On the Stability of Reduced-Rank Ridge Regression

1st David Degras
University of Massachusetts Boston, Inria, CEA
Boston, United States
ORCID: 0000-0002-7221-584X

2nd Thomas Chapalain
Inria, CEA, Paris-Saclay University
91191 Gif/Yvette, France
thomas.chapalain@ens-paris-saclay.fr

3rd Bertrand Thirion
Inria, CEA, Paris-Saclay University
91191 Gif/Yvette, France
ORCID: 0000-0001-5018-7895

***Abstract*—Reduced-rank ridge regression (RRR) is a simple yet effective approach to multi-task regression linked to partial least squares (PLS) approaches. As the stability of PLS and related techniques has been recently questioned, we investigate here the behavior of RRR in different data regimes. We find that with respect to out-of-sample prediction, RRR behaves better than alternatives such as ridge regression or PLS. We also show the empirical benefit of RRR in neuroimaging encoding experiments, where brain activity in multiple sites is explained by high-dimensional representation of stimulus content.**

***Index Terms*—multivariate regression, multitask learning, high dimension, shrinkage, neuroimaging, brain encoding**

I. Introduction

Multivariate regression, also known as multi-task regression, is a fundamental task in research domains such as chemometrics [20], genomics [3], econometrics [5], neuroimaging [8], [17], and more. Because response variables in multivariate regression often share patterns of dependence on predictor variables, conducting regression separately on each response (the so-called massive univariate approach) is suboptimal from a statistical perspective. This motivates the design of specific multivariate regression techniques that account for shared structure among predictors and responses. Common challenges in multivariate regression include multicollinearity among predictor and/or response variables; high data dimensionality, whereby the numbers of predictors and responses far exceed the number of observed samples; and high levels of measurement noise.

Several techniques can be used to tackle multivariate regression. Ridge regression shrinks the ordinary least squares (OLS) estimator towards zero to mitigate multicollinearity between predictors and reduce estimation variance. Although ridge regression is in essence a univariate method, it can be "made" (weakly) multivariate by forcing the ridge parameter to be common to all responses. Reduced-rank regression [15] handles the multivariate nature of responses more directly by imposing rank constraints on the matrix of regression coefficients. Reduced-rank ridge regression (RRR, [11]) combines the two previous ideas to improve the estimation. Principal component regression (PCR, [10]) and partial least squares regression (PLSR, [8]) utilize a suitable low-dimensional subspace of of the predictors. While the former usually selects a subspace of maximal variance in the predictor space, the latter seeks a predictor subspace of maximal covariation with the responses. Both approaches can be construed with in the framework of factor models. See also [4] for a comparison of ridge regression, PCR, and PLSR. Hyperparameter selection is most often done through cross-validation. Another type of approach, joint feature selection ([12], [13]), relies on the assumption that all responses are effectively predicted by a common small set of variables. This approach, of which lasso regression with ℓ_1/ℓ_2 penalties is a popular instance, can lead to more interpretable predictive models. As hyperparameter setting is more difficult and optimization more costly in practice, these approaches are not used on large datasets. Bayesian methods [19] and penalized maximum likelihood methods [2] have also been developed to suitably regularize regression solutions.

A number of comparative studies of multivariate regression techniques can be found in the literature ([4], [11]). However, to the best of our knowledge, systematic assessments of multivariate methods in the context of *high-dimensional data* are still scarce. In particular, recent research on partial least squares correlation and canonical correlation analysis in brain-behavior association studies [6] highlights the fact that very large sample sizes are needed to ensure the stability and generalizability of estimates - sample sizes that typically exceed those of most brain imaging studies. This leads us to conjecture that the same may be true in multivariate regression analyses. Accordingly, in this paper we set out to examine the effect of data dimension on mainstream multivariate regression techniques through experiments on artificial and real datasets. Our specific goals are to: (i) assess how quickly estimation performance degrades as the number of predictors increases, and (ii) determine which regression methods are more stable in high-dimensional regimes.

The remainder of this paper is organized as follows. In Section II, we detail the classical multivariate linear regression model and standard estimation techniques such as ordinary least squares, ridge regression, reduced-rank ridge regression, principal components regression, and partial least squares regression. In Section III, we report on a simulation study involving multiple data-generating models, data sizes, and the above estimators. In Section IV, we compare the multivariate regression methods on two large datasets from neuroimaging studies related to brain encoding. A discussion and concluding thoughts are presented in Section 5.

II. Models and estimators

Let $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ be observations of a vector $\mathbf{x} \in \mathbb{R}^p$ of predictor variables and associated vector $\mathbf{y} \in \mathbb{R}^q$ of response variables (or tasks). The multivariate (or multi-task) linear regression model expresses as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ is the $n \times q$ response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ the $n \times p$ predictor (or design) matrix, $\mathbf{B}$ a $p \times q$ matrix of regression coefficients to estimate, and $\mathbf{E}$ a noise matrix. It is often assumed that $\mathbf{X}$ is fixed and the entries of $\mathbf{E}$ are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$. In practice, the observed matrices $\mathbf{X}$ and $\mathbf{Y}$ are typically centered and scaled prior to analysis.

We now introduce common univariate and multivariate regression estimators. The ordinary least squares (OLS) estimator is defined as

$$\widehat{\mathbf{B}}_{OLS} = \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm (square root of sum of squared matrix entries). If $\mathbf{X}$ is not of full rank, e.g., when $p > n$, $(\mathbf{X}^\top\mathbf{X})^{-1}$ in (2) is understood as the Moore-Penrose pseudo-inverse of $\mathbf{X}^\top\mathbf{X}$.

The ridge regression estimator is defined as

$$\begin{aligned}\widehat{\mathbf{B}}_{ridge}(\lambda) &= \arg\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}. \end{aligned} \tag{3}$$

where $\lambda > 0$ is a regularization parameter. The limit case $\lambda = 0$ corresponds to the OLS estimator (2). For simplicity of exposition, the same ridge parameter is used for all q responses here. In some applications however, a separate ridge parameter is selected for each response. This may lead to finer adaptation but also to higher variability in estimation.

The reduced-rank ridge (RRR) estimator formulates as

$$\widehat{\mathbf{B}}_{RRR}(\lambda, r) = \arg\min_{\mathbf{B}:\mathrm{rank}(\mathbf{B})\leq r} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \tag{4}$$

with $\lambda \geq 0$ and $r \leq \min(p, q)$. If $r = \min(p, q)$, there is in fact no rank constraint and the RRR estimator reverts to the ridge estimator (3). If $\lambda = 0$, the RRR estimator boils down to a reduced-rank estimator. The general RRR estimator can be constructed in three steps: first, build the ridge estimator $\widehat{\mathbf{B}}_{ridge}(\lambda)$; second, calculate the truncated singular value decomposition (SVD) of rank r of $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}_{ridge}(\lambda) \approx \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top$; third, obtain $\widehat{\mathbf{B}}_{RRR}(\lambda, r)$ as $\widehat{\mathbf{B}}_{ridge}(\lambda)\mathbf{V}_r\mathbf{V}_r^\top$.

We next present principal components regression (PCR) and partial least squares regression (PLSR), both of which rely on dimension reduction and factor decomposition. To gain insights into these methods, we start by exposing the underlying factor model ([20]):

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E}, \\ \mathbf{Y} &= \mathbf{T}\mathbf{Q}^\top + \mathbf{F}. \end{aligned} \tag{5}$$

Here, $\mathbf{T}$ is a $n \times k$ matrix of independent factors, $\mathbf{P} \in \mathbb{R}^{p\times k}$ and $\mathbf{Q} \in \mathbb{R}^{q\times k}$ are factor loading matrices, and $\mathbf{E}$ and $\mathbf{F}$ are noise matrices. PCR and PLSR differ in the way they estimate $\mathbf{T}$, the first being based on the decomposition of $\mathbf{X}$ and the second on that of $\mathbf{X}^\top\mathbf{Y}$. We note that if $\mathbf{P}$ is of full rank, $\mathbf{T}$ equals $(\mathbf{X} - \mathbf{E})\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}$. Plugging this expression in the equation for $\mathbf{Y}$, model (5) can be written in the form (1).

We now turn to the computation of these methods. PCR consists in performing the principal components analysis (PCA) of $\mathbf{X}$ and then regressing $\mathbf{Y}$ on selected principal components. In addition to reducing the dimension of the predictor space, PCA produces uncorrelated composite variables, which helps numerically stabilizing the subsequent regression. A typical PCR strategy is to retain the first few principal components with highest variance, but other options are available [7]. After performing the SVD of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and extracting suitable submatrices $(\mathbf{U}_r, \mathbf{D}_r, \mathbf{V}_r)$ of rank r (note: these are not the matrices used to calculate the RRR estimator (4)), one sets $\widehat{\mathbf{T}}_{PCR} = \mathbf{U}_r$, $\widehat{\mathbf{P}}_{PCR} = \mathbf{V}_r\mathbf{D}_r$, and $\widehat{\mathbf{Q}}_{PCR} = \mathbf{Y}^\top\widehat{\mathbf{T}}_{PCR}$. The PCR estimator then writes

$$\widehat{\mathbf{B}}_{PCR}(r) = (\widehat{\mathbf{P}}_{PCR}\widehat{\mathbf{P}}_{PCR}^\top)^{-1}\widehat{\mathbf{P}}_{PCR}\widehat{\mathbf{Q}}_{PCR}^\top. \tag{6}$$

Finally, we present PLSR through its standard PLS2 implementation ([20]). In short, the algorithm proceeds by forming the cross-covariance matrix $\mathbf{X}^\top\mathbf{Y}$; extracting its first pair of singular vectors, say $\mathbf{w}_1 \in \mathbb{R}^p$ and $\mathbf{c}_1 \in \mathbb{R}^q$; projecting the data along these directions to get latent variables $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{u}_1 = \mathbf{Y}\mathbf{c}_1$; rescaling $\mathbf{t}_1$ to unit norm and regressing $\mathbf{u}_1$ on $\mathbf{t}_1$ to get the regression coefficient $b_1 = \mathbf{u}_1^\top\mathbf{t}_1$; calculating the vector $\mathbf{p}_1 = \mathbf{X}^\top\mathbf{t}_1$ of factor loadings for $\mathbf{X}$; and deflating $\mathbf{X}$ as follows: $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^\top$. The process is repeated until the required number r of PLS components has been obtained ($r \leq \min(n, p)$). The factor matrix and loading matrices are respectively estimated as $\widehat{\mathbf{T}} = (\mathbf{t}_1, \ldots, \mathbf{t}_r)$, $\widehat{\mathbf{P}} = (\mathbf{p}_1, \ldots, \mathbf{p}_r)$, and $\widehat{\mathbf{Q}} = (b_1\mathbf{c}_1, \ldots, b_r\mathbf{c}_r)$. By construction, the factors are orthogonal ($\widehat{\mathbf{Q}}^\top\widehat{\mathbf{Q}} = \mathbf{I}_r$) and the PLSR estimator equals

$$\widehat{\mathbf{B}}_{PLS}(r) = (\widehat{\mathbf{P}}_{PLS}\widehat{\mathbf{P}}_{PLS}^\top)^{-1}\widehat{\mathbf{P}}_{PLS}\widehat{\mathbf{Q}}_{PLS}^\top. \tag{7}$$

III. Numerical experiments

We have conducted simulations to compare the statistical performances of the estimators of Section II in models (1) and (5) at various data dimensionalities.

Our simulations feature three instances of model (1) studied in [11] (section 3.1.1). There, the predictor vector $\mathbf{x}$ is generated from a multivariate normal distribution with mean zero and Toeplitz covariance matrix $\mathbf{\Sigma} = (\rho^{|i-j|}) \in \mathbb{R}^{p\times p}$ where $\rho = 0.9$. The elements of $\mathbf{E}$ are i.i.d. as $N(0, \sigma^2)$ with $\sigma^2 = 0.25$. The three models differ by the rank of the regression matrix $\mathbf{B}$ which is either $\min(p, q)/2$ (model M_1), $\min(p, q)$ (full rank, model M_2), or 1 (model M_3). Our simulation study also includes the factor model (5) where the matrices $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{T}$, resp. error matrices $\mathbf{E}$ and $\mathbf{F}$, have their entries i.i.d. as $N(0, 1)$, resp. $N(0, \sigma^2)$ with $\sigma^2 = 0.25$. The total number of factors is $r + r_0$, with $r = 10$ predictive factors and $r_0 = 0.1p$ nonpredictive factors (corresponding columns of $\mathbf{Q}$ set to zero). In all considered models, the data dimensions n and q are set to 100 and 20, respectively, whereas p varies

in $\{10, 50, 100, 200, 500, 1000, 2000, 4000, 7000, 10000\}$. The wide range of values for p allows us to assess both standard regression settings where $p = \mathcal{O}(n)$ and high-dimensional settings where $p \gg n$.

For each model and each value of p, 100 simulations are conducted. In each simulation, 75% of the data are used for training (tuning parameter selection, model fitting) and 25% for testing (evaluation). The tuning parameters of each method (λ for ridge regression and RRR, rank r for RRR, PCR, and PLSR) are selected by grid search using 100 random splits of the training data. Denoting the training and testing data by $(\mathbf{X}_{train}, \mathbf{Y}_{train})$ and $(\mathbf{X}_{test}, \mathbf{Y}_{test})$, respectively, we assess the out-of-sample generalizability of a regression estimator $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}(\mathbf{X}_{train}, \mathbf{Y}_{train})$ with the coefficient of determination

$$R^2 = 1 - \frac{\|\mathbf{Y}_{test} - \mathbf{X}_{test}\widehat{\mathbf{B}}\|_F^2}{\|\mathbf{Y}_{test} - \overline{\mathbf{Y}}_{test}\|_F^2}. \quad (8)$$

The numerator of the fraction in (8) shows the squared prediction error whereas the denominator shows the response variance (up to a scaling factor) of the testing data. The quantity $\overline{\mathbf{Y}}_{test}$ is a centering matrix with constant rows equal to the row means of $\mathbf{Y}_{test}$. The R^2 coefficient is bounded above by 1 but it can take negative values as the prediction can get arbitrarily bad.

Figure 1 displays the average R^2 across 100 replications for each method, model, and ratio p/n. The shaded areas cover $\pm$ one standard deviation around the mean. In all models, all methods except OLS show an initial slight increase in R^2 followed by a decrease to 0 as p/n increases. RRR, PCR, and PLSR are the best methods with similar performances. Ridge regression tends to have lower performance when p/n is small to moderate. OLS performs substantially worse than other methods, especially when p/n is close to 1 and the OLS equations are nearly singular. Figure 2 enables a more precise comparison of the methods. It is obtained by rescaling the R^2 coefficients of Figure 1 by their maximum for each model and ratio p/n so that in each case, the best method has relative efficiency 100%. This figure confirms that RRR and PLSR have the best out-of-sample performance in all models and for all data dimensions. RRR dominates other methods in most of the range of p/n but has a drop in efficiency in favor of PLSR when p/n becomes very large (40 or more). PCR is the third best method in models M_1, M_2, M_3. Its relative efficiency in the factor model clearly deteriorates as the number $r_0 = 0.1p$ of nonpredictive factors increases with p/n. Ridge regression appears not to be competitive with RRR and PLSR in high-dimensional regimes. Its performance is significantly worse in model M_3 than in M_1 and M_2, showing its limited ability to recover low-rank structure in the regression matrix $\mathbf{B}$.

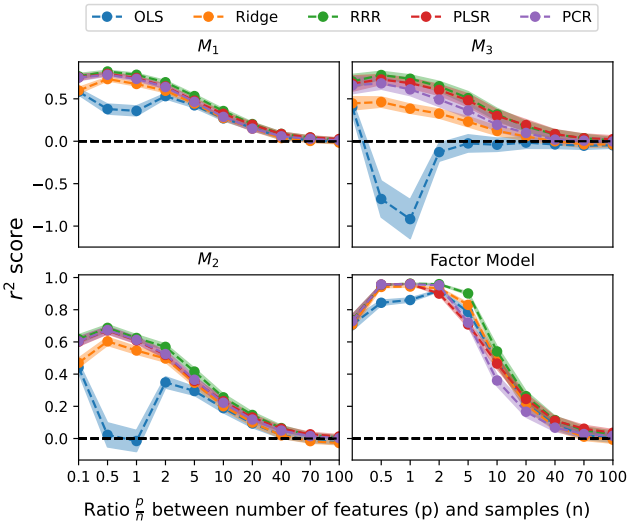

Fig. 1. Simulation study: out-of-sample predictive performance of regression methods as function of the ratio p/n. Five methods are considered in four simulation models. The performance of each method is measured by the coefficient of determination R^2 in (8). Solid lines show the R^2 averaged across 100 replications and shaded areas indicate its standard deviation.

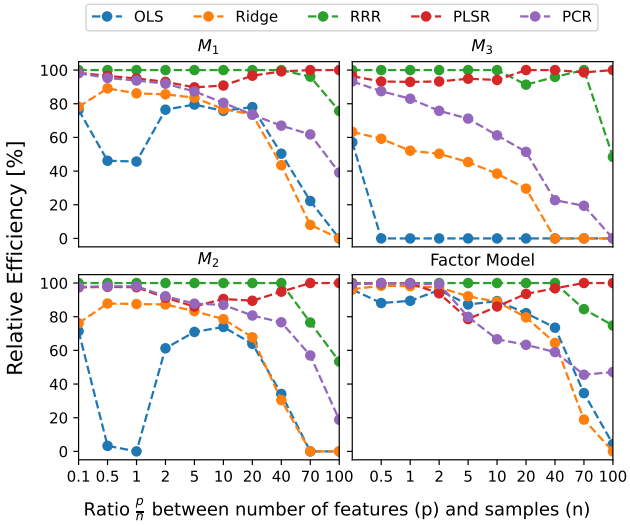

Fig. 2. Simulation study: relative efficiency of regression methods as function of the ratio p/n. Five methods are considered in four simulation models. For each ratio p/n, the coefficient of determination R^2 of each method is scaled by the R^2 of the best method and reported as a percentage.

IV. Application to Functional Neuroimaging Data

This section presents an application of the previous regression techniques to brain encoding tasks in functional neuroimaging. Specifically, the goal is to predict the brain response (as measured by functional magnetic resonance imaging, fMRI) to the viewing of natural images and short film sequences.

a) Datasets: The Natural Scenes Dataset (NSD) [1] is a recently published 7T fMRI dataset comprising a diverse collection of high-resolution fMRI measurements of 8 healthy participants viewing natural scenes taken from the MS-COCO dataset, corresponding to a grand total of 73 000 images. The associated measured fMRI data is z-scored within each NSD scan session and averaged across image repeats.

The Individual Brain Charting (IBC) [14] dataset is a 3T

fMRI dataset composed of a large amount of task-fMRI data acquired for a cohort of 12 healthy subjects. Here, we used specifically the BOLD fMRI signals recorded during the free-watching of the Raiders of the Lost Ark movie. The movie was cut into 10 segments lasting around 12 minutes each, corresponding to approximately 4800 brain volumes.

b) Methods: A parcellation of the fMRI brain images of each participant was performed using the Glasser atlas to keep only the region-of-interest (ROI) that showed reliable responses to images during the NSD experiment. These reliable ROI comprised the whole the visual cortex, resulting in a total of approximately 20,000 vertices of the left (LH) and right (RH) hemispheres respectively.

An encoding pipeline composed of a convolutional neural network (CNN) to extract the features representation of the visual input stimuli, coupled with a regression model (Ridge, RRR, PCR or PLSR respectively) that maps this representation to the brain data was used to fit individually each ROI fMRI data via nested cross-validation. The optimization hyperparameters of every regression method were selected within an inner-validation loop. Specifically, EfficientNet [18], resp. CorNet-RT [9], was used to obtain the features representation from the visual input stimuli of the NSD, resp. IBC.

For each participant in the NSD experiment, approximately 9000 different images were used as training set and 300 different left-out images used as a test set.

For the IBC dataset, the 4800 brain scans were split into the different sessions they were acquired from and a leave-one session out cross validation scheme was used to obtain the generalization performances of the regression methods.

To determine how well each regression model encoded brain responses, their predictive performances on left-out data were measured by the median noise-normalized encoding accuracy across all the vertices of all subjects and hemispheres:

$$\begin{aligned} \text{Score} = &\quad \text{median}\left\{\frac{R_1^2}{NC_1}, \cdots, \frac{R_v^2}{NC_v}\right\} \times 100 \\ R_v = &\quad \frac{\sum_t (G_{v,t} - \overline{G}_v)(P_{v,t} - \overline{P}_v)}{\sqrt{\sum_t (G_{v,t} - \overline{G}_v)^2 \sum_t (P_{v,t} - \overline{P}_v)^2}} \end{aligned} \tag{9}$$

where v is the index of vertices (over all subjects and hemispheres), t is the index of the test stimuli images, G and P correspond to, respectively, the ground truth and predicted fMRI test data, $\overline{G}$ and $\overline{P}$ are the ground truth and predicted fMRI test data averaged across test stimuli images, R is the Pearson correlation coefficient between G and P, and NC is the noise ceiling. For the NSD dataset, the noise ceiling was directly published alongside the paper's fMRI data whereas for the IBC dataset, the noise ceiling was estimated using the Shared-Response Model [16] (SRM) across all the participants.

c) Results: Brain encoding is a hard estimation problem due to its large feature space, high-dimensional output and low signal-to-noise ratio. As Figures 3 and 4 show, RRR outperforms alternatives in this task on both datasets and all

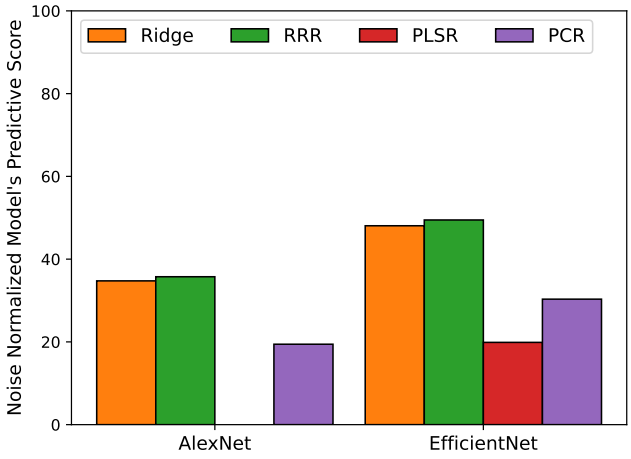

Fig. 3. Brain encoding task on the NSD dataset. Noise normalized prediction performances of ridge regression (Ridge), reduced-rank ridge regression (RRRR), principal components regression (PCR) and partial least squares regression (PLSR). Image features extracted by two convolutional neural network: AlexNet and EfficientNet. Some results are missing due to computational time issues.

CNN-derived image representations. The improvemement with respect to the second best method (ridge) is modest on the NSD dataset but much more marked on the IBC data. We presume that the superiority of RRR comes from two facts: (i) it can select a relevant covariate space based on the joint observation of $\mathbf{X}$ and $\mathbf{Y}$, unlike PCR, and (ii) it leverages the shared information between image voxels to denoise estimates (unlike ridge regression which operates in a univariate fashion) and thus leads to possibly small, yet systematic gains. We also note that in our data analyses, RRR, ridge regression, and PCR have all similar computation times whereas PLSR takes a much longer time to complete. (At the time of submitting this manuscript, the PLSR analysis of the IBC data based on the very large AlexNet features was still running.) This behavior might be related to implementation issues —we used the Scikit- Learn version for these experiments.

V. Discussion

This paper constitutes an initial investigation of the stability of multivariate regression techniques as data dimensions vary. This line of questioning is relevant to various scientific domains in which modern technology enables data collection at very large scales. Here, we have sought to assess the out-of-sample predictive ability of methods with a focus on high-dimensional settings. While it has been recently shown in [6] that multivariate association approaches such as canonical correlation analysis (CCA) or PLS scale poorly in high-dimensional settings, our results outline that predictive approaches retain decent performance, even in very hard settings when the number of features is much larger than the number of samples. Mutlivariate methods then typically outperform univariate models, such as ridge regression or OLS, as they use the shared information between regression problems to denoise the estimates. Interestingly, the same behavior is obtained with

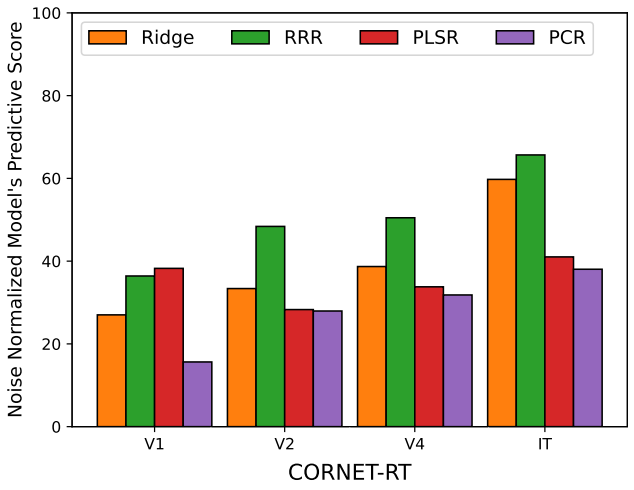

Fig. 4. Brain encoding task on the IBC dataset. Noise normalized prediction performances of ridge regression (Ridge), reduced-rank ridge regression (RRRR), principal components regression (PCR) and partial least squares regression (PLSR). V1, V2, V4 and IT correspond to the different hierarchical layers of the convolutional neural network CORNET-RT.

neuroimaging data: the bottleneck of current encoding models is the use voxel-specific estimates, where the signal in each voxel is noisy. Leveraging the distributed information across voxels proves beneficial. Among multivariate techniques, RRR behaves particularly well, as it performs covariate selection by considering the target information unlike PCR. This is particularly useful when the feature space is very large.

A limitation of the present study lies the restricted scope of its simulations. For example, we have kept the sample size n and number q of response variables fixed to isolate the effect of the predictor dimension p on out-of-sample predictive ability. This choice is motivated by [6] who show that in the related context of association studies, the performances of PLS and CCA approximately depend on the p/n ratio rather than on n and p separately. To gain a fuller picture of the methods' performance and stability, it will however be necessary to vary the dimensions (n, p, q) of models (1) and (5) in a much more systematic fashion, as well the true rank r of the underlying signal and the noise level σ^2. It would also be beneficial to include other techniques such as reduced-rank regression (without ridge penalty) in future analyses to further disentangle which aspects of a method explain its performances in a given data context. Another limitation of this research is the lack of a theoretical framework to better understand the connections between multivariate regression estimators and their statistical properties. To be sure, methods like ridge regression, PCR, and PLSR are not new and already have vast bodies of theory devoted to them. This literature is however spread across multiple domains; synthetic reviews of the relevant theory would be welcome. There are also many open questions on recent techniques such as RRR and how they relate to existing approaches.

References

[1] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[2] Belhassen Bayar, Nidhal Bouaynaya, and Roman Shterenberg. SMURC: High-dimension small-sample multivariate regression with covariance estimation. *IEEE journal of biomedical and health informatics*, 21(2):573–581, 2017.

[3] Anne-Laure Boulesteix and Korbinian Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44, 2006.

[4] lldiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[5] John Geweke. Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75(1):121–146, 1996.

[6] Markus Helmer, Shaun Warrington, Ali-Reza Mohammadi-Nejad, Jie Lisa Ji, Amber Howell, Benjamin Rosand, Alan Anticevic, Stamatios N Sotiropoulos, and John D Murray. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications biology*, 7(1):217–217, 2024.

[7] Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303, 1982.

[8] Anjali Krishnan, Lynne J. Williams, Anthony Randal McIntosh, and Hervé Abdi. Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, 56(2):455–475, 2011.

[9] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

[10] William F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

[11] Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining*, 4(6):612–622, 2011.

[12] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. Technical report, Statistics Department, UC Berkeley, 2006.

[13] Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77, 2010.

[14] Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, Véronique Joly-Testault, Gaëlle Médiouni-Cloarec, Christine Doublé, Bernadette Martins, Philippe Pinel, Evelyn Eger, Gaël Varoquaux, Christophe Pallier, Stanislas Dehaene, Lucie Hertz-Pannier, and Bertrand Thirion. Individual brain charting, a high-resolution fMRI dataset for cognitive mapping. *Scientific data*, 5(1):180105–180105, 2018.

[15] Gregory C. Reinsel, Rajabather Palani Velu, and Kun Chen. *Multivariate reduced-rank regression: theory, methods and applications*. Lecture Notes in Statistics ; volume 225. Springer, New York, NY, second edition. edition, 2023.

[16] Hugo Richard, Lucas Martin, Ana Luísa Pinho, Jonathan Pillow, and Bertrand Thirion. Fast shared response model for fMRI data. *CoRR*, abs/1909.12537, 2019.

[17] Ariel Rokem and Kendrick Kay. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience*, 9(12):giaa133, 2020.

[18] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019.

[19] George C. Tiao and Arnold Zellner. On the bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):277–285, 1964.

[20] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.