Prediction regions through Inverse Regression

Emilie Devijver, Emeline Perthame

Prediction regions through Inverse Regression *

Emilie Devijver

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France*
e-mail: emilie.devijver@kuleuven.be

and

Emeline Perthame

Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France
e-mail: emeline.perthame@pasteur.fr

Abstract: Predict 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 have already been derived for the well-known linear model, but recently, 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. To deal with high-dimensional data, inverse regression is used in this paper. 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 and make the problem tractable. 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.

MSC 2010 subject classifications: 62F12; 62F25; 62J05; 62E20.
Keywords and phrases: 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 linearly describe a response $\boldsymbol{Y} \in \mathbb{R}^L$ from regressors $\boldsymbol{X} \in \mathbb{R}^D$. The standard Gaussian linear model assumes that there exists $\boldsymbol{A}^\star \in \mathbb{R}^{L \times D}$ such that

$$\boldsymbol{Y} = \boldsymbol{A}^\star \boldsymbol{X} + \boldsymbol{\varepsilon} \tag{1}$$

*Institute of Engineering Univ. Grenoble Alpes
*This is an original survey paper

where the unobserved error term $\boldsymbol{\varepsilon} \sim \mathcal{N}_L(\mathbf{0}, \boldsymbol{\Sigma}^\star)$ is a Gaussian white noise.

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.

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 [20], the Dantzig selector [1], or the ridge estimator [11] 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.

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 $\boldsymbol{Y}$ on $\boldsymbol{X}$ in a space of lower dimension than D ; and the Sliced Inverse Regression (SIR) introduced in [15] 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 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, the 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 [7, 9, 10] 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, [17]). See [3] 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 Gaussian mixtures of regression models and applied to various data (planetology and spectra [5, 18]).

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 [5]), 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 also a performing method under some 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 [6]).

In this paper, we propose to address the multiple linear regression problem of Equation (1) 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 modeling 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, [13, 22, 24] derive confidence regions for slope coefficient and statistical testing of sparsity for linear model using several tools: relaxed projection [24], desparsifying Lasso [22] or through the computation of an approximate inverse of the Gram matrix [13]. Since those pioneer works, several articles provide extensions for more general models or estimators, as generalised linear model ([22] for convex loss function, [12] for subdifferential loss). We also refer to [16] for groups of variables and [19] 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 [2, 14], those results still remain difficult to be implemented.

In this paper, we propose to address the linear regression problem of Equation (1) by considering an inverse regression approach rather than sparse regression. We assume that the residuals of the inverse model are independent which reduce the number of parameters to estimate and overcome the dimensionality burden. In this modelling context, assessing confidence in predicted values is one major goal as deriving prediction regions is classical in regression, for the least square estimator for example. 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 this problem. Considering this approach, we get asymptotic

and non asymptotic distribution for parameters estimates, and then derive confidence regions for slope coefficients. Moreover, we derive asymptotic prediction regions which quantify uncertainty with prediction through an asymptotic normality theorem. 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 and non 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, we introduce the various elements of the modeling framework.

2.1. Inverse regression method

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

$$\mathbf{X}_i \sim \mathcal{N}_D(\mathbf{0}, \mathbf{\Gamma}^\star) \tag{2}$$
$$\mathbf{Y}_i | \mathbf{X}_i = \boldsymbol{A}^\star \mathbf{X}_i + \boldsymbol{\varepsilon}_i \tag{3}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N) \in \mathbb{R}^{L \times N}$ contains L responses for N subjects and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N) \in \mathbb{R}^{D \times N}$ contains D Gaussian 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 normally distributed, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \sim \mathcal{N}_L(\mathbf{0}, \mathbf{\Sigma}^\star)$. 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:

$$\mathbf{Y}_i \sim \mathcal{N}_L(\mathbf{0}, \mathbf{\Gamma}) \tag{4}$$
$$\mathbf{X}_i | \mathbf{Y}_i = \boldsymbol{A} \mathbf{Y}_i + \mathbf{e}_i \tag{5}$$

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 Gaussian 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 (4) and (5), which leads to assume that both response and covariates are centered.

Interestingly, forward parameters $(\boldsymbol{\Gamma}^\star, \boldsymbol{A}^\star, \boldsymbol{\Sigma}^\star)$ are expressed in function of the inverse parameters $(\boldsymbol{\Gamma}, \boldsymbol{A}, \boldsymbol{\Sigma})$ through the following mapping:

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

As Ψ is a one-to-one mapping, estimating the forward regression model, Equations (2)-(3), or the inverse regression model, Equations (4)-(5), is equivalent. One can also notice that Ψ is an involution. The advantage of the inverse approach appears when assumptions are made on the large residual covariance matrix $\boldsymbol{\Sigma}$ in the inverse regression problem of Equations (4)-(5). Indeed, assuming that $\boldsymbol{\Sigma}$ is diagonal drastically reduces the number of parameters to estimate, while keeping a general modelling. 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 $LD + L(L+1)/2 + D = 615$ by assuming that $\boldsymbol{\Sigma}$ is diagonal.

2.2. Estimation

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

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

Then, using Ψ, we get straightforwardly estimators for the forward regression:

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

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

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

The inverse regression trick allows to compute those estimators even when $D >> N$ as it requires the inversion of the $L \times L$ matrix $\boldsymbol{Y}^T\boldsymbol{Y}$ and not the inverse of $\boldsymbol{X}^T\boldsymbol{X}$. Moreover, inverse regression is not as computationally intensive as the least squares, because only small or diagonal matrices are inverted: $\widehat{\boldsymbol{\Gamma}}$ is of size L and $\widehat{\boldsymbol{\Sigma}}$ is diagonal.

2.3. Prediction of the response

Considering those estimators $(\widehat{\boldsymbol{A}}^\star, \widehat{\boldsymbol{\Gamma}}^\star, \widehat{\boldsymbol{\Sigma}}^\star)$, 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:

$$\widehat{\mathbf{Y}}_{N+1} = \widehat{\boldsymbol{A}}^\star \mathbf{x}_{N+1}.$$

In this article, we are interested in studying the uncertainty around this prediction which can be quantified by deriving prediction region. Moreover, we establish the exact distribution of $\widehat{\boldsymbol{\Gamma}}^\star$ and $\widehat{\boldsymbol{\Sigma}}^\star$ and the asymptotic normality of $\widehat{\boldsymbol{A}}^\star$ which is used to deduce prediction regions.

3. Theoretical study of the estimators

In this section, we assume that covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ are known. Moreover, $\boldsymbol{\Sigma}$ is supposed to be diagonal, which implies a diagonal + low rank decomposition for $\boldsymbol{\Gamma}^\star$. It allows correlations among covariates.

Under those assumptions, exact and asymptotic distribution of estimators are derived in this section for the forward regression.

3.1. Matrix normal distribution and Kronecker product

First we recall some properties about the matrix normal distribution and the tensor product. These results can be found in [8] chapter 2, but every important property is recalled in this paper as we use it extensively.

Definition 1 (Kronecker product)**.** *Let $A \in M_{m,n}(\mathbb{R})$ 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 is used to work with vectors instead of matrices.

Definition 2 (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.*

As we are interested in the distribution of matrix parameters, the matrix normal distribution is introduced.

Definition 3 (Matrix normal distribution)**.** *The random variable $X \in \mathbb{R}^{L \times D}$ is distributed according to a* matrix normal distribution *with mean X_0 and variances $U \in M_{L,L}(\mathbb{R})$ (among-row) and $V \in M_{D,D}(\mathbb{R})$ (among-column), denoted*

$$X \sim \mathcal{MN}_{LD}(X_0, U, V),$$

if and only if $vec(X) \sim \mathcal{N}_{LD}(vec(X_0), V \otimes U)$.

For this distribution, some interesting properties are derived.

Proposition 1. *The following equivalence holds:*

$$X \sim \mathcal{MN}_{LD}(X_0, U, V) \qquad \Leftrightarrow \qquad X^T \sim \mathcal{MN}_{DL}(X_0^T, V, U).$$

Proposition 2. *If $X \sim \mathcal{MN}_{LD}(X_0, U, V)$, the following properties hold for $A \in \mathcal{M}_{r,D}(\mathbb{R})$ and $B \in \mathcal{M}_{L,s}(\mathbb{R})$*

$$AXB \sim \mathcal{MN}_{rs}(AX_0B, AUA^T, B^TVB)$$
$$vec(AXB) = (B^T \otimes A)vec(X)$$

For $A \in \mathcal{M}_{r,D}(\mathbb{R}), B \in \mathcal{M}_{L,s}(\mathbb{R}), C \in \mathcal{M}_{r,D}(\mathbb{R}), D \in \mathcal{M}_{L,s}(\mathbb{R})$, the following holds:

$$\begin{aligned} Cov(vec(AXB), vec(CXD)) &= (B^TVD \otimes AUC^T) \\ Cov(vec(AXB), vec(CX^TD)) &= (B^T \otimes A)E(vec(X)vec(X)^T)T_{LD}^{-1}(D^T \otimes C) \\ &= (B^TV \otimes AU)T_{LD}^{-1}(D^T \otimes C) \text{ for } X \text{ centered} \end{aligned}$$

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. Distribution of matrices $\widehat{\mathbf{\Gamma}}^\star$ and $\widehat{\mathbf{\Sigma}}^\star$

In this section, distributiond of the predictors empirical covariance matrix $\widehat{\mathbf{\Gamma}}^\star$ and the residual covaiance matrix $\widehat{\mathbf{\Sigma}}^\star$ are studied.

As $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ are supposed to be known, an estimator of $\widehat{\mathbf{\Gamma}}^\star$ is deduced by pluging-in the estimator of $\boldsymbol{A}$ as followed:

$$\widehat{\mathbf{\Gamma}}^\star = \mathbf{\Sigma} + \widehat{\boldsymbol{A}}\mathbf{\Gamma}\widehat{\boldsymbol{A}}^\top.$$

The probability density function of $\widehat{\mathbf{\Gamma}}^\star$ is derived in the following theorem.

Theorem 1 (Distribution of $\widehat{\mathbf{\Gamma}}^\star$)**.** *Suppose $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ is a sequence of iid random variables satisfying the model defined in Equations (2)-(3). Suppose that $\widehat{\mathbf{\Gamma}}^\star$ is decomposed as $\mathbf{\Sigma} + \widehat{\boldsymbol{A}}\mathbf{\Gamma}\widehat{\boldsymbol{A}}^\top$ where $\widehat{\boldsymbol{A}}$ is the estimator defined Equation (7), then the probability density function of $\widehat{\mathbf{\Gamma}}^\star$ is defined as, for symmetric definite positive matrices structured as the sum of a diagonal and a low rank matrix:*

$$\pi^{\frac{-DL+L^2}{2}}\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}L\right)\}^{-1}det(\mathbf{\Sigma})^{-\frac{1}{2}L}det(\boldsymbol{B})^{-\frac{1}{2}D}etr\left(-\frac{1}{2}\mathbf{\Sigma}^{-1}\boldsymbol{A}\mathbf{Y}^T\mathbf{Y}\boldsymbol{A}^T\right)$$
$$etr\left(-\frac{1}{2}q\mathbf{\Sigma}(\mathbf{\Gamma}^\star - \mathbf{\Sigma})\right)\left(\prod_{\lambda>0}\lambda(\mathbf{\Gamma}^\star - \mathbf{\Sigma})\right)^{\frac{1}{2}(L-D-1)}\sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}L)_\kappa k!}$$
$$P_\kappa\left(\frac{1}{\sqrt{2}}\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}(\mathbf{Y}^T\mathbf{Y})^{\frac{1}{2}}(\mathbb{I}_L - q\boldsymbol{B})^{-\frac{1}{2}}, \boldsymbol{B}^{-1} - q\mathbb{I}_L, \frac{1}{2}\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Gamma}^\star - \mathbf{\Sigma})\mathbf{\Sigma}^{-\frac{1}{2}}\right)$$

where $\lambda(A)$ corresponds to the eigenvalues of A and $\boldsymbol{B} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{\Gamma}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, and $q > 0$ an arbitrary constant such that $\mathbb{I}_L - q\boldsymbol{B}$ is positive definite, and $\Gamma_L(\cdot)$, $etr(\cdot)$ and the Hayakawa polynomial $P_\kappa(\cdot,\cdot,\cdot)$ defined as in Appendix A.1.

Note that this distribution is related to a Wishart distribution with a rescaling related to $\boldsymbol{\Gamma}$ and a translation of $\boldsymbol{\Sigma}$. The proof is available in Appendix A.2 and mainly uses the law of the unconscious statistician and matricial computation.

Note that response and covariates play a symmetric role in inverse regression as their role are inverted for estimation. However, interestingly, the following theorem involves a standard Wishart-like distribution while the previous one involves a singular Wishart-like distribution even if they consist in finding the distribution of matrices with similar decomposition.

In the same way, the density distribution of residual empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}^\star$ is deduced.

Theorem 2 (Distribution of $\widehat{\boldsymbol{\Sigma}}^\star$)**.** *Suppose $((\boldsymbol{X}_1,\boldsymbol{Y}_1),\dots,(\boldsymbol{X}_N,\boldsymbol{Y}_N))$ is a sequence of iid random variables satisfying the model defined in Equations (2) and (3). Suppose that $\widehat{\boldsymbol{\Sigma}}^\star$ is decomposed as $(\boldsymbol{\Gamma}^{-1} + \widehat{\boldsymbol{A}}^\top\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{A}})^{-1}$ where $\widehat{\boldsymbol{A}}$ is the estimator defined Equation (7), then the probability density function of $\widehat{\boldsymbol{\Sigma}}^\star$ is defined as:*

$$\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}D\right)\}^{-1} det((\mathbf{Y}^T\mathbf{Y})^{-1})^{-\frac{1}{2}D} det(\boldsymbol{\Sigma}^\star)^{(L+1)}$$
$$etr\left(-\frac{1}{2}(\mathbf{Y}^T\mathbf{Y})\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right) etr\left(-\frac{1}{2}q\mathbf{Y}^T\mathbf{Y}\left((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1}\right)\right)$$
$$det((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1})^{\frac{1}{2}(D-L-1)} \sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}D)_\kappa k!}$$
$$P_\kappa\left(\frac{(1-q)^{-\frac{1}{2}}}{\sqrt{2}}(\mathbf{Y}^T\mathbf{Y})^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}, (1-q)\mathbb{I}_D, \frac{1}{2}(\mathbf{Y}^T\mathbf{Y})^{\frac{1}{2}}((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1})(\mathbf{Y}^T\mathbf{Y})^{\frac{1}{2}}\right)$$

where $\Gamma_L(\cdot)$ is the multivariate gamma function, $etr(\cdot)$ is the exponential of the trace of a matrix and $P_\kappa(\cdot,\cdot,\cdot)$ is the generalized Hayakawa polynomial. These notations are more precisely defined in Appendix A.1.

Proof is available in Appendix A.3 with a similar approach of Theorem 1. Note that confidence interval for covariance matrices $\boldsymbol{\Gamma}^\star$ and $\boldsymbol{\Sigma}^\star$ can be derived as the exact distribution of their estimators are known using the previous theorems. Moreover, the exact distribution of $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\Sigma}^\star}$ are known making the exact distribution of $\widehat{\boldsymbol{A}^\star}$ accessible. However, computing this distribution is strong analytically and algorithmically, so in the following section, we focus on the asymptotic normality of $\widehat{\boldsymbol{A}^\star}$.

3.3. Asymptotic normality of $\widehat{\boldsymbol{A}}^\star$

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

$$\widehat{\boldsymbol{A}} \sim \mathcal{MN}_{DL}(\boldsymbol{A}, \boldsymbol{\Sigma}, (\mathbf{Y}^T\mathbf{Y})^{-1}).$$

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.

From this, we derive the asymptotic normality 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 1. *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *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,*

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

Moreover, the covariance of this random matrix is given by the following:

$$\begin{aligned} &Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}))) = \\ &\quad \left((\boldsymbol{\Sigma}^{-1} + (\boldsymbol{A}^\star)^T\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star - 2\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{A}^\star) \otimes \boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}\boldsymbol{\Sigma}^\star\right) \\ &\quad + ((\boldsymbol{A}^\star)^T\boldsymbol{\Gamma}\boldsymbol{A}^\star \otimes \boldsymbol{A}^\star\boldsymbol{\Sigma}(\boldsymbol{A}^\star)^T) \\ &\quad - 2\left((\mathbf{I} \otimes \boldsymbol{\Sigma}^\star\boldsymbol{\Gamma}) + ((\boldsymbol{A}^\star)^T\boldsymbol{A}^T \otimes \boldsymbol{\Sigma}^\star\boldsymbol{\Gamma})\right) T_{LD}^{-1}((\boldsymbol{A}^\star)^T \otimes \boldsymbol{A}^\star). \end{aligned} \tag{13}$$

Proof of Lemma 1 is given in Appendix A.4.

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

Theorem 3 (Asymptotic normality of $\widehat{\boldsymbol{A}}^\star$)**.** *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *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* $\Theta(\boldsymbol{A})$*, then by Slutsky's Lemma we get the following:*

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

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 $R_N(\widehat{\boldsymbol{A}})$ is a rest term and $Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})$ is given in Lemma 1.

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 3, the linear combination with respect to $\hat{\boldsymbol{A}}$ defined in (12) is a multivariate Gaussian, centered.Using (13), we get the distribution of the vectorized vector $vec(\widehat{\boldsymbol{A}}^\star)$.

Limiting distribution (15) is get by using Slustky's Lemma, as $\widehat{\boldsymbol{A}}$ converges in probability to $\boldsymbol{A}$. □

This results is the key theorem of this article as it allows to derive confidence regions for $\boldsymbol{A}^\star$ and prediction regions. Wheres we consider the vectorize matrix $\widehat{\boldsymbol{A}}^\star$, formulae are explicit. Remark that the degree of freedom of the χ^2 distribution 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 $\boldsymbol{y}$ through the inverse regression method.

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

Theorem 4. *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *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

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

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

Note that this confidence region is a quadratic form as matrix parameters are considered. Then, he χ^2 distribtion is involved. Those explicit formulae allows to compute confidence regions in practice. Numeric performances stand in Section 5.

4.2. Prediction regions

Theorem 5. *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Then,*

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

where

$$\widetilde{\mathcal{PR}}_{\boldsymbol{y},\alpha} = \Big\{ y \in \mathbb{R}^L \ s.t. \quad (y - \widehat{\boldsymbol{A}}^\star \boldsymbol{x}_{N+1})^T (\Omega(\boldsymbol{A}^\star \boldsymbol{x}_{N+1}) + \boldsymbol{\Sigma}^\star)^{-1} (y - \widehat{\boldsymbol{A}}^\star \boldsymbol{x}_{N+1}) \leq \chi^2_L(1-\alpha) \Big\} \tag{17}$$

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

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

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

One can notice that the covariance matrix that is inverted in Equation (17) breaks down into 2 parts. The first one, $\Omega(\boldsymbol{A}^\star \boldsymbol{x}_{N+1})$, 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.

Moreover, as previously, every formula is explicit so numerical experiments are derived in Section 5.

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 and a regularized approach. A `R` code is available on authors' webpages to apply the 3 compared methods on simulated data, on the following webpage https://research.pasteur.fr/fr/member/emeline-perthame/.

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.

We consider a response with dimension L 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 three distinct designs namely a high-dimensional one ($N = 50, D = 100$), an asymptotic one ($N = 500, D = 100$) and an intermediate design ($N = 100, D = 100$) which allows to investigate situations with $N \leq D$ and $N > D$. 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 3 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 uniformly drawn into a uniform distribution on $(-2, 2)$. 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 $[-0.5, 0.5]$ and covariance matrix $\boldsymbol{\Gamma}$ is generated as in Case 2. The residual covariance matrix $\boldsymbol{\Sigma}$ is set to $\mathbb{I}_D$

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 [23] to our multivariate response framework.

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 refered 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 $N < D$. The accuracy of the method is assessed by computing the coverage (proportion of testing observations falling into the prediction region), the volume of the prediction

regions and the computation time 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 Table 1. This table 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 is compared to LSE when $N > D$ and to Lasso when $N \leq D$.

First, Table 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 larger coverage and smaller volume. Note that multivariate version of the Lasso is not implemented to our knowledge in `R` which makes IR a challenging method. Interestingly, IR, which does not suppose sparsity in the model, seems to be efficient on sparse design (Cases 1 and 2) regarding to both coverage and volume. Table 1 also illustrates that our results are asymptotic, meaning that performances of IR are good regarding volume and coverage for $N > D$. When $N < D$, the confidence level increases with N and is reached when $N > D$. Note that the confidence level is almost reached for $N = D$ which suggests that the asymptotic normality may be quickly reached. Compared to bootstrapped Lasso, IR 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 and N are computationally feasible. Computation time is reasonable while achieving challenging coverage and volume when both D and N are 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 1 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

			N = 50			N = 100			N = 500		
			Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
	IR	Coverage	0.88	0.87	0.84	0.92	0.93	0.93	0.95	0.94	0.95
	Lasso/LSE		0.86	0.88	0.86	0.90	0.92	0.88	0.94	0.95	0.95
L = 1	IR	Volume	1.26	1.26	1.25	1.28	1.28	1.27	1.29	1.29	1.29
	Lasso/LSE		1.31	1.28	1.53	1.26	1.25	1.35	1.45	1.45	1.44
	IR	CPU	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	Lasso/LSE		1.01	0.97	1.09	1.13	1.13	1.35	0.01	0.01	0.01
	IR	Coverage	0.86	0.84	0.86	0.91	0.91	0.90	0.94	0.94	0.95
	Lasso/LSE		0.86	0.86	0.77	0.89	0.90	0.81	0.95	0.95	0.95
L = 2	IR	Volume	1.90	1.90	1.92	1.96	1.94	1.99	2.00	1.96	2.03
	Lasso/LSE		2.07	2.05	3.00	1.95	1.93	2.31	2.51	2.50	2.55
	IR	CPU	0.09	0.09	0.10	0.09	0.09	0.09	0.09	0.15	0.09
	Lasso/LSE		2.06	2.11	2.27	2.53	2.61	2.82	0.01	0.01	0.01
	IR	Coverage	0.84	0.81	0.84	0.92	0.91	0.90	0.94	0.94	0.95
	Lasso/LSE		0.75	0.77	0.74	0.87	0.89	0.88	0.94	0.94	0.95
L = 5	IR	Volume	6.78	6.34	7.52	7.24	6.74	8.13	7.27	6.84	8.26
	Lasso/LSE		9.22	8.91	24.89	7.36	6.89	11.93	12.60	12.09	14.59
	IR	CPU	1.76	1.16	1.37	1.69	1.41	1.25	1.27	1.24	1.29
	Lasso/LSE		5.16	4.73	5.35	5.48	5.50	6.62	0.01	0.01	0.01

TABLE 1

Results of simulations study: prediction regions computed on datasets simulated under models described in Section 5.1. Coverage, volume and CPU time are computed for each method to compare performances. For large sample size, we compare IR with the LSE and for small size we compare IR with the bootstrapped Lasso. Each method is assessed 1000 times, and mean is computed.

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 Table 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 $L = 2$ and $D = 5$ 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 response. Violin plots of Figures 2 to 4 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 2 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 3. 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 4. The corresponding violin plots are centred around the true value.

Figure 5 displays the distribution of absolute prediction error $|\widehat{\mathbf{Y}} - \mathbf{Y}|$. Note that IR achieves interesting prediction accuracy most of prediction errors are

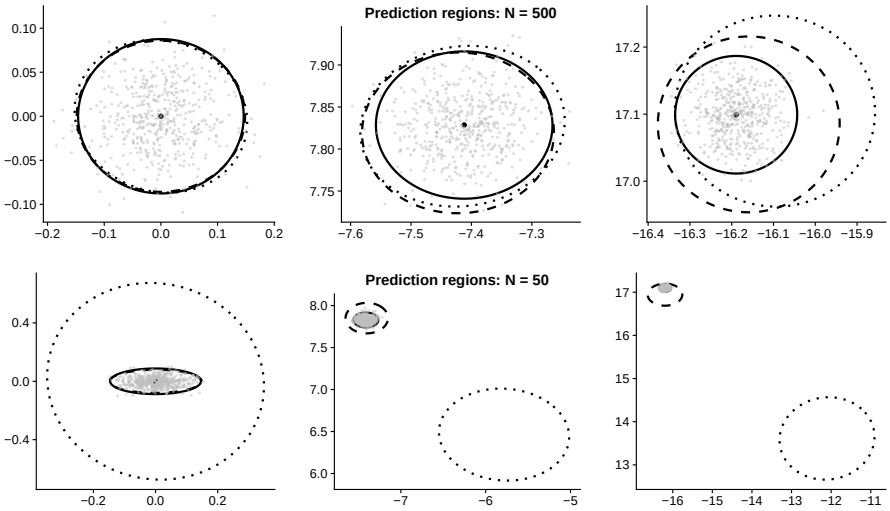

FIG 1. *Prediction regions for $L = 2$. Dotted line: LSE for $N = 500$ and Bootstrapped Lasso for $N = 50$, long dashed line: IR, solid line: true parameters, grey dots: 500 responses generated from the same covariate's profile.*

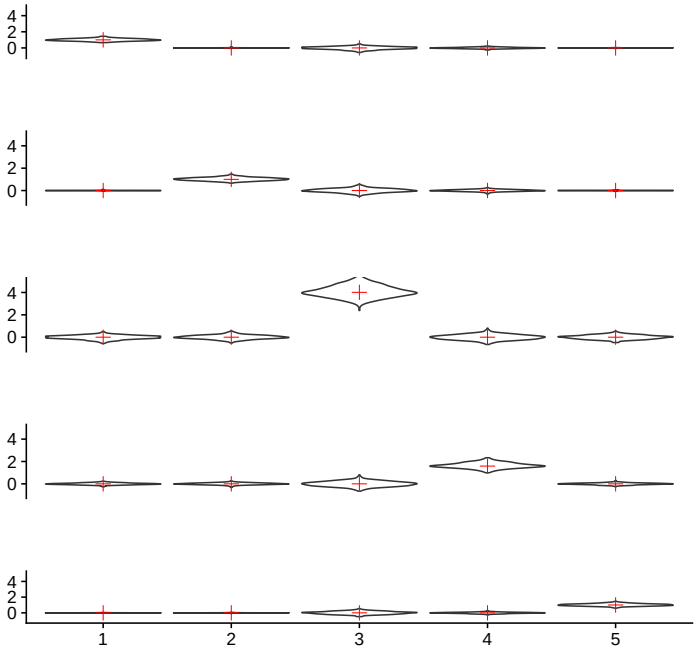

FIG 2. *Violin plots displaying the distribution of $\mathbf{\Gamma}^\star$ estimator for $L = 2, D = 5$ and Case 1. $\mathbf{\Gamma}^\star$ is diagonal, true values are located by red crosses.*

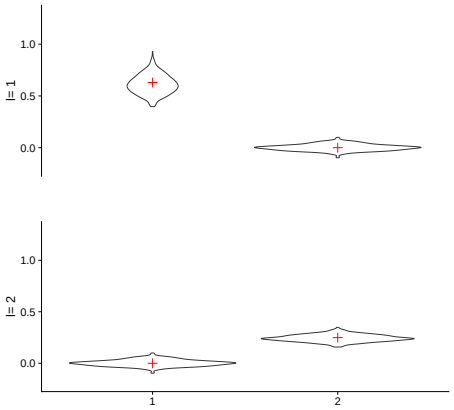

FIG 3. *Violin plots displaying the distribution of* $\boldsymbol{\Sigma}^\star$ *estimator for* $L = 2, D = 5$ *and Case 1.* $\boldsymbol{\Sigma}^\star$ *is diagonal, true values are located by red crosses.*

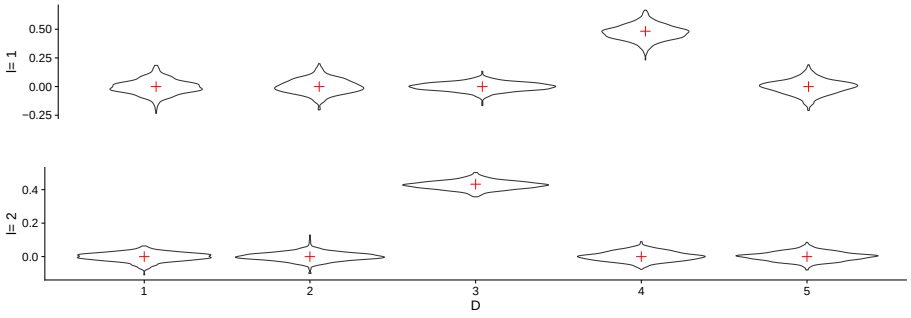

FIG 4. *Violin plots displaying the distribution of* $\boldsymbol{A}^\star$ *estimator for* $L = 2, D = 5$ *and Case 1.* $\boldsymbol{A}^\star$ *is sparse, with 2 non zero entries, true values are located by red crosses.*

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

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 $\mathbf{\Sigma}^\star$ for the 2nd response is smaller than residual variance of first response component.

6. Conclusion and further discussion

In this article, the properties of inverse regression are extensively investigated. 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 exact or asymptotic distributions and confidence regions are deduced. Last but not least, asymptotic prediction regions are derived, allowing to quantify the confidence in estimation.

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.

References

[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.

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

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

[4] N.A.S. Crowther. The exact non-central distribution of a quadratic form in normal vectors. *South African Statistical Journal*, 9:27–36, 1975.

[5] 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.

[6] 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.

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

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

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

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

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

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

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

[14] 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.

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

[16] 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.

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

[18] 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.

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

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

[21] Harald Uhlig. On singular wishart and singular multivariate beta distributions. *Ann. Statist.*, 22(1):395–405, 03 1994.

[22] 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.
[23] N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli, forthcoming paper*, 2017.
[24] 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.

Appendix A: Details for the proofs

A.1. Notations for the proof

In this section, we introduce some notations useful for the proofs.

1. Square root factorization of a positive definite matrix is denoted by $A^{\frac{1}{2}}$ such as $A = (A^{\frac{1}{2}})^2$
2. The imaginary number i is such as $i^2 = -1$
3. Exponential of the trace of a matrix denoted by $\mathrm{etr}(\cdot)$ returns the exponential of the sum of the diagonal terms of a matrix
4. For $L \in \mathbb{N}^\star$, the multivariate gamma function is denoted by $\Gamma_L(\cdot)$ and defined as
$$\Gamma_L(a) = \int_{A>0} \mathrm{etr}(-A)\det(A)^{a-\frac{1}{2}(L+1)}dA,$$
where the real part of a verifies $\mathrm{Re}(a) > \frac{1}{2}(L-1)$ and the integration space $A > 0$ refers to $L \times L$ symmetric positive definite matrices
5. Generalized Hayakawa polynomial introduced by [4] is denoted by $P_\kappa(\cdot,\cdot,\cdot)$ and defined for a complex matrix $T \in \mathcal{M}_{L,D}(\mathbb{C})$ and two real symmetric matrices $A \in \mathcal{M}_{D,D}(\mathbb{R})$ and $B \in \mathcal{M}_{L,L}(\mathbb{C})$ as
$$P_\kappa(T,A,B) = \pi^{-\frac{1}{2}DL}\int_U \mathrm{etr}(-(U+iT)(U+iT)^T)C_\kappa(-BUAU^T)dU$$
where $C_\kappa(S)$ is a zonal polynomial. For more details, we refer to [8].

A.2. Proof of Theorem 1

Proof. We are interested in the distribution of
$$\widehat{\mathbf{\Gamma}}^\star = \mathbf{\Sigma} + \widehat{\boldsymbol{A}}\mathbf{\Gamma}\widehat{\boldsymbol{A}}^\top.$$

From Proposition 3, we know that $\hat{\boldsymbol{A}} \sim \mathcal{MN}_{DL}(\boldsymbol{A}, \mathbf{\Sigma}, (\boldsymbol{Y}^T\boldsymbol{Y})^{-1})$.

Remark that $\widehat{\mathbf{\Gamma}}^\star$ is decomposed onto the sum of a diagonal matrix and a low rank matrix. This structure is general but involves a non invertible matrix.

First we focus on the distribution of $\widehat{\boldsymbol{A}}\mathbf{\Gamma}\widehat{\boldsymbol{A}}^\top$, where the randomness comes from $\widehat{\boldsymbol{A}}$. As $\widehat{\boldsymbol{A}} \in M_{D,L}(\mathbb{R})$ and $\mathbf{\Gamma} \in M_{L,L}(\mathbb{R})$, we know that the $D \times D$ matrix

$\widehat{\boldsymbol{A}}\boldsymbol{\Gamma}\widehat{\boldsymbol{A}}^\top$ is of rank L. Then the distribution cannot be related to a Wishart distribution (arguments used in Section A.3 can not be used).

Finally, combining arguments on quadratic form developed in [8] and singular Wishart distributions introduced in [21], we get the following density for $\boldsymbol{\Gamma}^\star$ defined for symmetric definite positive matrices structured as the sum of a diagonal and a low rank matrix:

$$\pi^{\frac{-DL+L^2}{2}}\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}L\right)\}^{-1}\det(\boldsymbol{\Sigma})^{-\frac{1}{2}L}\det(\boldsymbol{B})^{-\frac{1}{2}D}\mathrm{etr}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\mathbf{Y}^T\mathbf{Y}\boldsymbol{A}^T\right)$$
$$\mathrm{etr}\left(-\frac{1}{2}q\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^\star-\boldsymbol{\Sigma})\right)\left(\prod_{\lambda>0}\lambda(\boldsymbol{\Gamma}^\star-\boldsymbol{\Sigma})\right)^{\frac{1}{2}(L-D-1)}\sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}L)_\kappa k!}$$
$$P_\kappa\left(\frac{1}{\sqrt{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}(\mathbf{Y}^T\mathbf{Y})^{\frac{1}{2}}(\mathbb{I}_L-q\boldsymbol{B})^{-\frac{1}{2}},\boldsymbol{B}^{-1}-q\mathbb{I}_L,\frac{1}{2}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\Gamma}^\star-\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)$$

where $\lambda(A)$ corresponds to the eigenvalues of A and $\boldsymbol{B}=(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\boldsymbol{\Gamma}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, and $q>0$ an arbitrary constant such that $\mathbb{I}_L-q\boldsymbol{B}$ is positive definite, and $\Gamma_L(\cdot)$, $\mathrm{etr}(\cdot)$ and the Hayakawa polynomial $P_\kappa(\cdot,\cdot,\cdot)$ defined as in Section A.1. □

A.3. Proof of Theorem 2

Proof. The purpose of this proof is to derive the distribution of $\widehat{\boldsymbol{\Sigma}}^\star=(\boldsymbol{\Gamma}^{-1}+\widehat{\boldsymbol{A}}^T\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{A}})^{-1}$ knowing that $\widehat{\boldsymbol{A}}\sim\mathcal{MN}_{DL}\left(\boldsymbol{A},\boldsymbol{\Sigma},(\mathbf{Y}^T\mathbf{Y})^{-1}\right)$ from Proposition 3. First, using Chapter 7 of [8], we deduce that the quadratic form $\boldsymbol{S}_A=\widehat{\boldsymbol{A}}^T\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{A}}$ has the following density:

$$\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}D\right)\}^{-1}\det((\mathbf{Y}^\top\mathbf{Y})^{-1})^{-\frac{1}{2}D}\det(\boldsymbol{B})^{-\frac{1}{2}L}\mathrm{etr}\left(-\frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)$$
$$\mathrm{etr}\left(-\frac{1}{2}q\mathbf{Y}^\top\mathbf{Y}\boldsymbol{S}_A\right)\det(\boldsymbol{S}_A)^{\frac{1}{2}(D-L-1)}\sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}D)_\kappa k!}$$
$$P_\kappa\left(\frac{1}{\sqrt{2}}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbb{I}_D-q\boldsymbol{B})^{-\frac{1}{2}},\boldsymbol{B}^{-1}-q\mathbb{I}_D,\frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\boldsymbol{S}_A(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\right)$$

defined for $\boldsymbol{S}_A>0$, with $\boldsymbol{B}=\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}=\mathbb{I}_D$ as $\boldsymbol{\Sigma}$ is diagonal, $0<q<1$ an arbitrary constant, and $\Gamma_L(\cdot)$, $\mathrm{etr}(\cdot)$ and the Hayakawa polynomial $P_\kappa(\cdot,\cdot,\cdot)$ defined as in Section A.1. Therefore the density of $\boldsymbol{S}_A$ simplifies

$$\begin{aligned}&\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}D\right)\}^{-1}\det(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}D}\mathrm{etr}\left(-\frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)\\&\mathrm{etr}\left(-\frac{1}{2}q\mathbf{Y}^\top\mathbf{Y}\boldsymbol{S}_A\right)\det(\boldsymbol{S}_A)^{\frac{1}{2}(D-L-1)}\sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}D)_\kappa k!}\\&P_\kappa\left(\frac{(1-q)^{-\frac{1}{2}}}{\sqrt{2}}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}},(1-q)\mathbb{I}_D,\frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\boldsymbol{S}_A(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\right)\end{aligned}\tag{18}$$

Remark that this density is related to a Wishart distribution, because we consider the quadratic form associated to a Gaussian random variable, the transformation being through a linear application A leads to more complex formulae.

Then, transforming $\boldsymbol{T} = \boldsymbol{\Gamma}^{-1} + \boldsymbol{S}_A$ we obtain the density of $(\boldsymbol{\Sigma}^\star)^{-1}$ as a function of $\boldsymbol{T}$

$$f_{(\boldsymbol{\Sigma}^\star)^{-1}}(\mathbf{T}) = f_{\boldsymbol{S}_A}(\boldsymbol{T} - \boldsymbol{\Gamma}^{-1}), \text{defined for definite positive matrix } \boldsymbol{T}$$

where $f_{\boldsymbol{S}_A}$ refers to the density of $\boldsymbol{S}_A$ defined in Equation (18). Next, transforming $\boldsymbol{\Sigma}^\star = \boldsymbol{T}^{-1}$, with the Jacobian $J(\boldsymbol{T} \mapsto \boldsymbol{\Sigma}^\star) = \det(\boldsymbol{\Sigma}^\star)^{L+1}$ we obtain the density of $\boldsymbol{\Sigma}^\star$

$$f_{\boldsymbol{\Sigma}^\star}(\mathbf{T}) = f_{\boldsymbol{S}_A}((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1})\det(\boldsymbol{\Sigma}^\star)^{(L+1)} \tag{19}$$

defined for positive definite matrix $\boldsymbol{\Sigma}^\star$, which gives the following final density using notation of Equation (18):

$$\{2^{\frac{1}{2}DL}\Gamma_L\left(\frac{1}{2}D\right)\}^{-1}\det(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}D}\det(\boldsymbol{\Sigma}^\star)^{(L+1)}$$
$$\operatorname{etr}\left(-\frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)\operatorname{etr}\left(-\frac{1}{2}q\mathbf{Y}^\top\mathbf{Y}\left((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1}\right)\right)$$
$$\det((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1})^{\frac{1}{2}(D-L-1)}\sum_{k=0}^{\infty}\sum_{\kappa}\frac{1}{(\frac{1}{2}D)_\kappa k!}$$
$$P_\kappa\left(\frac{(1-q)^{-\frac{1}{2}}}{\sqrt{2}}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}, (1-q)\mathbb{I}_D, \frac{1}{2}(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}((\boldsymbol{\Sigma}^\star)^{-1} - \boldsymbol{\Gamma}^{-1})(\mathbf{Y}^\top\mathbf{Y})^{\frac{1}{2}}\right)$$

with functions $\Gamma_L(\cdot)$ and $\operatorname{etr}(\cdot)$ and the Hayakawa polynomial $P_\kappa(\cdot,\cdot,\cdot)$ defined as in Section A.1.

□

A.4. Proof of Lemma 1

Proof. We use the following lemma.

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

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

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

$$\begin{aligned} Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A}) = &(\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}[\widehat{\boldsymbol{A}} - \boldsymbol{A}]^\top\boldsymbol{\Sigma}^{-1} \\ &- (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\left([\widehat{\boldsymbol{A}} - \boldsymbol{A}]^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}[\widehat{\boldsymbol{A}} - \boldsymbol{A}]\right) \\ &\times (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1} \end{aligned} \tag{20}$$

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})) &= (\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)) &= ((\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)) &= ((\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)) &= (\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)) &= (\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)) &= ((\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.

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

□

Appendix B: Computation for univariate response - easier to understand

Whereas the method becomes less interesting for $L = 1$, because we reduce the problem to 1 dimension through the inversion method, we detail here the theoretical result for the scalar response case as computations are easier to derive and to understand. The only goal of this section is then to be pedagogical.

B.1. Asymptotic normality of $\widehat{A}^\star$

When we consider a real response, the Δ-method is used to deduce the distribution of $\widehat{\boldsymbol{A}}^\star \in \mathbb{R}^D$ from the distribution of $\widehat{\boldsymbol{A}} \in \mathbb{R}^D$. To highlight the univariate response, we denote $\gamma = \boldsymbol{\Gamma} \in \mathbb{R}$ and $s^\star = \boldsymbol{\Sigma}^\star \in \mathbb{R}$. First, let recall the distribution of the least square estimator.

Proposition 4 (Distribution of $\widehat{\boldsymbol{A}}$)**.** *Suppose $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ is a sequence of iid random variables satisfying the inverse regression model defined in Equations (4) and (5). Then, the following holds for the estimator defined in Equation (7).*

$$\widehat{\boldsymbol{A}} \sim \mathcal{N}_D(\boldsymbol{A}, \boldsymbol{\Sigma}(\mathbf{Y}^\top \mathbf{Y})^{-1}).$$

In Proposition 5, we define the function $g : \boldsymbol{A} \mapsto \boldsymbol{A}^\star$ and compute its gradient.

Proposition 5. *Let*

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

Then,

$$\nabla g(\boldsymbol{A}) = -2s^\star \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \boldsymbol{A}^\star + s^\star (\partial \boldsymbol{A})^\top \boldsymbol{\Sigma}^{-1}$$

where $(\partial \boldsymbol{A}) = \mathbb{I}_D$ is the differentiation of $\boldsymbol{A}$.

Proof. As g is the product of $s^\star$ and $\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1}$, we have

$$\nabla g = \partial(s^\star) \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} + s^\star \partial(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1})$$

As $L = 1$, we have $\partial A = \mathbf{I}_D$.
Next, we need to compute $\partial(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1})$:

$$\partial(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1}) = \partial(\boldsymbol{A})^\top \boldsymbol{\Sigma}^{-1}$$

Finally, we compute $\partial(s^\star)$: as $\partial(\gamma^{-1} + \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{A}) = 2\boldsymbol{\Sigma}^{-1} \boldsymbol{A}$,

$$\partial(s^\star) = -s^\star \partial(\gamma^{-1} + \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{A}) s^\star.$$

□

Theorem 6 (Asymptotic normality of $\widehat{\boldsymbol{A}^\star}$)**.** *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Let*

$$g : \mathbb{R}^{D\times 1} \to \mathbb{R}^{1\times D}$$
$$\boldsymbol{A} \mapsto \boldsymbol{A}^\star = s^\star \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Gamma}^{-1} + \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1}$$

and let $\nabla g(\boldsymbol{A})$ *the gradient of* g *evaluated in a matrix* $\boldsymbol{A}$*. Then, the following holds for the estimator* $\widehat{\boldsymbol{A}^\star} = g(\widehat{A})$ *defined in Equation* (9)*.*

$$\sqrt{N}(\widehat{\boldsymbol{A}^\star} - \boldsymbol{A}^\star) \underset{N\to+\infty}{\to} \mathcal{N}_D(0, \boldsymbol{\Gamma}^{-1} \nabla g(\boldsymbol{A})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{A})). \tag{21}$$

Moreover, let P *be the Cholesky decomposition of* $\boldsymbol{\Sigma}$*:* $\boldsymbol{\Sigma} = P^\top P$*. Then,*

$$\sqrt{N}\sqrt{\gamma}(\nabla g(\widehat{\boldsymbol{A}})P)^{-1}(\widehat{\boldsymbol{A}^\star} - \boldsymbol{A}^\star) \underset{N\to+\infty}{\to} \mathcal{N}_D(0, \mathbf{I}). \tag{22}$$

Proof. Equation (21) relies on the Δ-method applied to $\widehat{\boldsymbol{A}}$ which is Gaussian, as detailed in Proposition 4, through the function g defined in Proposition 5.

Equation (22) relies on Slutsky Lemma, which is used because $\widehat{\boldsymbol{A}}$ converges in probability to $\boldsymbol{A}$. □

B.2. Confidence region for $\widehat{\boldsymbol{A}^\star}$

From Theorem 6, confidence regions for $\widehat{\boldsymbol{A}^\star}$ are deduced using the following lemma, which makes the link between χ^2 distribution and multivariate Gaussian distribution.

Lemma 3. *If* $X \sim \mathcal{N}_k(\mu, \Sigma)$ *with* Σ *known, then a confidence region for* μ *at level* $1-\alpha$ *is* $\mathcal{R}_{\mu,\alpha}$*, with*

$$\mathcal{R}_{\mu,\alpha} = \left\{x \in \mathbb{R}^k \text{ such that } (x-\widehat{\mu})^\top \Sigma^{-1}(x-\widehat{\mu}) \le \chi^2_k(1-\alpha)\right\}$$

where $\chi^2_D(1-\alpha)$ *is the* $(1-\alpha)$ *quantile of the* χ^2 *distribution with* k *degrees of freedom.*

If $X \sim \mathcal{N}_k(\mu, \Sigma)$ *with* Σ *unknown, then a confidence region for* μ *at level* $1-\alpha$ *is* $\tilde{\mathcal{R}}_{\mu,\alpha}$*, with*

$$\tilde{\mathcal{R}}_{\mu,\alpha} = \left\{x \in \mathbb{R}^k \text{ such that } n(x-\widehat{\mu})^\top S^{-1}(x-\widehat{\mu}) \le T^2_{k,n-1}(1-\alpha)\right\}$$

with $S = 1/(n-1)\sum_i (x_i - \bar{x})(x_i-\bar{x})^\top$ *the empirical covariance, and* $T^2_{k,n-1}(1-\alpha)$ *the quantile of the Hotelling's* T^2 *distribution with parameters* k *and* $n-1$*.*

Then, we can construct an asymptotic confidence region for $\boldsymbol{A}^\star$ with level $1-\alpha$. Remark that combining Slutsky's lemma and Lemma 3 leads to a χ^2 distribution when the covariance is estimated as done for $\boldsymbol{A}^\star$.

Theorem 7 (Confidence region for $\boldsymbol{A}^\star$)**.** *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Then,*

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

where

$$\tilde{\mathcal{R}}_{\boldsymbol{A}^\star,\alpha} = \left\{\boldsymbol{a}^\star \in \mathbb{R}^D \text{ s.t. } \gamma(\boldsymbol{a}^\star - \widehat{\boldsymbol{A}}^\star)^\top (\nabla g(\widehat{\boldsymbol{A}})^\top \boldsymbol{\Sigma} \nabla g(\widehat{\boldsymbol{A}}))^{-1} (\boldsymbol{a}^\star - \widehat{\boldsymbol{A}}^\star) \leq \chi^2_D(1-\alpha)\right\}.$$

B.3. Prediction region

For a new profile $\boldsymbol{x}_{N+1}$, the prediction is get by $\widehat{\boldsymbol{y}}_{N+1} = \widehat{\boldsymbol{A}}^\star \boldsymbol{x}_{N+1}$ as described in Section 2.3. A prediction region is then deduced in the following theorem.

Theorem 8 (Prediction region for $\boldsymbol{y}$)**.** *Suppose* $((\boldsymbol{X}_1, \boldsymbol{Y}_1), \dots, (\boldsymbol{X}_N, \boldsymbol{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (2) *and* (3)*. Then,*

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

where

$$\widetilde{\mathcal{PR}}_{\boldsymbol{y},\alpha} = \left\{y \in \mathbb{R} \text{ s.t. } (y - \widehat{\boldsymbol{A}}^\star \boldsymbol{x}_{N+1})^\top v^{-1} (y - \widehat{\boldsymbol{A}}^\star \boldsymbol{x}_{N+1}) \leq \chi^2_D(1-\alpha)\right\}$$

with $v = \gamma \boldsymbol{x}_{N+1}^\top \nabla g(\widehat{\boldsymbol{A}})^\top \boldsymbol{\Sigma} \nabla g(\widehat{\boldsymbol{A}}) \boldsymbol{x}_{N+1} + s^\star$.