Journal of Machine Learning Research 23 (2022) 1-46 Submitted 2/21; Revised 12/21; Published 2/22

Sufficient reductions in regression with mixed predictors

Efstathia Bura (Corresponding author) EFSTATHIA.BURA@TUWIEN.AC.AT
Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
TU Wien
Vienna, 1040, Austria

Liliana Forzani LILIANA.FORZANI@GMAIL.COM
Facultad de Ingeniería Química
Universidad Nacional del Litoral
Researcher of CONICET
Santa Fe, Argentina

Rodrigo García Arancibia R.GARCIA.ARANCIBIA@GMAIL.COM
Instituto de Economía Aplicada Litoral-FCE-UNL
Universidad Nacional del Litoral
Researcher of CONICET
Santa Fe, Argentina

Pamela Llop LLOPPAMELA@GMAIL.COM
Facultad de Ingeniería Química
Universidad Nacional del Litoral
Researcher of CONICET
Santa Fe, Argentina

Diego Tomassi DIEGOTOMASSI@GMAIL.COM
Facultad de Ingeniería Química
Universidad Nacional del Litoral
Researcher of CONICET
Santa Fe, Argentina

Editor: Kenji Fukumizu

Abstract

Most data sets comprise of measurements on continuous and categorical variables. Yet, modeling high-dimensional mixed predictors has received limited attention in the regression and classification statistical literature. We study the general regression problem of inferring on a variable of interest based on high dimensional mixed continuous and binary predictors. The aim is to find a lower dimensional function of the mixed predictor vector that contains all the modeling information in the mixed predictors for the response, which can be either continuous or categorical. The approach we propose identifies sufficient reductions by reversing the regression and modeling the mixed predictors conditional on the response. We derive the maximum likelihood estimator of the sufficient reductions, asymptotic tests for dimension, and a regularized estimator, which simultaneously achieves variable (feature) selection and dimension reduction (feature extraction). We study the performance of the proposed method and compare it with other approaches through simulations and real data examples.

Keywords: High-dimensional, Multivariate Bernoulli, Regularization, Feature selection, Feature extraction

1. Introduction

Most data sets comprise of measurements on a mixture of categorical and continuous features. Examples abound in the biomedical and health sciences, neuro-imaging, genomics, finance, social media, and internet advertising. Genome-wide association studies GWAS are widely used in human genetics research to identify genes associated with complex diseases and in agricultural research to identify genes associated with quantitative traits such as yield and productivity (Dahl et al., 2016; Huang et al., 2010; Bermingham et al., 2015). In GWAS, single nucleotide polymorphism (SNPs) are genotyped for different groups of subjects and mixed linear model methodologies, where SNPs effects are modeled as random effects, are mainstream in genome-wide association studies (GWAS) (Zhang et al., 2010, 2021, e.g.), even though, a characterization of the biological mechanism for most quantitative traits remains elusive Dahl et al. (2016). In marketing, mixed media modeling is used to estimate the impact of various tactics on sales in order to forecast and come up with a better marketing strategy. Facebook is collecting data on *media mix*; i.e., factors that may have influence over sales, both continuous and discrete, and is using them to quantify the weight for each factor to create a model to predict marketing results for future strategy. In neuro-imaging, brain network analysis searches for insight into links between system-level properties and health outcomes. For example, Simpson et al. (2019) analyze the effects of multiple variables of interest and topological network features, both a mix of continuous and categorical variables, on the overall network structure in a multivariate modeling framework.

The first statistical approach to modeling the dependence structure of mixed data we found in the literature is the *location model* of Olkin and Tate (1961). The location model uses correlation as a measure of dependence and bypasses the mixed nature of the data by grouping the continuous variables using the categorical ones and requiring they be normally distributed with different means but same variance within the groups.

More recently, Markov Networks, or undirected graphical models, that encode pairwise conditional dependence relationships among random variables have been used to model multivariate mixed data. With few exceptions (Yang et al., 2014a,b, 2015; Chen et al., 2014), mixed continuous and categorical data are modeled with the Gaussian Graphical Model (GGM) in a manner similar to the location model. Binary variables are used to define the different categories and GGM requires the continuous variables be conditionally normal and pairwise conditionally independent within categories. References for GMMs for low-dimensional mixed data include Lauritzen and Wermuth (1989), Lauritzen (1996), Yuan and Lin (2007), Martin and Michael (2008), and in the high-dimensional setting, Cheng et al. (2014, 2017) and Lee and J.H. (2015). In particular, Cheng et al. (2017) proposed a simplified version of the conditional Gaussian distribution that reduces the number of parameters significantly while maintaining flexibility.

Both, GGM and the location model are unsupervised approaches for mixed data that do not include an output of interest. In the case of a categorical output, approaches for the treatment of mixed, in particular, binary and continuous input variables, include methods

based on nonparametric density estimation (Aitchison and Aitken, 1976), the use of logistic discrimination (Day and Kerridge, 1967), in which the probability of group membership is assumed to be a logistic function of the observed variates (Anderson, 1972, 1975), and a likelihood ratio classification rule (Krzanowski, 1975) based on the location model of Olkin and Tate (1961). Krzanowski (1993) surveys and summarizes the associated developments. More recently, the location model has been used in multiple imputation [see, e.g., Javaras and Van Dyk (2003), Van Buuren (2018, Ch. 4, Sec. 4.4)].

In this paper we study the general regression and classification problem with high-dimensional mixed predictors. Specifically, we consider the conditional distribution of

$$Y \mid (\mathbf{X}, \mathbf{H}), \tag{1}$$

where the response Y is either continuous or categorical, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is a vector of p continuous, and $\mathbf{H} = (H_1, H_2, \ldots, H_q)^T$ is a vector of q binary predictor variables. Our aim is to find a lower dimensional function of the mixed predictor vector $\mathbf{Z} = (\mathbf{X}^T, \mathbf{H}^T)^T$ that encapsulates *all* information the mixed predictors contain for the response Y. Specifically, our target is the identification of a function, other than the identity, $\mathbf{R} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ such that $F(Y \mid \mathbf{Z}) = F(Y \mid \mathbf{R}(\mathbf{Z}))$, where $F(\cdot|\cdot)$ denotes the conditional cumulative distribution function of the response given the predictors. Such a function $\mathbf{R}$ is called a *sufficient reduction* of the regression of Y on $\mathbf{Z}$.

This seemingly ambitious goal turns out to be surprisingly simple using the inventive tool of *inverse regression*. When Y and $\mathbf{Z}$ are both random, inverse regression is based on the equivalence of the following two statements [see Cook (2007), Bura et al. (2016), Bura and Forzani (2015)],

(i) $Y \mid \mathbf{Z} \overset{d}{=} Y \mid \mathbf{R}(\mathbf{Z})$

(ii) $\mathbf{Z} \mid (Y, \mathbf{R}(\mathbf{Z})) \overset{d}{=} \mathbf{Z} \mid \mathbf{R}(\mathbf{Z})$

where $\overset{d}{=}$ signifies equal in distribution. Statement (i) is an alternative definition of a sufficient reduction for the forward regression in (1) and (ii) is the usual definition of a sufficient *statistic* for a *parameter* Y indexing the distribution of the mixed $\mathbf{Z}$. The equivalence of (i) and (ii) obtains that if one considers Y as a *parameter*, the sufficient "statistic" for Y *is the sufficient reduction* for the regression of Y on $\mathbf{Z}$. In consequence, in order to find a sufficient reduction for *the forward regression of* Y *on* $\mathbf{Z}$ in (1), we can equivalently solve the *inverse problem* of finding a sufficient statistic for the regression of $\mathbf{Z}$ on Y.

In particular, if $\mathbf{Z} \mid Y$ is modeled with a distribution that allows the specification of a sufficient "statistic" for Y, then the reduction is exhaustive and *minimal* sufficient. This branch of sufficient dimension reduction is called *model-based* in contrast to *moment-based* approaches, such as sliced inverse regression (SIR, Li (1991)), sliced average variance estimation (SAVE, Cook and Weisberg (1991)), directional regression (DR, Li and Wang (2007)), among several others, that are based on inverse moments of $\mathbf{Z} \mid Y$ and typically obtain partial sufficient reductions.

Our approach exploits the factorization

$$F(\mathbf{X}, \mathbf{H} \mid Y) = F(\mathbf{X} \mid Y, \mathbf{H}) F(\mathbf{H} \mid Y), \tag{2}$$

which allows us to treat the continuous and binary predictors separately, while at the same time we account for their interdependence in their relationship with Y in Section 2. An advantageous aspect of (2) is that it provides an easier way to model and parametrize mixed data, since binary and continuous data can be modeled separately.

In Section 2.1 we model $\mathbf{X} \mid (Y, \mathbf{H})$ as multivariate normal and $\mathbf{H} \mid Y$ as multivariate Bernoulli in Section 2.2, in analogy to the Gaussian graphical model and the location model in unsupervised multivariate analysis of mixed data. We show that the resulting distribution (2) belongs to the exponential family, and derive sufficient reductions for the regression $Y \mid (\mathbf{X}, \mathbf{H})$ from the two separate regressions, $\mathbf{X} \mid (Y, \mathbf{H})$ and $\mathbf{H} \mid Y$ in Section 3. We compute the maximum likelihood estimator of the sufficient reduction in Sections 4.1 and 4.2, its asymptotic distribution in Section 4.4, and an asymptotic test for the dimension of the sufficient reduction in Section 4.5. We complete our treatment with a method for *simultaneous* sufficient dimension reduction and variable selection in Section 4.3.

Section 5 contains an extensive simulation study that demonstrates the competitive performance of our approach. Furthermore, we show the superior performance of our methods as compared with generalized linear models and a version of principal component regression that allows for mixed predictors in the analysis of three data sets in Section 6.

Even though our focus in this paper is the regression of the usually univariate Y on the mixed $\mathbf{Z}$ vector, our development results in a new multivariate regression method for a *mixed continuous and binary response*, on which we comment further as we conclude in Section 7.

2. The Model

Our approach combines the two types of graphical models, the multivariate Gaussian model for continuous data and the Ising model (Ising, 1925) for binary data, conditional on the appropriate arguments. By conditioning, we transform the graphical models into separate regressions of the multivariate continuous $\mathbf{X}$ on $(Y, \mathbf{H})$ and the multivariate binary $\mathbf{H}$ on Y and use the factorization in (2) to regress $(\mathbf{X}, \mathbf{H})$ on Y.

We start by specifying the notation we use throughout. The vec operator converts its matrix argument into a column vector. More precisely, if $\mathbf{G}$ is an $m \times n$ matrix then $\text{vec}(\mathbf{G})$ is an $mn \times 1$ vector obtained by stacking the columns of $\mathbf{G}$. The unvec operator is such that $\text{unvec}(\text{vec}(\mathbf{G})) = \mathbf{G}$. We let $k_q = q(q-1)/2$ and $m_p = p(p+1)/2$. The vech operator converts the lower half of a matrix including the main diagonal to a vector. That is, if $\mathbf{G}$ is a square $q \times q$ matrix then $\text{vech}(\mathbf{G})$ is a $m_q \times 1$ vector obtained by stacking the columns of the lower triangular part of $\mathbf{G}$ including the diagonal. There is a unique $\mathbf{D}_q \in \mathbb{R}^{q^2 \times q(q+1)/2}$ and $\mathbf{C}_q \in \mathbb{R}^{q(q+1)/2 \times q^2}$, such that $\text{vec}(\mathbf{G}) = \mathbf{D}_q \,\text{vech}(\mathbf{G})$ and $\text{vech}(\mathbf{G}) = \mathbf{C}_q \text{vec}(\mathbf{G})$ for any $\mathbf{G}$ symmetric $q \times q$ matrix.

The matrix $\mathbf{L}_q \in \mathbb{R}^{q \times q(q+1)/2}$ has entries 1 and 0, so that $\mathbf{L}_q \mathbf{C}_q$ is equal to $\mathbf{C}_q$except for replacing the value $1/2$ by zero throughout. The matrix $\mathbf{J}_q \in \mathbb{R}^{k_q \times q(q+1)/2}$ has entries 1 and 0, so that $\mathbf{J}_q \mathbf{C}_q$ is equal to $\mathbf{C}_q$ except for replacing the ones with zeros. A projection onto the columns of $\mathbf{b}$ is denoted by $\mathbf{P}_\mathbf{b}$ and the projection onto the orthogonal complement of $\mathbf{b}$ as $\mathbf{Q}_\mathbf{b}$.

2.1 The distribution of $\mathbf{X} \mid (\mathbf{H}, Y)$

We let the p-dimensional vector of continuous random variables $\mathbf{X}|(\mathbf{H}, Y)$ be multivariate normal with

$$\mathbf{X} \mid (\mathbf{H}, Y) \sim N\left(\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}), \boldsymbol{\Delta}\right), \tag{3}$$

where $\boldsymbol{\mu}_{\mathbf{X}} = \mathrm{E}_{\mathbf{X}}(\mathbf{X})$, $\boldsymbol{\mu}_{\mathbf{H}} = \mathrm{E}_{\mathbf{H}}(\mathbf{H})$, $\mathbf{f}_Y : \mathbb{R} \to \mathbb{R}^r$ is a known function of Y, $\bar{\mathbf{f}}_Y = \mathrm{E}_Y(\mathbf{f}_Y)$, $\mathbf{A} : p \times r$, and $\boldsymbol{\beta} : p \times q$, are unconstrained parameter matrices, and $\boldsymbol{\Delta}$ is a $p \times p$ positive definite covariance matrix. For example, if the response is continuous, $\mathbf{f}_Y$ can be a vector of polynomials of order r, or, in order to avoid multicollinearity, of a set of r orthonormal basis functions. If the response is categorical with values in one of h categories C_k, $k = 1, \ldots, h$, we set $r = h - 1$ and let the k-th element of $\mathbf{f}_Y$ to be $I(Y \in C_k)$, where I is the indicator function. To simplify notation, henceforth $\mathbf{f}_Y$ will signify the centered $\mathbf{f}_Y - \bar{\mathbf{f}}_Y$.

The probability density function of $\mathbf{X} \mid (\mathbf{H}, Y)$ in model (3) is

$$f(\mathbf{X} \mid \mathbf{H}, Y = y) = \frac{1}{\sqrt{2\pi}\sqrt{|\boldsymbol{\Delta}|}} \exp\Bigg\{ -\frac{1}{2}\Big((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})\Big)^T \boldsymbol{\Delta}^{-1}\Big((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})\Big)\Bigg\}. \tag{4}$$

Modelling the conditional distribution of the continuous predictors via (3) is a variation of the model used in *principal fitted components* (PFC) by Cook and Forzani (2008) by adding the multivariate categorical component $\mathbf{H}$ to the response Y. PFC is the first model-based SDR method that appeared in the literature. It leveraged the parametric inverse regression approach in Bura and Cook (2001) by adding the normality assumption to the linear model for the conditional mean in (3) and thereby obtaining the minimal sufficient reduction for the regression of Y on the continuous $\mathbf{X}$ and its maximum likelihood estimator. Specifically, PFC assumes model (3) with no categorical variables, so that $\mathrm{E}(\mathbf{X} \mid Y) = \boldsymbol{\mu}_{\mathbf{X}} + \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)$ and constant variance-covariance matrix $\boldsymbol{\Delta}$. Here we further build upon this formulation device by conditioning on the categorical predictors as well, and including them as binary regressors in the linear model for $\mathrm{E}(\mathbf{X} \mid \mathbf{H}, Y)$. Despite the difficulties this addition introduces, as will become evident in Section 3, the idea is similar and simple: use a linear model and all associated well studied tools to obtain statistically efficient and easily computable minimal sufficient reductions.

2.2 The distribution of $\mathbf{H} \mid Y$

The joint distribution of a random vector, whose elements are binary random variables, is modelled with the multivariate Bernoulli distribution [see Whittaker (2009); Dai (2012); Dai et al. (2013)]. Its probability mass function involves terms representing third and higher order moments of the random variables. The Ising model (Ising, 1925) is a simplified version of the multivariate Bernoulli distribution that includes up to second order interactions among the binary variables and is frequently used to alleviate the complexity of modeling. In the Ising model, the network structure can be identified from the coefficients of the interaction terms in the probability mass function.

Cheng et al. (2014) proposed a model for the conditional distribution of binary network data given covariates, which naturally incorporates covariate information into the Ising

model, allowing the strength of the connection to depend on the covariates. We adapt their model to regress the multivariate binary $\mathbf{H}$ on Y.

Let $\mathcal{H}$ = all possible combinations of $\mathbf{H} \in \{0,1\}^q$, $\mathbf{H}_{-j} = (H_1, \ldots, H_{j-1}, H_{j+1}, \ldots, H_q)$, $\mathbf{H}_{-i,-j} = (H_1, \ldots, H_{i-1}, H_{i+1}, \ldots, H_{j-1}, H_{j+1}, \ldots, H_q)$, $i, j = 1, \ldots, q$. The joint probability mass function of the q-dimensional vector of binary variables $\mathbf{H}$ conditional on Y is [see Cheng et al. (2014)]

$$\Pr(\mathbf{H} \mid Y = y) = \frac{1}{G(\mathbf{\Gamma}_y)} \exp\left\{ \operatorname{vech}^T(\mathbf{H}\mathbf{H}^T) \operatorname{vech}(\mathbf{\Gamma}_y) \right\}, \tag{5}$$

where $G(\mathbf{\Gamma}_y) = \sum_{\mathbf{H}\in\mathcal{H}} \exp\left(\operatorname{vech}^T(\mathbf{H}\mathbf{H}^T) \operatorname{vech}(\mathbf{\Gamma}_y)\right)$, and $\mathbf{\Gamma}_y = (\gamma_{ij}^y)$ is a $q \times q$ symmetric matrix with elements

$$\gamma_{jj}^y = \log\left(\frac{\Pr(H_j = 1 \mid \mathbf{H}_{-j} = 0, y)}{1 - \Pr(H_j = 1 | \mathbf{H}_{-j} = 0, y)} \right),$$

$$\gamma_{ij}^y = \log \frac{\Pr(H_i = 1, H_j = 1 \mid \mathbf{H}_{-i,-j} = 0, y) \Pr(H_i = 0, H_j = 0 \mid \mathbf{H}_{-i,-j} = 0, y)}{\Pr(H_i = 1, H_j = 0 \mid \mathbf{H}_{-i,-j} = 0, y) \Pr(H_i = 0, H_j = 1 \mid \mathbf{H}_{-i,-j} = 0, y)},$$

for $i \neq j$. A linear model with independent variables $\mathbf{f}_Y \in \mathbb{R}^r$ is a natural choice for each γ_{ij}^y,

$$\gamma_{ij}^y = \tau_{ij,0}^* + \boldsymbol{\tau}_{ij}^T \mathbf{f}_Y, \qquad i, j = 1, \ldots, q, \tag{6}$$

where $\boldsymbol{\tau}_{ij}^T = (\tau_{ij,1}, \ldots, \tau_{ij,r})$ is a vector of parameters independent of Y, and $\tau_{ij,0}^*$ is the intercept for each (i, j). The vector $\mathbf{f}_Y$ of functions in Y plays the role of covariates in the multivariate binary regression model formulation of Cheng et al. (2014). Here, $\mathbf{f}_Y$ is centered, and can be different from that in (4), even though, as will be seen later, choosing the same $\mathbf{f}_Y$ simplifies the formula for the joint distribution in (8) as well as the derivation of a sufficient reduction for the regression of Y on $(\mathbf{X}, \mathbf{H})$.

Next, we define the $q \times q$ matrices, $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_k$, $k = 1, \ldots, r$, as $[\boldsymbol{\tau}_0^*]_{ij} = \tau_{ij,0}^*$ and $[\boldsymbol{\tau}_k]_{ij} = \tau_{ij,k}$ with $i, j = 1, \ldots, q$ and $k = 1, \ldots, r$. We let $\boldsymbol{\tau}_0 = \operatorname{vech}(\boldsymbol{\tau}_0^*)$, a $q(q+1)/2$ vector, and $\boldsymbol{\tau} = (\operatorname{vech}(\boldsymbol{\tau}_1), \ldots, \operatorname{vech}(\boldsymbol{\tau}_r))$, a $q(q+1)/2 \times r$ matrix, so that the $q(q+1)/2$ vector $\operatorname{vech}(\mathbf{\Gamma}_y)$ is

$$\operatorname{vech}(\mathbf{\Gamma}_y) = \boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y.$$

Under (6), the probability mass function of $\mathbf{H} \mid Y$ in (5) is

$$\Pr(\mathbf{H} \mid Y = y) = \frac{1}{G(\mathbf{\Gamma}_y)} \exp\left\{ \operatorname{vech}^T(\mathbf{H}\mathbf{H}^T)(\boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y) \right\}, \tag{7}$$

with $G(\mathbf{\Gamma}_y) = \sum_{\mathbf{H}\in\mathcal{H}} \exp\left(\operatorname{vech}^T(\mathbf{H}\mathbf{H}^T)(\boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y)\right)$. Under (7) and (4), the joint distribution of the inverse regression $(\mathbf{X}, \mathbf{H} \mid Y)$ has probability density function

$$\begin{aligned} f(\mathbf{X}, \mathbf{H} \mid Y = y) &= f(\mathbf{X} \mid \mathbf{H}, Y = y) f(\mathbf{H} \mid Y = y) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{|\mathbf{\Delta}|}} \exp\Bigg\{ -\frac{1}{2}\Big((\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}) \Big)^T \\ &\qquad \mathbf{\Delta}^{-1}\Big((\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}) \Big) \Bigg\} \end{aligned}$$

$$\times \frac{1}{G(\mathbf{\Gamma}_y)} \exp\left\{ \text{vech}^T(\mathbf{H}\mathbf{H}^T) \left(\boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y\right) \right\}. \tag{8}$$

Our regression model for the mixed vector $\mathbf{Z}$ is similar to the regression model of Fitzmaurice and Laird (1997) with the difference that we do not allow $\boldsymbol{\mu}_{\mathbf{H}}$ to vary with Y in (4). This results in different maximum likelihood estimates for the parameters in (8) in Section 4.1.

3. Sufficient Reductions

We focus on the regression problem (1), where we aim to identify a reduction $\mathbf{R}(\mathbf{Z})$ such that $Y \mid \mathbf{Z} \overset{d}{=} Y \mid \mathbf{R}(\mathbf{Z})$. Since the latter is equivalent to $\mathbf{Z} \mid (Y, \mathbf{R}(\mathbf{Z})) \overset{d}{=} \mathbf{Z} \mid \mathbf{R}(\mathbf{Z})$, as discussed in the introduction, we derive the sufficient reduction $\mathbf{R}(\mathbf{Z})$ using (2).

Of central importance to our development is showing that the density of $(\mathbf{X}, \mathbf{H}) \mid Y$ in (8) belongs to the exponential family of distributions. In Appendix A.1, we express (8) as

$$f(\mathbf{X}, \mathbf{H} \mid Y = y) = h(\mathbf{X}, \mathbf{H}) \exp\left(\mathbf{T}^T(\mathbf{X}, \mathbf{H})\boldsymbol{\eta}_y - \psi(\boldsymbol{\eta}_y)\right), \tag{9}$$

which belongs to the natural exponential family of distributions [see, e.g., Morris (2006)]. In (9), $h(\mathbf{X}, \mathbf{H}) = (2\pi)^{-1/2}$, the sufficient statistic is

$$\mathbf{T}(\mathbf{X}, \mathbf{H}) = \begin{pmatrix} \mathbf{X} \\ \mathbf{H} \\ -\frac{1}{2}\mathbf{D}_p^T\mathbf{D}_p \,\text{vech}(\mathbf{X}\mathbf{X}^T) \\ \text{vec}(\mathbf{X}\mathbf{H}^T) \\ \mathbf{J}_q \,\text{vech}(\mathbf{H}\mathbf{H}^T) \end{pmatrix}, \tag{10}$$

the natural parameters are

$$\begin{aligned} \boldsymbol{\eta}_y = \begin{pmatrix} \boldsymbol{\eta}_{y1} \\ \boldsymbol{\eta}_{y2} \\ \boldsymbol{\eta}_{y3} \\ \boldsymbol{\eta}_{y4} \\ \boldsymbol{\eta}_{y5} \end{pmatrix} &= \begin{pmatrix} \mathbf{I}_p & \mathbf{f}_y^T \otimes \mathbf{I}_p & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I}_q & \mathbf{f}_y^T \otimes \mathbf{I}_q & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I}_{m_p} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{I}_{pq} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{I}_{k_q} & \mathbf{f}_y^T \otimes \mathbf{I}_{k_q} \end{pmatrix} \begin{pmatrix} \boldsymbol{\vartheta}_1 \\ \boldsymbol{\vartheta}_2 \\ \boldsymbol{\vartheta}_3 \\ \boldsymbol{\vartheta}_4 \\ \boldsymbol{\vartheta}_5 \end{pmatrix} \\ &:= \mathbf{F}_y\boldsymbol{\vartheta}, \end{aligned} \tag{11}$$

with $\boldsymbol{\vartheta}^T = (\boldsymbol{\vartheta}_1^T, \boldsymbol{\vartheta}_2^T, \boldsymbol{\vartheta}_3^T, \boldsymbol{\vartheta}_4^T, \boldsymbol{\vartheta}_5^T)^T$, where

$$\begin{aligned} \boldsymbol{\vartheta}_1 &= \begin{pmatrix} \boldsymbol{\vartheta}_{1,0} \\ \boldsymbol{\vartheta}_{1,1} \end{pmatrix} = \begin{pmatrix} \mathbf{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \mathbf{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} \\ \text{vec}(\mathbf{\Delta}^{-1}\mathbf{A}) \end{pmatrix} : (p + pr) \times 1, \\ \boldsymbol{\vartheta}_2 &= \begin{pmatrix} \boldsymbol{\vartheta}_{2,0} \\ \boldsymbol{\vartheta}_{2,1} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\beta}^T\mathbf{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\beta}^T\mathbf{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + \mathbf{L}_q\boldsymbol{\tau}_0 - \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\boldsymbol{\beta}^T\mathbf{\Delta}^{-1}\boldsymbol{\beta}) \\ \text{vec}(\mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\mathbf{\Delta}^{-1}\mathbf{A}) \end{pmatrix} : (q + qr) \times 1, \\ \boldsymbol{\vartheta}_3 &= \boldsymbol{\vartheta}_{3,0} = \text{vech}(\mathbf{\Delta}^{-1}) : k_p \times 1, \\ \boldsymbol{\vartheta}_4 &= \boldsymbol{\vartheta}_{4,0} = \text{vec}(\mathbf{\Delta}^{-1}\boldsymbol{\beta}) : pq \times 1, \\ \boldsymbol{\vartheta}_5 &= \begin{pmatrix} \boldsymbol{\vartheta}_{5,0} \\ \boldsymbol{\vartheta}_{5,1} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\boldsymbol{\beta}^T\mathbf{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{J}_q\boldsymbol{\tau}_0 \\ \text{vec}(\mathbf{J}_q\boldsymbol{\tau}) \end{pmatrix}, \end{aligned} \tag{12}$$

and

$$\begin{aligned}\psi(\boldsymbol{\eta}_y) &= -\frac{1}{2}\log|\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_{y3})| + \log(G(\boldsymbol{\Gamma}_y)) + \frac{1}{2}\boldsymbol{\eta}_{y1}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_{y3}))^{-1}\boldsymbol{\eta}_{y1} \quad (13)\\ &\coloneqq \psi_1(\boldsymbol{\eta}_y) + \psi_2(\boldsymbol{\eta}_y) + \psi_3(\boldsymbol{\eta}_y),\end{aligned}$$

with

$$\begin{aligned}G(\boldsymbol{\Gamma}_y) &= \sum_H \exp\Bigg[\left(\mathbf{J}_q C_q \text{vec}(\mathbf{H}\mathbf{H}^T)\right)^T\left(\boldsymbol{\eta}_{y5} + \mathbf{J}_q\frac{1}{2}\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_{y4}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_{y3}))^{-1}\bar{\boldsymbol{\eta}}_{y4})\right) \quad (14)\\ &\quad +\mathbf{H}^T\left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_{y4}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_{y4}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_{y3}))^{-1}\bar{\boldsymbol{\eta}}_{y4}))\right)\Bigg],\end{aligned}$$

where $\bar{\boldsymbol{\eta}}_{y4} = \text{unvec}(\boldsymbol{\eta}_{y4})$.

For any matrix $\mathbf{V}$, let $\mathcal{S}_{\mathbf{V}}$ denote the span of the columns of $\mathbf{V}$; that is, $\mathcal{S}_{\mathbf{V}} = \text{span}(\mathbf{V})$. Theorem 1 obtains the sufficient reduction for the regression of Y on $(\mathbf{X}, \mathbf{H})$ using a result from Bura et al. (2016).

Theorem 1 *Suppose that* $(\mathbf{X}, \mathbf{H}) \mid Y$ *has density given by (9). The minimal sufficient reduction for the regression* $Y \mid (\mathbf{X}, \mathbf{H})$ *is*

$$\mathbf{R}(\mathbf{X}, \mathbf{H}) = \boldsymbol{\alpha}_{\mathbf{a}}^T\left(\mathbf{T}(\mathbf{X}, \mathbf{H}) - \text{E}(\mathbf{T}(\mathbf{X}, \mathbf{H}))\right), \quad (15)$$

where $\mathbf{T}(\mathbf{X}, \mathbf{H})$ *is given by (10) and* $\boldsymbol{\alpha}_{\mathbf{a}}$ *is a basis for* $\mathcal{S}_{\mathbf{a}} = \text{span}\{\boldsymbol{\eta}_Y - \text{E}(\boldsymbol{\eta}_Y), Y \in \mathcal{Y}\}$*, with* $\boldsymbol{\eta}_Y$ *given in* (11)*.*

We prove Theorem 1 in Appendix A.2, where we see that the reduction in (15) is characterized by the coefficients of the basis for $\text{span}\left\{\boldsymbol{\eta}_Y - \text{E}(\boldsymbol{\eta}_Y), Y \in \mathcal{Y}\right\} = \text{span}(\mathbf{a})$ with

$$\mathbf{a} = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} \\ 0 \\ 0 \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix} = \begin{pmatrix} \text{unvec}(\boldsymbol{\vartheta}_{1,1}) \\ \text{unvec}(\boldsymbol{\vartheta}_{2,1}) \\ 0 \\ 0 \\ \text{unvec}(\boldsymbol{\vartheta}_{5,1}) \end{pmatrix}.$$

Since $\boldsymbol{\eta}_{y3}$ and $\boldsymbol{\eta}_{y4}$ do not depend on y, we denote them by $\boldsymbol{\eta}_3$ and $\boldsymbol{\eta}_4$, respectively, and Corollary 2 follows.

Corollary 2 *Suppose the density of* $(\mathbf{X}, \mathbf{H}) \mid Y$ *is given by* (9)*. A minimal sufficient dimension reduction for the regression of* Y *on* $(\mathbf{X}, \mathbf{H})$ *is given by*

$$\mathbf{R}(\mathbf{X}, \mathbf{H}) = \boldsymbol{\alpha}_{\mathbf{b}}^T\left(\mathbf{t}\left(\mathbf{X}, \mathbf{H}\right) - \text{E}\left(\mathbf{t}(\mathbf{X}, \mathbf{H})\right)\right), \quad (16)$$

where

$$\mathbf{t}(\mathbf{X}, \mathbf{H}) = \left(\mathbf{X}^T, \mathbf{H}^T, \left(\mathbf{J}_q\,\text{vech}(\mathbf{H}\mathbf{H}^T)\right)^T\right)^T, \quad (17)$$

and $\boldsymbol{\alpha}_{\mathbf{b}}$ *is a basis for* $\mathcal{S}_{\mathbf{b}} = \text{span}\{\mathbf{b}\}$ *with*

$$\mathbf{b} = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix} = \begin{pmatrix} \text{unvec}(\boldsymbol{\vartheta}_{1,1}) \\ \text{unvec}(\boldsymbol{\vartheta}_{2,1}) \\ \text{unvec}(\boldsymbol{\vartheta}_{5,1}) \end{pmatrix}. \quad (18)$$

As the reduction in (16) is not only sufficient but also *minimal*, we call it OPTIMAL SDR in the sequel.

Theorem 1 is consistent with previous work in model based sufficient dimension reduction. PFC (Cook and Forzani, 2008), in particular, is a special case when the predictors are all continuous. The PFC sufficient reduction is $\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathrm{E}(\mathbf{X})$, where $\boldsymbol{\alpha}_1 = \mathrm{span}(\boldsymbol{\Delta}^{-1}\mathbf{A})$, which agrees with Theorem 1 in the absence of $\mathbf{H}$. We summarize this case in Corollary 3.

Corollary 3 *When the predictor vector contains only continuous variables; that is, $q = 0$ and $\mathbf{Z} = \mathbf{X}$, the sufficient dimension reduction is*

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}_1^T \left(\mathbf{X} - \mathrm{E}(\mathbf{X})\right), \tag{19}$$

where $\mathcal{S}_{\boldsymbol{\alpha}_1} = \mathrm{span}(\boldsymbol{\alpha}_1) = \mathrm{span}(\boldsymbol{\Delta}^{-1}\mathbf{A})$, and $\mathbf{A} : p \times r$ in (3).

Corollary 4, on the other hand, obtains the minimal sufficient reduction in the important case of all binary predictors.

Corollary 4 *When the predictor vector contains only binary variables; that is, $p = 0$ and $\mathbf{Z} = \mathbf{H}$, the sufficient dimension reduction is*

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}_2^T \left(\mathbf{s}(\mathbf{H}) - \mathrm{E}\left(\mathbf{s}(\mathbf{H})\right)\right), \tag{20}$$

where

$$\mathbf{s}(\mathbf{H}) = \left(\mathbf{H}^T, \left(\mathbf{J}_q \,\mathrm{vech}(\mathbf{H}\mathbf{H}^T)\right)^T\right)^T, \tag{21}$$

and

$$\mathcal{S}_{\boldsymbol{\alpha}_2} = \mathrm{span}(\boldsymbol{\alpha}_2) = \mathrm{span}\begin{pmatrix} \mathbf{L}_q\boldsymbol{\tau} \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix}. \tag{22}$$

When the predictors are mixed, we derive a *sufficient but not minimal* reduction in Corollary (5), which we call SUB-OPTIMAL SDR.

Corollary 5 *Suppose that $(\mathbf{X}, \mathbf{H}) \mid Y$ has density (9). A sufficient reduction for the regression of Y on $(\mathbf{X}, \mathbf{H})$ is given by*

$$\mathbf{R}(\mathbf{X}, \mathbf{H}) = \boldsymbol{\alpha}_{\mathbf{c}}^T \left(\mathbf{w}(\mathbf{X}, \mathbf{H}) - \mathrm{E}(\mathbf{w}(\mathbf{X}, \mathbf{H}))\right), \tag{23}$$

with

$$\mathbf{w}(\mathbf{X}, \mathbf{H}) = \left(\mathbf{X}^T, \mathbf{H}^T, \left(\,\mathrm{vech}(\mathbf{H}\mathbf{H}^T)\right)^T\right)^T, \tag{24}$$

$$\mathrm{span}(\boldsymbol{\alpha}_{\mathbf{c}}) = \mathcal{S}_{\mathbf{c}} = \mathrm{span}\begin{pmatrix} \mathbf{c}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{c}_2 \end{pmatrix} = \mathrm{span}\begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} & 0 \\ -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} & 0 \\ 0 & \boldsymbol{\tau} \end{pmatrix}. \tag{25}$$

Predictor Distribution	Sufficient Reductions: optimal SDR	Sufficient Reductions: sub-optimal SDR
$(\mathbf{X}, \mathbf{H}) \mid Y$ with density (9)	$\boldsymbol{\alpha}_{\mathbf{b}}^T(\mathbf{t}(\mathbf{X},\mathbf{H}) - \mathrm{E}(\mathbf{t}(\mathbf{X},\mathbf{H})))$ $\mathbf{t}(\mathbf{X},\mathbf{H})$ in (17), $\mathcal{S}_{\mathbf{b}} = \mathrm{span}\{\mathbf{b}\}$, $\mathbf{b}$ in (18)	$\boldsymbol{\alpha}_{\mathbf{c}}^T(\mathbf{w}(\mathbf{X},\mathbf{H}) - \mathrm{E}(\mathbf{w}(\mathbf{X},\mathbf{H})))$ $\mathbf{w}(\mathbf{X},\mathbf{H})$ in (24), $\mathcal{S}_{\mathbf{c}}$ in (25), (27), and (28)
$\mathbf{X} \mid Y \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \boldsymbol{\Delta}\right)$	$\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))$ $\mathcal{S}_{\boldsymbol{\alpha}_1} = \mathrm{span}(\boldsymbol{\Delta}^{-1}\mathbf{A})$	
$\mathbf{H} \mid Y$ with mass function (7)	$\boldsymbol{\alpha}_2^T(\mathbf{s}(\mathbf{H}) - \mathrm{E}(\mathbf{s}(\mathbf{H})))$ $\mathbf{s}(\mathbf{H})$ in (21) $\mathcal{S}_{\boldsymbol{\alpha}_2}$ in (22)	

Table 1: Sufficient Reductions in Regressions with Mixed Predictors.

If $\mathrm{rank}(\mathbf{c}_1) = d_1 \leq \min\{r, p\}$ and $\mathrm{rank}(\mathbf{c}_2) = d_2 \leq \min\{r, q(q+1)/2\}$, then

$$\mathbf{c}_1 = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}\boldsymbol{\xi} \\ -\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi} \end{pmatrix}, \ \mathbf{c}_2 = \boldsymbol{\kappa}\boldsymbol{\iota}, \tag{26}$$

where $\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\xi}$, $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d_1}$, $\boldsymbol{\xi} \in \mathbb{R}^{d_1\times r}$, $\boldsymbol{\kappa} \in \mathbb{R}^{q(q+1)/2\times d_2}$ and $\boldsymbol{\iota} \in \mathbb{R}^{d_2\times r}$ are full rank matrices. Therefore,

$$\mathrm{span}(\mathbf{c}_1) = \mathrm{span}\{(\boldsymbol{\alpha}^T, -\boldsymbol{\alpha}^T\boldsymbol{\beta})^T\}, \tag{27}$$

$$\mathrm{span}(\mathbf{c}_2) = \mathrm{span}(\boldsymbol{\kappa}). \tag{28}$$

The weights of the sub-optimal reduction corresponding to the continuous part of the factorized joint density are obtained separately from those of the binary part. In some applications this separation may be of interest. For example, to predict the price of a certain good such as wine, continuous predictors may be related to the production process, such as alcohol content, acidity, aging time, etc., and binary predictors, such as whether it contains certain information on the label, are more related to sales strategies. Having two composite indicators (reductions), one that synthesizes quality (continuous part) and the other the marketing of the product (binary part), facilitates economic interpretability.

In Table 1, we summarize the results of this Section and tabulate the sufficient reductions for mixed normal and binary predictors.

4. Reduction Estimators and their Asymptotic Distribution

In this section we derive maximum likelihood estimators for our optimal and sub-optimal sufficient reductions, the asymptotic normality of the projection matrix of the optimal SDR, with which we also obtain asymptotic tests for dimension of both optimal and sub-optimal reductions. In addition, since identifying variables that are not associated with the outcome is important for both interpretation and for improving the predictive power

of a classifier or a regression model, we introduce a method to simultaneously obtain the sufficient reduction and carry out variable selection.

4.1 Parameter Estimation via Maximum Likelihood

We assume a random sample $(y_i, \mathbf{x}_i, \mathbf{h}_i)$, $i = 1, \ldots, n$, is drawn from the joint distribution of $(Y, \mathbf{X}, \mathbf{H})$ and that the conditional distribution models (5) and (3) hold. Finding the maximum likelihood estimators of the reductions derived in Section 3 requires first the estimation of the parameters $\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\tau}_0, \boldsymbol{\tau}$, in the joint density (8) with log-likelihood

$$\sum_{i=1}^{n} \log f_{\mathbf{X},\mathbf{H}}(\mathbf{x}_i, \mathbf{h}_i \mid y_i; \boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\tau}_0, \boldsymbol{\tau}). \tag{29}$$

We maximize (29) in two steps. First, we maximize $\sum_{i=1}^{n} \log f_{\mathbf{X}}(\mathbf{x}_i | y_i, \mathbf{h}_i; \boldsymbol{\Omega})$ to estimate the parameters $\boldsymbol{\Omega} = \{\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{H}}, \mathbf{A}, \boldsymbol{\beta}\}$. Since $\mathbf{X} \mid (\mathbf{H}, Y)$ follows a normal distribution, the maximum likelihood estimator (MLE) of $\boldsymbol{\Omega}$ is obtained from fitting a multivariate normal linear model of $\mathbf{X}$ on the centered $(\mathbf{H}, Y)$ via maximum likelihood estimation. The MLEs of $\mathbf{A}$ and $\boldsymbol{\beta}$ are $(\widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}}) = \mathbb{X}^T\mathbb{L}\left(\mathbb{L}^T\mathbb{L}\right)^{-1}$, where $\mathbb{X}$ denotes the $n \times p$ matrix with rows $(\mathbf{x}_i - \bar{\mathbf{x}})^T$, and $\mathbb{L}$ the $n \times (r+q)$ matrix with rows $\left((\mathbf{f}_{y_i} - \bar{\mathbf{f}}_y)^T, (\mathbf{h}_i - \bar{\mathbf{h}})^T\right)$, $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$, $\bar{\mathbf{f}}_y = \sum_{i=1}^{n} f_{y_i}/n$ and $\bar{\mathbf{h}} = \sum_{i=1}^{n} \mathbf{h}_i/n$. The MLE of the covariance matrix is $\widehat{\boldsymbol{\Delta}} = \left(\mathbb{X}^T - (\widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}})\mathbb{L}^T\right)\left(\mathbb{X}^T - (\widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}})\mathbb{L}^T\right)^T/n$.

Next, we estimate $\boldsymbol{\Upsilon} = (\boldsymbol{\tau}_0, \boldsymbol{\tau})$ maximizing the conditional log-likelihood function

$$\sum_{i=1}^{n} \log f_{\mathbf{H}}(\mathbf{h}_i \mid y_i; \boldsymbol{\Upsilon}).$$

Using parametrization (6), the joint probability mass function (5) can be written as

$$\Pr(\mathbf{H} \mid Y = y) = \exp\left(\sum_{j=1}^{q} \tau^*_{jj0} H_j + \sum_{j=1}^{q} \boldsymbol{\tau}_{jj}^T \mathbf{f}_y H_j + \sum_{1 \le j < j' \le q} \tau^*_{jj'0} H_j H_{j'} + \sum_{1 \le j < j' \le q} \boldsymbol{\tau}_{jj'}^T \mathbf{f}_y H_j H_{j'}\right) \frac{1}{G(\boldsymbol{\Gamma}_y)}.$$

Following Cheng et al. (2014), we consider a single binary variable H_j and condition over the rest $\mathbf{H}_{-j} = (H_1, \ldots, H_{j-1}, H_{j+1}, \ldots, H_q)$ to obtain

$$\log \frac{\Pr(H_j = 1 \mid \mathbf{H}_{-j}, Y)}{\Pr(H_j = 0 \mid \mathbf{H}_{-j}, Y)} = \tau^*_{jj0} + \boldsymbol{\tau}_{jj}^T \mathbf{f}_y + \sum_{j \ne j'} \tau^*_{jj'0} H_{j'} + \sum_{j < j'} \boldsymbol{\tau}_{jj'}^T \mathbf{f}_y H_{j'}. \tag{30}$$

Thus, the conditional log-odds for a specific binary variable H_j is linear in the parameters. Moreover, the conditional maximum likelihood estimators for these parameters can be obtained by fitting a logistic regression of H_j on $(\mathbf{f}_y, \mathbf{H}_{-j}, \mathbf{f}_y\mathbf{H}_{-j})$, so that we obtain estimators for $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}$ by fitting q univariate logistic regressions. In particular, for the sample points $(\mathbf{h}_i^T, y_i) = (h_{i1}, \ldots, h_{iq}, y_i)$, for each binary variable j $(j = 1, \ldots, q)$, the conditional

log-likelihood function is

$$\ell_j(\boldsymbol{\tau}_0, \boldsymbol{\tau}; \mathbf{h}_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \log \Pr(h_{ij} \mid \mathbf{h}_{i,-j}, y_i) = \frac{1}{n}\sum_{i=1}^{n} \left(h_{ij}\epsilon_{ij} - \log(1+\exp(\epsilon_{ij}))\right), \quad (31)$$

where $\mathbf{h}_{i,-j} = (h_{11}, \ldots, h_{i,j-1}, h_{i,j+1}, \ldots, h_{iq})$ and

$$\epsilon_{ij} = \log \frac{\Pr(h_{ij}=1 \mid \mathbf{h}_{i,-j}, y)}{\Pr(h_{ij}=0 \mid \mathbf{h}_{i,-j}, y_i)} = \tau^*_{jj0} + \boldsymbol{\tau}_{jj}^T \mathbf{f}_{y_i} + \sum_{j \neq j'} \tau^*_{jj'0} h_{ij'} + \sum_{j \neq j'} \boldsymbol{\tau}_{jj'}^T \mathbf{f}_{y_i} h_{ij'}.$$

To estimate $\boldsymbol{\Upsilon}$ we use the joint estimation algorithm proposed by Cheng et al. (2014) that maximizes $\sum_j \ell_j(\boldsymbol{\tau}_0, \boldsymbol{\tau}; \mathbf{h}_i, y_i)$.

4.2 Maximum Likelihood Estimation of the Reductions

To estimate the OPTIMAL SDR$\boldsymbol{\alpha}_\mathbf{b}$ in Corollary 2 and the SUB-OPTIMAL SDR$\boldsymbol{\alpha}_\mathbf{c}$ in Corollary 5, we need first to estimate $\mathbf{b}$ in (18) and $\mathbf{c}_1$ and $\mathbf{c}_2$ in (25). We use the ML estimators $(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}_\mathbf{H}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}_0, \widehat{\boldsymbol{\tau}})$ of the corresponding parameters in (29) in Section 4.1.

4.2.1 Optimal SDR

If $d = \dim(\mathcal{S}_\mathbf{b})$, with $d \leq \min\{r, p+q(q+1)/2\}$, the rank of $\mathbf{b}$ is also d with singular value decomposition

$$\mathbf{b} = \mathbf{U}^T \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}, \quad (32)$$

where $k_1 \geq \ldots \geq k_d > 0$ are the singular values of $\mathbf{b}$, $\mathbf{K} = \operatorname{diag}(k_1, \ldots, k_d)$, $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ is an $m \times m$ orthogonal matrix with $m = p + q(q+1)/2$, $\mathbf{U}_1 : m \times d$, $\mathbf{U}_0 : m \times (m-d)$, and $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ is an $r \times r$ orthogonal matrix with $\mathbf{R}_1 : r \times d$, $\mathbf{R}_0 : r \times (r-d)$. The submatrices satisfy $\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_0\mathbf{U}_0^T = \mathbf{I}_m$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_d$, $\mathbf{U}_0^T\mathbf{U}_0 = \mathbf{I}_{m-d}$, $\mathbf{U}_0^T\mathbf{U}_1 = \mathbf{0}$, $\mathbf{R}_1\mathbf{R}_1^T + \mathbf{R}_0\mathbf{R}_0^T = \mathbf{I}_r$, $\mathbf{R}_1^T\mathbf{R}_1 = \mathbf{I}_d$, $\mathbf{R}_0^T\mathbf{R}_0 = \mathbf{I}_{r-d}$, $\mathbf{R}_0^T\mathbf{R}_1 = \mathbf{0}$. Then,

$$\mathbf{b} = \mathbf{U}_1\mathbf{K}\mathbf{R}_1^T, \quad (33)$$

and, as a consequence, $\boldsymbol{\alpha}_\mathbf{b}$ in Corollary 2 can be set to $\mathbf{U}_1$. Plugging in the ML estimators $(\widehat{\boldsymbol{\Delta}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}})$ yields the ML estimator of $\mathbf{b}$,

$$\widehat{\mathbf{b}} = \begin{pmatrix} \widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{A}} \\ \mathbf{L}_q\widehat{\boldsymbol{\tau}} - \widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{A}} \\ \mathbf{J}_q\widehat{\boldsymbol{\tau}} \end{pmatrix} = \begin{pmatrix} \operatorname{unvec}(\widehat{\boldsymbol{\vartheta}}_{1,1}) \\ \operatorname{unvec}(\widehat{\boldsymbol{\vartheta}}_{2,1}) \\ \operatorname{unvec}(\widehat{\boldsymbol{\vartheta}}_{5,1}) \end{pmatrix}. \quad (34)$$

The singular value decomposition of the MLE of $\mathbf{b}$ is

$$\widehat{\mathbf{b}} = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\mathbf{K}}_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{K}}_0 \end{pmatrix} \widehat{\mathbf{R}}, \quad (35)$$

where $\widehat{\mathbf{K}}_1 = \operatorname{diag}(\widehat{k}_1, \ldots, \widehat{k}_d)$, $\widehat{\mathbf{K}}_0 = \operatorname{diag}(\widehat{k}_{d+1}, \ldots, \widehat{k}_{\min(m,r)})$, $\widehat{k}_i$ are the singular values of $\widehat{\mathbf{b}}$ in decreasing order, $\widehat{\mathbf{U}}$ is an $m \times m$ orthogonal matrix whose columns are the left singular

vectors of $\widehat{\mathbf{b}}$, and $\widehat{\mathbf{R}}$ is an $r \times r$ orthogonal matrix, whose columns are the right-singular vectors of $\widehat{\mathbf{b}}$. Let $\widehat{\mathbf{U}}_1$ be the first d columns of $\widehat{\mathbf{U}}$, $\widehat{\mathbf{R}}_1$ the first d columns of $\widehat{\mathbf{R}}^T$, and $\widehat{\mathbf{B}} = \widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T$. An estimator of $\mathbf{b}$ subject to $d = \dim(\mathcal{S}_{\mathbf{b}})$ is

$$\widehat{\mathbf{b}}^{(d)} = \widehat{\mathbf{U}}_1\widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T = \widehat{\mathbf{U}}_1\widehat{\mathbf{B}}, \tag{36}$$

and an estimator of the reduction $\boldsymbol{\alpha}_{\mathbf{b}}$ in Corollary 2 is

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{b}} = \widehat{\mathbf{U}}_1. \tag{37}$$

4.2.2 Sub-optimal SDR

To obtain an estimator for the space $\mathcal{S}_{\mathbf{c}}$ in (25) that gives the sub-optimal sufficient reduction (23), we set $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2)$, where $\mathbf{c}_1$ and $\mathbf{c}_2$ are given in (26), with $\text{rank}(\mathbf{c}_1) = d_1$ and $\text{rank}(\mathbf{c}_2) = d_2$. Plugging in the MLE $(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}_{\mathbf{H}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\tau}}_0, \widehat{\boldsymbol{\tau}})$ of the corresponding parameters in (29) from Section 4.1, we obtain estimators of $\mathbf{c}_1$ and $\mathbf{c}_2$,

$$\widehat{\mathbf{c}}_1 = \begin{pmatrix} \widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{A}} \\ -\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{A}} \end{pmatrix}, \quad \widehat{\mathbf{c}}_2 = \widehat{\tau}.$$

We consider their respective SVD decompositions as in Section 4.2.1. Let $\widehat{\mathbf{U}}_{c_1}$ denote the first d_1 left eigenvectors of $\widehat{\mathbf{c}}_1$, and $\widehat{\mathbf{U}}_{c_2}$ the first d_2 left eigenvectors of $\widehat{\mathbf{c}}_2$. Then, we define the estimator for the sub-optimal sufficient reduction in (23) as

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{c}} = \begin{pmatrix} \widehat{\mathbf{U}}_{\mathbf{c}_1} & 0 \\ 0 & \widehat{\mathbf{U}}_{\mathbf{c}_2} \end{pmatrix}.$$

4.3 Variable selection

We propose a method to combine computation of the sufficient reduction with variable selection by removing redundant variables from the reduction. This is carried out simultaneously by introducing structured regularization on matrix factorization. We exploit the factorization of the full rank maximum likelihood estimate $\widehat{\mathbf{b}}$ into a relevant full-rank factor $\mathbf{C} \in \mathbb{R}^{p+q(q+1)/2 \times d}$, which determines the reduction, and a matrix $\mathbf{B}$ that is immaterial.

The procedure is built upon the fact that the reduced rank estimator $\widehat{\mathbf{b}}^{(d)} = \widehat{\mathbf{U}}_1\widehat{\mathbf{B}}$ in (36) is also the solution to the least squares minimization problem

$$\min_{\mathbf{C} \in \mathbb{R}^{p+q(q+1)/2 \times d}, \mathbf{C}^T\mathbf{C} = \mathbf{I}, \mathbf{B} \in \mathbb{R}^{d \times r}} (\text{vec}(\widehat{\mathbf{b}}) - \text{vec}(\mathbf{CB}))^T(\text{vec}(\widehat{\mathbf{b}}) - \text{vec}(\mathbf{CB})), \tag{38}$$

where $\widehat{\mathbf{b}}$ is the maximum likelihood estimator of $\mathbf{b}$. The solution can be expressed as $\widehat{\mathbf{C}} = \widehat{\mathbf{U}}_1\mathbf{V}$, for some orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$, so that $\text{span}(\widehat{\mathbf{C}}) = \text{span}(\widehat{\mathbf{U}}_1)$.

All sufficient reductions in Section 4.2.1 are of the form $\mathbf{R}(\mathbf{X}, \mathbf{H}) = \mathbf{U}_1^T(\mathbf{t}(\mathbf{X}, \mathbf{H}) - \text{E}(\mathbf{t}(\mathbf{X}, \mathbf{H}))$. If t_j is the jth component of $\mathbf{t}(\mathbf{X}, \mathbf{H})$, and t_j is not associated with Y, the jth row of $\mathbf{U}_1$ is zero. Therefore, identifying predictors that are conditionally independent of Y corresponds to identifying the rows of $\mathbf{C}$ that contain only 0. This can be achieved using mixed-norm regularizers that are known to induce structured sparsity in the estimates (Bach et al., 2012).

The proposed procedure is as follows. For a fixed d, once we obtain $\widehat{\mathbf{b}}^{(d)} = \widehat{\mathbf{U}}_1\widehat{\mathbf{B}}$ in (36), we solve

$$\underset{\mathbf{C}\in\mathbb{R}^{(p+q(q+1)/2)\times d},\mathbf{C}^T\mathbf{C}=\mathbf{I}}{\arg\min} \left(\text{vec}(\widehat{\mathbf{b}}) - \text{vec}(\mathbf{C}\widehat{\mathbf{B}})\right)^T \left(\text{vec}(\widehat{\mathbf{b}}) - \text{vec}(\mathbf{C}\widehat{\mathbf{B}})\right) + \lambda\Omega(\mathbf{C}), \tag{39}$$

where $\Omega(\mathbf{C})$ is a mixed-norm regularizer which penalizes the rows of $\mathbf{C}$ in a similar manner to group-lasso. The specific form of $\Omega(\mathbf{C})$ depends on the type of predictor variables involved in the problem. We provide details in Appendix C.

4.4 Asymptotic distribution of the optimal sufficient reduction estimator

We derive the asymptotic distribution of the projection onto the column space of the estimated optimal sufficient reduction $\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}$ in (37), $\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}} = \widehat{\boldsymbol{\alpha}}_{\mathbf{b}}(\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}^T\widehat{\boldsymbol{\alpha}}_{\mathbf{b}})^{-1}\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}^T$. We use this result in the derivation of the asymptotic tests for dimension in Section 4.5. It can also be potentially used for inference about the sufficient dimension reduction, as, for example, computing confidence intervals for the prediction of future observations using results from Forzani et al. (2019).

Proposition 6 *Suppose that* $(\mathbf{X}, \mathbf{H}) \mid Y$ *has probability mass function (9) with the natural parameters* $\boldsymbol{\eta}_Y$ *satisfying (11) and that* $\mathbf{b}$ *has rank* d*. Then,*

$$\sqrt{n}\ \text{vec}\left(\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}} - \mathbf{P}_{\boldsymbol{\alpha}_{\mathbf{b}}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}}\right),$$

with

$$\mathbf{V}_{\widehat{\boldsymbol{\alpha}}_{\mathbf{b}}} = (\mathbf{I}_{m^2}\otimes\mathbf{K}_{mm})\left(\mathbf{b}^-\otimes\mathbf{Q}_{\mathbf{b}}\right)^T\mathbf{V}_{rcl}\left(\mathbf{b}^-\otimes\mathbf{Q}_{\mathbf{b}}\right)(\mathbf{I}_{m^2}\otimes\mathbf{K}_{mm}), \tag{40}$$

where $\mathbf{b}^-$ *is the Moore-Penrose generalized inverse of* $\mathbf{b}$*,*

$$\mathbf{V}_{rcl} = \mathbf{W}\mathbf{M}\mathbf{V}\mathbf{M}^T\mathbf{W}^T, \tag{41}$$

with

$$\mathbf{V}^{-1} = \text{E}\left(\mathbf{F}_y^T\mathbf{J}\mathbf{F}_y\right), \tag{42}$$

$\mathbf{F}_y$ *is defined in (11),* $\mathbf{J}$ *is the matrix of partial derivatives given by*

$$\mathbf{J} = \frac{\partial^2\psi(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_y\partial\boldsymbol{\eta}_y^T}, \tag{43}$$

$$\mathbf{M} = \begin{pmatrix} \mathbf{0}_{pr\times p} & \mathbf{I}_{pr} & \mathbf{0}_{pr\times q} & \mathbf{0}_{pr\times qr} & \mathbf{0}_{pr\times m_p} & \mathbf{0}_{pr\times qp} & \mathbf{0}_{pr\times k_q} & \mathbf{0}_{pr\times rk_q} \\ \mathbf{0}_{qr\times p} & \mathbf{0}_{qr\times pr} & \mathbf{0}_{qr\times q} & \mathbf{I}_{qr} & \mathbf{0}_{qr\times m_p} & \mathbf{0}_{qr\times qp} & \mathbf{0}_{qr\times k_q} & \mathbf{0}_{qr\times rk_q} \\ \mathbf{0}_{rk_q\times p} & \mathbf{0}_{rk_q\times pr} & \mathbf{0}_{rk_q\times q} & \mathbf{0}_{rk_q\times qr} & \mathbf{0}_{rk_q\times m_p} & \mathbf{0}_{rk_q\times qp} & \mathbf{0}_{rk_q\times k_q} & \mathbf{I}_{rk_q} \end{pmatrix}, \tag{44}$$

and

$$\mathbf{W} = \left(\mathbf{I}_r\otimes\begin{pmatrix}\mathbf{I}_p\\ \mathbf{0}_{q\times p}\\ \mathbf{0}_{k_q\times p}\end{pmatrix}, \mathbf{I}_r\otimes\begin{pmatrix}\mathbf{0}_{p\times q}\\ \mathbf{I}_q\\ \mathbf{0}_{k_q\times q}\end{pmatrix}, \mathbf{I}_r\otimes\begin{pmatrix}\mathbf{0}_{p\times k_q}\\ \mathbf{0}_{q\times k_q}\\ \mathbf{I}_{k_q}\end{pmatrix} \right). \tag{45}$$

The proof of Proposition 6 is given in Appendix B.

4.5 Tests for dimension

We propose two asymptotic tests for the dimension of the sufficient reduction in OPTIMAL SDR. We adapt these tests for the case of SUB-OPTIMAL SDR, in order to estimate the dimension of the continuous predictors separately from the binary predictors.

The dimension of the sufficient reduction coincides with the rank d of $\mathbf{b}$ in (18), which we estimate by sequentially testing the hypotheses

$$H_0 : \text{rank}(\mathbf{b}) = j \quad \text{vs.} \quad H_1 : \text{rank}(\mathbf{b}) > j, \tag{46}$$

for $j = 0, 1, \ldots, \min(r, m)$, where $m = p + q(q+1)/2$. For a fixed level α, the estimated rank is the smallest value of j for which the null is not rejected.

Bura and Yang (2011) proposed asymptotic tests for the rank of random matrices in sequential hypothesis testing. To construct the corresponding tests for dimension, we consider the singular value decomposition of $\mathbf{b}$ in (33) and $\widehat{\mathbf{b}}$ in (36) with $d = j$.

The first statistic we use to test (46) is $\Lambda_1(j) = n \sum_{i=j+1}^{\min(m,r)} \hat{k}_i^2$, where $\widehat{k}_i$'s are the singular values of $\widehat{\mathbf{b}}$ in descending order. Proposition 6 obtains the asymptotic normality of $\widehat{\mathbf{b}}$ with covariance $\mathbf{V}_{rlc}$ in (41). When $\text{rank}(\mathbf{b}) = j$,

$$\Lambda_1(j) \xrightarrow{\mathcal{D}} \sum_{i=1}^{s} \omega_i X_i^2, \tag{47}$$

where $s = \min(\text{rank}(\mathbf{V}_{rcl}), (r-j)(m-j))$, X_i^2 are independent chi-squared random variables with 1 degree of freedom, and the weights are the descending eigenvalues of $\mathbf{Q} = (\mathbf{R}_0^T \otimes \mathbf{U}_0^T)\mathbf{V}_{rcl}(\mathbf{R}_0 \otimes \mathbf{U}_0)$ [see Bura and Yang (2011)]. In practice, the weights ω_i, $i = 1, \ldots, s$, are replaced by $\widehat{\omega}_1 \geq \widehat{\omega}_2 \geq \ldots \geq \widehat{\omega}_s$, the descending eigenvalues of

$$\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T)\widehat{\mathbf{V}}_{rcl}(\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0), \tag{48}$$

where $\widehat{\mathbf{V}}_{rcl}$ is a consistent estimate of $\mathbf{V}_{rcl}$. This test rejects H_0 if $\Lambda_1(j) > q_\alpha$, where q_α is the $(1-\alpha)$ percentile of the distribution of $\sum_{i=1}^{s} \widehat{\omega}_i X_i^2$. We estimate q_α from the empirical distribution function of Λ_1, by generating 10000 realizations of $\sum_{i=1}^{s} \widehat{\omega}_i X_i^2$ and computing the empirical quantile $\widehat{q}_\alpha$.

The second is a Wald test with test statistic, $\Lambda_2(j) = n\text{vec}(\widehat{\mathbf{K}}_0)^T \widehat{\mathbf{Q}}^\dagger \text{vec}(\widehat{\mathbf{K}}_0)$, where $\widehat{\mathbf{K}}_0$ is defined in (35) and $\widehat{\mathbf{Q}}^\dagger$ is the Moore-Penrose inverse of $\widehat{\mathbf{Q}}$ in (48). Following Bura and Yang (2011), since $\widehat{\mathbf{b}}$ is asymptotically normal, if $j = \text{rank}(\mathbf{b})$, then $\Lambda_2(j) \xrightarrow{\mathcal{D}} \chi^2(s)$, where the degrees of freedom are $s = \min(\text{rank}(\mathbf{V}_{rcl}), (r-j)(m-j))$. The rejection region is $\Lambda_2(j) > \chi_\alpha^2(s)$, where $\chi_\alpha^2(s)$ is the $(1-\alpha)$ percentile of the $\chi^2(s)$ distribution.

5. Simulation Studies

We assess the performance of the proposed methods in estimating the sufficient reduction and its dimension, out-of-sample prediction, and variable selection in simulations.

In all our simulations, the response is generated from the uniform distribution on the integers $\{1, \ldots, r+1\}$, with $r = 5$, and we set $\mathbf{f}_y = I(y = j) - n_j/n$, where I is the indicator function, n denotes the total sample size and n_j the number of observations in category j,

for $j = 1, \ldots, r$. All reported results are based on sample sizes $n = 100, 200, 300, 500, 750$, and 100 repetitions. The `R` code we used in both simulations and real data analyses in Section 6 can be found at https://github.com/lforzani/SDR_mixed_predictions.

5.1 Estimation, prediction and dimension tests

We assess the accuracy of estimating $\text{span}(\boldsymbol{\alpha})$ with $\text{span}(\widehat{\boldsymbol{\alpha}})$ using $||\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\widehat{\boldsymbol{\alpha}}}||_2$ [see Ye and Lim (2016)]. The prediction error is computed as $||\mathbf{P}_{\boldsymbol{\alpha}^T(\mathbf{X}_N, \mathbf{H}_N)} - \mathbf{P}_{\widehat{\boldsymbol{\alpha}}^T(\mathbf{X}_N, \mathbf{H}_N)}||_2$, where $(\mathbf{X}_N, \mathbf{H}_N)$ is a new sample of size $N = 2000$ that is independent of the training sample. We estimate the sufficient reduction using the true d.

5.1.1 Continuous predictors

We generate p-variate continuous predictors as $\mathbf{X} \mid Y = y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta})$ with $\boldsymbol{\mu}_y = \mathbf{A}\mathbf{f}_y$ for $\mathbf{A} = \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{p \times d}$ of $\text{rank}(\boldsymbol{\alpha}) = d$ and $\boldsymbol{\xi} \in \mathbb{R}^{d \times r}$. We let $p = 20$ and $\mathbf{0}_l$, $\mathbf{1}_l$ denote the l-vectors of zeros and ones, respectively.

(a) For $d = 1$, we set $\boldsymbol{\xi} = \mathbf{1}_r^T$, $\boldsymbol{\alpha} = (\mathbf{0}_{p/2}^T, \mathbf{1}_{p/2}^T)^T$, $\boldsymbol{\Delta} = 5(\mathbf{I}_p + \rho\boldsymbol{\alpha}\boldsymbol{\alpha}^T)$ with $\rho = 0.55$.

(b) For $d = 2$, we set

$$\boldsymbol{\xi} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is an orthonormal basis of span $\left((\mathbf{0}_{p/2}^T, \mathbf{1}_{p/2}^T)^T, (\mathbf{0}_{p/2}^T, \mathbf{1}_{p/4}^T, -\mathbf{1}_{p/4}^T)^T\right)$, and $\boldsymbol{\Delta} = 5(\mathbf{I}_p + \rho_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^T + \rho_2\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2^T)$, for $\rho_1 = 0.55$ and $\rho_2 = 0.25$.

5.1.2 Binary predictors

We generate $q = 10$ binary predictors assuming that $\mathbf{H} \mid Y$ follows an Ising model with parameters $\{\boldsymbol{\tau}_0, \boldsymbol{\tau}\}$, where $\boldsymbol{\tau} = [\,\text{vech}(\boldsymbol{\tau}_1), \ldots, \text{vech}(\boldsymbol{\tau}_r)]$, $\boldsymbol{\tau}_j$ are $q \times q$ matrices and set $\boldsymbol{\tau}_0 = \mathbf{0}$.

(a) For $d = 1$, and $j = 1, \ldots, r$, $\boldsymbol{\tau}_j = 3 \times \mathbf{K}_1 / \sqrt{\sum_{ij}([\mathbf{K}_1]_{ij})}$ with

$$\mathbf{K}_1 = \left(\begin{array}{cccccc|c} 1 & 30 & 5 & 0 & 0 & 0 & \mathbf{0}_{1\times 4} \\ 30 & 1 & 10 & 0 & 0 & 0 & \vdots \\ 5 & 10 & 1 & 30 & 0 & 0 & \vdots \\ 0 & 0 & 30 & 1 & 30 & 0 & \vdots \\ 0 & 0 & 0 & 30 & 1 & 30 & \vdots \\ 0 & 0 & 0 & 0 & 30 & 1 & \vdots \\ 0 & 0 & 0 & 0 & 0 & 30 & \mathbf{0}_{1\times 4} \\ \hline \mathbf{0}_{3\times 1} & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 4} \end{array}\right).$$

(b) For $d = 2$, $\boldsymbol{\tau}_j = 3 \times \mathbf{K}_1 / \sqrt{\sum_{ij}([\mathbf{K}_1]_{ij})}$, for $j = 1, 3, 4, 5$, and

$$\boldsymbol{\tau}_2 = \frac{12}{\sqrt{6}} \times \begin{pmatrix} \mathbf{I}_6 & \mathbf{0}_{6\times 4} \\ \mathbf{0}_{4\times 6} & \mathbf{0}_{4\times 4} \end{pmatrix}.$$

5.1.3 Mixed predictors

(a) For $d = 1$, we use the same parameters as for continuous $\mathbf{X} \mid (\mathbf{H}, Y)$ and binary variables $\mathbf{H} \mid Y$ in Sections 5.1.1, 5.1.2, respectively. Moreover, we set $\boldsymbol{\mu}_H = 0$ and $\boldsymbol{\beta} = (\mathbf{1}_{p\times 6}/10, \mathbf{0}_{p\times 4}) \in \mathbb{R}^{p\times q}$, to induce sparsity in the binary predictors.

(b) For $d = 2$, we generate $\mathbf{H} \mid Y$ as in Section 5.1.2 with dimension 1 and $\mathbf{X} \mid (\mathbf{H}, Y)$ as in (a) with dimension 2.

In Figure 1, we plot on the y-axis the estimation error, $\|\mathbf{P}_{\widehat{\boldsymbol{\alpha}}} - \mathbf{P}_{\boldsymbol{\alpha}}\|_2$, and the prediction error, $\left\|\mathbf{P}_{\widehat{\boldsymbol{\alpha}}^T\mathbf{X}_N} - \mathbf{P}_{\boldsymbol{\alpha}^T\mathbf{X}_N}\right\|_2$, for OPTIMAL SDR versus the training sample size on the x-axis across all our simulation scenarios. For all types of predictors the prediction is smaller than the estimation error and both decrease as the sample size increases. Moreover, both increase as the dimension increases from 1 to 2 in the left and right panels, respectively, across types of predictors. When comparing types of predictors, continuous predictors exhibit higher estimation and prediction errors across sample sizes and mixed predictors result in the highest estimation and prediction errors.

For comparative purposes, we also include in Figure 1 the results from applying Principal Components Analysis (PCA), as unsupervised alternative, and Principal Fitted Component (PFC) that, although supervised, is designed only for continuous variables. In both cases, the reductions are on variables derived from the sufficient statistics in our SDR methods. That is, for mixed predictors, we apply PCA and PFC on (17), for continuous predictors $\mathbf{X}$ on themselves and for binary on (21). Figure 1 shows the clear advantage of OPTIMAL SDR for continuous and binary predictors. For mixed, even though PCA and PFC have an advantage since they are applied on the sufficient statistics and not on the predictors themselves, OPTIMAL SDR nevertheless performs better than both for $d = 1$ and roughly on par with PFC, which is also the closest to OPTIMAL SDR.

In Figure 2 we plot the estimation and prediction error of SUB-OPTIMAL SDR, where the continuous and binary variables are reduced separately. The pattern of behavior is consistent with that of OPTIMAL SDR in Figure 1, with the continuous variables inducing larger errors of both types across sample sizes and $d = 1, 2$. Again, the errors are smaller for dimension 1.

Under the same simulation settings, we also evaluate the performance of our simultaneous variable selection and dimension reduction method in Section 4.3. Selection of the hyperparameters (λ, γ) in (39) is carried out via 10-fold cross validation and minimizing the prediction error as optimization criterion. The procedure starts by estimating an upper bound λ_m so that the whole estimate vanishes for any $\lambda > \lambda_m$. We then set a grid of n_λ candidate values for λ, uniformly spaced on a logarithmic scale between 0 and λ_m. Here, we set $n_\lambda = 100$. For γ, we consider 11 values uniformly spaced in $[0, 1]$. In each fold, an initial full-rank estimate of the reduction is computed using the training set and then factorized using truncated SVD to compute $\widehat{\mathbf{B}}$ and an initial estimate for $\mathbf{C}$. The solution to (39) is computed for each pair of candidate values (λ_k, γ_k). The obtained reduction is applied to both the training and the test samples. Next, we fit a prediction model using the reduced training set and evaluate the prediction error on the reduced test sample. The average prediction error over the ten cross-validation folds is then computed for each candidate pair (λ_k, γ_k). We pick the combination that attains the smallest mean prediction error.

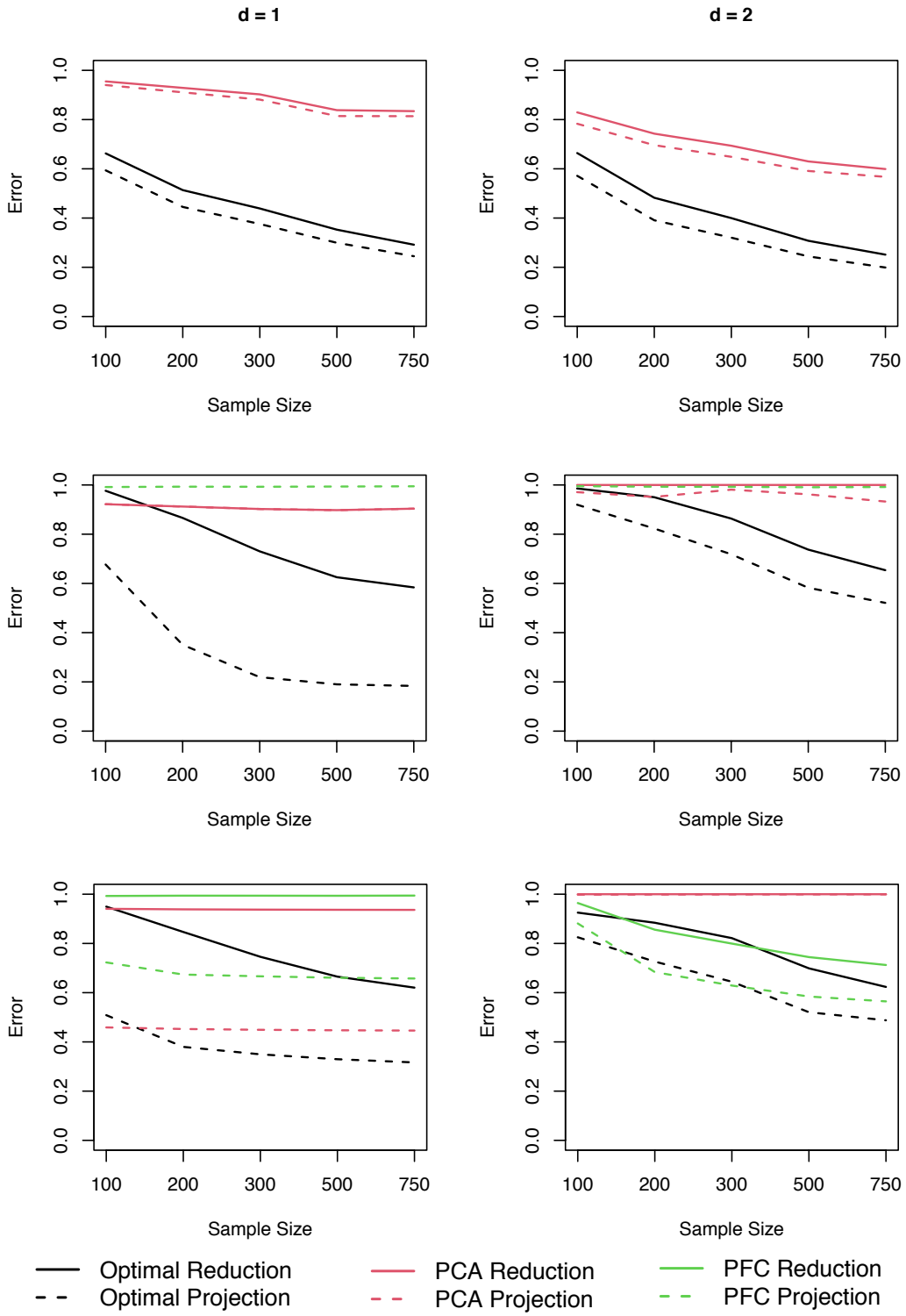

Figure 1: L_2 error of the estimation (Reduction) and out of sample prediction (Projection) of OPTIMAL SDR (black), PCA (red) and PFC (green) with continuous (first row), binary (second row) and mixed (last row) predictors for $d = 1$ (left panel) and $d = 2$ (right panel).

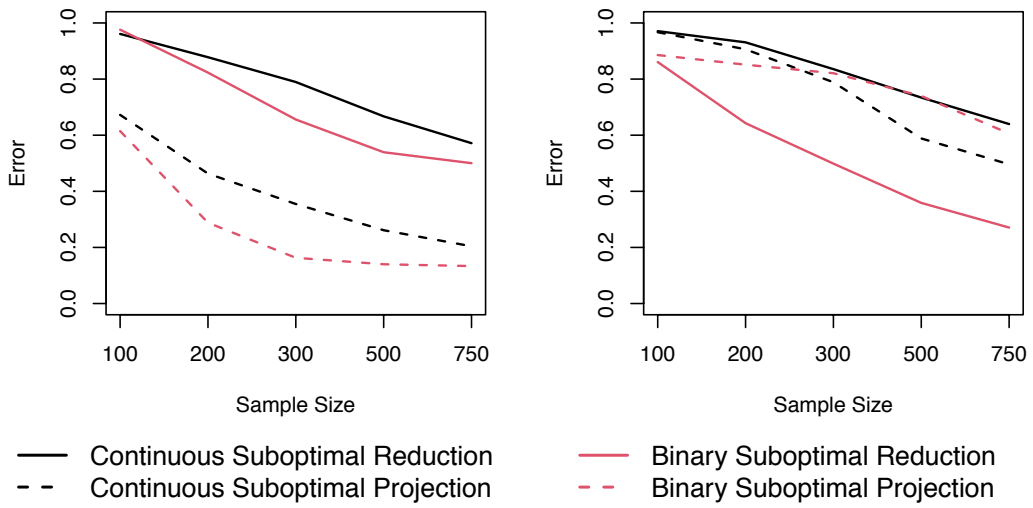

Figure 2: L_2 error of the estimation (Reduction) and out of sample prediction (Projection) of SUB-OPTIMAL SDR with mixed predictors with continuous (black) and binary (red) predictos for $d = 1$ (left panel) and $d = 2$ (right panel).

In Table 2, we report the proportion of variables correctly identified as non-relevant (true positives, TP) and the proportion of variables erroneously assessed as non-relevant (false negatives, FN). Between $d = 1$ and $d = 2$, TP is higher across sample sizes, whereas FN is lower. Both rates improve substantially as the sample size increases. When all predictors are continuous, both rates are lower across sample sizes. This is expected since the inclusion of a binary variable results in second order interaction effects in the reduction. Therefore, to rule out a binary variable both its own-coefficient and all the coefficients of its interaction terms must be zero. Overall, our regularized SDR approach achieves high true positive and small false negative rates for reasonable sample sizes.

In Table 3 we report the proportion of times out of 100 replications that the dimension d was correctly estimated based on the sequential tests of dimension in Section 4.5 for all our simulation settings. The sample size has a noticeable effect in the accuracy of the estimation of dimension, as expected since both tests are asymptotic. The weighted χ^2 test accuracy suffers more from increasing the dimension and all binary predictors as compared to that of the chi-squared test, across sample sizes. For mixed predictors, as well, the chi-squared test exhibits higher accuracy for both optimal and sub-optimal SDR across sample sizes.

Additionally, we check the robustness of our methods to non-normality of the continuous predictors. For this, we draw $\mathbf{X}$ from a multivariate non-central t-distribution with 5 degrees of freedom and the same mean and variance as the normal $\mathbf{X}$. The results are reported in Table 6 in Appendix D. We can see the methods continue to work well for heavy-tailed t predictors. In contrast, the unsupervised alternative, PCA, gives poorer results compared to PCA with normal predictors. As an aside, Table 6 provides further evidence of the much better performance of OPTIMAL SDR compared to PCA uniformly across normal and non-normal predictors.

			Sample Size				
Predictors	d	Rates	100	200	300	500	750
Continuous	1	TP	0.653	0.751	0.796	0.851	0.889
		FN	0.314	0.170	0.095	0.044	0.012
	2	TP	0.521	0.591	0.629	0.748	0.843
		FN	0.165	0.048	0.014	0.004	0.002
Binary	1	TP	0.188	0.310	0.400	0.55	0.623
		FN	0.167	0.117	0.045	0.015	0.018
	2	TP	0.255	0.300	0.368	0.458	0.528
		FN	0.048	0.020	0.012	0.000	0.000
Mixed	1	TP	0.632	0.592	0.589	0.671	0.674
		FN	0.493	0.333	0.196	0.200	0.170
	2	TP	0.596	0.656	0.639	0.583	0.610
		FN	0.451	0.413	0.325	0.163	0.124

Table 2: Accuracy of the regularized estimator in variable selection.

Predictors				Sample Size				
	Dimension	Test	Method	100	200	300	500	750
Continuous	$d = 1$	Weighted χ^2		0.60	0.82	0.83	0.89	0.95
		χ^2		0.00	0.99	0.99	1.00	0.98
	$d = 2$	Weighted χ^2		0.65	0.77	0.86	0.94	0.94
		χ^2		0.2	1.00	0.99	0.97	0.95
Binary	$d = 1$	Weighted χ^2		0	0.80	0.96	0.94	0.94
		χ^2		0	0.02	0.94	0.99	0.94
	$d = 2$	Weighted χ^2		0	0.02	0.30	0.66	0.96
		χ^2		0.02	0.20	0.92	0.96	0.94
Mixed	$d = 1$	Weighted χ^2	optimal SDR	0	0.2	0.45	0.64	0.94
		Weighted χ^2	sub-optimal SDR(cts)	0.68	0.84	0.89	0.90	0.95
		Weighted χ^2	sub-optimal SDR(bin)	0.5	0.75	0.87	0.94	0.95
		χ^2	optimal SDR	0	0.18	1	0.98	0.98
		χ^2	sub-optimal SDR(cts)	0	0.95	0.98	0.98	0.95
		χ^2	sub-optimal SDR(bin)	0	0.18	0.94	0.98	0.96
	$d = 2$	Weighted χ^2	optimal SDR	0	0.06	0.30	0.45	0.90
		Weighted χ^2	sub-optimal SDR(cts)	0.60	0.75	0.84	0.95	0.95
		Weighted χ^2	sub-optimal SDR(bin)	0	0.08	0.40	0.56	0.96
		χ^2	optimal SDR	0.08	0.36	0.64	0.92	0.93
		χ^2	sub-optimal SDR(cts)	0.12	0.98	0.99	0.96	0.95
		χ^2	sub-optimal SDR(bin)	0.22	0.30	0.96	0.96	0.95

Table 3: Proportion of correct dimension estimation under the simulation settings in Section 5.

6. Data Analyses

We compare our method with other approaches, such as generalized linear models and principal component regression, in two data applications. Specifically, we compare our methods with PCA and PCAmix in Sections 6.1 and 6.2. PCAmix (Chavent et al., 2012, 2014) is a version of PCA that accommodates mixed variables and implements *PCA with metrics*; i.e., Generalized Singular Value Decomposition (GSVD) of pre-processed data [see Chavent et al. (2014) for details]. PCAmix is ordinary standard PCA, when all variables are continuous, and standard multiple correspondence analysis (MCA), when all variables are categorical (Greenacre and Blasius (2006), Zhu et al. (2011), Camiz and Gomes (2013)).

6.1 Krzanowski Data Sets

Krzanowski (1975) studied the problem of discriminating between two groups in the presence of both binary and continuous explanatory variables. Krzanowski (1975) modeled the mixed predictors using the *location model* (Olkin and Tate, 1961) and proposed an allocation rule to two groups similar to Fisher's discriminant function. The location model transforms the q binary variables $H_1, \ldots, H_q$ to the corresponding 2^q-category multinomial vector and requires the continuous variables be conditionally normal in each of the 2^q categories with different means and same variance-covariance matrix. He showed that the simple linear discriminant function often gives satisfactory results, except when there is interaction between the mixed variables.

We analyze four of the five data sets in Krzanowski's paper which contains continuous and binary predictors and a binary response.

1. **Data Set 1:** Ten variables recorded on 40 patients who were surgically treated for renal hypertension. Seven of the variables were continuous and three binary. After one year, 20 patients were classified as improved and 20 as unimproved.

2. **Data Set 2:** Seven variables recorded on 93 patients suffering from jaundice. Four of the variables were continuous and three binary. The two groups were patients requiring medical and surgical treatment.

3. **Data Set 3:** Twelve variables recorded on 62 patients suffering from jaundice. Eight of the variables were continuous and four binary. The two groups were patients requiring medical and surgical treatment.

4. **Data Set 4:** Eleven variables recorded on 186 patients who underwent ablative surgery for advanced breast cancer between 1958 and 1965 at Guy's Hospital, London. Six of the variables were continuous and three binary. The two groups were patients for which the treatment was deemed to be successful and failure.

Some of the continuous variables were transformed to normality across all data sets. Since the response is binary, $\mathbf{f}_y$ in (8) is a vector of frequencies with $r = 1$, so that the dimension either SDR method can detect cannot exceed 1. We reduced the mixed predictors using our two methods, optimal SDR and sub-optimal SDR, and also PCA and PCAmix setting $d = 1$. In order to assess the classification accuracy of each method, the

Set		optimal SDR	sub-optimal SDR	Full	PCA	PCAmix	Location	Fisher	Logistic
1	MR	**0.250**	0.300	0.375	0.325	0.425	0.350	0.325	0.325
	AUC	**0.918**	**0.918**	0.885	0.675	0.575	-	-	-
2	MR	0.280	**0.204**	0.258	0.387	0.290	0.290	0.280	0.301
	AUC	0.857	**0.858**	0.837	0.513	0.469	-	-	-
3	MR	0.161	**0.145**	0.226	0.484	0.500	0.226	0.177	0.222
	AUC	0.949	**0.951**	0.944	0.623	0.646	-	-	-
4	MR	**0.296**	**0.290**	0.392	0.457	0.430	0.328	0.382	0.371
	AUC	**0.784**	**0.785**	0.738	0.544	0.572	-	-	-

Table 4: Leave-one-out misclassification rates and AUC values for four data sets in Krzanowski (1975).

reduced predictors serve as independent variables in a logistic regression model. For comparison, we also fit an *unreduced* logistic regression model with all the original predictors, which we refer to as Full.

In Table 4 we report the leave-one-out misclassification rate and the area under the receiver operator characteristics curve, AUC (Pepe, 2003, p. 67), with the smallest and largest values, respectively, in boldface. sub-optimal SDRemerges as the best method to summarize the mixed predictors with respect to misclassification error, followed by SDR Optimal that has better performance for data set 1. With respect to AUC, sub-optimal SDR is always the best.

In Table 4, we provide the leave-one-out misclassification rates of Fisher's LDA, logistic regression and Krzanowski's allocation rule based on the location model, as reported in Krzanowski (1975, Tab. 3). sub-optimal SDRexhibits better performance than Krzanowski's location model across data sets. optimal SDR performs the best in all data sets except for data set 2 where it is on par with Fisher's linear discriminant analysis. Moreover, the optimal and sub-optimal SDR misclassification rates are smaller than all other methods in Krzanowski (1975), as well as mixed nonparametric kernel methods (Vlachonikolis and Marriott, 1982). Taken all together, our SDR methods for mixed predictors consistently produce targeted data reductions that provide better fit and prediction.

6.2 Governance index application

Considerable social science and economics research is devoted to the construction of indexes for descriptive and predictive purposes (Vyas and Kumaranayake, 2006; Kolenikov and Angeles, 2009; Filmer and Scott, 2012; Merola and Baulch, 2014; Forzani et al., 2018; Duarte et al., 2021). An index is a statistical summary measure of change in a representative group of individual data points. It usually synthesizes the information contained in a set of p variables $\mathbf{X} \in \mathbb{R}^p$ via a linear combination, $\mathbf{R}(\mathbf{X}) = \boldsymbol{\omega}^T\mathbf{X} \in \mathbb{R}$, where $\boldsymbol{\omega}$ is the vector of weights of the composite index.

In this example, we study the impact of governance on economic growth in the twelve South American countries as measured by per capita *Gross Domestic Product* (GDP) using

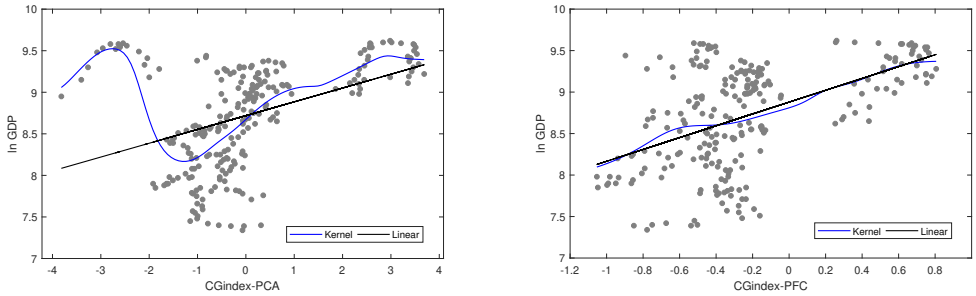

Figure 3: Log of per capita GDP versus Standard `PCA` and `PFC` based composite governance indexes.

the World Bank Governance Indicators.[1] The World Bank considers the following six aggregate indicators of governance that combine the views of a large number of enterprise, citizen and expert survey respondents: control of corruption (X_1); rule of law (X_2); regulatory quality (X_3); government effectiveness (X_4); political stability (X_5); voice and accountability (X_6). They are standardized to have mean zero and standard deviation one, with values from -2.5 to 2.5, approximately, where higher values correspond to better governance. All six are highly positively correlated, and are all positively correlated with the per capita GDP; i.e., economic growth is positively associated with better governance indicators.

Our aim is to build a *Composite Governance index* (CG) to predict Y, the logarithm of per capita *Gross Domestic Product* (GDP), measured in 2010 US dollars, over the period 1996 to 2018. Using the set of governance indicator variables, we start by constructing the CG index via standard Principal Component Analysis (PCA) and Principal Fitted Components (PFC) [see Corollary 3] setting $d = 1$ and $\mathbf{f}_y = \log(GDP)$ in (8).

In the left panel of Figure 3, we plot $\log(GDP)$ versus the CG indexes based on PCA, which is the standard approach in such index construction (Mazziotta and Pareto, 2019). In the right panel of Figure 3, the response is plotted versus the index based on PFC. Both plots indicate dependence of the response on the indexes but the nature of relationship is the data pattern is hard to understand. A linear trend appears stronger in the right panel, which is reflected in the better fit of the linear regression model (black) with $R^2 = 0.27$ versus 0.17 for PCA. However, the PCA-based index in nonparametric kernel regression (blue) results in better fit. Using the `np R` package, the value of the nonparametric version of R^2 is 0.32 for the PFC-based CG index, which is much lower than 0.54, the value for the PCA-based index.

In Figure 4, we plot $\log(GDP)$ versus the PCA and PFC composite governance indexes by country. The plots indicate that the PFC index gives a much better visualization of the relationship of $\log(GDP)$ within each country, suggesting that adjusting the index by country could improve its predictive performance.

We add country effect by introducing eleven binary variables $\mathbf{H}$. In Figure 5 we plot the log of GDP versus the CG index constructed by PCA for mixed variables (PCAmix)

1. Governance Indicators and per capita GDP data can be downloaded from *Worldwide Governance Indicators* and *The World Bank Data*, respectively.

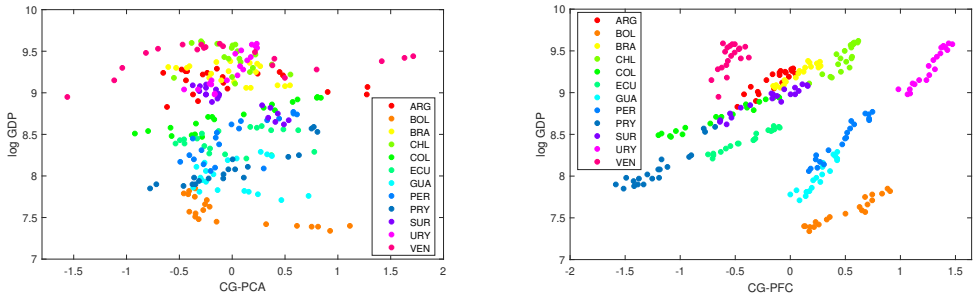

Figure 4: Log of per capita GDP versus standard PCA and standard PFC composite governance indexes by country.

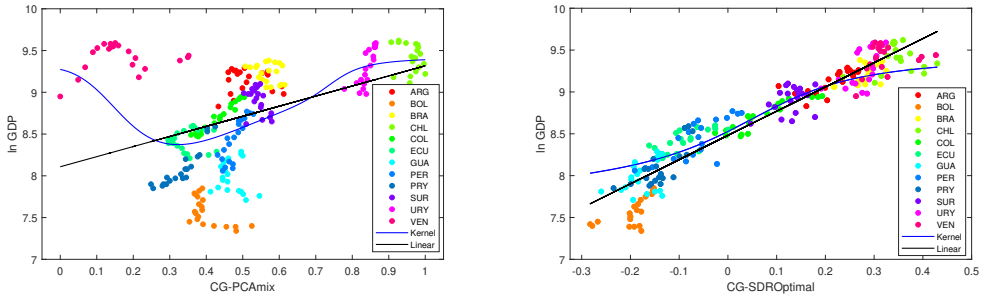

Figure 5: Log of per capita GDP versus Composite Governance index with country effect.

in the left panel and by our mixed OPTIMAL SDRapproach in the right panel. Hardly any difference between the plots in the left panels of Figures 3 and 5 is noticeable. The PCAMIX based CG index is very similar to the conventional PCA based CG that does not include country effect, with R^2 equal to 0.17 and 0.61 for the linear and nonparametric models, respectively. Moreover, neither PCA based CG index exhibits an easy to understand or model relationship with the response.

In contrast, a very clear and simple pattern appears in the right panel of Figure 5, where the response is plotted versus our OPTIMAL SDR based index. The pattern suggests modeling $\log(GDP)$ as a linear function of the GC index. This is a distinct improvement over PCA and PCAMIX (left panels of Figures 3 and 5) but also the SDR method PFC, which does not account for country effect (right panel of Figure 3). As a result, both the linear (black) and the kernel (blue) regression models for the regression of the log per capita GDP on the OPTIMAL SDR for mixed predictors based CG index have excellent fit with respective R^2 values of 0.91 and 0.93.

The average of the leave-one-out mean square prediction errors of the linear and kernel regression models in Table 5, provides an unbiased measure of predictive performance. The logarithm of the per capita GDP is regressed on the unsupervised CG indexes, constructed by PCA using only continuous predictors (PCA($\mathbf{X}$)) and its extension for mixed variables (PCAmix($\mathbf{X}, \mathbf{H}$)), and the supervised CG Indexes, constructed by PFC only on

Index Type	Method	Predictive Model	
		Linear	Non-Parametric
Unsupervised	PCA($\mathbf{X}$)	0.319	0.189
	PCAmix($\mathbf{X}, \mathbf{H}$)	0.320	0.209
Supervised	PFC($\mathbf{X}$)	0.292	0.282
	optimal SDR($\mathbf{X}, \mathbf{H}$)	0.029	0.028
	sub-optimal SDR($\mathbf{X}, \mathbf{H}$)	0.028	0.022

Table 5: Leave-one-out mean squared prediction errors for the per capita log GDP in South-American countries.

continuous predictors (PFC($\mathbf{X}$)) and our mixed predictor SDR methods, optimal SDRand sub-optimal SDR.

The leave-one-out mean squared prediction errors of the supervised PFC based CG index are smaller than both PCA and PCAmix for the linear model, even though PFC does not account for country effect. Nevertheless, when the kernel regression model is fitted, the PCA based index exhibits better performance than PFC. The dramatic drop in prediction error results from using optimal and sub-optimal SDR, as it is between 5 to 9 times smaller than the PCA, PCAmix and PFC errors for both the linear regression and the kernel regression models.

The regularized estimation of the optimal SDR reduction selects all five continuous predictors except for *rule of law*. *Political stability* and *voice and accountability* have the highest weights in the CG index. *Rule of law* is the most correlated with four of the other variables, with correlation coefficient values over 0.80. We stipulate that our method drops it as its relationship with GDP is mostly absorbed by the other four. The binary variables are all selected. That is, our method finds a significant country effect on GDP.

7. Discussion

Our approach falls within *model-based inverse regression* for sufficient dimension reduction (SDR) (Cook, 2007; Cook and Forzani, 2008; Bura and Forzani, 2015; Bura et al., 2016). Model-based SDR requires knowledge of the family of distributions of the inverse predictors in contrast to moment-based SDR, such as SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), or DR (Li and Wang, 2007), that impose conditions on the moments of the marginal distribution of the predictors. Because of this, our approach provides exhaustive identification and statistically *efficient* estimation of sufficient reductions for the conditional distribution of an output given mixed variables that contain all information in the mixed predictors for the output Y.

Furthermore, beyond the context of dimension reduction for the forward regression problem of Y on mixed predictors $\mathbf{Z}$, the modeling we use to accommodate the factorization in (2) in developing our SDR methods, is a new multivariate modeling approach for *response* vectors comprised of mixed variables. That is, if one were to only consider the multivariate regression of the mixed vector $\mathbf{Z} = (\mathbf{X}^T, \mathbf{H}^T)^T$ on some other variables, say $\mathbf{F}$, the models we use for the continuous and binary elements of $\mathbf{Z}$ in our development provide a new regression

tool for mixed responses. Specifically, since the joint distribution of $\mathbf{Z} \mid \mathbf{F}$ belongs to the exponential family (9), our approach yields *sufficient statistics* for the unknown natural parameters $\boldsymbol{\vartheta}$ in (11), as well as optimal (efficient) maximum likelihood estimators, in a similar manner to generalized linear modeling for univariate responses.

Acknowledgments

We would like to thank the associate editor and two anonymous referees for their comments and suggestions that helped us improve the paper. LF, RGA, PML and DT have been partially supported by the UNL grant CAID 503-20190-100022LI and by the ANPCYT grant PICT-2018-03005). EB would like to acknowledge support for this project from the Austrian Science Fund (FWF P 30690-N35) and the Vienna Science and Technology Fund (WWTF ICT19-018).

References

J. Aitchison and C.G.G. Aitken. Multivariate binary discrimination by the kernel method. Biometrika, 63(3):413–420, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335719.

J.A. Anderson. Separate sample logistic discrimination. Biometrika, 59(1):19–35, 1972. ISSN 00063444. URL http://www.jstor.org/stable/2334611.

J.A. Anderson. Quadratic logistic discrimination. Biometrika, 62(1):149–154, 1975. ISSN 00063444. URL http://www.jstor.org/stable/2334497.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. Statist. Sci., 27(4):450–468, 11 2012. URL 10.1214/12-STS394.

M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Scientific Reports, 5(1):2045–2322, 2015. URL https://doi.org/10.1038/srep10312.

E. Bura and R.D. Cook. Estimating the structural dimension of regressions via parametric inverse regression. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63(2):393–410, 2001. ISSN 13697412.

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. URL https://doi.org/10.1080/01621459.2014.914440.

E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction: A unifying approach. Journal of Multivariate Analysis, 102(1):130 – 142, 2011. ISSN 0047-259X. URL https://doi.org/10.1016/j.jmva.2010.08.007.

E. Bura, S. Duarte, and L. Forzani. Sufficient reductions in regressions with exponential family inverse predictors. Journal of the American Statistical Association, 111(515): 1313–1329, 2016. URL https://doi.org/10.1080/01621459.2015.1093944.

S. Camiz and G.C. Gomes. Joint correspondence analysis versus multiple correspondence analysis: a solution to an undetected problem. In Classification and data mining, Stud. Classification Data Anal. Knowledge Organ., pages 11–18. Springer, Heidelberg, 2013. URL https://doi.org/10.1007/978-3-642-28894-4_2.

M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco. Orthogonal rotation in pcamix. Advances in Data Analysis and Classification, 6:131–146, 2012. URL https://doi.org/10.1007/s11634-012-0105-3.

M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Multivariate analysis of mixed data: The r package pcamixdata, 2014.

S. Chen, D.M. Witten, and A. Shojaie. Selection and estimation for mixed graphical models. Biometrika, 102(1):47–64, 12 2014. ISSN 0006-3444. URL https://doi.org/10.1093/biomet/asu051.

J. Cheng, E. Levina, and J. Wang, P.and Zhu. A sparse Ising model with covariates. Biometrics, 70(4):943–953, 2014. ISSN 0006-341X. URL https://doi.org/10.1111/biom.12202.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. Journal of Computational and Graphical Statistics, 26(2):367–378, 2017. URL https://doi.org/10.1080/10618600.2016.1237362.

R.D. Cook. Fisher lecture: Dimension reduction in regression (with discussion). Statistical Science, 22:1–26, 2007. URL https://doi.org/10.1214/088342306000000682.

R.D. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. Statistical Science, 23:485–501, 2008. URL https://doi.org/10.1214/08-STS275.

R.D. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86:328–332, 1991. URL https://doi.org/10.2307/2290564.

Andrew Dahl, Valentina Iotchkova, Amelie Baud, Å sa Johansson, Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis, and Jonathan Marchini. A multiple-phenotype imputation method for genetic studies. Nature Genetics, 48(4):466–472, 2016. URL https://doi.org/10.1038/ng.3513.

B. Dai. Multivariate bernoulli distribution models. Technical report, Dept. Statistics, Univ. Wisconsin, Madison, WI 53706, July 2012. URL http://pages.stat.wisc.edu/~wahba/ftp1/tr1171.pdf.

B. Dai, S. Ding, and G. Wahba. Multivariate bernoulli distribution. Bernoulli, 19(4): 1465–1483, 09 2013. URL 10.3150/12-BEJSP10.

N.E. Day and D.F. Kerridge. A general maximum likelihood discriminant. Biometrics, 23(2): 313–323, 1967. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528164.

S. Duarte, L. Forzani, R. García Arancibia, P. Llop, and D. Tomassi. Socioeconomic index for income and poverty prediction: A sufficient dimension reduction approach. Review of Income and Wealth, 2021. URL https://doi.org/10.1111/roiw.12529.

D. Filmer and K. Scott. Assessing asset indices. Demography, 49:359–392, 2012. URL https://doi.org/10.1007/s13524-011-0077-5.

G.M. Fitzmaurice and N.M. Laird. Regression models for mixed discrete and continuous responses with potentially missing values. Biometrics, 53(1):110–122, 1997. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533101.

L. Forzani, R. García-Arancibia, P. Llop, and D. Tomassi. Supervised dimension reduction for ordinal predictors. Computational Statistics and Data Analysis, 125, 2018. URL https://doi.org/10.1016/j.csda.2018.03.018.

L. Forzani, D. Rodriguez, E.and Smucler, and M. Sued. Sufficient dimension reduction and prediction in regression: Asymptotic results. Journal of Multivariate Analysis, 171 (C):339–349, 2019. doi: 10.1016/j.jmva.2018.12.00. URL https://ideas.repec.org/a/eee/jmvana/v171y2019icp339-349.html.

M. Greenacre and J. Blasius, editors. Multiple correspondence analysis and related methods. Statistics in the Social and Behavioral Sciences Series. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-628-0; 1-58488-628-5. URL 10.1201/9781420011319.

X. Huang, X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, M. Li, D. Fan, Y. Guo, A. Wang, L. Wang, L. Deng, W. Li, Y. Lu, Q. Weng, K. Liu, and B. Han. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics, 42(11):961–967, 2010. URL https://doi.org/10.1038/ng.695.

E. Ising. Beitrag zur theorie des ferromagnetismus. Zeitschrift für Physik, 31(1):253–258, Feb 1925. ISSN 0044-3328. URL https://doi.org/10.1007/BF02980577.

K.N. Javaras and D.A. Van Dyk. Multiple imputation for incomplete data with semicontinuous variables. Journal of the American Statistical Association, 98(463):703–715, 2003. ISSN 01621459. URL http://www.jstor.org/stable/30045298.

S. Kolenikov and G. Angeles. Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? The Review of Income and Wealth, 55(1):128–165, 2009. URL https://doi.org/10.1111/j.1475-4991.2008.00309.x.

W.J. Krzanowski. Discrimination and classification using both binary and continuous variables. Journal of the American Statistical Association, 70(352):782–790, 1975. URL https://doi.org/10.2307/2285437.

W.J. Krzanowski. The location model for mixtures of categorical and continuous variables. Journal of Classification, 10(1):25–49, Jan 1993. ISSN 1432-1343. URL https://doi.org/10.1007/BF02638452.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Statist., 17(1):31–57, 03 1989. URL 10.1214/aos/1176347003.

S.L. Lauritzen. Graphical Models. Oxford University Press, Oxford, 1996.

J.D. Lee and Trevor J.H. Learning the structure of mixed graphical models. Journal of Computational and Graphical Statistics, 24(1):230–253, 2015. URL https://doi.org/10.1080/10618600.2014.900500.

B. Li and S. Wang. On directional regression for dimension reduction. Journal of the American Statistical Association, 102(479):997–1008, 2007. URL https://doi.org/10.1198/016214507000000536.

K.C. Li. Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association, 86:316–342, 1991. URL https://doi.org/10.2307/2290563.

J. Liu and J. Ye. Fast overlapping group lasso. 2010. URL arXiv:1009.0306v1.

J.W. Martin and I.J. Michael. Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning, 1(1–2):1–305, 2008. ISSN 1935-8237. URL http://dx.doi.org/10.1561/2200000001.

M. Mazziotta and A. Pareto. Use and misuse of pca for measuring well-being. Social Indicators Research, 142(2):451–476, Apr 2019. ISSN 1573-0921. URL 10.1007/s11205-018-1933-0.

G. Merola and B. Baulch. Using sparse categorical principal components to estimate asset indices new methods with an application to rural south east asia. 2014. URL https://doi.org/10.1111/rode.12568.

C.N. Morris. Natural Exponential Families. American Cancer Society, 2006. ISBN 9780471667193. URL 10.1002/0471667196.ess1759.pub2.

I. Olkin and R.F. Tate. Multivariate correlation models with mixed discrete and continuous variables. Ann. Math. Statist., 32(2):448–465, 06 1961. URL 10.1214/aoms/1177705052.

M.S. Pepe. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York, 2003.

S. L. Simpson, M. Bahrami, and P. J. Laurienti. A mixed-modeling framework for analyzing multitask whole-brain network data. Network Neuroscience, 3(2):307–324, 2019. URL https://doi.org/10.1162/netn_a_00065.

S. Van Buuren. Flexible imputation of missing data. CRC Press, 2nd edition, 2018.

I.G. Vlachonikolis and F.H.C. Marriott. Discrimination with mixed binary and continuous data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(1):23–31, 1982. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2347071.

S. Vyas and L. Kumaranayake. Constructing socio-economic status indices: How to use principal components analysis. Health Policy and Planning, 21(6):459–468, 2006. URL http://dx.doi.org/10.1093/heapol/czl029.

J. Whittaker. Graphical Models in Applied Multivariate Statistics. Wiley, 2009.

E. Yang, Y. Baker, P. Ravikumar, G. Allen, and Z. Liu. Mixed Graphical Models via Exponential Families. In Samuel Kaski and Jukka Corander, editors, Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, volume 33 of Proceedings of Machine Learning Research, pages 1042–1050, Reykjavik, Iceland, 22–25 Apr 2014a. PMLR. URL http://proceedings.mlr.press/v33/yang14a.html.

E. Yang, P. Ravikumar, G.I Allen, Y. Baker, Y.-W. Wan, and Z. Liu. A general framework for mixed graphical models, 2014b.

E. Yang, P. Ravikumar, G.I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. Journal of Machine Learning Research, 16(115):3813–3847, 2015. URL http://jmlr.org/papers/v16/yang15a.html.

K. Ye and L.-H. Lim. Schubert varieties and distances between subspaces of different dimensions. SIAM Journal on Matrix Analysis and Applications, 37(3):1176–1197, 2016. URL https://doi.org/10.1137/15M1054201.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68:49–67, 2006. URL https://doi.org/10.1111/j.1467-9868.2005.00532.x.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. Biometrika, 94(1):19–35, 2007. ISSN 00063444. URL http://www.jstor.org/stable/20441351.

Jin Zhang, Min Chen, Yangjun Wen, Yin Zhang, Yunan Lu, Shengmeng Wang, and Juncong Chen. A fast multi-locus ridge regression algorithm for high-dimensional genome-wide association studies. Frontiers in Genetics, 12:396, 2021. ISSN 1664-8021. doi: 10.3389/fgene.2021.649196. URL https://www.frontiersin.org/article/10.3389/fgene.2021.649196.

Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, Jianming Yu, Donna K. Arnett, Jose M. Ordovas, and Edward S. Buckler. Multivariate mean parameter estimation by using a partly exponential model. Nature Genetics, 42(4):355–360, 2010. ISSN 00359246. URL http://www.jstor.org/stable/2345860.

Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen. Effective supervised discretization for classification based on correlation maximization. IEEE International Conference on Information Reuse & Integration, pages 390–395, 2011. URL https://doi.org/10.1109/IRI.2011.6009579.

A. Proofs and Derivations for Section 3

A.1 Derivation of Eqn. (9)

From Eqn. (8), $f(\mathbf{X}, \mathbf{H} \mid Y = y)$ up to the constant $1/\sqrt{2\pi}$ is

$$\exp\Big\{ -\frac{1}{2}\Big((\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H}-\boldsymbol{\mu}_{\mathbf{H}})\Big)^T \boldsymbol{\Delta}^{-1}\Big((\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}) - \mathbf{A}\mathbf{f}_y - \boldsymbol{\beta}(\mathbf{H}-\boldsymbol{\mu}_{\mathbf{H}})\Big) + \text{vech}^T(\mathbf{H}\mathbf{H}^T)\,(\boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y) + \frac{1}{2}\log(|\boldsymbol{\Delta}|^{-1}) - \log(G(\boldsymbol{\Gamma}_y))\Big\}.$$

After some algebra and rearrangement of terms, we obtain

$$f(\mathbf{X}, \mathbf{H} \mid Y = y) = h(\mathbf{X}, \mathbf{H}) \exp\left(\mathbf{T}^T(\mathbf{X}, \mathbf{H})\boldsymbol{\eta}_y - \psi(\boldsymbol{\eta}_y)\right),$$

with $h(\mathbf{X}, \mathbf{H}) = (2\pi)^{-1/2}$,

$$\begin{aligned}\mathbf{T}^T(\mathbf{X}, \mathbf{H})\boldsymbol{\eta}_y = {} & \mathbf{X}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \mathbf{X}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + \mathbf{X}^T\boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y \\ & - \mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - \mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y \\ & - \frac{1}{2}\mathbf{X}^T\boldsymbol{\Delta}^{-1}\mathbf{X} + \mathbf{X}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\mathbf{H} \\ & - \frac{1}{2}\mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\mathbf{H} + \text{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}_0 + \text{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}\mathbf{f}_y, \end{aligned} \tag{49}$$

and

$$\begin{aligned}\psi(\boldsymbol{\eta}_y) = {} & \frac{1}{2}\boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \frac{1}{2}\mathbf{f}_y^T\mathbf{A}^T\boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y + \frac{1}{2}\boldsymbol{\mu}_{\mathbf{H}}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} \\ & + \boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y - \boldsymbol{\mu}_{\mathbf{X}}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - \boldsymbol{\mu}_{\mathbf{H}}^T\beta^T\boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y \\ & - \frac{1}{2}\log(|\boldsymbol{\Delta}|^{-1}) + \log(G(\boldsymbol{\Gamma}_y)).\end{aligned} \tag{50}$$

Since $\text{tr}(\mathbf{A}^T\mathbf{B}) = \text{vec}(\mathbf{A})^T\text{vec}(\mathbf{B})$, $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B})$ and $\mathbf{D}_q$ in Section 2 is such that $\text{vec}(\mathbf{A}) = \mathbf{D}_q\,\text{vech}(\mathbf{A})$, (49) equals

$$\begin{aligned}\mathbf{T}^T(\mathbf{X}, \mathbf{H})\,\boldsymbol{\eta}_y = {} & \mathbf{X}^T\left(\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + (\mathbf{f}_y^T \otimes \mathbf{I}_p)\text{vec}(\boldsymbol{\Delta}^{-1}\mathbf{A})\right) \\ & + \mathbf{H}^T\left(-\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - (\mathbf{f}_y^T \otimes \mathbf{I}_q)\text{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A})\right) \\ & - \frac{1}{2}(\mathbf{D}_p\mathbf{D}_p^T\,\text{vech}(\mathbf{X}\mathbf{X}^T))^T\,\text{vech}(\boldsymbol{\Delta}^{-1}) + \text{vec}(\mathbf{X}\mathbf{H}^T)^T\text{vec}(\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) \\ & + \text{vech}(\mathbf{H}\mathbf{H}^T)^T\left(-\frac{1}{2}\mathbf{D}_q^T\text{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \boldsymbol{\tau}_0 + (\mathbf{f}_y^T \otimes \mathbf{I}_{q(q+1)/2})\text{vec}(\boldsymbol{\tau})\right).\end{aligned}$$

Using the matrices $\mathbf{J}_q$ and $\mathbf{L}_q$ defined in Section 2, we obtain Eqns (10) and (11) from

$$\begin{aligned}\mathbf{T}^T(\mathbf{X}, \mathbf{H})\boldsymbol{\eta}_y = {} & \mathbf{X}^T\left(\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + (\mathbf{f}_y^T \otimes \mathbf{I}_p)\text{vec}(\boldsymbol{\Delta}^{-1}\mathbf{A})\right) \\ & + \mathbf{H}^T\left(-\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - (\mathbf{f}_y^T \otimes \mathbf{I}_q)\text{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A})\right) \\ & - \frac{1}{2}(\mathbf{D}_p\mathbf{D}_p^T\,\text{vech}(\mathbf{X}\mathbf{X}^T))^T\,\text{vech}(\boldsymbol{\Delta}^{-1}) + \text{vec}(\mathbf{X}\mathbf{H}^T)^T\text{vec}(\boldsymbol{\Delta}^{-1}\boldsymbol{\beta})\end{aligned}$$

$$
\begin{aligned}
&+ \mathbf{H}^T \left(-\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{L}_q\boldsymbol{\tau}_0 + (\mathbf{f}_y^T \otimes \mathbf{I}_q)\mathrm{vec}(\mathbf{L}_q\boldsymbol{\tau}) \right) \\
&+ (\mathbf{J}_q\,\mathrm{vech}(\mathbf{H}\mathbf{H}^T))^T \left(-\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{J}_q\boldsymbol{\tau}_0 + (\mathbf{f}_y^T \otimes \mathbf{I}_{k_q})\mathrm{vec}(\mathbf{J}_q\boldsymbol{\tau}) \right) \\
&= \mathbf{X}^T\boldsymbol{\eta}_{y1} + \mathbf{H}^T\boldsymbol{\eta}_{y2} - \frac{1}{2}(\mathbf{D}_p^T\mathbf{D}_p\mathrm{vec}(\mathbf{X}\mathbf{X}^T))^T\boldsymbol{\eta}_3 \\
&\qquad + \mathrm{vec}(\mathbf{X}\mathbf{H}^T)^T\boldsymbol{\eta}_4 + (\mathbf{J}_q\,\mathrm{vech}(\mathbf{H}\mathbf{H}^T))^T\boldsymbol{\eta}_{y5},
\end{aligned}
$$

where

$$\boldsymbol{\eta}_{y1} = \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + (\mathbf{f}_y^T \otimes \mathbf{I}_p)\mathrm{vec}(\boldsymbol{\Delta}^{-1}\mathbf{A}) = \mathbf{F}_{y1}\boldsymbol{\vartheta}_1,$$

with $\mathbf{F}_{y1} = (\mathbf{I}_p, \mathbf{f}_y^T \otimes \mathbf{I}_p)$, $\boldsymbol{\vartheta}_1 = (\boldsymbol{\vartheta}_{10}^T, \boldsymbol{\vartheta}_{11}^T)^T$, $\boldsymbol{\vartheta}_{10} = \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}$ and $\boldsymbol{\vartheta}_{11} = \mathrm{vec}(\boldsymbol{\Delta}^{-1}\mathbf{A})$,

$$
\begin{aligned}
\boldsymbol{\eta}_{y2} &= -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - (\mathbf{f}_y^T \otimes \mathbf{I}_q)\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}) \\
&\quad - \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{L}_q\boldsymbol{\tau}_0 + (\mathbf{f}_y^T \otimes \mathbf{I}_q)\mathrm{vec}(\mathbf{L}_q\boldsymbol{\tau}) \\
&= \mathbf{F}_{y2}\boldsymbol{\vartheta}_2,
\end{aligned}
$$

where $\mathbf{F}_{y2} = (\mathbf{I}_q, \mathbf{f}_y^T \otimes \mathbf{I}_q)$, $\boldsymbol{\vartheta}_2 = (\boldsymbol{\vartheta}_{20}^T, \boldsymbol{\vartheta}_{21}^T)^T$, with $\boldsymbol{\vartheta}_{20} = -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + \mathbf{L}_q\boldsymbol{\tau}_0 - \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta})$ and $\boldsymbol{\vartheta}_{21} = \mathrm{vec}(\mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A})$,

$$
\begin{aligned}
\boldsymbol{\eta}_3 &\coloneqq \boldsymbol{\eta}_{3y} = \mathrm{vech}(\boldsymbol{\Delta}^{-1}), \\
\boldsymbol{\eta}_4 &\coloneqq \boldsymbol{\eta}_{4y} = \mathrm{vec}(\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}),
\end{aligned}
$$

and,

$$
\begin{aligned}
\boldsymbol{\eta}_{y5} &= -\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{J}_q\boldsymbol{\tau}_0 + (\mathbf{f}_y^T \otimes \mathbf{I}_{k_q})\mathrm{vec}(\mathbf{J}_q\boldsymbol{\tau}) \\
&\coloneqq \mathbf{F}_{y5}\boldsymbol{\vartheta}_5,
\end{aligned}
$$

with $\mathbf{F}_{y5} = (\mathbf{I}_{k_q}, \mathbf{f}_y^T \otimes \mathbf{I}_{k_q})$ and $\boldsymbol{\vartheta}_5 = (\boldsymbol{\vartheta}_{50}^T, \boldsymbol{\vartheta}_{51}^T)^T$, $\boldsymbol{\vartheta}_{50} = -\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathbf{J}_q\boldsymbol{\tau}_0$ and $\boldsymbol{\vartheta}_{51} = \mathrm{vec}(\mathbf{J}_q\boldsymbol{\tau})$.

Eqn. (7) yields

$$G(\boldsymbol{\Gamma}_y) = \sum_H \exp\left[\,\mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\,(\boldsymbol{\tau}_0 + \boldsymbol{\tau}\mathbf{f}_y)\,\right].$$

Therefore, using $\mathbf{J}_q$, $\mathbf{L}_q$, $\mathbf{D}_q$ and $\mathbf{C}_q$ defined in Section 2, the notation defined above and the new notation $\bar{\boldsymbol{\eta}}_4 \coloneqq \mathrm{unvec}(\boldsymbol{\eta}_4)$, we get

$$
\begin{aligned}
G(\boldsymbol{\Gamma}_y) = \sum_H \exp\Bigg[&(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T \left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4) \right) \\
&+ \mathbf{H}^T \left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4) \right) \Bigg]
\end{aligned}
\tag{51}
$$

Finally, from Eqn. (50) using the matrix $\mathbf{D}_p$ defined in Section 2 and (51),

$$
\begin{aligned}
\psi(\boldsymbol{\eta}_y) &= \frac{1}{2}\boldsymbol{\eta}_{y1}^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \log G(\boldsymbol{\Gamma}_y) - \frac{1}{2}\log|\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3)| \\
&\coloneqq \psi_1(\boldsymbol{\eta}_y) + \psi_2(\boldsymbol{\eta}_y) + \psi_3(\boldsymbol{\eta}_y).
\end{aligned}
\tag{52}
$$

A.2 Proof of Theorem 1

Since the density of $(\mathbf{X}, \mathbf{H}) \mid Y$ belongs to the exponential family (Eqn. (9)), following Theorem 1 in Bura et al. (2016) we have that, the minimal sufficient reduction for the regression $Y \mid (\mathbf{X}, \mathbf{H})$ is given by

$$\mathbf{R}(\mathbf{X}, \mathbf{H}) = \boldsymbol{\alpha}_{\mathbf{a}}^T \left(\mathbf{T}(\mathbf{X}, \mathbf{H}) - \mathrm{E}(\mathbf{T}(\mathbf{X}, \mathbf{H})) \right),$$

with $\boldsymbol{\alpha}_{\mathbf{a}}$ is a basis for $\mathcal{S}_{\mathbf{a}} = \mathrm{span}\left\{ \boldsymbol{\eta}_Y - \mathrm{E}(\boldsymbol{\eta}_Y), Y \in \mathcal{Y} \right\}$, with $\boldsymbol{\eta}_Y$ given in (11). Therefore, from Eqns. (11) and (12) and since $\mathrm{E}(\mathbf{f}_Y) = \mathbf{0}$,

$$\boldsymbol{\eta}_y - \mathrm{E}(\boldsymbol{\eta}_y) = \begin{pmatrix} (\mathbf{f}_y^T \otimes \mathbf{I}_p)\mathrm{vec}(\boldsymbol{\Delta}^{-1}\mathbf{A}) \\ (\mathbf{f}_y^T \otimes \mathbf{I}_q)\mathrm{vec}(\mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}) \\ 0 \\ 0 \\ (\mathbf{f}_y^T \otimes \mathbf{I}_{k_q})\mathrm{vec}(\mathbf{J}_q\boldsymbol{\tau}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A}\mathbf{f}_y \\ (\mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A})\mathbf{f}_y \\ 0 \\ 0 \\ \mathbf{J}_q\boldsymbol{\tau}\mathbf{f}_y \end{pmatrix}.$$

Thus, $\mathrm{span}\left\{ \boldsymbol{\eta}_Y - \mathrm{E}(\boldsymbol{\eta}_Y), Y \in \mathcal{Y} \right\} = \mathrm{span}(\mathbf{a})$ with

$$\mathbf{a} = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}A \\ 0 \\ 0 \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix}.$$

A.3 Proof of Corollary 3

If follows from Corollary 2, since, in this case, $\boldsymbol{\vartheta}_{2,1} = 0$ and $\boldsymbol{\vartheta}_{5,1} = 0$.

A.4 Proof of Corollary 4

If follows from Corollary 2 since in this case $\boldsymbol{\vartheta}_{1,1} = 0$ and $\boldsymbol{\vartheta}_{2,1} = \mathbf{L}_q\boldsymbol{\tau}$.

A.5 Proof of Corollary 5

It suffices to show that $\mathrm{span}(\mathbf{b}) \subset \mathrm{span}(\boldsymbol{\alpha}_{\mathbf{c}})$. We observe that $\mathbf{b}$ can be written as

$$\mathbf{b} = \begin{pmatrix} \boldsymbol{\Delta}\mathbf{A} & 0 \\ -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} & \mathbf{L}_1\boldsymbol{\tau} \\ 0 & \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix} \begin{pmatrix} \mathbf{I}_r \\ \mathbf{I}_r \end{pmatrix} \coloneqq \widetilde{\mathbf{b}} \begin{pmatrix} \mathbf{I}_r \\ \mathbf{I}_r \end{pmatrix},$$

with

$$\mathrm{span}(\widetilde{\mathbf{b}}) = \mathrm{span}(\boldsymbol{\alpha}_{\mathbf{c}}).$$

As a consequence, $\mathrm{span}(\mathbf{b}) \subset \mathrm{span}(\boldsymbol{\alpha}_{\mathbf{c}})$, and therefore $\mathbf{R}(\mathbf{X}, \mathbf{H})$ in (23) is a sufficient dimension reduction, not necessarily minimal. The rest of the corollary follows immediately.

B. Proof of Proposition 6

In order to prove Proposition 6, we have to first study the asymptotic distribution of $\widehat{\mathbf{b}}$ in (34) in Section B.1 and prove some auxiliary lemmas in Section B.2.

B.1 Asymptotic distribution of $\widehat{\mathbf{b}}$

Proposition 7 *$\sqrt{n}\mathrm{vec}(\widehat{\mathbf{b}}-\mathbf{b}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{V}_{rcl})$ with*

$$\mathbf{V}_{rcl} = \mathbf{W}\mathbf{M}\mathbf{V}\mathbf{M}^T\mathbf{W}^T,$$

where $\mathbf{M}$, $\mathbf{W}$ and $\mathbf{V}$ are defined in Eqns. (44), (45) and (42), respectively.

Proof of Proposition 7. To obtain the asymptotic distribution of $\widehat{\mathbf{b}}$, we rewrite

$$\mathbf{b} = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{L}_q\boldsymbol{\tau} - \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A} \\ \mathbf{J}_q\boldsymbol{\tau} \end{pmatrix} = \begin{pmatrix} \mathrm{unvec}(\boldsymbol{\vartheta}_{1,1}) \\ \mathrm{unvec}(\boldsymbol{\vartheta}_{2,1}) \\ \mathrm{unvec}(\boldsymbol{\vartheta}_{5,1}) \end{pmatrix},$$

as follows. Letting $\widetilde{\mathbf{b}} \coloneqq \left(\boldsymbol{\vartheta}_{1,1}^T, \boldsymbol{\vartheta}_{2,1}^T, \boldsymbol{\vartheta}_{5,1}^T\right)^T$, $\widetilde{\mathbf{b}} = \mathbf{M}\boldsymbol{\vartheta}$, with $\mathbf{M}$ given in (44), which implies that

$$\mathrm{vec}(\mathbf{b}) = \mathbf{W}\widetilde{\mathbf{b}} = \mathbf{W}\mathbf{M}\boldsymbol{\vartheta}, \tag{53}$$

with $\mathbf{W}$ defined in (45). Then, from Eqn. (53) applied to $\widehat{\mathbf{b}}$,

$$\mathrm{vec}(\widehat{\mathbf{b}}) = \mathbf{W}\mathbf{M}\widehat{\boldsymbol{\vartheta}}. \tag{54}$$

To compute the asymptotic distribution of $\mathrm{vec}(\widehat{\mathbf{b}})$ we need to first obtain the asymptotic distribution of of $\widehat{\boldsymbol{\vartheta}}$, which is stated in the next lemma.

Lemma 8 *If* $\mathrm{avar}\left(\sqrt{n}\,\widehat{\boldsymbol{\vartheta}}\right) = \mathbf{V}$*, then*

$$\mathbf{V}^{-1} = \mathrm{E}\left(\mathbf{F}_y^T\mathbf{J}\mathbf{F}_y\right),$$

where $\mathbf{F}_y$ is defined in (11) and $\mathbf{J}$ is the matrix of partial derivatives given by (43)

$$\mathbf{J} = \frac{\partial^2\psi(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_y\partial\boldsymbol{\eta}_y^T}.$$

The asymptotic normality of $\widehat{\mathbf{b}}$ follows from Lemma 8. Its asymptotic variance is

$$\begin{aligned}\mathrm{avar}(\sqrt{n}\,\widehat{\mathbf{b}}) &= \mathbf{W}\mathbf{M}\mathrm{avar}\left(\sqrt{n}\,\widehat{\boldsymbol{\vartheta}}\right)\mathbf{M}^T\mathbf{W}^T \\ &= \mathbf{W}\mathbf{M}\mathbf{V}\mathbf{M}^T\mathbf{W}^T,\end{aligned}$$

as stated in Proposition 7.

Proof of Lemma 8. Since $\widehat{\boldsymbol{\vartheta}}$ is the maximum likehood estimator,

$$\mathbf{V} = \mathrm{avar}\left(\sqrt{n}\,\widehat{\boldsymbol{\vartheta}}\right) = -\left(\mathrm{E}\left[\frac{\partial^2\log f(\mathbf{X},\mathbf{H}\mid Y=y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]\right)^{-1}.$$

Plugging in Eqn. (9) $\boldsymbol{\eta}_y = \mathbf{F}_y\boldsymbol{\vartheta}$ from Eqn. (11) obtains

$$\log f(\mathbf{X},\mathbf{H}\mid Y=y) = \log h(\mathbf{X},\mathbf{H}) + \mathbf{T}^T(\mathbf{X},\mathbf{H})\boldsymbol{\eta}_y - \psi(\boldsymbol{\eta}_y)$$

$$= \log h(\mathbf{X}, \mathbf{H}) + \mathbf{T}^T(\mathbf{X}, \mathbf{H})\mathbf{F}_y\boldsymbol{\vartheta} - \psi(\mathbf{F}_y\boldsymbol{\vartheta}).$$

From (8), it follows that

$$\begin{aligned}
\frac{\partial \log f(\mathbf{X}, \mathbf{H} \mid Y = y)}{\partial \mathrm{vec}^T(\boldsymbol{\theta})} &= \mathbf{T}^T(\mathbf{X}, \mathbf{H})\mathbf{F}_y - \frac{\partial \psi(\boldsymbol{\eta}_y)}{\partial \boldsymbol{\eta}_y^T}\mathbf{F}_y, \\
\frac{\partial^2 \log f(\mathbf{X}, \mathbf{H} \mid Y = y)}{\partial \mathrm{vec}(\boldsymbol{\theta})\mathrm{vec}^T(\partial\boldsymbol{\theta})} & \\
&= -\mathbf{F}_y^T \frac{\partial^2 \psi(\boldsymbol{\eta}_y)}{\partial \boldsymbol{\eta}_y \partial \boldsymbol{\eta}_y^T}\mathbf{F}_y \\
&= -\mathbf{F}_y^T \mathbf{J} \mathbf{F}_y,
\end{aligned}$$

and therefore $\mathbf{V}^{-1} = \mathrm{E}\left(\mathbf{F}_y^T \mathbf{J} \mathbf{F}_y\right)$. ■

In order to compute $\mathbf{J}$, the first and second derivatives of $\psi(\boldsymbol{\eta}_y)$ with respect to $\boldsymbol{\eta}_y$ are required. This computation is carried out in Appendix E.

B.2 Auxiliary lemmas to prove Proposition 6

Lemma 9 *Let* $\widehat{\mathbf{H}} = \widehat{\mathbf{U}}_1\widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{K}^{-1}$. *Then,*

$$\sqrt{n}\mathrm{vec}(\widehat{\mathbf{H}} - \mathbf{U}_1) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, (\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_m)\mathbf{V}_{rlc}(\mathbf{R}_1\mathbf{K}^{-1} \otimes \mathbf{I}_m)\right),$$

where $\mathbf{V}_{rlc}$ *is defined in Eqn. (41),* $\widehat{\mathbf{U}}_1, \widehat{\mathbf{K}}_1$ *and* $\widehat{\mathbf{R}}_1$ *in Eqn. (18),* $\mathbf{U}_1, \mathbf{K}$ *and* $\mathbf{R}_1$ *in Eqn. (32).*

Proof By Eqn. (32), $\mathbf{b} = \mathbf{U}_1\mathbf{K}\mathbf{R}_1^T$, and by Eqn. (18), $\widehat{\mathbf{b}} = \widehat{\mathbf{U}}_1\widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T$. Then,

$$\begin{aligned}
\widehat{\mathbf{H}} - \mathbf{U}_1 &= \widehat{\mathbf{U}}_1\widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{K}^{-1} - \mathbf{U}_1 \\
&= \widehat{\mathbf{b}}\mathbf{R}_1\mathbf{K}^{-1} - \widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1} - \mathbf{U}_1 \\
&= (\widehat{\mathbf{b}} - \mathbf{b})\mathbf{R}_1\mathbf{K}^{-1} - \widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1},
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{n}\mathrm{vec}(\widehat{\mathbf{H}} - \mathbf{U}_1) &= \sqrt{n}\mathrm{vec}((\widehat{\mathbf{b}} - \mathbf{b})\mathbf{R}_1\mathbf{K}^{-1} - \widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1}) \\
&= \sqrt{n}(\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_m)\mathrm{vec}(\widehat{\mathbf{b}} - \mathbf{b}) - \mathrm{vec}(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1}). \qquad (55)
\end{aligned}$$

From Proposition 7,

$$\sqrt{n}(\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_m)\mathrm{vec}(\widehat{\mathbf{b}} - \mathbf{b}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}}), \qquad (56)$$

with $\boldsymbol{\Sigma}_{\mathbf{U}} = (\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V}_{rcl}(\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)^T = (\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V}_{rcl}(\mathbf{R}_1\mathbf{K}^{-1} \otimes \mathbf{I}_k)$. Also, since $\sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\right) = O_p(1)$ and $\mathbf{P}_{\mathbf{R}_1} = \mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})$,

$$\begin{aligned}
\sqrt{n}(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1}) &= \sqrt{n}(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T)\mathbf{P}_{\mathbf{R}_1}\mathbf{R}_1\mathbf{K}^{-1} \\
&= \sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\right)\left(\mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})\right)\mathbf{R}_1\mathbf{K}^{-1}
\end{aligned}$$

$$= \sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\right) O_p(n^{-1/2})\mathbf{R}_1\mathbf{K}^{-1}$$
$$= O_p(n^{-1/2}),$$

where we use that $\widehat{\mathbf{R}}_0^T\widehat{\mathbf{R}}_1 = \mathbf{0}$. As a consequence, $\sqrt{n}\,\mathrm{vec}(\widehat{\mathbf{U}}_0\widehat{\mathbf{K}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{K}^{-1}) \to 0$ in probability which, together with (56) in (55), obtain the result. ■

Lemma 10 *Let $\mathbf{\Gamma}$ be a matrix of dimension $p \times d$ with $d \leq p$ of full rank d and let $\mathbf{P}_{\mathbf{\Gamma}}$ be the orthogonal projection onto the columns of $\mathbf{\Gamma}$. Then,*

$$\frac{\partial \mathbf{P}_{\mathbf{\Gamma}}}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} = (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{Q}_{\mathbf{\Gamma}}). \tag{57}$$

Proof For a matrix $\mathbf{X}$ and $\mathbf{F}(\mathbf{X}) : m \times p$, $\mathbf{G}(\mathbf{X}) : p \times q$ differentiable functions of $\mathbf{X}$,

$$\frac{\partial \mathrm{vec}\,(\mathbf{F}(\mathbf{X})\mathbf{G}(\mathbf{X}))}{\partial \mathrm{vec}^T(\mathbf{X})} = (\mathbf{G}^T \otimes \mathbf{I}_m)\frac{\partial \mathrm{vec}\,(\mathbf{F}(\mathbf{X}))}{\partial \mathrm{vec}^T(\mathbf{X})} + (\mathbf{I}_q \otimes \mathbf{F})\frac{\partial \mathrm{vec}\,(\mathbf{G}(\mathbf{X}))}{\partial \mathrm{vec}^T(\mathbf{X})}. \tag{58}$$

For $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T$ and $\mathbf{G}(\mathbf{X}) = \mathbf{X}$ with $\mathbf{X} : p \times q$, applying (58) gives

$$\begin{aligned} \frac{\partial \mathrm{vec}(\mathbf{X}^T\mathbf{X})}{\partial^T \mathrm{vec}(\mathbf{X})} &= (\mathbf{I}_{q^2} + \mathbf{K}_{qq})(\mathbf{I}_q \otimes \mathbf{X}^T). \\ \frac{\partial \mathrm{vec}(\mathbf{X}^T\mathbf{X})^{-1}}{\partial^T \mathrm{vec}(\mathbf{X})} &= -((\mathbf{X}^T\mathbf{X})^{-1} \otimes (\mathbf{X}^T\mathbf{X})^{-1}))\frac{\partial \mathrm{vec}(\mathbf{X}^T\mathbf{X})}{\partial^T \mathrm{vec}(\mathbf{X})}. \end{aligned} \tag{59}$$

Applying (58) we have

$$\begin{aligned} \frac{\partial \mathrm{vec}\mathbf{P}_{\mathbf{\Gamma}}}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} &= \frac{\partial \mathrm{vec}(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} \\ &= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p)\frac{\partial \mathrm{vec}(\mathbf{\Gamma})}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} + (\mathbf{I}_p \otimes \mathbf{\Gamma})\frac{\partial \mathrm{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} \\ &= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \mathbf{\Gamma})\frac{\partial \mathrm{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \mathrm{vec}^T(\mathbf{\Gamma})}. \end{aligned}$$

Let

$$\mathbf{H} = \frac{\partial \mathrm{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \mathrm{vec}^T(\mathbf{\Gamma})}.$$

Then, by (58), (59) and (60),

$$\begin{aligned} \mathbf{H} &= (\mathbf{\Gamma} \otimes \mathbf{I}_d)\frac{\partial \mathrm{vec}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} + (\mathbf{I}_p \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{pd} \\ &= -(\mathbf{\Gamma} \otimes \mathbf{I}_d)((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}))(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{I}_d \otimes \Gamma^T) + (\mathbf{I}_p \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{pd}, \end{aligned}$$

which, in turn, yields

$$\frac{\partial \mathrm{vec}\mathbf{P}_{\mathbf{\Gamma}}}{\partial \mathrm{vec}^T(\mathbf{\Gamma})} = (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \mathbf{\Gamma})\big[- (\mathbf{\Gamma} \otimes \mathbf{I}_d)((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{I}_d \otimes \mathbf{\Gamma}^T)$$

$$
\begin{aligned}
& +(\mathbf{I}_p \otimes (\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1})\mathbf{K}_{pd}\big] \\
= \; & (\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \boldsymbol{\Gamma}(\boldsymbol{\Gamma}\boldsymbol{\Gamma})^{-1})\mathbf{K}_{pd} - (\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T) \\
& -(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1})\mathbf{K}_{dd}(\mathbf{I}_d \otimes \boldsymbol{\Gamma}^T) \\
= \; & (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{I}_p) - (\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T) \\
& -(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1})(\boldsymbol{\Gamma}^T \otimes \mathbf{I}_d)\mathbf{K}_{pd} \\
= \; & (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{I}_p) - (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{P}_{\boldsymbol{\Gamma}}) \\
= \; & (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{I}_p - \mathbf{P}_{\boldsymbol{\Gamma}}) \\
= \; & (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{Q}_{\boldsymbol{\Gamma}}).
\end{aligned}
$$

■

Lemma 11 *Assume that the $p \times d$, $d \leq p$, matrix $\widehat{\boldsymbol{\Gamma}}$ is asymptotically normal,*

$$\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{V}),$$

for a $p \times d$ matrix $\boldsymbol{\Gamma}$ of rank d. Then, $\sqrt{n}\mathrm{vec}(\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}})$ is asymptotically normal with mean 0 and variance-covariance matrix

$$(\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{Q}_{\boldsymbol{\Gamma}})\mathbf{V}((\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T \otimes \mathbf{Q}_{\boldsymbol{\Gamma}})(\mathbf{I}_{p^2} + \mathbf{K}_{pp}).$$

Proof Let $\mathbf{P}_{\boldsymbol{\Gamma}}$ be the projection onto the columns of $\boldsymbol{\Gamma}$ defined as $\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T$ and let g be the function defined in the subspace of the matrices $p \times d$ of full rank d such that $g(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T = \mathbf{P}_{\boldsymbol{\Gamma}}$. Lemma 10 implies

$$\nabla g(\boldsymbol{\Gamma}) = \frac{\partial \mathbf{P}_{\boldsymbol{\Gamma}}}{\partial \mathrm{vec}^T(\boldsymbol{\Gamma})} = (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1} \otimes \mathbf{Q}_{\boldsymbol{\Gamma}}).$$

Applying the Delta method, we obtain

$$\sqrt{n}\left(g(\widehat{\boldsymbol{\Gamma}}) - g(\boldsymbol{\Gamma})\right) \to \mathcal{N}\left(0, \nabla g(\boldsymbol{\Gamma})\mathbf{V}\nabla^T g(\boldsymbol{\Gamma})\right),$$

which completes the proof. ■

B.3 Proof of Proposition 6

By (37), $\widehat{\boldsymbol{\alpha}}_{\mathbf{b}} = \widehat{\mathbf{U}}_1$ and therefore $\boldsymbol{\alpha}_{\mathbf{b}} = \mathbf{U}_1$ and $\mathrm{span}(\widehat{\mathbf{U}}_1) = \mathrm{span}(\widehat{\mathbf{H}})$ with $\widehat{\mathbf{H}} = \widehat{\mathbf{U}}_1\widehat{\mathbf{K}}_1\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{K}^{-1}$ defined in Lemma 9. The same lemma gives the asymptotic distribution of $\widehat{\mathbf{H}}$ and therefore applying Lemma 11 with $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{H}}$ and $\boldsymbol{\Gamma} = \mathbf{H} = \mathbf{U}_1$ obtains the asymptotic distribution. Since $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_d$, it follows from Lemma 11 that the asymptotic variance is

$$(\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{U}_1\mathbf{K}^{-1}\mathbf{R}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1})\mathbf{V}_{rlc}(\mathbf{R}_1\mathbf{K}^{-1}\mathbf{U}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1}))(\mathbf{I}_{p^2} + \mathbf{K}_{pp})$$

By (33), $\mathbf{b} = \mathbf{U}_1\mathbf{K}\mathbf{R}_1^T$, and therefore $\mathbf{b}^- = \mathbf{R}_1\mathbf{K}^{-1}\mathbf{U}_1^T$ and the result follows. ■

C. Regularization term in variable selection

The specific form of $\Omega(\mathbf{C})$ in (39) depends on the type of predictor variables, as follows.

(a) When all predictors are continuous, we use the penalty $\Omega(\mathbf{C}) = \sum_{j=1}^{p} ||\mathbf{C}_j||_2$, with $\mathbf{C}_j$ the jth row of $\mathbf{C}$. In this case the sufficient reduction contains no interaction terms and each row of $\mathbf{C}$ corresponds to a single element of $\mathbf{X}$. Hence, by shrinking the jth row of $\mathbf{C}$ to 0, the computed reduction becomes insensitive to the measured value of X_j. When all predictors are continuous, under the assumed model the optimization problem is indeed fairly similar to group lasso (Yuan and Lin, 2006), as can be seen after rewriting (39) as

$$\underset{\mathbf{C}\in\mathbb{R}^{(p+q(q+1)/2)\times d}, \mathbf{C}^T\mathbf{C}=\mathbf{I}}{\arg\min} \|\mathrm{vec}(\widehat{\mathbf{b}}) - (\widehat{\mathbf{B}}^T \otimes \mathbf{I})\mathrm{vec}(\mathbf{C})\|_2^2 + \lambda \sum_{j=1}^{p} \|\mathbf{C}_j\|_2.$$

(b) When all predictors are binary, the sufficient reduction includes interaction effects H_iH_j. To discard the effect of a given binary variable, say H_j, we need to set all the entries in $\mathbf{C}$ related to H_j to zero. For a reduction of dimension d, there are d such entries related to the main effects and $d(q-1)$ related to the interaction terms. The grouping of the entries of $\mathbf{C}$ does not form a partition, since the entries affecting the interaction terms appear twice. For instance, assume for simplicity that $d = 1$. Parameter θ_{13} operates on variables H_1 and H_3 and then it enters the regularizer in groups $\{\eta_1, \theta_{12}, \theta_{13}, \ldots, \theta_{1q}\}$ and $\{\eta_3, \theta_{13}, \theta_{23}, \ldots, \theta_{q3}\}$. Both groups of parameters overlap at θ_{13}. Thus, the regularizer inducing the desired sparsity structure is a mixed-norm regularizer with overlapping groups, $\Omega(\mathbf{C}) = \sum_{g\in\mathcal{G}} \|\mathbf{C}_g\|_2$. Here, $g \subset \{1, \ldots, dq(q+1)/2\}$ indicates the subset of entries that affect the binary variable H_i and $\mathcal{G}$ is the collection of such groups. Moreover, each binary variable is associated with two groups, one derived from the main effects and one from the interaction terms, since they typically have rather different scales. The resulting regularized problem can be solved using algorithms for overlapping group lasso, as proposed, for example, in Liu and Ye (2010).

(c) When the predictors are mixed normal and binary, we combine the regularizers described in (a) and (b) in a single penalty $\Omega(\mathbf{C}) = \gamma \sum_{j=1}^{p} \|\mathbf{C}_j\|_2 + (1-\gamma) \sum_{g\in\mathcal{G}} \|\mathbf{C}_{G_i}\|_2$. The value of γ serves as a tuning weight for the amount of regularization in the continuous and binary parts, respectively. In SUB-OPTIMAL SDR, we carry out variable selection separately for the continuous and binary variables as described in (a) and (b).

D. Robustness under non-normality: Simulation results

				Continuous predictors						Mixed Predictors					
X	d	Method	$n =$	100	200	300	500	750	100	200	300	500	750
N	1	Optimal	Estimation	0.663	0.514	0.439	0.353	0.292	0.950	0.847	0.746	0.665	0.620
				(0.092)	(0.079)	(0.077)	(0.063)	(0.054)	(0.043)	(0.068)	(0.067)	(0.061)	(0.056)
			Prediction	0.594	0.445	0.376	0.300	0.245	0.509	0.380	0.350	0.330	0.317
				(0.096)	(0.075)	(0.070)	(0.056)	(0.047)	(0.162)	(0.053)	(0.041)	(0.035)	(0.026)
		PCA	Estimation	0.955	0.929	0.902	0.838	0.834	0.940	0.911	0.881	0.814	0.814
				(0.057)	(0.095)	(0.115)	(0.160)	(0.161)	(0.071)	(0.110)	(0.134)	(0.177)	(0.175)
			Prediction	0.940	0.911	0.881	0.814	0.814	0.459	0.453	0.449	0.447	0.446
				(0.071)	(0.110)	(0.134)	(0.177)	(0.175)	(0.020)	(0.016)	(0.014)	(0.011)	(0.008)
		NN	Optimal	Estimation	0.668	0.517	0.436	0.350	0.289	0.948	0.837	0.747	0.686	0.648
				(0.089)	(0.073)	(0.070)	(0.060)	(0.055)	(0.046)	0.067	0.067	(0.062)	(0.050)
			Prediction	0.604	0.454	0.378	0.300	0.247	0.569	0.486	0.465	0.456	0.449
				(0.093)	(0.071)	(0.066)	(0.054)	(0.048)	(0.112)	(0.076)	(0.041)	(0.035)	(0.028)
			PCA	Estimation	0.932	0.896	0.862	0.811	0.783	0.956	0.949	0.945	0.944	0.943
				(0.083)	(0.112)	(0.125)	(0.123)	(0.113)	(0.013)	(0.007)	(0.003)	(0.002)	(0.002)
			Prediction	0.906	0.860	0.819	0.760	0.731	0.557	0.515	0.496	0.489	0.485
				(0.101)	(0.133)	(0.144)	(0.134)	(0.119)	(0.102)	(0.045)	(0.017)	(0.012)	0.010
N	2	Optimal	Estimation	0.664	0.482	0.400	0.308	0.251	0.925	0.884	0.822	0.699	0.623
				(0.099)	(0.064)	(0.059)	(0.045)	(0.037)	(0.073)	(0.098)	(0.108)	(0.093)	(0.083)
			Prediction	0.572	0.391	0.320	0.244	0.199	0.825	0.726	0.645	0.521	0.488
				(0.108)	(0.062)	(0.051)	(0.039)	(0.032)	(0.127)	(0.154)	(0.156)	(0.121)	(0.110)
		PCA	Estimation	0.829	0.743	0.694	0.630	0.599	1.000	1.000	1.000	1.000	1.000
				(0.120)	(0.147)	(0.151)	(0.147)	(0.143)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
			Prediction	0.783	0.696	0.649	0.591	0.567	0.998	0.999	0.999	0.999	0.999
				(0.144)	(0.171)	(0.172)	(0.162)	(0.155)	(0.003)	0.002	0.001	0.001	0.001
		NN	Optimal	Estimation	0.735	0.555	0.472	0.386	0.319	0.920	0.858	0.774	0.661	0.583
				(0.091)	(0.072)	(0.066)	(0.060)	(0.055)	(0.069)	(0.096)	(0.102)	0.097	(0.079)
			Prediction	0.669	0.483	0.403	0.324	0.267	0.828	0.707	0.590	0.502	0.452
				(0.107)	(0.076)	(0.065)	(0.055)	(0.051)	(0.116)	(0.146)	(0.136)	(0.102)	(0.091)
			PCA	Estimation	0.953	0.927	0.908	0.874	0.857	1.000	1.000	1.000	1.000	1.000
				(0.058)	(0.073)	(0.089)	(0.119)	(0.118)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
			Prediction	0.938	0.909	0.888	0.852	0.836	0.999	0.999	0.999	0.999	0.999
				(0.075)	(0.088)	(0.106)	(0.136)	(0.135)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)

Table 6: Estimation and prediction errors (Euclidean norm) for normal (N) and non-central $t(5)$ (NN) distributed $\mathbf{X}$.

E. Computation of the matrix derivative in Eqn. (43)

The computation of $\mathbf{J}$ in Eqn. (42) (or Eqn. (43)) requires computing the derivative of $\psi(\boldsymbol{\eta}_y)$ with respect to $\boldsymbol{\eta}_y$.

E.1 General derivatives

Let

$$\begin{aligned}
\frac{\partial \text{vec}((\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1})}{\partial \text{vec}^T(\boldsymbol{\eta}_3)} &= \frac{\partial \text{vec}((\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1})}{\partial \text{vec}^T(\boldsymbol{\eta}_3)} \\
&= -\left[(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}) \otimes (\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1})\right] \frac{\partial \text{vec}((\text{unvec}(\mathbf{D}_p\eta_3)))}{\partial \text{vec}(\mathbf{D}_p\boldsymbol{\eta}_3)} \\
&= -\left[(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}) \otimes (\mathbf{D}_p^T\text{unvec}(\boldsymbol{\eta}_3))^{-1})\right] \mathbf{D}_p \\
&= -(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p. \\
\frac{\partial \boldsymbol{\eta}_{y1}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}}{\partial \boldsymbol{\eta}_{y1}^T} &= 2\boldsymbol{\eta}_{y1}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1} = 2\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}. \\
\frac{\partial^2 \boldsymbol{\eta}_{y1}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}}{\partial \boldsymbol{\eta}_{y1}\partial \boldsymbol{\eta}_{y1}^T} &= 2\boldsymbol{\Delta}.
\end{aligned}$$

Derivatives of $\psi_1(\boldsymbol{\eta}_3)$:

$$\begin{aligned}
\frac{\partial \psi_1(\boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_3^T} &= -\frac{1}{2}\text{vec}^T\left((\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\right)\mathbf{D}_p = -\frac{1}{2}\text{vec}^T(\boldsymbol{\Delta})\mathbf{D}_p. \\
\frac{\partial^2 \psi_1(\boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_3\partial \boldsymbol{\eta}_3^T} &= \frac{1}{2}\mathbf{D}_p^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1} \otimes (\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\mathbf{D}_p = \frac{1}{2}\mathbf{D}_p^T(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p.
\end{aligned}$$

Derivatives of $\psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)$: Let $\mathbf{K}_{pm} \in \mathbb{R}^{pm\times pm}$ be the unique matrix such that, for any symmetric $p \times m$ matrix $\mathbf{A}$, $\text{vec}(\mathbf{A}^T) = \mathbf{K}_{pm}\text{vec}(\mathbf{A})$. Then,

$$\begin{aligned}
\frac{\partial \psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_{y1}^T} &= \boldsymbol{\eta}_{y1}^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1} = \boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}. \\
\frac{\partial \psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_{y1}\partial \boldsymbol{\eta}_{y1}^T} &= \boldsymbol{\Delta}. \\
\frac{\partial \psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_3^T} &= -\frac{1}{2}\text{vec}^T(\boldsymbol{\eta}_{y1}\boldsymbol{\eta}_{y1}^T)((\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1} \otimes (\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1})\mathbf{D}_p^T \\
&= -\frac{1}{2}\text{vec}^T(\boldsymbol{\eta}_{y1}\boldsymbol{\eta}_{y1}^T)(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p \\
&= -\frac{1}{2}\text{vec}^T(\boldsymbol{\Delta}\boldsymbol{\eta}_{y1}\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\mathbf{D}_p. \\
\frac{\partial^2 \psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_3\partial \boldsymbol{\eta}_3^T} &= \mathbf{D}_p\left(\frac{I+\mathbf{K}_{pm}}{2}\right)(\boldsymbol{\Delta}\boldsymbol{\eta}_{y1}\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p^T \\
&= \mathbf{D}_p^T(\boldsymbol{\Delta}\boldsymbol{\eta}_{y1}\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p. \\
\frac{\partial^2 \psi_3(\boldsymbol{\eta}_{y1}, \boldsymbol{\eta}_3)}{\partial \boldsymbol{\eta}_{y1}\partial \boldsymbol{\eta}_3^T} &= -(\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})\mathbf{D}_p.
\end{aligned}$$

Derivatives of $\psi_2(\boldsymbol{\eta}_y)$: Let

$$\mathbf{M}_1 = \left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right),$$

$$\mathbf{M}_2 = \left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right).$$

For the sake of simplicity, we set $G(\boldsymbol{\Gamma}_y)$ in (14) to $\mathbf{S}$. Then, the first derivatives are

$$\begin{aligned}
\frac{\partial\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}^T} &= \frac{\mathbf{S}_1}{\mathbf{S}},\\
\frac{\partial\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y2}^T} &= \frac{\mathbf{S}_2}{\mathbf{S}},\\
\frac{\partial\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_3^T} &= -\frac{\mathbf{S}_{3_1}}{\mathbf{S}} - \frac{\mathbf{S}_{3_2}}{\mathbf{S}},\\
\frac{\partial\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_4^T} &= \frac{\mathbf{S}_{4_1}}{\mathbf{S}} + \frac{S_{4_2}}{\mathbf{S}},\\
\frac{\partial\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y5}^T} &= \frac{\mathbf{S}_5}{\mathbf{S}},
\end{aligned}$$

with,

$$\begin{aligned}
\mathbf{S}_1 &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta},\\
\mathbf{S}_2 &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\mathbf{H}^T,\\
\mathbf{S}_{3_1} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{C}_q\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \left((\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p + \frac{1}{2}\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p\right)\},\\
\mathbf{S}_{3_2} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5} + \frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2} + \bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1} + \frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4))\right)\right]\\
&\quad \text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p,
\end{aligned}$$

$$
\begin{aligned}
\mathbf{S}_{4_1} &= 2\sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(I_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}), \\
\mathbf{S}_{4_2} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \left(\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(I_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+(\mathbf{H}^T\otimes\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\right),
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{S}_5 &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right] \\
&\quad \mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T.
\end{aligned}
$$

Finally, the second derivatives are given by

$$
\begin{aligned}
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}\partial\boldsymbol{\eta}_{y1}^T} &= -\frac{\mathbf{S}_1^T\mathbf{S}_1}{\mathbf{S}^2}+\frac{\mathbf{S}_{11}}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}\partial\boldsymbol{\eta}_{y2}^T} &= -\frac{\mathbf{S}_1^T\mathbf{S}_2}{\mathbf{S}^2}+\frac{\mathbf{S}_{12}}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}\partial\boldsymbol{\eta}_3^T} &= \frac{\mathbf{S}_1^T(\mathbf{S}_{3_1}+\mathbf{S}_{3_2})}{\mathbf{S}^2}-\frac{(\mathbf{S}_{13_1}+\mathbf{S}_{13_2})}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}\partial\boldsymbol{\eta}_4^T} &= -\frac{\mathbf{S}_1^T(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})}{\mathbf{S}^2}+\frac{(\mathbf{S}_{14_1}+\mathbf{S}_{14_2})}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y1}\partial\eta_5^T} &= -\frac{\mathbf{S}_1^T\mathbf{S}_5}{\mathbf{S}^2}+\frac{\mathbf{S}_{15}}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y2}\partial\boldsymbol{\eta}_{y2}^T} &= \frac{-\mathbf{S}_2^T\mathbf{S}_2}{\mathbf{S}^2}+\frac{\mathbf{S}_{22}}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\eta_2\partial\boldsymbol{\eta}_3^T} &= \frac{\mathbf{S}_2^T(\mathbf{S}_{3_1}+\mathbf{S}_{3_2})}{\mathbf{S}^2}-\frac{(\mathbf{S}_{23_1}+\mathbf{S}_{23_2})}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y2}\partial\boldsymbol{\eta}_4^T} &= -\frac{\mathbf{S}_2^T(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})}{\mathbf{S}^2}+\frac{(\mathbf{S}_{24_1}+\mathbf{S}_{24_2})}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_{y2}\partial\boldsymbol{\eta}_{y5}^T} &= -\frac{\mathbf{S}_2^T\mathbf{S}_5}{\mathbf{S}^2}+\frac{\mathbf{S}_{25}}{\mathbf{S}}, \\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_3\partial\boldsymbol{\eta}_3^T} &= -\frac{(\mathbf{S}_{3_1}+\mathbf{S}_{3_2})^T(\mathbf{S}_{3_1}+\mathbf{S}_{3_2})}{\mathbf{S}^2}+\frac{\mathbf{S}_{33}}{\mathbf{S}},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_3\partial\boldsymbol{\eta}_4^T} &= \frac{(\mathbf{S}_{3_1}+\mathbf{S}_{3_2})^T(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})}{\mathbf{S}^2}-\frac{\mathbf{S}_{34}}{\mathbf{S}},\\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_4\partial\boldsymbol{\eta}_4^T} &= -\frac{(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})^T(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})}{\mathbf{S}^2}+\frac{\mathbf{S}_{44}}{\mathbf{S}},\\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_3\partial\boldsymbol{\eta}_{y5}^T} &= \frac{\mathbf{S}_5^T(S_{3_1}+\mathbf{S}_{3_2})}{\mathbf{S}^2}-\frac{(\mathbf{S}_{35_1}+\mathbf{S}_{35_2})}{\mathbf{S}},\\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\boldsymbol{\eta}_4\partial\boldsymbol{\eta}_{y5}^T} &= -\frac{\mathbf{S}_5^T(\mathbf{S}_{4_1}+\mathbf{S}_{4_2})^T}{\mathbf{S}^2}+\frac{(\mathbf{S}_{45_1}+\mathbf{S}_{45_2})}{\mathbf{S}},\\
\frac{\partial^2\psi_2(\boldsymbol{\eta}_y)}{\partial\eta_5\partial\boldsymbol{\eta}_{y5}^T} &= -\frac{\mathbf{S}_5^T\mathbf{S}_5}{\mathbf{S}^2}+\frac{\mathbf{S}_{55}}{\mathbf{S}},
\end{aligned}
$$

where,

$$
\begin{aligned}
\mathbf{S}_{11} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta},\\
\mathbf{S}_{12} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\mathbf{H}^T,\\
\mathbf{S}_{22} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\mathbf{H}\mathbf{H}^T,\\
\mathbf{S}_{15} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,\\
\mathbf{S}_{25} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \mathbf{H}\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,\\
\mathbf{S}_{55} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{J}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\mathrm{vec}(\bar{\boldsymbol{\eta}}_4^T(\mathrm{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]
\end{aligned}
$$

$$\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T)\text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,$$

$$\begin{aligned}
\mathbf{S}_{14_1} &= 2\sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \left[\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(I_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\right],\\
\mathbf{S}_{14_2} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \left[\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\left\{\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+(\mathbf{H}^T\otimes\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\right\}+(\mathbf{H}^T\otimes\boldsymbol{\Delta})\right],\\
\mathbf{S}_{24_1} &= 2\sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \mathbf{H}\text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(I_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}),\\
\mathbf{S}_{24_2} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^Tq\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \mathbf{H}\left(\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+(\mathbf{H}^T\otimes\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\right),\\
\mathbf{S}_{45_1} &= 2\sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad (I_q\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T)\text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,\\
\mathbf{S}_{45_2} &= \sum_H \exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right]\\
&\quad \left\{(\mathbf{I}_q\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H}+(\mathbf{H}\otimes\boldsymbol{\Delta}\boldsymbol{\eta}_{y1})\right\}\text{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T.
\end{aligned}$$

We let

$$[1]=\exp\left[(\mathbf{J}_q\mathbf{C}_q\text{vec}(\mathbf{H}\mathbf{H}^T))^T\left(\boldsymbol{\eta}_{y5}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right],$$

and,

$$[2]=\exp\left[\mathbf{H}^T\left(\boldsymbol{\eta}_{y2}+\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\boldsymbol{\eta}_{y1}+\frac{1}{2}\mathbf{L}_q\mathbf{D}_q^T\text{vec}(\bar{\boldsymbol{\eta}}_4^T(\text{unvec}(\mathbf{D}_p\boldsymbol{\eta}_3))^{-1}\bar{\boldsymbol{\eta}}_4)\right)\right].$$

Then,

$$
\begin{aligned}
\mathbf{S}_{13_1} &= \sum_H [1][2]\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\left((\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p+\frac{1}{2}\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p\right)\\
&\quad+\sum_H [1][2](\boldsymbol{\Delta}\otimes\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p,\\
\mathbf{S}_{13_2} &= \sum_H [1][2]\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p,\\
\mathbf{S}_{23_1} &= \sum_H [1][2]\mathbf{H}\left((\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta}\otimes\mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p+\frac{1}{2}\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p\right),\\
\mathbf{S}_{23_2} &= \sum_H [1][2]\mathbf{H}\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta}\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p,\\
\mathbf{S}_{35_1} &= \sum_H [1][2]\left(\mathbf{D}_p^T(\boldsymbol{\Delta}\boldsymbol{\eta}_1\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H})+\frac{1}{2}\mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H}\right)\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,\\
\mathbf{S}_{35_2} &= \sum_H [1][2]\mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T)\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T,\\
\mathbf{S}_{34} &= \sum_H [1][2]\Big(\mathbf{D}_p^T(\boldsymbol{\Delta}\boldsymbol{\eta}_1\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H})+\frac{1}{2}\mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H}\\
&\qquad+\mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T)\Big)\\
&\qquad\left[2\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+\left(\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+(\mathbf{H}^T\otimes\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\right)\right]\\
&\quad+\sum_H [1][2]SS+\sum_H [1][2]LL.
\end{aligned}
$$

where,

$$
\begin{aligned}
SS &= \mathbf{D}_p^T(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})(\mathbf{H}^T\otimes\mathbf{I}_{p^2})\left\{[\mathbf{K}_{pq}(\boldsymbol{\eta}_1\otimes\mathbf{I}_q)]\otimes\mathbf{I}_p\right\}+\mathbf{D}_p^T(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})(O^T\otimes\mathbf{I}_{p^2})KK,\\
O &= \mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T)),\\
LL &= \mathbf{D}_p^T(\boldsymbol{\Delta}\otimes\boldsymbol{\Delta})(G^T\otimes\mathbf{I}_{p^2})KK,\\
G &= \frac{1}{2}\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H},\\
KK &= \left\{(\mathbf{I}_q\otimes[(\mathbf{K}_{pq}\otimes\mathbf{I}_p)(\mathbf{I}_p\otimes\boldsymbol{\eta}_4)])+([(\mathbf{I}_q\otimes\mathbf{K}_{pq})(\boldsymbol{\eta}_4\otimes\mathbf{I}_q)]\otimes\mathbf{I}_p)\right\},\\
S_{44} &= \sum_H [1][2]\left[2(\mathbf{I}_q\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T)+(\mathbf{I}_q\otimes\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H}+(\mathbf{H}\otimes\boldsymbol{\Delta}\boldsymbol{\eta}_{y1})\right]\\
&\qquad\left[2\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+\left(\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\mathbf{I}_q\otimes\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})+(\mathbf{H}^T\otimes\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta})\right)\right]\\
&\qquad+\sum_H [1][2]HH+\sum_H [1][2]JJ,\\
HH &= 2(\mathbf{I}_q\otimes\boldsymbol{\Delta})(O^T\otimes\mathbf{I}_{qp})TT,\\
TT &= \{(\mathbf{I}_q\otimes\mathbf{K}_{q^2})(\mathrm{vec}(\mathbf{I}_q)\otimes\mathbf{I}_q)\otimes\mathbf{I}_p\},
\end{aligned}
$$

$$
\begin{aligned}
JJ &= (\mathbf{I}_q \otimes \boldsymbol{\Delta})(\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q \otimes \mathbf{I}_{pq})TT,\\
S_{33} &= \sum_H [1][2]\Big(\mathbf{D}_p^T(\boldsymbol{\Delta}\boldsymbol{\eta}_1 \otimes \boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4\mathbf{H}) + \frac{1}{2}\mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4 \otimes \boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{L}_q^T\mathbf{H}\\
&\qquad + \mathbf{D}_p^T(\boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4 \otimes \boldsymbol{\Delta}\bar{\boldsymbol{\eta}}_4)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q\mathrm{vec}(\mathbf{H}\mathbf{H}^T)\Big),\\
&\qquad \Big((\boldsymbol{\eta}_{y1}^T\boldsymbol{\Delta} \otimes \mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p + \frac{1}{2}\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta} \otimes \bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p\\
&\qquad + \mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta} \otimes \bar{\boldsymbol{\eta}}_4^T\boldsymbol{\Delta})\mathbf{D}_p\Big)\\
&\qquad + \sum_H [1][2]J_1 + \sum_H [1][2]J_2 + \sum_H [1][2]J_3,\\
S_T &= (\mathbf{I}_p \otimes G)(\boldsymbol{\Delta} \otimes \boldsymbol{\Delta})\mathbf{D}_p + (H \otimes \mathbf{I}_p)(\boldsymbol{\Delta} \otimes \boldsymbol{\Delta})\mathbf{D}_p,\\
G &= (\mathbf{K}_{p^2} \otimes \mathbf{I}_p)(\mathbf{I}_p \otimes \mathrm{vec}(\boldsymbol{\Delta})),\\
H &= (I_p \otimes \mathbf{K}_{p^2})(\mathrm{vec}(\boldsymbol{\Delta}) \otimes \mathbf{I}_p),\\
J_1 &= \mathbf{D}_p^T[(\boldsymbol{\eta}_1^T \otimes \mathbf{H}^T\bar{\boldsymbol{\eta}}_4^T) \otimes \mathbf{I}_{p^2}]S_T,\\
J_2 &= \frac{1}{2}\mathbf{D}_p^T[\mathbf{H}^T\mathbf{L}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T \otimes \bar{\boldsymbol{\eta}}_4^T)) \otimes \mathbf{I}_{p^2}]S_T\\
J_3 &= \mathbf{D}_p^T[(\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathbf{C}_q^T\mathbf{J}_q^T\mathbf{J}_q\mathbf{C}_q(\bar{\boldsymbol{\eta}}_4^T \otimes \bar{\boldsymbol{\eta}}_4^T)) \otimes \mathbf{I}_{p^2}]S_T.
\end{aligned}
$$