Random effects sufficient dimension reduction for clustered data

Linh H. Nghiem*

School of Mathematics and Statistics, University of Sydney, Australia

Francis K.C. Hui †

Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia

Abstract

Sufficient dimension reduction (SDR) is a popular class of regression methods which aim to find a small number of linear combinations of covariates that capture all the information of the responses (a central subspace). The majority of current methods for SDR focus on the setting of independent observations, while the few SDR techniques that have been developed for clustered data assume the linear transformation is identical across clusters. That is, they do not allow for heterogeneity between clusters when it comes to the sufficient dimension reduction process. In this article, we introduce the concept of random effect central subspaces, where cluster-specific central subspaces are assumed to be random following a distribution on the Grassmann manifold. This random effects distribution is the image of an exponential mapping from a Gaussian distribution on the tangent space of an overall fixed effects central subspace, and characterized by a covariance matrix capturing the heterogeneity between clusters. We incorporate the concept of random effect central subspaces into the principal fitted components model, and propose a two-stage algorithm for estimation and prediction of the random cluster-specific central subspaces. We demonstrate the consistency of the proposed estimators when the number of clusters grows while the cluster sizes remain bounded. Simulation studies demonstrate the superior performance of our proposed approach compared to both global and cluster-specific SDR methods. We apply the proposed method to study the relationship between the life expectancy of women with socioeconomic variables across countries. Results show log income per capita, infant mortality, and inequality primarily drive the two-dimensional overall fixed effects central subspace, although there is considerable variability between countries in how their cluster-specific central subspaces are driven by these predictors.

Keywords: EM algorithm, Grassmann manifold, mixed models, random effects, repeated measures

*Thanks to Andrew Wood and Janice Scealy for useful discussions.

†FKCH was supported by an Australia Research Council Discovery Project DP230101908.

1 Introduction

Sufficient dimension reduction (SDR, Ma and Zhu, 2013; Li, 2018) is to a class of statistical methods that assume the outcome depends on covariates via a small number of their linear combinations. These linear combinations are known as sufficient predictors, and retain the full regression information between the response and all the covariates, thereby overcoming the curse of dimensionality. The number of linear combinations is often known as the structural dimension. Since the pioneering works of Li and Duan (1989) and Li (1991), a vast literature has developed on different approaches to SDR, from inverse-moment-based and regression-based methods (e.g., Li, 1991; Cook and Forzani, 2008), forward regression (e.g., Xia et al., 2002), to semiparametric techniques. Much research has also been done to combine SDR with various aspects of statistical inference, such as SDR with variable selection for high dimensional sparse dimension reduction (e.g., Lin et al., 2018; Nghiem et al., 2023); we refer the reader to Ma and Zhu (2013); Rodrigues et al. (2022) for overviews of the SDR literature.

The vast majority of the SDR methods in the literature focus on the setting of independent observations. By contrast, the adaptation of SDR to clustered data settings remains relatively underdeveloped. Such settings are common in many disciplines including medical and social statistics (Verbeke and Molenberghs, 2009) and ecological and environmental studies (Bjork et al., 2018), where many scientific questions involve some degree of dimension reduction of the covariates. In correlated data settings, a very common method of analysis involves fitting mixed effects models or variations thereof (Verbeke and Molenberghs, 2009; Fitzmaurice et al., 2012), which combine fixed effects that are identical across clusters with random effects whose effects vary from one cluster to another. These random effects are assumed to come from some common (typically normal) distribution with a zero mean/location vector and a covariance matrix, the latter of which characterizes the degree of heterogeneity across the clusters. Mixed effect models can thus

be seen as a balance between a global fixed effects model that ignores all clustering/correlation structures in the data, and a cluster-specific fixed effects model that ignores possible shared information across clusters.

Among the few research that have been done on SDR for clustered data, all have effectively assumed a global fixed effects model approach to dimension reduction, i.e., the direction of the linear transformation is identical across all the clusters. Bi and Qu (2015) and Xu et al. (2016) employed a marginal estimation equation approach where working correlation matrices were included to account for temporal correlations arising within clusters for longitudinal data, while Hui and Nghiem (2022) proposed a finite mixture approach where the mixture proportions are modeled as known function of sufficient predictors, and random effects are added to the mixture means to account for within-cluster correlations. Such global fixed effects SDR is also assumed in Pfeiffer et al. (2021) and Song et al. (2023), who developed methods for matrix-valued predictors that are formed from the collection of all repeated measurements of covariates corresponding to each cluster. We also note the connected literature on single index models and variations thereof for clustered data (e.g., Pang and Xue, 2012; Tian and Qiu, 2023). All of these works again assume a global fixed effects model approach to dimension reduction: the random effects are a separate, independent component to the linear transformation of the covariates and thus to the sufficient predictors. On the other hand, such an assumption of the same direction for the linear transformation across all clusters may be restrictive, as it does not allow for heterogeneity between clusters when it comes to the sufficient dimension reduction operation itself.

In this article, we introduce the idea of random effects sufficient dimensions reduction, where heterogeneity across clusters of the dimension reduction operation is allowed by assuming the linear transformation of the covariates for each cluster is itself drawn from a common distribution. In turn, we have an overall fixed effects sufficient dimension reduction, and random cluster-

specific sufficient dimension reduction representing deviations away from this. Like standard mixed effects modeling, we characterize the degree of heterogeneity in the sufficient dimension operation between clusters by a random effects covariance matrix.

One immediate challenge in defining random effects SDR is that the direction of the linear transformation of the covariates is not unique, since it is invariant to any orthogonal rotation. Instead, similar to other SDR techniques, the estimation target is the subspace spanned by the columns of the linear transformation for each cluster, which in our setting is both the overall central subspace and the cluster-specific central subspaces. Because all of these central subspaces are elements on a Grassmann manifold, one approach to random effects SDR would be to define the distribution on this manifold. To date however, only a few such distributions have been proposed in the literature, especially when the structural dimension exceeds one. These distributions are often defined via another distribution on a Stiefel manifold, with notable examples being the matrix Bingham and the matrix Langevin distribution (Chikuse, 2003b,a). On the other hand, these distributions almost always contain intractable normalizing constants, making their parameter estimation, and thus random effects SDR using this approach, computationally burdensome. Recently, some distributions have been proposed to overcome the issue of intractable normalizing constraints (e.g., Scealy and Wood, 2019, 2022), but they are only applicable when the structural dimension is one, which is not usually the case for SDR.

To overcome the above challenge, we propose an alternative approach to random effects SDR as follows. Instead of placing a distribution on a Grassmann manifold directly, we construct the distribution of the cluster-specific central subspaces as the image of an exponential mapping of a distribution defined on a tangent space of the Grassmann manifold (Srivastava and Klassen, 2016)constructed at an overall fixed effects central subspace. This modeling approach has the advantage that the tangent subspace is a vector space, meaning we can assume a random effects dis-

tribution defined on the corresponding Euclidean space e.g., a matrix normal distribution with a covariance matrix characterizing the heterogeneity of sufficient predictors between clusters. Furthermore, we can leverage recent advances in Grassmann manifold computation (Zimmermann, 2017; Bendokat et al., 2020) to perform the exponential mapping efficiently. After defining a distribution for cluster-specific central subspace in this manner, we employ a likelihood-based inverse regression approach for SDR, specifically, the principal fitted components (PFC) model of Cook and Forzani (2008).

We propose a two-stage estimation algorithm for fitting the proposed model, where we first use a global fixed-effects SDR to estimate a parameter of the mean central subspace, and then apply the Monte-Carlo Expectation Maximization algorithm (Wei and Tanner, 1990) to estimate the remaining (identifiable) parameters, and predict the cluster-specific central subspaces. We prove the consistency of the proposed estimators when the number of clusters goes to infinity, while the cluster sizes can remain bounded. Simulation studies demonstrate the strong performance of our proposed model for random effects SDR, compared to both a global fixed-effects PFC that ignores the heterogeneity between clusters and a cluster-specific fixed-effects PFC model that does not borrow strength across clusters. Finally, we apply the proposed model to perform random effects SDR on a longitudinal dataset studying the relationship between the female life expectancy and various socioeconomic variables across different countries.

The remainder of this article is organized as follows. Section 2 offers a brief overview of relevant concepts and tools from differential geometry regarding the Grassmann manifold, and introduces the concept of random effect central subspaces. Section 3 proposes a model for random effects SDR, and a two-step procedure for parameter estimation. We prove the consistency of the proposed estimators in Section 4. Section 5 compares the proposed model with the two fixed-effects SDR models via simulation, while Section 6 discusses the selection of the structural dimension.

Section 7 presents an illustration of the proposed model on a longitudinal socioeconomic dataset, while Section 8 offers some concluding remarks.

2 Overview and key concepts

We begin by reviewing some key concepts related to Grassmann manifold and tangent spaces, before introducing the concept of random effects SDR.

2.1 Background on Grassmann manifold

A Grassmann manifold $\mathrm{Gr}(p,d)$ is the set of all linear subspaces with dimension d of $\mathbb{R}^p$. One way to represent a point on a Grassmann manifold is from a basis perspective (Bendokat et al., 2020). That is, a subspace $\mathcal{U} \in \mathrm{Gr}(p,d)$ is identified by a non-unique, semi-orthogonal matrix $\mathbf{U} \in \mathbb{R}^{p\times d}$, satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$, whose columns form a basis for $\mathcal{U}$. To simplify the notation between a space and its semi-orthogonal matrix, we will write $\mathcal{U} = [\mathbf{U}]$. A Grassmann manifold is often equipped with a Riemmann metric, which roughly speaking is an inner product defined on the tangent space of the manifold. In more detail, given a subspace $[\mathbf{U}] \in \mathrm{Gr}(p,d)$, the tangent space of $\mathrm{Gr}(p,d)$ at $[\mathbf{U}]$, denoted as $T_{[\mathbf{U}]}\,\mathrm{Gr}(p,d)$, is the collection of all possible directions of a curve on the manifold that passes through $[\mathbf{U}]$. That is, $T_{[\mathbf{U}]}\,\mathrm{Gr}(p,d) = \{\mathbf{V} \in \mathbb{R}^{p\times d} | \mathbf{U}^\top\mathbf{V} = \mathbf{0}\}$ is the set of all matrices orthogonal to $\mathbf{U}$. Equivalently, any matrix $\mathbf{V} \in T_{[\mathbf{U}]}\,\mathrm{Gr}(p,d)$ can be written in the form $\mathbf{V} = (\mathbf{I}_p - \mathbf{U}\mathbf{U}^\top)\mathbf{A}$, where $\mathbf{A}$ is any arbitrary matrix in $\mathbb{R}^{p\times d}$. Importantly, this tangent space is a vector space: for any two matrices $\mathbf{V}_1, \mathbf{V}_2 \in T_{[\mathbf{U}]}\,\mathrm{Gr}(p,d)$, we have $c_1\mathbf{V}_1 + c_2\mathbf{V}_2 \in T_{[\mathbf{U}]}\,\mathrm{Gr}(p,d)$ for any two scalars c_1 and c_2. As such, we can embed the tangent space with an inner product, $\Phi(\mathbf{V}_1,\mathbf{V}_2) = 2^{-1}\mathrm{trace}(\mathbf{V}_1^\top\mathbf{V}_2)$, which is known as a Riemann metric of the Grassmann manifold (Zimmermann, 2017).

Since the tangent space is a vector space, then we will define a probability distribution in this space whose variability subsequently determines the need to variability of points on the manifold. Furthermore, we will define a mapping between points on the tangent space and points on the Grassmann manifold. For the latter, we employ the exponential mapping defined as follows: given a point $[\mathbf{U}_1] \in \mathrm{Gr}(p,d)$ and a point $\mathbf{V} \in T_{[\mathbf{U}_1]}\mathrm{Gr}(p,d)$, the exponential map $\mathrm{Exp}_{[\mathbf{U}_1]}$ transfer a point $\mathbf{V} \in T_{[\mathbf{U}_1]}\mathrm{Gr}(p,d)$ to $[\mathbf{U}_2] \in \mathrm{Gr}(p,d)$. As shown in Bendokat et al. (2020), a semi-orthogonal basis for $[\mathbf{U}_2]$ is explicitly given by $\mathbf{U}_2 = \mathrm{Exp}_{[\mathbf{U}_1]}(\mathbf{V}) = \mathbf{U}_1\mathbf{M} + \mathbf{Q}\mathbf{N}$, where $\mathbf{Q}\mathbf{R} = \mathbf{V}$ is the QR decomposition of $\mathbf{V}$, $\mathbf{M} = \mathbf{D}\cos(\mathbf{\Theta})\mathbf{D}^\top$, and $\mathbf{N} = \mathbf{\Phi}\sin(\mathbf{\Theta})\mathbf{D}^\top$ where $\mathbf{\Phi}\mathbf{\Theta}\mathbf{D}^\top$ is the singular value decomposition (SVD) for $\mathbf{R}$. Note since $\mathbf{V}$ is orthogonal to $\mathbf{U}_1$, then so is $\mathbf{Q}$. Conversely, the inverse exponential mapping Exp^{-1} transfers a point $[\mathbf{U}_2] \in \mathrm{Gr}(p,d)$ to a point $\mathbf{V} \in T_{[\mathbf{U}_1]}\mathrm{Gr}(p,d)$. The explicit formula for this is $\mathbf{V} = \mathrm{Exp}^{-1}([\mathbf{U}_2]) = \mathbf{Q}^* \arctan(\mathbf{\Sigma}^*)\mathbf{D}^\top$ where $\mathbf{Q}^*\mathbf{\Sigma}^*\mathbf{D}^\top$ is the SVD of $(\mathbf{I}_p - \mathbf{U}\mathbf{U}_1)\mathbf{U}_2(\mathbf{U}_1^\top\mathbf{U}_2)^{-1}$. In these above formulas, the cosine, sine, and inverse tangent functions are applied point-wise to the diagonal elements of corresponding matrices.

2.2 Random effects sufficient dimension reduction

Consider a set of n independent clusters, such that for cluster $i = 1,\ldots,n$ we let y_{ij} denote the j-th measurement of the response for $j = 1,\ldots,m_i$, and $\mathbf{X}_{ij}$ denote a corresponding vector of p covariates. For each cluster, SDR implies that the response only depends on the covariates via a small number of their linear combinations,

$$y_{ij} \perp \mathbf{X}_{ij} | \mathbf{X}_{ij}^\top \mathbf{\Gamma}_i; \quad i = 1,\ldots,n; j = 1,\ldots,m_i, \tag{1}$$

where $\mathbf{\Gamma}_i \in \mathbb{R}^{p\times d}$ with $d < p$. Model (1) allows the directions $\mathbf{\Gamma}_i$ to potentially vary from cluster to another, with the idea being that the $\mathbf{\Gamma}_i$'s represent deviations from some overall direction

$\mathbf{\Gamma}_0$, which we define formally later on. Without loss of generality, assume each $\mathbf{\Gamma}_i$ is a semi-orthogonal matrix, $\mathbf{\Gamma}_i\mathbf{\Gamma}_i^\top = \mathbf{I}_d$ for $i = 1,\ldots,n$. Similar to standard SDR, only the subspaces spanned by each $\mathbf{\Gamma}_i$ is unique and identifiable. We refer to $[\mathbf{\Gamma}_i]$ as the central subspace for the ith cluster, which is a point on the Grassmann manifold $\text{Gr}(p,d)$.

In model (1), if the spaces $[\mathbf{\Gamma}_i]$ are identical for all n clusters, then there is no heterogeneity between clusters in terms of the sufficient dimension reduction operation i.e., $[\mathbf{\Gamma}_i] = [\mathbf{\Gamma}_0]$ for all i. As reviewed in Section 1, to our knowledge all the current literature on SDR for clustered data has been developed under such an approach. On the other hand, if the spaces $[\mathbf{\Gamma}_i]$ are assumed to be completely different from another, then in practice one would identify the central subspace for each cluster independently, and no information would be shared across clusters. In this case, the concept of an overall direction $[\mathbf{\Gamma}_0]$ which individual clusters deviate is not explicitly defined, and there is no borrowing of strength across clusters in the dimension reduction operation.

As a balance between the above two then, and analogous to the formulation mixed models discussed in Section 1, we propose obtaining the cluster-specific central subspaces $[\mathbf{\Gamma}_i]$ from an exponential mapping of an overall fixed effects central subspace $[\mathbf{\Gamma}_0]$ via a random velocity vector $\mathbf{V}_i$. That is, for $i = 1\ldots,n$,

$$\mathbf{\Gamma}_i = \text{Exp}_{[\mathbf{\Gamma}_0]}(\mathbf{V}_i), \mathbf{V}_i \in T_{[\mathbf{\Gamma}_0]}\,\text{Gr}(p,d). \tag{2}$$

We assume a random effects distribution for the $\mathbf{V}_i$'s, whose covariance matrix characterizes the variability of central subspace among the clusters. Critically, since $\mathbf{V}_i \in T_{[\mathbf{\Gamma}_0]}\,\text{Gr}(p,d)$, then from Section 2.1 we have that $\mathbf{V}_i$ is orthogonal to $\mathbf{\Gamma}_0$, and so any multivariate distribution imposed on $\mathbf{V}_i$ should only have a non-zero density in the complement subspace of $[\mathbf{\Gamma}_0]$. In this article, we explore one such distribution that satisfies these requirements, namely the singular matrix-valued normal (MN) distribution with zero mean vector and two covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Omega}$

that characterize the covariance among the rows and columns of $\mathbf{V}_i$, respectively. Specifically, $\mathbf{V}_i \sim \text{MN}_{p\times d}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, where the dimensions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are $p \times p$ and $d \times d$, respectively. The orthogonality between $\mathbf{V}_i$ and $\boldsymbol{\Gamma}_0$ implies that $\boldsymbol{\Sigma}$ is also orthogonal to $\boldsymbol{\Gamma}_0$, and $\boldsymbol{\Sigma}$ has rank at most $p-d$. The joint density of $\mathbf{V}_i$, defined on the subspace $\mathbf{V}_i^\top \boldsymbol{\Gamma}_0 = \mathbf{0}$, is given by

$$p(\mathbf{V}_i; \boldsymbol{\Sigma}, \boldsymbol{\Omega}) = \frac{\exp\left\{-\frac{1}{2}\,\text{tr}\left(\boldsymbol{\Omega}^{-1}\mathbf{V}_i^\top \boldsymbol{\Sigma}^{-}\mathbf{V}_i\right)\right\}}{(2\pi)^{pd/2}|\boldsymbol{\Lambda}|^{1/2}|\boldsymbol{\Omega}|^{1/2}}, \tag{3}$$

where $\boldsymbol{\Lambda}$ denotes an $(p-d)\times(p-d)$ diagonal matrix with elements containing the non-zero eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-}$ denotes the Moore-Penrose inverse of $\boldsymbol{\Sigma}$, and $\text{tr}(\cdot)$ and $|\cdot|$ denote the trace and determinant operators, respectively.

Based on the above random effects formulation, we have the following result, the proof of which is given in the appendix.

Lemma 1. *If the density function of $\mathbf{V}_i$ is symmetric about zero, then $[E(\boldsymbol{\Gamma}_i)] = [\boldsymbol{\Gamma}_0]$, where the expectation is taken with regards to a uniform probability measure on $\mathbb{R}^{p\times d}$.*

The above implies that by assuming a degenerate matrix-valued normal (MN) distribution, $[\boldsymbol{\Gamma}_0]$ is indeed the mean central subspace and thus represents an overall direction across all n clusters.

As noted in Gupta and Nagar (1999), for any scalar $s > 0$ we have $p(\mathbf{V}_i; \boldsymbol{\Sigma}, \boldsymbol{\Omega}) = p(\mathbf{V}_i; s\boldsymbol{\Sigma}, s^{-1}\boldsymbol{\Omega})$. That is, the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are only identifiable up to a scale. Thus without loss of generality, we set $\boldsymbol{\Omega}$ to be a correlation matrix. In this article, we will restrict $\boldsymbol{\Omega} = \mathbf{I}_d$, meaning that any two dimensions in the central subspace can vary independently from one another, and the variability of the central subspace among the clusters is solely characterized by $\boldsymbol{\Sigma}$. We leave the exploration of more complex structures of $\boldsymbol{\Omega}$ as an avenue of future research.

To conclude this section, we present a visualization of the variability of the cluster-specific cen-

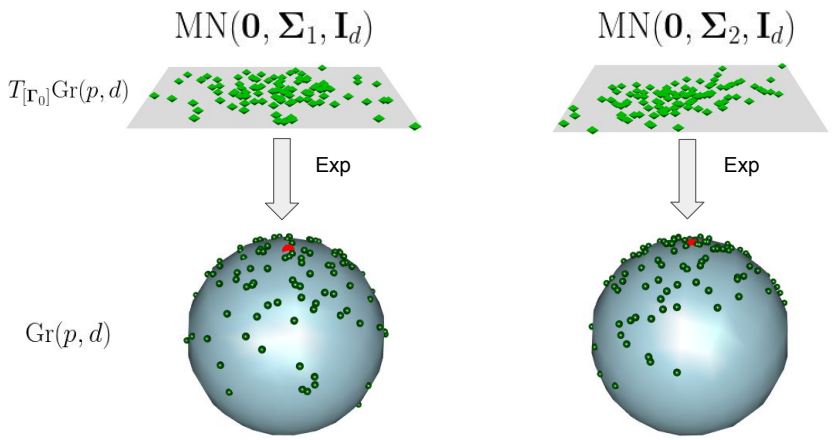

Figure 1: Visualization of random effect central subspaces based on the tangent space approach. In each panel, the overall fixed effects central subspace $[\mathbf{\Gamma}_0]$ is represented by the large red dot, while the planes on the top row represent the tangent space $T_{[\mathbf{\Gamma}_0]}$. Each point on the tangent plane represents one $\mathbf{V}_i$ generated from a matrix-valued normal distribution $\mathrm{MN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_d)$. Each green point on the sphere represents the corresponding $\mathbf{\Gamma}_i$ from the exponential map defined by equation (2). The left panel sets $\mathbf{\Sigma}_1 = \mathbf{K}\tilde{\mathbf{\Sigma}}_1\mathbf{K}$ with $\tilde{\mathbf{\Sigma}}_1 = 0.3\mathbf{I}_3$and $\mathbf{K} = \mathbf{I}_p - \mathbf{\Gamma}_0\mathbf{\Gamma}_0^\top$. The right panel sets $\mathbf{\Sigma}_2 = \mathbf{K}\tilde{\mathbf{\Sigma}}_2\mathbf{K}$ where $\tilde{\mathbf{\Sigma}}_2$ has an exchangeable structure with variance 0.3 and covariance 0.15.

tral subspace using the approach described above, in the simplest case with $d = 1$ and $p = 3$ i.e., a one-dimensional central subspace representing dimension reduction from three dimensions. In both panels of Figure 1, we set $\mathbf{\Gamma}_0 = (1, 0, 0)^\top$ and generate $n = 50$ points (clusters) $\mathbf{V}_i = \mathbf{K}\tilde{\mathbf{V}}_i$, where $\mathbf{K} = \mathbf{I}_3 - \mathbf{\Gamma}_0\mathbf{\Gamma}_0^\top$ and each $\tilde{\mathbf{V}}_i$ is sampled from the trivariate normal distribution $N_3(\mathbf{0}, \tilde{\mathbf{\Sigma}})$. This data generation process ensures each $\mathbf{V}_i$ is orthogonal to $\mathbf{\Gamma}_0$, and the corresponding covariance for $\mathbf{V}_i$ is given by $\mathbf{\Sigma} = \mathbf{K}\tilde{\mathbf{\Sigma}}\mathbf{K}$. Each panel in the figure corresponds to one choice for $\tilde{\mathbf{\Sigma}}$. The corresponding $\mathbf{\Gamma}_i$ is obtained from applying the exponential map (2) to each generated $\mathbf{V}_i$. We can see that different covariances on the tangent space lead to different variability patterns in the manifold, so the random effect covariance matrix $\mathbf{\Sigma}$ can be used (as a surrogate) to characterize heterogeneity among the cluster-specific central subspaces.

3 A random effects principal fitted components model

Based on the concept of random cluster-specific central subspaces introduced in the previous section, we now formulate a method for performing random effects SDR. Consider the set of observations $\{(\mathbf{X}_{ij}, y_{ij}); i = 1, \ldots, n,\ j = 1, \ldots, m_i\}$, with $N = \sum_{i=1}^{n} m_i$ denoting the total number of observations in the data, and recalling that the n clusters are assumed to be independent. Given an explicit parametric form (matrix normal) for the random effects distribution of the cluster-specific central subspaces, it is sensible to adopt a likelihood-based inverse-regression model for sufficient dimension reduction here. We propose to leverage the principal fitted components model (PFC) of Cook and Forzani (2008), such that for the i-th cluster we write

$$\mathbf{X}_{ij}|(y_{ij}, \boldsymbol{\Gamma}_i) = \boldsymbol{\mu}_i + \boldsymbol{\Gamma}_i \mathbf{v}_{ijy} + \boldsymbol{\varepsilon}_{ij},\ \boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Delta}), \tag{4}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^p$ denotes the conditional mean vector for the i-th cluster, $\mathbf{v}_{ijy} \in \mathbb{R}^d$ is an (unknown) function of y_{ij}, and $\boldsymbol{\Delta}$ is a $p \times p$ unstructured covariance matrix. Based on (4), the cluster-specific central subspace is spanned by $[\boldsymbol{\Theta}_i] = \boldsymbol{\Delta}^{-1}[\boldsymbol{\Gamma}_i]$ As in Cook and Forzani (2008); Cook and Li (2009); Bura and Forzani (2015) among others, when applying PFC we set $\mathbf{v}_{ijy} = \boldsymbol{\beta}\left\{\mathbf{f}_{yij} - E\left(\mathbf{f}_{yij}\right)\right\}$, where $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$ has rank $d \leq \min(p, r)$ and $\mathbf{f}_{ijy} \in \mathbb{R}^r$ is a known function of y_{ij} often chosen to be a reasonably flexible set of basis functions of y_{ij} e.g., piecewise polynomials. Also, note the $\boldsymbol{\mu}_i$'s in (4) play the role of cluster-specific intercept terms, and are treated as fixed parameters; this approach is similar to fixed effect models in econometrics (e.g., Hsiao et al., 2002).

Substituting our choice of $\mathbf{v}_{ijy}$ into (4) and combining with the developments in Section 2.2, we now formally define our approach to random effects SDR, which we refer to as random effect

PFC or RPFC,

$$\begin{aligned}
&\mathbf{X}_{ij}|(y_{ij}, \boldsymbol{\Gamma}_i) = \boldsymbol{\mu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\beta} \mathbf{f}_{ijy} + \boldsymbol{\varepsilon}_{ij}, \quad \boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Delta}), \\
&\boldsymbol{\Gamma}_i = \mathrm{Exp}_{[\boldsymbol{\Gamma}_0]}(\mathbf{V}_i),\ \mathbf{V}_i \overset{iid}{\sim} \mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_d); \quad i = 1, \ldots, n,
\end{aligned} \tag{5}$$

where the ε_{ij}'s are assumed to be independent of $\boldsymbol{\Gamma}_i$. All the $\boldsymbol{\Gamma}_i$'s are images of an exponential map from a subspace $[\boldsymbol{\Gamma}_0]$ on the Grassmann manifold. Furthermore, by Lemma 1 we have that the overall fixed effects central subspace is given by $[\boldsymbol{\Theta}_0] = [E(\boldsymbol{\Theta}_i)] = \boldsymbol{\Delta}^{-1}[\boldsymbol{\Gamma}_0]$. Also, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$ in our formulation are constrained to be the same across clusters: while it is possible to vary them by cluster, empirically we found that doing so leads to overfitting the data and unstable fitting. This is perhaps not too surprising given with clustered data, the cluster sizes m_i's are usually small relative to the number of clusters n. Besides, with RPFC our primary aim is on estimating the overall central subspace $[\boldsymbol{\Theta}_0]$, the covariance matrix $\boldsymbol{\Sigma}$ characterizing the heterogeneity between clusters, and predicting the random cluster-specific central subspaces $[\boldsymbol{\Theta}_i]$.

From equation (4), the marginal log-likelihood function of the RPFC model is given by

$$\begin{aligned}
\ell(\boldsymbol{\Gamma}_0, \boldsymbol{\mu}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{n} \log \left[\int \left\{ \prod_{j=1}^{m_i} \mathrm{N}(\mathbf{X}_{ij}; \boldsymbol{\mu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\beta} \mathbf{f}_{ijy}, \boldsymbol{\Delta}) \right\} \mathrm{MN}(\mathbf{V}_i; \mathbf{0}, \boldsymbol{\Gamma}, \mathbf{I}_d) d\mathbf{V}_i \right] \\
&= \sum_{i=1}^{n} \log \ell_i(\boldsymbol{\Gamma}_0, \boldsymbol{\mu}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}),
\end{aligned}$$

where the integral is over all the points on the tangent space $T_{[\boldsymbol{\Gamma}_0]}\,\mathrm{Gr}(d, p)$, and $\boldsymbol{\Gamma}_i = h(\boldsymbol{\Gamma}_0, \mathbf{V}_i)$ with $h(\cdot, \cdot)$ defined as in Section 2.1. Next, we establish a result regarding the likelihood function's invariance to transformations of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\beta}$.

Proposition 1. *For any orthogonal matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top \boldsymbol{A} = \mathbf{I}_d$, we have* $\ell(\boldsymbol{\Gamma}_0, \boldsymbol{\mu}_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}) = \ell(\boldsymbol{\Gamma}_0 \boldsymbol{A}, \boldsymbol{\mu}_i, \boldsymbol{A}^\top \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\Sigma})$.

Proposition 1 implies that for the RPFC model, only the parameters $\boldsymbol{\mu}_i, \boldsymbol{\Delta}, \boldsymbol{\Sigma}$ are identifiable. On

the other hand, although $\boldsymbol{\Gamma}_0$ is not identifiable its span $[\boldsymbol{\Gamma}_0]$ *is* identifiable, since this span is invariant to any orthogonal rotation $\boldsymbol{\Gamma}_0\mathbf{A}$. Consequently, only the cluster-specific $[\boldsymbol{\Gamma}_i]$ is identifiable, not the actual $\boldsymbol{\Gamma}_i$. The coefficient $\boldsymbol{\beta}$ is also not identifiable, but this is often not a parameter of interest in SDR.

3.1 Estimation for fixed d

We propose a two-stage procedure to fit the RPFC model in (5), assuming the structural dimension d is fixed; we come back to the issue of selecting d in Section 6. In the first stage, we estimate $[\boldsymbol{\Gamma}_0]$ by fitting a so-called global PFC model (GPFC) which ignores the clustering i.e., $\mathbf{X}_{ij}|y_{ij} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\beta}}\mathbf{f}_{ijy} + \tilde{\boldsymbol{\varepsilon}}_{ij},\ \boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, \tilde{\boldsymbol{\Delta}})$. That is, we maximize

$$\ell_g(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Delta}}) = -\frac{N}{2}\log|\tilde{\boldsymbol{\Delta}}| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\mathbf{X}_{ij} - \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\beta}}\mathbf{f}_{ijy}\right)^\top \tilde{\boldsymbol{\Delta}}^{-1}\left(\mathbf{X}_{ij} - \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\beta}}\mathbf{f}_{ijy}\right), \quad (6)$$

the details of which can be found in Cook and Forzani (2008). Theorem 1 in Section 4 proves that the maximum likelihood estimator $[\hat{\boldsymbol{\Gamma}}_0^{\text{GPFC}}]$ of $\tilde{\boldsymbol{\Gamma}}$ is a consistent estimator for $[\boldsymbol{\Gamma}_0]$, providing a justification for this approach. To ease the notation, we write $\hat{\boldsymbol{\Gamma}}_0^{\text{GPFC}}$ simply as $\hat{\boldsymbol{\Gamma}}_0$. On the other hand, note GPFC only consistently estimates $[\boldsymbol{\Gamma}_0]$, but not the overall fixed effects central subspace $\boldsymbol{\Delta}^{-1}[\boldsymbol{\Gamma}_0]$. Moreover, equation 6 clearly does not offer predictions of the cluster-specific central subspaces.

In the second stage of the estimation procedure, we begin by removing the cluster-specific intercepts $\boldsymbol{\mu}_i$ (which are nuisance parameters in the context of SDR) by subtracting each observation from its cluster mean. This leads to a revised form of the RPFC model, $\mathbf{Z}_{ij} = \mathbf{X}_{ij} - \bar{\mathbf{X}}_i = \boldsymbol{\Gamma}_i\boldsymbol{\beta}\mathbf{h}_{ijy} + \boldsymbol{\epsilon}_{ij}$ where $\bar{\mathbf{X}}_i = m_i^{-1}\sum_{j=1}^{n}\mathbf{X}_{ij}$, $\mathbf{h}_{ijy} = \mathbf{f}_{ijy} - m_i^{-1}\sum_{j=1}^{m_i}\mathbf{f}_{ijy}$, and the errors are defined as $\boldsymbol{\epsilon}_{ij} = \boldsymbol{\varepsilon}_{ij} - m_i^{-1}\sum_{j=1}^{m_i}\boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, (1-m_i^{-1})\boldsymbol{\Delta})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Next, to form the likelihood of the centered observations $\mathbf{Z}_{ij}$, observe that $\sum_{j=1}^{m_i}\mathbf{Z}_{ij} = \mathbf{0}$ and

so it suffices to form the likelihood based on $(m_i - 1)$ observations. Furthermore, the conditional covariance of any pair $(\mathbf{Z}_{ij}, \mathbf{Z}_{ij'})$ is $-m_i^{-1}\boldsymbol{\Delta}$ for $j, j' = 1, \ldots, m_i$ and $j \neq j'$. With this in mind, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$ and $\mathbf{Z}_i$ be the $p \times (m_i - 1)$ matrix whose jth column is given by $\mathbf{Z}_{ij}$. Similarly, let $\mathbf{H}_i$ be the $r \times (m_i - 1)$ matrix whose jth column is $\mathbf{h}_{ijy}$. It follows that, conditional on $\boldsymbol{\Gamma}_i$ and $\boldsymbol{y}_i$, the matrix $\mathbf{Z}_i$ follows a matrix normal distribution $\mathbf{Z}_i|(\mathbf{y}_i, \boldsymbol{\Gamma}_i) \sim \text{MN}_{p\times(m_i-1)}(\boldsymbol{\Gamma}_i\boldsymbol{\beta}\mathbf{H}_i, \boldsymbol{\Delta}, \mathbf{L}_i)$, where $\mathbf{L}_i = \mathbf{I}_{m_i-1} - m_i^{-1}\mathbf{J}_{m_i}$ where $\mathbf{J}_{m_i}$ is an $m_i \times m_i$ matrix of ones.

In the second stage of the algorithm, we use a Monte-Carlo expectation-maximization (MCEM, Wei and Tanner, 1990) algorithm to estimate the remaining parameters $\boldsymbol{\Psi} = \{\boldsymbol{\beta}^\top, \text{vech}(\boldsymbol{\Delta})^\top, \text{vech}(\boldsymbol{\Sigma})^\top\}^\top$ in the RPFC model. Write the complete data log-likelihood of the (centered) RFPC model, given $\hat{\boldsymbol{\Gamma}}_0$, as

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \left[\log\left\{\text{MN}(\mathbf{Z}_i;\ \boldsymbol{\Gamma}_i\boldsymbol{\beta}\mathbf{H}_i, \boldsymbol{\Delta}, \mathbf{L}_i)\right\} + \log\left\{\text{MN}(\mathbf{V}_i; \mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_d)\right\}\right] = \sum_{i=1}^{n} \ell_{ci}(\boldsymbol{\Psi}),$$

where $\hat{\boldsymbol{\Gamma}}_0$ is implicit in the construction of each $\boldsymbol{\Gamma}_i$ i.e., $\boldsymbol{\Gamma}_i = \text{Exp}_{[\hat{\boldsymbol{\Gamma}}_0]}(\mathbf{V}_i)$. The MCEM algorithm iterates between the following two-steps:

E-step: Let $\boldsymbol{\Psi}^{(0)}$ denote the estimates at the current iteration of the MCEM algorithm. Given independence of the clusters, the Q-function is then defined as
$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = \sum_{i=1}^{n} \int \ell_{ci}(\boldsymbol{\Psi}) p(\mathbf{V}_i; \mathbf{Z}_i, \mathbf{y}_i, \boldsymbol{\Psi}^{(0)})\, d\mathbf{V}_i$, where $p(\mathbf{V}_i; \mathbf{Z}_i, \mathbf{y}_i, \boldsymbol{\Psi}^{(0)})$ generically denotes the conditional distribution of the random effects given the observed data and current estimates. Like many applications of mixed effects models, the expectation in the Q-function does not possess a closed form, and so we utilize Monte-Carlo integration instead to approximate this. Suppose we sample T values $\mathbf{V}^t \sim \text{MN}(\mathbf{0}, \boldsymbol{\Sigma}^{(0)}, \mathbf{I}_d); t = 1, \ldots, T$. In the simulation study and application later on, we set $T = 400$. Then defining $\boldsymbol{\Gamma}^t = \text{Exp}_{[\hat{\boldsymbol{\Gamma}}_0]}(\mathbf{V}^t)$, we construct the

weights

$$\tilde{w}_{it}^{(0)} = \exp\Bigg[-\frac{m_i}{2}\log|\boldsymbol{\Delta}^{(0)}| - \frac{p}{2}\log|\mathbf{L}_i| \\ -\frac{1}{2}\operatorname{tr}\left\{\mathbf{L}_i^{-1}\left(\mathbf{Z}_{ij} - \boldsymbol{\Gamma}^t\boldsymbol{\beta}^{(0)}\mathbf{H}_i\right)^\top\left(\boldsymbol{\Delta}^{(0)}\right)^{-1}\left(\mathbf{Z}_i - \boldsymbol{\Gamma}^t\boldsymbol{\beta}^{(0)}\mathbf{H}_{iy}\right)\right\}\Bigg],$$

and $w_{it}^{(0)} = (\sum_{i=1}\tilde{w}_{it}^{(0)})^{-1}\tilde{w}_{it}^{(0)}$, such that the $w_{it}^{(0)}$'s are normalized to sum to one for each cluster. Following this, the Q-function above is approximated as

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}) \approx \sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}\left[\log\left\{\mathrm{MN}(\mathbf{Z}_i;\boldsymbol{\Gamma}^t\boldsymbol{\beta}\mathbf{H}_{iy},\boldsymbol{\Delta},\mathbf{L}_i)\right\} + \log\left\{\mathrm{MN}(\mathbf{V}^t;\mathbf{0},\boldsymbol{\Sigma},\mathbf{I}_d)\right\}\right].$$

M-step: We update the remaining parameters in the RPFC model as $\boldsymbol{\Psi}^{(1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)})$, and achieve this via a series of conditional updates. First, the update for $\boldsymbol{\Delta}$ can be straightforwardly obtained in closed-form

$$\boldsymbol{\Delta}^{(1)} = \frac{1}{N-n}\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}(\mathbf{Z}_i - \boldsymbol{\Gamma}^t\boldsymbol{\beta}^{(0)}\mathbf{H}_{iy})\mathbf{L}_i^{-1}(\mathbf{Z}_{ij} - \boldsymbol{\Gamma}^t\boldsymbol{\beta}^{(0)}\mathbf{H}_{iy})^\top.$$

Next, $\boldsymbol{\Delta}^{(1)}$, then from setting the derivative of the approximated Q-function with respect to $\boldsymbol{\beta}$ equal to zero, we obtain

$$\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}\boldsymbol{\Gamma}^{t\top}\left(\boldsymbol{\Delta}^{(1)}\right)^{-1}\boldsymbol{\Gamma}^t\boldsymbol{\beta}\mathbf{F}_{iy} = \left\{\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}\boldsymbol{\Gamma}^{t\top}\left(\boldsymbol{\Delta}^{(0)}\right)^{-1}\mathbf{G}_{iy}\right\},$$

where $\mathbf{F}_{iy} = \mathbf{H}_{iy}\mathbf{L}_i^{-1}\mathbf{H}_{iy}^\top$ and $\mathbf{G}_{iy} = \mathbf{Z}_i\mathbf{L}_i^{-1}\mathbf{H}_{iy}^\top$. By recognizing this equation has the form $\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{A}_{it}\boldsymbol{\beta}\mathbf{F}_{iy} = \mathbf{E}$, where $\mathbf{A}_{it} = w_{it}^{(0)}\boldsymbol{\Gamma}^{t\top}\left(\boldsymbol{\Delta}^{(0)}\right)^{-1}\boldsymbol{\Gamma}^t$ and $\mathbf{E} = \sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}\boldsymbol{\Gamma}^{t\top}\left(\boldsymbol{\Delta}^{(0)}\right)^{-1}\mathbf{G}_{iy}^\top$, then we can equivalently write it as $\boldsymbol{\Xi}\mathrm{vec}(\boldsymbol{\beta}) = \mathrm{vec}(\mathbf{E})$ where $\boldsymbol{\Xi} = \sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{F}_{iy}\otimes\mathbf{A}_{it}$. It follows that $\mathrm{vec}(\boldsymbol{\beta}^{(1)}) = \boldsymbol{\Xi}^{-1}\mathrm{vec}(\mathbf{E})$, and a closed-form update is obtained.

Finally, if $\boldsymbol{\Sigma}$ is not assumed to have additional structure besides being orthogonal to $\boldsymbol{\Gamma}_0$, then from straightforward algebra we can show that maximizing the approximated Q-function leads to the update $\boldsymbol{\Sigma}^{(1)} = \sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{(0)}\mathbf{V}^t(\mathbf{V}^t)^\top$. Otherwise, if $\boldsymbol{\Sigma}$ is structured and characterized by a

set of parameters Φ, then we can update these parameters correspondingly. For instance, in the simulation studies later on we consider settings where $\boldsymbol{\Sigma} = \sigma^2\mathbf{K}$ and $\mathbf{K} = \mathbf{I}_p - \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^\top$. In such case, it is straightforward to show that the update for σ^2 is given by

$$(\sigma^2)^{(1)} = \frac{1}{d(p-d)}\sum_{t=1}^{T}\text{tr}\left(w_{it}^{(0)}\widehat{\mathbf{K}}\mathbf{V}^t\mathbf{V}^{t^\top}\right) = \frac{1}{d(p-d)}\sum_{t=1}^{T}\text{tr}\left\{w_{it}^{(0)}\left(\mathbf{I}_p - \hat{\boldsymbol{\Gamma}}_0\hat{\boldsymbol{\Gamma}}_0^\top\right)\mathbf{V}^t\mathbf{V}^{t^\top}\right\}.$$

We iterate between the Monte-Carlo E-step and M-step until convergence, which we can assess (say) by successive changes in the marginal log-likelihood function, $\ell(\boldsymbol{\Psi}|\hat{\boldsymbol{\Gamma}}_0) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\Psi}|\hat{\boldsymbol{\Gamma}}_0)$, where $\ell_i(\boldsymbol{\Psi}|\hat{\boldsymbol{\Gamma}}_0) = \log\{f(\mathbf{Z}_i|\hat{\boldsymbol{\Gamma}}_0,\boldsymbol{\Psi})\} = \log(\int \text{MN}(\mathbf{Z}_i;\boldsymbol{\Gamma}_i\boldsymbol{\beta}\mathbf{H}_{iy},\boldsymbol{\Delta},\mathbf{L}_i)\,\text{MN}(\mathbf{V}_i;\mathbf{0},\boldsymbol{\Sigma},\mathbf{I}_d)d\mathbf{V}_i)$ for $i = 1,\ldots,n$, being smaller than some certain tolerance value e.g., $\left|\ell(\boldsymbol{\Psi}^{(1)}|\hat{\boldsymbol{\Gamma}}_0) - \ell(\boldsymbol{\Psi}^{(0)}|\hat{\boldsymbol{\Gamma}}_0)\right| \leq 10^{-3}$. Note $\ell(\boldsymbol{\Psi}|\hat{\boldsymbol{\Gamma}}_0)$ is straightforwardly approximated using the proposed normalized weights in the MCEM algorithm, $\ell(\boldsymbol{\Psi}^{(0)}|\hat{\boldsymbol{\Gamma}}_0) \approx \sum_{i=1}^{n}\sum_{t=1}^{T}w_{it}^{(0)}$. Let $\hat{\boldsymbol{\Psi}} = \{\hat{\boldsymbol{\beta}}^\top, \text{vech}(\hat{\boldsymbol{\Delta}})^\top, \text{vech}(\hat{\boldsymbol{\Sigma}})^\top\}^\top$ denote the estimator of $\boldsymbol{\Psi}$ upon convergence of the two-stage estimation procedure. Then the estimate of the overall fixed effects central subspace is then given by $[\hat{\boldsymbol{\Theta}}_0] = \hat{\boldsymbol{\Delta}}^{-1}[\hat{\boldsymbol{\Gamma}}_0]$.

After model fitting, we propose to predict the cluster-specific random effect central subspaces as follows. For the i-th cluster, we first compute a prediction of $\mathbf{V}_i$ on the tangent space $T_{[\boldsymbol{\Gamma}_0]}\text{Gr}(p,d)$ as the mean of the conditional distribution $\widehat{\mathbf{V}}_i = \int \mathbf{V}_i p(\mathbf{V}_i;\mathbf{X}_{ij},\boldsymbol{y}_i,\hat{\boldsymbol{\Psi}})$. Note this can be easily calculated as $\widehat{\mathbf{V}}_i \approx \sum_{t=1}^{T}w_{it}^{(\infty)}\mathbf{V}^t$, where $\mathbf{V}^t \sim \text{MN}(\mathbf{0},\hat{\boldsymbol{\Sigma}},\mathbf{I}_d)$ and $w_{it}^{(\infty)}$ denotes the normalized weights $w_{it}^{(0)}$ of the MECM algorithm evaluated at $\hat{\boldsymbol{\Psi}}$. We then obtain a prediction of the cluster-specific central subspace as $\hat{\boldsymbol{\Gamma}}_i = \text{Exp}_{\hat{\boldsymbol{\Gamma}}_0}(\tilde{V}_i)$ and $[\hat{\boldsymbol{\Theta}}_i] = \hat{\boldsymbol{\Delta}}^{-1}[\hat{\boldsymbol{\Gamma}}_i]$, for $i = 1,\ldots,n$. It is possible to construct other predictors of the cluster-specific central subspace by using alternative choices of $\widehat{\mathbf{V}}_i$ e.g., the mode or median of $p(\mathbf{V}_i;\mathbf{X}_{ij},\boldsymbol{y}_i,\widehat{\boldsymbol{\Psi}})$. In our empirical exploration, we experimented with several choices and found that using the mean of the conditional distribution tended to be the most stable and accurate.

4 Asymptotic Theory

In this section, we establish the consistency of the proposed estimators from Section 3.1 for the identifiable parameters of the RPFC model. We focus on the situation when the number of clusters $n \to \infty$, and the cluster size m_i is finite and bounded for all $i = 1, \ldots, n$.

Let $\widetilde{\boldsymbol{\Psi}}^* = (\boldsymbol{\Gamma}_0^{*\top}, \boldsymbol{\beta}^{*\top}, \text{vech}(\boldsymbol{\Delta}^*)^\top, \text{vech}(\boldsymbol{\Sigma}^*)^{*\top})$ denote the true parameter value of $\boldsymbol{\Gamma}_0$, $\boldsymbol{\beta}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ in (5). Then we first prove the consistency of $[\hat{\boldsymbol{\Gamma}}_0]$ from fitting the global PFC model in the first stage of the estimation procedure. We require the following conditions.

Condition 1. $\widehat{\boldsymbol{\Sigma}}_{ff} = N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{f}_{ij}\boldsymbol{f}_{ij}^\top \to \boldsymbol{\Sigma}_{ff}$, *where* $\boldsymbol{\Sigma}_{ff}$ *is a* $r \times r$ *positive definite matrix, when* $N \to \infty$.

Condition 2. $\widehat{\boldsymbol{\Sigma}}_{xx} = N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathbf{X}_{ij}\mathbf{X}_{ij}^\top \xrightarrow{p} \boldsymbol{\Sigma}_{xx}$, *where* $\boldsymbol{\Sigma}_{xx}$ *is a* $p \times p$ *positive definite matrix when* $N \to \infty$.

The above two conditions are mild since they essentially require $\mathbf{X}_{ij}$ and $\mathbf{f}_{yij}$ to have finite (marginal) variances.

Theorem 1. *Let* $\hat{\boldsymbol{\Gamma}}_0^{GPFC}$ *denote the maximum likelihood estimate of* $\tilde{\boldsymbol{\Gamma}}$ *from the GPFC model i.e., obtained by maximizing equation* (6)*. Assume Conditions 1 and 2 are satisfied. Then* $[\hat{\boldsymbol{\Gamma}}_0^{GPFC}] \xrightarrow{p} [\boldsymbol{\Gamma}_0^*]$ *as* $N \to \infty$.

Theorem 1 only requires the total number of observations $N \to \infty$, which is satisfied when the number of clusters $n \to \infty$. We point out that only $[\hat{\boldsymbol{\Gamma}}_0^{\text{GPFC}}]$, i.e., the space spanned by $\hat{\boldsymbol{\Gamma}}_0^{\text{GPFC}}$, is consistent for $[\hat{\boldsymbol{\Gamma}}_0]$. That is, the actual estimate $\hat{\boldsymbol{\Gamma}}_0^{\text{GPFC}}$ converges to $\boldsymbol{\Gamma}_0^*\mathbf{A}$ for some orthogonal $d \times d$ matrix $\mathbf{A}$. This result is somewhat analogous to how in the linear mixed effects model, the ordinary least square estimator (which ignores the clustered nature of the data) is a consistent (though statistically less efficient) estimator for the true fixed effects coefficients (Verbeke and

Molenberghs, 2009).

Next, we establish consistency of the estimated covariance matrices $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Delta}}$ obtained from the second stage of estimation procedure. Note consistency of $\hat{\mathbf{\Delta}}$ is required for the consistency of the estimated fixed effect central subspace. In the development below, and leveraging the result of and discussion around Theorem 1 above, we will assume the true $\mathbf{\Gamma}_0^*$ is known up to an orthogonal rotation $\mathbf{\Gamma}_0^*\mathbf{A}$. Let $\mathbf{\Psi}^* = (\mathbf{A}^\top \boldsymbol{\beta}^*, \text{vech}(\mathbf{\Delta}^*)^\top, \text{vech}(\mathbf{\Sigma}^*)^\top)$, and $s = \dim(\mathbf{\Psi}^*) = \dim(\tilde{\mathbf{\Psi}}^*)$. In this case, we will write the marginal log-likelihood (in the second stage) as $\ell(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\mathbf{A}) = \sum_{i=1}^n \ell_i(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\mathbf{A})$. Note the independence of the clusters, then without loss of generality, we assume $\ell_1(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\mathbf{A})$ satisfies the following regularity conditions.

Condition 3. *For any $d \times d$ orthogonal matrix $\boldsymbol{A}$, the true parameter $\mathbf{\Psi}^*$ is an interior point of a compact parameter space, and $\ell_1(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\boldsymbol{A})$ is distinct as a function of $\mathbf{\Psi}$.*

Condition 4. *For any $d \times d$ orthogonal matrix $\boldsymbol{A}$ and $\mathbf{\Psi}$ in an open set containing $\mathbf{\Psi}^*$, the $s \times s$ information matrix $\mathcal{I}(\mathbf{\Psi})$ with elements $\iota_{jk}(\mathbf{\Psi}) = E\left(\partial^2 \ell_1(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\boldsymbol{A})/\partial\Psi_j\partial\Psi_k\right)$ for $j,k = 1,\ldots,s$ is positive definite with all its eigenvalues bounded away from zero and infinity.*

Condition 5. *For any $d \times d$ orthogonal matrix $\boldsymbol{A}$ for all $\mathbf{\Psi}$ in an open set that contains $\mathbf{\Psi}^*$, the third derivatives $\partial^3 \ell_1(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\boldsymbol{A})/\partial\Psi_j\partial\Psi_k\partial\Psi_l$ exists, and there exists functions $M_{ijk}(\mathbf{Z})$ such that $|\partial^3 \ell_1(\mathbf{\Psi}|\mathbf{\Gamma}_0^*\boldsymbol{A})/\partial\Psi_j\partial\Psi_k\partial\Psi_l| \leq M_{ijk}(\mathbf{Z}_1)$, where $E_{\mathbf{\Psi}^*}\{M_{ijk}(\mathbf{Z})\} < \infty$ for $j,k,l = 1,\ldots,s$.*

These conditions are necessary to establish the following two results regarding the proposed estimators, and are analogous to conditions often made when studying the asymptotic properties of mixed models (e.g., Nie, 2007; Hui et al., 2017). Note in the SDR setting, since $\mathbf{\Gamma}_0^*$ is only identifiable up to an orthogonal rotation, then we need to impose conditions on any such possible rotation.

Theorem 2. *Assume Conditions 3-5 holds. If the cluster size satisfy $m_i \geq 2$ for all $i = 1,\ldots,n$*

then it holds that $\hat{\boldsymbol{\Delta}} \xrightarrow{p} \boldsymbol{\Delta}^*$ *and* $\hat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}^*$ *as* $n \to \infty$.

Theorem 2 establishes the consistency of the two covariance matrices that characterize the fixed effect central subspace and the heterogeneity among clusters. Note the result only requires the number of clusters n to diverge and each cluster to have at least two measurements; it does not require each cluster size m_i to diverge. This is in line with classical likelihood-based theory for mixed effects models, where model parameters can be consistently estimated even if the cluster sizes are bounded (e.g., Nie, 2007; Hui et al., 2017). The result also establishes the consistency of the covariance matrix $\hat{\boldsymbol{\Sigma}}$, which is key to characterizing the heterogeneity of the random cluster-specific central subspaces.

Corrolary 1. *Assume Conditions 1-5 hold. Then* $\boldsymbol{\Delta}^{-1}[\hat{\boldsymbol{\Gamma}}_0] \xrightarrow{p} [\boldsymbol{\Theta}_0^*]$ *when* $n \to \infty$.

Corrollary 1 is a direct application of the Slutsky's theorem, and guarantees consistent estimation of the overall fixed effects central subspace.

5 Simulation study

We performed a numerical study to assess the finite performance of the RPFC model i.e., for estimating the overall fixed effects central subspace, the variability of central subspaces across clusters, and predicting the random cluster-specific central subspaces. We simulated independent clustered data from the two inverse models as follows: for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, we first generated $y_{ij} \sim \mathrm{N}(0, 1)$ and set $\boldsymbol{\Delta}$ to be an AR(1) correlation matrix with autocorrelation parameter 0.5. Next, we generated $\boldsymbol{\Gamma}_i$ following (2), with $\boldsymbol{\Gamma}_0$ constructed from the QR decomposition of a random $p \times d$ matrix whose elements are generated from the uniform distribution between $(-1, 1)$, and the $p \times d$ matrices $\mathbf{V}_i$'s independently generated from the singular matrix normal distribution $\mathrm{MN}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_d)$, where $\boldsymbol{\Sigma} = \mathbf{K}\tilde{\boldsymbol{\Sigma}}\mathbf{K}$ and $\mathbf{K} = \mathbf{I}_p - \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^\top$. Finally, given $\boldsymbol{V}_i$ and

$\boldsymbol{\Gamma}_0$, we simulated the covariates as

(M1) $\mathbf{X}_{ij} = \boldsymbol{\Gamma}_i v_{yij} + \boldsymbol{\varepsilon}_{ij}, v_{yij} = y_{ij} + (1/2)y_{ij}^2 + (1/3)y_{ij}^3; \;\; \boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Delta})$,

(M2) $\mathbf{X}_{ij} = \boldsymbol{\Gamma}_i v_{yij} + \boldsymbol{\varepsilon}_{ij}, v_{yij} = [y_{ij} + (1/2)y_{ij}^2 + (1/3)y_{ij}^3,\ y_{ij}]; \;\; \boldsymbol{\varepsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Delta})$.

Note the structural dimension for the overall and cluster-specific central subspaces is $d = 1$ and $d = 2$ for models M1 and M2, respectively, and is assumed to be known. In both models, we set $p = 7$ covariates and considered two settings for the random effects covariance matrix, $\boldsymbol{\Sigma}$: (a) we set and assumed in the fitting process an isotropic structure, $\boldsymbol{\Sigma} = \sigma^2\mathbf{K}$ and $\sigma^2 \in \{0.04, 0.10, 0.50\}$; (b) we generated $\boldsymbol{\Sigma} = \mathbf{K}\tilde{\boldsymbol{\Sigma}}\mathbf{K}$ with three possible structures of $\tilde{\boldsymbol{\Sigma}}$, namely a diagonal form $\tilde{\boldsymbol{\Sigma}} = 0.3\mathbf{I}_p$, an AR(1) form for $\tilde{\boldsymbol{\Sigma}}$ with variance set to 0.3 and autocorrelation parameter set to 0.5, and an exchangeable structure for $\tilde{\boldsymbol{\Sigma}}$ where all diagonal elements are set to 0.3 and all off-diagonal elements set to 0.2. For both models and settings, we then fitted the RPFC model in (5) assuming the random effects covariance matrix to be unstructured i.e., no assumption about $\boldsymbol{\Sigma}$ is made except the requirement $\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0 = \mathbf{0}$.

For each combination of models M1 and M2 and the two settings of $\boldsymbol{\Sigma}$, the true mean and cluster-specific central subspaces are given by $[\boldsymbol{\Theta}_0] = \boldsymbol{\Delta}^{-1}[\boldsymbol{\Gamma}_0]$ and $[\boldsymbol{\Theta}_i] = \boldsymbol{\Delta}^{-1}[\boldsymbol{\Gamma}_i]$, respectively. We set the number of clusters to $n \in \{100, 500\}$, and for each n simulated the cluster sizes m_i randomly to be integers between 10 and 15 inclusive. We generated 200 simulated datasets per simulation setting. For each simulated dataset, we compare the performance of RPFC with the following two alternatives:

- Global fixed-effects PFC (GPFC) model: This method ignores the clustered nature of the data, instead assuming the central subspace to be the same across all clusters and applying a single PFC model to the entire dataset. The estimator for both the overall central subspace and all cluster-specific central subspaces is hence the same and given by

$\hat{\boldsymbol{\Theta}}_0^{\text{GPFC}} = \hat{\boldsymbol{\Theta}}_i^{\text{GPFC}} = \left(\hat{\boldsymbol{\Delta}}^{\text{GPFC}}\right)^{-1}[\hat{\boldsymbol{\Gamma}}_0]$, where we note $\hat{\boldsymbol{\Gamma}}_0$ is given by 6. By construction, GPFC does not produce estimators of random effect covariance matrix $\boldsymbol{\Sigma}$.

- Separate fixed-effects PFC (SPFC) model: This method fits a separate PFC model to cluster. That is, for $i = 1, \ldots, n$ we maximize

$$\ell_i(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Gamma}}_i, \tilde{\boldsymbol{\beta}}_i, \tilde{\boldsymbol{\Delta}}_i) = -\frac{m_i}{2}\log|\tilde{\boldsymbol{\Delta}}_i| - \frac{1}{2}\sum_{j=1}^{m_i}\left(\mathbf{X}_{ij} - \tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\Gamma}}_i\tilde{\boldsymbol{\beta}}_i\mathbf{f}_{ijy}\right)^\top \tilde{\boldsymbol{\Delta}}_i^{-1}\left(\mathbf{X}_{ij} - \tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\Gamma}}_i\tilde{\boldsymbol{\beta}}_i\mathbf{f}_{ijy}\right). \quad (7)$$

  Let $\hat{\boldsymbol{\Delta}}_i^{\text{SPFC}}$ and $\hat{\boldsymbol{\Gamma}}_i^{\text{SPFC}}$ denote the resulting estimates for $\tilde{\boldsymbol{\Delta}}_i$ and $\tilde{\boldsymbol{\Gamma}}_i$, respectively. Then the cluster-specific central subspace from SPFC is given by $[\hat{\boldsymbol{\Theta}}_i^{\text{SPFC}}] = \left(\hat{\boldsymbol{\Delta}}_i^{\text{SPFC}}\right)^{-1}[\hat{\boldsymbol{\Gamma}}_i^{\text{SPFC}}]$. Furthermore, a reasonable estimator for the overall central subspace is given by the intrinsic sample mean of these estimates. That is, $[\hat{\boldsymbol{\Theta}}_0^{\text{SPFC}}] = \text{argmin}_{[\boldsymbol{\Theta}_0]}\sum_{i=1}^n \Phi(\hat{\boldsymbol{\Theta}}_i^{\text{SPFC}}, \boldsymbol{\Theta}_0)$ where $\Phi(\cdot,\cdot)$ is the Riemann metric on the Grassmann manifold reviewed in Section 2.1. Finally, an estimate of $\boldsymbol{\Sigma}$ for SPFC can be obtained by first performing an inverse exponential mapping (see Section 2.1) from $[\hat{\boldsymbol{\Theta}}_i^{\text{SPFC}}]$ to the tangent space of $\text{Gr}(p,d)$ at $[\hat{\boldsymbol{\Theta}}_0^{\text{SPFC}}]$. Letting $\widehat{\boldsymbol{W}}_i$ denote the image of that map, then we have $\hat{\boldsymbol{\Sigma}}^{\text{SPFC}} = n^{-1}\sum_{i=1}^n \widehat{\boldsymbol{W}}_i\widehat{\boldsymbol{W}}_i^\top$.

For all three inverse regression estimators compared, we constructed f_{yij} from polynomial bases with degree $r = 4$ and centered them within each cluster such that $\sum_{j=1}^{m_i} f_{yij} = 0$. We assessed performance using the following three measures: (1) for estimating the overall central subspace, we calculated the Frobenius norm of the difference between the projection matrix formed from the estimate, and the projection matrix formed from the corresponding true value. That is, for each simulated dataset we computed $\|\mathcal{P}(\hat{\boldsymbol{\Theta}}_0) - \mathcal{P}(\boldsymbol{\Theta}_0)\|_F$, where $\mathcal{P}(\boldsymbol{C}) = \boldsymbol{C}(\boldsymbol{C}^\top\boldsymbol{C})^{-1}\boldsymbol{C}$ for a generic matrix $\boldsymbol{C}$, and $\|\cdot\|_F$ denotes the Frobenius norm; (2) for estimating the random effects covariance matrix, we computed $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$; (3) for predicting the cluster-specific central subspace, we computed the average Frobenius form $n^{-1}\sum_{i=1}^n \|\mathcal{P}(\hat{\boldsymbol{\Theta}}_i) - \mathcal{P}(\boldsymbol{\Theta}_i)\|_F$.

5.1 Results

When the random effect covariance $\mathbf{\Sigma}$ is (known to be) isotropic in structure, Table 1 demonstrates that RPFC had the overall best performance. For estimating the overall central subspace $[\mathbf{\Theta}_0]$, SPFC had the poorest performance, while when the clusters relative homogeneous (e.g., $\sigma^2 = 0.04$) RPFC still produced a lower estimation error than GPFC. This latter result is interesting because while both RPFC and GPFC use the same estimator for $\mathbf{\Gamma}_0$, the former estimates the overall central subspace better due to its superior performance at recovering the residual covariance matrix $\mathbf{\Delta}$. When $\sigma^2 = 0.50$, RPFC had worse performance than GPFC and SPFC when $n = 100$, but its performance greatly improved with increasing n and tended to be similar to GPFC for the two larger number of clusters tested. Turning to estimation of $\mathbf{\Sigma}$, RPFC consistently outperformed SPFC, and its estimation error decreased with increasing n while this does not occur for SPFC. Finally, for predicting the cluster-specific central subspaces, RPFC consistently outperformed SPFC especially when σ^2 was small. This reflected the former's capacity to borrow strength across clusters, which in turn improved the overall performance at predicting the $[\mathbf{\Theta}_i]$'s. When the amount of heterogeneity between clusters increased, RPFC was still able to predict $[\mathbf{\Theta}_i]$ better than SPFC, although the differences between the two methods became smaller. Note the performance relating to prediction of the cluster-specific central subspaces does not tend to decrease substantially when the number of clusters n increases: this is a consequence of the cluster sizes m_i being bounded as n grows in our simulation design.

In the second setting when the random effects covariance matrix $\mathbf{\Sigma}$ was misspecified and assumed to be unstructured, Table 2 demonstrates that RPFC retains the best overall performance. Not surprisingly, compared to the previous setting when $\mathbf{\Sigma}$ is (known to be) isotropic, the error from estimating an unstructured $\mathbf{\Sigma}$ tended to be larger. The performance of RPFC depended on the form of $\tilde{\mathbf{\Sigma}}$: for estimating the mean central subspace, RPFC exhibited the best performance when

Table 1: Simulation results for setting (a) with an isotropic random effects covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{K}$. The methods compared include the random effects PFC (RPFC), global PFC (GPFC), and separate PFC (SPFC) models. Performance is assessed in terms of estimating the overall central subspace $[\mathbf{\Theta}_0]$, the random effects covariance matrix $\tilde{\mathbf{\Sigma}}$, and predicting the cluster-specific central subspaces $[\mathbf{\Theta}_i]$. In each case, the mean Frobenius error (averaged across the 200 simulated datasets) is shown, while the corresponding standard errors is given in parentheses. The lowest values for each setting in each row is highlighted.

			$[\mathbf{\Theta}_0]$			$\mathbf{\Sigma}$		$[\mathbf{\Theta}_i]$	
σ^2	Model	n	RPFC	GPFC	SPFC	RPFC	SPFC	RPFC	SPFC
0.04	M1	100	**0.25 (0.08)**	0.34 (0.10)	1.27 (0.26)	**0.01 (0.00)**	0.40 (0.06)	**0.74 (0.11)**	1.15 (0.04)
		500	**0.12 (0.04)**	0.28 (0.08)	1.14 (0.23)	**0.01 (0.00)**	0.40 (0.06)	**0.71 (0.10)**	1.14 (0.04)
		1000	**0.09 (0.03)**	0.27 (0.07)	1.10 (0.21)	**0.01 (0.00)**	0.39 (0.05)	**0.72 (0.09)**	1.14 (0.04)
	M2	100	0.83 (0.34)	**0.82 (0.29)**	2.46 (0.35)	**0.03 (0.01)**	0.45 (0.04)	**1.08 (0.11)**	1.52 (0.05)
		500	**0.41 (0.15)**	0.52 (0.12)	2.45 (0.34)	**0.02 (0.01)**	0.46 (0.04)	**1.04 (0.08)**	1.52 (0.04)
		1000	**0.30 (0.10)**	0.45 (0.11)	2.35 (0.41)	**0.02 (0.01)**	0.45 (0.06)	**1.01 (0.11)**	1.49 (0.09)
0.1	M1	100	**0.40 (0.13)**	0.62 (0.17)	1.35 (0.22)	**0.04 (0.01)**	0.31 (0.05)	**0.77 (0.06)**	1.17 (0.03)
		500	**0.22 (0.06)**	0.57 (0.14)	1.23 (0.23)	**0.03 (0.01)**	0.32 (0.05)	**0.75 (0.05)**	1.17 (0.03)
		1000	**0.24 (0.25)**	0.59 (0.21)	1.26 (0.35)	**0.04 (0.01)**	0.32 (0.05)	**0.77 (0.12)**	1.18 (0.09)
	M2	100	**1.25 (0.45)**	**1.25 (0.36)**	2.44 (0.34)	**0.08 (0.03)**	0.39 (0.04)	**1.25 (0.08)**	1.55 (0.03)
		500	**0.68 (0.26)**	0.89 (0.19)	2.41 (0.32)	**0.04 (0.01)**	0.39 (0.04)	**1.20 (0.06)**	1.55 (0.02)
		1000	**0.50 (0.18)**	0.83 (0.14)	2.48 (0.35)	**0.03 (0.01)**	0.40 (0.04)	**1.18 (0.05)**	1.55 (0.02)
0.5	M1	100	2.04 (0.42)	**1.79 (0.35)**	1.82 (0.31)	**0.61 (0.07)**	0.81 (0.03)	**0.85 (0.03)**	1.21 (0.02)
		500	1.72 (0.52)	**1.65 (0.39)**	1.74 (0.31)	**0.53 (0.10)**	0.80 (0.03)	**0.84 (0.02)**	1.20 (0.01)
		1000	**1.44 (0.52)**	1.49 (0.35)	1.63 (0.30)	**0.47 (0.11)**	0.80 (0.03)	**0.83 (0.01)**	1.21 (0.01)
	M2	100	2.52 (0.37)	**2.32 (0.34)**	2.51 (0.34)	**0.76 (0.09)**	0.83 (0.05)	**1.47 (0.02)**	1.59 (0.02)
		500	2.31 (0.36)	**2.29 (0.31)**	2.51 (0.31)	**0.69 (0.08)**	0.82 (0.04)	**1.47 (0.01)**	1.59 (0.01)
		1000	**2.20 (0.34)**	**2.20 (0.31)**	2.46 (0.31)	**0.66 (0.07)**	0.82 (0.04)	**1.47 (0.01)**	1.59 (0.01)

$\tilde{\mathbf{\Sigma}}$ had an exchangeable structure, but for estimating the random effect covariance matrix $\mathbf{\Sigma}$, RPFC had the lowest estimation error when $\tilde{\mathbf{\Sigma}}$ was diagonal in structure. The estimation errors for both the mean central subspace and random effect covariance decreased noticeably when the number of clusters n increased for RPFC, but not for SPFC. Finally, for predicting cluster-specific central subspaces, RPFC again had consistently lower errors than SPFC across all three structures of $\tilde{\mathbf{\Sigma}}$ tested.

Table 2: Simulation results for setting (b) with an unstructured random effects covariance matrix $\boldsymbol{\Sigma}$. The methods compared include the random effects PFC (RPFC), global PFC (GPFC), and separate PFC (SPFC) models. Performance is assessed in terms of estimating the mean central subspace $[\boldsymbol{\Theta}_0]$, the random effects covariance matrix $\tilde{\boldsymbol{\Sigma}}$, and predicting the cluster-specific central subspaces $[\boldsymbol{\Theta}_i]$. In each case, mean average Frobenius error (averaged across the 200 simulated datasets) is shown, while the corresponding standard errors is given in parentheses. The lowest values for each setting in each row is highlighted.

			$[\boldsymbol{\Theta}_0]$			$\boldsymbol{\Sigma}$		$[\boldsymbol{\Theta}_i]$	
$\tilde{\boldsymbol{\Sigma}}$	Model	n	RPFC	GPFC	SPFC	RPFC	SPFC	RPFC	SPFC
Diagonal	M1	100	**1.22 (0.44)**	1.34 (0.37)	1.64 (0.32)	**0.49 (0.13)**	**0.49 (0.06)**	**0.88 (0.03)**	1.21 (0.02)
		500	**0.69 (0.22)**	1.14 (0.29)	1.47 (0.25)	**0.38 (0.04)**	0.46 (0.05)	**0.86 (0.01)**	1.20 (0.01)
		1000	**0.54 (0.14)**	1.10 (0.27)	1.44 (0.22)	**0.36 (0.04)**	0.45 (0.04)	**0.86 (0.01)**	1.20 (0.01)
	M2	100	**2.02 (0.38)**	2.06 (0.31)	2.50 (0.34)	**0.46 (0.08)**	0.55 (0.07)	**1.46 (0.03)**	1.60 (0.02)
		500	**1.57 (0.44)**	1.83 (0.31)	2.49 (0.33)	**0.37 (0.06)**	0.55 (0.06)	**1.43 (0.02)**	1.59 (0.01)
		1000	**1.28 (0.37)**	1.66 (0.29)	2.43 (0.34)	**0.33 (0.05)**	0.55 (0.07)	**1.42 (0.02)**	1.59 (0.01)
AR(1)	M1	100	**1.14 (0.42)**	1.37 (0.39)	1.55 (0.30)	**0.79 (0.19)**	**0.79 (0.08)**	**0.99 (0.04)**	1.24 (0.02)
		500	**0.63 (0.21)**	1.22 (0.33)	1.40 (0.27)	**0.69 (0.14)**	0.79 (0.06)	**0.99 (0.03)**	1.24 (0.01)
		1000	**0.52 (0.15)**	1.19 (0.32)	1.33 (0.23)	**0.70 (0.16)**	0.79 (0.06)	**0.99 (0.03)**	1.23 (0.01)
	M2	100	**1.98 (0.40)**	2.09 (0.37)	2.50 (0.32)	**0.68 (0.13)**	0.74 (0.09)	**1.52 (0.03)**	1.63 (0.02)
		500	**1.49 (0.46)**	1.83 (0.37)	2.45 (0.32)	**0.62 (0.13)**	0.74 (0.09)	**1.49 (0.03)**	1.62 (0.01)
		1000	**1.24 (0.42)**	1.76 (0.35)	2.49 (0.34)	**0.60 (0.12)**	0.75 (0.09)	**1.47 (0.03)**	1.62 (0.01)
Exchangeable	M1	100	**0.93 (0.41)**	1.38 (0.38)	1.52 (0.32)	**1.16 (0.49)**	1.27 (0.22)	**0.99 (0.07)**	1.23 (0.02)
		500	**0.55 (0.20)**	1.31 (0.34)	1.34 (0.25)	**1.07 (0.54)**	1.28 (0.20)	**0.99 (0.08)**	1.23 (0.02)
		1000	**0.45 (0.19)**	1.23 (0.34)	1.25 (0.21)	**1.07 (0.54)**	1.27 (0.21)	**0.99 (0.08)**	1.23 (0.01)
	M2	100	**1.74 (0.43)**	1.99 (0.38)	2.44 (0.34)	**1.02 (0.34)**	1.09 (0.27)	**1.48 (0.06)**	1.62 (0.02)
		500	**1.32 (0.39)**	1.90 (0.34)	2.50 (0.34)	**0.96 (0.35)**	1.11 (0.25)	**1.43 (0.07)**	1.60 (0.02)
		1000	**1.07 (0.39)**	1.77 (0.39)	2.44 (0.36)	**0.92 (0.36)**	1.11 (0.26)	**1.42 (0.07)**	1.60 (0.02)

6 Selecting the structural dimension

For the standard PFC model without random effects, Cook and Forzani (2008) proposed to select the structural dimension d by either a likelihood ratio test or via an information criteria. In this article, we adopt the latter approach when it comes to selecting d for RPFC (see also Ma and Zhang, 2015; Luo and Li, 2021, for examples of where information criteria are employed to select the structural dimension in SDR). In particular, we propose two computationally efficient approaches that allow for the selection of d to be made *prior* to fitting the (second stage of the) RPFC model.

In the first approach, since the structural dimension is equal to the dimension of $[\hat{\boldsymbol{\Gamma}}_0]$, which

is estimated via a GPFC model in the first stage of the estimation procedure in Section 3.1, we consider formulating an information criterion directly from this GPFC model. Let $w \in \{0, 1, \ldots, \min(r, p)\}$ be a candidate for d. Then after applying GPFC model with this candidate choice, let $\ell_g(w) = \ell_g(w; \hat{\boldsymbol{\mu}}^{\text{GPFC}}, \hat{\boldsymbol{\Gamma}}^{\text{GPFC}}, \hat{\boldsymbol{\beta}}^{\text{GPFC}}, \tilde{\boldsymbol{\Delta}}^{\text{GPFC}})$ be the corresponding value of the maximized log-likelihood function in (6). Noting the corresponding number of parameters involved is $h(w) = p(p+3)/2 + rw + w(p-w)$, we construct an information criterion of the form $-2\ell_g(w) + \tau h(w)$ where we consider the model complexity as $\tau = 2$ or $\tau = \log(N)$ corresponding to the global Akaike information criterion (GAIC) and the global Bayesian information criterion (GBIC) respectively. We select d that minimizes either GAIC or GBIC.

In the second approach, by noting that the structural dimension is the same across all clusters in the RPFC model, we consider selecting d via the SPFC model discussed in Section 5. Specifically, after fitting SPFC with a dimension candidate w, let $\ell_i(w; \hat{\boldsymbol{\mu}}_i^{\text{SPFC}}, \hat{\boldsymbol{\Gamma}}_i^{\text{SPFC}}, \hat{\boldsymbol{\beta}}_i^{\text{SPFC}}, \hat{\boldsymbol{\Delta}}_i^{\text{SPFC}})$ denote the corresponding value of the maximized log-likelihood function for the i-th cluster i.e., the maximized value of (7). Then an information criterion for the i-th cluster can be defined as $-2\ell_i(w) + \tau_i h(w)$, and we select d by minimizing $\sum_{i=1}^{n} \{-2\ell_i(w) + \tau_i h(w)\}$. As in the first approach, we consider a specific Akaike information criterion (SAIC) by setting all $\tau_i = 2$, and a specific Bayesian information criterion (SBIC) by setting all $\tau_i = \log(m_i)$.

To reiterate, we adopt these methods to select d given their computational efficiency: instead of fitting the RPFC model to each candidate d using the two-stage estimation procedure outlined in Section 3.1, we select d using either the GPFC and SPFC models, which are computationally very scalable to fit.

To assess the finite sample performance of the above procedure, we conduct a simulation study where clustered data were generated from the two inverse regression models as in Section 5, but focusing on the second setting where the random effects covariance matrix $\boldsymbol{\Sigma}$ is assumed

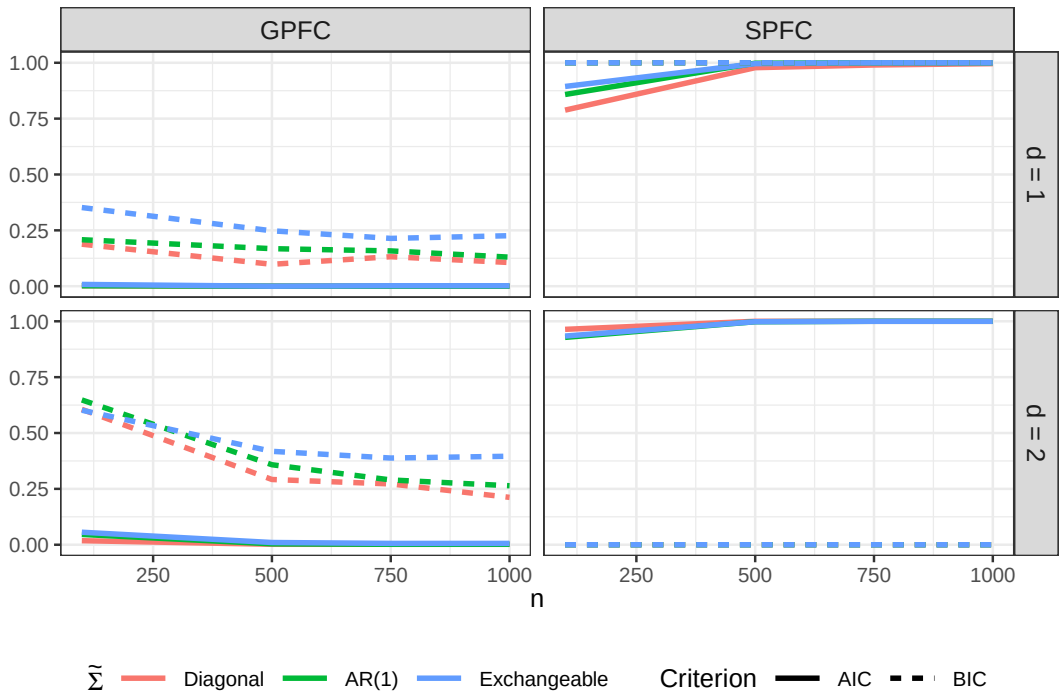

Figure 2: Proportion of 500 simulated datasets where the proposed information criteria methods selected the correct structural dimension for the RPFC model. The top and bottom panels correspond to models M1 and M2, whose true d equals 1 and 2 respectively.

to be unstructured but takes on one of three possible forms (diagonal, an autoregressive form, and an exchangeable structure). Recall the true structural dimension of models M1 and M2 is $d = 1$ and 2, respectively. Figure 2 shows that SAIC exhibits the best performance among the four considered criterion for selecting d. It is also the only method whose performance tended to improve when the number of clusters n increases. The two information criteria derived from the GPFC model exhibited poor performance, possibly because GAIC and GBIC are constructed from a model that ignores the heterogeneity among clusters and thus effectively underfits the data. Finally, the SBIC consistently chose $\hat{d} = 1$, meaning its model complexity penality was likely too severe. We leave theoretical investigation of the proposed information criteria for selecting d in the RPFC model as an avenue for future research.

7 Application

We applied RPFC to perform random effects sufficient dimension reduction on socioeconomic data extracted from the Gapminder database. Briefly, the data contain multiple socioeconomic variables of $n = 117$ countries collected in the years 1990–2015, and are available at `https://open-numbers.github.io/`. In this analysis, we focused on modeling the relationship between life expectancy of women across the countries as a function of $p = 6$ predictors: X_1 = log income per capita, X_2 = sex ratio i.e., number of females per 100 males across all age groups, X_3 = infant mortality rate per 1000 new births, X_4 = emissions consumption per person, X_5 = the average children per woman, and X_6 = income inequality via Gini index. We treat each country as an independent cluster, with the number of repeated measurements m_i ranging from 21 to 26.

Let y_{ij} and $\mathbf{X}_{ij}$ denote the j-th measurement for the life expectancy of women, and for the six predictors, of the i-th country respectively. Using SAIC, the number of structural dimensions was chosen to be $\hat{d} = 2$. Applying RPFC with this choice produced the resulting semi-orthogonal basis matrix characterizing the estimate of the overall fixed effects subspace

$$\hat{\boldsymbol{\Theta}}_0^\top = \begin{bmatrix} -0.993 & -0.047 & 0.055 & 0.007 & 0.044 & -0.086 \\ 0.114 & -0.216 & 0.576 & 0.084 & 0.100 & -0.769 \end{bmatrix}$$

We see that across all clusters/countries, the first sufficient predictor is influenced mostly by log income per capita, and the second sufficient predictor is primarily driven by infant mortality and

income inequality. RPFC produced an estimated random effects covariance matrix as

$$\hat{\boldsymbol{\Sigma}} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \begin{bmatrix} 0.004 & -0.001 & -0.002 & 0.000 & 0.004 & 0.004 \\ -0.001 & 0.004 & 0.005 & 0.000 & -0.008 & -0.001 \\ -0.002 & 0.005 & 0.031 & -0.000 & -0.050 & -0.002 \\ 0.000 & 0.000 & -0.000 & 0.002 & -0.000 & 0.011 \\ 0.004 & -0.008 & -0.050 & -0.000 & 0.080 & 0.002 \\ 0.004 & -0.001 & -0.002 & 0.011 & 0.002 & 0.060 \end{bmatrix} \end{array}.$$

Examining its diagonal elements, it is interesting to note that infant mortality (X_3), children per woman (X_5), and income inequality (X_6) exhibited the largest values, suggesting they were responsible for driving heterogeneity among countries in terms of the cluster-specific central subspaces. To further study the extent of this heterogeneity, we predicted the random cluster-specific central subspaces for each country given by RPFC, and summarized this by reporting the importance of each predictor for each subspace based on the corresponding diagonal elements of the estimated projection matrix. That is, let $[\hat{\boldsymbol{\Theta}}_i] = \hat{\boldsymbol{\Delta}}^{-1}[\hat{\boldsymbol{\Gamma}}_i]$ denote the prediction of the cluster-specific central subspace for the i-th country. We then computed the $p \times p$ projection matrix $\mathcal{P}(\hat{\boldsymbol{\Theta}}_i)$ and use its diagonal elements to evaluate the importance of each covariate (across both structural dimensions) for the i-th country. Note each diagonal element of a projection matrix is between zero and one, and the higher the value the more important a covariate is (see also Tan et al., 2018; Nghiem et al., 2023, for examples of this metric's usage elsewhere in SDR).

Figure 3 displays the resulting importance of each predictor for select countries using RPFC, as well as the corresponding variable importance for the overall fixed effects central subspace (bottom row). Consistent with the estimated $\hat{\boldsymbol{\Theta}}_0$ above, the cluster-specific central subspaces

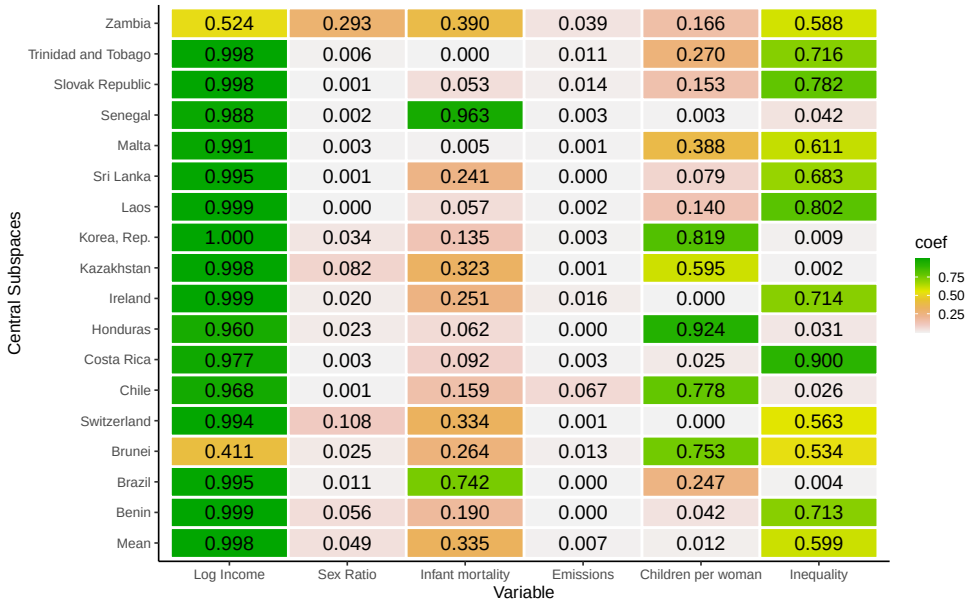

Figure 3: Importance of the $p = 6$ predictors in driving cluster-specific central subspaces for select countries, as well as for the overall fixed effects central space (bottom row), as obtained from RPFC. An orange-green color gradient is used to indicate variable importance, with dark green representing stronger importance.

depended most importantly on log income per capita, infant mortality, and income inequality. However, the importance of these three remaining predictors varied greatly from one country to another. For instance, infant mortality was important in Senegal but not in Mauritius and Nicaragua, while the number of children per woman was important in Honduras but not in Ireland, Switzerland, and Poland.

For a point of comparison, we constructed the same metrics by applying SPFC, and results analogous to those presented in Figure 3 can be found in the Appendix. Moreover, Figure 4 compares the distributions of each variable's importance across cluster-specific central subspaces from the two approaches. Interestingly, the distributions given by SPPC tended to exhibit more variability than those given by RPFC for the first four predictors, but the trend is somewhat reversed for the average childern per woman and income inequality. While sex ratio was very important in SPFC across most countries, it tended to be the least important for RPFC. On the other hand, log income was highly important for RPFC but its importance varied greatly from one country to another for

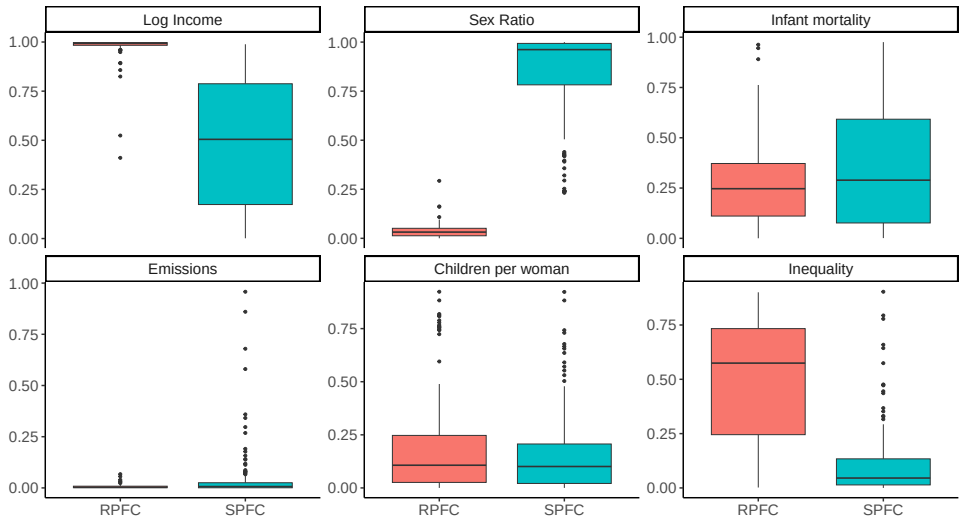

Figure 4: Boxplots of the importance of the six predictors for RPFC and SPFC, as measured by the corresponding diagonal elements in the projection matrix associated with cluster-specific central subspaces, across all $n = 117$ countries.

SPFC. The relative patterns of variability for RPFC were consistent with what was seen in the estimate of $\hat{\mathbf{\Sigma}}$, e.g., log income, sex ratio, and emissions consumption per person exhibited less variability across countries compared with the other three variables.

Finally, in the Supplementary Materials, we present scatterplots of the response versus cluster-specific sufficient predictors i.e., $\mathbf{Z}_i = \mathbf{X}_i^\top \hat{\mathbf{\Theta}}_i$, for selected countries. Results show that female life expectancy exhibits a strong, sometimes close to linear relationship with one or both of the two sufficient predictors across many countries. In some countries such as Malta and Canada, the first sufficient predictor is more informative about the response than the second one. This is reversed though for other countries such as Bangladesh, while in some other countries e.g., Laos, both sufficient predictors are strongly and close to linearly associated with female life expectancy. Overall, these results (further) demonstrate the heterogeneity across countries in terms of the sufficient dimension reduction.

8 Discussion

We proposed random effects sufficient dimension reduction in clustered data settings, where the cluster-specific central subspaces are assumed to follow a common distribution on a Grassmann manifold. By defining the distribution as the image of an exponential mapping of a distribution defined on the tangent space of the Grassmann manifold at an overall fixed effect central subspace, this facilitates computational efficiency and straightforward interpretation similar to the traditional mixed effect models. We incorporated random effects SDR in the context of the principal fitted component (PFC) model, and proposed a two-step fitting procedure for the resulting random effects PFC model to estimate the overall central subspace and the covariance matrix characterizing the heterogeneity between clusters, and to predict the random cluster-specific central subspaces. Simulation studies show the random effects PFC model has superior performance to both a global fixed-effects PFC model ignoring all the clustering nature of the data, and a cluster-specific fixed-effects PFC model that does not share information across clusters. We applied RPFC to analyze the relationship between female life expectancy and various socioeconomic variables across different countries over years.

Multiple extensions can be made to enhance and broaden the applicability of RPFC and random effects SDR in general. One can generalize the inverse predictor $\mathbf{X}_{ij}|y_{ij}$ from being normally distributed to being from an exponential family, such as those developed in Cook and Li (2009). This extension would allow an SDR model on the covariates with mixed types of variables, such as categorical, count, and continuous. Theoretically, future work could study the prediction error of the cluster-specific central spaces e.g., the behavior of $n^{-1}\sum_{i=1}^{n}\|\mathcal{P}(\hat{\boldsymbol{\Theta}}_i)-\mathcal{P}(\boldsymbol{\Theta}_i)\|_F$ when the cluster sizes m_i are also growing n, establish the order consistency of SAIC (say) in selecting d, and develop inferential properties for the random effects covariance matrix $\boldsymbol{\Sigma}$ that account for potential boundary issues). Finally, in terms of modeling the random effect central subspaces, one

can explore more structures for the covariance matrix $\mathbf{\Sigma}$ on the tangent space of the Grassmann manifold at the fixed effect central subspace.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains the proof of all theoretical results in Sections 2-4 and additional results for the application in Section 7.

References

Bendokat, T., R. Zimmermann, and P.-A. Absil (2020). A grassmann manifold handbook: Basic geometry and computational aspects. *arXiv preprint arXiv:2011.13699*.

Bi, X. and A. Qu (2015). Sufficient dimension reduction for longitudinal data. *Statistica Sinica 25*, 787–807.

Bjork, J. R., F. K. C. Hui, R. B. O'Hara, and J. M. Montoya (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular ecology 27*, 2714–2724.

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

Chikuse, Y. (2003a). Concentrated matrix langevin distributions. *Journal of Multivariate Analysis 85*(2), 375–394.

Chikuse, Y. (2003b). *Statistics on special manifolds*, Volume 1. Springer.

Cook, R. D. and L. Forzani (2008). Principal Fitted Components for Dimension Reduction in Regression. *Statistical Science 23*(4), 485 – 501.

Cook, R. D. and L. Li (2009). Dimension reduction in regressions with exponential family predictors. *Journal of Computational and Graphical Statistics 18*(3), 774–791.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012). *Applied longitudinal analysis*. John Wiley & Sons.

Gupta, A. K. and D. K. Nagar (1999). *Matrix variate distributions*, Volume 104. CRC Press.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of econometrics 109*(1), 107–150.

Hui, F. K. C., S. Mueller, and A. H. Welsh (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica 27*, 501–518.

Hui, F. K. C. and L. H. Nghiem (2022). Sufficient dimension reduction for clustered data via finite mixture modelling. *Australian & New Zealand Journal of Statistics 64*(2), 133–157.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. Florida: CRC Press.

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

Li, K.-C. and N. Duan (1989). Regression analysis under link violation. *The Annals of Statistics 17*, 1009–1052.

Lin, Q., Z. Zhao, and J. S. Liu (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics 46*, 580–610.

Luo, W. and B. Li (2021). On order determination by predictor augmentation. *Biometrika 108*(3), 557–574.

Ma, Y. and X. Zhang (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika 102*, 409–420.

Ma, Y. and L. Zhu (2013). A review on dimension reduction. *International Statistical Review 81*, 134–150.

Nghiem, L. H., F. K. C. Hui, S. Müller, and A. H. Welsh (2023). Sparse sliced inverse regression via cholesky matrix penalization. *Statistica Sinica 33*, 1–33.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *Journal of Statistical Planning and Inference 137*, 1787–1804.

Pang, Z. and L. Xue (2012). Estimation for the single-index models with random effects. *Computational Statistics & Data Analysis 56*, 1837–1853.

Pfeiffer, R. M., D. B. Kapla, and E. Bura (2021). Least squares and maximum likelihood estimation of sufficient reductions in regressions with matrix-valued predictors. *International journal of data science and analytics 11*, 11–26.

Rodrigues, S. A., R. Huggins, and B. Liquet (2022). Central subspaces review: methods and applications. *Statistics Surveys 16*(none), 210 – 237.

Scealy, J. and A. T. Wood (2019). Scaled von mises–fisher distributions and regression models for paleomagnetic directional data. *Journal of the American Statistical Association*.

Scealy, J. L. and A. T. Wood (2022). Score matching for compositional distributions. *Journal of the American Statistical Association*, 1–13.

Song, M., E. Bura, R. Parzer, and R. M. Pfeiffer (2023). Structured time-dependent inverse regression (stir). *Statistics in Medicine 42*(9), 1289–1307.

Srivastava, A. and E. P. Klassen (2016). *Functional and shape data analysis*, Volume 1. Springer.

Tan, K. M., Z. Wang, T. Zhang, H. Liu, and R. D. Cook (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika 105*(4), 769–782.

Tian, Z. and P. Qiu (2023). Multivariate single index modeling of longitudinal data with multiple responses. *Statistics in Medicine*.

Verbeke, G. and G. Molenberghs (2009). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B 64*, 363–410.

Xu, K., W. Guo, M. Xiong, L. Zhu, and L. Jin (2016). An estimating equation approach to dimension reduction for longitudinal data. *Biometrika 103*, 189–203.

Zimmermann, R. (2017). A matrix-algebraic algorithm for the riemannian logarithm on the stiefel manifold under the canonical metric. *SIAM Journal on Matrix Analysis and Applications 38*(2), 322–342.