

RESEARCH ARTICLE OPEN ACCESS

# Large Multi-Response Linear Regression Estimation Based on Low-Rank Pre-Smoothing

Xinle Tian<sup>1</sup> | Alex Gibberd<sup>2</sup>  | Matthew A. Nunes<sup>1</sup>  | Sandipan Roy<sup>1</sup>

<sup>1</sup>Department of Mathematical Sciences, University of Bath, Bath, UK | <sup>2</sup>School of Mathematical Sciences, Lancaster University, Lancaster, UK

**Correspondence:** Matthew A. Nunes ([man54@bath.ac.uk](mailto:man54@bath.ac.uk))

**Received:** 1 September 2025 | **Revised:** 23 March 2026 | **Accepted:** 25 March 2026

**Keywords:** low-rank approximation | multi-response model | parameter estimation | prediction | pre-smoothing

## ABSTRACT

Pre-smoothing is a technique aimed at increasing the signal-to-noise ratio in data to improve subsequent estimation and model selection in regression problems. However, pre-smoothing has thus far been limited to the univariate response regression setting. However, there are many scientific applications in which interest lies in multi-response regression problems, particularly when the number of responses is large. Motivated by this setting, this article proposes a technique for data pre-smoothing based on low-rank approximation. We establish theoretical results on the performance of the proposed methodology, which show that in this large-response setting, the proposed technique outperforms ordinary least squares estimation with the mean squared error criterion, whilst being computationally more efficient than alternative approaches such as reduced rank regression. We quantify our estimator's benefit empirically in a number of simulated experiments. We also demonstrate our proposed low-rank pre-smoothing technique on real data arising from the environmental and biological sciences.

## 1 | Introduction

In multivariate data analysis, a common task in analyzing multi-response data is to model the relationships between the vector of responses of interest and a number of explanatory variables. Such data are observed in a multitude of scientific applications, including ecology [1], genetics [2, 3], chemometrics [4, 5], bioinformatics [6], population health [7], as well as economics and finance [8, 9]. Whilst component-wise analysis may be an option, analyzing each component separately fails to consider the dependence between the components of the responses; this can result in estimators of model parameters that are less efficient than those that account for potential correlation between responses (see, e.g., References [10, 11]), and thus multivariate linear models are often more appropriate. See [12] for a comprehensive introduction to classical parametric models. Parameter estimation in classical multi-response regression settings is achieved with the ordinary least squares (OLS) estimator, with

the Gauss-Markov theorem showing this estimator to be the best linear unbiased estimator (BLUE) [13, chapter 7].

In many analysis settings, the mean squared error (MSE) criterion is commonly used to measure the quality of a prediction technique, incorporating the well-known bias-variance trade-off in assessing estimators. One way to achieve a smaller MSE in univariate regression settings is to use so-called *pre-smoothing*, in which response data is smoothed in some way prior to estimation via OLS [14–16]. The intuition behind this approach is that this first step increases the signal-to-noise ratio in the data, thus reducing the variance contribution in the bias-variance trade-off and consequently reducing parameter uncertainty. Other authors have also noted that pre-smoothing can help with model selection in uni-response [17]. The pre-smoothing paradigm has subsequently been used in other modeling settings, such as with functional data [18–20], as well as semi-parametric and survival models [21–23]. Despite these contributions, pre-smoothing has

---

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). Statistical Analysis and Data Mining published by Wiley Periodicals LLC.

not been explored in the multi-response regression setting. The work in this article is particularly motivated by settings in which the number of responses is large, a situation commonly found in applications such as genomics [24] and spatiotemporal modeling [25].

To this end, this article proposes an alternative to the OLS estimator for multi-response regression models using data pre-smoothing, in the spirit of [15, 16] in the one-dimensional setting. In what follows, we illustrate this methodology with a low-rank approximation to achieve the pre-smoothing; hence we term our proposed technique *low-rank pre-smoothing* (LRPS). Our proposed low-rank pre-smoothing technique is naturally “model-free” in the sense that it does not need additional covariate information such as the nonparametric pre-smoothers proposed in the one-dimensional setting; see also the examples of Aerts et al. [17, section 2.3]. We investigate how LRPS can be used in the linear regression setting and demonstrate that it achieves a lower MSE than the OLS estimator and shows particular improvements when the number of responses is large or the signal-to-noise ratio is small. Our technique is computationally efficient since estimation can be performed directly in the latent space defined by the low-rank approximation. The proposed methodology can be seen as an alternative to the popular reduced rank regression (RRR) technique for such settings and complements other multi-response regression methods in the literature such as principal component regression (PCR) and partial least squares regression (PLS).

This article is structured as follows. In Section 2, we provide a detailed description of our low-rank pre-smoothing estimator. In particular, the theoretical properties of the estimator are explored in Section 2.1. We then present the results of the simulation studies in Section 3 to examine the effectiveness of the proposed LRPS estimator. We illustrate the practical use of our proposed method in prediction tasks of real data arising in biology and environmental science in Section 4. Finally, we make some concluding remarks in Section 5.

## 2 | Low-Rank Pre-Smoothing in Multi-Response Regression

Consider the multivariate multiple outcome linear regression model

$$Y = XB + E, \quad (1)$$

where  $Y = (y_1, \dots, y_n)^\top$  and  $E = (e_1, \dots, e_n)^\top$  are  $n \times q$  dimensional random matrices representing the outcomes and errors respectively. We study the fixed design case whereby  $X \in \mathbb{R}^{n \times p}$ , and  $B \in \mathbb{R}^{p \times q}$  are parameters to be estimated. We further assume that  $e_i$  is independent from  $e_j$  for all  $j \neq i$ , where  $i, j = 1, 2, \dots, n$ , so rows are assumed to be independent and identical draws from a distribution  $G$  and have constant covariance  $\text{Cov}(e_i) = \Sigma_e$ . In the case where  $e_i \stackrel{iid}{\sim} N_q(0, \Sigma_e)$ , the model (1) can be alternatively written as

$$\text{vec}(Y) \sim \mathcal{N}_{nq}(\text{vec}(XB), \Sigma_e \otimes I_p). \quad (2)$$

When  $n > p$  the maximum-likelihood estimator for  $B$  under (2) is well defined, and is given by the OLS solution  $\hat{B} = S_X^{-1} S_{XY}$  where

$S_X = n^{-1} X^\top X$  and  $S_{XY} = n^{-1} X^\top Y$ . This estimator has the distributional property (see e.g., Mardia et al. [26, chapter 6]) that

$$\text{vec}(\hat{B}) \sim \mathcal{N}_{pq}(\text{vec}(B), n^{-1} \Sigma_e \otimes S_X^{-1}), \quad (3)$$

and as a consequence, the estimator is consistent and unbiased [27, chapter 4].

Whilst asymptotically the OLS maintains the best unbiased linear estimate (BLUE) properties, the  $pq$  dimensional quantity  $\text{vec}(B)$  is still relatively high-dimensional. As such, performance in finite samples may result in noisy estimates, with considerable variance. The aim of this paper is to propose an estimator that enjoys superior finite-sample (and asymptotic) performance when estimating  $B$ . Our proposal is straightforward, and aims to reduce the proportion of the noise, that is, variance of errors, that contributes to the regression problem at hand. We now formally define our proposed estimator.

**Definition 1.** The LRPS estimator.

Let  $S_Y = n^{-1} Y^\top Y$  and consider its eigendecomposition  $\hat{V} \hat{D} \hat{V}^\top = S_Y$ . The Low-Rank Pre-Smoothing (LRPS) estimator for  $B$  is then given by

$$\tilde{B} = S_X^{-1} S_{XY} \hat{V}_k \hat{V}_k^\top, \quad (4)$$

where  $\hat{V} = (\hat{V}_k, \hat{V}_{\setminus k})$ , with  $\hat{V}_k = (\hat{v}_1, \dots, \hat{v}_k)$  being the first  $k \leq q$  eigenvectors, and  $\hat{V}_{\setminus k}$  the remaining  $q - k$  eigenvectors. We assume the ordering of  $k$  matches that of the eigenvalues  $\hat{D} = \text{diag}(\hat{d}_1, \dots, \hat{d}_q)$  and  $\hat{d}_k \geq \hat{d}_{k+1}$ .

The LRPS estimator can be seen as the OLS estimator applied to a “pre-smoothed” set of outcomes  $\tilde{Y} = Y \hat{W}_k$ , with  $\hat{W}_k = \hat{V}_k \hat{V}_k^\top$  acting to project  $Y$  onto the subspace spanned by  $\hat{V}_k$ , and then  $\hat{V}_k^\top$  projecting back into the original dimensions of  $Y$ . The estimator is simple to compute, with the truncated eigenvectors  $\hat{V}_k$  found in  $\mathcal{O}(kq^2)$  computational time. Furthermore, via the Eckhardt-Young theorem [28] we have that  $\tilde{Y}$  is the best rank  $k$  approximation of  $Y$  under the least-squares (Frobenius) loss. The method is feasible as  $\hat{V}_k$  are the empirical eigenvectors associated with the observed  $Y$ , and not on a population, or model-dependent quantity. Our motivation is somewhat similar to that of Cristobal Cristobal et al. [15], Janssen et al. [16] who propose non-parametric pre-smoothing of an outcome before performing the regression. However, in our case, we have a multi-outcome (multivariate) setting, and the smoothing is performed via a specific device, that is the projection onto the most influential eigenvectors of the outcome. We note here that other suitable multivariate non-parametric data smoothers could be used for pre-smoothing, however, they would need other (covariate) information as inputs to the smoother, and potentially suffer from the curse of dimensionality in high-dimensional multi-response data regimes.

At this point, we should note that a similar methodology can be employed within the class of methods known as reduced rank-regression [29]. We elaborate on the similarities and differences between LRPS and these methods later in Section 2.2, however as the name suggests, reduced rank-regression aims to find a low-rank approximation of  $B$ , whereas LRPS makes no such

explicit assumption, and indeed, we can set  $k \geq \min(n, p)$  in LRPS to enable  $\tilde{B}$  to remain full rank. A further difference between the methods is that LRPS does not require knowledge of  $X$  prior to the choice of eigenvectors, as  $V_k$  is obtained directly from  $Y$ . However, the optimal choice of  $k$  will in general depend on the shape of the spectrum (eigenvalues) of  $XB$ ,  $E$ , and the dimensionality  $p, n, q$ . This aspect of our proposed method is explored further in Section 3.

## 2.1 | Theoretical Properties of the LRPS Estimator

In this section, we discuss the asymptotic distribution for the LRPS estimator and the potential efficiency gains relative to OLS. Our analysis is in the standard dimensional setting where  $p$  and  $q$  are assumed fixed with  $p < n$ , and we let  $n \rightarrow \infty$ . To study the estimator in this setting, we consider the population covariance  $\Sigma_y = n^{-1}(XB)^\top(XB) + \Sigma_e = B^\top S_X B + \Sigma_e$ , its eigendecomposition  $\Sigma_y = VDV^\top$ , and the smoothing operator  $W_k = V_k V_k^\top$  based on the first  $k$  eigenvectors.

**Theorem 1.** *Asymptotic distribution of LRPS.*

Under model (1), suppose the moments of the errors  $\{e_1, \dots, e_n\}$  exist (and are finite) up to the fourth order. Assuming that  $\lim_{n \rightarrow \infty} S_X = \Sigma_X$ , then we have

$$n^{1/2} \text{vec}(\tilde{B} - BW_k) \xrightarrow{D} \mathcal{N}_{pq}(\mathbf{0}, W_k \Sigma_e W_k \otimes \Sigma_X^{-1}) \text{ as } n \rightarrow \infty. \quad (5)$$

The proof of the above can be found in Appendix A.1, and relies on the consistency of the sample eigenvectors  $\text{plim}_{n \rightarrow \infty} \hat{V}_k = V_k$  [30, 31], where  $\text{plim}$  denotes as convergence in probability; without loss of generality, we also assume that the design matrix  $X$  (and data matrix  $Y$ ) have been mean-centered. The result shows  $\tilde{B}$  is a root- $n$  consistent estimator of the corresponding low-rank projection of  $B$ ,  $BW_k$ ; we note here that whilst motivated via the OLS estimator under the Gaussian assumption in (2), the distributional result of Theorem 1 holds with only mild moment assumptions on the error distribution  $G$  (see Appendix A.1). Furthermore, this illustrates that whilst the LRPS estimator may be a biased estimator of the coefficient matrix  $B$  (for  $k < q$ ), it enjoys better mean squared error (MSE) performance than the OLS in some important scenarios.

Let  $\text{MSE}(\tilde{B}) := \mathbb{E}_E(\|\tilde{B} - B\|_F^2)$  where  $\|\cdot\|$  denotes as Frobenius norm. We take expectations over the error matrix  $E$  with  $n$  rows and  $q$  columns, for example, as in the model (1). For a large  $n$ , if the asymptotic distribution of Theorem 1 holds, the MSE of LRPS can be approximated as

$$\text{MSE}(\tilde{B}) \approx \underbrace{\frac{\bar{\alpha}_X}{n} \text{Tr}(\widehat{W}_k \Sigma_e)}_{\text{Var}(\tilde{B})} + \underbrace{\text{Tr}\left\{\left(I_q - \widehat{W}_k\right) B^\top B\right\}}_{\text{Bias}^2(\tilde{B})}, \quad (6)$$

where  $\bar{\alpha}_X := \text{Tr}(S_X^{-1})$ . The first term represents the variance of the estimator, and the second, the square of the bias.<sup>1</sup>

Note that the form of the mean squared error expression in (6) is, as expected, dependent on the chosen number of low rank components,  $k$ . This recovers the MSE for the OLS when  $k = q$ ,

in which case the bias disappears as  $\widehat{W}_q = I_q$ . The variance reduction relative to OLS is given by

$$\frac{\text{Var}(\tilde{B})}{\text{Var}(\hat{B})} \approx \frac{\|\Sigma_e^{1/2} \widehat{V}_k\|_F^2}{\|\Sigma_e^{1/2}\|_F^2} \leq \frac{\sum_{i=1}^k \gamma_i^2(\Sigma_e^{1/2})}{\sum_{i=1}^q \gamma_i^2(\Sigma_e^{1/2})},$$

where  $\gamma_i(\Sigma_e^{1/2})$  denotes the  $i$ th largest singular value of  $\Sigma_e^{1/2}$ . Thus if the spectrum of the errors, which refers to the spread of singular values of the error covariance matrix, is relatively flat, then the reduction can be of order  $k/q$ , a substantial benefit when  $q$  is large. On the other hand, if the spectrum of the errors is spiked, for example, if  $\Sigma_e$  were geometrically decaying, then the variance reduction benefits of LRPS may be minimal. In terms of bias, we see this is related to the proportion of the signal (as expressed via  $B$  through  $X$ ) projected into the orthocomplement of the space spanned by the first  $k$  singular vectors of  $Y$ . The eigenvectors  $\widehat{V}_k$  will in general not be optimal for representing the signal  $XB$ , however, are chosen based on a combination of this signal, and the error  $EV_k$ .

## 2.2 | Relation to Reduced-Rank-Regression

In the case when  $k \leq \min(p, n)$  it is of interest to understand the performance of LRPS in terms of estimation bias, in particular, in comparison to reduced-rank-regression (RRR) methods [29, 32] which aims to solve the problem

$$\bar{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \|Y - XB\|_F^2 \quad \text{s.t. } r(B) \leq k,$$

for a given  $k \leq \min(p, q)$ , where  $r(V)$  denotes the rank of  $V$ . A classical approach to this problem is to minimize  $B$  after projecting onto the space spanned by the rank- $k$  approximation of  $\hat{Y} = HY$  where  $H = n^{-1}XS_X^{-1}X^\top$ .

**Definition 2.** Reduced-rank regression.

Consider the decomposition:

$$U_k G_k U_k^\top + U_{\setminus k} G_{\setminus k} U_{\setminus k}^\top = \frac{1}{n} Y^\top H H Y = \hat{B}^\top S_X \hat{B},$$

where  $U_k, G_k$  are matrices representing the first  $k$  eigenvectors and eigenvalues of the decomposition respectively. The Reduced-Rank Regression (RRR) estimator is then defined as

$$\bar{B} = S_X^{-1} S_{XY} U_k U_k^\top. \quad (7)$$

The form of the estimator follows from noting that when  $n > p$  we have

$$\bar{B} = \arg \min_{r(B) \leq k_*} \left\{ \|Y - X\hat{B}\|_F^2 + \|XB - X\hat{B}\|_F^2 \right\}.$$

Applying theorem 2.1 in Reinsel et al. [32] to optimize over the second term in the equation above leads to (7). We thus see that LRPS and RRR are closely related, the difference being the choice of basis used in the projection step. When  $n > p$ , then RRR represents the optimal solution in the Gaussian error setting (where it attains the MLE and thus BLUE property), when the low-rank  $k_* = r(B) \leq \min(q, p)$  assumption holds. Importantly, our proposed LRPS method can also be applied when  $k \geq p$ ,

**TABLE 1** | Comparison of OLS, RRR, LRPS, and PCR, where we define  $U_k, V_k, Q_k$  be the top  $k$  right singular vectors of  $X\hat{B}_{OLS}, Y$  and  $X$  respectively.

Method	Coefficient estimator
OLS	$\hat{B}_{OLS} = (X^\top X)^{-1} X^\top Y$
RRR	$\hat{B}_{RRR} = \hat{B}_{OLS} U_k U_k^\top$
LRPS	$\hat{B}_{LRPS} = \hat{B}_{OLS} V_k V_k^\top$
PCR	$\hat{B}_{PCR} = (Z_k^\top Z_k)^{-1} Z_k^\top Y$ where $Z_k = XQ_k$

whereas RRR falls back to the OLS solution (for  $k \geq \min(p, q)$ ) we obtain  $\bar{B} = \hat{B} U_k U_k^\top = \hat{B}$ ). In this setting, RRR and OLS may have high-variance, whilst LRPS can still benefit from variance reduction.

### 2.3 | Other Related Work

As well as RRR, it is also natural to draw a connection to principal component regression (PCR) [33, 34], which likewise relies on low-dimensional representations using a least squares objective. In PCR, the predictors are first projected onto the leading principal components of the design matrix  $X$ , and the response is then regressed onto this reduced set of components. The resulting estimator thus depends on the principal subspace of  $X$ , rather than directly on the response as in LRPS. Table 1 provides a high-level overview of the three methods, LRPS, RRR and PCR in comparison to OLS.

One may also consider the James-Stein (JS) shrinkage estimator [35], which shrinks the OLS estimator toward zero through a scale-dependent shrinkage factor. Unlike approaches such as LRPS or RRR, the JS estimator does not explicitly impose a low-rank structure on either the response matrix  $Y$  or the covariates  $X$ , but instead achieves regularization solely via global shrinkage of the OLS coefficients. We provide empirical comparison of the methods in Table 1 in Section 3.

### 2.4 | Computational Considerations

In this section, we consider the computational complexity of the LRPS estimator, and compare it to that of RRR to motivate the regression situations in which LRPS will provide benefits.

**Proposition 1.** *Computational cost of LRPS.*

For the regression model (1), the computational complexity of the LRPS estimator is<sup>2</sup>

$$O(np^2 + p^3 + k((n+p)(p+q) + \min\{p, q\})).$$

The details of the computation of the above expression can be found in Appendix A.2. In essence, the LRPS estimator can be computed directly in the latent space, and thus involves matrix multiplications of lower dimension to that of RRR.

*Comparison with RRR.* Using similar arguments to those for the derivation of the result in Proposition 1 (Appendix A.2), the

complexity of RRR using Equation (7) is

$$O(np^2 + p^3) + O(nqk) + O(npq) + O(p^2q) \\ + O(npq) + O(q^2k) + O(pq^2).$$

Both LRPS and RRR algorithms perform two common operations, namely computing the inverse of  $S_X$  of complexity  $O(np^2 + p^3)$ , and performing the truncated SVD of a matrix of dimension  $n \times q$  of complexity  $O(nqk)$  ( $Y$  for LRPS,  $X\hat{B}$  for RRR). These are the first two terms in the computation above and in Appendix A.2 for the LRPS estimator. Removing this from their respective computations, the terms corresponding to the two estimators can be more directly compared:

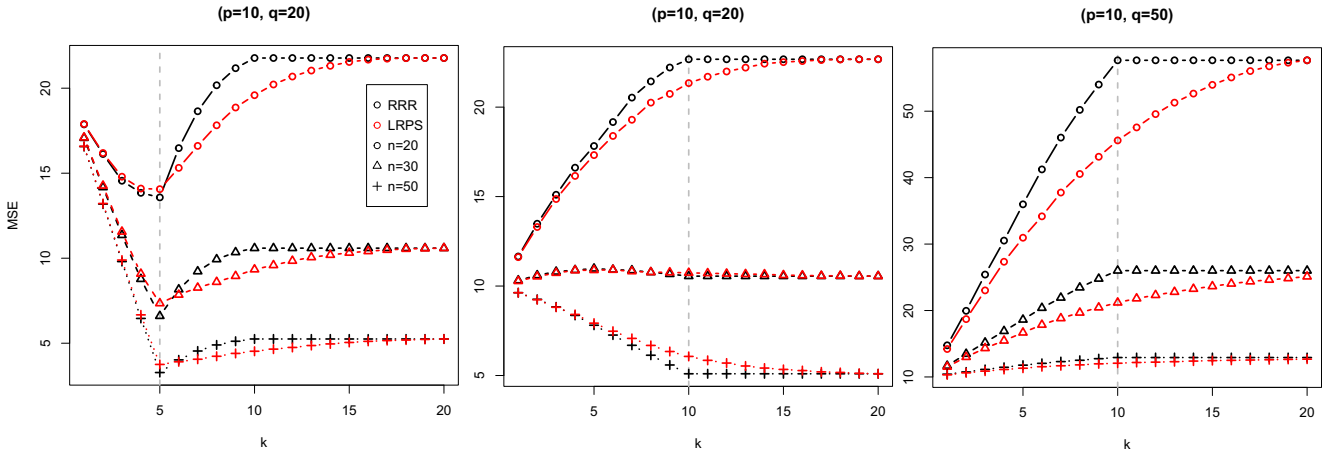
$$\begin{array}{cccccc} \text{RRR} : & O(npq) & O(p^2q) & O(pq^2) & O(q^2k) & O(npq) \\ & \downarrow & \downarrow & \downarrow & \downarrow & \\ \text{LRPS} : & O(npk) & O(p^2k) & O(pqk) & O(\min\{p, q\}k) & \end{array}$$

For the first three terms, there is a scaling factor (reduction) of  $\frac{k}{q}$  from RRR to LRPS, a significant reduction for the fourth term, and the fifth term of RRR is additional to the computation. In light of this, LRPS is expected to be more computationally efficient than RRR if  $k$  is chosen to be low (due to the first three terms), and for large  $q$  settings as signified by the fourth term.

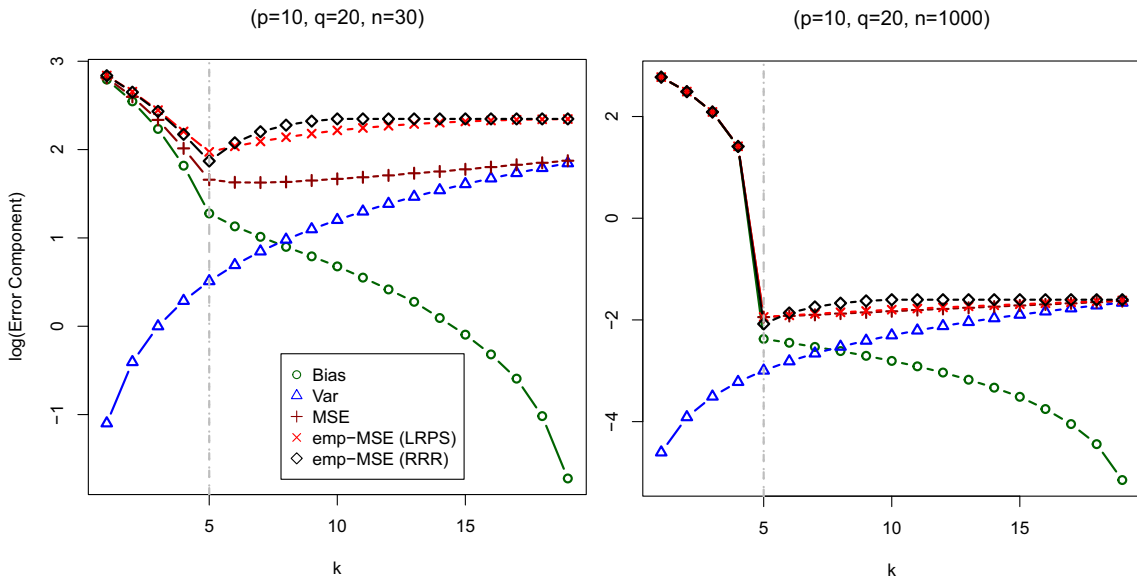
### 2.5 | Empirical Analysis and Finite-Sample Behavior

To give some intuition on the differences between the three estimators, LRPS, RRR, and OLS, we consider a simple simulation study in which we empirically track the MSE, the bias, and the variance of the estimators as a function of  $k$ . We consider two scenarios, one where the true rank  $r(B) = k_* = p/2$  is small relative to  $\min(p, q)$ , and the other where  $k_* = \min(p, q)$ . For simplicity, we let  $p = q = 10$ ,  $k_* = \{5, 10\}$  and consider three sample sizes  $n = \{20, 30, 50\}$  to assess the finite-sample behavior of the estimators. In each case, we assume the model (2), we simulate  $X$  and  $E$  as a set of isotropic Gaussian random vectors (independent rows). Further details on the simulation design can be found in Section 3, where we present results from a more comprehensive study.

The results in Figure 1 shed some light on the similarities and differences between LRPS and RRR. When the true model is low-rank, for example, when  $k_* = 5$ , we see that for all sample sizes the minimum MSE is attained at  $r(\bar{B}) = k_*$  and RRR performs slightly better in terms of estimation error at this minimum. On the other hand, when the low-rank assumption no longer holds ( $k_* = 10$ ), we see that LRPS outperforms RRR in the small sample size setting, whereas RRR becomes superior as the sample size  $n$  grows. Indeed, in the small sample setting, a rank-1 approximation is preferred by both approaches, however, LRPS still maintains better performance. In the scenario where  $q$  is large relative to  $p$  (Figure 1, right panel), LRPS exhibits superior estimation performance in the full-rank scenario, consistently over a range of sample sizes. This latter point is important as the setting where  $q > p$  often occurs in practice. For instance, this is often observed in datasets from air pollution monitoring stations, where measurements of several pollutants are collected across



**FIGURE 1** | Performance of LRPS and RRR in terms of empirical  $\widehat{\text{MSE}} = m^{-1} \sum_{s=1}^m \|Y^{(s)} - X^{(s)} B^{(s)}\|_F^2$  where  $Y^{(s)}, X^{(s)}$  and  $B^{(s)}$  represent the data, and estimate (LRPS in red, and RRR in black) for simulation  $s$ , and  $m = 100$  is the number of simulations.



**FIGURE 2** | The bias<sup>2</sup>( $\tilde{B}$ ), Var( $\tilde{B}$ ), and MSE( $\tilde{B}$ ) values extrapolated from the LRPS asymptotic distribution in Theorem 1, compared to the finite-sample (empirical) MSE for both LRPS and RRR. In this simulation,  $k_* = 5$  and other than the sample size, the conditions are the same as those given in the left panel of Figure 1.

multiple locations. This is discussed later in our application of LRPS to real data.

A characteristic property of LRPS is that the estimator responds to changes in  $k_* > \min(p, q)$ , whereas RRR falls back to the OLS estimator in this regime. To investigate the bias-variance tradeoff occurring due to the pre-smoothing we plot the bias<sup>2</sup>( $\tilde{B}$ ), Var( $\tilde{B}$ ), and MSE( $\tilde{B}$ ) components from (6), and compare them with the empirical MSE achieved in a finite sample, in this case, letting  $n = 30$  and  $n = 1000$ . The results are presented in Figure 2. As one may expect, extrapolating from the asymptotic to the finite sample setting may not be appropriate for small samples, but the approximation becomes tight as  $n$  increases. Comparing the dark red (+) and red (x) lines in Figure 2, we note that they essentially overlap in the larger  $n = 1000$  setting. We now provide more insight into the performance of LRPS through in-depth simulations.

### 3 | Simulation Study

In this section, we evaluate the performance of our proposed technique by using simulated examples to assess the estimation of the regression coefficient matrix, as well as out-of-sample predictive performance. The data  $Y, X$  are simulated according to (1), where we fix the design of the covariates to be  $X \sim \mathcal{N}_{n \times p}(0, I_p \otimes I_n)$ . We vary the specification of the true coefficient matrix  $B$ , and the covariance structure of the errors  $\Sigma_e$ . The choices here cover a range of scenarios that are of interest in practice and help further elucidate the behavior of LRPS (and RRR) when the spectrum (eigenstructure) of the signal  $XB$  and noise  $E$  vary. The various conditions for  $B$  and  $\Sigma_e$  are given below.

#### Coefficient Matrix Conditions:

1. *Sparse and low-rank:* Let  $s$  be the expected rank of the coefficient matrix. We first select active elements for

$B'_{ij}$  by sampling from the Bernoulli distribution,  $B'_{ij} \sim \text{Bernoulli}(s^{-1})$  for  $j = 1, \dots, s$ . The non-zero coefficients in each row are then scaled (by dividing by the row-average) to maintain  $\|B_{i\cdot}\|_2 = 1$ .

2. *Dense and unknown rank*: Two random matrices  $A_L \in \mathbb{R}^{p \times p}$ ,  $A_R \in \mathbb{R}^{q \times q}$  are generated with standard normal, and Uniform(0, 1) entries respectively. Let  $L$  and  $R$  be the left and right singular vectors of  $A_L$  and  $A_R$ , the coefficient matrix is then generated according to

$$B = A_L \text{diag}(d_1, \dots, d_{\min(p,q)}) A_R^T,$$

where  $d_i = \lambda^i$  exhibits geometric decay.

3. *Dense and known rank*: The matrix  $B$  is constructed as in Condition 2, however, the diagonal singular values are replaced with  $d_i = \lambda$  for  $i = 1, \dots, k_*$  and  $d_i = 0$  for  $k_* + 1, \dots, \min(p, q)$ . This creates a matrix which has rank  $k_*$ . Note, the examples demonstrated in Section 2.5 also use matrices generated in this manner.

*Noise Conditions:*

1. *Isotropic IID noise*: The covariance of the errors is given by  $\Sigma_e = \sigma^2 I_q$ .
2. *Geometric Decaying Toeplitz*: We let  $\Sigma_e = \text{Toeplitz}(\rho)$  such that  $\Sigma_{e^{i,j}} = \sigma^2 \rho^{|i-j|}$ .

Note that simulations in this section as well as the examples in Section 2.5 were produced in the R statistical computing environment [37].

### 3.1 | Estimation Error Results

This section examines the estimation error incurred by LRPS and RRR in the scenarios described above. As in the previous section, we focus on understanding the performance of the methods over a range of  $k$ , and examine how this performance is dependent on the composition of the signal and noise. For simplicity, we fix  $n = 100$  and report performance as measured by the empirical mean squared error from the true coefficient matrix  $B$ , averaged across  $m = 100$  experiments.

*Large  $q$ , moderate  $p$ :*

In this case, we consider when  $q = 100 > p = 10$  and thus the range of  $k$  with which we can apply LRPS is extended relative to RRR (which we recall limits  $k \leq \min(p, q)$ ). In this scenario, we should expect that the structure of the error covariance plays a key role, and given the number of outcomes, the regression problem inherits a lower signal-to-noise as  $\|\Sigma_e\|_F \asymp q$  whilst  $\|XB\|_F \asymp \sqrt{pq}$ . We run three experiments in this setting:

- The first probes what we may expect to happen if  $B$  were sparse, under  $B$  condition 1. In this case, the rank of the matrix is bounded by  $s$ , which we choose to be  $p/2 = 5$ , since sparsity is a common assumption in high-dimensional regression, we wanted to investigate how methods that previously would focus explicitly on a low-rankness assumption can also be applied in this setting.

- Second, we investigate the case where  $B$  is dense, but with decaying singular values (Condition 2), and we consider  $\lambda = \{0.9, 0.5, 0.2\}$  to study the impact of the decay rate.
- Third, we examine the impact of a spiked error covariance via a geometrically decaying correlation structure with values  $\rho = \{0.9, 0.5, 0.2\}$ . The specification of  $B$  is given via Condition 3 with a flat spectrum ( $\lambda = 1$ ).

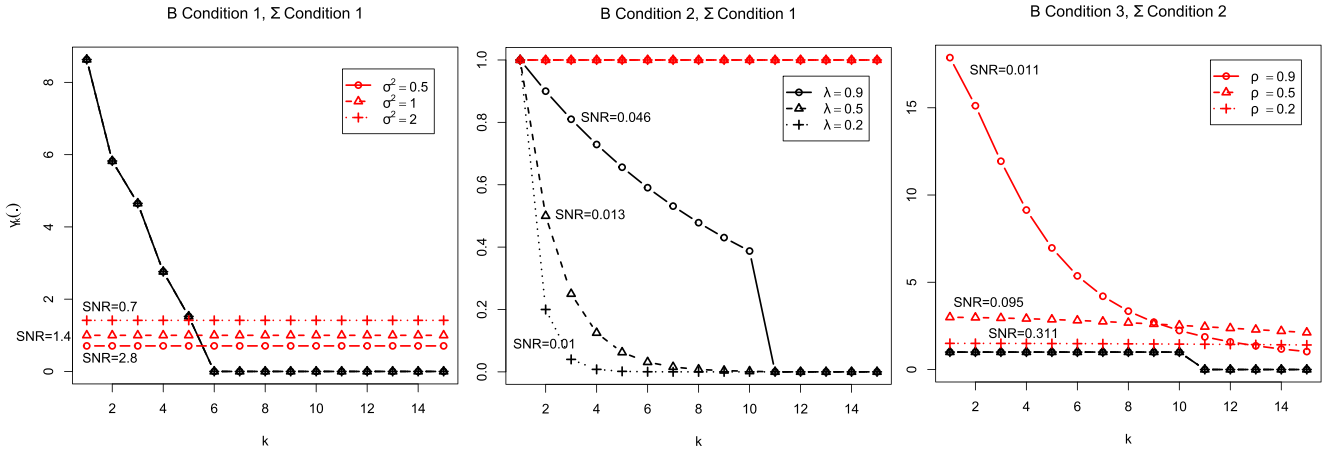
The last two experiments fix  $\sigma^2 = 1$ , whereas the first fixes the design of  $B$  and varies  $\sigma^2 = \{0.5, 1, 2\}$  to directly investigate the impact of the error variance on estimation performance. Figure 2 summarizes the distribution of  $\gamma_j^2(B)$  and  $\gamma_j(\Sigma_e)$ , where for each  $j = 1, \dots, q$  we take the average over  $m = 100$  experiments.

The results of the experiment are presented in Figure 4, with experiments 1–3 running from left to right. The figure shows a range of empirical behaviors for the RRR and LRPS estimators and illustrates both estimators' potential benefits beyond the OLS estimation. The best rank approximation, that is, the  $k$  which minimizes the estimation error is a function of both the specification of the signal (via  $B$ ), and the noise structure—in general, a higher signal-to-noise ratio mandates the choice of a smaller  $k$ . Intuitively, this observation makes sense, as in the high-noise settings, we should desire to constrain the estimator more to avoid excessive variance, that is, the variance reduction benefits of the low-rank projection dominate the error. Observe, at least when  $q > p$ , that LRPS is often superior to RRR in these approximations, for example, when comparing the rank-1 approximation performance in the middle and right plots.

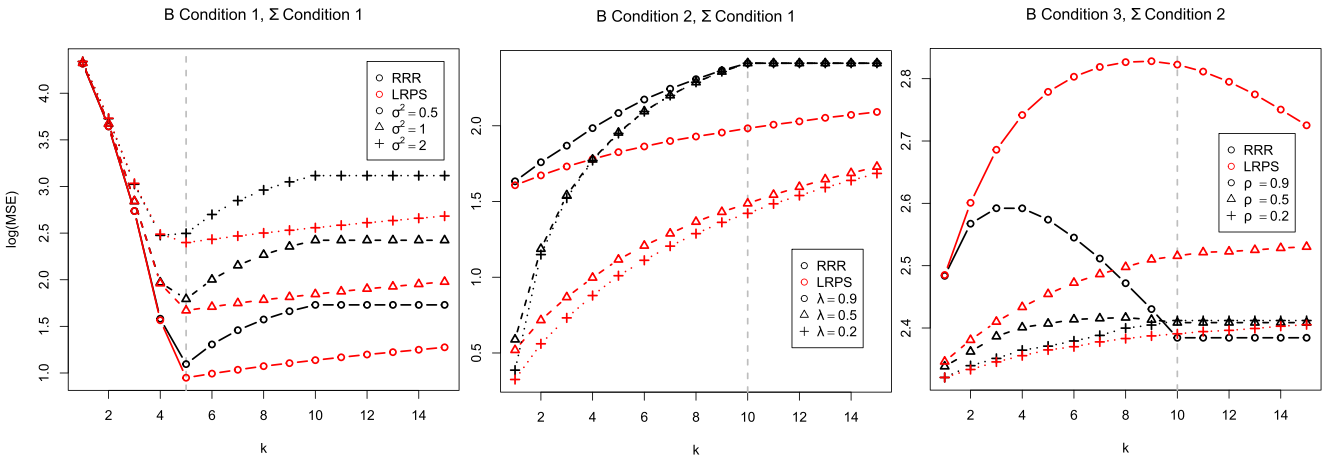
*Moderate  $q$ , large  $p$ :*

In contrast to the high  $q$  setting, in this simulation we investigate the relatively high-dimensional estimation problem where  $q = 10$  is fixed and we set  $p \in \{20, 40, 60\}$ , we fix the sample size at  $n = 100$ . In this situation, the true rank of  $B$  is limited to be less than or equal to  $q = 10$ . Specifically, we simulate  $B$  matrices according to Condition 3, and set the rank  $k_* = 5$ . We vary the SNR according by adopting  $\Sigma$ -Condition 1, and letting  $\sigma^2 \in \{0.5, 1, 2\}$ . The results of the experiment are seen in Figure 5. We see that both RRR and LRPS perform well relative to the OLS, and whilst RRR outperforms LRPS for  $k \leq k_*$  the difference between the two estimators is small. In this case, the best approximating model, that is, the  $k$  which minimizes the MSE, depends not only on the SNR mediated by  $\sigma^2$ , but also the variance in the estimators due to the higher number of covariates  $p > q$ . As  $p$  approaches  $n$  from below, we see that again the rank-1 approximation achieves the best performance for both LRPS and RRR. These results highlight the importance of performing the pre-smoothing (projection) to reduce noise, either when the dimensionality of covariates is high, and/or when the SNR is low. Generally when  $k \geq \hat{k} = \arg \min_k \text{MSE}(k)$ , we see that LRPS outperforms RRR, and the difference in performance can be substantial in the large  $q$  setting.

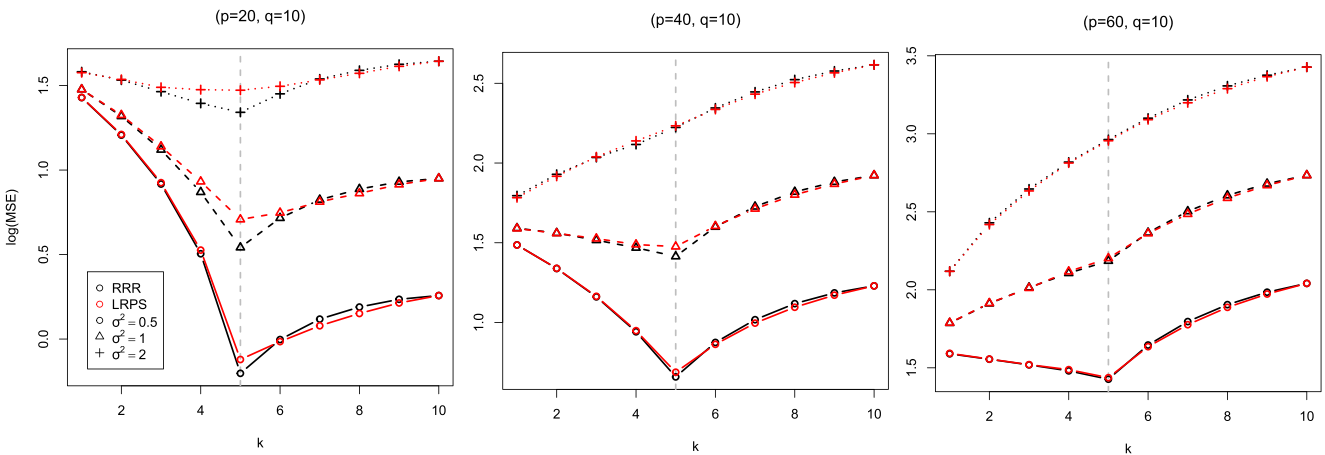
*Sensitivity to heavy-tailed errors.* To conclude our investigations into empirical estimation error, we also examine the performance of LRPS compared with RRR under heavy-tailed error settings. Specifically, multivariate heavy tail errors are generated from a multivariate  $t$ -distribution with 3 degrees of freedom (using the



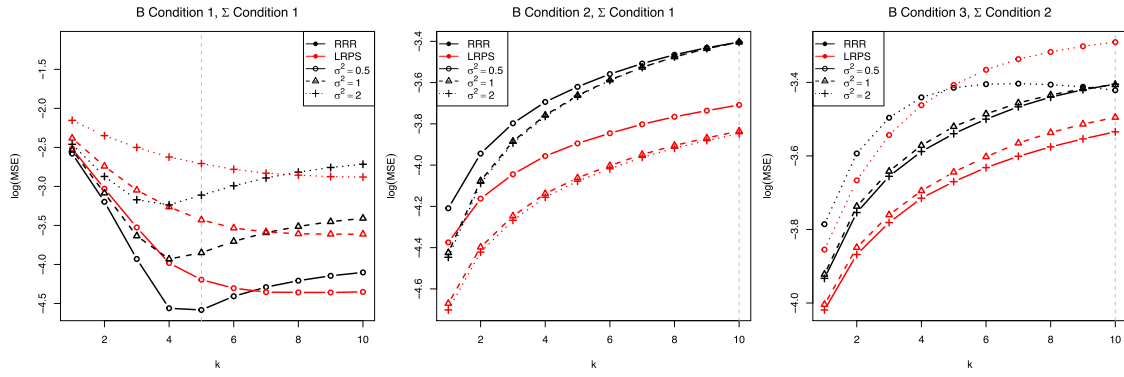
**FIGURE 3** | Distribution of singular values for the signal  $\gamma_j(B)$  and noise  $\gamma_j^{1/2}(\Sigma_e)$ , in black and red respectively. The signal-to-noise ratio  $\text{SNR} := \|XB\|_F / \|\Sigma_e\|_F$  is given next to each line, and the lines represent the average value of the spectrum over  $m = 100$  simulations.



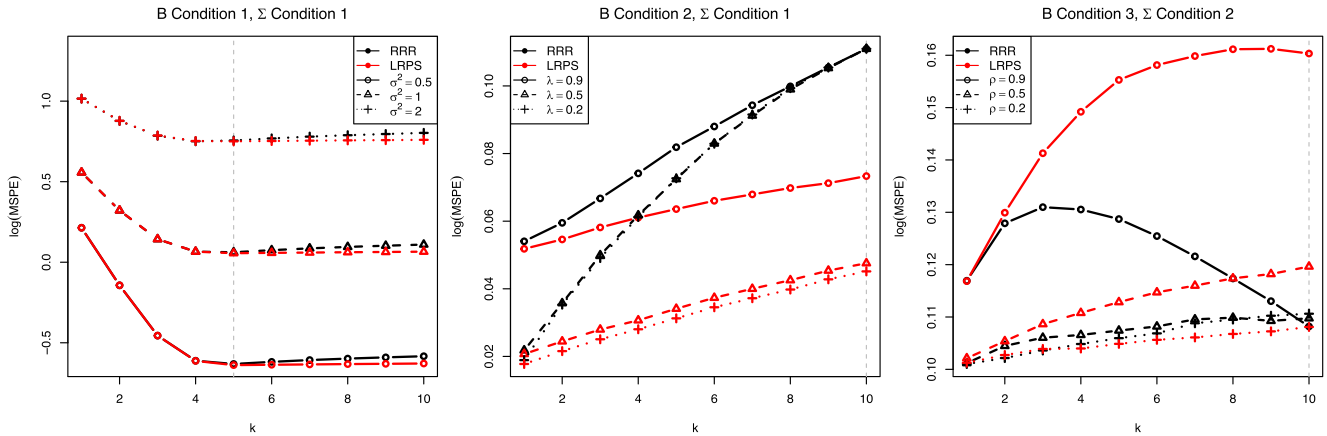
**FIGURE 4** | Comparison of LRPS and RRR performance as a function of  $k$  in the case where  $p = 10$ ,  $q = 100$ ,  $n = 100$ . Black lines denote the performance of RRR and red those of LRPS. In the middle and right plots, we fix  $\sigma^2 = 1$  and vary the eigenvalues of the signal and noise respectively by adjusting  $\lambda$  and  $\rho$ . In  $B$  Condition 3, we set all the singular values of  $B$  to be one, and maintain that  $r(B) = p$  indicated by the dashed line, in the left plot, we set the sparsity to be  $s = 5$ .



**FIGURE 5** | Comparison of LRPS and RRR performance as a function of  $k$  in the case where  $q = 10$ ,  $n = 100$  and  $p \in \{20, 40, 60\}$ . The covariance of errors is given by  $\Sigma$ -Condition 1 with  $\sigma^2 \in \{0.5, 1, 2\}$ , the true rank of  $B$  is denoted by the vertical gray line.



**FIGURE 6** | Comparison of coefficient estimation error across LRPS and RRR under a heavy-tailed error. All results are averaged over 100 simulations with  $p = 10$ ,  $q = 100$ , and  $n = 100$  with different  $B$  conditions.



**FIGURE 7** | Sensitivity plots of out-of-sample prediction risk versus  $k$  for LRPS and RRR in the case where  $p = 10$ ,  $q = 100$ ,  $n = 100$ . Black lines denote the performance of RRR, and red denotes the performance of LRPS.

`mvtnorm` R package [38]). The simulation setup follows the large  $q = 100$ , moderate  $p = 10$  and  $n = 100$  design while varying the structure of the coefficient matrix  $B$ . The error covariance is taken to be either diagonal or Toeplitz, matching the two different noise conditions for the covariance structures used in the Gaussian setting. The results are shown in Figure 6.

In the left panel, corresponding to a low-rank sparse coefficient matrix  $B$ , RRR performs well when the rank is correctly specified but suffers a notable loss in accuracy when the rank is overestimated. By contrast, in the two right panels, where  $B$  is dense, LRPS consistently performs better than RRR across the range of ranks considered. Overall, the performance gains of LRPS relative to OLS and RRR appear robust to the case of heavy-tailed errors.

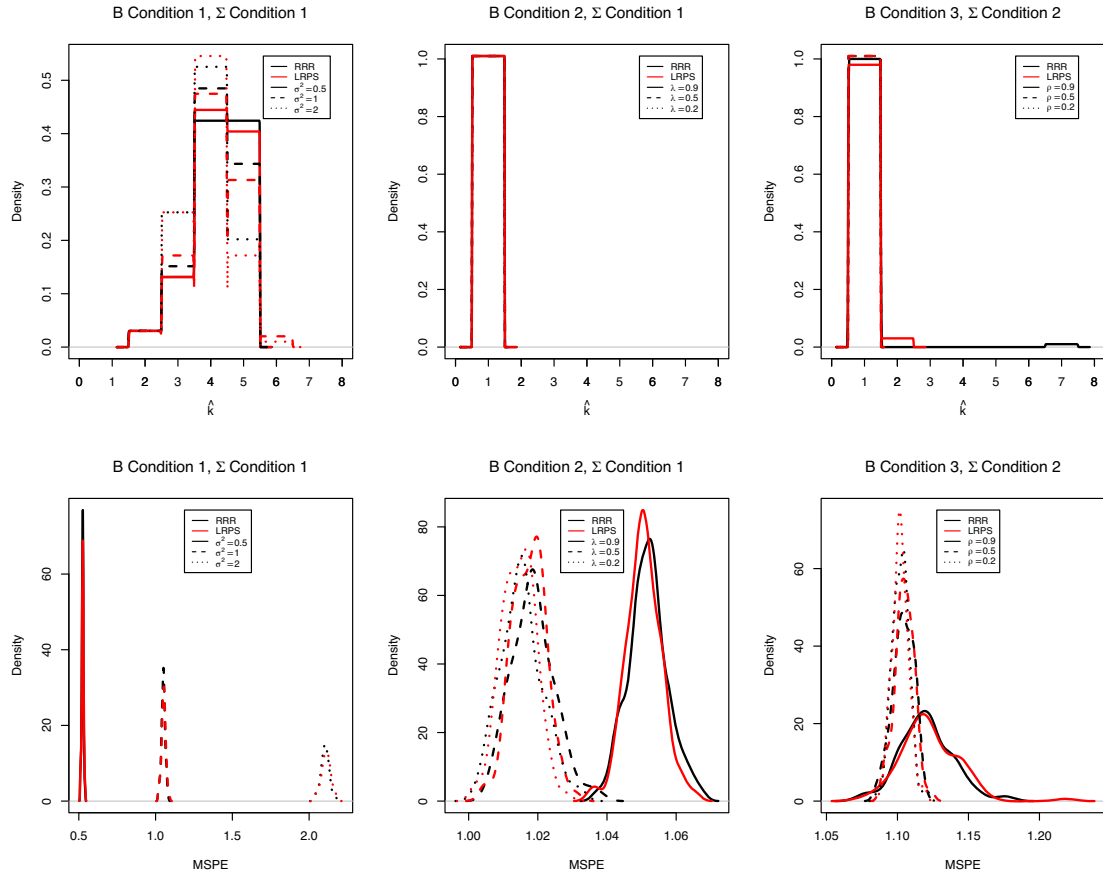
### 3.2 | Predictive Performance

For LRPS and RRR, we see that the choice of  $k$  is important. In this section, we demonstrate the effectiveness of cross-validation for choosing  $k$ , and further evaluate the predictive performance of the LRPS and RRR estimators, in an out-of-sample setting.

Before considering methods to select  $k$ , we first consider the sensitivity of the estimation performance to mis-specification of  $k$  in some examples. To this end, Figure 7 presents plots of the

mean squared predictive error as a function of the rank parameter  $k$  for LRPS and RRR, under the large  $q$ , moderate  $p$  setting corresponding to Figure 4. The performance trends observed in the predictive plots are consistent with those seen in the estimation of the coefficient matrix, as shown in Figure 4. Notably, in scenarios where the error spectrum is relatively flat and LRPS is expected to perform well, it outperforms RRR (the left two panels). In contrast, when the error covariance  $\Sigma_e$  in settings where the error covariance exhibits a spiked spectrum (the right panel), LRPS performs less favorably, highlighting the limitations of the method under such conditions.

In general, the selection of how strong we wish our constraints to be in statistical inference may depend on the task at hand. For example, in the lasso ( $\ell_1$  regularized regression), if one wishes to perform an optimal selection of non-zero coefficients, the specification of the regularization (constraint) may be different from that for producing optimal predictions. In that setting, typically cross-validation will overestimate the support of the regression function compared to more conservative information criteria. Moving back to the rank-restricted setting with RRR and LRPS, we may encounter a similar position—that is, do we construct an information criterion for the selection of  $k$ , or do we use some in-sample empirical measure of prediction error, for example, cross-validation. As we are focusing on global constraints (i.e., working with linear projections across all variates), we here



**FIGURE 8** | Top: Distribution of  $\hat{k}$  (chosen by 2-fold cross-validation); bottom: Distribution of prediction error. Both distributions are estimated across 100 replications. Prediction error is based on applying the singular vectors from the training set  $n = 100$  to the out-of-sample data. The simulation settings mirror those in Figures 3 and 4, with RRR given by the black lines, and LRPS given by those in red.

focus on predictive performance rather than parametric inference when deciding the appropriate specification of  $k$ , and to this end, adopt a simple cross-validation procedure.

Specifically, we propose to apply 2-fold cross-validation, such that we minimize

$$\text{cvMSPE}(k) := \frac{1}{2} \left( \|Y^{(1)} - X^{(1)} \hat{B}^{(2)} W_k^{(2)}\|_F^2 + \|Y^{(2)} - X^{(2)} \hat{B}^{(1)} W_k^{(1)}\|_F^2 \right), \quad (8)$$

where  $Y^{(1)}$  indicates data obtained from the first half of the data and  $Y^{(2)}$  the second half of the data. Note, we use the eigenvectors from the corresponding training dataset, that is,  $W_k^{(j)} = V_k^{(j)} V_k^{(j)\top}$  are based on the first  $k$  singular vectors of  $Y^{(j)} \in \mathbb{R}^{n/2 \times q}$ , where  $j \in \{1, 2\}$ . The  $\text{cvMSPE}(k)$  represents a popular in-sample measure of the out-of-sample performance, in particular giving (for large  $n$ ) a good approximation to a form of predictive risk [39].

In the following experiments, we consider the estimates  $\hat{k}$  according to

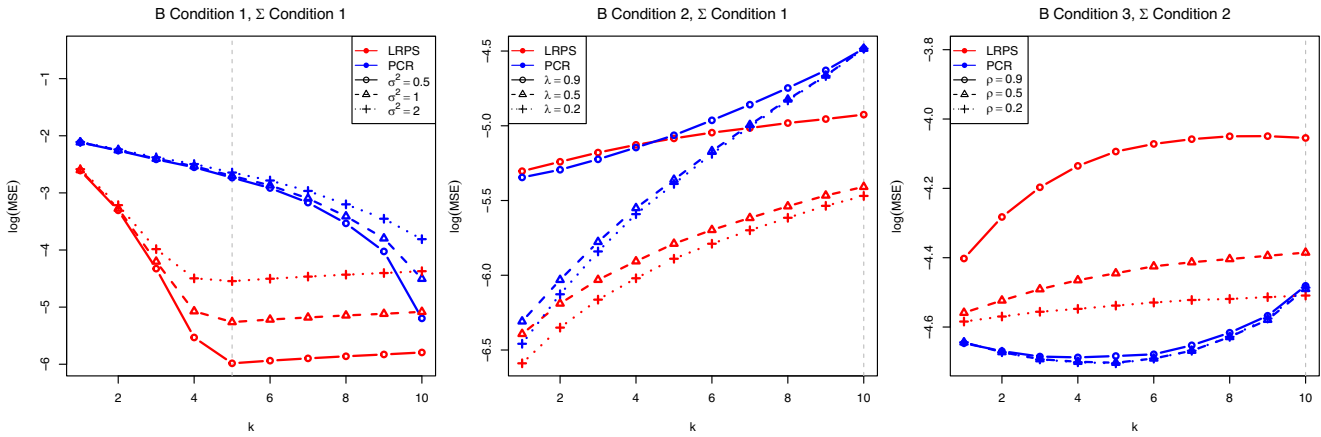
$$\hat{k} = \arg \min_{k=1,2,\dots} \text{cvMSPE}(k),$$

for both RRR and LRPS. Note, the version stated in (8) can be readily applied to RRR by simply replacing  $V_k$  with  $U_k$  (from Definition 2). To evaluate the performance of the selection

criteria, we consider the empirical distribution of  $\hat{k}$  alongside the out-of-sample expected prediction error  $\mathbb{E}_{(y', x')} \|y' - x' \hat{B}_{\hat{k}} V_{\hat{k}}^\top\|_F^2$  where  $y', x'$  are assumed to be generated from the same generating process that generated the training data, that is,  $y' \sim \mathcal{N}_q(0, B^\top \Sigma_X B + \Sigma_e)$  and  $x' \sim \mathcal{N}_p(0, I_p)$ .

The results of our experiment are presented in Figure 8. The simulation settings are exactly the same as those considered in Figure 4. In this case, the difference between LRPS and RRR is negligible—both LRPS and RRR estimators appear to adapt its rank (and model complexity) via the proposed cross-validation procedure, that is,  $\hat{k}$  depends on the simulation design (e.g., signal-to-noise ratio, noise/signal distribution).

It is worth noting that, in this study, we adopt a 2-fold cross-validation scheme for simplicity; however, this is not the only approach for selecting the rank parameter  $k$ . In other applications,  $k$  may be chosen adaptively for each task or fixed across different experimental settings. For example, in time-series neuroscience applications, the optimal rank can be selected using rolling-window or blocked cross-validation schemes that respect temporal dependence. In multi-subject studies, the rank parameter may also be tuned at the subject level or at the group level, depending on whether the primary interest lies in individual-specific or population-level dynamics.



**FIGURE 9** | Comparison of LRPS and PCR performance as a function of  $k$  in the case where  $p = 10$ ,  $q = 100$ ,  $n = 100$ . Red lines denote LRPS and blue lines denote PCR.

### 3.3 | Performance Comparison With PCR

When  $p$  is large, it is natural to consider PCR, as it captures the most relevant structure in  $XB$ . While a comprehensive comparison among LRPS, RRR, and PCR is of interest and will be pursued in future work, we focus here solely on a comparison between LRPS and PCR. Using the same experimental settings as in Figure 4, we fix  $q = 100$ ,  $n = 100$ , and  $p = 10$ , and vary the conditions of  $B$  and  $\Sigma_e$  through different values of  $\lambda$  and  $\rho$ . Under these settings, we examine and compare the coefficient estimation performance of LRPS and PCR.

Figure 9 demonstrates the performance of LRPS relative to PCR across various settings. The results exhibit patterns similar to those observed in the comparison between LRPS and RRR. In scenarios where LRPS outperforms RRR, it also maintains an advantage over PCR. Conversely, in settings where LRPS is expected to perform worse than RRR, its performance likewise falls short when compared with PCR. Overall, these findings reinforce the same conclusion, such that in cases with a flat error spectrum, LRPS continues to achieve the best performance.

### 3.4 | Computational Cost Investigation

To evaluate and compare the computational efficiency of the three methods OLS, RRR, and LRPS, we conduct a series of simulations by systematically varying one of the dimensions using scaling factors from 1 to 5 (sample size  $n$ , response dimension  $q$  or covariate dimension  $p$ ) while keeping the others fixed. For simplicity, we fix  $k = 1$  for both LRPS and RRR.

#### *The case when $p > q$ :*

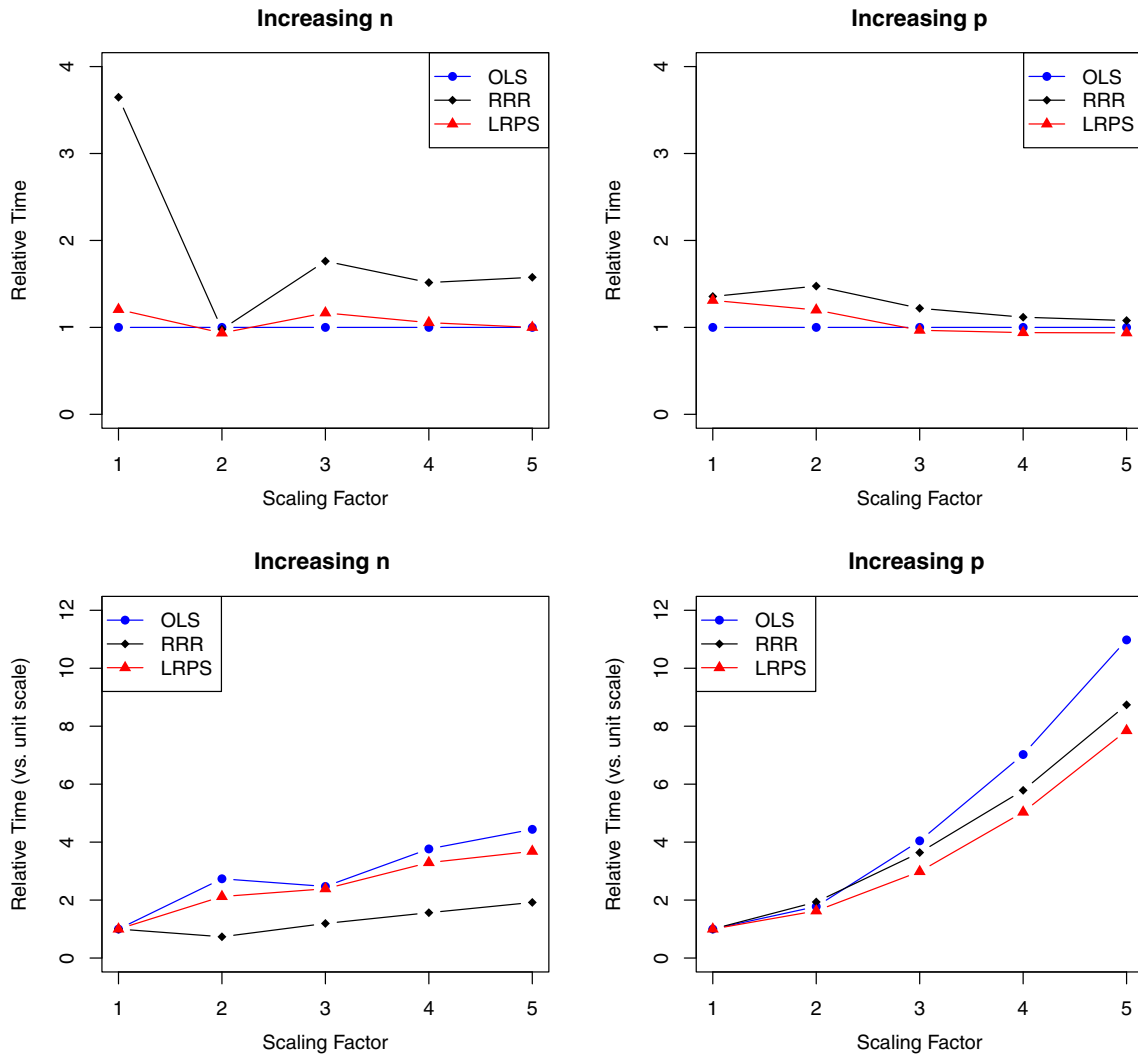
In the  $p > q$  case, we fix  $q = 10$  and let either  $n$  or  $q$  increase with the scaling factor, starting from  $n = 200$  and  $p = 20$ . For each setting, we generated simulation datasets and measured the average computation time over 100 repetitions for each method. The top plots in Figure 10 display the average computation time relative to OLS, where the computational time of OLS is normalized to 1 at each scaling level. In other words, the plotted values for RRR and LRPS reflect their runtime as a multiple of the corresponding

OLS runtime. This illustrates the relative computational cost of adopting one of the low-rank regression methods as data dimensionality increases. It is evident that as  $n$  or  $p$  increase, LRPS consistently outperforms RRR in terms of computational time relative to OLS (Figure 10, top-left and top-right panels). When examining the within method computation time relative to the runtime for the initial simulation setup, we observe that there is a mildly worse scaling with increasing  $n$  (Figure 10, bottom-left). Computational cost of LRPS increases at a faster rate than that of RRR. However, when examining the scaling of the methods with increasing  $p$ , whilst all methods have a quadratic scaling behavior, LRPS increases at a slower rate than both OLS and RRR.

#### *The case when $q > p$ :*

In the  $q > p$  case, we always fix  $p = 30$  and let either  $n$  or  $q$  increase with the scaling factor, starting from  $n = 60$  and  $q = 100$ . Similar to above, for each setting and each method, we calculate the average computation time over 100 simulated datasets. The top plots in Figure 11 display the average computation time relative to OLS. Notably, as  $n$  or  $q$  increases, LRPS consistently outperforms RRR in terms of computational efficiency, demonstrating faster runtimes relative to OLS. The bottom plots show the computational cost of each method relative to its own runtime at the scaling factor of 1. From the plots, we observe that the growth in computation time for LRPS is consistently lower than that of RRR, indicating better scalability with respect to increasing data dimension. In particular, both the OLS and LRPS illustrate linear growth with  $q$ , whereas RRR has a mild quadratic scaling. These simulations confirm the observations in Section 2.4, where RRR was shown to have additional computational complexity compared to LRPS.

We also recorded raw empirical runtime and memory usage of LRPS when compared to other multiresponse regression methods. More specifically, we evaluate the OLS, RRR, LRPS and PCR techniques investigated in this section in a range of regression scenarios. The raw wall clock computation times are shown in Table B1 in Appendix B for when  $q$  is fixed, examining performance as  $n$  or  $p$  increases, whilst Table B2 fixes  $p$  and examines performance as  $n$  or  $q$  increases. For the first setting, LRPS shows a slower increase in runtime compared to other methods



**FIGURE 10** | Top panels: Computation time of RRR and LRPS relative to OLS across varying data dimensions, where OLS time is normalized to 1 at each scaling level. Bottom panels: Computation time of OLS, RRR and LRPS relative to their own runtime at scaling factor 1 under varying data dimensions. Left: Varying  $n$  with  $q = 10$  and  $p = 20$ ; Right: Varying  $p$  with  $n = 200$  and  $q = 10$ .

as the sample size increases; as  $p$  increases all methods are comparable. When  $p$  is fixed, PCR appears the most stable as  $n$  increases, whereas when  $q$  increases LRPS and PCR are comparable. Tables B3 and B4 report the corresponding memory consumption for experiments examined in Tables B1 and B2; the simulations indicate that all methods show stable memory usage across different regression problems.

## 4 | Application to Real Data

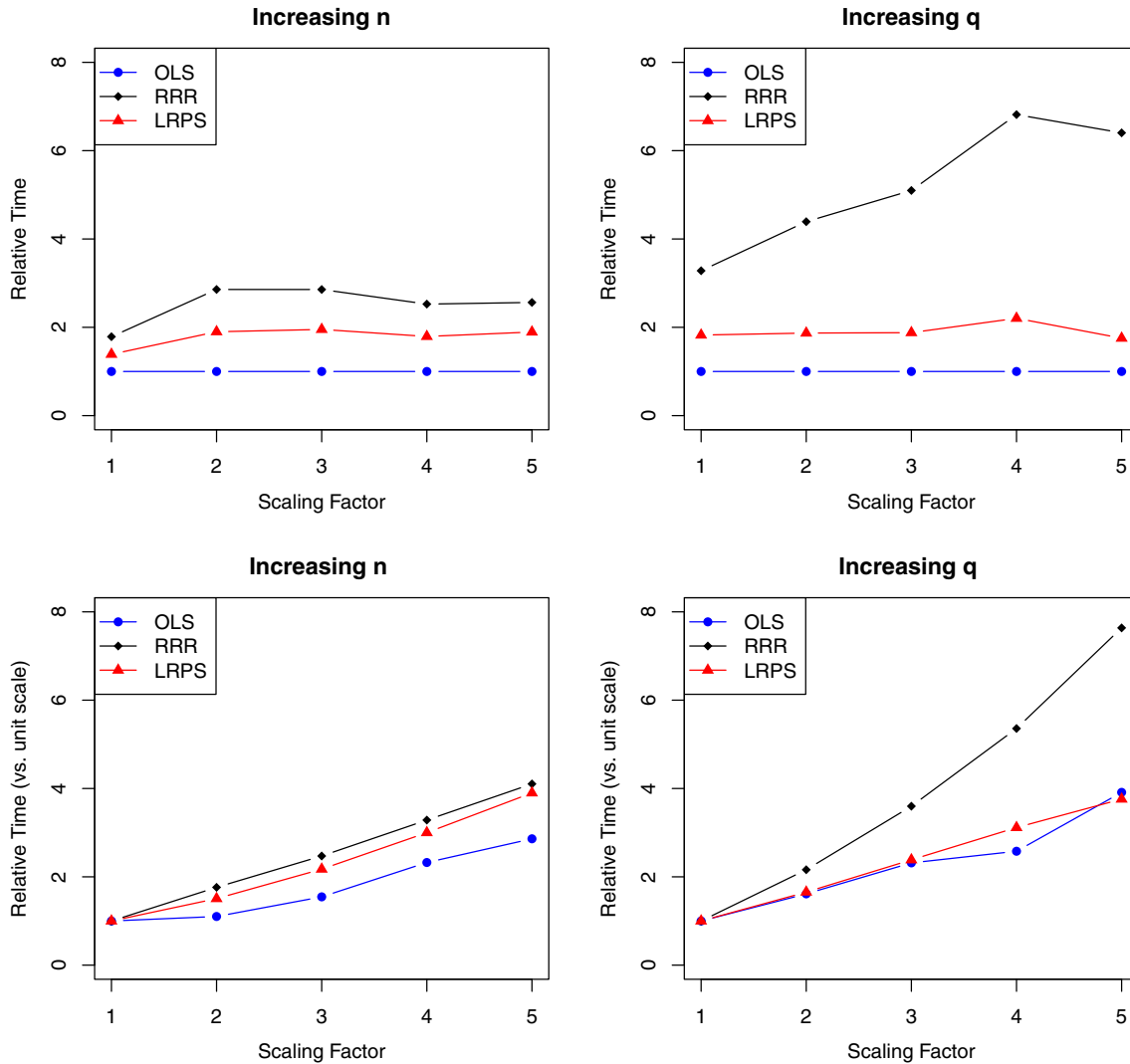
In this section, we demonstrate our proposed LRPS method to study datasets arising from two scientific applications, the first considering air pollution and the second gene activation data.

### 4.1 | Air Pollution Example

Particulate matter up to  $2.5 \mu m$  (PM2.5), is a type of air pollution that poses a significant threat to human health and the environment. These tiny particles are made up of a complex mixture of

solids, chemicals, and liquid droplets that can penetrate the respiratory system and cause respiratory and cardiovascular problems. It is recognized that controlling such particular matters is of vital importance for the public health [40, 41]. In this example, we consider the possible association between PM2.5 emissions and recordings of other pollutant gases, namely: ozone ( $O_3$ ), sulfur dioxide ( $SO_2$ ), carbon monoxide (CO), and nitrogen dioxide ( $NO_2$ ). To study this association, we consider three datasets:

- *UK*: Pollution across seven cities ( $p = 7$ ) in the United Kingdom (UK). There are daily recordings of the pollutants for a total of  $n = 400$  data-points,  $q = 28$  (the number of sites  $\times$  number of polluting gases). This data is available from <https://uk-air.defra.gov.uk/>. We use the first 80% observations  $n_{\text{train}} = 320$  data points to train the model, and the remaining 20% observations  $n_{\text{test}} = 80$  for evaluation.
- *Beijing*: This multi-site air-quality dataset [25] can be found at the UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets> and has been used in the literature in different contexts, for example, Ma et al. [42]. In



**FIGURE 11** | Top panels: Computation time of RRR and LRPS relative to OLS across varying data dimensions. OLS time is normalized to 1 at each scaling level. Bottom panels: Computation time of OLS, RRR and LRPS relative to their own runtime at scaling factor 1 under varying data dimensions. Left: Varying  $n$  with  $p = 30$  and  $q = 100$ ; Right: Varying  $q$  with  $n = 60$  and  $p = 30$ .

this case, we have hourly air pollutant measurements from  $p = 12$  outdoor monitoring sites in Beijing, collected from March 2013 to February 2017. We focus on the air pollutant measurements for January 2017, resulting in a dataset with  $n = 743$  ( $n_{\text{train}} = 594$ ,  $n_{\text{test}} = 149$ ) observations and  $q = 48$  outcomes. It should be noted that this dataset contains some missing values, which for simplicity we replace by the value of the previous observation in the missing variable.

- *USA*: The last dataset is collected from  $p = 37$  monitoring sites in the United States of America (USA) from January 2017 to April 2019, resulting in a total of  $n = 728$  ( $n_{\text{train}} = 578$ ,  $n_{\text{test}} = 150$ ) observations (available from <https://www.epa.gov/outdoor-air-quality-data>). This is the largest dataset in this study, with  $q = 148$  outcomes.

For each dataset we take first-differences (across rows of  $Y, X$ ) and standardize all variables so they have mean zero and standard deviation 1. This preprocessing follows established practice in previous studies [42, 43] and is adopted to promote stationarity

of the data. In each case, we estimate the regression parameters  $B$  on the training dataset, and tune  $k$  via the 2-fold cross-validation procedure employed previously. The remaining  $n_{\text{test}}$  data points (20%) are used for evaluation via the mean-square prediction error

$$\text{MSPE} = \frac{1}{qn_{\text{test}}} \left\| X_{\text{test}} \hat{B} - Y_{\text{test}} \right\|_F^2.$$

The results of the prediction errors and the rank  $k$  used in the low-rank pre-smoothing model are presented in Table 2.

In all cases, LRPS provides better predictive performance than the more traditional approaches. Generally, LRPS and RRR outperform OLS. Whilst the differences in performance are not vast, the results demonstrate that low-rank approximations may generally be useful, and it is interesting that in these cases the rank selected is always lower than  $\min(p, q)$ . In the case of the USA, we see that the rank selected is  $k = 1, 2$  in RRR and LRPS respectively. This behavior was seen in our synthetic experiments in the high  $q$  setting, for example, see *B* Condition 2,  $\Sigma$  Condition 1 results in Figure 4.

**TABLE 2** | Prediction error and optimal rank for the pollution datasets described in the main text.

Dataset	Method	Prediction	
		error (MSPE)	Rank
UK ( $p = 7, q = 28$ )	LRPS	<b>0.738</b>	5
	RRR	0.749	6
	OLS	0.751	N/A
Beijing ( $p = 12, q = 48$ )	LRPS	<b>1.145</b>	12
	RRR	1.154	6
	OLS	1.175	N/A
USA ( $p = 37, q = 148$ )	LRPS	<b>0.911</b>	2
	RRR	0.916	1
	OLS	0.925	N/A

Note: The bold values indicate the best performing methods in each setting.

## 4.2 | Gene Association Example

In addition to the air pollution datasets, we applied our proposed methodology to a genetic association dataset from Wille et al. [24]. This dataset is from a microarray experiment designed to explore regulatory mechanisms within the isoprenoid gene network of *Arabidopsis thaliana*, also known as *thale cress* or *mouse-ear cress*. Isoprenoids are known to play numerous essential biochemical roles in plants. To track gene expression levels,  $n = 118$  GeneChip microarray experiments were conducted. The dataset includes  $p = 39$  predictor genes from two isoprenoid biosynthesis pathways, MVA and MEP, while the response variables consist of the expression of  $q = 795$  genes across 56 metabolic pathways. To reduce skewness in the distributions, all variables are log-transformed before fitting the linear models. We use the first  $n_{\text{train}} = 96$  observations as the training set, and then calculate the prediction error using the remaining  $n_{\text{test}} = 22$  observations.

The results are presented in Table 3, and demonstrate the benefits of the low-rank approximation similar to the air-pollution dataset. In this case, we have a very large  $q$  compared to  $p$ , and both are relatively large when compared to the number of data-points  $n$ . It is worth remarking that our holdout dataset in this case is only  $n_{\text{test}} = 22$  observations, however, the performance is also averaged across the number of outcomes  $q = 795$  which is relatively large. The analysis here can be considered relatively “high-dimensional” in nature. Both LRPS and RRR choose  $k = 2$  as the best approximating rank, potentially illustrating behavior akin to  $B$  Condition 1/2 and  $\Sigma$  Condition 1 in our synthetic experiments, that is, where the error spectrum was relatively flat, but the signal was concentrated in terms of the top singular values. Whilst the difference between LRPS and RRR is not large, LRPS maintains the better performance (in-line with the synthetic experiments and the other application), and both low-rank methods considerably outperform OLS.

## 4.3 | Sensitivity to Rank $k$

Figure 4 presents a sensitivity analysis of the rank parameter  $k$  with respect to predictive performance for both the LRPS and RRR methods. In datasets with a small number of observations,

**TABLE 3** | Prediction error and optimal rank for the genetic association dataset described in the main text.

Method	Prediction error (MSPE)	Rank
LRPS	<b>0.414</b>	2
RRR	0.418	2
OLS	0.481	N/A

Note: The bold values indicate the best performing methods in each setting.

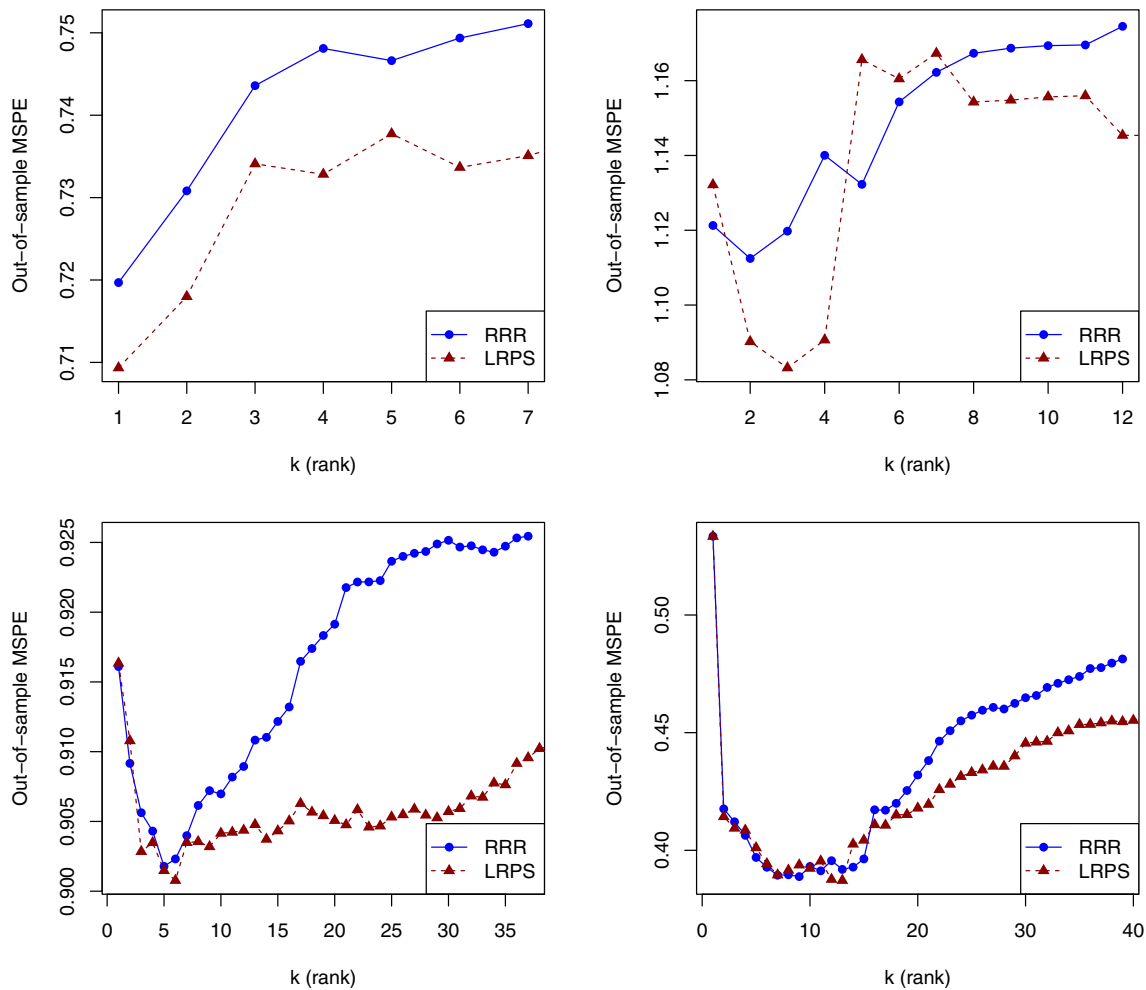
such as the UK dataset, the optimal performance may occur at very low ranks (e.g.,  $k = 1$ ), yet cross-validation can tend to over-estimate the rank in this regime. In contrast, for larger datasets, cross-validation is no longer constrained to selecting a rank of 1 and instead is able to identify higher-rank structures that improve performance for the application (Figure 12).

## 5 | Conclusions

This article proposes *Low-Rank Pre-Smoothing*, a technique for multi-response regression settings. When viewed as a projection, the method can be seen as an initial step to smooth observed data to increase its signal-to-noise ratio, after which a traditional estimation method such as ordinary least squares (OLS) can be performed for estimation and subsequent prediction. Our technique outperforms the traditional OLS estimator in terms of mean squared estimation and out-of-sample prediction error, driven by sacrificing a small amount of bias in favor of reducing the estimation uncertainty. This improved performance was demonstrated through extensive simulated examples, our methodology showing particular benefits when a large number of responses is observed and/or when the signal-to-noise ratio is low. We also illustrated the utility of our proposed methodology on a number of real datasets. We believe our proposed LRPS estimator provides a simple low-rank regression method that nicely complements existing methods, and can provide a further multi-response regression alternative to a wide range of practitioners.

Reflecting on the limitations of the technique, we do note that in some circumstances it may indeed be preferable to use an alternative low-rank regression technique, such as reduced rank regression (RRR) or principal component regression (PCR), and in general the best method will depend on the true generative process. In particular, one may consider using an alternative methodology when the signal  $XB$  aligns predominantly in a direction that is not represented by  $V_k$ , as in these cases, whilst the variance may be significantly reduced by LRPS, there is also significant bias. On the other hand, when  $q$  is modest, variance savings may be limited by LRPS, and RRR or PCR may be more beneficial as they focus more closely on the space spanned by the design matrix. In practice, one can choose between the alternative methods by assessing out-of-sample prediction error via cross-validation procedures, which could in principle be extended to choose between low-rank regression methods, alongside an appropriate choice of  $k$ .

Whilst the motivation for the work in this article arose from parameter estimation and prediction in the classical multi-response setting, we envisage that the tools developed here have the potential to make similar gains in other settings,



**FIGURE 12** | Sensitivity plots of out-of-sample prediction risk versus  $k$  for LRPS and RRR on real datasets. Top-left: UK dataset; Top-right: Beijing dataset; Bottom-left: US dataset; Bottom-right: Genetic dataset.

such as model selection in multi-response regression, nonlinear regression models, and time series forecasting. In particular, one non-trivial extension to the LRPS technique would be in the high-dimensional data setting, which would require development of penalized (LASSO and ridge) regression techniques using LRPS.

#### Author Contributions

All authors have contributed equally to this work.

#### Acknowledgments

The authors have nothing to report.

#### Funding

The authors have nothing to report.

#### Conflicts of Interest

The authors declare no conflicts of interest.

#### Data Availability Statement

The datasets used to support the work in this article are available from the links provided in the text of Section 4.

#### Endnotes

<sup>1</sup>Note that our approximate statement here misuses the asymptotic properties in the non-asymptotic setting. A more precise statement could potentially be obtained via application of Berry-Esseen type limit theorems on the sample eigenvectors; however, this requires further technical assumptions and we prefer to leave finite-sample inference for future work.

<sup>2</sup>The power of the cubic term ( $p^3$ ) in the expression for the complexity can be reduced to  $p^\gamma$ , for some  $\gamma < 3$ , depending on the algorithm used to compute the inverse of  $S_X$  (see e.g., [36]), but this is not crucial to our discussion here.

<sup>3</sup>Note that the setup for the theorem in [31] assumes independent and identically distributed data  $Z_1, \dots, Z_n$ . However, the arguments in the proof, predominantly borrowed from [46], only require the root- $n$  consistency of  $\hat{\Omega}$  as an estimator for  $\Omega$ . Thus the theorem applies under the more general setting of independent, but non-identically distributed data as in the regression model (1).

#### References

1. M. J. Anderson and A. J. Underwood, "Effects of Gastropod Grazers on Recruitment and Succession of an Estuarine Assemblage: A Multivariate and Univariate Approach," *Oecologia* 109 (1997): 442–453.
2. S. Sivakumaran, F. Agakov, E. Theodoratou, et al., "Abundant Pleiotropy in Human Complex Diseases and Traits," *American Journal of Human Genetics* 89, no. 5 (2011): 607–618.

3. J. Liu and V. D. Calhoun, "A Review of Multivariate Analyses in Imaging Genetics," *Frontiers in Neuroinformatics* 8 (2014): 29.
4. B. Skagerberg, J. F. MacGregor, and C. Kiparissides, "Multivariate Data Analysis Applied to Low-Density Polyethylene Reactors," *Chemometrics and Intelligent Laboratory Systems* 14, no. 1–3 (1992): 341–356.
5. A. J. Burnham, J. F. MacGregor, and R. Viveros, "Latent Variable Multivariate Regression Modeling," *Chemometrics and Intelligent Laboratory Systems* 48, no. 2 (1999): 167–180.
6. S. Kim, K.-A. Sohn, and E. P. Xing, "A Multivariate Regression Approach to Association Analysis of a Quantitative Trait Network," *Bioinformatics* 25, no. 12 (2009): i204–i212.
7. S. Bonnini and M. Borghesi, "Relationship Between Mental Health and Socio-Economic, Demographic and Environmental Factors in the COVID-19 Lockdown Period—A Multivariate Regression Analysis," *Mathematics* 10, no. 18 (2022): 3237.
8. N. Gospodinov, R. Kan, and C. Robotti, "Spurious Inference in Reduced-Rank Asset-Pricing Models," *Econometrica* 85, no. 5 (2017): 1613–1628.
9. L. Petrella and V. Raponi, "Joint Estimation of Conditional Quantiles in Multivariate Linear Regression Models With an Application to Financial Distress," *Journal of Multivariate Analysis* 173 (2019): 70–84.
10. M. S. Srivastava and T. K. S. Solanky, "Predicting Multivariate Response in Linear Regression Model," *Communications in Statistics: Simulation and Computation* 32, no. 2 (2003): 389–409.
11. A. J. Izenman, *Modern Multivariate Statistical Techniques*, vol. 1 (Springer, 2008).
12. T. W. Anderson, *Introduction to Multivariate Statistical Analysis*, 3rd ed. (Wiley, 2003).
13. R. A. Johnson, D. W. Wichern, and R. Johnson, *Applied Multivariate Statistical Analysis*, 6th ed. (Prentice Hall, 2007).
14. P. Faraldo Roca and W. González Manteiga, "On Efficiency of a New Class of Linear Regression Estimates Obtained by Preliminary Non-Parametric Estimation," in *New Perspectives in Theoretical and Applied Statistics* (Springer, 1987), 229–242.
15. J. A. Cristobal Cristobal, P. Faraldo Roca, and W. Gonzalez Manteiga, "A Class of Linear Regression Parameter Estimators Constructed by Nonparametric Estimation," *Annals of Statistics* 15, no. 2 (1987): 603–609.
16. P. Janssen, J. Swanepoel, and N. Veraverbeke, "Efficiency of Linear Regression Estimators Based on Presmoothing," *Communications in Statistics – Theory and Methods* 30, no. 10 (2001): 2079–2097.
17. M. Aerts, N. Hens, and J. S. Simonoff, "Model Selection in Regression Based on Pre-Smoothing," *Journal of Applied Statistics* 37, no. 9 (2010): 1455–1472.
18. D. B. Hitchcock, G. Casella, and J. G. Booth, "Improved Estimation of Dissimilarities by Presmoothing Functional Data," *Journal of the American Statistical Association* 101, no. 473 (2006): 211–222.
19. D. B. Hitchcock, J. G. Booth, and G. Casella, "The Effect of Pre-Smoothing Functional Data on Cluster Analysis," *Journal of Statistical Computation and Simulation* 77, no. 12 (2007): 1043–1055.
20. F. Ferraty, W. González Manteiga, A. Martínez Calvo, and P. Vieu, "Presmoothing in Functional Linear Regression," *Statistica Sinica* 22 (2012): 69–94.
21. E. Musta, V. Patilea, and I. Van Keilegom, "A Presmoothing Approach for Estimation in the Semiparametric Cox Mixture Cure Model," *Bernoulli* 28, no. 4 (2022): 2689–2715.
22. L. Tedesco, J. Beyhum, and I. Van Keilegom, "Instrumental Variable Estimation of the Proportional Hazards Model by Presmoothing," *Electronic Journal of Statistics* 19, no. 1 (2025): 656–717.
23. E. Musta, V. Patilea, and I. Van Keilegom, "Regression Estimation Using Surrogate Responses Obtained by Presmoothing," *Statistica Neerlandica* 79, no. 1 (2024): e12351.
24. A. Wille, P. Zimmermann, E. Vranová, et al., "Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in *Arabidopsis thaliana*," *Genome Biology* 5 (2004): 1–13.
25. X. Liang, T. Zou, B. Guo, et al., "Assessing Beijing's PM2.5 Pollution: Severity, Weather Impact, APEC and Winter Heating," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471, no. 2182 (2015): 20150257.
26. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis. Probability and Mathematical Statistics*, 10th ed. (Academic Press, 1995).
27. W. H. Greene, *Econometric Analysis*, 8th ed. (Pearson Education India, 2018).
28. I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (2016): 20150202.
29. A. J. Izenman, "Reduced-Rank Regression for the Multivariate Linear Model," *Journal of Multivariate Analysis* 5, no. 2 (1975): 248–264.
30. H. Neudecker and A. M. Wesselman, "The Asymptotic Variance Matrix of the Sample Correlation Matrix," *Linear Algebra and Its Applications* 127 (1990): 589–599.
31. E. Bura and R. Pfeiffer, "On the Distribution of the Left Singular Vectors of a Random Matrix and Its Applications," *Statistics & Probability Letters* 78, no. 15 (2008): 2275–2280.
32. G. Reinsel, R. Velu, and K. Chen, *Multivariate Reduced-Rank Regression: Theory and Applications. Lecture Notes in Statistics* (Springer New York, 2022).
33. W. F. Massy, "Principal Components Regression in Exploratory Statistical Research," *Journal of the American Statistical Association* 60, no. 309 (1965): 234–256.
34. I. Jolliffe, "Principal Component Analysis," in *International Encyclopedia of Statistical Science* (Springer, 2011), 1094–1096.
35. W. James and C. Stein, "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (University of California Press, 1961), 361–379.
36. Z. Tonks, G. Sankaran, and J. Davenport, "Fast Matrix Operations in Computer Algebra," in *2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (Springer, 2017), 67–70.
37. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2024), <https://www.R-project.org/>.
38. A. Genz and F. Bretz, *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics* (Springer-Verlag, 2009).
39. S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *Journal of the American Statistical Association* 119, no. 546 (2024): 1434–1445.
40. U. Franck, S. Odeh, A. Wiedensohler, B. Wehner, and O. Herbarth, "The Effect of Particle Size on Cardiovascular Disorders-The Smaller the Worse," *Science of the Total Environment* 409, no. 20 (2011): 4217–4221.
41. V. C. Pun, F. Kazemiparkouhi, J. Manjourides, and H. H. Suh, "Long-Term PM2.5 Exposure and Respiratory, Cancer, and Cardiovascular Mortality in Older US Adults," *American Journal of Epidemiology* 186, no. 8 (2017): 961–969.
42. X. Ma, Q. Zhou, and X. Zi, "Multiple Change Points Detection in High-Dimensional Multivariate Regression," *Journal of Systems Science and Complexity* 35, no. 6 (2022): 2278–2301.

43. C. Zou, Y. Ke, and W. Zhang, "Estimation of Low Rank High-Dimensional Multivariate Linear Models for Multi-Response Data," *Journal of the American Statistical Association* 117, no. 538 (2022): 693–703.

44. K. M. Abadir and J. R. Magnus, *Matrix Algebra*, vol. 1 (Cambridge University Press, 2005).

45. H. Cramér, *Mathematical Methods of Statistics*, vol. 26 (Princeton University Press, 1999).

46. Z. Ratsimalahelo, "Rank Test Based on Matrix Perturbation Theory," Technical Report, EERI Research Paper Series, 2001.

47. P. Billingsley, *Convergence of Probability Measures*, 2nd ed. (John Wiley & Sons, 2013).

## Appendix A

### Proof of Theoretical Results

In this appendix, we provide details of the proof of Theorem 1 and Proposition 1 described in the main text.

### Proof of Theorem 1

We begin by establishing a result which be useful in the proof of the theorem. Let  $Y$  be a random data matrix from model (1) (where the fourth order moments of the errors are finite). Note that in what follows, we assume without loss of generality that  $X$  and  $Y$  are mean-centered, so that  $S_y = n^{-1}Y^T Y$  is the sample covariance matrix, and define  $\Sigma_y = n^{-1}(XB)^T(XB) + \Sigma_e$ .

**Lemma 1.** *Let  $S_y$  and  $\Sigma_y$  be as above. Then*

$$\sqrt{n}\text{vec}(S_y - \Sigma_y) \xrightarrow{D} \mathcal{N}_{q^2}(0, \mathbf{V}_y), \quad (\text{A1})$$

where

$$\mathbf{V}_y = \mathbb{E}[ee^T \otimes ee^T] - \text{vec}(\Sigma_e) \otimes \text{vec}(\Sigma_e)^T, \quad (\text{A2})$$

in which  $e \in \mathbb{R}^{q \times 1}$  denotes a generic row of  $E$ .

*Proof.* Denote by  $S_e = n^{-1}E^T E$  the sample covariance of the error vectors  $E$  with corresponding population covariance matrix  $\Sigma_e$ . Since the fourth order moments of the error distribution are assumed finite and the rows of  $E$  are assumed to be independent and identically distributed under model (1), theorem 1 in [30] (with zero mean random variables) establishes that  $S_e$  is a root- $n$  consistent estimator for  $\Sigma_e$ :

$$\sqrt{n}\text{vec}(S_e - \Sigma_e) \xrightarrow{D} \mathcal{N}_{q^2}(0, \mathbf{V}_y), \quad (\text{A3})$$

where and  $e$  is as above.

Let  $Z := n^{-1}(XB)^T(XB) = n^{-1}B^T(X^T X)B \in \mathbb{R}^{q \times q}$  and  $W := n^{-1}(XB)^T E \in \mathbb{R}^{q \times q}$ . Using the linear model (1), the sample covariance matrix of  $Y$ ,  $S_y$ , can be expressed as

$$S_y = n^{-1}Y^T Y = n^{-1}((XB)^T(XB) + (XB)^T E + E^T(XB) + E^T E). \quad (\text{A4})$$

Then

$$\Sigma_y := \mathbb{E}(S_y) = n^{-1}((XB)^T(XB)) + n^{-1}\mathbb{E}((XB)^T E) + n^{-1}\mathbb{E}(E^T(XB)) + \Sigma_e.$$

Because  $\mathbb{E}(E) = 0$  (and thus  $\mathbb{E}(W) = 0$ ), this simplifies to

$$\Sigma_y = n^{-1}(XB)^T(XB) + \Sigma_e. \quad (\text{A5})$$

Then combining (A4) and (A5), we get

$$\begin{aligned} \underbrace{\sqrt{n}\text{vec}(S_y - \Sigma_y)}_C &= \sqrt{n}\text{vec}((Z + W + W^T + S_e) - \Sigma_y) \\ &= \sqrt{n}\text{vec}((Z + W + W^T + S_e) - (Z + \Sigma_e)) \\ &= \underbrace{\sqrt{n}\text{vec}((W + W^T))}_{(A_1)} + \underbrace{\sqrt{n}\text{vec}(S_e - \Sigma_e)}_{(A_2)}. \end{aligned} \quad (\text{A6})$$

Using the properties of the vec operator [44],

$$\text{vec}(W) = \text{vec}(n^{-1}(XB)^T E) = \underbrace{(I_q \otimes n^{-1}(XB)^T)}_A \text{vec}(E),$$

where  $\otimes$  denotes the Kronecker product, so

$$\begin{aligned} \text{Var}(\text{vec}(W)) &= \text{Var}(A\text{vec}(E)) = A\text{Var}(\text{vec}(E))A^T \\ &= (I_q \otimes n^{-1}(XB)^T)(\Sigma_e \otimes I_n)(I_q \otimes n^{-1}(XB)^T)^T \\ &= (I_q \Sigma_e I_q) \otimes (n^{-2}(XB)^T I_n (XB)) \\ &= \Sigma_e \otimes n^{-1}Z, \end{aligned}$$

using the transpose and mixed-product properties of the Kronecker product. Assuming that (with mean-centered  $X$ )  $\lim_{n \rightarrow \infty} S_X = \Sigma_x$  (where  $S_X = n^{-1}X^T X$ ) for some fixed positive definite matrix  $\Sigma_x$  (as is customary in regression), we have

$$\mathbb{E}(\text{vec}(W)) = 0 \quad \text{and}$$

$$\text{Var}(\text{vec}(W)) = \Sigma_e \otimes n^{-1}Z = (\Sigma_e \otimes n^{-1}B^T S_X B) \rightarrow 0.$$

Hence considering the first term  $A_1$  in (A6), we have

$$\text{vec}((W + W^T)) = (\text{vec}(W)) + (\text{vec}(W^T)) \xrightarrow{P} 0,$$

thus considering (A6) and using (the multivariate version of) Slutsky's theorem (see e.g., [45, section 20.6]), we have

$$\sqrt{n}\text{vec}(S_y - \Sigma_y) \xrightarrow{D} \mathcal{N}_{q^2}(0, \mathbf{V}_y),$$

with

$$\mathbf{V}_y = \mathbb{E}[ee^T \otimes ee^T] - \text{vec}(\Sigma_e) \otimes \text{vec}(\Sigma_e)^T. \quad \square$$

We now present a theorem and a corollary without proof, both taken from [31], which are used as intermediary results to establish Theorem 1.

**Theorem.** *Let  $\hat{\Omega}$  be a  $p \times q$ -dimensional random matrix that is a function of a random sample of  $p$ -dimensional data  $Z_1, \dots, Z_n$ . Assume that  $\hat{\Omega}_n$  is asymptotically normally distributed as*

$$\sqrt{n}\text{vec}(\hat{\Omega}_n - \Omega) \xrightarrow{D} \mathcal{N}_{pq}(0, \mathbf{V}).$$

*Let  $r = \text{rank}(\Omega)$  with  $d \leq \min(p, q)$ , and let the singular value decomposition of  $\Omega$  be*

$$\Omega = L^T \begin{pmatrix} D & 0 \\ \mathbf{0} & 0 \end{pmatrix} R,$$

where  $L^T = (L_1, L_0)$  is of order  $p \times p$  with  $L_1 : p \times r$ ,  $L_0 : p \times (p-r)$ .  $R^T = (R_1, R_0)$  is orthogonal with  $R_1 : q \times r$ ,  $R_0 : q \times (q-r)$ . The  $r$  left singular vectors  $U_1$  of  $\Omega$  correspond to its  $r$  non-zero singular values,  $d_1 \geq \dots \geq d_r$ . Analogously, the SVD of  $\hat{\Omega}$  is

$$\hat{\Omega} = \hat{L}^T \begin{pmatrix} \hat{D}_1 & \mathbf{0} \\ \mathbf{0} & \hat{D}_0 \end{pmatrix} \hat{R}, \quad (\text{A7})$$

with  $\widehat{L}^\top = (\widehat{L}_1, \widehat{L}_0)$ ,  $\widehat{R}^\top = (\widehat{R}_1, \widehat{R}_0)$ , and the partition conforms with the partition of  $\Omega$ .

Assume that  $\lambda_r > \epsilon > 0$ , that is, the minimum positive singular value of  $\Omega$  is well separated from zero. Then as  $n \rightarrow \infty$

a.  $\widehat{D}_1 \xrightarrow{P} D_1$  and  $\widehat{D}_0 \xrightarrow{P} 0$ .

b.  $\widehat{R}_1 \xrightarrow{P} R_1$  and  $\widehat{L}_1 \xrightarrow{P} L_1$ .

**Corollary.** The matrix  $\widehat{L}_1$  of left singular vectors in (A7) is asymptotically normal with

$$n^{1/2} \text{vec}(\widehat{L}_1 - L_1) \xrightarrow{D} N_{pd}(\mathbf{0}, \Sigma_L), \quad (\text{A8})$$

where

$$\Sigma_L = (D^{-1} R_1^\top \otimes I_p) \mathbf{V} (R_1 D^{-1} \otimes I_p). \quad (\text{A9})$$

The assumption on the minimum positive singular value ensures that  $\Lambda^{-1}$  is bounded, necessary for the proof of the theorem (see [31]).

We now apply the theorem and corollary to our setting.

Let  $V \Lambda V^\top$  be the compact form SVD of the quantity  $\Sigma_y$  from Lemma 1 (equivalent to its eigendecomposition since it is positive semi-definite), and suppose  $\Sigma_y$  has rank  $r$ . Suppose also that the random matrix  $Y$  has the compact form decomposition  $Y = \widehat{U} \widehat{D} \widehat{U}^\top$ , where the diagonal entries of  $\widehat{D}$  are  $\{\widehat{d}_1, \dots, \widehat{d}_r\}$ . Then  $S_y$  can be written as

$$n^{-1} Y^\top Y = n^{-1} \widehat{V} \widehat{D} \widehat{U}^\top \widehat{V} \widehat{D} \widehat{U}^\top = \widehat{V} \widehat{\Lambda} \widehat{V}^\top,$$

due to the orthogonality of  $\widehat{U}$ , and where  $\widehat{\Lambda} = \text{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_r)$  is a diagonal matrix containing the descending eigenvalues  $\widehat{\lambda}_i = \widehat{d}_i^2/n, i = 1, \dots, r$ .

Now using the result established in Lemma 1, the asymptotic normality assumption of the theorem is satisfied for  $S_y$  as an estimator of  $\Sigma_y$ ,<sup>3</sup> hence  $\widehat{\Lambda} \xrightarrow{P} \Lambda$ , and the corollary gives that the matrix  $\widehat{V}$  of eigenvectors of  $S_y$  are asymptotically normal with

$$n^{1/2} \text{vec}(\widehat{V} - V) \xrightarrow{D} N_{qr}(\mathbf{0}, \Sigma_{\widehat{V}}) \quad \text{where} \\ \Sigma_{\widehat{V}} = (\Lambda^{-1} V^\top \otimes I_q) \mathbf{V}_y (V \Lambda^{-1} \otimes I_q).$$

Hence considering the subset of first  $k$  eigenvectors and corresponding eigenvalues ( $k < r$ ), we have that

$$n^{1/2} \text{vec}(\widehat{V}_k - V_k) \xrightarrow{D} N_{qk}(\mathbf{0}, \Sigma_{\widehat{V}_k}), \quad (\text{A10})$$

with

$$\Sigma_{\widehat{V}_k} = (\Lambda_k^{-1} V_k^\top \otimes I_q) \mathbf{V}_y (V_k \Lambda_k^{-1} \otimes I_q),$$

where  $\widehat{V}_k, V_k$  and  $\Lambda_k$  correspond to the quantities above, curtailed to  $k$  eigenvectors/eigenvalues.

In view of the relationship between the LRPS estimator and the OLS estimator  $\widehat{B} \widehat{V}_k \widehat{V}_k^\top$ , we have  $\text{vec}(\widehat{B} \widehat{V}_k \widehat{V}_k^\top) = (\widehat{V}_k \widehat{V}_k^\top \otimes I_p) \text{vec}(\widehat{B})$ , and the corresponding expression with the true eigenvectors also holds.

Now considering the quantity  $\text{vec}(\widehat{V}_k \widehat{V}_k^\top)$ , due to the consistency result for  $\widehat{V}_k$  in (A10), the continuous mapping theorem (see, e.g., Reference [47]) then means that  $\widehat{V}_k \widehat{V}_k^\top$  is a consistent estimator of  $V_k V_k^\top$ , that is,  $\widehat{V}_k \widehat{V}_k^\top \xrightarrow{P} V_k V_k^\top$  (see also the proof of theorem 1 in [31]). Using the distributional property (3) and applying Slutsky's theorem gives that

$$n^{1/2} \text{vec}(\widehat{B} - B V_k V_k^\top) = n^{1/2} \text{vec}(\widehat{B}(\widehat{V}_k \widehat{V}_k^\top - V_k V_k^\top) + (\widehat{B} - B) V_k V_k^\top) \\ \xrightarrow{D} \mathcal{N}(\mathbf{0}, V_k V_k^\top \Sigma_e V_k V_k^\top \otimes S_X^{-1}),$$

where the variance term is derived as follows:

$$\text{Var}((V_k V_k^\top \otimes I_p) \text{vec}(n^{1/2} \widehat{B})) = (V_k V_k^\top \otimes I_p) n \text{Var}(\text{vec}(\widehat{B})) (V_k V_k^\top \otimes I_p)^\top \\ = (V_k V_k^\top \otimes I_p) (\Sigma_e \otimes S_X^{-1}) (V_k V_k^\top \otimes I_p) \\ = V_k V_k^\top \Sigma_e V_k V_k^\top \otimes S_X^{-1},$$

using the mixed-product property of Kronecker products.

Note that this result will also be valid if the distributional result (3) is asymptotic, as in the case when the error distribution  $G$  in the regression model (1) is not Gaussian.

### Proof of Proposition 1

The optimal computational complexity of the LRPS estimator will be different depending on whether  $q > p$  or  $p > q$ , as seen below.

*The case when  $p > q$ :*

The LRPS estimator can be computed as follows:

$$\widehat{B} = \underbrace{(X^\top X)^{-1}}_{O(np^2+p^3)} \underbrace{X^\top U_k}_{O(npk)} \underbrace{D_k V_k^\top}_{O(qk)}, \\ \underbrace{\hspace{10em}}_{O(p^2k)} \\ \underbrace{\hspace{10em}}_{O(pqk)}$$

After performing the truncated SVD of  $Y$ , the operations required to compute the LRPS estimator are: projection, low-rank reconstruction, and two matrix multiplications, contributing  $O(npk)$ ,  $O(qk)$ ,  $O(p^2k)$  and  $O(pqk)$  respectively, where we note that the matrix multiplication with  $D_k$  can be done efficiently since it is diagonal. This leads to a total complexity of

$$O(np^2 + p^3) + O(nqk) + O(npk) + O(qk) + O(p^2k) + O(pqk).$$

*The case when  $q > p$ :*

Similarly, in this case the LRPS estimator can be calculated as follows:

$$\widehat{B} = \underbrace{(X^\top X)^{-1}}_{O(np^2+p^3)} \underbrace{X^\top U_k}_{O(npk)} \underbrace{D_k V_k^\top}_{O(qk)}, \\ \underbrace{\hspace{10em}}_{O(p^2k)} \\ \underbrace{\hspace{10em}}_{O(pk)} \\ \underbrace{\hspace{10em}}_{O(pqk)}$$

leading to a total complexity of

$$O(np^2 + p^3) + O(nqk) + O(npk) + O(pk) + O(p^2k) + O(pqk).$$

## Appendix B

### Additional Computational Comparisons

In this section, we present additional simulations to assess runtimes and memory usage for different multiresponse regression methods as outlined in the main text. Below we report the wall clock computation times for fixed  $q$  and fixed  $p$  (Tables B1 and B2) and the memory consumption (Tables B3 and B4), where  $S1 - S5$  indicates a scaling factor to increase the size of the regression problem by the dimension ( $n, p$  or  $q$ ) under consideration (see Section 3.4).

**TABLE B1** | Average runtime (s) for regression methods described in the text for different dimension sizes, for fixed  $q$ .

Method	Increasing $n$ runtime ( $\times 10^{-3}$ s)					Increasing $p$ runtime ( $\times 10^{-3}$ s)				
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
OLS	0.15	0.24	0.27	0.44	0.59	0.15	0.31	0.58	0.99	1.59
RRR	0.27	0.47	0.58	0.79	0.91	0.25	0.41	0.75	1.12	1.69
LRPS	0.69	0.36	0.39	0.48	0.54	0.23	0.34	0.98	0.94	1.88
PCR	0.46	0.71	0.87	1.42	1.25	0.35	0.76	1.76	1.34	1.94

Note: Left: Increasing  $n$ ; Right: Increasing  $p$ .

**TABLE B2** | Average runtime (s) for regression methods described in the text for different dimension sizes, for fixed  $p$ .

Method	Increasing $n$ runtime ( $\times 10^{-3}$ s)					Increasing $q$ runtime ( $\times 10^{-3}$ s)				
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
OLS	0.27	0.40	0.61	0.84	1.04	0.24	0.28	0.35	0.64	0.71
RRR	0.70	1.06	1.59	2.14	2.71	0.62	1.33	2.11	3.17	4.46
LRPS	0.47	0.68	1.06	1.58	1.92	0.33	0.65	0.84	1.07	1.30
PCR	0.80	0.51	0.62	0.60	0.83	0.37	0.37	0.29	0.33	0.39

Note: Left: Increasing  $n$ ; Right: Increasing  $q$ .

**TABLE B3** | Average memory usage (MB) for multiresponse regression methods described in the text for different dimension sizes, for fixed  $q$ .

Method	Increasing $n$ memory (kB)					Increasing $p$ memory (kB)				
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
OLS	1.28	0.70	0.70	0.70	0.70	0.72	0.72	0.72	0.72	0.72
RRR	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
LRPS	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
PCR	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70

Note: Left: Increasing  $n$ ; Right: Increasing  $p$ .

**TABLE B4** | Average memory usage (MB) for multiresponse regression methods described in the text for different dimension sizes, for fixed  $p$ .

Method	Increasing $n$ memory (kB)					Increasing $q$ memory (kB)				
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
OLS	0.93	0.70	0.70	0.70	0.70	0.93	0.93	0.93	0.93	0.93
RRR	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
LRPS	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
PCR	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70

Note: Left: Increasing  $n$ ; Right: Increasing  $q$ .