,\mathrm{vec}(\mathbf{B}))'\boldsymbol{\Delta}^{-1}(\mathrm{vec}(\mathbf{X}_i) - \mathbf{M}_i \,\mathrm{vec}(\mathbf{B})) + c_1 \\ &= -\frac{1}{2}\sum_{i=1}^{n}\mathrm{vec}(\mathbf{B})'\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\mathbf{M}_i \,\mathrm{vec}(\mathbf{B}) - 2\,\mathrm{vec}(\mathbf{B})'\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\,\mathrm{vec}(\mathbf{X}_i) + c_2,\end{aligned}$$

where c_1, c_2 are constants not depending on $\mathbf{B}$. Therefore,

$$\frac{\partial \ell(\mathbf{B}|\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_n)}{\partial\,\mathrm{vec}(\mathbf{B})} = \sum_{i=1}^{n}\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\mathbf{M}_i \,\mathrm{vec}(\mathbf{B}) - \mathbf{M}_i'\boldsymbol{\Delta}^{-1}\,\mathrm{vec}(\mathbf{X}_i).$$

By the first order condition, the ML estimator is given by

$$\widehat{\mathbf{B}}_{\text{ML}} = \mathrm{unvec}\left((\sum_{i=1}^{n}\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\mathbf{M}_i)^{-1}\left(\sum_{i=1}^{n}\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\,\mathrm{vec}(\mathbf{X}_i)\right)\right), \tag{3.18}$$

with the $dHp \times dHp$ covariance matrix

$$\mathrm{Cov}(\mathrm{vec}(\widehat{\mathbf{B}}_{\text{ML}})|\mathbf{Y}) = \left(\sum_{i=1}^{n}\mathbf{M}_i'\boldsymbol{\Delta}^{-1}\mathbf{M}_i\right)^{-1}.$$

Both the estimator and its covariance matrix include the unknown error covariance $\boldsymbol{\Delta}$. We compute $\widehat{\mathbf{B}}_{\text{ML}}$ in a two-step procedure. We first use $\widehat{\mathbf{B}}_{\text{same}}$ or $\widehat{\mathbf{B}}_{\text{diff}}$ to estimate $\boldsymbol{\Delta}$ by the residual covariance $\widehat{\boldsymbol{\Delta}}$, as described in the next Section 3.4. Then we use this $\widehat{\boldsymbol{\Delta}}$ to calculate $\widehat{\mathbf{B}}_{\text{ML}}$ in (3.18).

Models (3.10) and (3.13) can be seen as generalized multivariate linear regression models, so the OLS estimators derived in those two previous sections do not agree with the ML estimator given here. However, for the restrictive case $\boldsymbol{\Delta} = \boldsymbol{\Delta}_p \otimes \mathbf{I}_T$ one can show that $\widehat{\mathbf{B}}_{\text{ML}}$ in (3.18) would agree with $\widehat{\mathbf{B}}_{\text{diff}}$ in (3.15).

3.4. Estimation of the Error Covariance

An estimate for $\boldsymbol{\Delta}$ is needed for calculating the reduction or the ML estimator of $\mathbf{B}$, as well as estimating the covariance of any of the $\widehat{\mathbf{B}}$. The latter is, for example, needed for variable selection tests.

From now on let $\widehat{\mathbf{B}}$ be either $\widehat{\mathbf{B}}_{\text{same}}$ or $\widehat{\mathbf{B}}_{\text{diff}}$ introduced in the previous section, depending on whether all individuals are measured at equal time points or not. Then we can define the fitted values and residuals as

$$\begin{aligned} \widehat{\mathbf{X}}_i &= (\mathbf{S}_i \otimes \mathbf{G}(y_i))\widehat{\mathbf{B}}, \\ \widehat{\mathbf{E}}_i &= \mathbf{X}_i - \widehat{\mathbf{X}}_i, \\ \iff \operatorname{vec}(\widehat{\mathbf{E}}_i) &= \operatorname{vec}(\mathbf{X}_i) - (\mathbf{I}_p \otimes \mathbf{S}_i \otimes \mathbf{G}(y_i)) \operatorname{vec}(\widehat{\mathbf{B}}). \end{aligned}$$

If we let

$$\widehat{\mathbf{E}} = \begin{pmatrix} \operatorname{vec}(\widehat{\mathbf{E}}_1) \\ \vdots \\ \operatorname{vec}(\widehat{\mathbf{E}}_n) \end{pmatrix} \in \mathbb{R}^{n \times Tp}, \tag{3.19}$$

the OLS based estimate of $\boldsymbol{\Delta}$ is

$$\widehat{\boldsymbol{\Delta}}_{\text{OLS}} = \frac{1}{n-q} \widehat{\mathbf{E}}' \widehat{\mathbf{E}}, \tag{3.20}$$

where $q = \operatorname{rank}(\mathbf{I}_p \otimes \mathbf{Z}) = pdH$, if each $\mathbf{S}_i$ has full rank d and $\mathbf{G}(\mathbf{Y})$ has full rank H. We give a short argument to show $\operatorname{rank}(\mathbf{Z}) = dH$. Let $\mathbf{z} \in \ker(\mathbf{Z}) \subseteq \mathbb{R}^{dH}$ and let $\mathbf{z}_l \in \mathbb{R}^H$ denote the l-th sub-vector of $\mathbf{z}$ for $l = 1, \ldots, d$. Then, for each $i = 1, \ldots, n$,

$$\begin{aligned} & (\mathbf{S}_i' \otimes \mathbf{G}(y_i))\mathbf{z} = 0, \\ \iff & \sum_{l=1}^{d} (\mathbf{S}_i)_{ls} (\mathbf{G}(y_i)\mathbf{z}_l) = 0, \quad \forall s \in \{1, \ldots, T\}. \end{aligned}$$

Since $\mathbf{S}_i$ has full rank, this implies $\mathbf{G}(y_i)\mathbf{z}_l = 0$ for each $l = 1, \ldots, d$. This holds for every i, so also $\mathbf{G}(\mathbf{Y})z_l = 0$ for every l. If $\mathbf{G}(\mathbf{Y})$ has full rank, this gives $\mathbf{z}_l = 0$ for each l and therefore $\mathbf{z} = 0$. So by the rank-nullity theorem, $\operatorname{rank}(\mathbf{Z}) = dH$.

3.4.1. Separable Error Covariance

If we assume a Kronecker structure for $\boldsymbol{\Delta}$, such that $\boldsymbol{\Delta} = \boldsymbol{\Delta}_p \otimes \boldsymbol{\Delta}_T$ for some $\boldsymbol{\Delta}_p \in \mathbb{R}^{p \times p}, \boldsymbol{\Delta}_T \in \mathbb{R}^{T \times T}$, the number of estimated parameters decreases from $pT(pT+1)/2$ to $p(p+1)/2 + T(T+1)/2$.

For estimation, we use the procedure from [17], that was already used for the LSIR-algorithm in Section 2.4. Let $\hat{\mathbf{e}}^i_{\cdot j}$ be the j-th column of $\widehat{\mathbf{E}}_i$, $\hat{\mathbf{e}}^i_{t\cdot}$ the t-th row of $\widehat{\mathbf{E}}_i$ (as

column/vector), and $\bar{\mathbf{e}}_{.j}, \bar{\mathbf{e}}_{t.}$ the corresponding means over all observations. Then

$$\widehat{\mathbf{\Delta}}_{pt} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{e}}^i_{t.} - \bar{\mathbf{e}}_{t.})(\hat{\mathbf{e}}^i_{t.} - \bar{\mathbf{e}}_{t.})', \quad \widehat{\mathbf{\Delta}}_{p} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{\Delta}}_{pt} \in \mathbb{R}^{p\times p},$$
$$\widehat{\mathbf{\Delta}}_{Tj} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{e}}^i_{.j} - \bar{\mathbf{e}}_{.j})(\hat{\mathbf{e}}^i_{.j} - \bar{\mathbf{e}}_{.j})', \quad \widehat{\mathbf{\Delta}}_{T} = \frac{1}{p}\sum_{j=1}^{p}\widehat{\mathbf{\Delta}}_{Tj} \in \mathbb{R}^{T\times T},$$

and set $\widehat{\mathbf{\Delta}}_{\text{kron}} = \widehat{\mathbf{\Delta}}_p \otimes \widehat{\mathbf{\Delta}}_T$.

3.4.2. ML Estimate of Error Covariance

In this section, we again assume a normal distribution of the errors $\mathbf{E}_i$ in (3.8), as in Section 3.3.2. The vectorized predictors in (3.9) have the distribution

$$\text{vec}(\mathbf{X}_i)|(Y_i = y_i) \overset{indep.}{\sim} N(\mathbf{M}_i \,\text{vec}(\mathbf{B}), \mathbf{\Delta}), i = 1, \ldots, n,$$

where $\mathbf{M}_i = (\mathbf{I}_p \otimes \mathbf{S}_i' \otimes \mathbf{G}(y_i)) \in \mathbb{R}^{Tp\times dHp}$. For any coefficient $\mathbf{B}$, the likelihood L of $\mathbf{\Delta}$ then satisfies

$$\begin{aligned}
&L(\mathbf{\Delta}|\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \propto \\
&\quad\propto \det(\mathbf{\Delta})^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))'\mathbf{\Delta}^{-1}(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))\right\} = \\
&\quad= \det(\mathbf{\Delta})^{-n/2}\exp\left\{-\frac{1}{2}\,\text{tr}\left(\mathbf{\Delta}^{-1}\sum_{i=1}^{n}(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))'\right)\right\} = \\
&\quad= \det(\mathbf{\Delta})^{-n/2}\exp\left\{-\frac{1}{2}\,\text{tr}(\mathbf{\Delta}^{-1}\mathbf{U})\right\},
\end{aligned}$$

for $\mathbf{U} := \sum_{i=1}^{n}(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))(\text{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\text{vec}(\mathbf{B}))'$ depending on $\mathbf{B}$, where $\propto$ signifies proportional to. We assume that $\mathbf{U}$ is positive definite and let $\mathbf{V} = \mathbf{U}^{1/2}\mathbf{\Delta}^{-1}\mathbf{U}^{1/2}$. Then,

$$\begin{aligned}
L(\mathbf{\Delta}|\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_n) &\propto \det(\mathbf{\Delta})^{-n/2}\exp\{-\frac{1}{2}\,\text{tr}(\mathbf{\Delta}^{-1}\mathbf{U})\} = \\
&= \det(\mathbf{U})^{-n/2}\det(\mathbf{V})^{n/2}\exp\{-\frac{1}{2}\,\text{tr}(\mathbf{V})\}.
\end{aligned}$$

For fixed $\mathbf{U}$, maximizing the likelihood is equivalent to finding the matrix $\mathbf{V}$ that maximizes

$$\det(\mathbf{V})^{n/2}\exp\{-\frac{1}{2}\,\text{tr}(\mathbf{V})\} = \prod_{j=1}^{Tp}\lambda_j^{n/2}\exp\{-\frac{1}{2}\lambda_j\},$$

where the λ_js denote the eigenvalues of $\mathbf{V}$. Maximizing with respect to the eigenvalues we obtain $\lambda_j = n$ for all $j = 1, \ldots, Tp$ and therefore $\widehat{\mathbf{V}}_{\text{ML}} = n\mathbf{I}_{Tp}$. Using the definition of $\mathbf{V}$

this yields

$$\widehat{\boldsymbol{\Delta}}_{\mathrm{ML}} = \frac{1}{n}\mathbf{U} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\mathrm{vec}(\mathbf{B}))(\mathrm{vec}(\mathbf{X}_i) - \mathbf{M}_i\,\mathrm{vec}(\mathbf{B}))',$$

for any fixed $\mathbf{B}$. Since the ML estimator of $\mathbf{B}$ itself depends on $\boldsymbol{\Delta}$, we use the estimator

$$\widehat{\boldsymbol{\Delta}}_{\mathrm{ML}} = \frac{1}{n}\widehat{\mathbf{E}}'\widehat{\mathbf{E}}, \tag{3.21}$$

where the residuals are calculated from $\widehat{\mathbf{B}}$ as in (3.19). This estimate can then be used to calculate $\widehat{\mathbf{B}}_{\mathrm{ML}}$.

The ML estimate of $\boldsymbol{\Delta}$ differs from the OLS based estimate only in the scaling constant. Simulations performed later on suggest that the scaling by n yields a less biased estimate (see Figures 4.1 and 4.2). In general, we will therefore use this adapted ML estimator (3.21) to estimate $\boldsymbol{\Delta}$.

One open problem is to prove consistency of this covariance error estimator. For this asymptotic property does not matter whether the ML based or the OLS based scaling is used. In [3, Theorem 4], Bura and Forzani show asymptotic normality (implying consistency) of the residual covariance matrix for a homeostatic model. However, there are requirements on the mean that would only be satisfied for our model, if it were an exact Multivariate Linear Regression Model. In [12], Hoadley states general conditions for consistency of MLestimators for the case of independent, but not identically distributed, observations. These conditions are by far not trivial to check for our model, so we rely on the usefulness of the error covariance estimator suggested by the simulations.

3.5. Estimation of the Reduction

Using the estimators derived in the previous sections, the estimation of the reduction (3.6) is straightforward. We will use the empirical covariance matrix of the vectorized predictors as estimate for $\boldsymbol{\Sigma}_x$, i.e.

$$\widehat{\boldsymbol{\Sigma}}_x = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{vec}(\mathbf{X}_i)\,\mathrm{vec}(\mathbf{X}_i)' \in \mathbb{R}^{Tp\times Tp}.$$

Then, for a given $k = \mathrm{rank}(\mathbf{B}') = \mathrm{rank}(\mathbf{B})$, we estimate $\mathrm{span}(\mathbf{B}')$ by

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right),$$

where $(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k)$ is the $p \times k$ matrix of the first k left singular vectors of $\widehat{\mathbf{B}}'$.

Thus, the estimator of the sufficient reduction in (3.6) for $\mathbf{X}_i$ in (3.8), $i = 1, \ldots, n$, is

$$\widehat{R}_p(\mathbf{X}_i) = \mathrm{unvec}\left(\widehat{\boldsymbol{\Sigma}}_x^{-1}\,\mathrm{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right) \in \mathbb{R}^{T\times k}. \tag{3.22}$$

By Corollary 3.4 in [6], $\widehat{\mathbf{\Sigma}}_x$ and $\widehat{\mathbf{\Delta}}$, the sample covariance matrix of the residuals, can be used interchangeably in (3.22). For the reduction, it does not matter which exact scaling of the sample residual covariance matrix is used, since it just scales the reduced observations by the inverse of this scalar.

In the simulations, we compare the predictive performance of these reductions for using $\widehat{\mathbf{\Sigma}}_x$ and $\widehat{\mathbf{\Delta}}$, and also for estimating both these covariance matrices by a Kronecker structure. Also, we estimate the span of $\mathbf{B}'$ from $\widehat{\mathbf{B}}_{\text{ML}}$ and compare the performance.

In practice, the true rank k is unknown and d, H are parameters of choice. To not lose information, one should choose $dH \geq k$. This would be ensured by setting $dH \geq p$. However, this might not always be possible. For example, in the binary case $H = 1$, because otherwise $\mathbf{G}(\mathbf{Y})$ in (3.4) cannot have full rank. Also, d cannot exceed T to ensure the full (row-)rank of $\mathbf{S}$.

3.6. Variable Selection

In this section we introduce a testing procedure for our STIR-model to test for the importance of a marker $j \in \{1, \dots, p\}$. We can test the null $\mathcal{H}_0 : \beta^h_{l,j} = 0$ for all $l = 1 \dots, d, h = 1, \dots, H$, which is equivalent to

$$\mathcal{H}_0 : \mathbf{B}_{.j} = \mathbf{A}_j \operatorname{vec}(\mathbf{B}) = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \mathbf{B}_{.j} \neq 0 \tag{3.23}$$

for the full row rank matrix

$$\mathbf{A}_j = \begin{pmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_{dH} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{dH \times dHp}, \tag{3.24}$$

where $\mathbf{0} \in \mathbb{R}^{dH \times dH}$ and the identity matrix is positioned as the j-th $(dH \times dH)$-block.

Let $\mathbf{C} = \mathbf{I}_p \otimes (\mathbf{S}\mathbf{S}')^{-1}\mathbf{S} \otimes (\mathbf{G}(\mathbf{Y})'\mathbf{G}(\mathbf{Y}))^{-1}\mathbf{G}(\mathbf{Y})'$. Under the STIR-model from Definition 3.1 for equal time points for each individual and assuming normality of the errors, we have

$$\operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}}) \sim N_{dHp}(\operatorname{vec}(\mathbf{B}), \mathbf{C}(\mathbf{\Delta} \otimes \mathbf{I}_n)\mathbf{C}') \tag{3.25}$$

Therefore, letting $\widehat{\mathbf{B}}_{.j} = \mathbf{A}_j \operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}})$ leads to $\widehat{\mathbf{B}}_{.j} \sim N_{dH}(\mathbf{B}_{.j}, \mathbf{A}_j\mathbf{C}(\mathbf{\Delta} \otimes \mathbf{I}_n)\mathbf{C}'\mathbf{A}_j')$. Since this covariance matrix still has full rank, under the null

$$(\mathbf{A}_j \operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}}))'(\mathbf{A}_j\mathbf{C}(\mathbf{\Delta} \otimes \mathbf{I}_n)\mathbf{C}'\mathbf{A}_j')^{-1}(\mathbf{A}_j \operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}})) \sim \chi^2(dH).$$

Provided $\widehat{\mathbf{\Delta}}$ is consistent for $\mathbf{\Delta}$, a simple application of Slutsky's theorem yields that

$$W = (\mathbf{A}_j \operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}}))'(\mathbf{A}_j\mathbf{C}(\widehat{\mathbf{\Delta}} \otimes \mathbf{I}_n)\mathbf{C}'\mathbf{A}_j')^{-1}(\mathbf{A}_j \operatorname{vec}(\widehat{\mathbf{B}})) \sim_{H_0} \chi^2(dH). \tag{3.26}$$

The null (3.23) is rejected at level α, if $W > \chi^2_\alpha(dH)$.

In practice, Corollary 3.4 is useful to compute this test statistic, because the variance-covariance matrix of $\widehat{\mathbf{B}}_{.j}$ is exactly the (j, j) sub-matrix of $\operatorname{Cov}(\operatorname{vec}(\widehat{\mathbf{B}}_{\text{same}})|\mathbf{Y})$. From this corollary, we can also see that the covariance matrix of $\widehat{\mathbf{B}}_{.j}$ still has full rank, if $\mathbf{S}, \mathbf{\Delta}_{jj}$, and $\mathbf{G}(\mathbf{Y})$ have full ranks d, T and H.

Similar tests can be performed using $\widehat{\mathbf{B}}_{\text{diff}}$ or $\widehat{\mathbf{B}}_{\text{ML}}$. However, we will mostly use this more simple estimator in the simulations.

4. Simulations

We perform extensive simulation studies to assess the performance of the STIR method. We will generate the data from this STIR model, as well as from a model satisfying the main assumption of the LSIR algorithm (2.17) for both the case of a binary and continuous response variable Y. Each scenario is repeated for fixed error covariance and for error covariance depending on Y through a multiplicative scalar. We assess the estimation accuracy of estimating $\mathbf{B}$ and $\mathbf{\Delta}$, and report predictive performance for fitting a regression model on the reduced predictors. We compare different versions of our method to LSIR and to using no reduction. This analysis mostly agrees with what was done in [22], but also examines the performance when using the ML estimator for $\mathbf{B}$.

4.1. Data Generation

We start by generating a response Y. In the binary case, we draw $y_i \overset{iid}{\sim} Bern(0.5)$, $i = 1, \dots, n$ for $n = 500$. Some settings will also use $n = 2000$ observations. For the continuous case we instead use $y_i \overset{iid}{\sim} N(0, 0.1), i = 1, \dots, n$. For the error covariances not depending on Y we will use $\mathbf{\Delta} = AR^1_{Tp}(\rho)$ with $\rho = 0.8$, i.e. an autoregressive structure where $(\mathbf{\Delta})_{jk} = \rho^{|j-k|}, j, k = 1, \dots, Tp$. If we let the error covariance depend on Y, we set $\mathbf{\Delta}_0 = 0.1 \cdot AR^1(0.8), \mathbf{\Delta}_1 = AR^1(0.8)$ in the binary case and $\mathbf{\Delta}_{y_i} = \exp(-\min(\sqrt{10}|y_i|, 2)) \cdot AR^1_{Tp}(0.8)$ in the continuous case. When generating from a error covariance matrix, which is dependent on Y, we expect that not all information on Y is covered by the first moment of $\mathbf{X}|Y$.

4.1.1. STIR

For binary Y we have to set $H = 1$, because otherwise $\mathbf{G}(\mathbf{Y})$ in (3.4) cannot have full rank. We use

$$\mathbf{G}(\mathbf{Y}) = \big(y_1 - \bar{y}, \dots, y_n - \bar{y}\big)' \in \mathbb{R}^{n \times 1},$$

which is already centered when taking the expectation w.r.t. Y. For continuous response Y, we can choose H arbitrarily and use Fourier basis elements as functions of Y. The rows of $\mathbf{G}(\mathbf{Y})$ are then given by

$$\mathbf{G}(y_i) = (\cos(2\pi y_i), \sin(2\pi y_i), \cos(2\pi 2 y_i), \sin(2\pi 2 y_i), \cos(2\pi 3 y_i), \dots) \in \mathbb{R}^{1 \times H}, \tag{4.1}$$

for $i = 1, \dots, n$ and we then center the columns of $\mathbf{G}(\mathbf{Y}) \in \mathbb{R}^{n \times H}$ by their empirical mean.

We consider fixed, equal time points (either $\mathcal{T} = \{\exp(t/6 - T/6), t = 1, \ldots, T\}$ or $\mathcal{T} = \{t/T, t = 1, \ldots, T\}$) and random, unequal time points, where, for each individual, T time points are drawn from a uniform distribution on $(0, 1)$. These time points are only drawn once for each setting and then stay the same for all replications of that setting. We use the polynomial basis $s_j(t) = t^j, j = 1, \ldots, d$ for time basis functions. The rows of the resulting $\mathbf{S}$ matrix are then centered by the mean over all time points (T time points for equal time case and nT time points for unequal time case).

With all the components specified, we can then generate the predictors $\mathbf{X}_i$ from (3.8), where we use multivariate normal errors $\text{vec}(\mathbf{E}_i) \overset{iid}{\sim} N(0, \boldsymbol{\Delta})$ or $\text{vec}(\mathbf{E}_i) \overset{iid}{\sim} N(0, \boldsymbol{\Delta}_{y_i})$ for $i = 1, \ldots, n$.

4.1.2. LSIR

As model satisfying the mean structure of the LSIR section (2.17), we will use and generate from

$$\mathbf{X}_i = \boldsymbol{\psi}\mathbf{G}_{y_i}\boldsymbol{\phi}' + \mathbf{E}_i \in \mathbb{R}^{T\times p}, \quad i = 1, \ldots, n, \tag{4.2}$$

with $\boldsymbol{\psi} \in \mathbb{R}^{T\times r}$, $\boldsymbol{\phi} \in \mathbb{R}^{p\times m}$ and some $r \times m$ matrix $\mathbf{G}_{y_i}$, and $\text{vec}(\mathbf{E}) \sim N(0, \boldsymbol{\Delta})$ or $\text{vec}(\mathbf{E}) \sim N(0, \boldsymbol{\Delta}_y)$.

We choose the matrix $\mathbf{G}_{y_i}$, such that $\text{vec}(\mathbf{G}_{y_i})$ agrees with $\mathbf{G}(y_i)$, the i-th row of $\mathbf{G}(\mathbf{Y})$, from the previous subsection for both the binary and continuous case.

4.2. Performance Measures

In this section we define how we compare and measure the performance of our proposed STIR method to other methods. We assess how well the parameters $\mathbf{B}$ and $\boldsymbol{\Delta}$ are estimated and how useful the reduced predictors of our method are to predict new responses.

4.2.1. Estimation Accuracy

To see whether $\widehat{\mathbf{B}}$ is a good point estimator, we consider the average of $\|\mathbf{B} - \widehat{\mathbf{B}}\|_F$ over multiple replications for the scenarios where we generate from our model with true parameter $\mathbf{B}$, where $||.||_F$ denotes the Frobenius norm. We also take a look at $\|\mathbf{B} - \sum_{j=1}^{n_{rep}} \widehat{\mathbf{B}}^j/n_{rep}\|_F$ and $\|\boldsymbol{\Delta} - \sum_{j=1}^{n_{rep}} \widehat{\boldsymbol{\Delta}}^j/n_{rep}\|_F$ for an increasing number of replications to check whether our estimators are unbiased. Here the superscript j denotes the estimator resulting from the j-th replication. For all three measures, we compare the OLS based estimators to the ML based ones.

Since the reduction in our model does not need a point estimate of the true $\mathbf{B}$, but rather of $\text{span}(\mathbf{B}')$, we use the the principal angle to assess closeness of the subspace spanned by $\widehat{\mathbf{B}}'$ to the subspace spanned by $\mathbf{B}'$. The principal angle $0 \leq \theta \leq \pi/2$ can be expressed as

$\cos\theta = \|\mathbf{P}_{\mathbf{B}'}\mathbf{P}_{\widehat{\mathbf{B}}'}\| = \|\mathbf{P}_{\widehat{\mathbf{B}}'}\mathbf{P}_{\mathbf{B}'}\|$, where $\|\cdot\|$ is the spectral norm and $\mathbf{P}_{\mathbf{A}}$ is the orthogonal projection onto the column space of a matrix $\mathbf{A}$ [13]. For two subspaces $\text{span}(\mathbf{A}) \subseteq \mathbb{R}^m$ and $\text{span}(\mathbf{B}) \subseteq \mathbb{R}^m$ with the same dimension, the principal angle θ is zero if and only if $\text{span}(\mathbf{A}) \cap \text{span}(\mathbf{B}) \neq 0$, and $\theta = \pi/2$ if and only if $\text{span}(\mathbf{A})$ is orthogonal to $\text{span}(\mathbf{B})$. We report $1 - \cos(\theta)$, which is zero when $\text{span}(\widehat{\mathbf{B}}') = \text{span}(\mathbf{B}')$.

4.2.2. Modeling and Predicting the Response

We compare the predictive performance of the following reductions. From the STIR model and for $\text{span}\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right) = \text{span}(\widehat{\mathbf{B}}')$ with different k, we use

$$\begin{aligned}
\widehat{R}_{\text{delta}}(\mathbf{X}_i) &= \text{unvec}\left(\widehat{\boldsymbol{\Delta}}^{-1}\,\text{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right), \\
\widehat{R}_{\text{del-kron}}(\mathbf{X}_i) &= \text{unvec}\left((\widehat{\boldsymbol{\Delta}}_p^{-1} \otimes \widehat{\boldsymbol{\Delta}}_T^{-1})\,\text{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right), \\
\widehat{R}_{\text{sigma}}(\mathbf{X}_i) &= \text{unvec}\left(\widehat{\boldsymbol{\Sigma}}_x^{-1}\,\text{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right), \\
\widehat{R}_{\text{sig-kron}}(\mathbf{X}_i) &= \text{unvec}\left((\widehat{\boldsymbol{\Sigma}}_p^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_T^{-1})\,\text{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right),
\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_p$ and $\widehat{\boldsymbol{\Sigma}}_T$ are estimated as in the first step of the LSIR algorithm 2.11, as well as not using any scaling

$$\widehat{R}_{\text{unscaled}}(\mathbf{X}_i) = \mathbf{X}_i\left(\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_k\right).$$

For $\text{span}\left(\widehat{\mathbf{V}}_1, \ldots, \widehat{\mathbf{V}}_k\right) = \text{span}(\widehat{\mathbf{B}}_{\text{ML}}{}')$, we also use

$$\widehat{R}_{\text{delta}}(\mathbf{X}_i) = \text{unvec}\left(\widehat{\boldsymbol{\Delta}}_x^{-1}\,\text{vec}(\mathbf{X}_i)\right)\left(\widehat{\mathbf{V}}_1, \ldots, \widehat{\mathbf{V}}_k\right).$$

All these reductions $\widehat{R}(\mathbf{X}_i)$ are $T \times k$ matrices. When we apply LSIR to a continuous Y, we categorize the observed responses into 10 categories based on the deciles of a $N(0, 0.1)$-distribution. The reduction from LSIR is given by

$$\widehat{R}_{\text{LSIR}}(\mathbf{X}_i) = \widehat{\boldsymbol{\psi}}'\mathbf{X}_i\widehat{\boldsymbol{\phi}} \in \mathbb{R}^{r\times m}.$$

We also compare the performance to using no reduction at all, i.e. $\widehat{R}_{\text{unred}}(\mathbf{X}_i) = \mathbf{X}_i \in \mathbb{R}^{T\times p}$.

For any of these reductions let $\widetilde{\mathbf{X}}_i = \widehat{R}(\mathbf{X}_i)$ be the reduced predictors for $i = 1, \ldots, n$. With those, we model the mean of the response Y through a link function g as

$$g(\mathbb{E}[Y \mid \widetilde{\mathbf{X}}_i]) = \mu + \text{vec}(\boldsymbol{\theta})'\,\text{vec}(\widetilde{\mathbf{X}}_i), \tag{4.3}$$

where $\boldsymbol{\theta}$ is a parameter matrix corresponding to the $\mathbf{X}_i$s. Sometimes we will also allow

$$g(\mathbb{E}[Y \mid \widetilde{\mathbf{X}}_i]) = \mu + \text{vec}(\boldsymbol{\theta})'\,\text{vec}(\widetilde{\mathbf{X}}_i) + \text{vec}(\widetilde{\mathbf{X}}_i)'\boldsymbol{\Theta}\,\text{vec}(\widetilde{\mathbf{X}}_i), \tag{4.4}$$

where $\boldsymbol{\Theta}$ is a symmetric parameter matrix for the interaction and quadratic terms among the components of the reduction. For binary Y, we use the logit link function and obtain ML estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ from fitting a logit model. For continuous Y, we use the identity link to fit a linear regression model and obtain least squares estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$.

For a continuous response Y we will also apply a generalized additive model (GAM) [10],

$$g(\mathbb{E}(Y \mid \widetilde{\mathbf{X}}_i)) = \beta_0 + \sum_{j=1}^{Tk} f_j(\mathrm{vec}(\widetilde{\mathbf{X}}_i)_j). \tag{4.5}$$

where $\mathrm{vec}(\widetilde{\mathbf{X}}_i)_j$ is the j-th element of $\mathrm{vec}(\widetilde{\mathbf{X}}_i)$ and the f_js are smooth functions, for $j = 1, \ldots, Tk$, $i = 1, \ldots, n$. We fit model (4.5) using the `mgcv` package [24] in `R` with REML smoothness estimation.

We then generate 100 new data samples, calculate the reductions of those predictors and use these regression models to predict the new responses from the new reduced predictors.

For a binary response Y, we will assess this performance by the AUC (area under the curve) between the 100 true new responses and the predicted probabilities from the logit model on the reduced predictors. In `R`, we use the `ROCR` package [21] to calculate the AUC.

For a continuous response Y, we assess the performance by the empirical correlation between the true new responses and the predictions from the regression model for the new reduced predictors.

4.3. Simulation Results

Here we state the exact specifications of the scenarios to generate the data for binary and continuous response and give the results.

4.3.1. Binary Response

In the binary case we have to choose $r = m = 1$ for LSIR and $H = 1$ for STIR. All considered settings are listed in Table 4.1. Settings STIR-1,2 and LSIR-1,2 aim to show the difference between same and different time points for the individuals. Settings STIR-1,2 consider large T and smaller p, while STIR-3,4,5,6 do the opposite. Starting from STIR-4 with equally spaced time points, settings STIR-5,6 aim to show the difference between having the markers available 3 times or once, but with a 3 times stronger signal. The scaling of the true $\mathbf{B}$ matrices was chosen in a way to obtain interpretable and meaningful AUC values.

Estimation Accuracy

Table A.1 in the appendix shows the estimation accuracy of the STIR estimates to the true $\mathbf{B}$ for the scenarios where we generate from the STIR model. We see that in each

Table 4.1.: Simulation scenarios for $y_i \sim Bern(0.5), i = 1, \ldots, 500$. For setting STIR-1 - STIR-6 $rank(\mathbf{B}) = 2$, for settings LSIR-1 and LSIR-2 $rank(\mathbf{B}) = 1$. *different time points for each observation in the sample.

Setting	Data generation	True B	Error dist.	Time points t	(T, p, d)
STIR-1	model (3.1)	$\mathbf{B} = 0.1 \cdot \begin{pmatrix} 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \\ 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \end{pmatrix} \in \mathbb{R}^{dH \times p}, H = 1$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(10, 3, 4)
STIR-2	model (3.1)	$\mathbf{B} = 0.1 \cdot \begin{pmatrix} 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \\ 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	ind.*, runif(0,1)	(10, 3, 4)
STIR-3	model (3.1)	$\mathbf{B} = 0.1 \cdot \begin{pmatrix} 1 & 0.7 & 0.1 & 1 & 0.7 & \cdots & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.5 & 0.3 & \cdots & 0.1 \\ 1 & 0.7 & 0.1 & 1 & 0.7 & \cdots & 0.1 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(4, 15, 3)
STIR-4	model (3.1)	$\mathbf{B} = B_4 = \begin{pmatrix} 1 & 0 & 0.5 & 0 & 0.5 & 0 & 1 \\ 0 & 0.25 & 0.5 & 0.25 & 0.5 & 0.25 & 0 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(3, 7, 2)
STIR-5	model (3.1)	$\mathbf{B} = \begin{pmatrix} B_4 & B_4 & B_4 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(3, 21, 2)
STIR-6	model (3.1)	$\mathbf{B} = 3B_4 = 3\begin{pmatrix} 1 & 0 & 0.5 & 0 & 0.5 & 0 & 1 \\ 0 & 0.25 & 0.5 & 0.25 & 0.5 & 0.25 & 0 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(3, 7, 2)
LSIR-1	model (4.2)	$\phi = (1, 0.5, 0.1)'$ $\psi = (1, \ldots, 8)'/8$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(8, 3, 4)
LSIR-2	model (4.2)	$\phi = (1, 0.5, 0.1)'$ $\psi = (1, \ldots, 8)'/8$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	ind.*, runif(0,1)	(8, 3, 4)

setting the estimation accuracy increases for increasing number of observations n, because the average of $\|\mathbf{B} - \hat{\mathbf{B}}\|_F$ decreases. In general, the ML estimate is slightly more accurate than the OLS estimate and both estimators perform better when generating with error covariance depending on Y (a) than when generating from a fixed $\boldsymbol{\Delta}$ (b). This effect is likely caused by the less noisy errors for controls, i.e. $\boldsymbol{\Delta}_0 = 0.1 \cdot AR^1(0.8)$.

For scenarios STIR-1 and 2, the principle angle is almost 0, so the span of the true $\mathbf{B}'$ is estimated very well. In the other STIR scenarios we do see higher numbers of $1 - \cos(\theta)$ in Table A.2, which do decrease for increasing number of observations. This result is not surprising, as for larger p and smaller T, the estimation of span($\mathbf{B}'$) becomes more difficult. Here we only analyzed the OLS based estimator $\widehat{\mathbf{B}}$.

In Figure 4.1 we check whether our estimates for $\mathbf{B}$ and $\boldsymbol{\Delta}$ are unbiased in scenarios STIR-1 and 4 with fixed error covariance not depending on Y (b). We can see that $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}_{\text{ML}}$ perform both well and on a par and seem to be unbiased. Furthermore, the scaling factor used in the ML estimator $\boldsymbol{\Delta}$ leads to a less biased estimate of $\boldsymbol{\Delta}$ compared to the OLS based scaling factor.

Predictive Performance

Next, we look at the predictive performance of the logit models fitted to the reduced predictors. The predictions are evaluated via AUC and Tables 4.2 and 4.3 show the results averaged over 100 replications for our STIR reduction using the error covariance estimate $\widehat{\boldsymbol{\Delta}}$ as scaling in the reduction, the STIR reduction with error covariance estimated as a Kronecker structure used for scaling, the STIR reduction based on $\widehat{\mathbf{B}}_{\text{ML}}$ using $\widehat{\boldsymbol{\Delta}}$, the

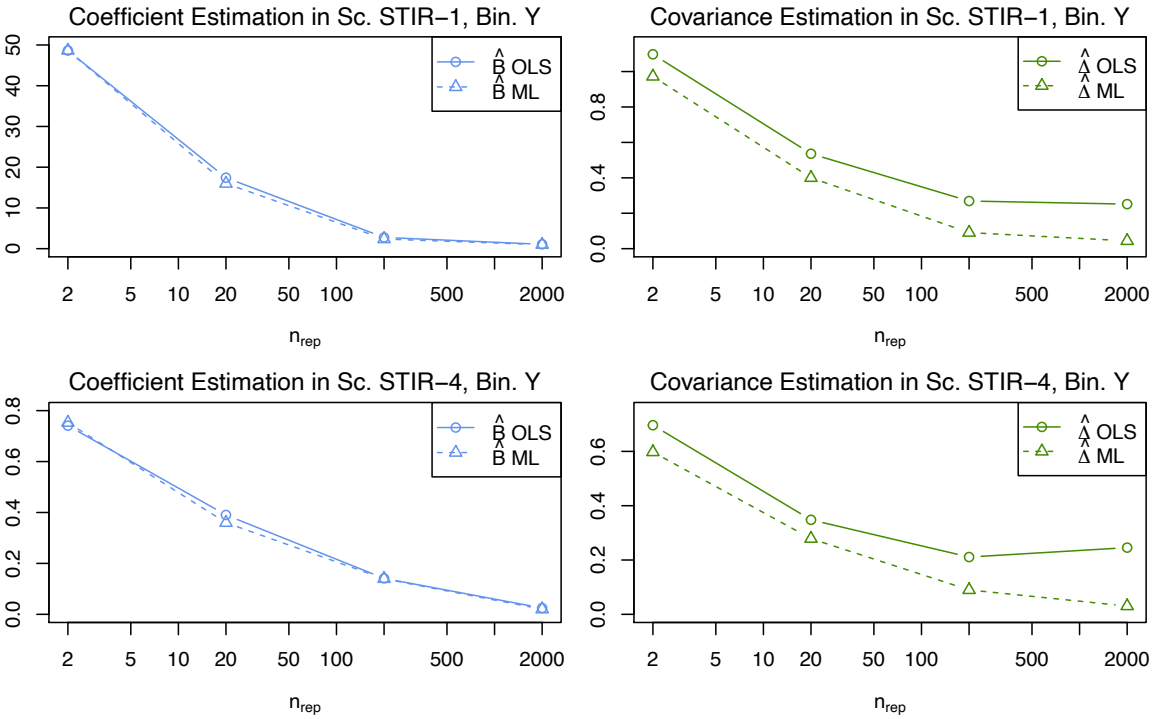

Figure 4.1.: $\|\mathbf{B} - \sum_{j=1}^{n_{rep}} \widehat{\mathbf{B}}^j/n_{rep}\|_F$ and $\|\mathbf{\Delta} - \sum_{j=1}^{n_{rep}} \widehat{\mathbf{\Delta}}^j/n_{rep}\|_F$ is shown for $n = 500$ and increasing number of replications for scenarios STIR-1 and 4 for binary Y with fixed $\mathbf{\Delta}$ independent of Y. We compare the OLS based estimate to the ML based one.

LSIR algorithm and no reduction. The specification of k in these tables only refers to the first 3 columns, not LSIR and Unreduced.

We again see that in general the AUC values are higher, which corresponds to better predictive performance, in the settings where the error covariance matrix depends on Y. Again, this might be due to the fact that the errors are smaller for approximately half of the observations.

Comparing STIR-1 and 2, we see that same time points for the individuals leads to slightly better predictive performance. In LSIR-1 and 2 the performances are very much alike. The performance in scenario STIR-6 is better than in STIR-5, so increasing the signal of given markers leads to better predictive models than using more markers of similar importance.

Across all scenarios, the performance of using $\widehat{\mathbf{B}}$ to calculate the reduction is very similar to the one obtained by $\widehat{\mathbf{B}}_{\mathrm{ML}}$, so in practice (e.g. for the data analysis in Chapter 5) we will use the more easily derived OLS based estimator $\widehat{\mathbf{B}}$.

Our method seems to be fairly competitive compared to LSIR. In all scenarios, even LSIR-1 and 2, where we generate the data from a model corresponding to the LSIR algorithm, our method matches or beats the performance of LSIR even with $k = 1$ or $k = 2$.

In general, it seems that using the unreduced predictors in the logit model is hard to

beat. However, in certain settings (STIR-1,2,3 a) with a second moment effect of Y, i.e. when the error covariance depends on Y, the STIR method has a significantly higher predictive power when including second order terms (squares and interactions), e.g. AUC of 0.971 compared to 0.668 of Unreduced without second order terms. Even with $n = 2000$ observations, the unreduced model including all second order terms does not come close to that performance (AUC of 0.855). In these settings, our method is able to capture the relation of the predictors to the response even with $k = 1$ and a moderate number of observations $n = 500$ when including second order terms, overwhelmingly beating an unreduced model and LSIR.

In the appendix, Tables A.3 and A.4 show the corresponding results for our STIR reduction using all the different matrices for scaling that were mentioned in Section 4.2.2. Across these different reductions, using $\widehat{\mathbf{\Delta}}$ for scaling performs the best and we use this for the comparison with the other methods. For scenarios STIR-1,2 and 3 the differences are minor, but for the others the differences are quite remarkable, e.g. in STIR-4 a using $\widehat{\mathbf{\Delta}}$ for scaling leads to an AUC of 0.925 compared to 0.755 when using $\widehat{\mathbf{\Sigma}}_x$ or 0.891 with no scaling. Especially using $\widehat{\mathbf{\Sigma}}_x$ can lead to poor results in some scenarios.

Table 4.2.: AUC values for 100 new binary Y-observations and values predicted by logit model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR 1-4.

Reduction Covariance	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR ML $\widehat{\mathbf{\Delta}}$	LSIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	Unreduced
Scenario STIR-1 a $(T, p, d) = (10, 3, 4)$					
k=1	0.605 (0.067)	0.536 (0.068)	0.61 (0.065)	0.578 (0.064)	0.637 (0.053)
k=1, 2.ord.terms	0.982 (0.019)	0.977 (0.022)	0.983 (0.017)	0.791 (0.045)	0.633 (0.078)
same w. n=2000	0.996 (0.005)	0.996 (0.006)	0.996 (0.006)	0.811 (0.046)	0.998 (0.005)
k=2	0.631 (0.053)	0.569 (0.069)	0.63 (0.052)		
k=3	0.637 (0.053)	0.637 (0.053)	0.637 (0.053)		
Scenario STIR-1 b					
k=1	0.577 (0.069)	0.524 (0.067)	0.576 (0.067)	0.539 (0.058)	0.594 (0.056)
k=1, 2.ord.terms	0.54 (0.063)	0.507 (0.062)	0.542 (0.064)	0.539 (0.058)	0.508 (0.055)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.562 (0.064)	0.575 (0.061)	0.54 (0.06)
k=2	0.589 (0.061)	0.544 (0.068)	0.587 (0.058)		
k=3	0.594 (0.056)	0.594 (0.056)	0.594 (0.056)		
Scenario STIR-2 a $(T, p, d) = (10, 3, 4)$					
k=1	0.576 (0.064)	0.515 (0.066)	0.588 (0.074)	0.551 (0.065)	0.6 (0.062)
k=1, 2.ord.terms	0.978 (0.022)	0.982 (0.018)	0.98 (0.018)	0.791 (0.046)	0.608 (0.088)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.562 (0.064)	0.575 (0.061)	0.54 (0.06)
k=2	0.601 (0.06)	0.554 (0.073)	0.598 (0.066)		
k=3	0.6 (0.062)	0.6 (0.062)	0.6 (0.062)		
Scenario STIR-2 b					
k=1	0.564 (0.061)	0.523 (0.063)	0.558 (0.06)	0.528 (0.065)	0.567 (0.063)
k=1, 2.ord.terms	0.53 (0.061)	0.51 (0.064)	0.529 (0.064)	0.528 (0.065)	0.501 (0.064)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.562 (0.064)	0.575 (0.061)	0.54 (0.06)
k=2	0.566 (0.058)	0.529 (0.064)	0.568 (0.066)		
k=3	0.567 (0.063)	0.567 (0.063)	0.567 (0.063)		
Scenario STIR-3 a $(T, p, d) = (4, 15, 3)$					
k=1	0.542 (0.086)	0.502 (0.077)	0.542 (0.083)	0.568 (0.064)	0.668 (0.06)
k=1, 2.ord.terms	0.971 (0.017)	0.969 (0.018)	0.973 (0.014)	0.787 (0.05)	0.543 (0.082)
same w. n=2000	0.975 (0.015)	0.972 (0.015)	0.975 (0.015)	0.797 (0.049)	0.855 (0.056)
k=2	0.562 (0.078)	0.504 (0.072)	0.564 (0.078)		
k=3	0.665 (0.057)	0.561 (0.067)	0.671 (0.06)		
Scenario STIR-3 b					
k=1	0.524 (0.076)	0.502 (0.067)	0.525 (0.073)	0.539 (0.057)	0.618 (0.055)
k=1, 2.ord.terms	0.521 (0.073)	0.506 (0.067)	0.52 (0.073)	0.538 (0.058)	0.502 (0.054)
same w. n=2000	0.539 (0.065)	0.501 (0.064)	0.539 (0.065)	0.57 (0.065)	0.509 (0.049)
k=2	0.54 (0.07)	0.501 (0.059)	0.543 (0.071)		
k=3	0.616 (0.056)	0.535 (0.061)	0.624 (0.055)		
Scenario STIR-4 a $(T, p, d) = (3, 7, 2)$					
k=1	0.925 (0.085)	0.794 (0.065)	0.924 (0.086)	0.858 (0.037)	0.983 (0.015)
k=1, 2.ord.terms	0.985 (0.016)	0.964 (0.018)	0.985 (0.016)	0.896 (0.033)	0.941 (0.036)
same w. n=2000	0.992 (0.009)	0.971 (0.016)	0.992 (0.01)	0.9 (0.031)	1 (0.001)
k=2	0.987 (0.012)	0.866 (0.043)	0.987 (0.012)		
Scenario STIR-4 b					
k=1	0.825 (0.123)	0.698 (0.078)	0.832 (0.117)	0.788 (0.042)	0.954 (0.018)
k=1, 2.ord.terms	0.82 (0.125)	0.689 (0.082)	0.828 (0.121)	0.788 (0.042)	0.794 (0.047)
same w. n=2000	0.915 (0.05)	0.74 (0.059)	0.918 (0.045)	0.778 (0.047)	0.925 (0.027)
k=2	0.956 (0.018)	0.793 (0.047)	0.956 (0.018)		

Table 4.3.: AUC values for 100 new binary Y-observations and values predicted by logit model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR 5, 6 and LSIR 1, 2.

Reduction Covariance	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR ML $\widehat{\mathbf{\Delta}}$	LSIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	Unreduced
Scenario STIR-5 a $(T, p, d) = (3, 21, 2)$					
k=1	0.988 (0.043)	0.871 (0.072)	0.985 (0.056)	0.957 (0.025)	1 (0.002)
k=1, 2.ord.terms	0.994 (0.011)	0.974 (0.017)	0.993 (0.012)	0.961 (0.023)	0.547 (0.071)
same w. n=2000	0.994 (0.011)	0.974 (0.017)	0.996 (0.008)	1 (0.002)	0.986 (0.014)
k=2	1 (0.001)	0.955 (0.026)	1 (0.001)		
Scenario STIR-5 b					
k=1	0.911 (0.11)	0.735 (0.095)	0.921 (0.096)	0.897 (0.033)	0.992 (0.006)
k=1, 2.ord.terms	0.908 (0.112)	0.73 (0.096)	0.918 (0.098)	0.897 (0.033)	0.545 (0.07)
same w. n=2000	0.908 (0.112)	0.73 (0.096)	0.98 (0.013)	0.982 (0.015)	0.883 (0.037)
k=2	0.997 (0.004)	0.895 (0.034)	0.997 (0.003)		
Scenario STIR-6 a $(T, p, d) = (3, 7, 2)$					
k=1	1 (0)	0.998 (0.005)	1 (0)	1 (0.001)	1 (0)
k=1, 2.ord.terms	1 (0)	0.998 (0.005)	1 (0)	0.999 (0.002)	0.996 (0.008)
same w. n=2000	1 (0)	0.999 (0.003)	1 (0)	0.999 (0.003)	1 (0)
k=2	1 (0)	0.998 (0.006)	1 (0)		
Scenario STIR-6 b					
k=1	1 (0)	0.983 (0.01)	1 (0)	0.992 (0.006)	1 (0)
k=1, 2.ord.terms	1 (0.001)	0.982 (0.011)	1 (0.001)	0.992 (0.006)	0.997 (0.004)
same w. n=2000	1 (0)	0.983 (0.01)	1 (0)	0.99 (0.007)	1 (0)
k=2	1 (0)	0.993 (0.006)	1 (0)		
Scenario LSIR-1 a $(T, p, d) = (8, 3, 4)$					
k=1	0.849 (0.102)	0.669 (0.123)	0.851 (0.095)	0.846 (0.049)	0.941 (0.027)
k=1, 2.ord.terms	0.983 (0.015)	0.983 (0.015)	0.984 (0.015)	0.889 (0.043)	0.875 (0.054)
same w. n=2000	0.997 (0.005)	0.996 (0.005)	0.996 (0.005)	0.89 (0.034)	1 (0.002)
k=2	0.934 (0.031)	0.86 (0.083)	0.928 (0.035)		
k=3	0.941 (0.027)	0.941 (0.027)	0.941 (0.027)		
Scenario LSIR-1 b					
k=1	0.799 (0.083)	0.635 (0.105)	0.795 (0.102)	0.773 (0.051)	0.884 (0.034)
k=1, 2.ord.terms	0.766 (0.094)	0.606 (0.1)	0.765 (0.105)	0.773 (0.051)	0.661 (0.064)
same w. n=2000	0.813 (0.073)	0.643 (0.098)	0.793 (0.084)	0.779 (0.049)	0.851 (0.042)
k=2	0.869 (0.041)	0.771 (0.092)	0.867 (0.04)		
k=3	0.884 (0.034)	0.884 (0.034)	0.884 (0.034)		
Scenario LSIR-2 a $(T, p, d) = (8, 3, 4)$					
k=1	0.855 (0.091)	0.656 (0.132)	0.856 (0.083)	0.852 (0.046)	0.946 (0.026)
k=1, 2.ord.terms	0.986 (0.013)	0.983 (0.016)	0.982 (0.018)	0.891 (0.033)	0.878 (0.055)
same w. n=2000	0.998 (0.005)	0.997 (0.004)	0.998 (0.003)	0.895 (0.035)	0.999 (0.003)
k=2	0.93 (0.037)	0.842 (0.087)	0.927 (0.036)		
k=3	0.946 (0.026)	0.946 (0.026)	0.946 (0.026)		
Scenario LSIR-2 b					
k=1	0.786 (0.087)	0.619 (0.101)	0.796 (0.089)	0.777 (0.048)	0.887 (0.03)
k=1, 2.ord.terms	0.757 (0.089)	0.591 (0.089)	0.763 (0.095)	0.777 (0.048)	0.663 (0.058)
same w. n=2000	0.811 (0.077)	0.647 (0.104)	0.802 (0.085)	0.783 (0.048)	0.851 (0.038)
k=2	0.869 (0.045)	0.771 (0.079)	0.865 (0.043)		
k=3	0.887 (0.03)	0.887 (0.03)	0.887 (0.03)		

4.3.2. Continuous Response

The settings used for continuous Y are listed in Table 4.4. Similarly to the binary case, settings STIR-1,2 and LSIR-1,2 aim to show the difference between same and different time points for the individuals. Settings STIR-1,2 consider large T and smaller p, while STIR-3,4,5,6 do the opposite. Starting from STIR-4 with equally spaced time points, settings STIR-5,6 aim to show the difference between having the markers available 3 times or once, but with a 3 times stronger signal. In the STIR settings we will use different values for r and m for LSIR.

Table 4.4.: Simulation scenarios for $y_i \sim N(0, 0.1), i = 1, \ldots, 500$. For all settings $rank(\mathbf{B}) = 2$. *different time points for each observation in the sample.

Setting	Data generation	True B	Error dist.	Time points t	(T, p, H, d)
STIR-1	model (3.1)	$\mathbf{B} = \begin{pmatrix} 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \\ 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \end{pmatrix} \in \mathbb{R}^{dH \times p}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(10, 3, 2, 2)
STIR-2	model (3.1)	$\mathbf{B} = \begin{pmatrix} 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \\ 1 & 0.7 & 0.1 \\ 0.5 & 0.3 & 0.1 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	ind.*, runif(0,1)	(10, 3, 2, 2)
STIR-3	model (3.1)	$\mathbf{B} = \begin{pmatrix} 1 & 0.7 & 0.1 & 1 & 0.7 & \cdots & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.5 & 0.3 & \cdots & 0.1 \\ 1 & 0.7 & 0.1 & 1 & 0.7 & \cdots & 0.1 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(3, 15, 3, 1)
STIR-4	model (3.1)	$\mathbf{B} = B_4 = \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 0.5 & 1 & 0.5 & 1 & 0.5 & 0 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(2, 7, 2, 1)
STIR-5	model (3.1)	$\mathbf{B} = \begin{pmatrix} B_4 & B_4 & B_4 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(2, 21, 2, 1)
STIR-6	model (3.1)	$\mathbf{B} = 3B_4 = 3\begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 0.5 & 1 & 0.5 & 1 & 0.5 & 0 \end{pmatrix}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	t/T	(2, 7, 2, 1)
LSIR-1	model (4.2)	$\psi = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \phi = \begin{pmatrix} 0.1 & 1 \\ -0.05 & -1 \\ \vdots & \vdots \\ 0.1 & 1 \\ -0.05 & -1 \end{pmatrix} \in \mathbb{R}^{8 \times 2}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	$\exp(t/6 - T/6)$	(3, 8, 3, 2)
LSIR-2	model (4.2)	$\psi = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \phi = \begin{pmatrix} 0.1 & 1 \\ -0.05 & -1 \\ \vdots & \vdots \\ 0.1 & 1 \\ -0.05 & -1 \end{pmatrix} \in \mathbb{R}^{8 \times 2}$	(a) $N(0, \Delta_y)$ (b) $N(0, \Delta)$	ind.*,runif(0,1)	(3, 8, 3, 2)

Estimation Accuracy

As in the binary case, we first take a look at the estimation accuracy of the parameter estimates estimate to the true $\mathbf{B}$ for the scenarios where we generate from the STIR model, which is given in Table A.5 in the appendix.

Exactly as in the binary case, the average of $\|\mathbf{B} - \hat{\mathbf{B}}\|_F$ decreases for increasing number of observations in all settings, the ML estimate is slightly more accurate than the OLS estimate and both estimators perform better when the error covariance depends on Y. Since for the a version of the scenarios, we scale the fixed $\mathbf{\Delta}$ by a real and Y-dependent constant smaller than one, this effect is likely again caused by smaller noise in the data.

The take-away from the analysis of the principle angle between span($\widehat{\mathbf{B}}'$) and span($\mathbf{B}'$) is the same as in the binary case as well. Table A.6 shows that for scenarios STIR-1 and 2, the principle angle is almost 0, and in the other more difficult STIR scenarios we see decreasing numbers of $1 - \cos(\theta)$ for increasing sample size.

Figure 4.2 shows that also in the continuous case $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}_{\text{ML}}$ seem to be unbiased, but here the ML estimate has a slightly better accuracy. The figure also suggests that the ML estimate for $\boldsymbol{\Delta}$ is less biased than the OLS estimate, as $\|\boldsymbol{\Delta} - \sum_{j=1}^{n_{rep}} \widehat{\boldsymbol{\Delta}}^j / n_{rep}\|_F$ goes to zero for increasing number of repetitions for the ML estimate, while it reaches a plateau for the other scaling factor.

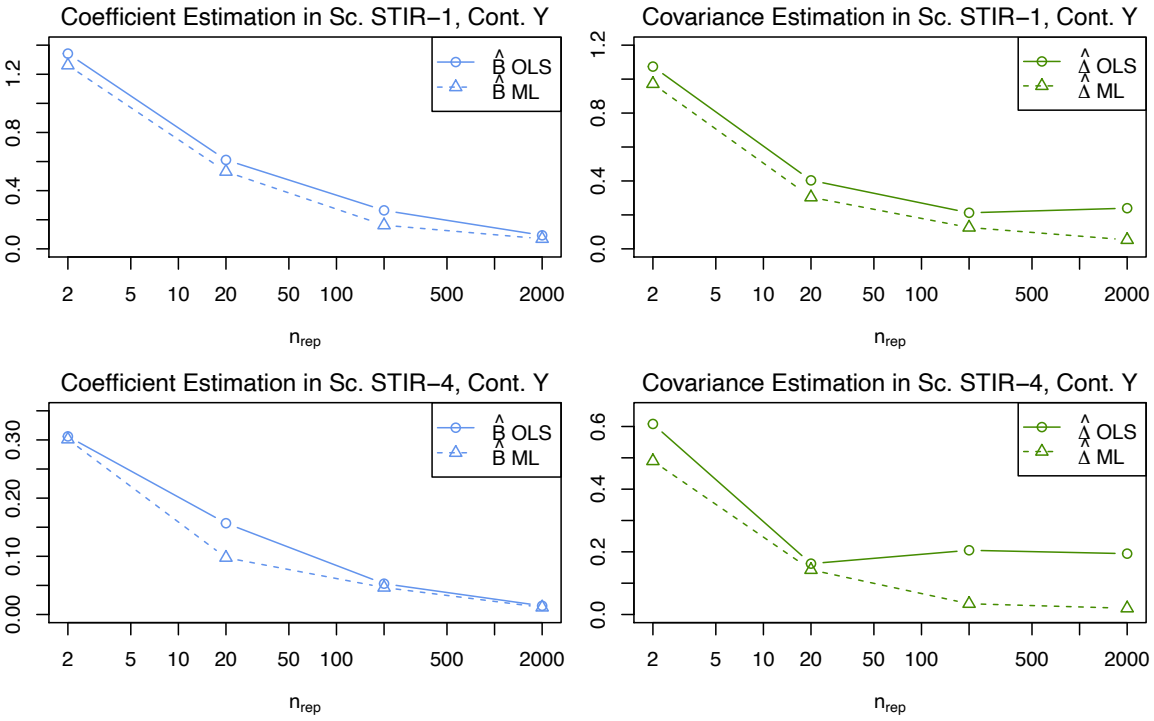

Figure 4.2.: $\|\mathbf{B} - \sum_{j=1}^{n_{rep}} \widehat{\mathbf{B}}^j / n_{rep}\|_F$ and $\|\boldsymbol{\Delta} - \sum_{j=1}^{n_{rep}} \widehat{\boldsymbol{\Delta}}^j / n_{rep}\|_F$ is shown for $n = 500$ and increasing number of replications for scenarios STIR-1 and 4 for continuous Y with fixed $\boldsymbol{\Delta}$ independent of Y. We compare the OLS based estimate to the ML based one.

Predictive Performance

The following tables show the prediction performance of the regression models (linear and gam) fitted to the reduced predictors. It is measured by the empirical correlation between true new responses and their predicted values.

The two main Tables 4.5 and 4.6, as well as Tables A.7 and A.8 in the appendix, use the same column structure as in the binary setting. The specification of k in these tables again

only refers to the first 3 columns, not LSIR and Unreduced, while r and m only refer to LSIR.

Across all settings we again see a slightly better predictive performance, indicated by a higher correlation, for scenarios 'a', where the error covariance matrix depends on Y, compared to 'b', where it does not.

Comparing STIR-1 and 2, there is hardly any difference in performance between using same or different time points for the individuals in the sample. However, in LSIR-1 and 2 the difference for our method is striking, e.g. our method with $k = 1$ achieves a correlation of 0.97 for equal time points (Scenario LSIR-1 a) and 0.268 for different time points (Scenario LSIR-2 a). So our method seems to be very sensitive to the modeling of time. As for binary Y, the performance in scenario STIR-6 is better than in STIR-5, so increasing the signal of given markers leads to better predictive models than using more markers of similar importance.

Here, the performance of using $\widehat{\mathbf{B}}$ to calculate the reduction is again similar to the one obtained by $\widehat{\mathbf{B}}_{\mathrm{ML}}$, but not as alike as in the binary case. Using $\widehat{\mathbf{B}}$ yields better correlations in scenario LSIR-2 (e.g. 0.268 versus 0.188 for $k = 1$), while $\widehat{\mathbf{B}}_{\mathrm{ML}}$ yields slightly better correlations in scenarios STIR-1,2 and 3.

Our method seems to be fairly competitive compared to LSIR. In all scenarios except LSIR-2 our method beats or matches the performance of LSIR even with $k = 1$ or $k = 2$. Using a gam regression model instead of the linear model did only improve the performance slightly in settings STIR-1-4 a (e.g. from 0.475 to 0.502 in STIR-1 a).

The unreduced predictors in the linear model are again hard to beat, but our method is able to at least match the performance in every setting when using $k = 2$ (e.g. 0.218 compared to 0.134 for STIR-3 b), with the only exception of scenario LSIR-2.

Looking at the effect of using different scaling matrices in the reduction for our method in Tables A.7 and A.8 in the appendix, we see that again using $\widehat{\boldsymbol{\Delta}}$ for scaling did perform the best. For all STIR scenarios, using a Kronecker product structure deteriorates the performance drastically. However, in scenario LSIR-2 we do see the opposite, e.g. a correlation of 0.223 for our method with $k = 1$ when using $\widehat{\boldsymbol{\Delta}}$ compared to 0.62 for $\widehat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}_p \otimes \hat{\boldsymbol{\Delta}}_T$.

To summarize, our method did not achieve as astonishing results as in the binary case for continuous Y, but is still competitive to LSIR and using no reduction.

Table 4.5.: Correlation between 100 new continuous Y-observations and values predicted by linear model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR 1-4.

Reduction Covariance	STIR $\hat{\mathbf{\Delta}}$	STIR $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR ML $\hat{\mathbf{\Delta}}$	LSIR $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	Unreduced
Scenario STIR-1 a $(T, p, H, d) = (10, 3, 2, 2)$					
k=1=r=m	0.475 (0.116)	0.312 (0.147)	0.487 (0.103)	0.412 (0.09)	0.504 (0.092)
same w. gam	0.502 (0.123)	0.316 (0.149)	0.516 (0.106)	0.458 (0.096)	0.506 (0.089)
k=1=r=m w.2.ord.	0.463 (0.126)	0.288 (0.159)	0.476 (0.113)	0.459 (0.097)	0.032 (0.089)
same w. n=2000	0.557 (0.097)	0.451 (0.114)	0.56 (0.096)	0.471 (0.106)	0.478 (0.099)
k=2,r=3,m=1	0.508 (0.093)	0.454 (0.091)	0.506 (0.094)	0.411 (0.092)	
k=3=r,m=2	0.504 (0.092)	0.504 (0.092)	0.504 (0.092)	0.419 (0.088)	
r=7,m=2				0.43 (0.092)	
Scenario STIR-1 b					
k=1=r=m	0.421 (0.127)	0.242 (0.145)	0.425 (0.123)	0.374 (0.097)	0.463 (0.1)
same w. gam	0.428 (0.131)	0.233 (0.153)	0.433 (0.128)	0.399 (0.107)	0.447 (0.1)
k=1=r=m w.2.ord.	0.371 (0.138)	0.191 (0.147)	0.369 (0.146)	0.402 (0.107)	0.022 (0.101)
same w. n=2000	0.491 (0.108)	0.361 (0.129)	0.496 (0.107)	0.416 (0.11)	0.377 (0.122)
k=2,r=3,m=1	0.469 (0.098)	0.412 (0.099)	0.469 (0.1)	0.372 (0.098)	
k=3=r,m=2	0.463 (0.1)	0.463 (0.1)	0.463 (0.1)	0.379 (0.094)	
r=7,m=2				0.387 (0.095)	
Scenario STIR-2 a $(T, p, H, d) = (10, 3, 2, 2)$					
k=1=r=m	0.474 (0.109)	0.326 (0.147)	0.502 (0.091)	0.391 (0.097)	0.492 (0.094)
same w. gam	0.494 (0.112)	0.327 (0.151)	0.523 (0.092)	0.429 (0.103)	0.493 (0.091)
k=1=r=m w.2.ord.	0.462 (0.112)	0.3 (0.159)	0.489 (0.092)	0.432 (0.102)	0.035 (0.084)
same w. n=2000	0.527 (0.098)	0.432 (0.115)	0.53 (0.095)	0.445 (0.116)	0.441 (0.095)
k=2,r=3,m=1	0.498 (0.093)	0.45 (0.108)	0.498 (0.093)	0.392 (0.094)	
k=3=r,m=2	0.492 (0.094)	0.492 (0.094)	0.492 (0.094)	0.394 (0.1)	
r=7,m=2				0.404 (0.105)	
Scenario STIR-2 b					
k=1=r=m	0.427 (0.126)	0.26 (0.161)	0.452 (0.109)	0.346 (0.11)	0.454 (0.097)
same w. gam	0.428 (0.13)	0.251 (0.155)	0.457 (0.111)	0.365 (0.118)	0.443 (0.1)
k=1=r=m w.2.ord.	0.378 (0.123)	0.2 (0.149)	0.405 (0.116)	0.369 (0.118)	0.023 (0.108)
same w. n=2000	0.475 (0.1)	0.365 (0.12)	0.478 (0.1)	0.4 (0.117)	0.353 (0.094)
k=2,r=3,m=1	0.465 (0.098)	0.398 (0.113)	0.461 (0.096)	0.348 (0.104)	
k=3=r,m=2	0.454 (0.097)	0.454 (0.097)	0.454 (0.097)	0.359 (0.101)	
r=7,m=2				0.365 (0.101)	
Scenario STIR-3 a $(T, p, H, d) = (3, 15, 3, 1)$					
k=1=r=m	0.166 (0.135)	0.059 (0.112)	0.19 (0.137)	0.148 (0.117)	0.162 (0.122)
same w. gam	0.194 (0.163)	0.06 (0.116)	0.222 (0.159)	0.169 (0.138)	0.16 (0.109)
k=1=r=m w.2.ord.	0.19 (0.166)	0.06 (0.117)	0.224 (0.156)	0.175 (0.135)	-0.004 (0.086)
same w. n=2000	0.306 (0.141)	0.161 (0.108)	0.308 (0.141)	0.188 (0.115)	0.113 (0.093)
k=2,r=1,m=3	0.239 (0.128)	0.128 (0.107)	0.242 (0.126)	0.131 (0.116)	
k=3,r=2,m=3	0.206 (0.131)	0.096 (0.107)	0.199 (0.124)	0.111 (0.117)	
r=2,m=10				0.125 (0.119)	
Scenario STIR-3 b					
k=1=r=m	0.101 (0.136)	0.023 (0.111)	0.124 (0.138)	0.122 (0.125)	0.134 (0.125)
same w. gam	0.105 (0.151)	0.028 (0.11)	0.115 (0.15)	0.117 (0.141)	0.129 (0.118)
k=1=r=m w.2.ord.	0.097 (0.133)	0.022 (0.11)	0.121 (0.14)	0.125 (0.141)	-0.007 (0.097)
same w. n=2000	0.24 (0.151)	0.078 (0.117)	0.254 (0.153)	0.16 (0.115)	0.083 (0.115)
k=2,r=1,m=3	0.218 (0.126)	0.105 (0.108)	0.223 (0.127)	0.107 (0.114)	
k=3,r=2,m=3	0.179 (0.131)	0.073 (0.112)	0.176 (0.125)	0.092 (0.117)	
r=2,m=10				0.09 (0.122)	
Scenario STIR-4 a $(T, p, H, d) = (2, 7, 2, 1)$					
k=1=r=m	0.412 (0.124)	0.024 (0.113)	0.411 (0.126)	0.167 (0.158)	0.666 (0.057)
same w. gam	0.457 (0.128)	0.024 (0.133)	0.453 (0.128)	0.162 (0.185)	0.663 (0.057)
k=1=r=m w.2.ord.	0.457 (0.128)	0.03 (0.127)	0.457 (0.13)	0.166 (0.18)	0.651 (0.059)
same w. n=2000	0.445 (0.131)	0.032 (0.142)	0.444 (0.132)	0.117 (0.171)	0.719 (0.048)
k=2,r=1,m=3	0.668 (0.056)	0.398 (0.081)	0.668 (0.057)	0.408 (0.08)	
r=2,m=4				0.454 (0.081)	
Scenario STIR-4 b					
k=1=r=m	0.362 (0.123)	0.019 (0.118)	0.364 (0.127)	0.154 (0.132)	0.621 (0.062)
same w. gam	0.387 (0.138)	0.035 (0.14)	0.387 (0.142)	0.165 (0.151)	0.615 (0.063)
k=1=r=m w.2.ord.	0.389 (0.137)	0.02 (0.124)	0.39 (0.14)	0.166 (0.142)	0.566 (0.072)
same w. n=2000	0.401 (0.133)	0.041 (0.128)	0.4 (0.135)	0.095 (0.149)	0.655 (0.058)
k=2,r=1,m=3	0.625 (0.062)	0.351 (0.099)	0.625 (0.062)	0.359 (0.095)	
r=2,m=4				0.408 (0.091)	

Table 4.6.: Correlation between 100 new continuous Y-observations and values predicted by linear model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR 5 - 6 and LSIR 1 - 2

Reduction Covariance	STIR $\hat{\boldsymbol{\Delta}}$	STIR $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}_p \otimes \hat{\boldsymbol{\Delta}}_T$	STIR ML $\hat{\boldsymbol{\Delta}}$	LSIR $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}_p \otimes \hat{\boldsymbol{\Delta}}_T$	Unreduced
Scenario STIR-5 a $(T, p, H, d) = (2, 21, 2, 1)$					
k=1=r=m	0.489 (0.136)	0.073 (0.152)	0.489 (0.135)	0.289 (0.221)	0.822 (0.033)
same w. gam	0.507 (0.143)	0.065 (0.211)	0.508 (0.142)	0.317 (0.245)	0.817 (0.034)
k=1=r=m w.2.ord.	0.505 (0.137)	0.075 (0.188)	0.505 (0.136)	0.325 (0.239)	0.051 (0.099)
same w. n=2000	0.515 (0.147)	0.091 (0.182)	0.515 (0.146)	0.303 (0.241)	0.788 (0.041)
k=2,r=1,m=4	0.824 (0.033)	0.588 (0.068)	0.824 (0.033)	0.591 (0.069)	
r=2,m=10				0.652 (0.058)	
Scenario STIR-5 b					
k=1=r=m	0.465 (0.142)	0.064 (0.159)	0.466 (0.142)	0.278 (0.219)	0.794 (0.038)
same w. gam	0.462 (0.163)	0.052 (0.198)	0.462 (0.162)	0.285 (0.245)	0.786 (0.039)
k=1=r=m w.2.ord.	0.475 (0.148)	0.062 (0.186)	0.476 (0.149)	0.298 (0.238)	0.045 (0.112)
same w. n=2000	0.494 (0.144)	0.07 (0.17)	0.494 (0.141)	0.255 (0.222)	0.733 (0.048)
k=2,r=1,m=4	0.799 (0.038)	0.538 (0.071)	0.799 (0.038)	0.539 (0.073)	
r=2,m=10				0.601 (0.068)	
Scenario STIR-6 a $(T, p, H, d) = (2, 7, 2, 1)$					
k=1=r=m	0.586 (0.128)	0.056 (0.171)	0.585 (0.129)	0.407 (0.253)	0.923 (0.017)
same w. gam	0.589 (0.13)	0.043 (0.227)	0.588 (0.132)	0.418 (0.272)	0.923 (0.018)
k=1=r=m w.2.ord.	0.589 (0.129)	0.053 (0.203)	0.588 (0.131)	0.427 (0.266)	0.936 (0.013)
same w. n=2000	0.57 (0.13)	0.063 (0.22)	0.57 (0.131)	0.29 (0.259)	0.951 (0.01)
k=2,r=1,m=3	0.924 (0.017)	0.795 (0.035)	0.924 (0.017)	0.808 (0.038)	
r=2,m=4				0.839 (0.038)	
Scenario STIR-6 b					
k=1=r=m	0.565 (0.129)	0.042 (0.161)	0.565 (0.13)	0.328 (0.225)	0.909 (0.019)
same w. gam	0.565 (0.132)	0.041 (0.205)	0.564 (0.134)	0.331 (0.257)	0.907 (0.02)
k=1=r=m w.2.ord.	0.567 (0.129)	0.054 (0.187)	0.567 (0.131)	0.346 (0.242)	0.913 (0.017)
same w. n=2000	0.56 (0.132)	0.067 (0.206)	0.56 (0.133)	0.247 (0.23)	0.933 (0.013)
k=2,r=1,m=3	0.91 (0.019)	0.756 (0.042)	0.91 (0.019)	0.771 (0.043)	
r=2,m=4				0.81 (0.044)	
Scenario LSIR-1 a $(T, p, H, d) = (3, 8, 3, 2)$					
k=1	0.97 (0.007)	0.971 (0.006)	0.97 (0.007)	0.973 (0.007)	0.973 (0.007)
same w. gam	0.973 (0.006)	0.974 (0.006)	0.973 (0.006)	0.972 (0.006)	0.972 (0.007)
k=1 w.2.ord.	0.97 (0.007)	0.971 (0.006)	0.97 (0.007)	0.916 (0.02)	0.916 (0.02)
same w. n=2000	0.973 (0.007)	0.973 (0.007)	0.973 (0.007)	0.97 (0.007)	0.97 (0.007)
k=2	0.971 (0.006)	0.972 (0.006)	0.971 (0.006)		
k=3	0.971 (0.006)	0.972 (0.006)	0.971 (0.006)		
Scenario LSIR-1 b					
k=1	0.965 (0.008)	0.967 (0.008)	0.965 (0.008)	0.968 (0.008)	0.968 (0.008)
same w. gam	0.966 (0.008)	0.968 (0.008)	0.966 (0.008)	0.968 (0.008)	0.967 (0.008)
k=1 w.2.ord.	0.965 (0.008)	0.966 (0.008)	0.965 (0.009)	0.908 (0.021)	0.908 (0.021)
same w. n=2000	0.97 (0.008)	0.97 (0.007)	0.97 (0.008)	0.965 (0.008)	0.965 (0.008)
k=2	0.966 (0.008)	0.967 (0.007)	0.966 (0.008)		
k=3	0.967 (0.008)	0.967 (0.007)	0.967 (0.007)		
Scenario LSIR-2 a $(T, p, H, d) = (3, 8, 3, 2)$					
k=1	0.268 (0.162)	0.696 (0.273)	0.188 (0.155)	0.975 (0.006)	0.975 (0.006)
same w. gam	0.262 (0.166)	0.695 (0.274)	0.186 (0.156)	0.974 (0.006)	0.974 (0.006)
k=1 w.2.ord.	0.248 (0.167)	0.689 (0.281)	0.175 (0.157)	0.921 (0.018)	0.921 (0.018)
same w. n=2000	0.497 (0.233)	0.898 (0.132)	0.281 (0.207)	0.971 (0.008)	0.971 (0.008)
k=2	0.732 (0.096)	0.961 (0.011)	0.568 (0.162)		
k=3	0.773 (0.092)	0.964 (0.01)	0.631 (0.15)		
Scenario LSIR-2 b					
k=1	0.223 (0.15)	0.62 (0.248)	0.169 (0.142)	0.971 (0.006)	0.971 (0.006)
same w. gam	0.217 (0.152)	0.616 (0.253)	0.166 (0.141)	0.97 (0.006)	0.969 (0.006)
k=1 w.2.ord.	0.202 (0.153)	0.61 (0.258)	0.151 (0.144)	0.91 (0.02)	0.91 (0.02)
same w. n=2000	0.343 (0.221)	0.784 (0.217)	0.219 (0.191)	0.965 (0.009)	0.965 (0.009)
k=2	0.7 (0.114)	0.952 (0.015)	0.569 (0.163)		
k=3	0.75 (0.105)	0.957 (0.013)	0.622 (0.151)		

4.3.3. Wald Test

In this Section weassess the power of the Wald test for variable selection derived in Section 3.6 when generating from our model. Therefore, we consider the binary settings STIR-1 and 4, where we test for the first marker corresponding to the first column of the true $\mathbf{B}$, and the continuous settings STIR-1 and 4, where we test for the third marker corresponding to the third column of the true $\mathbf{B}$. In each setting we multiply the corresponding column of $\mathbf{B}$ by a scaling factor $c > 0$ and compute the power of the test as number of rejected test statistics at level $\alpha = 0.05$ over the number of simulations, which is chosen as 500.

Figure 4.3 shows the rejection rates of the four scenarios for varying scaling factor c, which measures the degree of violation of the null hypothesis. In all four scenarios, the α-level under the null (i.e. $c = 0$) is met and the power increases for higher c and for a higher number of observations ($n = 2000$ versus $n = 500$). Also, the power is higher for fixed error covariances not depending on Y (b). A possible explanation for this is that the test statistic uses the estimated variance-covariance matrix of the vectorized coefficient, which does assume a fixed $\mathbf{\Delta}$, so in the 'a' scenarios this covariance is misspecified.

From this analysis, the variable selection test seems to work fine. We tried to find a comparable test, e.g. testing whether all T coefficients of the selected predictor were equal to zero with a T degrees of freedom Wald test in the logit model for binary Y and a linear F-test in the linear regression model for continuous Y. However, the null hypothesis of these tests seems to not correspond to our null hypothesis and the tests do not meet the α-level under our null.

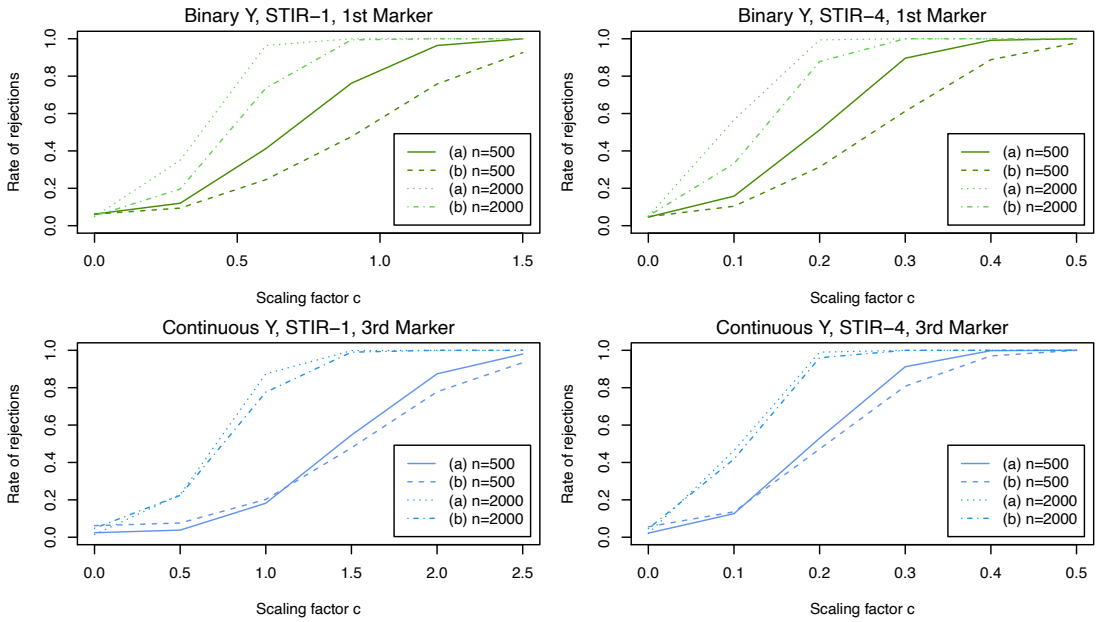

Figure 4.3.: Rejection rates over 500 replications of Wald test testing significance of one marker for STIR scenarios.

5. Data

In this final chapter we apply the STIR method to a real data set and see how it compares to a standard logit model on the unreduced predictors and to the LSIR reduction in predictive performance.

We analyze a medical data set on glioma, a type of brain cancer. The original data are from a study examining the relationship of fourteen serially measured biomarkers with glioma risk in individuals sampled from active component US military personnel [1]. Here, we use a resampled version of this data set, that was provided by R. Pfeiffer.

In this data set we have $n_1 = 131$ cases ($Y = 1$) of glioma and $n_0 = 109$ healthy control subjects ($Y = 0$) with no cancer, giving a total sample size of $n = 240$. The fourteen available markers are: interleukin(IL)-12p40, IL-15, IL-16, IL-7, IL-8, monocyte chemoattractant protein (MCP1), thymus and activation regulated chemokine (TARC), placental growth factor (PLGF), tumor necrosis factor alpha (TNFa), vascular endothelial growth factor (VEGF), hepatocyte growth factor (HGF), transforming growth factor beta (TGFb1), interferon gamma (IFNγ) and IL-10.

All these markers were measured in serum at three time points prior to glioma diagnosis for cases, or the date of selection for controls. A typical space of time between the measurements for the study participants is two years. Also, the age at diagnosis or selection is given. We analyse two different versions of the data set. The first is the full version, where we log-log transform IL-8, log-transform all other markers and include the age at each measurement. For IL-10 we have to add a constant to make all values positive before log-transforming. In the second version, we exclude age and the two markers IL-8 and IL-10, because these two markers do not seem to follow a normal distribution.

As in the simulations, we measure the predictive performance by the AUC, but here we have to apply a cross-validation procedure. To obtain empirical confidence intervals for the resulting AUCs, we bootstrap n_1 cases and n_0 controls with replacement 100 times and for each resulting data set we use the following 10-fold cross-validation procedure. For each of the 10 folds we estimate the STIR and LSIR reductions from the data withholding that fold. With the reductions of the predictors used to find these reductions we fit a logit model to the corresponding responses in a similar way to the simulations, where we also include all second order terms of the predictors, i.e. all two-way interactions and the square of each predictor. Then we calculate the reductions for the 'new' observations in the fold and predict their responses from the logit model. This way we have valid predictions (i.e. from a model that has not yet seen this observation) for each observation after going through all folds and an AUC can be obtained. We report the mean and a 95% confidence interval of the AUCs over all bootstrapped data sets. As baseline comparison, we again also use a logit model on the unreduced predictors following the same cross-validation procedure.

The individual time points are given as time before diagnosis/selection at each measurement as negative values. We tried many different transformations of those time points and used different basis functions of time, e.g. polynomial or Fourier basis.

Table 5 shows the results for transformed time points, such that the time point of the first measurement is one for each individual and the difference to the next ones is the actual time passed between the measurements (in years). We also applied the STIR method assuming equal time points $t = 1/3, 2/3$ and 1 for each individual and used a B-spline basis with degree $d = 2$ in both cases (using the function `bs()` from the base `R` package `splines`). The term 'quadratic effects' indicates that all second order terms were included in the logit models. Doing this for the unreduced predictors, the number of variables in the logit model would exceed n and is therefore infeasible.

For both versions of the data set, the STIR reduction always performs better when using equal time points for all individuals. A possible explanation for this is that the estimation of $\mathbf{B}$ is computationally more efficient for equal time points. In the larger data set, which includes age at each measurement, LSIR reaches the overall best performance of 0.672. STIR reaches an AUC of 0.639 for equal time points and $k = 2$, beating the unreduced predictors' AUC of 0.609. On the smaller data set, excluding age and non-normal predictors, STIR with equal time and $k = 2$ performs similar to the unreduced predictors with AUCs of 0.625 and 0.622, which is higher than LSIR's AUC of 0.597.

In summary, the STIR-method can be seen to be fairly competitive to LSIR. However, we do not see such an outstanding performance of STIR as in the binary simulations, when the relation between the predictors and the response was not just in the mean, but also in the second moment of the predictors, which seems not to be the case in this glioma data set. Also, the prediction task for this data set is quite challenging in general, as can be seen by the relatively low AUC values of the unreduced logit models.

Table 5.1.: AUC estimates and 95% confidence intervals (CIs) for the glioma data, where $(p, T) = (12, 3)$ or $(p, T) = (15, 3)$ and the total sample size is $n = 240$. We use equal time points $(1/3, 2/3, 1)$ and unequal time points and model **S** using spline basis functions.

Reduction	**AUC (95%CI)**	**AUC (95%CI)**
	Unequal time points	**Equal time points**
Data including age, IL-8 and IL-10, $p = 15$		
STIR $(d = 2, k = 1)$	0.572 (0.49,0.666)	0.585 (0.509,0.672)
STIR, quadratic effects $(d = 2, k = 1)$	0.509 (0.405,0.604)	0.525 (0.406,0.617)
STIR $(d = 2, k = 2)$	0.612 (0.527,0.69)	0.639 (0.543,0.745)
STIR, quadratic effects $(d = 2, k = 2)$	0.533 (0.432,0.634)	0.534 (0.418,0.633)
LSIR (r=1,m=1)	0.672 (0.581,0.751)	
LSIR, quadratic effects (r=1,m=1)	0.669 (0.581,0.748)	
Unreduced	0.609 (0.539,0.695)	
Data with excluded predictors, $p = 12$		
STIR $(d = 2, k = 1)$	0.574 (0.501,0.667)	0.578 (0.491,0.647)
STIR, quadratic effects $(d = 2, k = 1)$	0.517 (0.433,0.62)	0.524 (0.388,0.598)
STIR $(d = 2, k = 2)$	0.603 (0.532,0.691)	0.625 (0.529,0.725)
STIR, quadratic effects $(d = 2, k = 2)$	0.538 (0.443,0.644)	0.536 (0.41,0.643)
LSIR (r=1,m=1)	0.597 (0.504,0.713)	
LSIR, quadratic effects (r=1,m=1)	0.586 (0.48,0.705)	
Unreduced	0.622 (0.526,0.706)	

6. Conclusion

In this thesis we propose a new sufficient dimension reduction approach for longitudinally measured predictors and a real response. Many of the usual existing reduction techniques, where the longitudinal structure is ignored, can suffer from a loss of information, since the time structure is ignored.

The STIR model assumes the conditional mean of the predictors given the response to consist of linear combinations of functions of time and functions of the response. The reduction derived in STIR only reduces the markers, while accounting for the time effect through the modeling of the mean. Advantages of the model are that different time points for individuals can be modeled and the reduction of markers allows an easier interpretation of the reduction.

In the simulations we see that the STIR method is sensitive to the specific modeling of the time points, e.g. there is a big difference in predictive performance between scenarios LSIR-1 and 2 in the continuous response case. Other limitations of the proposed STIR method are that finding a satisfying modeling of time can be difficult and is not straightforward, and that the reduction is not minimal.

In STIR, using the ML estimator $\widehat{\mathbf{B}}_{\mathrm{ML}}$ for estimating the parameter $\mathbf{B}$ does not significantly change the predictive performance compared to the OLSbased estimator.

Throughout most of the simulation settings and on a real data set, STIR is competitive to LSIR and to standard regression methods using the vectorized unreduced predictors (logit model in binary case and linear regression in continuous case).

In simulation settings for binary response regressions, where the error covariance matrix depends on Y, STIR excels in predictive performance, beating the other methods by far. For example in binary scenario STIR-2 a it reaches a high AUC of over 0.95 for $k = 1$ when including all second order terms, while LSIR does only achieve an AUC of under 0.8 and a standard logit model reaches an AUC of around 0.6.

In general, the reduction space in STIR covers more than just the first moment. In those scenarios it covers almost all of the modeling information, which the standard logit model is not able to pick up even for significantly large sample sizes.

A. Appendix

Lemma A.1 (Inverse Regression)**.** *For two random quantities* $\mathbf{X} \in \mathbb{R}^p, Y \in \mathbb{R}$ *with a (joint) distribution we have that*

$$F(Y|\mathbf{X}) = F(Y|R(\mathbf{X})) \iff F(\mathbf{X}|Y, R(\mathbf{X})) = F(\mathbf{X}|R(\mathbf{X})).$$

Proof. We give a proof idea for the case that $\mathbf{X}$ and Y have a joint continuous distribution. In form of densities the statement equivalently reads as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \frac{f_{Y,\mathbf{X}}(y,\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \overset{!}{=} \frac{f_{Y,R(\mathbf{X})}(y,R(\mathbf{x}))}{f_{R(\mathbf{X})}(R(\mathbf{x}))} = f_{Y|R(\mathbf{X})}(y|R(\mathbf{x})) \iff \tag{A.1}$$

$$f_{\mathbf{X}|R(\mathbf{X}),Y}(\mathbf{x}|R(\mathbf{x}),y) = \frac{f_{Y,\mathbf{X},R(\mathbf{X})}(y,\mathbf{x},R(\mathbf{x}))}{f_{Y,R(\mathbf{X})}(y,R(\mathbf{x}))} \overset{!}{=} \frac{f_{\mathbf{X},R(\mathbf{X})}(\mathbf{x},R(\mathbf{x}))}{f_{R(\mathbf{X})}(R(\mathbf{x}))} = f_{\mathbf{X}|R(\mathbf{X})}(\mathbf{x}|R(\mathbf{x})). \tag{A.2}$$

Noting that $f_{Y,\mathbf{X},R(\mathbf{X})}(y,\mathbf{x},\mathbf{r}) = \begin{cases} 0 & \mathbf{r} \neq R(\mathbf{x}) \\ f_{Y,\mathbf{X}}(y,\mathbf{x}) & \mathbf{r} = R(\mathbf{x}) \end{cases}$, the equivalence is then easy to show. □

Lemma A.2 (Properties of Kronecker Product)**.** *The Kronecker product is bilinear and associative. Additionally it satisfies the following useful properties for matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ *of compatible dimensions. We do not give the (rather easy) proofs here.*

- $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$,
- $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ *for invertible* $\mathbf{A}$ *and* $\mathbf{B}$,
- $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$,
- $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \operatorname{vec}(\mathbf{B})$, *often applied with* $\mathbf{A}$ *or* $\mathbf{C}$ *being the identity,*
- $\operatorname{rank}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{rank}(\mathbf{A}) \operatorname{rank}(\mathbf{B})$.

The following facts about matrix derivatives are taken from [16, Section 2.5].

Lemma A.3. *For matrices of compatible dimensions* $\mathbf{A}, \mathbf{B}, \mathbf{X}$ *it holds that*

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AXB}) = \mathbf{A}'\mathbf{B}', \tag{A.3}$$

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AX}'\mathbf{B}) = \mathbf{BA}, \tag{A.4}$$

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AX}'\mathbf{BX}) = \mathbf{BXA} + \mathbf{B}'\mathbf{XA}'. \tag{A.5}$$

A.1. Additional Tables for Simulations

A.1.1. Binary Response

Table A.1.: Estimation accuracy $\|\mathbf{B} - \hat{\mathbf{B}}\|_F$ for binary Y, $y_i \sim Bern(0.5), i = 1, \ldots, n$. Results are means and standard deviations (in parentheses) over 500 replications.

	STIR-1	STIR-2	STIR-3	STIR-4	STIR-5	STIR-6
$n = 200$						
OLS (a)	52.998 (21.909)	31.302 (13.252)	161.614 (29.365)	1.953 (0.499)	3.386 (0.519)	1.953 (0.499)
OLS (b)	71.332 (28.744)	42.314 (18.114)	218.311 (38.985)	2.672 (0.685)	4.609 (0.699)	2.672 (0.685)
ML (a)	53.042 (21.996)	18.15 (7.568)	171.289 (31.84)	1.952 (0.498)	3.517 (0.553)	1.952 (0.498)
ML (b)	69.744 (28.633)	25.907 (10.744)	226.227 (41.71)	2.634 (0.681)	4.656 (0.716)	2.634 (0.681)
$n = 500$						
OLS (a)	33.952 (13.562)	19.735 (8.05)	101.941 (19.67)	1.214 (0.333)	2.144 (0.336)	1.214 (0.333)
OLS (b)	45.539 (18.554)	26.145 (10.619)	137.283 (26.536)	1.617 (0.448)	2.877 (0.465)	1.617 (0.448)
ML (a)	32.874 (13.74)	11.223 (4.456)	104.463 (19.765)	1.191 (0.323)	2.145 (0.337)	1.191 (0.323)
ML (b)	44.12 (18.538)	15.45 (6.019)	139.13 (26.584)	1.572 (0.434)	2.855 (0.46)	1.572 (0.434)
$n = 2000$						
OLS (a)	17.002 (6.86)	9.935 (4.334)	50.536 (9.566)	0.611 (0.162)	1.087 (0.167)	0.611 (0.162)
OLS (b)	22.782 (9.666)	13.468 (5.863)	69.021 (12.379)	0.826 (0.209)	1.456 (0.215)	0.826 (0.209)
ML (a)	16.345 (6.5)	5.816 (2.454)	50.646 (9.738)	0.595 (0.155)	1.055 (0.166)	0.595 (0.155)
ML (b)	21.627 (9.274)	7.837 (3.242)	69.035 (12.447)	0.803 (0.207)	1.41 (0.213)	0.803 (0.207)
$n = 10000$						
OLS (a)	7.664 (3.089)	4.434 (1.792)	22.735 (4.201)	0.272 (0.071)	0.485 (0.072)	0.272 (0.071)
OLS (b)	10.322 (4.188)	5.96 (2.438)	30.847 (5.695)	0.359 (0.094)	0.65 (0.095)	0.359 (0.094)
ML (a)	7.3 (3.065)	2.438 (1.015)	22.719 (4.158)	0.263 (0.072)	0.47 (0.068)	0.263 (0.072)
ML (b)	9.796 (4.109)	3.298 (1.394)	30.788 (5.683)	0.347 (0.092)	0.629 (0.093)	0.347 (0.092)

Table A.2.: $1 - \cos(\theta)$ for binary Y, $y_i \sim Bern(0.5), i = 1, \ldots, n$, where θ is the principle angle between span($\mathbf{B}'$) and span($\widehat{\mathbf{B}}'$). Results are means over 500 replications.

	STIR-1	STIR-2	STIR-3	STIR-4	STIR-5	STIR-6
$n = 200$						
OLS (a)	-5.666e-12	-5.873e-13	1.967e-01	7.223e-03	1.079e-02	8.096e-04
OLS (b)	-3.749e-12	-7.240e-13	2.694e-01	1.302e-02	1.911e-02	1.480e-03
$n = 500$						
OLS (a)	-8.331e-13	-2.998e-13	1.079e-01	2.870e-03	4.427e-03	3.197e-04
OLS (b)	-5.460e-12	-1.936e-13	1.660e-01	5.345e-03	7.890e-03	5.982e-04
$n = 2000$						
OLS (a)	-8.826e-13	-3.989e-13	3.249e-02	7.400e-04	1.071e-03	8.238e-05
OLS (b)	-7.534e-13	-2.196e-13	5.605e-02	1.321e-03	1.915e-03	1.472e-04
$n = 10000$						
OLS (a)	-8.106e-13	-1.283e-13	6.806e-03	1.502e-04	2.120e-04	1.669e-05
OLS (b)	-3.260e-13	-1.343e-13	1.200e-02	2.691e-04	3.895e-04	2.993e-05

Table A.3.: AUC values for 100 new binary Y-observations and values predicted by logit model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR-1-4.

Reduction Scaling	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR $\widehat{\mathbf{\Sigma}}_x$	STIR $\widehat{\mathbf{\Sigma}}_x = \hat{\mathbf{\Sigma}}_p \otimes \hat{\mathbf{\Sigma}}_T$	STIR $\mathbf{I}_{Tp}$
Scenario STIR-1 a $(T, p, d) = (10, 3, 4)$					
k=1	0.605 (0.067)	0.536 (0.068)	0.597 (0.07)	0.536 (0.068)	0.542 (0.071)
k=1, 2.ord.terms	0.982 (0.019)	0.977 (0.022)	0.981 (0.019)	0.977 (0.022)	0.977 (0.021)
same w. n=2000	0.996 (0.005)	0.996 (0.006)	0.997 (0.004)	0.996 (0.006)	0.996 (0.007)
k=2	0.631 (0.053)	0.569 (0.069)	0.628 (0.055)	0.569 (0.069)	0.577 (0.067)
k=3	0.637 (0.053)	0.637 (0.053)	0.637 (0.053)	0.637 (0.053)	0.637 (0.053)
Scenario STIR-1 b					
k=1	0.577 (0.069)	0.524 (0.067)	0.573 (0.069)	0.524 (0.067)	0.525 (0.068)
k=1, 2.ord.terms	0.54 (0.063)	0.507 (0.062)	0.536 (0.063)	0.507 (0.062)	0.509 (0.06)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.56 (0.067)	0.518 (0.061)	0.52 (0.063)
k=2	0.589 (0.061)	0.544 (0.068)	0.587 (0.061)	0.544 (0.068)	0.549 (0.068)
k=3	0.594 (0.056)	0.594 (0.056)	0.594 (0.056)	0.594 (0.056)	0.594 (0.056)
Scenario STIR-2 a $(T, p, d) = (10, 3, 4)$					
k=1	0.576 (0.064)	0.515 (0.066)	0.571 (0.065)	0.515 (0.065)	0.517 (0.067)
k=1, 2.ord.terms	0.978 (0.022)	0.982 (0.018)	0.978 (0.021)	0.981 (0.021)	0.98 (0.021)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.56 (0.067)	0.518 (0.061)	0.52 (0.063)
k=2	0.601 (0.06)	0.554 (0.073)	0.599 (0.06)	0.554 (0.073)	0.558 (0.072)
k=3	0.6 (0.062)	0.6 (0.062)	0.6 (0.062)	0.6 (0.062)	0.6 (0.062)
Scenario STIR-2 b					
k=1	0.564 (0.061)	0.523 (0.063)	0.561 (0.061)	0.523 (0.064)	0.526 (0.062)
k=1, 2.ord.terms	0.53 (0.061)	0.51 (0.064)	0.527 (0.061)	0.51 (0.064)	0.511 (0.062)
same w. n=2000	0.562 (0.067)	0.518 (0.061)	0.56 (0.067)	0.518 (0.061)	0.52 (0.063)
k=2	0.566 (0.058)	0.529 (0.064)	0.565 (0.058)	0.529 (0.064)	0.534 (0.065)
k=3	0.567 (0.063)	0.567 (0.063)	0.567 (0.063)	0.567 (0.063)	0.567 (0.063)
Scenario STIR-3 a $(T, p, d) = (4, 15, 3)$					
k=1	0.542 (0.086)	0.502 (0.077)	0.524 (0.085)	0.502 (0.076)	0.506 (0.078)
k=1, 2.ord.terms	0.971 (0.017)	0.969 (0.018)	0.972 (0.017)	0.969 (0.018)	0.97 (0.017)
same w. n=2000	0.975 (0.015)	0.972 (0.015)	0.975 (0.015)	0.972 (0.015)	0.974 (0.016)
k=2	0.562 (0.078)	0.504 (0.072)	0.54 (0.077)	0.504 (0.072)	0.518 (0.079)
k=3	0.665 (0.057)	0.561 (0.067)	0.661 (0.057)	0.561 (0.067)	0.629 (0.068)
Scenario STIR-3 b					
k=1	0.524 (0.076)	0.502 (0.067)	0.513 (0.078)	0.502 (0.067)	0.507 (0.07)
k=1, 2.ord.terms	0.521 (0.073)	0.506 (0.067)	0.512 (0.073)	0.505 (0.067)	0.508 (0.067)
same w. n=2000	0.539 (0.065)	0.501 (0.064)	0.533 (0.064)	0.501 (0.064)	0.511 (0.069)
k=2	0.54 (0.07)	0.501 (0.059)	0.526 (0.071)	0.501 (0.059)	0.509 (0.063)
k=3	0.616 (0.056)	0.535 (0.061)	0.613 (0.057)	0.535 (0.061)	0.578 (0.061)
Scenario STIR-4 a $(T, p, d) = (3, 7, 2)$					
k=1	0.925 (0.085)	0.794 (0.065)	0.755 (0.124)	0.781 (0.069)	0.891 (0.088)
k=1, 2.ord.terms	0.985 (0.016)	0.964 (0.018)	0.966 (0.02)	0.964 (0.018)	0.977 (0.018)
same w. n=2000	0.992 (0.009)	0.971 (0.016)	0.977 (0.017)	0.971 (0.016)	0.985 (0.011)
k=2	0.987 (0.012)	0.866 (0.043)	0.985 (0.013)	0.866 (0.043)	0.978 (0.016)
Scenario STIR-4 b					
k=1	0.825 (0.123)	0.698 (0.078)	0.706 (0.13)	0.69 (0.078)	0.779 (0.11)
k=1, 2.ord.terms	0.82 (0.125)	0.689 (0.082)	0.697 (0.132)	0.682 (0.082)	0.776 (0.11)
same w. n=2000	0.915 (0.05)	0.74 (0.059)	0.83 (0.091)	0.736 (0.061)	0.862 (0.062)
k=2	0.956 (0.018)	0.793 (0.047)	0.954 (0.018)	0.793 (0.047)	0.933 (0.022)

Table A.4.: AUC values for 100 new binary Y-observations and values predicted by logit model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR-5,6 and LSIR-1,2.

Reduction Scaling	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR $\widehat{\mathbf{\Sigma}}_x$	STIR $\widehat{\mathbf{\Sigma}}_x = \hat{\mathbf{\Sigma}}_p \otimes \hat{\mathbf{\Sigma}}_T$	STIR $\mathbf{I}_{Tp}$
Scenario STIR-5 a $(T, p, d) = (3, 21, 2)$					
k=1	0.988 (0.043)	0.871 (0.072)	0.635 (0.092)	0.82 (0.081)	0.974 (0.044)
k=1, 2.ord.terms	0.994 (0.011)	0.974 (0.017)	0.952 (0.022)	0.967 (0.02)	0.992 (0.012)
same w. n=2000	0.994 (0.011)	0.974 (0.017)	0.952 (0.022)	0.967 (0.02)	0.992 (0.012)
k=2	1 (0.001)	0.955 (0.026)	0.994 (0.01)	0.955 (0.026)	1 (0.001)
Scenario STIR-5 b					
k=1	0.911 (0.11)	0.735 (0.095)	0.614 (0.106)	0.704 (0.095)	0.861 (0.122)
k=1, 2.ord.terms	0.908 (0.112)	0.73 (0.096)	0.61 (0.105)	0.7 (0.096)	0.857 (0.125)
same w. n=2000	0.908 (0.112)	0.73 (0.096)	0.61 (0.105)	0.7 (0.096)	0.857 (0.125)
k=2	0.997 (0.004)	0.895 (0.034)	0.992 (0.006)	0.895 (0.034)	0.997 (0.003)
Scenario STIR-6 a $(T, p, d) = (3, 7, 2)$					
k=1	1 (0)	0.998 (0.005)	0.746 (0.077)	0.987 (0.015)	1 (0)
k=1, 2.ord.terms	1 (0)	0.998 (0.005)	0.957 (0.021)	0.996 (0.007)	1 (0)
same w. n=2000	1 (0)	0.999 (0.003)	0.963 (0.02)	0.997 (0.006)	1 (0.002)
k=2	1 (0)	0.998 (0.006)	1 (0)	0.998 (0.006)	1 (0)
Scenario STIR-6 b					
k=1	1 (0)	0.983 (0.01)	0.792 (0.082)	0.966 (0.021)	1 (0.001)
k=1, 2.ord.terms	1 (0.001)	0.982 (0.011)	0.787 (0.084)	0.964 (0.021)	0.999 (0.004)
same w. n=2000	1 (0)	0.983 (0.01)	0.823 (0.064)	0.971 (0.016)	1 (0)
k=2	1 (0)	0.993 (0.006)	1 (0)	0.993 (0.006)	1 (0)
Scenario LSIR-1 a $(T, p, d) = (8, 3, 4)$					
k=1	0.849 (0.102)	0.669 (0.123)	0.756 (0.121)	0.666 (0.123)	0.704 (0.125)
k=1, 2.ord.terms	0.983 (0.015)	0.983 (0.015)	0.985 (0.017)	0.983 (0.014)	0.981 (0.014)
same w. n=2000	0.997 (0.005)	0.996 (0.005)	0.996 (0.005)	0.996 (0.005)	0.996 (0.006)
k=2	0.934 (0.031)	0.86 (0.083)	0.909 (0.056)	0.859 (0.084)	0.877 (0.076)
k=3	0.941 (0.027)	0.941 (0.027)	0.941 (0.027)	0.941 (0.027)	0.941 (0.027)
Scenario LSIR-1 b					
k=1	0.799 (0.083)	0.635 (0.105)	0.743 (0.1)	0.634 (0.104)	0.66 (0.105)
k=1, 2.ord.terms	0.766 (0.094)	0.606 (0.1)	0.703 (0.109)	0.605 (0.1)	0.63 (0.1)
same w. n=2000	0.813 (0.073)	0.643 (0.098)	0.757 (0.088)	0.641 (0.098)	0.665 (0.1)
k=2	0.869 (0.041)	0.771 (0.092)	0.849 (0.056)	0.77 (0.092)	0.793 (0.08)
k=3	0.884 (0.034)	0.884 (0.034)	0.884 (0.034)	0.884 (0.034)	0.884 (0.034)
Scenario LSIR-2 a $(T, p, d) = (8, 3, 4)$					
k=1	0.855 (0.091)	0.656 (0.132)	0.77 (0.108)	0.654 (0.132)	0.695 (0.13)
k=1, 2.ord.terms	0.986 (0.013)	0.983 (0.016)	0.981 (0.017)	0.983 (0.016)	0.978 (0.017)
same w. n=2000	0.998 (0.005)	0.997 (0.004)	0.998 (0.003)	0.997 (0.004)	0.997 (0.004)
k=2	0.93 (0.037)	0.842 (0.087)	0.901 (0.061)	0.84 (0.088)	0.869 (0.082)
k=3	0.946 (0.026)	0.946 (0.026)	0.946 (0.026)	0.946 (0.026)	0.946 (0.026)
Scenario LSIR-2 b					
k=1	0.786 (0.087)	0.619 (0.101)	0.74 (0.096)	0.618 (0.1)	0.648 (0.106)
k=1, 2.ord.terms	0.757 (0.089)	0.591 (0.089)	0.708 (0.096)	0.59 (0.089)	0.62 (0.091)
same w. n=2000	0.811 (0.077)	0.647 (0.104)	0.767 (0.093)	0.646 (0.104)	0.668 (0.106)
k=2	0.869 (0.045)	0.771 (0.079)	0.853 (0.057)	0.77 (0.079)	0.792 (0.079)
k=3	0.887 (0.03)	0.887 (0.03)	0.887 (0.03)	0.887 (0.03)	0.887 (0.03)

A.1.2. Continuous Response

Table A.5.: Estimation accuracy $\|\mathbf{B} - \hat{\mathbf{B}}\|_F$ for continuous Y, $y_i \sim N(0, 0.1), i = 1, \ldots, n$. Results are means and standard deviations (in parentheses) over 500 replications.

	STIR-1	STIR-2	STIR-3	STIR-4	STIR-5	STIR-6
$n = 200$						
OLS (a)	4.032 (1.418)	3.275 (1.191)	10.076 (4.312)	0.87 (0.209)	1.531 (0.26)	0.87 (0.209)
OLS (b)	5.19 (1.812)	4.135 (1.511)	13.736 (5.25)	1.112 (0.274)	1.959 (0.319)	1.112 (0.274)
ML (a)	3.705 (1.32)	2.24 (0.831)	9.076 (3.955)	0.791 (0.196)	1.425 (0.242)	0.791 (0.196)
ML (b)	4.723 (1.678)	2.841 (0.993)	12.456 (4.869)	1.01 (0.261)	1.824 (0.293)	1.01 (0.261)
$n = 500$						
OLS (a)	2.512 (0.895)	2.066 (0.703)	5.662 (1.806)	0.537 (0.136)	0.957 (0.133)	0.537 (0.136)
OLS (b)	3.218 (1.125)	2.635 (0.905)	8.039 (2.288)	0.685 (0.175)	1.234 (0.171)	0.685 (0.175)
ML (a)	2.244 (0.8)	1.376 (0.463)	4.824 (1.55)	0.492 (0.123)	0.868 (0.125)	0.492 (0.123)
ML (b)	2.911 (1.044)	1.759 (0.623)	6.886 (2.021)	0.622 (0.157)	1.12 (0.155)	0.622 (0.157)
$n = 2000$						
OLS (a)	1.232 (0.411)	1.048 (0.344)	2.579 (0.544)	0.267 (0.061)	0.471 (0.062)	0.267 (0.061)
OLS (b)	1.576 (0.53)	1.331 (0.445)	3.734 (0.735)	0.344 (0.079)	0.609 (0.085)	0.344 (0.079)
ML (a)	1.096 (0.372)	0.679 (0.216)	2.184 (0.479)	0.245 (0.056)	0.421 (0.061)	0.245 (0.056)
ML (b)	1.404 (0.487)	0.866 (0.282)	3.171 (0.656)	0.313 (0.072)	0.545 (0.082)	0.313 (0.072)
$n = 10000$						
OLS (a)	0.561 (0.196)	0.447 (0.158)	1.141 (0.217)	0.122 (0.029)	0.212 (0.029)	0.122 (0.029)
OLS (b)	0.723 (0.251)	0.581 (0.208)	1.655 (0.31)	0.157 (0.036)	0.272 (0.038)	0.157 (0.036)
ML (a)	0.488 (0.161)	0.298 (0.1)	0.967 (0.191)	0.11 (0.026)	0.188 (0.026)	0.11 (0.026)
ML (b)	0.627 (0.211)	0.38 (0.13)	1.398 (0.28)	0.141 (0.033)	0.241 (0.034)	0.141 (0.033)

Table A.6.: $1-\cos(\theta)$ for continuous Y, $y_i \sim N(0, 0.1), i = 1, \ldots, n$, where θ is the principle angle between span($\mathbf{B}'$) and span($\widehat{\mathbf{B}}'$). Results are means over 500 replications.

	STIR-1	STIR-2	STIR-3	STIR-4	STIR-5	STIR-6
$n = 200$						
OLS (a)	-9.780e-15	-9.209e-15	3.075e-02	8.936e-03	1.561e-02	1.011e-03
OLS (b)	-1.312e-14	-1.294e-14	4.681e-02	1.439e-02	2.431e-02	1.654e-03
$n = 500$						
OLS (a)	-1.065e-14	-9.504e-15	1.170e-02	3.480e-03	6.421e-03	3.898e-04
OLS (b)	-1.005e-14	-9.312e-15	1.824e-02	5.448e-03	1.015e-02	6.094e-04
$n = 2000$						
OLS (a)	-9.349e-15	-1.065e-14	3.059e-03	9.580e-04	1.584e-03	1.064e-04
OLS (b)	-6.523e-15	-1.039e-14	4.811e-03	1.501e-03	2.537e-03	1.675e-04
$n = 10000$						
OLS (a)	-3.011e-14	-1.207e-14	6.320e-04	1.921e-04	3.234e-04	2.140e-05
OLS (b)	-1.238e-14	-9.507e-15	1.002e-03	2.958e-04	5.063e-04	3.290e-05

Table A.7.: Correlation between 100 new continuous Y-observations and values predicted by linear model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR-1-4.

Reduction Scaling	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR $\widehat{\mathbf{\Sigma}}_x$	STIR $\widehat{\mathbf{\Sigma}}_x = \hat{\mathbf{\Sigma}}_p \otimes \hat{\mathbf{\Sigma}}_T$	STIR $\mathbf{I}_{Tp}$
Scenario STIR-1 a $(T, p, H, d) = (10, 3, 2, 2)$					
k=1	0.475 (0.116)	0.312 (0.147)	0.413 (0.147)	0.306 (0.148)	0.334 (0.144)
same w. gam	0.502 (0.123)	0.316 (0.149)	0.418 (0.153)	0.309 (0.149)	0.341 (0.152)
k=1=r=m w.2.ord.	0.463 (0.126)	0.288 (0.159)	0.393 (0.17)	0.282 (0.16)	0.312 (0.162)
same w. n=2000	0.557 (0.097)	0.451 (0.114)	0.546 (0.103)	0.45 (0.114)	0.462 (0.111)
k=2	0.508 (0.093)	0.454 (0.091)	0.5 (0.093)	0.453 (0.091)	0.463 (0.09)
k=3	0.504 (0.092)	0.504 (0.092)	0.504 (0.092)	0.504 (0.092)	0.504 (0.092)
Scenario STIR-1 b					
k=1	0.421 (0.127)	0.242 (0.145)	0.355 (0.151)	0.238 (0.145)	0.261 (0.148)
same w. gam	0.428 (0.131)	0.233 (0.153)	0.348 (0.157)	0.229 (0.153)	0.256 (0.15)
k=1=r=m w.2.ord.	0.371 (0.138)	0.191 (0.147)	0.3 (0.162)	0.187 (0.147)	0.213 (0.152)
same w. n=2000	0.491 (0.108)	0.361 (0.129)	0.47 (0.117)	0.358 (0.13)	0.375 (0.124)
k=2	0.469 (0.098)	0.412 (0.099)	0.461 (0.098)	0.411 (0.099)	0.422 (0.098)
k=3	0.463 (0.1)	0.463 (0.1)	0.463 (0.1)	0.463 (0.1)	0.463 (0.1)
Scenario STIR-2 a $(T, p, H, d) = (10, 3, 2, 2)$					
k=1	0.474 (0.109)	0.326 (0.147)	0.426 (0.136)	0.321 (0.148)	0.346 (0.141)
same w. gam	0.494 (0.112)	0.327 (0.151)	0.432 (0.137)	0.321 (0.153)	0.35 (0.141)
k=1=r=m w.2.ord.	0.462 (0.112)	0.3 (0.159)	0.409 (0.142)	0.294 (0.161)	0.321 (0.153)
same w. n=2000	0.527 (0.098)	0.432 (0.115)	0.523 (0.104)	0.431 (0.116)	0.442 (0.114)
k=2	0.498 (0.093)	0.45 (0.108)	0.496 (0.093)	0.45 (0.108)	0.46 (0.104)
k=3	0.492 (0.094)	0.492 (0.094)	0.492 (0.094)	0.492 (0.094)	0.492 (0.094)
Scenario STIR-2 b					
k=1	0.427 (0.126)	0.26 (0.161)	0.378 (0.151)	0.256 (0.162)	0.28 (0.158)
same w. gam	0.428 (0.13)	0.251 (0.155)	0.373 (0.152)	0.246 (0.156)	0.274 (0.156)
k=1=r=m w.2.ord.	0.378 (0.123)	0.2 (0.149)	0.324 (0.146)	0.195 (0.15)	0.219 (0.148)
same w. n=2000	0.475 (0.1)	0.365 (0.12)	0.469 (0.102)	0.364 (0.12)	0.377 (0.119)
k=2	0.465 (0.098)	0.398 (0.113)	0.461 (0.098)	0.397 (0.114)	0.411 (0.108)
k=3	0.454 (0.097)	0.454 (0.097)	0.454 (0.097)	0.454 (0.097)	0.454 (0.097)
Scenario STIR-3 a $(T, p, H, d) = (3, 15, 3, 1)$					
k=1	0.166 (0.135)	0.059 (0.112)	0.099 (0.135)	0.056 (0.111)	0.117 (0.14)
same w. gam	0.194 (0.163)	0.06 (0.116)	0.099 (0.141)	0.057 (0.115)	0.115 (0.138)
k=1=r=m w.2.ord.	0.19 (0.166)	0.06 (0.117)	0.106 (0.146)	0.058 (0.116)	0.127 (0.162)
same w. n=2000	0.306 (0.141)	0.161 (0.108)	0.283 (0.142)	0.159 (0.108)	0.283 (0.134)
k=2	0.239 (0.128)	0.128 (0.107)	0.236 (0.126)	0.128 (0.107)	0.221 (0.127)
k=3	0.206 (0.131)	0.096 (0.107)	0.205 (0.13)	0.096 (0.107)	0.18 (0.122)
Scenario STIR-3 b					
k=1	0.101 (0.136)	0.023 (0.111)	0.046 (0.117)	0.022 (0.11)	0.056 (0.122)
same w. gam	0.105 (0.151)	0.028 (0.11)	0.045 (0.128)	0.027 (0.109)	0.045 (0.118)
k=1=r=m w.2.ord.	0.097 (0.133)	0.022 (0.11)	0.042 (0.12)	0.022 (0.109)	0.042 (0.129)
same w. n=2000	0.24 (0.151)	0.078 (0.117)	0.183 (0.157)	0.075 (0.116)	0.194 (0.146)
k=2	0.218 (0.126)	0.105 (0.108)	0.215 (0.125)	0.105 (0.108)	0.199 (0.123)
k=3	0.179 (0.131)	0.073 (0.112)	0.18 (0.13)	0.073 (0.112)	0.151 (0.125)
Scenario STIR-4 a $(T, p, H, d) = (2, 7, 2, 1)$					
k=1	0.412 (0.124)	0.024 (0.113)	0.272 (0.137)	0.024 (0.112)	0.295 (0.133)
same w. gam	0.457 (0.128)	0.024 (0.133)	0.294 (0.16)	0.023 (0.132)	0.295 (0.131)
k=1=r=m w.2.ord.	0.457 (0.128)	0.03 (0.127)	0.309 (0.162)	0.029 (0.126)	0.335 (0.151)
same w. n=2000	0.445 (0.131)	0.032 (0.142)	0.299 (0.172)	0.031 (0.142)	0.323 (0.166)
k=2	0.668 (0.056)	0.398 (0.081)	0.666 (0.056)	0.398 (0.081)	0.62 (0.065)
Scenario STIR-4 b					
k=1	0.362 (0.123)	0.019 (0.118)	0.247 (0.131)	0.018 (0.118)	0.25 (0.126)
same w. gam	0.387 (0.138)	0.035 (0.14)	0.256 (0.161)	0.033 (0.139)	0.242 (0.131)
k=1=r=m w.2.ord.	0.389 (0.137)	0.02 (0.124)	0.266 (0.161)	0.019 (0.124)	0.264 (0.159)
same w. n=2000	0.401 (0.133)	0.041 (0.128)	0.283 (0.158)	0.041 (0.128)	0.288 (0.154)
k=2	0.625 (0.062)	0.351 (0.099)	0.622 (0.063)	0.351 (0.099)	0.572 (0.07)

Table A.8.: Correlation between 100 new continuous Y-observations and values predicted by linear model fit. Results are means and standard deviations (in parentheses) over 100 replications for each setting. Settings STIR-5,6 and LSIR-1,2

Reduction Scaling	STIR $\widehat{\mathbf{\Delta}}$	STIR $\widehat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}_p \otimes \hat{\mathbf{\Delta}}_T$	STIR $\widehat{\mathbf{\Sigma}}_x$	STIR $\widehat{\mathbf{\Sigma}}_x = \hat{\mathbf{\Sigma}}_p \otimes \hat{\mathbf{\Sigma}}_T$	STIR $\mathbf{I}_{Tp}$
Scenario STIR-5 a $(T, p, H, d) = (2, 21, 2, 1)$					
k=1	0.489 (0.136)	0.073 (0.152)	0.228 (0.182)	0.072 (0.149)	0.414 (0.152)
same w. gam	0.507 (0.143)	0.065 (0.211)	0.211 (0.204)	0.063 (0.207)	0.411 (0.146)
k=1 w.2.ord.	0.505 (0.137)	0.075 (0.188)	0.23 (0.204)	0.071 (0.186)	0.434 (0.154)
same w. n=2000	0.515 (0.147)	0.091 (0.182)	0.252 (0.208)	0.091 (0.182)	0.445 (0.168)
k=2	0.824 (0.033)	0.588 (0.068)	0.809 (0.037)	0.588 (0.068)	0.8 (0.036)
Scenario STIR-5 b					
k=1	0.465 (0.142)	0.064 (0.159)	0.235 (0.18)	0.061 (0.159)	0.389 (0.154)
same w. gam	0.462 (0.163)	0.052 (0.198)	0.208 (0.193)	0.048 (0.195)	0.383 (0.147)
k=1 w.2.ord.	0.475 (0.148)	0.062 (0.186)	0.224 (0.203)	0.059 (0.185)	0.402 (0.16)
same w. n=2000	0.494 (0.144)	0.07 (0.17)	0.257 (0.196)	0.07 (0.169)	0.419 (0.164)
k=2	0.799 (0.038)	0.538 (0.071)	0.782 (0.044)	0.538 (0.071)	0.769 (0.045)
Scenario STIR-6 a $(T, p, H, d) = (2, 7, 2, 1)$					
k=1	0.586 (0.128)	0.056 (0.171)	0.157 (0.185)	0.046 (0.168)	0.44 (0.153)
same w. gam	0.589 (0.13)	0.043 (0.227)	0.122 (0.199)	0.043 (0.222)	0.44 (0.152)
k=1 w.2.ord.	0.589 (0.129)	0.053 (0.203)	0.133 (0.199)	0.044 (0.196)	0.442 (0.161)
same w. n=2000	0.57 (0.13)	0.063 (0.22)	0.141 (0.231)	0.062 (0.221)	0.433 (0.172)
k=2	0.924 (0.017)	0.795 (0.035)	0.92 (0.017)	0.795 (0.035)	0.906 (0.02)
Scenario STIR-6 b					
k=1	0.565 (0.129)	0.042 (0.161)	0.155 (0.176)	0.036 (0.157)	0.422 (0.151)
same w. gam	0.565 (0.132)	0.041 (0.205)	0.137 (0.195)	0.029 (0.203)	0.418 (0.156)
k=1 w.2.ord.	0.567 (0.129)	0.054 (0.187)	0.138 (0.2)	0.043 (0.181)	0.42 (0.165)
same w. n=2000	0.56 (0.132)	0.067 (0.206)	0.158 (0.22)	0.067 (0.206)	0.423 (0.167)
k=2	0.91 (0.019)	0.756 (0.042)	0.904 (0.021)	0.756 (0.042)	0.889 (0.023)
Scenario LSIR-1 a $(T, p, H, d) = (3, 8, 3, 2)$					
k=1	0.97 (0.007)	0.971 (0.006)	0.948 (0.019)	0.971 (0.006)	0.97 (0.006)
same w. gam	0.973 (0.006)	0.974 (0.006)	0.948 (0.02)	0.971 (0.006)	0.973 (0.006)
k=1 w.2.ord.	0.97 (0.007)	0.971 (0.006)	0.948 (0.02)	0.971 (0.006)	0.97 (0.006)
same w. n=2000	0.973 (0.007)	0.973 (0.007)	0.969 (0.007)	0.973 (0.007)	0.971 (0.007)
k=2	0.971 (0.006)	0.972 (0.006)	0.96 (0.009)	0.972 (0.006)	0.971 (0.007)
k=3	0.971 (0.006)	0.972 (0.006)	0.962 (0.008)	0.972 (0.006)	0.971 (0.006)
Scenario LSIR-1 b					
k=1	0.965 (0.008)	0.967 (0.008)	0.932 (0.032)	0.966 (0.008)	0.965 (0.008)
same w. gam	0.966 (0.008)	0.968 (0.008)	0.932 (0.032)	0.966 (0.008)	0.966 (0.008)
k=1 w.2.ord.	0.965 (0.008)	0.966 (0.008)	0.931 (0.032)	0.966 (0.008)	0.965 (0.008)
same w. n=2000	0.97 (0.008)	0.97 (0.007)	0.965 (0.009)	0.97 (0.008)	0.968 (0.007)
k=2	0.966 (0.008)	0.967 (0.007)	0.954 (0.01)	0.967 (0.007)	0.966 (0.008)
k=3	0.967 (0.008)	0.967 (0.007)	0.957 (0.009)	0.967 (0.007)	0.966 (0.008)
Scenario LSIR-2 a $(T, p, H, d) = (3, 8, 3, 2)$					
k=1	0.268 (0.162)	0.696 (0.273)	0.204 (0.17)	0.689 (0.274)	0.812 (0.226)
same w. gam	0.262 (0.166)	0.695 (0.274)	0.2 (0.17)	0.688 (0.276)	0.81 (0.232)
k=1 w.2.ord.	0.248 (0.167)	0.689 (0.281)	0.183 (0.175)	0.682 (0.283)	0.805 (0.239)
same w. n=2000	0.497 (0.233)	0.898 (0.132)	0.464 (0.219)	0.895 (0.135)	0.939 (0.096)
k=2	0.732 (0.096)	0.961 (0.011)	0.684 (0.115)	0.961 (0.011)	0.97 (0.007)
k=3	0.773 (0.092)	0.964 (0.01)	0.736 (0.11)	0.964 (0.01)	0.971 (0.007)
Scenario LSIR-2 b					
k=1	0.223 (0.15)	0.62 (0.248)	0.171 (0.153)	0.613 (0.249)	0.726 (0.243)
same w. gam	0.217 (0.152)	0.616 (0.253)	0.163 (0.157)	0.608 (0.255)	0.724 (0.248)
k=1 w.2.ord.	0.202 (0.153)	0.61 (0.258)	0.146 (0.163)	0.602 (0.261)	0.718 (0.255)
same w. n=2000	0.343 (0.221)	0.784 (0.217)	0.315 (0.205)	0.779 (0.218)	0.856 (0.184)
k=2	0.7 (0.114)	0.952 (0.015)	0.647 (0.129)	0.952 (0.015)	0.962 (0.01)
k=3	0.75 (0.105)	0.957 (0.013)	0.708 (0.125)	0.957 (0.013)	0.965 (0.009)

Bibliography

[1] A. Brenner, P. Inskip, J. Rusiecki, C. Rabkin, J. Engels, and R. Pfeiffer. Serially measured pre-diagnostic levels of serum cytokines and risk of brain cancer in active component military personnel. *British Journal of Cancer*, 119(7):893–900, 2018.

[2] E. Bura. Lecture notes in introduction to sufficient dimension reduction, May 2020.

[3] E. Bura and L. Forzani. Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, 110(509):420–434, 2015.

[4] E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142, 2011.

[5] D. R. Cook. *Regression Graphics: Ideas for studying regressions through graphics*. Wiley, New York, 1998.

[6] R. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4), Nov 2008.

[7] R. Cook and S. Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[8] M. Eaton. *Multivariate Statistics: A Vector Space Approach*. Probability and Statistics Series. Wiley, 1983. `https://books.google.at/books?id=1CvvAAAAMAAJ`.

[9] M. L. Eaton. *Multivariate Statistics: A Vector Space Approach*, volume 53 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007. `https://projecteuclid.org/euclid.lnms/1196285102`.

[10] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

[11] N. E. Helwig. Lecture notes in multivariate linear regression. `http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf`, Jan. 2017.

[12] B. Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics*, 42(6):1977–1991, 1971.

[13] I. C. F. Ipsen and C. D. Meyer. The angle between complementary subspaces. *The American Mathematical Monthly*, 102(10):904–911, 1995.

[14] A. J. Izenman. *Modern Multivariate Statistical Techniques*. Springer-Verlag New York, 2008.

[15] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[16] K. B. Petersen and M. S. Pedersen. The matrix cookbook. `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`, Nov. 2012. Version 20121115.

[17] R. M. Pfeiffer, L. Forzani, and E. Bura. Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine*, 31(22):2414–2427, 2012.

[18] R. M. Pfeiffer, D. B. Kapla, and E. Bura. Least squares and maximum likelihood estimation of sufficient reductions in regressions with matrix-valued predictors. *International Journal of Data Science and Analytics*, pages 1–16, 2020.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[20] Y. Shao, R. Cook, and S. Weisberg. Marginal tests with sliced average variance estimation. *Biometrika*, 94(2):285–296, 2007.

[21] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005.

[22] M. Song, E. Bura, R. Parzer, and R. M. Pfeiffer. Structured time-dependent inverse regression (stir). unpublished, 2021.

[23] S. Weisberg. Dimension reduction regression in r. *Journal of Statistical Software*, 7(1):1–22, 2002.

[24] S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.

[25] P. Zeng and Y. Zhu. An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101:271–290, Jan. 2010.