Prediction regions through Inverse Regression

Emilie Devijver, Emeline Perthame

▶ **To cite this version:**

Emilie Devijver, Emeline Perthame. Prediction regions through Inverse Regression. 2019. hal-01840234v2

HAL Id: hal-01840234

https://hal.science/hal-01840234v2

Preprint submitted on 27 Jun 2019 (v2), last revised 11 Dec 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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: Francis Bach, David Blei, and Bernhard Schölkopf

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 normality

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. 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)). See 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. Note that this inversion regression approach has been studied to estimate mixtures of regression models and applied to various data (planetology and spectra (Deleforge et al., 2015; 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 $\mathbf{X} \sim \mathcal{E}_d(\mu, \Sigma, \phi)$ *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:*

- $E(\mathbf{X}) = \mu$*,*
- $Var(\mathbf{X}) = \frac{E(\mathcal{R}^2)}{k}\Sigma = -2\phi'(0)$ *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 and can be represented stochastically by*

$$\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 = \mu^* + \mathcal{R}^* C_{22}\mathbf{U}^{(r-k)}$$

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

$$\mathcal{R}^* = \mathcal{R}\sqrt{1-\beta}|(\mathcal{R}\sqrt{\beta}\mathbf{U}^{(k)} = C_{11}^{-1}(\mathbf{x}_1 - \mu_1))$$
$$\mu^* = \mu_2 + C_{21}C_{11}^{-1}(\mathbf{x}_1 - \mu_1)$$

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.*

- *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

We propose to address the following linear regression problem with random regressors also known as generative model:

$$\mathbf{X}_i \sim \mathcal{E}_D(\mathbf{0}, \mathbf{\Gamma}^\star, \phi) \text{ with } \text{rank}(\mathbf{\Gamma}^\star) = D \tag{1}$$
$$\mathbf{Y}_i|\mathbf{X}_i = \boldsymbol{A}^\star\mathbf{X}_i + \boldsymbol{\varepsilon}_i \tag{2}$$

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, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \sim \mathcal{E}_L(\mathbf{0}, \mathbf{\Sigma}^\star, \phi)$ with $\text{rank}(\mathbf{\Sigma}^\star) = L$. 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 (1) and (2). 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:

$$\mathbf{Y}_i \sim \mathcal{E}_L(\mathbf{0}, \mathbf{\Gamma}, \phi) \text{ with } \text{rank}(\mathbf{\Gamma}) = L \tag{3}$$
$$\mathbf{X}_i|\mathbf{Y}_i = \boldsymbol{A}\mathbf{Y}_i + \mathbf{e}_i \tag{4}$$

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 $\mathbf{\Sigma}$. 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 (3) and (4), which leads to assume that both response and covariates are centered.

Interestingly, 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 mapping:

$$\begin{aligned}\Psi : (\mathbf{\Gamma}, \boldsymbol{A}, \mathbf{\Sigma}) &\mapsto (\mathbf{\Gamma}^\star, \boldsymbol{A}^\star, \mathbf{\Sigma}^\star) \\ &= (\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}).\end{aligned} \tag{5}$$

Details to prove this one-to-one mapping are given in Appendix, relying on the conditional distribution introduced previously. As the mapping Ψ is an involution, the forward regression model (1)-(2) is equivalent to the inverse regression model (3)-(4), hence the generative property of the model. The advantage of the inverse approach appears when structure is assumed on the large residual covariance matrix $\mathbf{\Sigma}$ in the inverse regression problem. Indeed, 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 $LD+L(L+1)/2+D = 615$ assuming that $\mathbf{\Sigma}$ is diagonal), while keeping a general modelling: it implies a diagonal + low rank decomposition for $\mathbf{\Gamma}^\star$ through the mapping Ψ. So, the residuals of the inverse model are assumed to be not correlated while structured correlations are allowed among covariates in the forward model. This is a strength of this model as in practice, correlated predictors often occur on real data.

2.3 Estimation

Considering the inverse model defined in Equations (3)-(4), the least squares estimators are:

$$\begin{aligned} \widehat{\mathbf{\Gamma}} &= \frac{1}{N-1}\mathbf{Y}^\top\mathbf{Y}; \\ \widehat{\boldsymbol{A}} &= (\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}\top\mathbf{X}; \\ \forall j \in \{1,\dots,D\}, \widehat{\mathbf{\Sigma}}_{j,j} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_{i,j} - [\widehat{\boldsymbol{A}}\mathbf{Y}_i]_j)^2. \end{aligned} \tag{6}$$

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

$$\widehat{\mathbf{\Gamma}}^\star = \widehat{\mathbf{\Sigma}} + \widehat{\boldsymbol{A}}\widehat{\mathbf{\Gamma}}\widehat{\boldsymbol{A}}^\top; \tag{7}$$

$$\widehat{\boldsymbol{A}}^\star = (\widehat{\mathbf{\Gamma}}^{-1} + \widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^\top\widehat{\mathbf{\Sigma}}^{-1}; \tag{8}$$

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

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 (2) 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 normality 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 $A \in \mathcal{M}_{r,D}(\mathbb{R})$ and $B \in \mathcal{M}_{L,s}(\mathbb{R})$*

$$A\mathbf{X}B \sim \mathcal{ME}_{rs}(A\mathbf{X}_0B, AUA^T \otimes B^TVB, \phi);$$
$$vec(A\mathbf{X}B) = (B^T \otimes A)vec(\mathbf{X}).$$

For $A \in M_{r,D}(\mathbb{R}), B \in M_{L,s}(\mathbb{R}), C \in M_{r,D}(\mathbb{R}), D \in M_{L,s}(\mathbb{R})$ 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 normality of $\widehat{\boldsymbol{A}}^\star$

In order to derive the asymptotic normality 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.

Proposition 9 (Distribution of $\widehat{\boldsymbol{A}}$) *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* (1) *and* (2)*, or equivalently in Equations* (3) *and* (4)*. Then,*

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

This result is an extension of the least square estimator in the multivariate linear model to the multiple multivariate linear model. The proof is straightforward using that $\hat{\boldsymbol{A}}$ is a linear combination of $\mathbf{X}$.

From this, we derive the asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$. A matricial 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 10 *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* (1) *and* (2)*. Let*

$$\begin{aligned} g : \mathbb{R}^{D \times L} &\to \mathbb{R}^{L \times D} \\ \boldsymbol{A} &\mapsto \boldsymbol{A}^\star = \boldsymbol{\Sigma}^\star \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{10}$$

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

$$Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}) = \boldsymbol{\Sigma}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A})^T \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star - \boldsymbol{A}^\star (\widehat{\boldsymbol{A}} - \boldsymbol{A}) \boldsymbol{A}^\star. \tag{11}$$

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

$$\begin{aligned} \frac{-1}{2\phi'(0)} Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}))) = & \\ & \left((\boldsymbol{\Sigma}^{-1} + (\boldsymbol{A}^\star)^T \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star - 2\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star) \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma} \boldsymbol{\Sigma}^\star\right) \\ & + ((\boldsymbol{A}^\star)^T \boldsymbol{\Gamma} \boldsymbol{A}^\star \otimes \boldsymbol{A}^\star \boldsymbol{\Sigma} (\boldsymbol{A}^\star)^T) \\ & - 2\left((\mathbf{I} \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma}) + ((\boldsymbol{A}^\star)^T \boldsymbol{A}^T \otimes \boldsymbol{\Sigma}^\star \boldsymbol{\Gamma})\right) T_{LD}^{-1} ((\boldsymbol{A}^\star)^T \otimes \boldsymbol{A}^\star). \end{aligned} \tag{12}$$

Proof of Lemma 10 is given in Appendix 6.

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

Theorem 11 (Asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$) *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 (1) and (2). Let*

$$\begin{aligned} g : \mathbb{R}^{D\times L} &\to \mathbb{R}^{L\times D} \\ \boldsymbol{A} &\mapsto \boldsymbol{A}^\star = \boldsymbol{\Sigma}^\star \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{13}$$

Then, the following holds for the estimator $\widehat{\boldsymbol{A}}^\star$ defined in Equation (8),

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

where $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$ defined in Equation (12).

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

$$\sqrt{N}(vec(\widehat{\boldsymbol{A}}^\star) - vec(\boldsymbol{A}^\star))^T \Theta(\widehat{\boldsymbol{A}})^{-1}(vec(\widehat{\boldsymbol{A}}^\star) - vec(\boldsymbol{A}^\star)) \underset{N\to+\infty}{\to} F(\sqrt{.}); \tag{14}$$

where F is the distribution function of the random variable $\mathcal{R}$ involved in the stochastic representation of $vec(\boldsymbol{A}^\star)$.

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 (13), 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 10.

Then,

$$\sqrt{N}(\widehat{\boldsymbol{A}}^\star - \boldsymbol{A}^\star) = \sqrt{N}Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}) + \sqrt{N}R_N(\widehat{\boldsymbol{A}}). \tag{15}$$

The last term in (15) converges to 0 in probability, and by Proposition 9, the linear combination with respect to $\hat{\boldsymbol{A}}$ defined in (11) is a matrix variate elliptical distribution, centered. Using (12), we get the distribution of the vectorized vector $vec(\widehat{\boldsymbol{A}}^\star)$.

Asymptotic distribution of the quadratic form Equation (14) 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. As an example, assume that both $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian (setting $\phi = \exp(-u/2)$). Theorem 11 gives that $\widehat{\boldsymbol{A}^\star}$ is asymptotically Gaussian. Moreover, remark that the quadratic form associated to $vec(\widehat{\boldsymbol{A}^\star})$ asymptotically follows a χ^2 distribution with LD degrees of freedom which depends on the size of the response and the covariates in the same way.

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 12 *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 (1) and (2). 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

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

with $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$ defined in Equation (12) and Q the quantile function of $F(\sqrt{.})$.

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

4.2 Prediction regions

Theorem 13 *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 (1) and (2). Then,*

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

where

$$\widetilde{\mathcal{PR}}_{\mathbf{Y},\alpha} = \big\{\mathbf{y} \in \mathbb{R}^L \text{ s.t.} \quad (\mathbf{y} - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1})^T (\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1}) + -2\phi'(0)\boldsymbol{\Sigma}^\star)^{-1} (\mathbf{y} - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1}) \leq Q(1-\alpha)\Big\}; \tag{16}$$

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

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

where $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$ defined in Equation (12).

One can notice that the covariance matrix that is inverted in Equation (16) 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. In the special case of Gaussian setting, the quadratic form in Equation (16) associated to prediction $\mathbf{y}_{n+1}$ asymptotically follows a χ^2 distribution with L degrees of freedom. This closed-form formula is used hereafter in numerical experiments

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 (5). For each combination of dimension, we focus on the 4 following scenarii:

(Case 1) Sparse regression coefficients and independent responses: $\boldsymbol{A}$ is a $D \times L$ matrix with 90% of zero entries randomly drawn. The 10% nonzero remaining coefficients are drawn from a uniform distribution on $(-2, 2)$ for $D = 100$ and $(-0.5, 0.5)$ for $D = 1000$. Matrix $\boldsymbol{\Gamma}$ of covariances between response terms is set to $\mathbb{I}_L$. The residual covariance matrix of inverse regression $\boldsymbol{\Sigma}$ is set to $\mathbb{I}_D$. Note that a diagonal $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Gamma}$ 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 $\boldsymbol{\Gamma}$ is generated as in Case 2. The residual covariance matrix $\boldsymbol{\Sigma}$ is set to $\mathbb{I}_D$.

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

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

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:

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

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 5 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.

Datasets are generated under a linear regression model as defined in Equations (1)-(2). 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 for designs with $N \leq D$. 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, results used to generate this figure are available in Table 1 in Appendix 6. This figure presents the results for varying sample sizes and designs 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 LSE when $N > D$ and to Lasso when $N \leq D$ (in grey).

First, Figure 1 demonstrates that IR performs as well as a variable selection method. Indeed, its performances are similar or even better than Lasso for multivariate response: IR achieves better coverages and smaller volumes in all cases. 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 also efficient 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 LSE performs better as no assumption is needed. Figure 1 illustrates that our results are asymptotic, meaning that performances of IR are good regarding volume and coverage for large N (as the confidence level is almost reached for $N = D$, the asymptotic normality may be, in practice, quickly reached with respect to N). The confidence level increases with N and 0.95 is reached when $N > D$. Compared to bootstrapped Lasso, IR

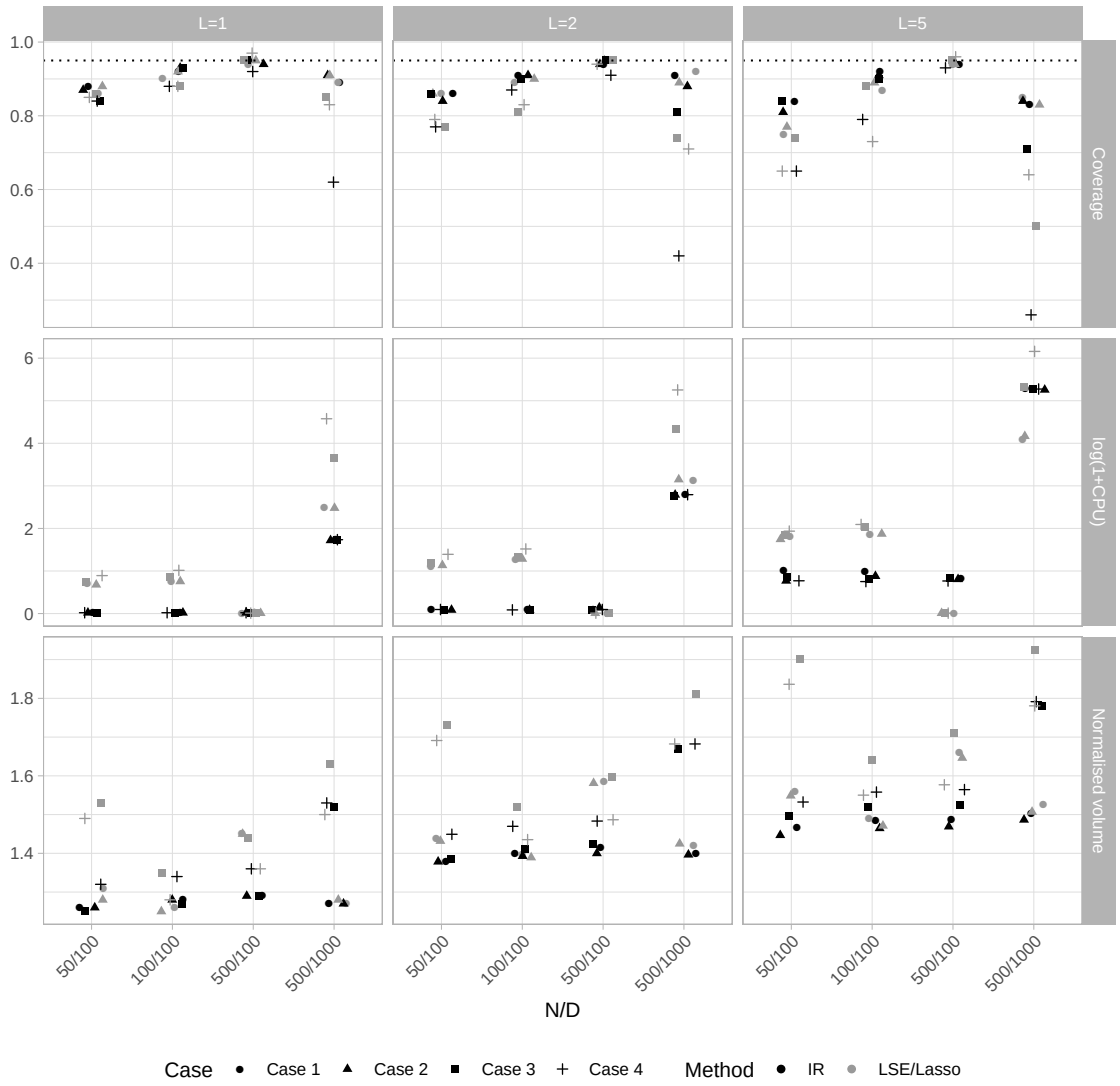

Figure 1: Results of simulations study for Gaussian distribution: prediction regions computed on datasets simulated under models described in Section 5.1. Coverage, normalised volume and CPU time are computed for each method to compare performances. For sample sizes larger than number of covariates, we compare IR (in black) with the LSE (in grey) and for sample sizes smaller than number of covariates we compare IR (in black) with the bootstrapped Lasso (in grey). 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.

approach is significantly faster as our method does not rely on resampling. At last, this table shows that IR works well in high-dimension as large D is computationally feasible. 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 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.

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.

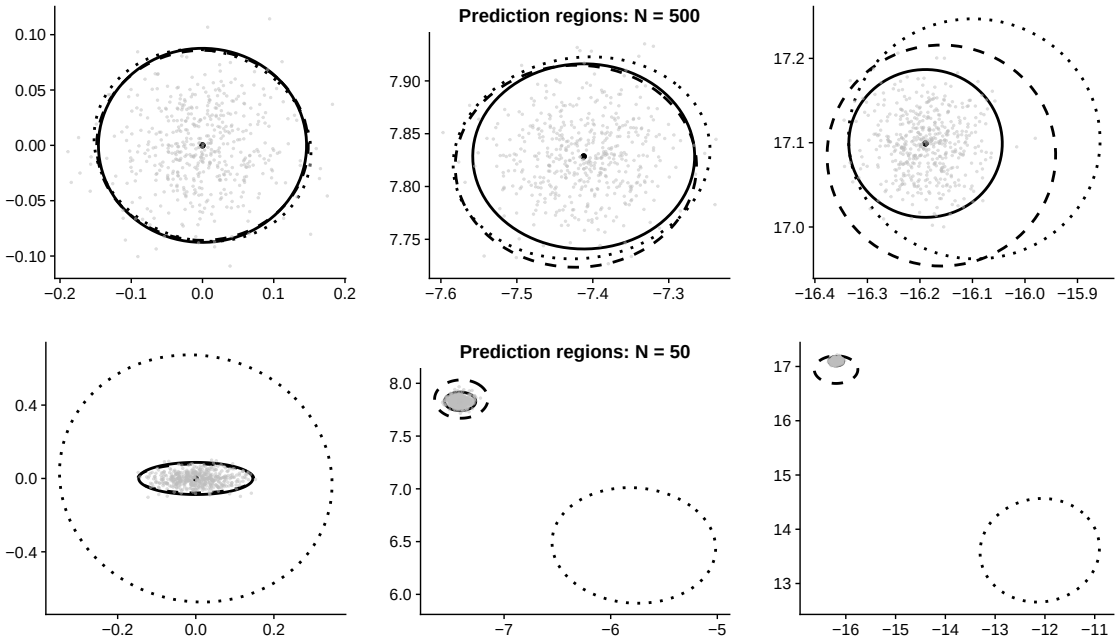

Figure 2: Prediction regions for $L = 2$. Dotted line: LSE for $N = 500$ and Bootstrapped Lasso for $N = 50$, long dashed line: IR, solid line: true parameters, grey dots: 500 responses generated from the same covariate's profile. On the left panel, median covariate's profile are considered. In the middle 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.

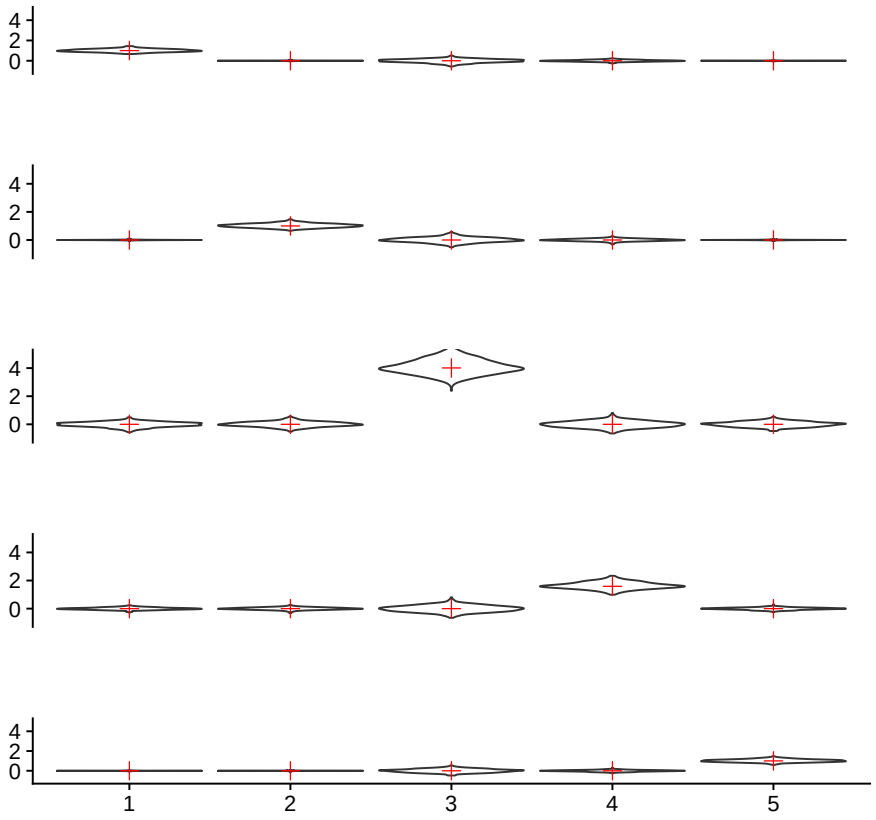

Figure 3: Violin plots displaying the distribution of $\mathbf{\Gamma}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\mathbf{\Gamma}^\star$ is diagonal, true values are located by red crosses.

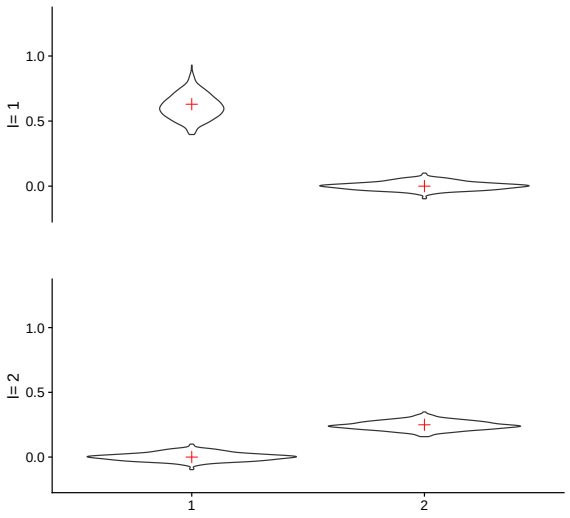

Figure 4: Violin plots displaying the distribution of $\boldsymbol{\Sigma}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\boldsymbol{\Sigma}^\star$ is diagonal, true values are located by red crosses.

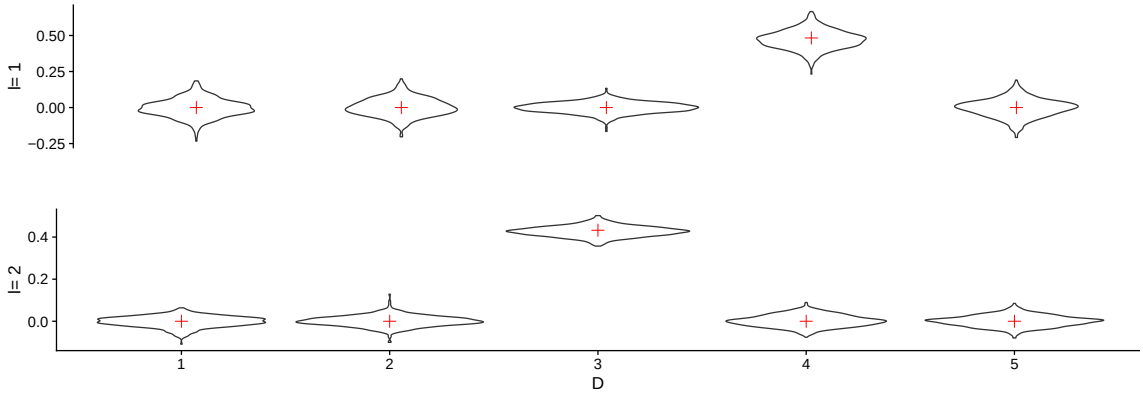

Figure 5: Violin plots displaying the distribution of $\boldsymbol{A}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\boldsymbol{A}^\star$ is sparse, with 2 non zero entries, true values are located by red crosses.

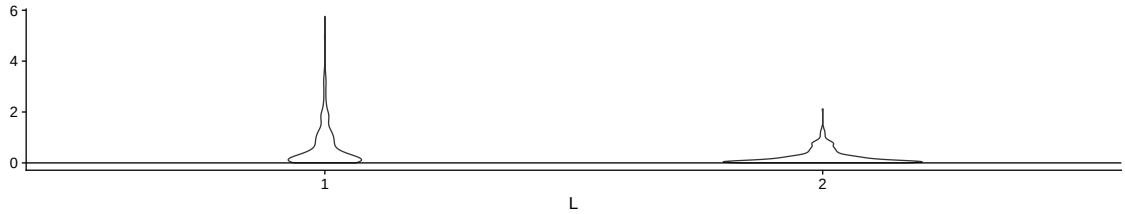

Figure 6: Violin plots displaying the distribution of the absolute prediction error $|\widehat{\mathbf{Y}} - \mathbf{Y}|$ for $L = 2, D = 5$ and Case 1.

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.

Appendix A: details for the proofs

One-to-one mapping

With no loss of generality, the generative model Equations (3)-(4) is equivalent to assuming that $(\mathbf{X}, \mathbf{Y})$ is a random vector with distribution

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

Using properties of conditional distribution of elliptical distributions, we get the following marginal for $\mathbf{X}$ and conditional for $\mathbf{Y}$ distributions

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

We therefore define

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

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

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

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

Proof of Lemma 10

We use the following lemma.

Lemma 14 *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 10:

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

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

$$\begin{aligned}Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}}-\boldsymbol{A}) &= (\boldsymbol{\Gamma}^{-1}+\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}[\widehat{\boldsymbol{A}}-\boldsymbol{A}]^\top\boldsymbol{\Sigma}^{-1}-(\boldsymbol{\Gamma}^{-1}+\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\\ &\times\left([\widehat{\boldsymbol{A}}-\boldsymbol{A}]^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}[\widehat{\boldsymbol{A}}-\boldsymbol{A}]\right)(\boldsymbol{\Gamma}^{-1}+\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}.\end{aligned} \tag{17}$$

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

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

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})) &= -2\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)) &= -2\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)) &= -2\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)) &= -2\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)) &= -2\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)) &= -2\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:

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

Appendix B: results of simulation study

The table hereafter details numerical values displayed in Figure 1.

References

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.

			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
L = 1	IR	Coverage	0.88	0.87	0.84	0.84	0.92	0.93	0.93	0.88	0.95	0.94	0.95	0.92	0.89	0.91	0.85	0.62
	Lasso/LSE		0.86	0.88	0.86	0.85	0.90	0.92	0.88	0.88	0.94	0.95	0.95	0.97	0.89	0.91	0.85	0.83
	IR	Volume	1.26	1.26	1.25	1.32	1.28	1.28	1.27	1.34	1.29	1.29	1.29	1.36	1.27	1.27	1.52	1.53
	Lasso/LSE		1.31	1.28	1.53	1.49	1.26	1.25	1.35	1.28	1.45	1.45	1.44	1.36	1.27	1.28	1.63	1.5
	IR	CPU	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	4.65	4.59	4.6	4.67
	Lasso/LSE		1.01	0.97	1.09	1.44	1.13	1.13	1.35	1.76	0.01	0.01	0.01	0.01	11.19	10.93	38.08	96.11
L = 2	IR	Coverage	0.86	0.84	0.86	0.77	0.91	0.91	0.90	0.87	0.94	0.94	0.95	0.91	0.91	0.88	0.81	0.42
	Lasso/LSE		0.86	0.86	0.77	0.79	0.89	0.90	0.81	0.83	0.95	0.95	0.95	0.94	0.92	0.89	0.74	0.71
	IR	Volume	1.90	1.90	1.92	2.10	1.96	1.94	1.99	2.16	2.00	1.96	2.03	2.20	1.96	1.95	2.79	2.83
	Lasso/LSE		2.07	2.05	3.00	2.86	1.95	1.93	2.31	2.06	2.51	2.50	2.55	2.21	2.02	2.03	3.28	2.83
	IR	CPU	0.09	0.09	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.15	0.09	0.10	15.43	15.28	14.94	15.33
	Lasso/LSE		2.06	2.11	2.27	3.02	2.53	2.61	2.82	3.56	0.01	0.01	0.01	0.01	22.05	22.33	75.58	190.1
L = 5	IR	Coverage	0.84	0.81	0.84	0.65	0.92	0.91	0.90	0.79	0.94	0.94	0.95	0.93	0.83	0.84	0.71	0.26
	Lasso/LSE		0.75	0.77	0.74	0.65	0.87	0.89	0.88	0.73	0.94	0.94	0.95	0.96	0.85	0.83	0.5	0.64
	IR	Volume	6.78	6.34	7.52	8.45	7.24	6.74	8.13	9.18	7.27	6.84	8.26	9.37	7.65	7.26	17.86	18.45
	Lasso/LSE		9.22	8.91	24.89	20.88	7.36	6.89	11.93	8.95	12.60	12.09	14.59	9.75	8.27	7.77	26.47	17.9
	IR	CPU	1.76	1.16	1.37	1.16	1.69	1.41	1.25	1.12	1.27	1.24	1.29	1.15	195.01	190.78	192.79	194.16
	Lasso/LSE		5.16	4.73	5.35	5.93	5.48	5.50	6.62	7.11	0.01	0.01	0.01	0.01	58.67	63.54	202.78	471.46

Table 1: Results of simulation study displayed in Figure 1.

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

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.

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.

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

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.

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.

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

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.

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.

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

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.

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.