Likelihood-based Sufficient Dimension Reduction

R.Dennis Cook* and Liliana Forzani†

University of Minnesota and Instituto de Matemática Aplicada del Litoral

September 25, 2008

Abstract

We obtain the maximum likelihood estimator of the central subspace under conditional normality of the predictors given the response. Analytically and in simulations we found that our new estimator can preform much better than sliced inverse regression, sliced average variance estimation and directional regression, and that it seems quite robust to deviations from normality.

Key Words: Central subspace, Directional regression, Grassmann manifolds, Sliced inverse regression, Sliced average variance estimation.

*School of Statistics, University of Minnesota, Minneapolis, MN, 55455. email: dennis@stat.umn.edu.
Research for this article was supported in part by grant DMS-0704098 from the U.S. National Science Foundation. Part of this work was completed while both authors were in residence at the Isaac Newton Institute for Mathematical Sciences, Cambridge, U.K.

†Facultad de Ingeniería Química, Universidad Nacional del Litoral and Instituto Matemática Aplicada del Litoral, CONICET, Güemes 3450, (3000) Santa Fe, Argentina. email: liliana.forzani@gmail.com.
The authors are grateful to Bing Li, Penn State University, for providing his directional regression code, to Marcela Morvidone from the Lutheries team, Acoustique et Musique of the Institut Jean Le Rond D'Alembert-Universite Pierre et Marie Curie, Paris, for providing the data for the birds-cars-planes illustration, and to the Editor for his proactive efforts.

1 Introduction

Since the introduction of sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (SAVE; Cook and Weisberg, 1991) there has been considerable interest in dimension reduction methods for the regression of a real response Y on a random vector $\mathbf{X} \in \mathbb{R}^p$ of predictors. A common goal of SIR, SAVE and many other dimension reduction methods is to estimate the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ (Cook, 1994, 1998), which is defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ with the property that Y is conditionally independent of $\mathbf{X}$ given the projection of $\mathbf{X}$ onto $\mathcal{S}$. Informally, these methods estimate the fewest linear combinations of the predictor that contain all the regression information on the response. SIR uses a sample version of the first conditional moment $\mathrm{E}(\mathbf{X}|Y)$ to construct an estimator of $\mathcal{S}_{Y|\mathbf{X}}$, while SAVE uses sample first and second $\mathrm{E}(\mathbf{X}\mathbf{X}^T|Y)$ conditional moments. Other dimension reduction methods are also based on the first two conditional moments and as a class we refer to them as F2M methods.

Although SIR and SAVE have found wide-spread use in application, they nevertheless both have known limitations. In particular, the subspace $\mathcal{S}_{\mathrm{SIR}}$ estimated by SIR is typically a proper subset of $\mathcal{S}_{Y|\mathbf{X}}$ when the response surface is symmetric about the origin. SAVE was developed in response to this limitation and it provides exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$ under mild conditions (Li and Wang, 2007; Shao, Cook and Weisberg, 2007), but its ability to detect linear trends is generally inferior to SIR's. For these reasons, SIR and SAVE have been used as complementary methods, with satisfactory results often obtained by informally combining their estimated directions (see, for example, Cook and Yin, 2001; Bura and Pfeiffer, 2003; Li and Li, 2004; Pardoe, Yin and Cook, 2007). Several authors, in an effort to develop methodology that retains the advantages of SIR and SAVE while avoiding their limitations, have proposed alternative F2M methods. These include combinations of SIR and SIRII (Li, 1991) and combinations of SIR and SAVE (Ye and Weiss, 2003; Zhu, Ohtaki and Li, 2005). Cook and Ni (2005) proposed a

method (IRE) of estimating $\mathcal{S}_{\mathrm{SIR}}$ that is asymptotically efficient among methods based on first conditional moments. Although IRE can be much more efficient that SIR in estimation, it nevertheless shares SIR's scope limitations.

Recently, Li and Wang (2007) proposed a novel F2M method called directional regression (DR). They showed that DR, like SAVE, provides exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$ under mild conditions, and they argued that it is more accurate than or competitive with all of the previous F2M dimension reduction proposals. They also concluded that the class of F2M methods can be expected to yield results of merit in practice, except perhaps when the regression surface undulates, necessitating the use of higher-order conditional moments for exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$.

In this article we take a substantial step forward in the development of F2M dimension reduction methods. Our new method provides exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$ under the same mild conditions as DR and SAVE. However, unlike the previous methods we employ a likelihood-based objective function to acquire the reduced dimensions. Consequently, when the likelihood is accurate our new method – called LAD (likelihood acquired directions) – inherits properties and methods from general likelihood theory. The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ can be estimated using likelihood ratio testing or an information criterion like AIC or BIC, and conditional independence hypotheses involving the predictors can be tested straightforwardly. While likelihood-based estimation can be sensitive to deviations from the underlying assumptions, we demonstrate that LAD has good robustness properties and can be much more accurate than DR, which is reportedly the "best" of the known F2M methods. We show in particular that LAD provides an asymptotically optimal F2M method in a sense described herein.

The advantages of the full likelihood approach developed herein could be anticipated from the work of Zhu and Hastie (2003) and Pardoe et al. (2007). Zhu and Hastie (2003) used a marginal pseudo-likelihood approach to sequentially identify optimal discriminat-

ing directions for non-normal discriminant analysis. Pardoe et al. (2007) showed that for normal data, and in a population sense, the subspace identified by the Zhu-Hastie sequential likelihood method and the subspace identified by SAVE are one and the same. Thus it was to be expected that the full maximum likelihood estimator of the the central subspace under normality would prove to have advantages over SIR, SAVE, DR and the Zhu-Hastie method under the same assumptions.

The rest of the article is organized as follows. Section 2 is devoted to population results. We develop LAD estimation in Section 3. In Section 4 we compare DR, LAD, SAVE and SIR, and discuss the robustness of LAD and its relationship with a method for discriminant analysis proposed by Zhu and Hastie (2003). Inference methods for d and for contributing variables are considered in Sections 5 and 6. Section 7 contains an illustration of how the proposed methodology might be used in practice. Proofs and other supporting material are given in the appendices.

For positive integers p and q, $\mathbb{R}^{p\times q}$ stands for the class of real $p \times q$ matrices. For $\mathbf{A} \in \mathbb{R}^{p\times p}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, $\mathbf{A}\mathcal{S} \equiv \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$. A *semi-orthogonal matrix* $\mathbf{A} \in \mathbb{R}^{p\times q}$, $q < p$, has orthogonal columns, $\mathbf{A}^T\mathbf{A} = I_q$. A *basis matrix* for a subspace $\mathcal{S}$ is a matrix whose columns form a basis for $\mathcal{S}$. For $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathcal{S}_{\mathbf{B}} \equiv \text{span}(\mathbf{B})$ denotes the subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{B}$. If $\mathbf{B} \in \mathbb{R}^{p\times q}$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is symmetric and positive definite, then the projection onto $\mathcal{S}_{\mathbf{B}}$ relative to $\mathbf{\Sigma}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})} \equiv \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{\Sigma}$. $\mathbf{P}_{\mathcal{S}}$ indicates the projection onto the subspace $\mathcal{S}$ in the usual inner product, and $\mathbf{Q}_{\mathcal{S}} = \mathbf{I} - \mathbf{P}_{\mathcal{S}}$. The orthogonal complement $\mathcal{S}^{\perp}$ of a subspace $\mathcal{S}$ is constructed with respect to the usual inner product, unless indicated otherwise. A tilde $\tilde{}$ over a parameter indicates its sample version and a hat $\hat{}$ indicates its maximum likelihood estimator (MLE).

2 Population Results

A q dimensional subspace $\mathcal{S}$ of $\mathbb{R}^p$ is a *dimension reduction subspace* if $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Equivalently, if $\boldsymbol{\alpha}$ is a basis matrix for a subspace $\mathcal{S}$ of $\mathbb{R}^p$ and $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}$ then again $\mathcal{S}$ is a dimension reduction subspace. Under mild conditions the intersection of all dimension reduction subspaces is itself a dimension reduction subspace and then is called the central subspace and denoted by $\mathcal{S}_{Y|\mathbf{X}}$. While the central subspace is a well-defined parameter in almost all regressions, methods for estimating it depend on additional structure.

Let S_Y denote the support of Y, which may be continuous, discrete or categorical in this section. For notational convenience, we frequently use $\mathbf{X}_y$ to denote a random vector distributed as $\mathbf{X}|(Y = y)$, $y \in S_Y$. The full notation $\mathbf{X}|(Y = y)$ will be used when it seems useful for clarity. Further, let $\boldsymbol{\mu}_y = \mathrm{E}(\mathbf{X}_y)$, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, $\boldsymbol{\Delta}_y = \mathrm{var}(\mathbf{X}_y) > 0$, $\boldsymbol{\Delta} = \mathrm{E}(\boldsymbol{\Delta}_Y)$ and $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$. It is common practice in the literature on F2M methods to base analysis on the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. This involves no loss of generality at the population level since central subspaces are equivariant under full rank linear transformations of the predictors, $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$. It also facilitates the development of moment-based methods since $\boldsymbol{\Sigma}$ can be replaced with its sample version for use in practice. However, the $\mathbf{Z}$ scale is not convenient for maximum likelihood estimation since it "hides" $\boldsymbol{\Sigma}$ in the standardized predictor, and the MLE of $\boldsymbol{\Sigma}$ is not necessarily its sample version. As a consequence, we stay in the original scale of $\mathbf{X}$ throughout this article, except when making connections with previous methodology.

The following theorem gives necessary and sufficient conditions for a dimension reduction subspace when $\mathbf{X}_y$ is normally distributed.

Theorem 1 *Assume that* $\mathbf{X}_y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, $y \in S_Y$. *Let* $\mathcal{M} = \mathrm{span}\{\boldsymbol{\mu}_y - \boldsymbol{\mu}|y \in S_Y\}$. *Then* $\mathcal{S} \subseteq \mathbb{R}^p$ *is a dimension reduction subspace if and only if (a)* $\boldsymbol{\Delta}^{-1}\mathcal{M} \subseteq \mathcal{S}$ *and (b)* $\mathbf{Q}_{\mathcal{S}}\boldsymbol{\Delta}_Y^{-1}$ *is a non-random matrix.*

The next proposition gives conditions that are equivalent to condition (b) from The-

orem 1. It is stated in terms of a basis matrix $\boldsymbol{\alpha}$, but the results do not depend on the particular basis selected.

Proposition 1 *Let $\boldsymbol{\alpha}$ be a basis matrix for $\mathcal{S} \subseteq \mathbb{R}^p$ and let $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \in \mathbb{R}^{p\times p}$ be a full rank matrix with $\boldsymbol{\alpha}^T\boldsymbol{\alpha}_0 = 0$. Then condition (b) of Theorem 1 and the following five statements are equivalent. For all $y \in S_Y$,*

(i) $\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1} = \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}$,

(ii) $\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}$ *and* $\boldsymbol{\Delta}_y(I_p - \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)})$ *are constant matrices,*

(iii) $\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)} = \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}$ *and* $\boldsymbol{\Delta}_y(\mathbf{I}_p - \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}) = \boldsymbol{\Delta}(\mathbf{I}_p - \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})})$,

(iv) $\boldsymbol{\Delta}_y = \boldsymbol{\Delta} + \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}(\boldsymbol{\Delta}_y - \boldsymbol{\Delta})\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}$,

(v) $\boldsymbol{\Delta}_y^{-1} = \boldsymbol{\Delta}^{-1} + \boldsymbol{\alpha}\{(\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})^{-1} - (\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\}\boldsymbol{\alpha}^T$.

This proposition does not require normal distributions. With or without normality, condition (b) of Theorem 1 constrains the covariance matrices $\boldsymbol{\Delta}_y$ so that $\mathbf{Q}_{\mathcal{S}}\boldsymbol{\Delta}_y^{-1} = \mathbf{C}$, $y \in S_Y$, where $\mathbf{C}$ is a constant matrix. This moment constraint is equivalent to the five statements of Proposition 1 without regard to the distribution of $\mathbf{X}_y$, provided that the required inverses exist. For instance, starting with condition (b) of Theorem 1 we have $\mathbf{Q}_{\mathcal{S}} = \mathbf{C}\boldsymbol{\Delta}_y$ which implies that $\mathbf{Q}_{\mathcal{S}} = \mathbf{C}\boldsymbol{\Delta}$ and thus that $\mathbf{C} = \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Delta}^{-1}$, leading to condition (i) of Proposition 1. Nevertheless, some useful interpretations still arise within the normal family. With normal populations, $\mathrm{var}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y) = \boldsymbol{\Delta}_y\{\mathbf{I}_p - \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}\}$ (Cook, 1998, p. 131). Thus, condition (ii) of Proposition 1 requires that $\mathrm{var}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, Y)$ be non-random.

Condition (iii) says that the centered means $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y) - \boldsymbol{\mu}_y = \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}(\mathbf{X} - \boldsymbol{\mu}_y)$ must all lie in the same subspace $\mathcal{S}_{\boldsymbol{\Delta}\boldsymbol{\alpha}}$. Together, Theorem 1 and Proposition 1 imply that the deviations $\boldsymbol{\Delta}_y - \boldsymbol{\Delta}$, $y \in S_Y$, must have common invariant subspace $\mathcal{S}_{\boldsymbol{\Delta}\boldsymbol{\alpha}}$ and the translated conditional means $\boldsymbol{\mu}_y - \boldsymbol{\mu}$ must fall in that same subspace.

Results to this point are in terms of dimension reduction subspaces, $\mathcal{S}$ in Theorem 1 and span($\boldsymbol{\alpha}$) in Proposition 1. However, MLEs seem most easily derived in terms of orthonormal bases. The next proposition, which will facilitate finding MLEs in the next section, gives a characterization of a dimension reduction subspace in terms of semi-orthogonal basis matrices.

Proposition 2 *Assume that* $\mathbf{X}_y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, $y \in S_Y$. *Let* $\boldsymbol{\alpha}$ *be a semi-orthogonal basis matrix for* $\mathcal{S} \subseteq \mathbb{R}^p$ *and let* $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \in \mathbb{R}^{p\times p}$ *be an orthogonal matrix. Then* $\mathcal{S}$ *is a dimension reduction subspace if and only if the following two conditions are satisfied. For all* $y \in S_Y$,

1. $\boldsymbol{\alpha}^T\mathbf{X}|(Y=y) \sim N(\boldsymbol{\alpha}^T\boldsymbol{\mu} + \boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y, \boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})$, *for some* $\boldsymbol{\nu}_y \in \mathbb{R}^{\dim(\mathcal{S})}$,
2. $\boldsymbol{\alpha}_0^T\mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, Y=y) \sim N(\mathbf{H}\boldsymbol{\alpha}^T\mathbf{X} + (\boldsymbol{\alpha}_0^T - \mathbf{H}\boldsymbol{\alpha}^T)\boldsymbol{\mu}, \mathbf{D})$ *with* $\mathbf{D} = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)^{-1}$ *and* $\mathbf{H} = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}$.

We see from this theorem that if $\mathcal{S}$ is a dimension reduction subspace with basis $\boldsymbol{\alpha}$, then the distribution of $\boldsymbol{\alpha}^T\mathbf{X}|(Y = y)$ can depend on y, while the distribution of $\boldsymbol{\alpha}_0^T\mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, Y = y)$ cannot. Conversely, if these two distributional conditions hold, then $\mathcal{S} = \text{span}(\boldsymbol{\alpha})$ is a dimension reduction subspace.

The central subspace exists when $\mathbf{X}_y$ is normally distributed (Cook, 1998, Prop. 6.4). Consequently it can be characterized as the intersection of all subspaces $\mathcal{S}$ satisfying Theorem 1. Let $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. We use $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ to denote a semi-orthogonal basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$. A subspace $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathbb{R}^p$ with dimension $d \leq p$ corresponds to a hyperplane through the origin, which can be generated by a $p \times d$ basis matrix. The set of such planes is called a Grassmann manifold $\mathcal{G}_{(d,p)}$ in $\mathbb{R}^p$. The dimension of $\mathcal{G}_{(d,p)}$ is $pd - d^2 = d(p-d)$, since a plane is invariant under nonsingular right-side linear transformations of its basis matrix (Chikuse, 2003).

In the next section we use the model developed here with $\boldsymbol{\alpha} = \boldsymbol{\eta}$ to derive the MLEs. We refer to this as the LAD model.

3 Estimation of $\mathcal{S}_{Y|\mathbf{X}}$ when d is Specified

3.1 LAD maximum likelihood estimators

The population foundations of SIR, SAVE, DR and other F2M methods do not place constraints on S_Y. However, the methods themselves do require a discrete or categorical response. To maintain consistency with previous F2M methods we assume in this section that the response takes values in the support $S_Y = \{1, 2, \ldots, h\}$. When the response is continuous it is typical to follow Li (1991) and replace it with a discrete version constructed by partitioning its range into h slices. Slicing is discussed in Section 3.3.

We assume that the data consist of n_y independent observations of $\mathbf{X}_y$, $y \in S_Y$. The following proposition summarizes maximum likelihood estimation when d is specified. The choice of d is considered in Section 5. In preparation, let $\tilde{\mathbf{\Sigma}}$ denote the sample covariance matrix of $\mathbf{X}$, let $\tilde{\mathbf{\Delta}}_y$ denote the sample covariance matrix for the data with $Y = y$, and let $\tilde{\mathbf{\Delta}} = \sum_{y=1}^{h} f_y \tilde{\mathbf{\Delta}}_y$, where f_y is the fraction of cases observed with $Y = y$.

Theorem 2 *Under the LAD model the MLE of $\mathcal{S}_{Y|\mathbf{X}}$ maximizes over $\mathcal{S} \in \mathcal{G}_{(d,p)}$ the log likelihood function*

$$L_d(\mathcal{S}) = -\frac{np}{2}(1 + \log(2\pi)) + \frac{n}{2}\log|\mathbf{P}_{\mathcal{S}}\tilde{\mathbf{\Sigma}}\mathbf{P}_{\mathcal{S}}|_0 - \frac{n}{2}\log|\tilde{\mathbf{\Sigma}}| - \frac{1}{2}\sum_{y=1}^{h} n_y \log|\mathbf{P}_{\mathcal{S}}\tilde{\mathbf{\Delta}}_y\mathbf{P}_{\mathcal{S}}|_0 \quad (1)$$

where $|\mathbf{A}|_0$ indicates the product of the non-zero eigenvalues of a positive semi-definite symmetric matrix $\mathbf{A}$. The MLE of $\mathbf{\Delta}^{-1}$ is $\widehat{\mathbf{\Delta}}^{-1} = \tilde{\mathbf{\Sigma}}^{-1} + \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}^T\tilde{\mathbf{\Delta}}\widehat{\boldsymbol{\eta}})^{-1}\widehat{\boldsymbol{\eta}}^T - \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}^T\tilde{\mathbf{\Sigma}}\widehat{\boldsymbol{\eta}})^{-1}\widehat{\boldsymbol{\eta}}^T$, where $\widehat{\boldsymbol{\eta}}$ is any semi-orthogonal basis matrix for the MLE of $\mathcal{S}_{Y|\mathbf{X}}$. The MLE $\widehat{\mathbf{\Delta}}_y$ of $\mathbf{\Delta}_y$ is constructed by substituting $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathbf{\Delta}}$ and $\widehat{\boldsymbol{\eta}}^T\tilde{\mathbf{\Delta}}_y\widehat{\boldsymbol{\eta}}$ for the corresponding quantities on the right of the equation in part (iv) of Proposition 1.

Using the results of this theorem it can be shown that the MLE of $\mathbf{\Sigma}$ is $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Delta}} + \mathbf{P}^T_{\widehat{\boldsymbol{\eta}}(\widehat{\mathbf{\Delta}})}\widehat{\mathbf{M}}\mathbf{P}_{\widehat{\boldsymbol{\eta}}(\widehat{\mathbf{\Delta}})}$, where $\widehat{\mathbf{M}}$ is the sample version of $\mathrm{var}(\boldsymbol{\mu}_Y)$.

If $\mathcal{S}_{Y|\mathbf{X}} = \mathbb{R}^p$ ($d = p$) then the log likelihood (1) reduces to the usual log likelihood for fitting separate means and covariance matrices for the h populations. We refer to this as the *full model*. If $\mathcal{S}_{Y|\mathbf{X}}$ is equal to the origin ($d = 0$) then (1) becomes the log likelihood for fitting a common mean and common covariance matrix to all populations. This corresponds to deleting the two terms of (1) that depend on $\mathcal{S}$. Following Shapiro (1986, Prop. 3.2) we found the analytic dimension of the parameter space for the LAD model by computing the rank of the Jacobian of the parameters. For $h > 1$ this rank is $D = p+(h-1)d+p(p+1)/2+d(p-d)+(h-1)d(d+1)/2$. In reference to the parameters of the model representation in Proposition 2, this count can be explained as follows. The first addend of D corresponds the unconstrained overall mean $\boldsymbol{\mu} \in \mathbb{R}^p$, and the second gives the parameter count for the $\boldsymbol{\nu}_y \in \mathbb{R}^d$, which are constrained by $\sum_y f_y \boldsymbol{\nu}_y = 0$ so that they are identified. The third addend corresponds to the positive definite symmetric matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$, and the fourth to the dimension of $\mathcal{G}_{(d,p)}$. Given these parameters, condition (iv) of Proposition 1 says that $\text{span}(\boldsymbol{\Delta}_y - \boldsymbol{\Delta})$ is in the d-dimensional subspace $\mathcal{S}_{\boldsymbol{\Delta}\boldsymbol{\eta}}$ and thus $\boldsymbol{\Delta}_y - \boldsymbol{\Delta}$ can be represented as $\boldsymbol{\delta}\mathbf{M}_y\boldsymbol{\delta}^T$, where $\boldsymbol{\delta} \in \mathbb{R}^{p\times d}$ is a basis matrix for $\mathcal{S}_{\boldsymbol{\Delta}\boldsymbol{\eta}}$, $\mathbf{M}_y \in \mathbb{R}^{d\times d}$ is symmetric and $\sum_y f_y \mathbf{M}_y = 0$. The parameter count for the $\mathbf{M}_y$'s is the final addend in D.

Properties of the MLEs from Theorem 2 depend on the nature of the response and characteristics of the model itself. If Y is categorical and d is specified, as assumed throughout this section, then the MLE of any identified function of the parameters in Theorem 2 is asymptotically unbiased and has minimum asymptotic variance out of the broad class of F2M estimators constructed by minimizing a discrepancy function. We refer to estimators with this property as *asymptotically efficient F2M estimators.* This form of asymptotic efficiency relies on Shapiro's (1986) theory of over-parameterized structural models. The connections between LAD and Shapiro's theory are outlined in Appendix A.5.

3.2 Numerical optimization

We were unable to find a closed-form solution to $\arg\max L_d(\mathcal{S})$, and so it was necessary to use numerical optimization. Using Newton-Raphson iteration on $\mathcal{G}_{(d,p)}$, we adapted Lippert's *sg_min 2.4.1* computer code (www-math.mit.edu/∼lippert/sgmin.html) for Grassmann optimization with analytic first derivatives and numerical second derivatives. In our experience $L_d(\mathcal{S})$ may have multiple local maxima, which seems common for log likelihoods defined on Grassmann manifolds. A standard way to deal with multiple maxima is to use an estimator that is one Newton-Raphson iteration step away from a $\sqrt{n}$-consistent estimator (See, for example, Small, Wang and Yang, 2000). Since SAVE and DR are both $\sqrt{n}$-consistent (Li and Wang, 2007), we started with the one that gave the largest likelihood and then iterated until convergence. We argue later that this LAD estimator of $\mathcal{S}_{Y|\mathbf{X}}$ dominates DR and SAVE. Nevertheless, DR and SAVE are ingredients in our method, since addressing the problem of multiple local maxima would be more difficult without a $\sqrt{n}$-consistent estimator to start iteration.

3.3 Slicing

To facilitate a discussion of slicing, we use W to denote a continuous response, assuming that $\mathbf{X}|(W = w)$ is normal and satisfies the LAD model with central subspace $\mathcal{S}_{W|\mathbf{X}}$. We continue to use Y with support $S_Y = \{1, \ldots, h\}$ to denote the sliced version of W. It is known that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{W|\mathbf{X}}$ with equality when h is sufficiently large. For instance, if $\mathbf{\Delta}_w$ is constant, then $h \geq d + 1$ is necessary to estimate $\mathcal{S}_{W|\mathbf{X}}$ fully. We assume that $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{W|\mathbf{X}}$ throughout this section, so slicing results in no loss of scope. Two additional issues arise when loss of scope is not worrisome: (a) Can we still expect good performance from LAD with h fixed? (b) What are the consequences of varying h?

It can be shown that for any fixed h, the mean $\boldsymbol{\mu}_y$ and covariance $\mathbf{\Delta}_y$ corresponding to the sliced response still satisfy conditions (a) and (b) of Theorem 1 with $\mathcal{S} = \mathcal{S}_{Y|\mathbf{X}}$, but

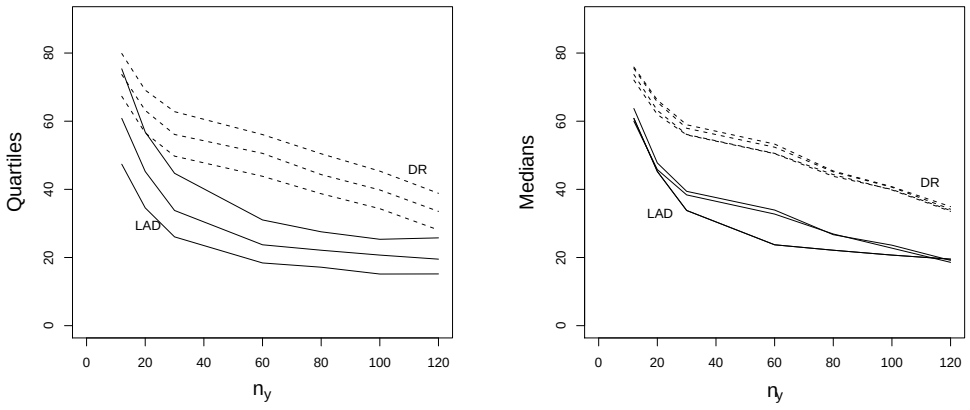
the distribution of $\mathbf{X}_y$ is not generally normal. The LAD estimator of $\mathcal{S}_{Y|\mathbf{X}}$ is still $\sqrt{n}$-consistent (Shapiro, 1986; Appendix A.5), but non-normality mitigates the asymptotic efficiency that holds when $\mathbf{X}_y$ is normal since the third and fourth moments of $\mathbf{X}_y$ may no longer behave as specified under the model. However, we expect that LAD will still perform quite well relative to the class of F2M estimators when $\boldsymbol{\mu}_w$ and $\boldsymbol{\Delta}_w$ vary little within each slice y, because then $\mathbf{X}_y$ should be nearly symmetric and the fourth moments of $\mathbf{Z}_y = \boldsymbol{\Delta}_y^{-1/2}(\mathbf{X}_y - \boldsymbol{\mu}_y)$ should not be far from those of a standard normal random vector.

Efficiency can depend also on the number of slices, h. Although much has been written on choosing h since Li's (1991) pioneering work on SIR, no widely accepted rules have emerged. The general consensus seems to be in accord with Li's original conclusions: h doesn't matter much, provided that it is large enough to allow estimation of d and that there are sufficient observations per slice to estimate the intra-slice parameters, $\boldsymbol{\mu}_y$ and $\boldsymbol{\Delta}_y$ in our LAD models. Subject to this informal condition, we tend to use a small number of slices, say $5 \leq h \leq 15$. Comparing the estimates of $\mathcal{S}_{Y|\mathbf{X}}$ for a few values within this range can be a useful diagnostic on the choice of h. Only rarely do we find that the choice matters materially.

4 Comparison of F2M Methods with d Specified

4.1 Assuming normality

For a first illustration we simulated observations from a simple LAD model using $\boldsymbol{\Delta}_y = \mathbf{I}_p + \sigma_y^2 \boldsymbol{\eta}\boldsymbol{\eta}^T$ with $p = 8$, $\boldsymbol{\eta} = (1, 0 \ldots, 0)^T$, $h = 3$, $\sigma_1 = 1$, $\sigma_2 = 4$ and $\sigma_3 = 8$. The use of the identity matrix $\mathbf{I}_p$ in the construction of $\boldsymbol{\Delta}_y$ was for convenience only since the results are invariant under full rank transformations. The predictors were generated according to $\mathbf{X}_y = \mu_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} + \sigma_y \boldsymbol{\eta}\epsilon$, where $(\boldsymbol{\varepsilon}^T, \epsilon) \sim N(0, \mathbf{I}_{p+1})$, with $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^1$, $\mu_1 = 6$, $\mu_2 = 4$ and $\mu_3 = 2$. Figure 1a shows the quartiles from 400 replications of the



a. Normal error b. Other errors

Figure 1: Quartiles (a) and medians (b) of the angle between $\mathcal{S}_{Y|\mathbf{X}}$ and its estimate versus sample size.

angle between an estimated basis and $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\eta})$ for several sample sizes and two methods, LAD (solid lines) and DR (dashed lines). LAD dominates DR at all sample sizes. Figure 1b is discussed in the next section.

4.2 Assuming linearity and constant covariance conditions

SIR, SAVE and DR do not require conditional normality, but instead use two weaker conditions on the marginal distribution of the predictors: (a) $\text{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a linear function of $\mathbf{X}$ (*linearity condition*) and (b) $\text{var}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a nonrandom matrix (*constant covariance condition*). We forgo discussion of these conditions since they are well known and widely regarded as mild, and were discussed in detail by Li and Wang (2007). They expressed these conditions in the standardized scale of $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, but these are equivalent to the $\mathbf{X}$ scale conditions used here.

The linearity and constant covariance conditions guarantee that SIR, SAVE and DR provide consistent estimators of a subspace of $\mathcal{S}_{Y|\mathbf{X}}$. In particular, they imply that $\text{span}(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y) \subseteq \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$, which is the population basis for SAVE represented in the $\mathbf{X}$

scale. Thus we can define the population SAVE subspace in the $\mathbf{X}$ scale as $\mathcal{S}_{\mathrm{SAVE}} = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Sigma} - \boldsymbol{\Delta}_h)$. We next argue that we can expect good results from LAD without assuming normality, but requiring the weaker conditions used for SAVE and DR. This involves considering the robustness to deviations from normality of the estimator defined by (1).

Holding f_y fixed as $n \to \infty$, $L_d(\mathcal{S})/n$ converges to the population function

$$K_d(\mathcal{S}) = -\frac{p}{2}(1 + \log(2\pi)) + \frac{1}{2}\log|\mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}}|_0 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{y=1}^{h} f_y \log|\mathbf{P}_{\mathcal{S}}\boldsymbol{\Delta}_y\mathbf{P}_{\mathcal{S}}|_0.$$

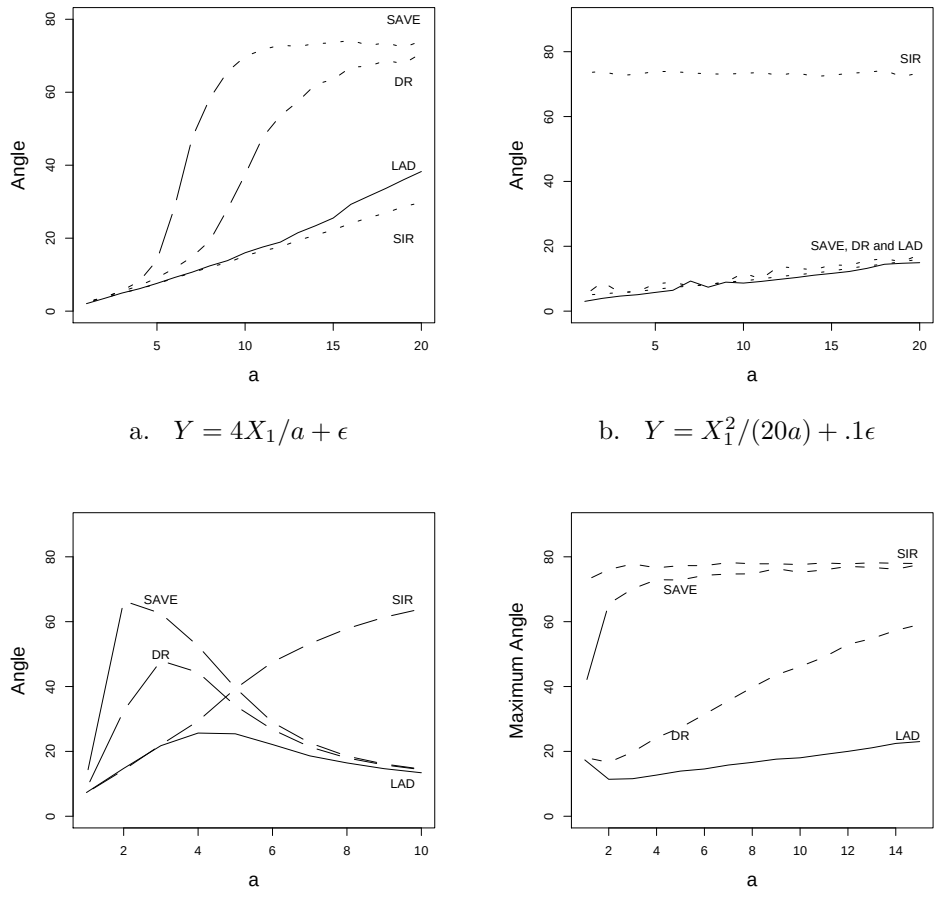
The population LAD subspace is then $\mathcal{S}_{\mathrm{LAD}} = \arg\max_{\mathcal{S}\in\mathcal{G}_{(d,p)}} K_d(\mathcal{S})$. The next proposition requires no conditions other than the convergence of $L_d(\mathcal{S})/n$ to $K_d(\mathcal{S})$.

Proposition 3 $\mathcal{S}_{\mathrm{LAD}} = \mathcal{S}_{\mathrm{SAVE}}$.

This proposition indicates that LAD and SAVE estimate the same subspace, even when the distribution of $\mathbf{X}_y$ is non-normal and the linearity and constant covariance conditions fail. Proposition 3 may be of little practical importance if there is no useful connection with $\mathcal{S}_{Y|\mathbf{X}}$, the subspace we would like to estimate. Let $\mathcal{S}_{\mathrm{DR}}$ denote the subspace estimated by directional regression. We know from Li and Wang (2007) that $\mathcal{S}_{\mathrm{SAVE}} = \mathcal{S}_{\mathrm{DR}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ under the linearity and constant covariance conditions and that these three subspaces are equal under mild additional conditions. It follows from Proposition 3 that, under these same conditions, $\mathcal{S}_{\mathrm{LAD}} = \mathcal{S}_{\mathrm{SAVE}} = \mathcal{S}_{\mathrm{DR}} = \mathcal{S}_{Y|\mathbf{X}}$. The moment relations of Theorem 1 still hold in this setting, but $\mathbf{X}_y$ may no longer be normal. As in Section 3.3, we still have a $\sqrt{n}$-consistent estimator, but non-normality can mitigate the asymptotic efficiency that holds when $\mathbf{X}_y$ is normal. If $\mathbf{X}_y$ is substantially skewed or the fourth moments of $\mathbf{Z}_y$ deviate substantially from those of a standard normal random vector then better estimators may exist. Pursuit of improved methods non-parametrically will likely require large sample sizes for the estimation of third and fourth moments.

Li and Wang (2007) showed that DR can achieve considerable gains in efficiency over SAVE and other F2M methods. We next use simulation results to argue that LAD can perform much better than DR. Recall that Figure 1a shows LAD can be much more efficient that DR when $\mathbf{X}_y$ is normally distributed. Using the same simulation setup, Figure 1b shows the median over 400 replication of the angle between $\widehat{\boldsymbol{\eta}}$ and $\mathcal{S}_{Y|\mathbf{X}}$ for standard normal, t_5, χ^2_5 and uniform $(0,1)$ error $(\boldsymbol{\varepsilon}^T, \epsilon)$ distributions. It follows from Cook and Yin (2001, Prop. 3) that the linearity and constant covariance conditions hold for this simulation scenario and consequently $\mathcal{S}_{\text{SAVE}} = \mathcal{S}_{\text{DR}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The final condition for equality (Li and Wang, 2007, condition b of Theorem 3) can be verified straightforwardly and thus we conclude that $\mathcal{S}_{\text{SAVE}} = \mathcal{S}_{\text{DR}} = \mathcal{S}_{Y|\mathbf{X}}$. The lower most LAD curve of Figure 1b corresponds to the normal errors. The other results for both LAD and DR are so close that the individual curves were not marked. These results sustain our previous conclusion that normality is not essential for the likelihood-based objective function (1) to give good results in estimation. DR did not exhibit any advantages.

It is well-known that SIR is generally better than SAVE at finding linear trends in the mean function $\mathrm{E}(Y|\mathbf{X})$, while SAVE does better at finding quadratic structure. Simple forward quadratic models have often been used as test cases to illustrate this phenomenon and compare methods (see, for example, Cook and Weisberg, 1991). Here we present results from the following four simulation models to provide further contrast between SIR, SAVE, DR and LAD. For $n = 500$ we first generated $\mathbf{X} \sim N(0, \mathbf{I}_p)$ and $\epsilon \sim N(0,1)$ and then generated Y according to the following four models: (1) $Y = 4X_1/a + \epsilon$, (2) $Y = X_1^2/(20a) + .1\epsilon$, (3) $Y = X_1/(10a) + aX_1^2/100 + .6\epsilon$, and (4) $Y = .4a(\boldsymbol{\beta}_1^T\mathbf{X})^2 + 3\sin(\boldsymbol{\beta}_2^T\mathbf{X}/4) + .2\epsilon$. For simulation models 1, 2 and 3, $p = 8$ and $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{(1,0,\ldots,0)^T\}$. For model 4, $p = 20$, and $\mathcal{S}_{Y|\mathbf{X}}$ is spanned by $\boldsymbol{\beta}_1 = (1,1,1,0,\ldots,0)^T$ and $\boldsymbol{\beta}_2 = (1,0,\ldots,0,1,3)^T$. With $a = 1$ model 4 is identical to simulation model I used by Li and Wang (2007). The conditional distribution of $\mathbf{X}_y$ is normal for model 1, but



a. $Y = 4X_1/a + \epsilon$ b. $Y = X_1^2/(20a) + .1\epsilon$

c. $Y = X_1/(10a) + aX_1^2/100 + .6\epsilon$ d. $Y = .4a(\boldsymbol{\beta}^T\mathbf{X})^2 + 3\sin(\boldsymbol{\beta}_2^T\mathbf{X}/4) + .2\epsilon$.

Figure 2: Comparison of SIR, SAVE, DR and LAD: Plots of the average angle or average maximum angle between $\mathcal{S}_{Y|\mathbf{X}}$ and its estimates for four regression models at selected values of a. Solid lines give the LAD results.

non-normal for the other three models. Figures 2a, b and c show plots of the average angle over 400 replications between $\mathcal{S}_{Y|\mathbf{X}}$ and its estimates for $h = 5$ and $a = 1, \ldots, 10$. Since $d = 2$ for model 4 we summarized each simulation run with the maximum angle between $\mathcal{S}_{Y|\mathbf{X}}$ and the subspace generated by the estimated directions with $h = 10$. The vertical axis of Figure 2d is the average maximum angle over the 400 replicates.

In Figure 2a (model 1) the strength of the linear trend decreases as a increases. Here the methods perform essentially the same for strong linear trends (small a). SAVE

and DR deteriorate quickly as a increases, with DR performing better. LAD and SIR perform similarly, with SIR doing somewhat better for large a. Since $\mathbf{\Delta}_y$ is constant, LAD overfits by estimating individual covariance matrices. SIR uses only first conditional moments and thus is not susceptible to this type of overfitting, which may account for SIR's advantage when $\mathbf{\Delta}_y$ is constant and the linear effect is small (large a).

In model 2 $\text{cov}(\mathbf{X}, Y) = 0$, and the strength of the quadratic term decreases as a increases. This is the kind of setting in which it is known that SIR estimates a proper subset of $\mathcal{S}_{Y|\mathbf{X}}$, in this case the origin. The simulation results for this model are shown in Figure 2b, where SAVE, DR and LAD perform similarly, with LAD doing slightly better at all values of a.

In model 3, which has both linear and quadratic components in X_1, the strength of the linear trend decreases and the strength of the quadratic trend increases as a increases. We see from Figure 2c that SIR, SAVE and DR perform as might be expected from the previous plots, while LAD always does at least as well as the best of these methods and does better for middle values of a.

Model 4 has a linear trend in $\boldsymbol{\beta}_2^T X_2$ and a quadratic in $\boldsymbol{\beta}_1^T X_1$. As suggested by Figure 2d, SIR cannot find the quadratic direction and so its maximum angle is always large. At small values of a the contributions of the linear and quadratic terms to the mean function are similar and DR and LAD perform similarly. As a increases the quadratic term dominates the mean function, making it hard for SAVE and DR to find the linear effect $\boldsymbol{\beta}_2^T X_2$. However, LAD does quite well at all value of a. Finally, we repeated the simulations for models 1, 2 and 3 with $h = 10$ slices and normal and non-normal (t_5,χ_5, $U(0, 1)$) error distributions, finding qualitatively similar behavior.

4.3 Robustness of $\widehat{\mathcal{S}}_{Y|\mathbf{X}}$ to non-normality

The previous simulations indicate that normality is not essential for (1) to provide useful estimates of $\mathcal{S}_{Y|\mathbf{X}}$. In this section we give an explanation for why this might be so.

Recalling that $\boldsymbol{\eta}$ is a semi-orthogonal basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$ and that $(\boldsymbol{\eta}, \boldsymbol{\eta}_0)$ is an orthogonal matrix, the possibly non-normal density J of $(\boldsymbol{\eta}^T\mathbf{X}, \boldsymbol{\eta}_0^T\mathbf{X})|Y$ can be represented as $J(\boldsymbol{\eta}^T\mathbf{X}, \boldsymbol{\eta}_0^T\mathbf{X}|Y) = k(\boldsymbol{\eta}^T\mathbf{X}|Y)g(\boldsymbol{\eta}_0^T\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$, where the density g does not depend on Y because $Y \perp\!\!\!\perp \boldsymbol{\eta}_0^T\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X}$. When $\mathbf{X}|Y$ is normal the densities k and g are implied by Proposition 2. The log likelihood L_d based on this decomposition can be represented broadly as $L_d = L^{(k)} + L^{(g)}$, where $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ and the superscripts k and g indicate the density from which that portion of the log likelihood is derived. Keeping $\boldsymbol{\eta}$ fixed, we assume that $\max L_d = \max L^{(k)} + \max L^{(g)}$. For example, this is true when, for fixed $\boldsymbol{\eta}$, the parameters of k and g are defined on a product space $\boldsymbol{\Theta}_k \times \boldsymbol{\Theta}_g$ so $L^{(k)}$ and $L^{(g)}$ can be maximized independently. This product space structure holds for the normal model and was used implicitly when deriving the MLEs in Appendix A.4. We thus have the partially maximized log likelihood $L_d(\mathcal{S}) = L^{(k)}(\mathcal{S}) + L^{(g)}(\mathcal{S})$, which is to be maximized over $\mathcal{G}_{(d,g)}$. For the normal model $L_d(\mathcal{S})$ was given in (1).

Repeating the above argument under the assumption that $Y \perp\!\!\!\perp \mathbf{X}$ gives the density decomposition $J_0(\boldsymbol{\eta}^T\mathbf{X}, \boldsymbol{\eta}_0^T\mathbf{X}) = k_0(\boldsymbol{\eta}^T\mathbf{X})g(\boldsymbol{\eta}_0^T\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ and partially maximized log likelihood $L_0(\mathcal{S}) = L^{(k_0)}(\mathcal{S}) + L^{(g)}(\mathcal{S})$. Since $Y \perp\!\!\!\perp \mathbf{X}$, $L_0(\mathcal{S})$ is a constant function of $\mathcal{S}$ and thus can be subtracted from $L_d(\mathcal{S})$, giving $L_d(\mathcal{S}) - L_0(\mathcal{S}) = L^{(k)}(\mathcal{S}) - L^{(k_0)}(\mathcal{S})$, which does not depend on g. Consequently, the MLE of $\mathcal{S}_{Y|\mathbf{X}}$ can be represented as $\arg\max L_d(\mathcal{S}) = \arg\max\{L^{(k)}(\mathcal{S}) - L^{(k_0)}(\mathcal{S})\}$. This says that we do not need g to estimate $\mathcal{S}_{Y|\mathbf{X}}$ alone, provided $L^{(k)}$ and $L^{(g)}$ can be maximized independently while holding $\boldsymbol{\eta}$ fixed.

Diaconis and Freedman (1984) show that almost all projections of high dimensional data are approximately normal. Thus when d is small relative to p it may be reasonable

to approximate $k(\boldsymbol{\eta}^T\mathbf{X}|Y)$ and $k_0(\boldsymbol{\eta}^T\mathbf{X})$ with compatible normal densities, leading to estimates of $\mathcal{S}_{Y|\mathbf{X}}$ that are the same as those from (1).

Zhu and Hastie (2003) proposed an exploratory nonparametric method for discriminant analysis based on a certain likelihood ratio $\mathrm{LR}(\boldsymbol{\alpha})$ as a function of a single discriminant direction $\boldsymbol{\alpha} \in \mathbb{R}^p$. Their method, which was based on reasoning by analogy from Fisher's linear discriminant, proceeds sequentially by first finding $\boldsymbol{\alpha}_1 = \arg\max \mathrm{LR}(\boldsymbol{\alpha})$. Subsequent directions $\boldsymbol{\alpha}_j \in \mathbb{R}^p$ are then defined as $\boldsymbol{\alpha}_j = \arg\max \mathrm{LR}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha}^T\boldsymbol{\Phi}\boldsymbol{\alpha}_k = 0$, $k = 1, \ldots, j-1$, where $\boldsymbol{\Phi}$ is a user-specified inner product matrix. Assuming normality of $\mathbf{X}|Y$, Pardoe, et al. (2007, Prop. 3) demonstrated that *in the population* the Zhu-Hastie method and SAVE produce the same subspace. More fundamentally, it follows by definition of LR that $\log\{\mathrm{LR}(\boldsymbol{\alpha})\} = L^{(k)}(\mathcal{S}_{\boldsymbol{\alpha}}) - L^{(k_0)}(\mathcal{S}_{\boldsymbol{\alpha}})$. Consequently, when $\mathbf{X}|Y$ is normal and $d = 1$ maximizing $\mathrm{LR}(\boldsymbol{\alpha})$ is equivalent to maximizing $L_1(\mathcal{S})$ (1). With its close connection to Fisher's linear discriminant and its reliance on simple likelihood ratios, the Zhu-Hastie method is grounded in familiar statistical concepts and thus provides simple intuitive insight into the workings of the full likelihood estimator developed in Section 3. However, although the likelihood and MLE in Theorem 2 are not as intuitive initially, the full-likelihood approach has the advantages of being compatible with familiar information-based stopping criteria, and avoids the sequential optimization and dependence on user-specified inputs of the Zhu-Hastie method.

5 Choice of d

In this section we consider ways in which $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ can be chosen in practice, distinguishing the true value d from the value d_0 used in fitting. The hypothesis $d = d_0$ can be tested by using the likelihood ratio statistic $\Lambda(d_0) = 2\{\hat{L}_p - \hat{L}_{d_0}\}$, where $\hat{L}_p$ denotes the value of the maximized log likelihood for the full model with $d_0 = p$ and $\hat{L}_{d_0}$ is the maximum value of the log likelihood (1). Under the null hypothesis $\Lambda(d_0)$

is distributed asymptotically as a chi-squared random variable with degrees of freedom $(p-d_0)\{(h-1)(p+1)+(h-3)d_0+2(h-1)\}/2$, for $h \geq 2$ and $d_0 < p$ (Shapiro, 1986 and Appendix A.5). The statistic $\Lambda(d_0)$ can be used in a sequential testing scheme to choose d: Using a common test level and starting with $d_0 = 0$, choose the estimate of d as the first hypothesized value that is not rejected. This method for dimension selection is common in dimension reduction literature (see Cook, 1998, p. 205, for background).

A second approach is to use an information criterion like AIC or BIC. BIC is consistent for d while AIC is minimax-rate optimal (Burnham and Anderson, 2002). For $d \in \{0, \ldots, p\}$, the dimension is selected that minimizes the information criterion $IC(d_0) = -2\hat{L}_{d_0} + h(n)g(d_0)$, where $g(d_0)$ is the number of parameters to be estimated as a function of d_0, in our case $p + (h-1)d_0 + d_0(p-d_0) + (h-1)d_0(d_0+1)/2 + p(p+1)/2$, and $h(n)$ is equal to $\log n$ for BIC and 2 for AIC. This version of AIC is a simple adaptation of the commonly occurring form for other models.

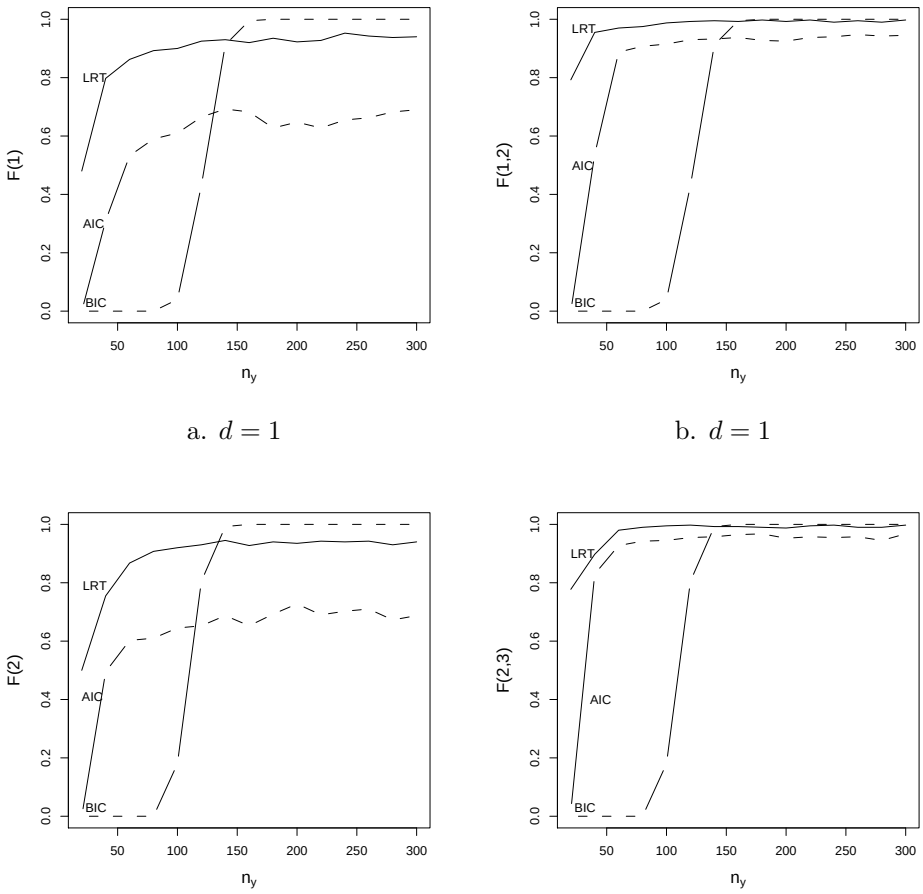
Consider inference on d in the simulation model with $d = 1$ introduced in Section 4.1. Figures 3a and b give the fractions $F(1)$ and $F(1,2)$ of 500 replications in which the indicated procedure selected $d = 1$ and $d = 1$ or 2 versus n_y. BIC gave the best results for large n_y, but the likelihood ratio test (LRT) also performed well and may be a useful choice when the sample size is not large.

Figures 3c and d display results for inference on d in the simulation model $\mathbf{X}_y = \boldsymbol{\eta}\boldsymbol{\mu}_y + \boldsymbol{\varepsilon} + \boldsymbol{\eta}\mathbf{A}_y^{1/2}\boldsymbol{\epsilon}$ with $d = 2$, $p = 8$, $h = 3$, $\boldsymbol{\eta}^T = ((1,0,\ldots,0)^T, (0,1,0,\ldots,0)^T)$, $\boldsymbol{\Delta}_y = \mathbf{I}_p + \boldsymbol{\eta}\mathbf{A}_y\boldsymbol{\eta}^T$, $\boldsymbol{\mu}_1 = (6,2)^T$, $\boldsymbol{\mu}_2 = (4,4)^T$, $\boldsymbol{\mu}_3 = (6,2)^T$, and

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} 8 & 1 \\ 1 & 2 \end{pmatrix}.$$

Again, BIC performs the best for large samples, but LRT has advantages otherwise.

Although deviations from normality seem to have little impact on estimation of $\mathcal{S}_{Y|\mathbf{X}}$



a. $d = 1$ b. $d = 1$

c. $d = 2$ d. $d = 2$

Figure 3: Inference about d: $F(i)$, $F(i, j)$ are the fraction of runs in which the estimated d was one of the arguments. Results with same value of d are from the same simulation model.

when d is known, they can have a pronounced effect on the estimate of d. In such cases the permutation test proposed by Cook and Weisberg (2001) and developed by Cook and Yin (2001) can serve as an effective substitute for the LRT or an information criterion. To confirm that the permutation test applies straightforwardly in the present setting, Table 1 shows the percentage of time $d = 1$ was selected by the LRT and permutation test methods in 200 replications of the simulation model of Figure 1 with $n = 40$ and four error distributions. Results for the LRT under normality and all results for the

permutation test method are within the expected binomial error at the nominal level. As expected the LRT with χ^2_5 and t_5 error distributions exhibits clear deviations from the nominal. It is also possible to derive the asymptotic distribution of the likelihood ratio statistic under non-normal error distributions. However, the permutation test is a straightforward and reliable method for choosing d when normality is at issue, and it avoids the task of assessing if the sample size is large enough for the asymptotic results to be useful.

Table 1: Percentage of time the nominal 5% likelihood ratio test (LRT) and permutation test (PT) methods chose $d = 1$ in 200 replications with $n = 40$ and four error ϵ distributions.

	Error Distribution			
Method	$N(0,1)$	$U(0,1)$	χ^2_5	t_5
LRT	96.5	92.5	47.5	38.5
PT	93.5	96.0	94.5	96.5

6 Testing Variates

With d fixed *a priori* or after estimation, it may be of interest to test an hypothesis that a selected subspace $\mathcal{H}$ of dimension $\ell \leq p - d$ is orthogonal to $\mathcal{S}_{Y|\mathbf{X}}$ in the usual inner product. The restriction on ℓ is to ensure that the dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is still d under the hypothesis. Letting $\mathbf{H}_0 \in \mathbb{R}^{p\times\ell}$ be a semi-orthogonal basis matrix for $\mathcal{H}$, the hypothesis can be restated as $\mathbf{P}_{\mathbf{H}_0}\mathcal{S}_{Y|\mathbf{X}} = 0$ or $\mathbf{P}_{\mathbf{H}_1}\boldsymbol{\eta} = \boldsymbol{\eta}$, where $(\mathbf{H}_0, \mathbf{H}_1)$ is an orthogonal matrix. For instance, to test the hypothesis that a specific subset of ℓ variables is not directly involved in the reduction $\boldsymbol{\eta}^T\mathbf{X}$, set the columns of $\mathbf{H}_0$ to be the corresponding ℓ columns of $\mathbf{I}_p$.

The hypothesis $\mathbf{P}_{\mathbf{H}_0}\mathcal{S}_{Y|\mathbf{X}} = 0$ can be tested by using a standard likelihood test. The test statistic is $\Lambda_d(\mathbf{H}_0) = 2(\hat{L}_d - \hat{L}_{d,\mathbf{H}_0})$, where $\hat{L}_d$ is the maximum value of the log

likelihood (1), and $\hat{L}_{d,\mathbf{H}_0}$ is the maximum value of (1) with $\mathcal{S}_{Y|\mathbf{X}}$ constrained by the hypothesis. Under the hypothesis the statistic $\Lambda_d(\mathbf{H}_0)$ is distributed asymptotically as a chi-squared random variable with $d\ell$ degrees of freedom. The maximized log likelihood $\hat{L}_{d,\mathbf{H}_0}$ can be obtained by maximizing over $\mathcal{S} \in \mathcal{G}_{(d,p-\ell)}$ the constrained log likelihood

$$L_d(\mathcal{S}) = -\frac{np}{2}(1+\log(2\pi)) + \frac{n}{2}\log|\mathbf{P}_{\mathcal{S}}\mathbf{H}_1^T\tilde{\mathbf{\Sigma}}\mathbf{H}_1\mathbf{P}_{\mathcal{S}}|_0 - \sum_{g=1}^{h}\frac{n_y}{2}\log|\mathbf{P}_{\mathcal{S}}\mathbf{H}_1^T\tilde{\mathbf{\Delta}}_y\mathbf{H}_1\mathbf{P}_{\mathcal{S}}|_0, \quad (2)$$

where $\mathbf{H}_1 \in \mathbb{R}^{p\times(p-\ell)}$ is a basis matrix for $\mathcal{H}^{\perp}$. When testing that a specific subset of ℓ variables is not directly involved in the reduction, the role of $\mathbf{H}_1$ in (2) is to select the parts of $\tilde{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Delta}}_y$ that correspond to the other variables.

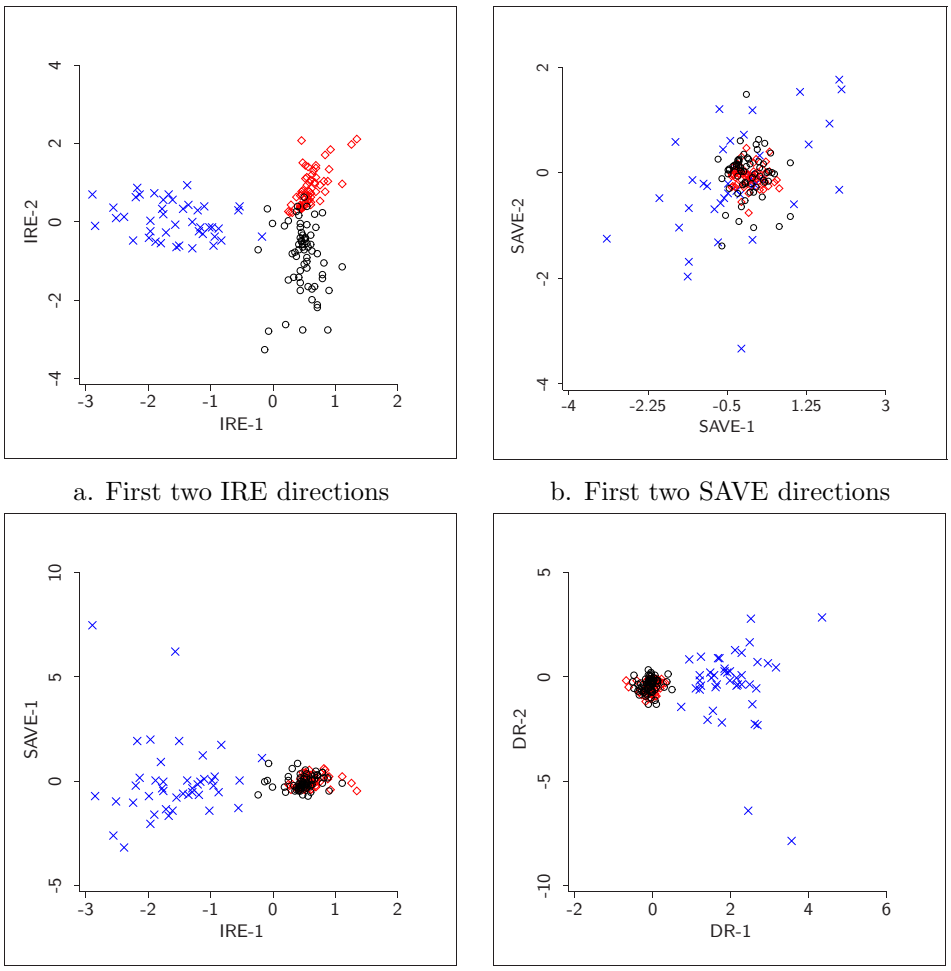
7 Is it a bird, a plane or a car?

This illustration is from a pilot study to assess the possibility of distinguishing birds, planes and cars by the sounds they make, the ultimate goal being the construction of sonic maps that identify both the level and source of sound. A two-hour recording was made in the city of Ermont, France, and then 5 second snippets of sounds were selected. This resulted in 58 recordings identified as birds, 43 as cars and 64 as planes. Each recording was processed and ultimately represented by 13 SDMFCCs (Scale Dependent Mel-Frequency Cepstrum Coefficients). The 13 SDMFCCs were obtained as follows: the signal was decomposed using a Gabor dictionary (a set of Gabor frames with different window sizes) through a matching pursuit algorithm. Each atom of the dictionary depends on time, frequency and scale. The algorithm gave for each signal a linear combination of the atoms of the dictionary. A weighted histogram of the coefficients of the decomposition was then calculated for each signal. The histogram had two dimensions in terms of frequency and scale, and for each frequency-scale pair the amplitude of the coefficients that falls in that bin were added. After that the two-dimensional cosine discrete

transform of the histogram was calculated, resulting in the 13 SDMFCCs.

We focus on reducing the dimension of the 13-dimensional feature vector, which may serve as a preparatory step for developing a classifier. Figure 4a shows a plot of the first and second IRE predictors (Cook and Ni, 2005) marked by sound source, cars (blue $\times$'s), planes (black $\circ$'s) and birds (red $\diamond$'s). Since there are three sound sources, IRE can provide only two directions for location separation. Application of predictor tests associated with IRE gave a strong indication that only four of the 13 predictors are needed to describe the location differences of Figure 4a.

A plot of the first two SAVE predictors is shown in Figure 4b. To allow greater visual resolution, three remote cars were removed from this plot, but not from the analysis or any other plot. Figure 4b shows differences in variation but no location separation is evident. This agrees with the general observation that SAVE tends to overlook location separation in the presence of strong variance differences. Here, as in Figures 4c and 4d, planes and birds are largely overplotted. The plot of the first IRE and SAVE predictors given in Figure 4c shows separation in location and variance for cars from planes and birds. The first two DR predictors in Figure 4d show similar results. Incorporating a third SAVE or DR direction in these plots adds little to the separation between birds and planes. In contrast to the results for IRE, SAVE and DR, the plot of the first two LAD predictors shown in Figure 5 exhibits strong separation in both location and variation. In fact, the first two LAD predictors perfectly separates the sound sources, suggesting that they may be sufficient for discrimination. The first five DR predictors are needed to fit linearly the first LAD predictor with $R^2 \approx .95$, while the first 11 DR predictors are needed to fit the second LAD predictor with $R^2 \approx .95$. Clearly, LAD and DR give quite different representations of the data.



a. First two IRE directions

b. First two SAVE directions

c. First IRE and first SAVE directions

d. First and second DR directions

Figure 4: Plots of IRE, SAVE and DR predictors for the birds-planes-cars example. Birds, red $\diamond$'s; planes, black $\circ$'s; cars, blue $\times$'s.

8 Discussion

Many dimension reduction methods have been proposed since the original work on SIR and SAVE. Mostly these are based on nonparametric or semi-parametric method-of-moment arguments, leading to various spectral estimates of $\mathcal{S}_{Y|\mathbf{X}}$. Minimizing assumptions while still estimating $\mathcal{S}_{Y|\mathbf{X}}$ consistently has been a common theme in their development. Little attention was devoted directly to efficiency. The approach we propose achieves asymptotic F2M efficiency and all results we have indicate that its performance

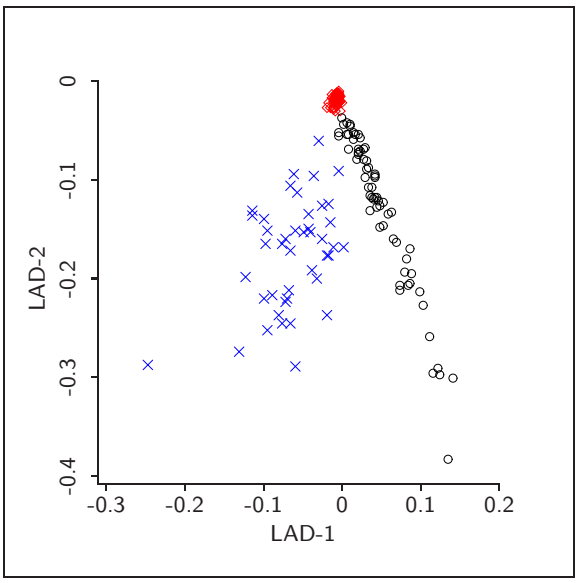


Figure 5: Plot of the first two LAD directions for the birds-planes-cars example. Birds, red $\diamond$'s; planes, black $\circ$'s; cars, blue $\times$'s.

is competitive with or superior to all other F2M methods. We emphasized LAD's performance relative to that of DR since, judging from the report of Li and Wang (2007), DR is a top F2M method.

In addition to producing apparently superior dimension reduction methodology, our work also renewed our appreciation for classical likelihood-based reasoning and we believe that it will find a central place in the development of future methodology.

A Appendix: Proofs and Justifications

In order to prove various results we need an identity from Rao (1973, p. 77). Let $\mathbf{B} \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix, and let $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \in \mathbb{R}^{p\times p}$ be a full rank matrix with $\boldsymbol{\alpha}^T\boldsymbol{\alpha}_0 = 0$. Then

$$\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T + \mathbf{B}^{-1}\boldsymbol{\alpha}_0(\boldsymbol{\alpha}_0^T\mathbf{B}^{-1}\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\alpha}_0^T\mathbf{B}^{-1} = \mathbf{B}^{-1}. \tag{3}$$

As a consequence of (3) we have

$$\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\mathbf{B})} = \mathbf{P}_{\boldsymbol{\alpha}_0(\mathbf{B}^{-1})}. \quad (4)$$

Additionally, if $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$ is orthogonal then

$$|\boldsymbol{\alpha}_0^T \mathbf{B} \boldsymbol{\alpha}_0| = |\mathbf{B}||\boldsymbol{\alpha}^T \mathbf{B}^{-1} \boldsymbol{\alpha}|, \quad (5)$$

$$(\boldsymbol{\alpha}_0^T \mathbf{B}^{-1} \boldsymbol{\alpha}_0)^{-1} = \boldsymbol{\alpha}_0^T \mathbf{B} \boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_0^T \mathbf{B} \boldsymbol{\alpha} (\boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha})^{-1} \boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha}_0, \quad (6)$$

$$-(\boldsymbol{\alpha}_0^T \mathbf{B}^{-1} \boldsymbol{\alpha}_0)^{-1} (\boldsymbol{\alpha}_0^T \mathbf{B}^{-1} \boldsymbol{\alpha}) = (\boldsymbol{\alpha}_0^T \mathbf{B} \boldsymbol{\alpha})(\boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha})^{-1}. \quad (7)$$

A.1 Proof of Proposition 1

We begin by showing that condition b of Theorem 1 implies (i). We then show that each conclusion of Proposition 1 implies the next, ending by showing that (v) implies condition b of Theorem 1.

Condition b of Theorem 1 implies (i): $\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1} = \mathbf{C} \Rightarrow \boldsymbol{\alpha}_0^T = \mathbf{C}\boldsymbol{\Delta}_y \Rightarrow \boldsymbol{\alpha}_0^T = \mathbf{C}\boldsymbol{\Delta} \Rightarrow \mathbf{C} = \boldsymbol{\alpha}_0^T \boldsymbol{\Delta}^{-1}$. Conclusion (i) implies (ii) by from application of (4) with $\mathbf{B} = \boldsymbol{\Delta}_y$:

$$\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)} = \boldsymbol{\alpha}_0 (\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1} \boldsymbol{\alpha}_0)^{-1} \boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1} = \mathbf{C}_1, \quad (8)$$

$$(\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)})\boldsymbol{\Delta}_y = \boldsymbol{\alpha}_0 (\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1} \boldsymbol{\alpha}_0)^{-1} \boldsymbol{\alpha}_0^T = \mathbf{C}_2, \quad (9)$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are constant matrices since $\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1}$ is constant by hypothesis (i).

If (ii) is true then (8) and (9) must hold. This implies that $\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}_y^{-1}$ is constant and thus equal to $\boldsymbol{\alpha}_0^T \boldsymbol{\Delta}^{-1}$. Conclusion (iii) follows by application of (4) with $\mathbf{B} = \boldsymbol{\Delta}$. Condition (iv) follows from (iii) by replacing $\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}$ with $\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}$ in the second condition of (iii) and rearranging terms: $\boldsymbol{\Delta}_y - \boldsymbol{\Delta} = \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}(\boldsymbol{\Delta}_y - \boldsymbol{\Delta})\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}$. Conclusion (v) follows from (iv) by direct multiplication. Finally, multiplying (v) on the left by $\boldsymbol{\alpha}_0$ immediately gives condition b of Theorem 1.

A.2 Proof of Theorem 1

By definition $\mathrm{var}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y) = (\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)})\boldsymbol{\Delta}_y$ and $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y) = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\omega}_y + \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}(\mathbf{X} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\omega}_y)$, where $\boldsymbol{\omega}_y = \boldsymbol{\Gamma}^T(\boldsymbol{\mu}_y - \boldsymbol{\mu})$, $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ and $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix whose columns form a basis for $\mathcal{M}$. Consequently $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y)$ and $\mathrm{var}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}, y)$ are constant if and only if $(\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)})\boldsymbol{\Delta}_y$ and $\mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}$ are constant and $\mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}$. Using Proposition 1 these three conditions are equivalent to $\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}$ being constant and $\mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}\boldsymbol{\Gamma} = \mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}$. Now, $\mathbf{P}^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}\boldsymbol{\Gamma} = \boldsymbol{\Gamma} \Leftrightarrow \mathbf{P}_{\boldsymbol{\alpha}(\boldsymbol{\Delta})}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) = \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma} \Leftrightarrow \mathrm{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) \subseteq \mathrm{span}(\boldsymbol{\alpha})$.

A.3 Proof of Proposition 2

Let $\boldsymbol{\rho}_y = \boldsymbol{\alpha}_0^T\boldsymbol{\mu} + \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y + (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T(\mathbf{X} - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y)$. Since $\mathbf{X}|y$ is normal, $\boldsymbol{\alpha}_0^T\mathbf{X}|(\boldsymbol{\alpha}^T\mathbf{X}, y) \sim N(\boldsymbol{\rho}_y, \boldsymbol{\Theta}_y)$, with $\boldsymbol{\Theta}_y = \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha}_0$. Assume that $\mathcal{S}$ is a dimension reduction subspace. The first conclusion of Proposition 2 follows immediately. Using (6), (7), Theorem 1 and Proposition 1(i) we have $\boldsymbol{\Theta}_y = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}\boldsymbol{\alpha}_0)^{-1}$ and $\boldsymbol{\rho}_y = \mathbf{H}\boldsymbol{\alpha}^T\mathbf{X} + (\boldsymbol{\alpha}_0^T - \mathbf{H}\boldsymbol{\alpha}^T)\boldsymbol{\mu}$, which are equivalent to the second conclusion of Proposition 2.

Assume that the distributions of Proposition 2 hold. Using the forms for $\boldsymbol{\Theta}_y$ and $\boldsymbol{\rho}_y$ we have $\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0$ and $(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})^{-1} = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}$. Using these plus (7) we get

$$\begin{aligned}\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}\boldsymbol{\alpha} &= -(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}\boldsymbol{\alpha}_0)(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}_y\boldsymbol{\alpha})^{-1} \\ &= -(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1} = \boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}.\end{aligned}$$

It follows that $\mathbf{Q}_{\mathcal{S}}\boldsymbol{\Delta}_Y^{-1}$ is constant. Using Proposition 2(1) implies $\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X}) = \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\nu}_y$ and therefore $\mathcal{S}$ is a dimension reduction subspace.

A.4 Proof of Theorem 2

Recalling that $\boldsymbol{\eta}$ is a semi-orthogonal basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$, the log likelihood based on the representation of the distribution of $(\boldsymbol{\eta}^T\mathbf{X}, \boldsymbol{\eta}_0^T\mathbf{X}|Y)$ given in Proposition (2) can be written as

$$\begin{aligned} L_d = & -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta}| \\ & -\frac{1}{2}\sum_y n_y[\boldsymbol{\eta}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\eta}\boldsymbol{\nu}_y)]^T(\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta})^{-1}[\boldsymbol{\eta}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\eta}\boldsymbol{\nu}_y)] \\ & -\frac{1}{2}\sum_y n_y(\bar{\mathbf{X}}_y - \boldsymbol{\mu})^T\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu}) \\ & -\sum_y \frac{n_y}{2}\operatorname{tr}\{\boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta})^{-1}\} - \sum_y \frac{n_y}{2}\operatorname{tr}\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\tilde{\boldsymbol{\Delta}}_y\} \end{aligned} \tag{10}$$

where $\mathbf{K} = (\boldsymbol{\eta}_0 - \boldsymbol{\eta}\mathbf{H}^T)$, and $\mathbf{H}$ and $\mathbf{D}$ were defined in Proposition 2. Consider the fourth term T_4 of (10), the only one that involves the $\boldsymbol{\nu}$'s. For any quantity $\mathbf{a}_y$, let $\bar{\mathbf{a}} = \sum_y f_y\mathbf{a}_y$, where $f_y = n_y/n$. We use a Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^d$ to minimize $T_4/n = \sum_y f_y(\mathbf{Z}_y - \bar{\mathbf{B}}\boldsymbol{\nu}_y)^T\mathbf{B}_y^{-1}(\mathbf{Z}_y - \bar{\mathbf{B}}\boldsymbol{\nu}_y) + \boldsymbol{\lambda}^T\bar{\boldsymbol{\nu}}$ subject to the constraint $\bar{\boldsymbol{\nu}} = 0$, where $\mathbf{Z}_y = \boldsymbol{\eta}^T(\bar{\mathbf{X}}_y - \boldsymbol{\mu})$, $\mathbf{B}_y = \boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta}$, and $\bar{\mathbf{B}} = \boldsymbol{\eta}^T\boldsymbol{\Delta}\boldsymbol{\eta}$. Differentiating with respect to $\boldsymbol{\nu}_y$ we get $-2f_y\bar{\mathbf{B}}\mathbf{B}_y^{-1}\mathbf{Z}_y + 2f_y\bar{\mathbf{B}}\mathbf{B}_y^{-1}\bar{\mathbf{B}}\boldsymbol{\nu}_y + f_y\boldsymbol{\lambda} = 0$. Equivalently, $-2f_y\mathbf{Z}_y + 2f_y\bar{\mathbf{B}}\boldsymbol{\nu}_y + f_y\mathbf{B}_y\bar{\mathbf{B}}^{-1}\boldsymbol{\lambda} = 0$. Adding over y the second term is 0, giving the Lagrangian $\boldsymbol{\lambda} = 2\bar{\mathbf{Z}}$. Substituting back and solving for $\boldsymbol{\nu}_y$, $\boldsymbol{\nu}_y = \bar{\mathbf{B}}^{-1}(\mathbf{Z}_y - \mathbf{B}_y\bar{\mathbf{B}}^{-1}\bar{\mathbf{Z}})$. Substituting into T_4 we get the optimized version

$$\tilde{T_4}/n \;=\; \sum_y f_y\bar{\mathbf{Z}}^T\bar{\mathbf{B}}^{-1}\mathbf{B}_j\mathbf{B}_j^{-1}\mathbf{B}_j\bar{\mathbf{B}}^{-1}\bar{\mathbf{Z}} = \bar{\mathbf{Z}}^T\bar{\mathbf{B}}^{-1}\bar{\mathbf{Z}} = (\boldsymbol{\eta}^T\bar{\mathbf{X}} - \boldsymbol{\eta}^T\boldsymbol{\mu})^T\bar{\mathbf{B}}^{-1}(\boldsymbol{\eta}^T\bar{\mathbf{X}} - \boldsymbol{\eta}^T\boldsymbol{\mu}).$$

To find the maximum for $\boldsymbol{\mu}$ we consider

$$\partial L_d/\partial\boldsymbol{\mu} \;=\; n\boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Delta}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}) + n\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}). \tag{11}$$

Using (4) and the definitions of $\mathbf{H}$ and $\mathbf{P}_{\boldsymbol{\eta}(\boldsymbol{\Delta})}$, we have

$$\begin{aligned}
\mathbf{K}^T &= \boldsymbol{\eta}_0^T - (\boldsymbol{\eta}_0^T\boldsymbol{\Delta}\boldsymbol{\eta})(\boldsymbol{\eta}^T\boldsymbol{\Delta}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T = \boldsymbol{\eta}_0^T(\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\eta}(\boldsymbol{\Delta})}) = \boldsymbol{\eta}_0^T\mathbf{P}_{\boldsymbol{\eta}_0(\boldsymbol{\Delta}^{-1})} \\
&= (\boldsymbol{\eta}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_0)^{-1}\boldsymbol{\eta}_0^T\boldsymbol{\Delta}^{-1} \\
\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T &= (\boldsymbol{\eta}_0 - \boldsymbol{\eta}\mathbf{H}^T)^T\mathbf{D}^{-1}(\boldsymbol{\eta}_0^T - \mathbf{H}\boldsymbol{\eta}^T) = \boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_0(\boldsymbol{\eta}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_0)^{-1}\boldsymbol{\eta}_0^T\boldsymbol{\Delta}^{-1}. \quad (12)
\end{aligned}$$

Plugging (12) into (11) and using (3) we get $\partial L_d/\partial\boldsymbol{\mu} = n\boldsymbol{\Delta}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$. Then L_d is maximized on $\boldsymbol{\mu}$ when $\boldsymbol{\mu} = \bar{\mathbf{X}}$ and, with $\tilde{\boldsymbol{\Sigma}}_y = \tilde{\boldsymbol{\Delta}}_y + (\bar{\mathbf{X}}_y - \bar{\mathbf{X}})(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})^T$,

$$\begin{aligned}
L_d &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta}| \\
&\quad -\frac{1}{2}\sum_y n_y \operatorname{tr}\{\boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta})^{-1}\} - \frac{1}{2}\sum_y n_y \operatorname{tr}\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\tilde{\boldsymbol{\Sigma}}_y\}.
\end{aligned}$$

Now, the MLE for $\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta}$ will be such that $\widehat{\boldsymbol{\eta}^T\boldsymbol{\Delta}_y\boldsymbol{\eta}} = \boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}$ and therefore

$$\begin{aligned}
L_d &= -\frac{np}{2}\log 2\pi - \frac{nd}{2} - \frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}| \\
&\quad -\frac{1}{2}\sum_y n_y \operatorname{tr}\{\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\tilde{\boldsymbol{\Sigma}}_y\}.
\end{aligned}$$

To find the MLE for $\mathbf{H}$, recall that $\mathbf{K} = \boldsymbol{\eta}_0 - \boldsymbol{\eta}\mathbf{H}^T$ and consider

$$\frac{\partial L_d}{\partial \mathbf{H}} = -\sum_y n_y\mathbf{D}^{-1}\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\eta} + \sum_y n_y\mathbf{D}^{-1}\mathbf{H}\boldsymbol{\eta}^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\eta}.$$

This gives the maximum at $\widehat{\mathbf{H}} = (\sum_y n_y\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\eta})(\sum_y n_y\boldsymbol{\eta}^T\tilde{\boldsymbol{\Sigma}}_y\boldsymbol{\eta})^{-1} = \boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\eta}(\boldsymbol{\eta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\eta})^{-1}$, where $\tilde{\boldsymbol{\Sigma}} = \sum_y f_y\tilde{\boldsymbol{\Sigma}}_y$. The maximum over $\mathbf{D}$ will be at, using (3),

$$\begin{aligned}
\widehat{\mathbf{D}} &= (\boldsymbol{\eta}_0^T - \widehat{\mathbf{H}}\boldsymbol{\eta}^T)\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\eta}_0^T - \widehat{\mathbf{H}}\boldsymbol{\eta}^T)^T \\
&= [(\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_0)^{-1}\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}]\tilde{\boldsymbol{\Sigma}}[(\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_0)^{-1}\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}]^T = (\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_0)^{-1}.
\end{aligned}$$

Using (5) we get the log-likelihood in $\boldsymbol{\eta}$ as

$$\begin{aligned} L_d &= -\frac{np}{2}(1+\log 2\pi) + \frac{n}{2}\log|\boldsymbol{\eta}_0^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_0| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}| \\ &= -\frac{np}{2}(1+\log 2\pi) + \frac{n}{2}\log|\boldsymbol{\eta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\eta}| - \frac{n}{2}\log|\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\eta}^T\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\eta}|. \end{aligned}$$

The partially maximized log likelihood (1) now follows since $|\mathbf{P}_{\mathcal{S}_{\boldsymbol{\eta}}}\widehat{\boldsymbol{\Sigma}}\mathbf{P}_{\mathcal{S}_{\boldsymbol{\eta}}}|_0 = |\boldsymbol{\eta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}|$.

It can be seen that specifying values for $\boldsymbol{\eta}$, $\mathbf{A} = \boldsymbol{\eta}^T\boldsymbol{\Delta}\boldsymbol{\eta}$, $\mathbf{H}$ and $\mathbf{D}$ uniquely determines $\boldsymbol{\Delta}$. From the MLEs of those quantities, we can obtain the MLE for $\boldsymbol{\Delta}^{-1}$ as follows. Using (12) with (3) gives $\boldsymbol{\Delta}^{-1} = \boldsymbol{\eta}\mathbf{A}^{-1}\boldsymbol{\eta}^T + \mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T$. The MLE for $\boldsymbol{\Delta}^{-1}$ can now be obtained by substituting the previous estimators for $\boldsymbol{\eta}$, $\mathbf{A}$, $\mathbf{H}$ and $\mathbf{D}$ on the right hand side. With $\widehat{\mathbf{K}} = \widehat{\boldsymbol{\eta}}_0 - \widehat{\boldsymbol{\eta}}\widehat{\mathbf{H}}^T$ and using a previous form for $\widehat{\mathbf{D}}$ this estimator can be written as

$$\begin{aligned} \widehat{\boldsymbol{\Delta}}^{-1} &= \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}^T\tilde{\boldsymbol{\Delta}}\widehat{\boldsymbol{\eta}})^{-1}\widehat{\boldsymbol{\eta}}^T + \widehat{\mathbf{K}}(\widehat{\mathbf{K}}^T\tilde{\boldsymbol{\Sigma}}\widehat{\mathbf{K}})^{-1}\widehat{\mathbf{K}}^T \\ &= \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}^T\tilde{\boldsymbol{\Delta}}\widehat{\boldsymbol{\eta}})^{-1}\widehat{\boldsymbol{\eta}}^T + \tilde{\boldsymbol{\Sigma}}^{-1} - \widehat{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}^T\tilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\eta}})^{-1}\widehat{\boldsymbol{\eta}}^T. \end{aligned}$$

A.5 Asymptotic Efficiency

In this appendix we establish our connection with Shapiro's (1986) theory of over-parameterized structural models and discuss the conditions necessary for application of his results. This is not intended to be a comprehensive review. We assume throughout this appendix that $\mathcal{S}_{\text{LAD}} = \mathcal{S}_{Y|\mathbf{X}}$. This assumption holds under normality of $\mathbf{X}_y$ and under the weaker conditions discussed in Section 4.2.

In our context, Shapiro's $\widehat{\mathbf{x}}$ is the vector of length $ph + p(p+1)h/2$ consisting of the h means $\bar{\mathbf{X}}_y$ followed by $\text{vech}(\tilde{\boldsymbol{\Delta}}_y)$, $y = 1, \ldots, h$, where vech is the operator that maps a symmetric $p \times p$ matrix to $\mathbb{R}^{p(p+1)/2}$ by stacking its unique elements. Shapiro's $\boldsymbol{\xi}$ is defined in the same way using the population means $\boldsymbol{\mu}_y$ and variances $\boldsymbol{\Delta}_y$. Then $\sqrt{n}(\widehat{\mathbf{x}} - \boldsymbol{\xi}_0)$ is asymptotically normal with mean 0 and covariance matrix $\boldsymbol{\Gamma} > 0$, where $\boldsymbol{\xi}_0$ denotes the

true value of $\boldsymbol{\xi}$ and $\boldsymbol{\Gamma}$ depends on the distribution of $\mathbf{X}_y$. The structure of $\boldsymbol{\Gamma}$ is conveniently viewed in blocks corresponding to the asymptotic variances 'avar' and covariances 'acov' of the $\bar{\mathbf{X}}_y$'s and $\text{vech}(\tilde{\boldsymbol{\Delta}}_y)$'s. The diagonal blocks are of the form $\text{avar}(\bar{\mathbf{X}}_y) = f_y^{-1}\boldsymbol{\Delta}_y$ and $\text{avar}\{\,\text{vech}(\tilde{\boldsymbol{\Delta}}_y)\} = f_y^{-1}\mathbf{H}(\boldsymbol{\Delta}_y^{1/2} \otimes \boldsymbol{\Delta}_y^{1/2})\text{var}(\mathbf{Z}_y \otimes \mathbf{Z}_y)(\boldsymbol{\Delta}_y^{1/2} \otimes \boldsymbol{\Delta}_y^{1/2})\mathbf{H}^T$, where $\mathbf{Z}_y = \boldsymbol{\Delta}_y^{-1/2}(\mathbf{X}_y - \boldsymbol{\mu}_y)$ and $\mathbf{H}$ is the linear transformation $\text{vech}(\boldsymbol{\Delta}_y) = \mathbf{H}\,\text{vec}(\boldsymbol{\Delta}_y)$. The off-diagonal blocks are all 0 except for $\text{acov}\{\,\text{vech}(\tilde{\boldsymbol{\Delta}}_y), \bar{\mathbf{X}}_y\} = f_y^{-1}\mathbf{H}\text{E}\{(\mathbf{X}_y - \boldsymbol{\mu}_y) \otimes (\mathbf{X}_y - \boldsymbol{\mu}_y)(\mathbf{X}_y - \boldsymbol{\mu}_y)^T\}$.

The next step is to define Shapiro's $\boldsymbol{\xi} = \mathbf{g}(\boldsymbol{\theta})$ to connect with LAD. This is conveniently done by using the reparameterization $\boldsymbol{\delta} = \boldsymbol{\Delta}\boldsymbol{\eta}$. Then from Theorem 1 and part (iv) of Proposition 1 we have $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \boldsymbol{\delta}\boldsymbol{\nu}_y$, $\sum_{y=1}^h f_y\boldsymbol{\nu}_y = 0$, and $\boldsymbol{\Delta}_y = \boldsymbol{\Delta} + \boldsymbol{\delta}\mathbf{M}_y\boldsymbol{\delta}^T$, where the $\mathbf{M}_y$'s are symmetric $d \times d$ matrices with $\sum_{y=1}^h f_y\mathbf{M}_y = 0$. Let $\boldsymbol{\theta}$ consist of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\nu}_1 \ldots, \boldsymbol{\nu}_{h-1}$, $\text{vech}(\boldsymbol{\Delta})$, $\boldsymbol{\delta}$, $\text{vech}(\mathbf{M}_1), \ldots, \text{vech}(\mathbf{M}_{h-1})$, with parameter space $\boldsymbol{\Theta}$ being the product of the parameter spaces for the individual components. The parameter $\boldsymbol{\delta} \in \mathbb{R}^{p\times d}$ is not identified and thus $\boldsymbol{\xi}$ is over-parameterized. Since the elements of $\mathbf{g}$ are analytic functions they are twice continuously differentiable on $\boldsymbol{\Theta}$ and every point in $\boldsymbol{\Theta}$ is regular, except perhaps on a set of Lebesgue measure 0 (Shapiro, 1986, Section 3).

A discrepancy function $F(\widehat{\mathbf{x}}, \boldsymbol{\xi})$ for fitting $\boldsymbol{\xi} = \mathbf{g}(\boldsymbol{\theta})$ must have the properties that $F \geq 0$, $F = 0$ if and only if $\widehat{\mathbf{x}} = \boldsymbol{\xi}$ and F is twice continously differentiable in $\widehat{\mathbf{x}}$ and $\boldsymbol{\xi}$. The LAD discrepancy function is defined as $F_{\text{LAD}}(\widehat{\mathbf{x}}, \boldsymbol{\xi}) = (2/n)\{L_p(\widehat{\mathbf{x}}|\widehat{\mathbf{x}}) - L_d(\boldsymbol{\xi}|\widehat{\mathbf{x}})\}$, where L_d is as given in (10). To emphasize its connection with $\boldsymbol{\xi}$, L_d can also be written, apart from additive constants, as

$$L_d(\boldsymbol{\xi}|\widehat{\mathbf{x}}) \;=\; -\sum_{y=1}^{h}(n_y/2)\{\log|\boldsymbol{\Delta}_y| + \text{tr}(\tilde{\boldsymbol{\Delta}}_y\boldsymbol{\Delta}_y^{-1}) + (\bar{\mathbf{X}}_y - \boldsymbol{\mu}_y)^T\boldsymbol{\Delta}_y^{-1}(\bar{\mathbf{X}}_y - \boldsymbol{\mu}_y)\}.$$

It can be seen from the properties of L_d that F_{LAD} satisfies the conditions necessary for a discrepancy function. For instance, since F_{LAD} is an analytic function of $\widehat{\mathbf{x}}$ and $\boldsymbol{\xi}$ it is twice

continuously differentiable in its arguments. All arguments that minimize $F_{\rm LAD}(\widehat{\mathbf{x}}, \mathbf{g}(\boldsymbol{\theta}))$ are unique except for $\boldsymbol{\delta}$ which is over-parameterized: If $\boldsymbol{\delta}_1$ minimizes $F_{\rm LAD}$ and $\text{span}(\boldsymbol{\delta}_1) = \text{span}(\boldsymbol{\delta}_2)$ then $\boldsymbol{\delta}_2$ also minimizes $F_{\rm LAD}$. Identified and estimable functions of $\boldsymbol{\theta}$ are of the form $k(\boldsymbol{\theta}) = t\{\mathbf{g}(\boldsymbol{\theta})\}$. Then $k(\widehat{\boldsymbol{\theta}})$ is unique for any $\widehat{\boldsymbol{\theta}} = \arg\min F_{\rm LAD}(\widehat{\mathbf{x}}, \mathbf{g}(\boldsymbol{\theta}))$ and is a $\sqrt{n}$-consistent estimator of $k(\boldsymbol{\theta})$. Also, $nF_{\rm LAD}(\widehat{\mathbf{x}}, \mathbf{g}(\widehat{\boldsymbol{\theta}}))$ is equal to the likelihood ratio statistic Λ used in Section 5.

Let $\mathbf{V} = (1/2)\partial^2 F_{\rm LAD}(\widehat{\mathbf{x}}, \boldsymbol{\xi})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T$, evaluated the point $(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0)$. This block diagonal matrix is equal to the Fisher information matrix for $\boldsymbol{\xi}$ based on the full model. The block diagonal elements of $\mathbf{V}^{-1}$ have one of two forms: $f_y^{-1}\boldsymbol{\Delta}_y$ and $2f_y^{-1}\mathbf{H}(\boldsymbol{\Delta}_y \otimes \boldsymbol{\Delta}_y)\mathbf{H}^T$. Now $\mathbf{V}^{-1} = \boldsymbol{\Gamma}$ is a sufficient condition for LAD to give asymptotically efficient F2M estimators (Shapiro, 1986, eq. 5.1.). If $\mathbf{X}_y$ is normal then this relation holds and it follows that the LAD estimator of any identified function of $\boldsymbol{\theta}$ has the smallest asymptotic variances out of the class of minimum discrepancy estimators based on $\widehat{\mathbf{x}}$. If $\mathbf{X}_y$ is not normal then the agreement between $\mathbf{V}^{-1}$ and $\boldsymbol{\Gamma}$ depends only on $\text{acov}\{\text{vech}(\tilde{\boldsymbol{\Delta}}_y), \bar{\mathbf{X}}_y\}$ and $\text{avar}\{\text{vech}(\tilde{\boldsymbol{\Delta}}_y)\}$, since $\text{avar}(\bar{\mathbf{X}}_y) = f_y^{-1}\boldsymbol{\Delta}_y$ is the same as the corresponding element of $\mathbf{V}^{-1}$. If $\mathbf{X}_y$ is symmetric for each $y \in S_Y$ then $\text{acov}\{\text{vech}(\tilde{\boldsymbol{\Delta}}_y), \bar{\mathbf{X}}_y\} = 0$ and asymptotic efficiency depends only on the relation between the fourth moments of $\mathbf{Z}_y$ and those of a standard normal random vector.

A.6 Proof of Proposition 3

To show that $\mathcal{S}_{\rm SAVE} = \mathcal{S}_{\rm LAD}$ we use (5) to write

$$\begin{aligned} K_d(\mathcal{S}) &= c + \frac{1}{2}\log|\mathbf{B}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{B}_0| - \sum_{y=1}^{h}\frac{f_y}{2}\log|\mathbf{B}_0^T\boldsymbol{\Delta}_y^{-1}\mathbf{B}_0| - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \sum_{g=1}^{h}\frac{f_y}{2}\log|\boldsymbol{\Delta}_y| \\ &\leq c - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \sum_{g=1}^{h}\frac{f_y}{2}\log|\boldsymbol{\Delta}_y|, \end{aligned}$$

where $(\mathbf{B}, \mathbf{B}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and $\mathcal{S} = \text{span}(\mathbf{B})$. The inequality follows since $\log|\mathbf{B}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{B}_0| \leq \log|\mathbf{B}_0^T\boldsymbol{\Delta}^{-1}\mathbf{B}_0|$ and the function $\log|\mathbf{B}_0^T\boldsymbol{\Delta}^{-1}\mathbf{B}_0|$ is convex in $\boldsymbol{\Delta}$. Let $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ denote an orthogonal matrix with the columns of $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ forming a basis for $\mathcal{S}_{\text{SAVE}}$. The desired conclusion will follow if we show that

$$\frac{1}{2}\log|\boldsymbol{\beta}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0| - \sum_{g=1}^{h}\frac{f_y}{2}\log|\boldsymbol{\beta}_0^T\boldsymbol{\Delta}_y^{-1}\boldsymbol{\beta}_0| = 0, \tag{13}$$

since $K_d(\mathcal{S})$ will then attain its upper bound at $\mathcal{S} = \mathcal{S}_{\text{SAVE}}$.

It follows from the definition of $\boldsymbol{\beta}$ that for each $y \in S_Y$ there is a vector $\boldsymbol{\omega}_y$ so that $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y) = \boldsymbol{\beta}\boldsymbol{\omega}_y$. Consequently, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y) = \mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y)$. Thus $\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y = \mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma})}^T(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y) = \mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma})}^T(\boldsymbol{\Sigma} - \boldsymbol{\Delta}_y)\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma})}$. From this it can be verified by direct multiplication that $\boldsymbol{\Delta}_y^{-1} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{\beta}\{(\boldsymbol{\beta}^T\boldsymbol{\Delta}_y\boldsymbol{\beta})^{-1} - (\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\}\boldsymbol{\beta}^T$. Substituting this $\boldsymbol{\Delta}_y^{-1}$ into the left side of (13) shows that (13) holds.

References

Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel inference.* New York: Wiley.

Chikuse, Y. (2003), *Statistics on Special Manifolds.* New York: Springer.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences*, 18-25. Alexandria, VA: American Statistical Association.

Cook, R. D. (1998). *Regression Graphics.* New York: Wiley.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association* **100**, 410–428.

Cook, R. D. and Weisberg, S. (1991) Discussion of "Sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association* **86**, 328–332.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion), *Australia New Zealand Journal of Statistics* **43** 147-199.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of Statistics* **12**, 793–815.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20**, 3406–3412.

Li, B. and Wang S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association* **102**, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.

Pardoe, I., Yin, X. and Cook, R.D. (2007). Graphical tools for quadratic discriminant analysis. *Technometrics* **49**, 172–183.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, second ed. New York: Wiley.

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

Shapiro, A. (1986) Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81**, 142–149.

Small, C. G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statistical Science* **15**, 313–332.

Ye, Z. and Weiss, R. (2003). Using the Bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968-978.

Zhu, L., Ohtaki, M. and Li, Y. (2005). On hybrid methods of inverse regression-based algorithms. *Computational Statistics and Data Analysis* **51**, 2621-2635.

Zhu, M. and Hastie, T.J. (2003). Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics* **12**, 101-120.