Journal of Machine Learning Research 21 (2020) 1-24 Submitted 6/19; Revised 5/20; Published 5/20

Prediction regions through Inverse Regression

Emilie Devijver EMILIE.DEVIJVER@UNIV-GRENOBLE-ALPES.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Emeline Perthame EMELINE.PERTHAME@PASTEUR.FR
Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France

Editor: Animashree Anandkumar

Abstract

Predicting a new response from a covariate is a challenging task in regression, which raises new question since the era of high-dimensional data. In this paper, we are interested in the inverse regression method from a theoretical viewpoint. Theoretical results for the well-known Gaussian linear model are well-known, but the curse of dimensionality has increased the interest of practitioners and theoreticians into generalization of those results for various estimators, calibrated for the high-dimension context. We propose to focus on inverse regression. It is known to be a reliable and efficient approach when the number of features exceeds the number of observations. Indeed, under some conditions, dealing with the inverse regression problem associated to a forward regression problem drastically reduces the number of parameters to estimate, makes the problem tractable and allows to consider more general distributions, as elliptical distributions. When both the responses and the covariates are multivariate, estimators constructed by the inverse regression are studied in this paper, the main result being explicit asymptotic prediction regions for the response. The performances of the proposed estimators and prediction regions are also analyzed through a simulation study and compared with usual estimators.

Keywords: Inverse regression, Prediction regions, Confidence regions, High-dimension, Asymptotic distribution

1. Introduction

In a multiple (several response variables) and multivariate (several predictors) regression framework, one wants to describe a response $\mathbf{Y} \in \mathbb{R}^L$ from regressors $\mathbf{X} \in \mathbb{R}^D$. When considering a high number of predictors, the number of parameters could be quickly larger than the sample size, making the estimates impossible to compute in practice or/and providing bad performances for estimators such as lack of stability. This phenomena is generally referred as curse of dimensionality. Several tricks have been proposed in the literature to cope with this issue.

For Gaussian linear models, one of the most famous method is variable selection based on regularized regression, which reduces the dimension of the regression problem to the subset of the most relevant features. Methods include the Lasso (Tibshirani, 1994), the Dantzig selector (Candes and Tao, 2007), or the Ridge estimator (Hoerl and Kennard, 1970) to refer to the most popular. These widely used methods are designed to account for univariate response and few implementations exist for multivariate response, considering then independent response terms. Some extensions have been proposed for generalized linear models, as introduced for example in Bühlmann and van de Geer (2011).

Another way to deal with high dimensional data consists in dimension reduction techniques which extract components or latent variables that summarize the information of a large dataset into a small dimension space. For example, the Principal Component Regression (PCR) selects a subset of principal components for regression and focuses on hyperplanes; the Partial Least Square regression

(PLS) projects the predicted variables and looks for latent variables, correlated to both response and covariates, in order to perform the regression of $\mathbf{Y}$ on $\mathbf{X}$ in a space of lower dimension than D ; and the Sliced Inverse Regression (SIR) introduced in Li (1991) restricts the regressors to few projections by inverting the role of predictors and response. SIR is based on a prior linear dimension reduction by considering the covariance matrix of the inverse expectation $\mathbb{E}(\mathbf{X}|\mathbf{Y})$ (hence the name of the method). The main assumption of SIR relies on Linearity Design Condition, satisfied by elliptical distributions, among which Gaussian distribution, Student distribution and Laplace distribution for the most known. The eigenvectors of this covariance matrix are computed in order to find a subspace that retains the information on $\mathbf{Y}$ contained by the predictors. Many extensions have been proposed, some of them very recently, to data with array (tensor)-valued predictors Ding and Cook (2015), to non linear regression through score function Babichev and Bach (2018), to a sparse estimator Lin et al. (2019) to cite just a few. Some theoretical results have been derived for SIR, as in Hsing and Carroll (1992) where they proved the consistency and asymptotic normality for the 2-slices estimator, and extended for K slices in Saracco (1997). However, the number of axes to retain must be specified beforehand, which is one of the main drawbacks of those methods. Even if procedures have been proposed to choose this parameter (e.g., cross validation, elbow rule, or other heuristics), results are still sensitive to this choice.

More precisely, in the context of regression with random predictors, several authors proposed reduction dimension techniques based on the joint distribution of both predictors and response (George and Oman, 1996; Helland, 1992; Helland and Almøy, 1994) to identify components used to reduce the dimension of predictors matrix. Interestingly, while the regression of interest (referred as *forward* regression in the literature) usually models the conditional distribution of response given predictors $\mathbf{Y}|\mathbf{X}$, some authors explored the properties of inverse models, meaning that the conditional distribution of predictors is studied given the response $\mathbf{X}|\mathbf{Y}$ (referred as *inverse* regression (Oman, 1991)). It has first been introduced for simple linear regression in the Gaussian setting in Krutchkoff (1967); Hunter and Lamboy (1981); Williams (1969) in which it is compared to the least square approach. Its prediction properties have also been studied in Parker et al. (2010) and we found applications of inverse regression to calibration problem (Kannan et al. (2007)) in industrial sector. See also Cook (2007) for an interesting overview of these techniques. The goal of inverse regression techniques is to preserve the information on the regression of interest by studying the inverse conditional distribution as it is directly related to the forward conditional distribution of interest. It consists in inverting the role of response and covariates in the regression model to estimate parameters, taking benefit of the large number of regressors as observations and of the small size of the response. Some works have also considered this method combined with mixture model to propose a nonlinear estimator, for the specific case of Gaussian distribution Deleforge et al. (2015) or Student distribution Perthame et al. (2018).

Whereas variable selection methods are mainly used for high-dimensional data, the inverse regression approach is particularly interesting in three specific frameworks. First, when $D >> N$, if a large number of covariates is known to have an impact on the response (e.g. in planetology (Deleforge et al., 2015)), selecting variables is not relevant while inverse regression is effective. Secondly, when dealing with large dimension for both sample size and number of predictors (N and D large), inverse regression is efficient under few weak assumptions: it avoids the inversion of a large empirical covariance matrix which is time consuming in practice even if it is invertible in theory. Thirdly, inverse regression has the advantage to allow multiple response potentially correlated, which is more and more frequent with real data (e.g. in biology with measurement of multiple phenotypes (El Behi et al., 2017)). Moreover, a more general paradigm is considered for theoretical results, by considering data is modelled within elliptically contoured distributions. It encompasses (among others) the very classical Gaussian as well as Student and contaminated Gaussian distributions, which own good properties for extreme values or outliers.

In this paper, we propose to address the multiple linear regression problem under an inverse regression approach. We study first the theoretical properties of the estimators of the inverse regression model. Then we focus on a prediction purpose by deriving prediction regions. Indeed, under the linear modelling framework, one can predict a new response from a new covariate using the estimator of regression coefficient matrix $\boldsymbol{A}^\star$. Provided that an estimator of $\boldsymbol{A}^\star$ is available, it is relevant to quantify uncertainty around this prediction. This paper focuses on both confidence region for parameters estimates and prediction regions in high-dimensional settings.

Note that few theoretical confidence intervals have been derived in high dimensional context. For Lasso based estimators, Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014) derive confidence regions for slope coefficient and statistical testing of sparsity for linear model using several tools: relaxed projection Zhang and Zhang (2014), desparsifying Lasso van de Geer et al. (2014) or through the computation of an approximate inverse of the Gram matrix Javanmard and Montanari (2014). Since those pioneer works, several articles have provided extensions for more general models or estimators, as generalised linear model (van de Geer et al. (2014) for convex loss function, Janková and van de Geer (2015) for subdifferential loss). We also refer to Meinshausen (2015) for groups of variables and Stucky and van de Geer (2017) for linear regression models with structured sparsity, among others. However, those results rely on strong assumptions on the design and although some authors consider more practical aspects (Chao et al., 2015; Lee et al., 2016), those results still remain difficult to be implemented.

In this paper, we propose to address the linear regression problem for elliptical distributions under an inverse regression approach rather than sparse regression. We assume that the residuals of the inverse model are not correlated which reduces the number of parameters to estimate and overcome the dimensionality burden, while allowing both covariates and residuals of the forward model to be dependent. In this modelling context, assessing confidence in predicted values is one major goal in the manner of least square estimator. However, when the number of predictors becomes too large, least square method suffers from the curse of dimensionality, has bad performances and is computationally intensive while inverse regression approach tackles those problems. Under our framework, we get asymptotic distribution for parameters estimates, and then derive confidence regions for slope coefficients. Moreover, we derive a theorem to quantify prediction uncertainty through asymptotic prediction regions. Then, the properties of parameters estimates are illustrated in an intensive simulation study through finite distance examples.

The paper is organised as follows. In Section 2, the inverse regression model is introduced, as well as the estimation and prediction procedure. Asymptotic distribution of parameters estimates are derived in Section 3. Then, confidence region of slope coefficients and prediction regions are established in Section 4. The finite-sample performance of the proposed confidence and prediction regions are investigated in Section 5, which also includes a comparison with existing methods namely least squares and Lasso. The paper concludes by a discussion in Section 6.

2. Inverse regression model

In this section, the various elements of the modelling framework are introduced.

2.1. Elliptical distributions

Introduced in Li (1991), the inverse regression relies on the Linearity Design Condition (LDC) which relates the covariates to elliptical distribution. First, recall the definition of elliptical distributions.

Definition 1 *Let* $\mathbf{X}$ *be a* d*-dimensional random vector.* $\mathbf{X}$ *is said to be* elliptically distributed *if and only if there exist a vector* $\mu \in \mathbb{R}^d$*, a positive semidefinite matrix* $\Sigma \in \mathbb{R}^{d\times d}$ *and a function* $\phi : \mathbb{R}_+ \to \mathbb{R}$ *such that the characteristic function* $t \mapsto \varphi_{\mathbf{X}-\mu}(t)$ *of* $\mathbf{X} - \mu$ *corresponds to* $t \mapsto \phi(t'\Sigma t), t \in \mathbb{R}^d$*. We write* $\mathbf{X} \sim \mathcal{E}_d(\mu, \Sigma, \phi)$*.*

In this article, we focus on the characterisation provided by the following theorem (we refer to Cambanis et al. (1981) for more details).

Theorem 2 *Let $\mathbf{X}$ be a d-dimensional random vector. $\mathbf{X} \sim \mathcal{E}_d(\mu, \Sigma, \phi)$ with rank$(\Sigma) = k$ if and only if*

$$\mathbf{X} = \mu + \mathcal{R}\Lambda\mathbf{U}^{(k)}$$

where the equality holds in distribution, and where $\mathbf{U}^{(k)}$ is a k-dimensional random vector uniformly distributed on the unit hypersphere with $k-1$ dimensions $\mathcal{S}^{k-1}$, $\mathcal{R}$ is a non-negative random variable with distribution function F related to ϕ being stochastically independent of $\mathbf{U}^{(k)}$, $\mu \in \mathbb{R}^d$ and $\Lambda \in \mathbb{R}^{d\times k}$ with rank$(\Lambda) = k$.

Multivariate normal distribution, multivariate t-distribution and multivariate Laplace distribution are some examples of elliptical distributions. In the following proposition we summarise the main properties we will use in this paper. We refer to Frahm (2004) for details.

Proposition 3 *Let $\mathbf{X} \sim \mathcal{E}_d(\mu, \Sigma, \phi)$ with rank$(\Sigma) = k$. The following hold:*

- *$E(\mathbf{X}) = \mu$,*
- *$Var(\mathbf{X}) = \frac{E(\mathcal{R}^2)}{k}\Sigma = -2\phi'(0)\Sigma$ if ϕ is differentiable at 0,*
- *An affine transformation of an elliptic random variable is also elliptic.*
- *Conditional distribution: let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is a k-dimensional sub-vector of $\mathbf{X}$, and let $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} \in \mathbb{R}^{d\times d}$. Provided the conditional random vector $\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1$ exists, it is also elliptically distributed: $\mathbf{X}_2|(\mathbf{X}_1 = \mathbf{x}_1) \sim \mathcal{E}_{d-k}(\mu^*, \Sigma^*, \phi^*)$. Moreover, it can be represented stochastically by*

  $$\mathbf{X}_2|(\mathbf{X}_1 = \mathbf{x}_1) = \mu^* + \mathcal{R}^*\mathbf{U}^{(r-k)}\Gamma^*$$

  where the equality holds in distribution, and $\mathbf{U}^{(r-k)}$ is uniformly distributed on $\mathcal{S}^{r-k-1}$, and

  $$\begin{aligned}\mathcal{R}^* &= \mathcal{R}\sqrt{1-\beta}|(\mathcal{R}\sqrt{\beta}\mathbf{U}^{(k)} = \Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1))\\ \mu^* &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1)\\ \Sigma^* &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{aligned}$$

  where $\beta \sim Beta(k/2, (r-k)/2)$ and $\mathcal{R}, \beta, \mathbf{U}^{(k)}$ and $\mathbf{U}^{(r-k)}$ are supposed to be mutually independent, and $\Sigma^* = (\Gamma^*)^t\Gamma^*$.
- *The sum of independent elliptically distributed random vector with the same dispersion matrix Σ is elliptical too (Hult and Lindskog, 2002).*

2.2. Inverse regression method

For a vector of responses $\mathbf{Y}_i \in \mathbb{R}^L$ and covariates $\mathbf{X}_i \in \mathbb{R}^D$, let introduce the following generative model:

$$\begin{bmatrix}\mathbf{X}_i\\ \mathbf{Y}_i\end{bmatrix} \sim \mathcal{E}_{D+L}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}; \begin{bmatrix}\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top & \boldsymbol{A}\mathbf{\Gamma}\\ (\boldsymbol{A}\mathbf{\Gamma})^\top & \mathbf{\Gamma}\end{bmatrix}; \phi\right); \tag{1}$$

Using Proposition 3, this joint distribution leads to the linear regression problem we aim to address in this paper, characterized by the following marginal and conditional distributions:

$$\mathbf{X}_i \sim \mathcal{E}_D(\mathbf{0}, \mathbf{\Gamma}^\star, \phi) \text{ with } \mathrm{rank}(\mathbf{\Gamma}^\star) = D \tag{2}$$

$$\mathbf{Y}_i|\mathbf{X}_i = \boldsymbol{A}^\star\mathbf{X}_i + \boldsymbol{\varepsilon}_i \tag{3}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N) \in \mathbb{R}^{L\times N}$ corresponds to L responses for N subjects and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N) \in \mathbb{R}^{D\times N}$ contains D elliptical centered predictors with covariance matrix $\mathbf{\Gamma}^\star$. The error term $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N)$ is an unobserved $L \times N$ matrix with independent columns elliptically distributed, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \sim \mathcal{E}_L(\mathbf{0}, \mathbf{\Sigma}^\star, \phi_{\boldsymbol{\varepsilon}})$ with $\mathrm{rank}(\mathbf{\Sigma}^\star) = L$, independent from the regressors. The $L \times D$ matrix of slope coefficients is denoted by $\boldsymbol{A}^\star$. When D is large or/and when the number of observations N is smaller than D, the so-called least square estimate of $\boldsymbol{A}^\star$ is not numerically computable for the *forward regression* defined in Equations (2) and (3). Indeed, it requires the inversion of the possibly large matrix $\mathbf{X}^\top\mathbf{X}$ which is not invertible when $D > N$ and computationally intensive for large D when $N > D$. An interesting and relatively simple approach to handle this high dimensional problem is to consider the *inverse regression* problem, derived as well from the joint distribution in Equation (1):

$$\mathbf{Y}_i \sim \mathcal{E}_L(\mathbf{0}, \mathbf{\Gamma}, \phi) \text{ with } \mathrm{rank}(\mathbf{\Gamma}) = L \tag{4}$$

$$\mathbf{X}_i|\mathbf{Y}_i = \boldsymbol{A}\mathbf{Y}_i + \mathbf{e}_i \tag{5}$$

where $\boldsymbol{A}$ is a $D \times L$ matrix of slope coefficients of the *inverse regression* and $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$ is a $D \times N$ matrix of unobserved centered elliptical random noise with residual covariance matrix $\mathbf{\Sigma}$, of type $\phi_{\mathbf{e}}$, independent from $\mathbf{Y}$. The inverse regression approach consists in inverting the response and the covariates in the model and performing regression of response on covariates. While least squares estimate is not computable in high dimension for forward regression, it turns out that dealing with the inverse regression problem, under some assumptions on the noise $\mathbf{e}$ detailed hereafter, drastically reduces the number of parameters and makes the problem tractable.

Note that no intercept is considered in models (4) and (5), which leads to assume that both response and covariates are centered.

The key point of the model introduced in Equation (1) is that forward parameters $(\mathbf{\Gamma}^\star, \boldsymbol{A}^\star, \mathbf{\Sigma}^\star)$ are expressed in function of the inverse parameters $(\mathbf{\Gamma}, \boldsymbol{A}, \mathbf{\Sigma})$ through the following one-to-one mapping :

$$\begin{aligned}\Psi : (\mathbf{\Gamma}, \boldsymbol{A}, \mathbf{\Sigma}) \mapsto &(\mathbf{\Gamma}^\star, \boldsymbol{A}^\star, \mathbf{\Sigma}^\star)\\ =&(\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}).\end{aligned} \tag{6}$$

Details to prove the involution between forward and inverse regression are given in Appendix, relying on the joint distribution introduced previously. As the mapping Ψ is an involution, the forward regression parameters in model (2)-(3) map to the inverse regression parameters in model (4)-(5). The advantage of the inverse approach appears when structure is assumed on large covariance matrix $\mathbf{\Gamma}^\star$. Indeed, the linear regression problem we address in this paper is not computable when the number of predictors D becomes large. When $D > N$, the LSE is not computable as the matrix $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ is singular and when $N > D$, even if the problem is analytically tractable, computation of the LSE involves the inversion of a large $D \times D$ matrix. To handle this issue, a solution would be to decrease the dimension. Three solutions can be proposed: either using variable selection (such as model selection or lasso approaches) ; or using projection methods (such as PLS, PCR or SIR approaches) if all variables are active in the prediction, if the user does not assume that its regression problem is sparse, or if its purpose is just prediction, with no necessarily variable selection ; or at last, in the same paradigm, one can decrease the dimension of the problem by reducing the number of parameters to estimate with respect to structure assumptions on the modelling. A drastic way to decrease the number of parameters would be to make assumptions on the large covariance matrix involved in the problem $\mathbf{\Gamma}^\star$ and assume that it is diagonal. This would result in a strong assumption on dependence structure of predictors (no correlation allowed) but would drastically reduces the number of parameters and make the problem computable. Our suggestion is to assume a structured dependence among predictors which would be flexible enough to allow correlations among predictors, but constrained enough to efficiently decrease the dimension. For example, a block structure or a Toeplitz matrix would be convenient. Nevertheless, it turns out that the low rank decomposition, also known as factor model we assume for $\mathbf{\Gamma}^\star = \mathbf{\Sigma} + \mathbf{A\Gamma A^T}$ provides a good approximation and is flexible enough to handle various applied problems while being numerically tractable (Leek and Storey (2007, 2008); Carvalho et al. (2008); Sun and Cai (2009); Blum et al. (2010); Causeur et al. (2011); Teschendorff et al. (2011); Perthame et al. (2015)). Notice that a direct consequence of this assumption on $\mathbf{\Gamma}^\star$ in the forward regression problem is to assume that the residuals of the inverse regression are not correlated, which is the assumption we make in this paper. Assuming that $\mathbf{\Sigma}$ is diagonal drastically reduces the number of parameters to estimate: for example, if $D = 100$ and $L = 5$, the number of parameters to estimate goes from $LD + L(L+1)/2 + D(D+1)/2 = 5565$ in the full model to $LD + L(L+1)/2 + D = 615$ under our model. The robustness to this assumption is assessed in the Case 4 of the simulation study in Section 5.

2.3. Estimation

Considering the inverse model defined in Equations (4)-(5), the least squares estimators are:

$$\begin{aligned} \widehat{\mathbf{\Gamma}} &= \frac{1}{N-1}\mathbf{Y}^\top\mathbf{Y}; \\ \widehat{\boldsymbol{A}}^T &= (\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}; \\ \forall j \in \{1,\dots,D\}, \widehat{\mathbf{\Sigma}}_{j,j} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_{i,j} - [\widehat{\boldsymbol{A}}\mathbf{Y}_i]_j)^2. \end{aligned} \tag{7}$$

Using mapping Ψ, it is straightforward to deduce estimators for the forward regression:

$$\widehat{\mathbf{\Gamma}}^\star = \widehat{\mathbf{\Sigma}} + \widehat{\boldsymbol{A}}\widehat{\mathbf{\Gamma}}\widehat{\boldsymbol{A}}^\top; \tag{8}$$

$$\widehat{\boldsymbol{A}}^\star = (\widehat{\mathbf{\Gamma}}^{-1} + \widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Sigma}}^{-1}; \tag{9}$$

$$\widehat{\mathbf{\Sigma}}^\star = (\widehat{\mathbf{\Gamma}}^{-1} + \widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{A}})^{-1}. \tag{10}$$

The inverse regression trick allows to numerically compute those estimators even when $D >> N$ as it requires the inversion of the $L \times L$ matrix $\mathbf{Y}^T\mathbf{Y}$ and not the inverse of $\mathbf{X}^T\mathbf{X}$ as in the forward regression. Assuming $\mathbf{\Sigma}$ is diagonal avoids also to invert a full $D \times D$ matrix. To ensure invertible estimator of $\mathbf{\Gamma}$, it only requires the response dimension to be smaller than the sample size.

2.4. Prediction of the response

Considering those estimators $(\widehat{\boldsymbol{A}}^\star, \widehat{\mathbf{\Gamma}}^\star, \widehat{\mathbf{\Sigma}}^\star)$, a new response $\widehat{\mathbf{Y}}_{N+1}$ is predicted for a new observed profile $\mathbf{x}_{N+1}$ from Model (3) and defined by:

$$\widehat{\mathbf{Y}}_{N+1} = \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}_{N+1}) = \widehat{\boldsymbol{A}}^\star\mathbf{x}_{N+1}.$$

The purpose of this article is to study the uncertainty around this prediction which can be quantified by deriving prediction region. Moreover, we deeply study the theoretical properties of the estimators of this model and establish the exact distribution of $\widehat{\mathbf{\Gamma}}^\star$ and $\widehat{\mathbf{\Sigma}}^\star$ and the asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$ which is involved into prediction region computation.

3. Theoretical study of the estimators

In this section, we assume that covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ are both known and that $\mathbf{\Sigma}$ is diagonal as previously stated. In this section, asymptotic distribution of estimators for the forward regression are derived.

3.1. Matrix variate elliptical distribution and Kronecker product

First we recall some properties about the matrix variate elliptical distribution and the tensor product. These results can be found in Gupta and Nagar (2000) Chapter 2, but some properties are recalled in this paper as we use them extensively.

Definition 4 (Kronecker product) *Let $A \in M_{m,n}(\mathbb{R})$ be a $m \times n$ matrix with real elements denoted by $(a_{i,j})$ with $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ and $B \in M_{p,q}(\mathbb{R})$. Then, the Kronecker product $A \otimes B$ is the $mp \times nq$ block matrix:*

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots, & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & A_{mn}B \end{pmatrix}.$$

The vectorization hereafter is used to work with vectors instead of matrices.

Definition 5 (Vectorization) *The vectorization $vec(A)$ of a matrix A is a linear transformation which converts the matrix into a column vector, by stacking the columns of the matrix on top of one another.*

Gupta and Varga (1994) define the matrix variate elliptical distribution through characteristic function and derive a theorem to characterize this distribution using vec operator which is more convenient to derive results in this paper (Theorems 2.1 and 2.3 in Gupta and Varga (1994)).

Definition 6 (Matrix variate elliptical distribution) *The random variable $\mathbf{X} \in \mathbb{R}^{L \times D}$ is distributed according to a* matrix variate elliptical distribution *with mean $\mathbf{X}_0$ and variances $U \in M_{L,L}(\mathbb{R})$ (among-row) and $V \in M_{D,D}(\mathbb{R})$ (among-column), denoted*

$$\mathbf{X} \sim \mathcal{ME}_{L,D}(\mathbf{X}_0, U \otimes V, \phi),$$

if and only if $vec(\mathbf{X}) \sim \mathcal{E}_{LD}(vec(\mathbf{X}_0), V \otimes U, \phi)$.

For this distribution, some interesting properties are derived.

Proposition 7 *The following equivalence holds:*

$$\mathbf{X} \sim \mathcal{ME}_{L,D}(\mathbf{X}_0, U \otimes V, \phi) \qquad \Leftrightarrow \qquad \mathbf{X}^T \sim \mathcal{ME}_{D,L}(\mathbf{X}_0^T, V \otimes U, \phi).$$

Proposition 8 *If $\mathbf{X} \sim \mathcal{ME}_{LD}(\mathbf{X}_0, U \otimes V, \phi)$, the following properties hold for $A \in \mathcal{M}_{r,L}(\mathbb{R})$ and $B \in \mathcal{M}_{D,s}(\mathbb{R})$*

$$\begin{aligned} A\mathbf{X}B &\sim \mathcal{ME}_{r,s}(A\mathbf{X}_0 B, AUA^T \otimes B^T V B, \phi); \\ vec(A\mathbf{X}B) &= (B^T \otimes A) vec(\mathbf{X}). \end{aligned}$$

For $A \in M_{r,L}(\mathbb{R}), B \in M_{D,s}(\mathbb{R}), C \in M_{r,L}(\mathbb{R}), D \in M_{D,s}(\mathbb{R})$ and if $\mathbf{X}$ has finite second order moments, the following holds:

$$\begin{aligned} Cov(vec(A\mathbf{X}B), vec(C\mathbf{X}D)) &= -2\phi'(0)(B^T V D \otimes AUC^T); \\ Cov(vec(A\mathbf{X}B), vec(C\mathbf{X}^T D)) &= -2\phi'(0)(B^T \otimes A)E(vec(\mathbf{X})vec(\mathbf{X})^T)T_{LD}^{-1}(D^T \otimes C) \\ &= -2\phi'(0)(B^T V \otimes AU)T_{LD}^{-1}(D^T \otimes C) \text{ for } \mathbf{X} \text{ centered.} \end{aligned}$$

where T_{LD} is the commutation matrix, transforming the vectorized form of a matrix of size $L \times D$ into the vectorized form of its transpose.

3.2. Asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$

In order to derive the asymptotic distribution of the forward regression coefficients $\widehat{\boldsymbol{A}}^\star$, the distribution of the inverse regression coefficients matrix $\widehat{\boldsymbol{A}}$ is described at first. We start by describing the conditional distribution of $\hat{\boldsymbol{A}}$ conditionally to $\mathbf{Y}_1, \dots, \mathbf{Y}_n$.

Proposition 9 (Distribution of $\widehat{\boldsymbol{A}}$ conditionally to $\mathbf{Y}$) *Suppose* $((\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_N, \mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*, or equivalently in Equations* (4) *and* (5)*. Then, conditionally to* $\mathbf{Y}$*,*

$$\widehat{\boldsymbol{A}} - \boldsymbol{A} \sim \mathcal{ME}_{D,L}(\mathbf{0}, \frac{-1}{2\phi'(0)} \boldsymbol{\Sigma} \otimes (\mathbf{Y}^T \mathbf{Y})^{-1}, \phi).$$

This result is an extension of the least square estimator in the multivariate linear model to the multiple multivariate linear model. The proof uses that $\hat{\boldsymbol{A}}$ is a linear combination of $\mathbf{X}$ and is given in Appendix 7. However, deriving confidence regions for forward regression based on a distribution conditioned on the responses would provide conditional confidence regions. To get rid of the conditioning, we provide the asymptotic distribution of $\widehat{\boldsymbol{A}}$.

Proposition 10 (Distribution of $\widehat{\boldsymbol{A}}$) *Suppose* $((\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_N, \mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*, or equivalently in Equations* (4) *and* (5)*. Then,*

$$\sqrt{N}(\widehat{\boldsymbol{A}} - \boldsymbol{A}) \underset{n \to +\infty}{\to} \mathcal{MN}_{D,L}(\mathbf{0}, \frac{\phi'_{\mathbf{e}}(0)}{\phi'(0)} \boldsymbol{\Sigma} \otimes \boldsymbol{\Gamma}^{-1}).$$

This result is an application of the Central Limit Theorem, and the proof is given in Appendix 7.

From this, we derive the asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$. A matrix version of the Δ-method is used, which involves the differential of the function $g : \boldsymbol{A} \mapsto \boldsymbol{A}^\star$ and the corresponding asymptotic variance of $\widehat{\boldsymbol{A}}^\star$. They are first computed in the following lemma.

Lemma 11 *Suppose* $((\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_N, \mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Let*

$$\begin{aligned} g : \mathbb{R}^{D \times L} &\to \mathbb{R}^{L \times D} \\ \boldsymbol{A} &\mapsto \boldsymbol{A}^\star = \boldsymbol{\Sigma}^\star \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{11}$$

Then the differential of this function at point $(\widehat{\boldsymbol{A}} - \boldsymbol{A})$ *is,*

$$Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}) = \boldsymbol{\Sigma}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A})^T \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star - \boldsymbol{A}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A}) \boldsymbol{A}^\star. \tag{12}$$

Moreover, the covariance of this random matrix is given by the following, if ϕ *is differentiable in 0,*

$$\begin{aligned} &\frac{\phi'(0)}{\phi'_{\mathbf{e}}(0)} Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}))) = \\ &\qquad \big((\boldsymbol{\Sigma}^{-1} + (\boldsymbol{A}^\star)^T \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star - 2 \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star) \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma} \boldsymbol{\Sigma}^\star\big) \\ &\qquad + ((\boldsymbol{A}^\star)^T \boldsymbol{\Gamma} \boldsymbol{A}^\star \otimes \boldsymbol{A}^\star \boldsymbol{\Sigma} (\boldsymbol{A}^\star)^T) \\ &\qquad - 2 \big((\mathbf{I} \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma}) + ((\boldsymbol{A}^\star)^T \boldsymbol{A}^T \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma})\big) T_{LD}^{-1} ((\boldsymbol{A}^\star)^T \otimes \boldsymbol{A}^\star). \end{aligned} \tag{13}$$

Proof of Lemma 11 is given in Appendix. Those explicit formulae allow to compute confidence regions in practice using the so-called plug-in estimator of $\Theta(\mathbf{A})$. Numeric performances stand in Section 5.

Finally, the following theorem, which is the key of this paper, details the asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$.

Theorem 12 (Asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$) *Suppose* $((\mathbf{X}_1,\mathbf{Y}_1),\dots,(\mathbf{X}_N,\mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3). *Let*

$$\begin{aligned} g:\mathbb{R}^{D\times L} &\to \mathbb{R}^{L\times D} \\ \boldsymbol{A} &\mapsto \boldsymbol{A}^\star = \boldsymbol{\Sigma}^\star \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Gamma}^{-1}+\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{14}$$

Then, the following holds for the estimator $\widehat{\boldsymbol{A}}^\star$ *defined in Equation* (9),

$$\sqrt{N}(vec(\widehat{\boldsymbol{A}}^\star)-vec(\boldsymbol{A}^\star)) \underset{N\to+\infty}{\to} \mathcal{N}_{DL}(\mathbf{0},\Theta(\boldsymbol{A}));$$

where $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A})))$ *defined in Equation* (13).

Moreover, $\Theta(\widehat{\boldsymbol{A}})$ *is a consistent estimator of* $\Theta(\boldsymbol{A})$, *and*

$$\sqrt{N}(vec(\widehat{\boldsymbol{A}}^\star)-vec(\boldsymbol{A}^\star))^T\Theta(\widehat{\boldsymbol{A}})^{-1}(vec(\widehat{\boldsymbol{A}}^\star)-vec(\boldsymbol{A}^\star)) \underset{N\to+\infty}{\to} \chi^2_{DL}. \tag{15}$$

where χ^2_q *denotes the* χ^2*-quantile function with* q *degrees of freedom.*

Proof The matrix version of the Δ-method is a second order Taylor expansion of $g:\boldsymbol{A}\mapsto\boldsymbol{A}^\star$. Therefore, for $\boldsymbol{A}\in M_{D,L}(\mathbb{R})$ and g defined by Equation (14), the Taylor expansion leads to

$$\widehat{\boldsymbol{A}}^\star = g(\widehat{\boldsymbol{A}}) = g(\boldsymbol{A}) + Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A}) + R_N(\widehat{\boldsymbol{A}});$$

with $R_N(\widehat{\boldsymbol{A}})$ a rest term that vanishes to 0 when $N\to+\infty$ and $Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A})$ is given in Lemma 11.

Then,

$$\sqrt{N}(\widehat{\boldsymbol{A}}^\star-\boldsymbol{A}^\star) = \sqrt{N}Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A}) + \sqrt{N}R_N(\widehat{\boldsymbol{A}}). \tag{16}$$

The last term in (16) converges to 0 in probability, and by Proposition 10, the linear combination with respect to $\hat{\boldsymbol{A}}$ defined in (12) is a matrix variate Gaussian distribution, centered. Using (13), we get the distribution of the vectorized vector $vec(\widehat{\boldsymbol{A}}^\star)$.

Asymptotic distribution of the quadratic form Equation (15) is deduced using both Slutsky's theorem and Corollary 1 of Cambanis et al. (1981). ■

This result is the key theorem of this article as it provides closed-form expressions to derive confidence regions for $\boldsymbol{A}^\star$ and prediction regions. This theorem is very general: we assume that $(\mathbf{X},\mathbf{Y})$ follows any matrix elliptical distribution; then, the asymptotic distribution of the regression coefficients estimated by the inverse method, is Gaussian. Even if we can get the asymptotic normality from the least square estimator for $\boldsymbol{A}^\star$, the goal of this paper is to show that, for large number of covariates and large number of responses, the inverse regression has better results than the least square estimators. Numerical experiments in accord with this are available in Section 5.

4. Confidence regions and predictions regions

In this section, we provide confidence regions for $vec(\boldsymbol{A}^\star)$ and prediction regions for $\mathbf{Y}$ through the inverse regression method.

4.1. Confidence regions for $\boldsymbol{A}^\star$

Theorem 13 *Suppose* $((\mathbf{X}_1,\mathbf{Y}_1),\dots,(\mathbf{X}_N,\mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3). *Then, a confidence region for* $\boldsymbol{A}^\star$ *is*

$$P\left(vec(\boldsymbol{A}^\star)\in\tilde{\mathcal{R}}_{vec(\boldsymbol{A}^\star),\alpha}\right) \underset{n\to+\infty}{\to} 1-\alpha;$$

where

$$\tilde{\mathcal{R}}_{vec(\boldsymbol{A}^\star),\alpha} = \Big\{ \boldsymbol{a}^\star \in M_{L,D}(\mathbb{R}) \ s.t.\ (vec(\boldsymbol{a}^\star - \widehat{\boldsymbol{A}}^\star))^T \Theta(\boldsymbol{A})^{-1} (vec(\boldsymbol{a}^\star - \widehat{\boldsymbol{A}}^\star)) \leq \chi^2_{DL}(1-\alpha) \Big\}\,;$$

with $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$ *defined in Equation* (13) *and* χ^2_q *the* χ^2*-quantile function with* q *degrees of freedom.*

Those explicit formulae allow to compute confidence regions in practice. Numeric performances stand in Section 5.

4.2. Prediction regions

Theorem 14 *Suppose* $((\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_N, \mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Then,*

$$P\left(\mathbf{Y}_{N+1} \in \widetilde{\mathcal{PR}}_{\mathbf{Y},\alpha}\right) \underset{N\to+\infty}{\to} 1-\alpha;$$

where

$$\begin{aligned} \widetilde{\mathcal{PR}}_{\mathbf{Y},\alpha} = \big\{ &\mathbf{y} \in \mathbb{R}^L \ s.t. \\ &(\mathbf{y} - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1})^T (\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1}) + \boldsymbol{\Sigma}^\star)^{-1} (\mathbf{y} - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1}) \leq \chi^2_L(1-\alpha) \Big\}\,; \end{aligned} \tag{17}$$

where $\Omega(\widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1})$ *is the following* $L \times L$ *covariance matrix;*

$$\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1}) = (\mathbb{I}_L \otimes \mathbf{X}^T_{N+1}) \Theta(\boldsymbol{A}) (\mathbf{X}^T_{N+1} \otimes \mathbb{I}_L);$$

where $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$ *defined in Equation* (13)*.*

One can notice that the covariance matrix that is inverted in Equation (17) breaks down into 2 parts. The first one, $\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1})$, represents the variance of the prediction which depends on the estimation accuracy of $\boldsymbol{A}^\star$ while the second part, $\boldsymbol{\Sigma}^\star$, is the variance inherited from the residuals.

5. Simulations

The goal of this section is to compute the prediction regions derived from the theoretical results presented in Section 4. For several designs regarding the sample size, the dimension, the sparsity and several covariance patterns, we study the coverage, the volume of the interval and the computation time. For comparison, we also compute prediction intervals deduced from the least square estimator or a regularized approach (depending on the dimension we consider)[1].

5.1. Simulation design

In order to assess the impact of data dimension and design complexity on different estimation methods of prediction regions, we perform a simulation study. The Gaussian setting is considered. The response dimension L is varying in $\{1, 2, 5\}$. Indeed, when $L = 1$ or 2, prediction regions are easily graphically displayable which is useful to visualize methods. We focus on four distinct designs: for $D = 100$, we consider a high-dimensional one with $N = 50$, an asymptotic one with $N = 500$ and an intermediate design with $N = 100$. We also study a design with $D = 1000$ and $N = 100$ to assess our method when D is large. Data are simulated according to an inverse regression model and forward parameters are deduced from Equation (6). For each combination of dimension, we focus on the 4 following scenarii:

[1] The `R` code to use the 3 compared methods on simulated data is available at https://research.pasteur.fr/fr/member/emeline-perthame/.

(Case 1) Sparse regression coefficients and independent responses: $\mathbf{A}$ is a $D \times L$ matrix with 3% of its rows with non zero entries, uniformly drawn into a uniform distribution on (-3, 3) for D = 100 and (-1, 1) for D = 1000. This setting ensures that the sparsity level of $\mathbf{A}^\star$ is 3% regardless to L, as proposed in van de Geer et al. (2014). Coefficients are drawn from a uniform distribution on $(-2, 2)$ for $D = 100$ and $(-1, 1)$ for $D = 1000$. Matrix $\mathbf{\Gamma}$ of covariances between response terms is set to $\mathbb{I}_L$. The residual covariance matrix of inverse regression $\mathbf{\Sigma}$ is set to $\mathbb{I}_D$. Note that a diagonal $\mathbf{\Sigma}$ and a sparse $\boldsymbol{A}$ under the inverse model lead to a sparse matrix of regression coefficients for forward regression $\boldsymbol{A}^\star$.

(Case 2) Sparse regression coefficients and correlated responses: same as previous scenario except that $\mathbf{\Gamma}$ is a full covariance matrix generated according to a factor model such as dependence among response terms is rather strong.

(Case 3) Full matrix of regression coefficients and correlated responses: coefficient matrix $\boldsymbol{A}$ is full with entries uniformly sampled in $[-0.5, 0.5]$ for $D = 100$ and $[-0.125, 0.125]$ for $D = 1000$ (to ensure similar SNRs) and covariance matrix $\mathbf{\Gamma}$ is generated as in Case 2. The residual covariance matrix $\mathbf{\Sigma}$ is set to $\mathbb{I}_D$.

(Case 4) Unconstrained residual covariance matrix $\mathbf{\Sigma}$: same as previous scenario except that $\mathbf{\Sigma}$ is a full covariance matrix generated according to a factor model such as dependence among residuals of inverse regression terms is medium. Note that this scenario violates our assumption that $\mathbf{\Sigma}$ is diagonal and allows to assess the robustness of our approach to this limitation.

Cases 1 and 2 mimic designs with a low number of active covariates to predict the response. Therefore, they are favourable for the lasso, which performs variable selection. In Case 2, responses are correlated, which might affect the lasso. Case 3 falls in our inverse regression paradigm so it favours IR and LSE (when computable) but not the lasso. At last, Case 4 violates the assumption we made so only LSE is favoured (when computable), as it does not make any assumption on the structure of the design. Note that the amplitude of coefficients in $\boldsymbol{A}$ differs from one case to another. This amplitude is adjusted in order to make scenarii comparable regarding to the signal to noise ratio (SNR) criterion defined as:

$$\text{SNR} = \frac{1}{L}\text{trace}(\boldsymbol{A}^\star \mathbf{\Gamma}^\star (\boldsymbol{A}^\star)^T (\mathbf{\Sigma}^\star)^{-1});$$

where trace refers to the sum of diagonal entries of a matrix. In this simulation setting, for all cases and all values of L, the SNR varies between 8 and 10 which is rather (reasonably) high. Note that we extended the well-known SNR definition of Verzelen and Gassiat (2017) to our multivariate response framework. However, note that even if SNR values are similar for the different settings, the evolution of performance with respect to L is difficult to analyze because the difficulty of the regression problem depends on L.

Datasets are generated under a linear regression model as defined in Equations (2)-(3). For each simulated design, 1 000 learning datasets with dimension (N, D) are generated as well as 1 000 corresponding testing observations. Note that the computation of prediction regions for inverse model involves the computation of a commutation matrix. To compute such matrices, we used the fast routine implemented in the function `commutation.matrix` available in the `R` package `matrixcalc`.

We compare the prediction regions derived from the 3 following methods: the proposed method based on inverse regression referred as IR in the following, the so-called least square estimator (LSE) for designs with $N > D$ and a Lasso prediction interval based on bootstrap. To our knowledge, no computation of prediction regions for the lasso is implemented into a R package. In our simulation study, we therefore implemented our own version of prediction regions for the lasso, based on a bootstrap approach. N observations are sampled with replacement within the training set and

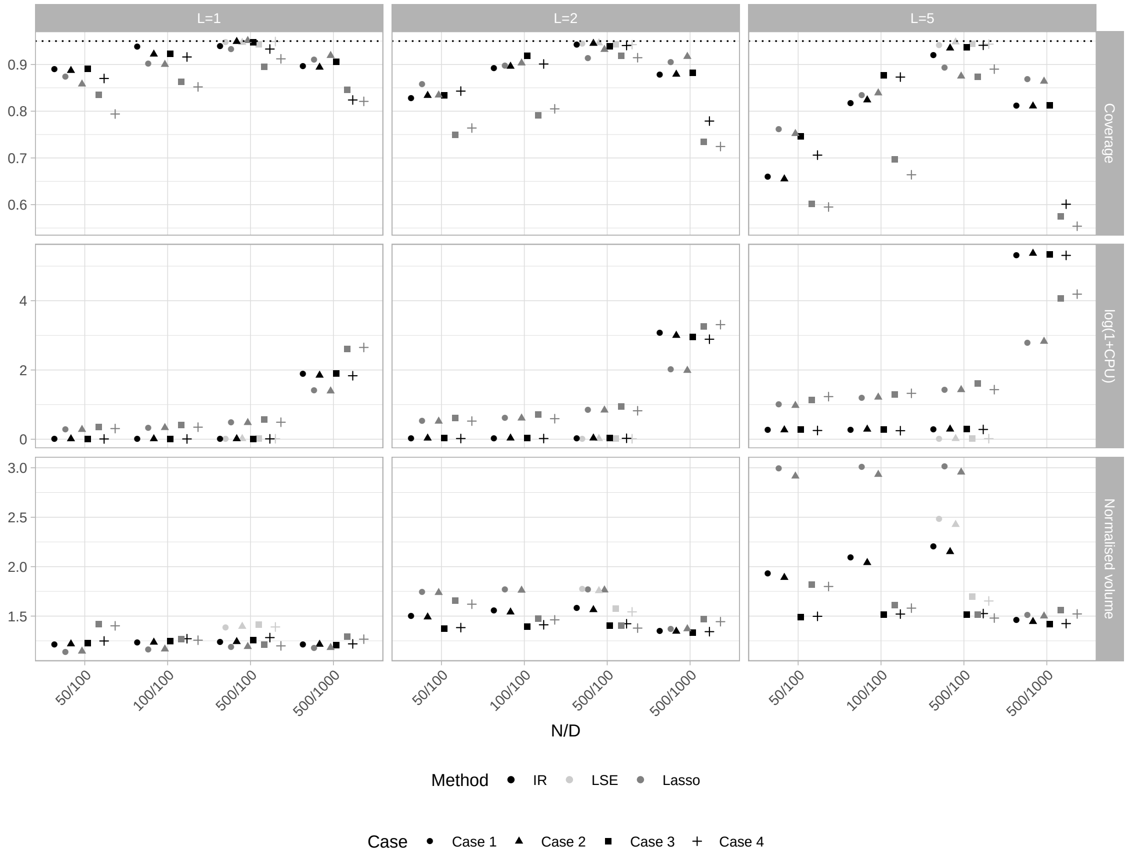

Figure 1: Results of simulations study for Gaussian distribution: prediction regions computed on datasets simulated under models described in Section 5.1. Coverage, normalised volume and CPU time are computed for each method to compare performances. IR (in black) is compared to the bootstrapped Lasso (in dark grey) and to LSE (in light grey) for designs with $N > D$. Each method is assessed 1000 times, and mean is computed and reported on the graph with colour for the method and dot shape for the 4 cases. We added a small amount of random variation to the horizontal location of each point in order to avoid overlap.

lasso variable selection is applied using `glmnet` R package. For each bootstrapped training set, the response of testing set is predicted. By repeating this procedure B = 100 times, the distribution of the prediction is estimated. The prediction region is deduced using the $\alpha/2$ and $1-\alpha/2$ quantiles of this distribution. For more details, the code is available online on the author's webpage. The accuracy of the method is assessed by computing the coverage (proportion of testing observations falling into the prediction region), the normalised volume of the prediction regions (defined as the Lth root of the volume) and the computation time (on log scale) required to compute the prediction region on a MacBook Pro - 2,9 GHz Intel Core i5 processor - RAM 16 Go with programs written in `R`. In this simulation study, the level of confidence for prediction regions is set to 95%.

5.2. Results of the intensive simulation study

The results of this simulation study are presented in Figure 1 (corresponding numerical values are available in Table 1 in Appendix). This figure presents the results for varying sample and design sizes in column, and coverage, volume and time computation in row for varying methods and response dimension. For each scenario, IR (in black) is compared to Lasso (in dark grey) and to LSE (in light gray) when $N > D$.

First, Figure 1 demonstrates that the proposed method reached a valid asymptotic coverage and performs well with respect to LSE and Lasso. Indeed, its performances regarding coverage and volume are similar or even better than Lasso for multivariate response. As expected, the confidence level increases with the ratio N/D. Note that multivariate version of the Lasso is not implemented to our knowledge in R which makes IR a challenging method. Interestingly, IR does not suppose sparsity in the model but seems to be competitive with the Lasso on sparse design (Cases 1 and 2) regarding to both coverage and volume. Under Case 3, data are generated under the underlying model of IR, with $\boldsymbol{\Sigma}$ diagonal and no sparsity assumption on $\boldsymbol{A}$, so IR performs particularly well. Case 4 is the most difficult for our procedure, as it violates our assumption that $\boldsymbol{\Sigma}$ is diagonal. Performances are rather good compared with the Lasso, but, when it is computable ($N > D$), LSE performs better as no assumption is needed. However, IR has always a smallest volume than Lasso and LSE, so it seems to have correctly detected the shape of the ellipsoid. Remark that when $D > N$, especially for $D = 1000$, inference seems a challenging problem and coverage is far from 95%. Coverage is even poorer for Case 4 as none of the methods seems to compute a reasonable prediction region. Compared to bootstrapped Lasso, IR approach is significantly faster as our method does not rely on resampling. Computation time is reasonable while achieving challenging coverage and volume when D is large. Whatever the design, note that the volume of prediction regions increases with L, meaning the underlying space dimension. It is interesting to notice that, by normalising the volume by the dimension, the volume stays almost constant across the situations studied.

Figure 2 displays a graphical representation of prediction regions for Case 1 which are ellipses when $L = 2$. We consider two sample sizes, $N = 50$ and $N = 500$. Dotted line represents ellipses computed by LSE when $N = 500$ and Lasso when $N = 50$, long dashed line represents ellipses computed by IR and solid line represents true prediction regions computed with true parameters used for simulation. Grey dots are 500 replications of responses from the same covariate's profile representing the residual variance. Three specific profiles of covariates are considered: on the left panel, prediction ellipse for the median covariate's profile is computed, which is an easy situation. When $N = 500$, both LSE and IR provide similar ellipses, close to the true one. When $N = 50$, IR's ellipse is close to the true one while Lasso correctly predicts the response but the volume of the ellipse is larger. For the middle panel, a covariate's profile corresponding to quantile 0.35 is generated, making the computation of the prediction ellipse more complex. When sample size is large, LSE and IR are competitive regarding to true ellipse and equivalent. When $N = 50$, the ellipse computed with IR is larger than the theoretical one. The bootstrapped Lasso fails in prediction, which confirms the lower coverages observed in Figure 1. At last, for the right panel, an even more extreme profile associated to quantile 0.2 is generated, making the computation less reliable. When $N = 500$, the volume of ellipses computed by LSE and IR gets even larger as the covariate's profile gets far from the mean. Notice that LSE and IR again achieve similar ellipses in this setting. When $N = 50$, conclusions of the middle panel apply as well.

To conclude, we would advice to use IR when no variable selection is expected, whereas considering a high-dimensional dataset. Moreover, IR may consider correlated responses, which is not the case for usual estimators in high-dimension. Remark also that assumptions are weakest when L is larger, because it describes the number of factors in the decomposition of the covariance matrix. So we would also advice to use IR if the response is multivariate. About the robustness to diagonal residual covariance matrix, we think that, even if it would be very interesting to derive it, a theoretical result is out of the scope of this paper. The robustness is tested in practice, and we denote that

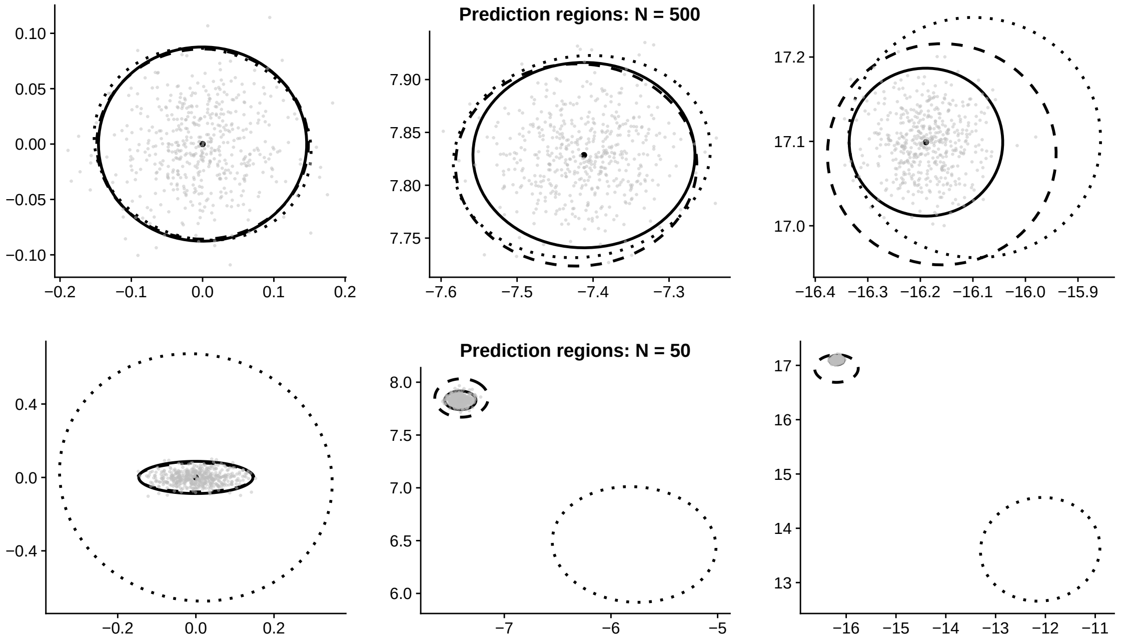

Figure 2: Prediction regions for $L = 2$. Dotted line: LSE for $N = 500$ and Bootstrapped Lasso for $N = 50$, long dashed line: IR, solid line: true parameters, grey dots: 500 responses generated from the same covariate's profile. On the left panel, median covariate's profile are considered. In the middle panel, a covariate's profile corresponding to quantile 0.35 is generated. On the right panel, a profile associated to quantile 0.2 is generated. Thus, more on the right, more difficult it is.

for reasonable dimensions (not the largest one), IR performs good. When the number of covariates is too large, and the assumption is violated, performance is not good as one may expect.

5.3. Study of estimation accuracy

In this section, we focus on the first setting (Case 1) with $L = 2$, $D = 5$ and $N = 100$ in order to visualise the ability of inverse regression to estimate parameters $(\boldsymbol{A}^\star, \boldsymbol{\Gamma}^\star, \boldsymbol{\Sigma}^\star)$ and to predict the response. Violin plots of Figures 3 to 5 display the distribution of the estimators in black and the true value of the parameter in red. Regarding the estimation of the $D \times D$ matrix $\boldsymbol{\Gamma}^\star$, Figure 3 demonstrates that IR is able to retrieve the diagonal structure of the true matrix. Note that the estimation is more variable for diagonal terms. Same remarks hold for the estimation of the $L \times L$ matrix $\boldsymbol{\Sigma}^\star$, see Figure 4. Regarding estimation of $\boldsymbol{A}^\star$, it is interesting to notice that IR partially retrieves the sparse structure of the true parameter. Indeed, all values in $\boldsymbol{A}^\star$ are zero except the 4th coefficient of the first row, and the 3rd value of the second row in Figure 5. The corresponding violin plots are centred around the true value.

Figure 6 displays the distribution of absolute prediction error $|\widehat{\mathbf{Y}} - \mathbf{Y}|$. Note that IR achieves interesting prediction accuracy as most of prediction errors are close to 0. Prediction error of the second response seems easier to predict than the first component which is not surprising as the residual variance in matrix $\boldsymbol{\Sigma}^\star$ for the 2nd response is smaller than residual variance of first response component.

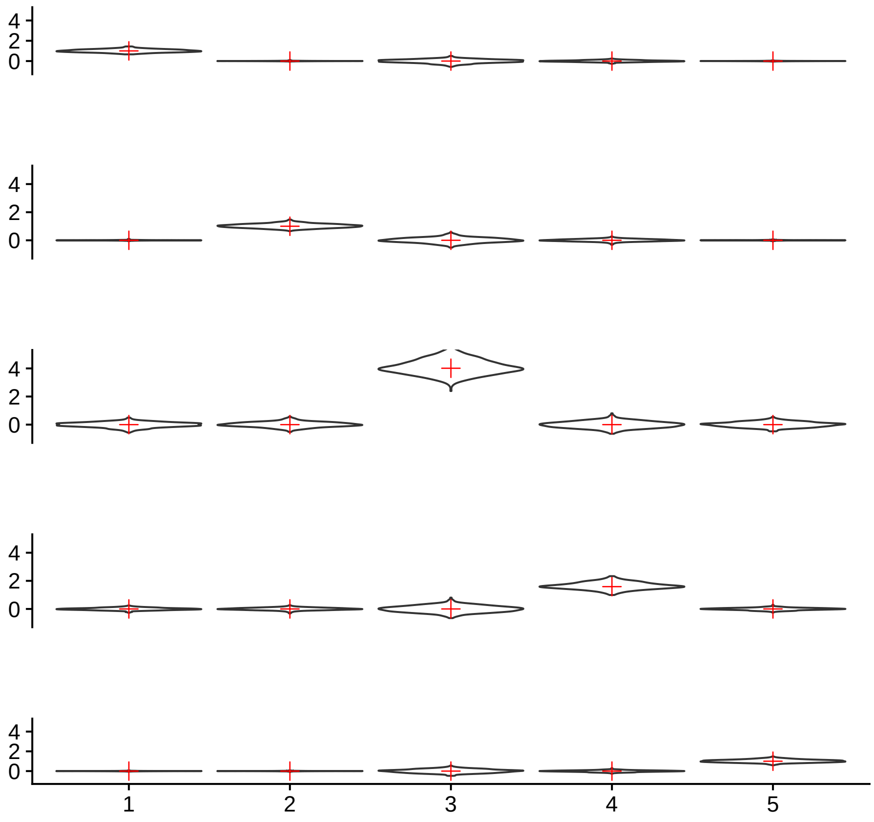

Figure 3: Violin plots displaying the distribution of $\mathbf{\Gamma}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\mathbf{\Gamma}^\star$ is diagonal, true values are located by red crosses.

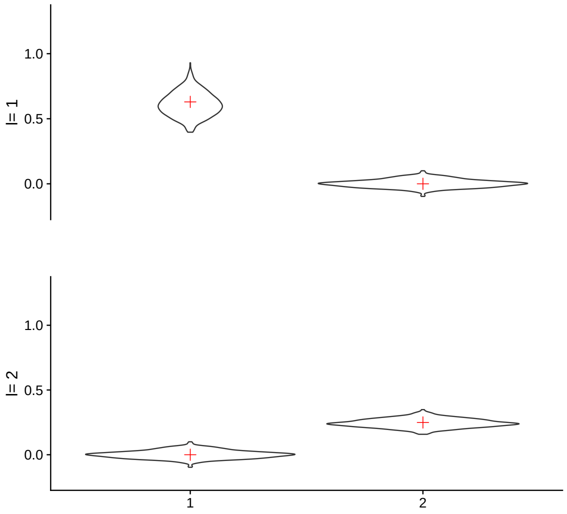

Figure 4: Violin plots displaying the distribution of $\mathbf{\Sigma}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\mathbf{\Sigma}^\star$ is diagonal, true values are located by red crosses.

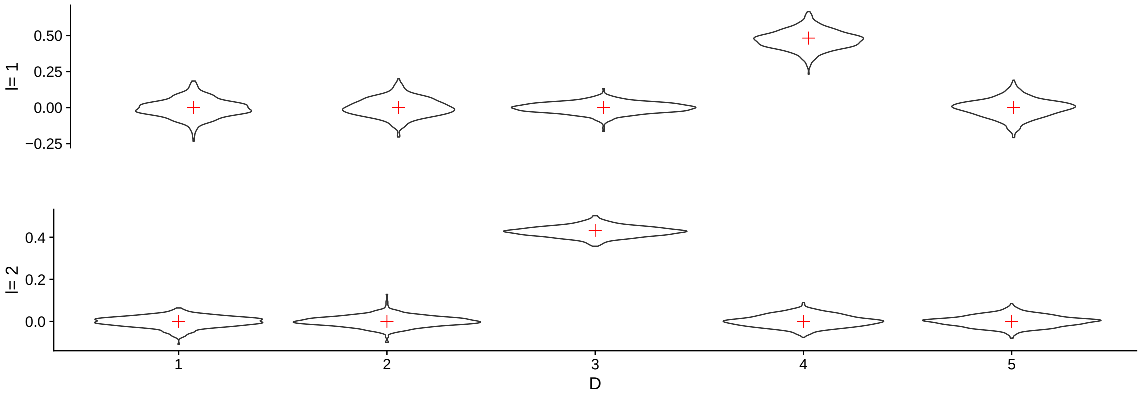

Figure 5: Violin plots displaying the distribution of $\boldsymbol{A}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\boldsymbol{A}^\star$ is sparse, with 2 non zero entries, true values are located by red crosses.

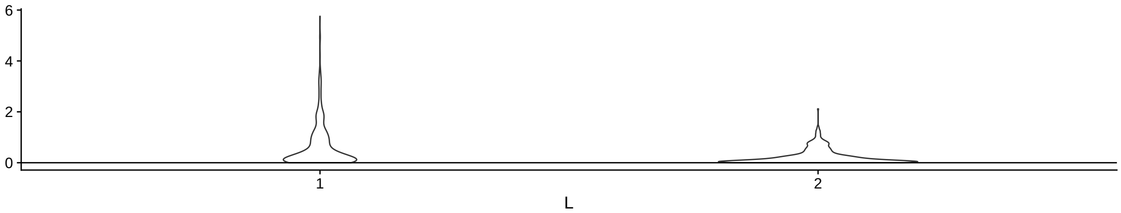

Figure 6: Violin plots displaying the distribution of the absolute prediction error $|\widehat{\mathbf{Y}} - \mathbf{Y}|$ for $L = 2, D = 5$ and Case 1.

6. Conclusion and further discussion

In this article, the properties of inverse regression are extensively investigated under the general framework of elliptical distributions. Inverse regression addresses linear regression issues with random multivariate predictors and multiple responses. The characteristic of this model is that it inverts the role of covariates and response. By making weak assumptions on the residual covariance matrix of the inverse regression, this model allows to consider settings with both large sample size and covariates dimension, as an alternative to least square methods or regularized methods. Explicit estimators of model parameters are derived, for which asymptotic distributions and confidence regions are deduced. Last but not least, asymptotic prediction regions are derived, allowing to quantify the confidence in prediction.

In an intensive simulation study, we present inverse regression as an alternative to variable selection when the sample size is small regarding to the dimension of covariates. Indeed, inverse regression achieves interesting coverage for reasonable time computation. Although our results are asymptotic, performances are challenging for finite sample and illustrates how this model can be used in practice.

A future work could be the extension of this model to generalized linear model by considering other distributions of the noise of the inverse model.

7. Acknowledgments

The authors acknowledge the Editor and the reviewers for their interesting comments which improved the manuscript, both on the theory and the numerical section. E. Devijver also sincerely acknowledges R. Molinier for fruitful mathematical discussions and for its eternal rigor, which have been necessary to define things properly.

Appendix A: details for the proofs

Relation between forward and inverse regression

The joint distribution defined in (1) leads to the marginal and the conditional distributions of Equations (2)-(5) and to a mapping between their mean and variance parameters. Indeed, using conditioning properties of elliptical distributions, we get the following marginal for $\mathbf{X}$ and conditional for $\mathbf{Y}$ distributions

$$\begin{aligned}\mathbf{X} &\sim \mathcal{E}_D(\mathbf{0}, \mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top, \phi);\\ \mathbf{Y}|\mathbf{X} &\sim \mathcal{E}_L(\mathbf{\Gamma}^\top\boldsymbol{A}^\top(\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top)^{-1}\mathbf{X}, \mathbf{\Gamma} - \mathbf{\Gamma}^\top\boldsymbol{A}^\top(\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\mathbf{\Gamma}, \tilde{\phi}).\end{aligned}$$

where $\tilde{\phi}$ is defined in Cambanis et al. (1981) and not detailed here. For more details, we refer to this key paper, as the explicit expression of the characteristic function is not used in this article.

We therefore define

$$\begin{aligned}\mathbf{\Gamma}^\star &= \mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top;\\ \mathbf{\Sigma}^\star &= \mathbf{\Gamma} - \mathbf{\Gamma}^\top\boldsymbol{A}^\top(\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}\mathbf{\Gamma} = (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1};\end{aligned}$$

using Woodbury matrix identity. Lastly we define $\boldsymbol{A}^\star$ which gives the relations between forward and inverse regression

$$\begin{aligned}\boldsymbol{A}^\star &= \mathbf{\Gamma}^\top\boldsymbol{A}^\top(\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top)^{-1}\\ &= \mathbf{\Gamma}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1} - \mathbf{\Gamma}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\\ &= \left[\mathbf{\Gamma}(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}) - \mathbf{\Gamma}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}\right](\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\\ &= (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1};\end{aligned}$$

using Woodbury identity matrix again and the symmetric property of $\mathbf{\Gamma}$.

Proof of Proposition 9

Regression coefficients $\boldsymbol{A}$ are estimated using the so-called least square estimator in the inverse regression model

$$\widehat{\boldsymbol{A}}^\top = (\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}.$$

Conditionally to random vectors $(\mathbf{Y}_i)_{i=1}^N$, $\widehat{\boldsymbol{A}}$ is a linear combination of the elliptically distributed random vectors $(\mathbf{X}_i)_{i=1}^N$. The conditional distribution of $\widehat{\boldsymbol{A}}$ hence follows:

$$\widehat{\boldsymbol{A}}^\top|\mathbf{Y}_1, \dots, \mathbf{Y}_N \sim \mathcal{ME}(\boldsymbol{A}, \frac{-1}{2\phi'(0)}(\mathbf{Y}^T\mathbf{Y})^{-1} \otimes \mathbf{\Sigma}, \phi).$$

The distribution is stable because we consider an affine transformation of matrix elliptical variables from family defined by ϕ. The conditional expectation and conditional variance are deduced directly from the formula.

Proof of Proposition 10

$$\widehat{\boldsymbol{A}}^T = \boldsymbol{A}^T + \left(\frac{\mathbf{Y}^T\mathbf{Y}}{N}\right)^{-1}\frac{\mathbf{Y}^T\mathbf{e}}{N}.$$

As $\mathbf{Y}$ is elliptic, $\frac{\mathbf{Y}^T\mathbf{Y}}{N}$ converges to $-2\phi'(0)\mathbf{\Gamma}$, we get that

$$\widehat{\boldsymbol{A}}^T \underset{N\to+\infty}{\to} \boldsymbol{A}^T + (-2\phi'(0)\mathbf{\Gamma})^{-1}\lim_{N\to+\infty}\frac{\mathbf{Y}^T\mathbf{e}}{N}.$$

If we denote $\mathbf{W}_i = \mathbf{Y}_i^T \mathbf{e}_i$ a matrix with L rows and D columns, we get that $\mathrm{E}(\mathbf{W}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{W}) = 4\phi'(0)\phi'_{\mathbf{e}}(0)\mathbf{\Gamma} \otimes \mathbf{\Sigma}$, by the law of total variance. Using the central limit theorem, we get that

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_i\right) = \frac{\mathbf{Y}^T\mathbf{e}}{\sqrt{N}} \underset{n\to+\infty}{\to} \mathcal{MN}_{L,D}(\mathbf{0}, 4\phi'(0)\phi'_{\mathbf{e}}(0)\mathbf{\Gamma} \otimes \mathbf{\Sigma}).$$

Thus,

$$(-2\phi'(0)\mathbf{\Gamma})^{-1}\frac{\mathbf{Y}^T\mathbf{e}}{\sqrt{n}} \underset{n\to+\infty}{\to} \mathcal{MN}_{L,D}(\mathbf{0}, \frac{\phi'_{\mathbf{e}}(0)}{\phi'(0)}\mathbf{\Gamma}^{-1} \otimes \mathbf{\Sigma}).$$

Finally,

$$\sqrt{N}(\widehat{\boldsymbol{A}} - \boldsymbol{A}) \underset{n\to+\infty}{\to} \mathcal{MN}_{D,L}(\mathbf{0}, \frac{\phi'_{\mathbf{e}}(0)}{\phi'(0)}\mathbf{\Sigma} \otimes \mathbf{\Gamma}^{-1}).$$

Proof of Lemma 11

We use the following lemma.

Lemma 15 *If* $\|\boldsymbol{A}\| \leq 1$, *then* $(\mathbb{I} - \boldsymbol{A})^{-1} = \mathbb{I} + \boldsymbol{A} + \boldsymbol{A}^2 + o(||\boldsymbol{A}||^2)$.

Then, we prove Lemma 11:

$$\begin{aligned}g(\boldsymbol{A} + h\boldsymbol{M}) - g(\boldsymbol{A}) &= h(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{M}^\top\mathbf{\Sigma}^{-1}\\ &- h(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}(\boldsymbol{M}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{M})(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}\mathbf{\Sigma}^{-1} + O(h^2);\\ Dg(\boldsymbol{A}).\boldsymbol{M} &= (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{M}^\top\mathbf{\Sigma}^{-1}\\ &- (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}(\boldsymbol{M}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{M})(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}\mathbf{\Sigma}^{-1}.\end{aligned}$$

Next, remember that $\boldsymbol{M} = (\widehat{\boldsymbol{A}} - \boldsymbol{A})$, we have:

$$\begin{aligned}Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}) &= (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}[\widehat{\boldsymbol{A}} - \boldsymbol{A}]^\top\mathbf{\Sigma}^{-1} - (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\\ &\times \left([\widehat{\boldsymbol{A}} - \boldsymbol{A}]^\top\mathbf{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}[\widehat{\boldsymbol{A}} - \boldsymbol{A}]\right)(\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}\mathbf{\Sigma}^{-1}.\end{aligned} \tag{18}$$

Then, we compute the covariance. We decompose it as the following.

$$\begin{aligned}Cov(\mathrm{vec}(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}))) =& var(\mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1})) + var(\mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star))\\ &+ var(\mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{A}^\star))\\ &- 2cov(\mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1}), \mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star))\\ &- 2cov(\mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1}), \mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{A}^\star)).\\ &- 2cov(\mathrm{vec}(\mathbf{\Sigma}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})^\top\mathbf{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star), \mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{A}^\star)).\end{aligned}$$

Then, we want to compute each term explicitly.

$$\begin{aligned}
var(\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1})) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star);\\
var(\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star)) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}((\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star);\\
var(\mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})\boldsymbol{A}^\star)) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}((\boldsymbol{A}^\star)^\top\boldsymbol{\Gamma}\boldsymbol{A}^\star\otimes\boldsymbol{A}^\star\boldsymbol{\Sigma}(\boldsymbol{A}^\star)^\top);\\
cov(\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1}),\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star)) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star);\\
cov(\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1}),\mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})\boldsymbol{A}^\star)) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}(\mathbf{I}\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})T_{LD}^{-1}((\boldsymbol{A}^\star)^\top\otimes\boldsymbol{A}^\star);\\
cov(\mathrm{vec}(\boldsymbol{\Sigma}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star),\mathrm{vec}(\boldsymbol{A}^\star(\widehat{\boldsymbol{A}}-\boldsymbol{A})\boldsymbol{A}^\star)) &= \frac{\phi'_{\mathsf{e}}(0)}{\phi'(0)}((\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})T_{LD}^{-1}((\boldsymbol{A}^\star)^\top\otimes\boldsymbol{A}^\star).
\end{aligned}$$

Putting everything together, we get the following:

$$\begin{aligned}
\frac{\phi'(0)}{\phi'_{\mathsf{e}}(0)}Cov(\mathrm{vec}(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A}))) =&(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star)+((\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star)\\
&+((\boldsymbol{A}^\star)^\top\boldsymbol{\Gamma}\boldsymbol{A}^\star\otimes\boldsymbol{A}^\star\boldsymbol{\Sigma}(\boldsymbol{A}^\star)^\top)-2(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star)\\
&-2(\mathbf{I}\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})T_{LD}^{-1}((\boldsymbol{A}^\star)^\top\otimes\boldsymbol{A}^\star)\\
&-2((\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})T_{LD}^{-1}((\boldsymbol{A}^\star)^\top\otimes\boldsymbol{A}^\star)\\
=&\big((\boldsymbol{\Sigma}^{-1}+(\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star-2\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star)\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star\big)\\
&+((\boldsymbol{A}^\star)^\top\boldsymbol{\Gamma}\boldsymbol{A}^\star\otimes\boldsymbol{A}^\star\boldsymbol{\Sigma}(\boldsymbol{A}^\star)^\top)\\
&-2\left((\mathbf{I}\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})+((\boldsymbol{A}^\star)^\top\boldsymbol{A}^\top\otimes\boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})\right)T_{LD}^{-1}((\boldsymbol{A}^\star)^\top\otimes\boldsymbol{A}^\star).
\end{aligned}$$

Appendix B: results of simulation study

The table hereafter details numerical values displayed in Figure 1.

References

D. Babichev and F. Bach. Slice inverse regression with score functions. *Electron. J. Statist.*, 12(1): 1507–1543, 2018.

Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11:368, 2010.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-20192-9.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368 – 385, 1981. ISSN 0047-259X.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

C.M. Carvalho, Chang J., J.E. Lucas, J.R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association: Applications and Case Studies*, 103: 484, 2008.

			N = 50				N = 100				N = 500				N = 500, D = 1000			
			Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
	IR	Coverage	**0.89**	**0.89**	**0.89**	**0.87**	**0.94**	**0.92**	**0.92**	**0.92**	0.94	0.95	0.95	0.93	0.90	0.89	**0.91**	0.82
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	**0.95**	**0.95**	**0.94**	**0.95**	NA	NA	NA	NA
	Lasso		0.87	0.86	0.84	0.79	0.90	0.90	0.86	0.85	0.93	0.95	0.90	0.92	**0.91**	**0.92**	0.85	0.82
L = 1	IR	Volume	1.21	1.22	**1.23**	**1.25**	1.23	1.24	**1.25**	1.27	1.24	1.24	1.26	1.28	1.21	1.21	**1.21**	**1.22**
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	1.39	1.40	1.41	1.39	NA	NA	NA	NA
	Lasso		**1.14**	**1.14**	1.42	1.40	**1.17**	**1.17**	1.27	**1.26**	**1.19**	**1.19**	**1.21**	**1.19**	**1.18**	**1.18**	1.29	1.27
	IR	CPU	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.90	1.85	1.90	1.83
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	0.02	0.02	0.02	0.01	NA	NA	NA	NA
	Lasso		0.28	0.28	0.35	0.31	0.33	0.33	0.41	0.35	0.48	0.48	0.56	0.48	1.41	1.39	2.61	2.65
	IR	Coverage	0.83	0.83	**0.83**	**0.84**	0.89	0.90	**0.92**	**0.90**	**0.94**	0.94	**0.94**	**0.94**	0.88	0.88	**0.88**	**0.78**
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	**0.94**	**0.95**	**0.94**	**0.94**	NA	NA	NA	NA
	Lasso		**0.86**	0.83	0.75	0.76	**0.90**	0.90	0.79	0.80	0.91	0.93	0.92	0.92	**0.90**	**0.92**	0.73	0.72
L = 2	IR	Volume	**1.50**	**1.49**	**1.37**	**1.38**	**1.56**	**1.54**	**1.39**	**1.41**	**1.58**	**1.56**	**1.40**	1.42	**1.35**	**1.35**	**1.33**	**1.34**
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	1.77	1.76	1.57	1.54	NA	NA	NA	NA
	Lasso		1.74	1.74	1.65	1.62	1.77	1.76	1.48	1.46	1.77	1.76	**1.40**	**1.38**	1.37	1.37	1.47	1.44
	IR	CPU	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.03	3.07	3.00	2.95	2.89
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	0.02	0.02	0.02	0.01	NA	NA	NA	NA
	Lasso		0.53	0.52	0.61	0.52	0.61	0.61	0.71	0.59	0.85	0.84	0.95	0.82	2.02	1.99	3.25	3.31
	IR	Coverage	0.66	0.66	**0.75**	**0.71**	0.82	0.82	**0.88**	**0.87**	0.92	0.94	0.94	0.94	0.81	0.81	**0.81**	**0.60**
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	**0.94**	**0.95**	**0.94**	**0.94**	NA	NA	NA	NA
	Lasso		**0.76**	**0.75**	0.60	0.60	**0.83**	**0.84**	0.70	0.66	0.89	0.88	0.87	0.88	**0.87**	**0.86**	0.57	0.55
L = 5	IR	Volume	**1.93**	**1.89**	**1.49**	**1.50**	**2.09**	**2.04**	**1.51**	**1.52**	**2.20**	**2.15**	1.52	1.53	**1.46**	**1.44**	**1.42**	**1.42**
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	2.48	2.43	1.70	1.65	NA	NA	NA	NA
	Lasso		2.99	2.92	1.82	1.80	3.01	2.93	1.61	1.58	3.02	2.96	**1.51**	**1.48**	1.51	1.50	1.56	1.52
	IR	CPU	0.28	0.27	0.28	0.25	0.28	0.29	0.28	0.25	0.29	0.29	0.29	0.28	5.32	5.37	5.34	5.31
	LSE		NA	NA	NA	NA	NA	NA	NA	NA	0.02	0.02	0.02	0.02	NA	NA	NA	NA
	Lasso		1.00	0.97	1.13	1.23	1.20	1.21	1.29	1.33	1.43	1.43	1.62	1.43	2.78	2.83	4.07	4.19

Table 1: Results of simulation study displayed in Figure 1 (best results for coverage and volume are in bold).

D. Causeur, C. Friguet, M. Houée, and M. Kloareg. Factor analysis for multiple testing (FAMT): an R package for large-scale significance testing under dependence. *Journal of Statistical Software*, 40(14):1–19, 2011.

S.K. Chao, Y. Ning, and H. Liu. On high dimensional post-regularization prediction intervals. Technical report, arXiv, 2015.

D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

A. Deleforge, F. Forbes, and R. Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015.

S. Ding and R. D. Cook. Tensor sliced inverse regression. *Journal of Multivariate Analysis*, 133:216 – 231, 2015. ISSN 0047-259X.

M. El Behi, C. Sanson, C. Bachelin, L. Guillot-Noël, J. Fransson, B. Stankoff, E. Maillart, N. Sarrazin, V. Guillemot, H. Abdi, I. Cournu-Rebeix, B. Fontaine, and V. Zujovic. Adaptive human immunity drives remyelination in a mouse model of demyelination. *Brain*, 4(170):967–980, 2017.

G. Frahm. *Generalized Elliptical Distributions: Theory and Applications.* PhD thesis, Universitt zu Kln, 1 2004.

E. I. George and S.D. Oman. Multiple-shrinkage principal component regression. *The Statistician*, 45:111–124, 1996.

A. K. Gupta and T. Varga. A new class of matrix variate elliptically contoured distributions. *Journal of the Italian Statistical Society*, 3(2):255–270, Jun 1994. ISSN 1613-981X.

A.K. Gupta and D.K. Nagar. *Matrix variate distributions.* Chapman & HALL/CRC, 2000.

I.S. Helland. Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society, Series B*, 54:637–347, 1992.

I.S. Helland and T. Almøy. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89:583–591, 1994.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. ISSN 00401706.

T. Hsing and R. J. Carroll. An asymptotic theory for sliced inverse regression. *Ann. Statist.*, 20(2): 1040–1061, 06 1992.

H. Hult and F. Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34:587–608, 09 2002.

W. G. Hunter and W. F. Lamboy. A bayesian analysis of the linear calibration problem. *Technometrics*, 23(4):323–328, 1981.

J. Janková and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Statist.*, 9(1):1205–1229, 2015.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, January 2014. ISSN 1532-4435.

N. Kannan, J. P. Keating, and R. L. Mason. A comparison of classical and inverse estimators in the calibration problem. *Communications in Statistics - Theory and Methods*, 36(1):83–95, 2007.

R. G. Krutchkoff. Classical and inverse regression methods of calibration. *Technometrics*, 9(3): 425–439, 1967.

J.D. Lee, D.L. Sun, Y. Sun, and J.E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016.

J. T. Leek and J. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105:18718–18723, 2008.

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

Q. Lin, Z. Zhao, and J. S. Liu. Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 0(0):1–33, 2019.

N. Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945, 2015. ISSN 1467-9868.

S.D. Oman. Random calibration with many measurements: An application of stein estimation. *Technometrics*, 33:187–195, 1991.

P. A. Parker, G. G. Vining, S. R. Wilson, J. L. Szarka III, and N. G. Johnson. The prediction properties of classical and inverse regression for the simple linear calibration problem. *Journal of Quality Technology*, 42(4):332–347, 2010.

E. Perthame, C. Friguet, and D. Causeur. Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, pages 1–14, 2015.

E. Perthame, F. Forbes, and A. Deleforge. Inverse regression approach to robust nonlinear high-to-low dimensional mapping. *Journal of Multivariate Analysis*, 163:1–14, 2018.

J. Saracco. An asymptotic theory for sliced inverse regression. *Communications in Statistics - Theory and Methods*, 26(9):2141–2171, 1997.

B. Stucky and S. van de Geer. Asymptotic confidence regions for highdimensional structured sparsity. Technical report, arXiv, 2017.

W. Sun and T.-T. Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society, Series B*, 71(2):1–32, 2009.

A.E. Teschendorff, J. Zhuang, and M. Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27:11: 1496–1505, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli, forthcoming paper*, 2017.

E. J. Williams. A note on regression methods in calibration. *Technometrics*, 11(1):189–192, 1969.

C.-H. Zhang and S.S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 217–242, 2014. ISSN 1467-9868.