JDSSV

Journal of Data Science, Statistics, and Visualisation

May 2025, Volume V, Issue V. *doi: 10.52933/jdssv.v5i5.138*

Sparse Data-Driven Random Projection in Regression for High-Dimensional Data

Roman Parzer
TU Wien

Peter Filzmoser
TU Wien

Laura Vana-Gür
TU Wien

Abstract

We examine the linear regression problem in a challenging high-dimensional setting with correlated predictors where the degree of sparsity of the coefficients is unknown and can vary from sparse to dense. In this setting, we propose a combination of probabilistic variable screening with random projection tools as a computationally efficient approach. In particular, we introduce a new data-driven random projection for dimension reduction in linear regression, which is motivated by a theoretical bound on the gain in expected prediction error over conventional random projections when using information about the true coefficient. The variables to be included in the projection are screened by considering the correlation of the predictors. To reduce the dependence on fine-tuning choices, we aggregate over an ensemble of linear models. A threshold parameter is introduced to obtain a higher degree of sparsity, which can be chosen together with the number of models in the ensemble by cross-validation. In extensive simulations, we compare the proposed method with other random projection tools and with well-known methods, and show that it is competitive in terms of prediction in a variety of scenarios with different sparsity and predictor covariance settings, while most competitors are targeted at either sparse or dense settings. Finally, we illustrate the method on two data applications.

Keywords: High-dimensional regression, dimension reduction, random projection, screening.

1. Introduction

Recent advances in technology have allowed more and more quantities to be tracked and stored, leading to a huge increase in the amount of data, making available datasets more complex and larger than ever, both in dimension and size. We consider a standard linear regression setting, where the response variable is given by

$$y_i = \mu + x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where n is the number of observations, μ is a deterministic intercept, the x_i are iid observations of p-dimensional covariates or predictors with a common covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown parameter vector, and the ε_is are iid error terms with $\mathbb{E}[\varepsilon_i] = 0$ and constant $\mathrm{Var}(\varepsilon_i) = \sigma^2$ independent of the x_is. We are interested in studying the case where $p > n$ or even $p \gg n$.

In this paper, we tackle the challenge of high-dimensional linear regression with correlated covariates by introducing an ensemble method that integrates a novel random projection, specifically designed for the regression problem, together with a variable screening step pre-projection. *Sparse* methods, such as Tibshirani (1996)'s LASSO (ℓ_1 penalized regression), the adaptive LASSO of Zou 2006, Zou and Hastie 2005's elastic net which combines the ℓ_1 penalty with an ℓ_2 penalty to tackle the inability of the LASSO to handle multicollinearity, the ordered weighted ℓ_1-norm regression of Figueiredo and Nowak 2014, which can handle strongly correlated predictors or, alternatively, Bayesian shrinkage priors (see Gruber and Kastner 2023 for a discussion), excel in settings where the true β sparse, while regression techniques such as partial least squares (PLS; Fornell and Cha 1994) achieve their best performance in *dense* settings, i.e., where many predictors contribute information about the response (Cook and Forzani 2019). Our proposed approach, on the other hand, can adapt to various degrees of sparsity in the true coefficients and provides a computationally efficient alternative to both *sparse* and *dense* methods. This flexibility makes the method relevant in real-world scenarios where the level of sparsity is often unknown and may vary significantly across applications. Moreover, the novel data-driven random projection accommodates various types of correlation settings.

Random projection linearly maps a set of points in high dimensions into a much lower-dimensional space. The method is theoretically grounded in the lemma of Johnson and Lindenstrauss (1984) (*JL*), who proved the existence of a linear map that approximately preserves pairwise distances for a set of points in high dimensions in a much lower-dimensional space. Possible applications are low-rank approximations (Clarkson and Woodruff 2013), data reduction for high n (e.g., Geppert et al. 2015; Ahfock et al. 2021), or data privacy (e.g., Zhou et al. 2007), but it has also been applied in the context of regression problems, see e.g., Maillard and Munos (2009), Guhaniyogi and Dunson (2015), Mukhopadhyay and Dunson (2020). However, employed random projections that satisfy the JL lemma are data-agnostic and generally rely on sampling the entries from a (sub)-Gaussian distribution. In this paper, we propose a new random projection better suited for linear regression, which takes the variables' effect on the response into consideration while also accounting for the correlation among the predictors. More specifically, we construct a sparse random projection that uses the high-dimensional ordinary least squares projection (HOLP) estimator of Wang and Leng (2016) in the construction of the matrix. The proposed construction ensures that the information about the predictor-response relationship can be recovered in the high-dimensional space based on the solution learned from the lower-dimensional space. The approach is motivated by a theoretical result, where we provide a theoretical bound on the expected gain in prediction error when using a projection that incorporates information about the true β coefficients compared to a conventional random projection.

The combination of random projection with a variable screening step pre-projection is motivated by the fact that, for very large p, random projection can suffer from noise accumulation, as too many irrelevant predictors are being considered for prediction purposes (Mukhopadhyay and Dunson 2020). To alleviate this issue, we do not project all predictors onto the lower-dimensional space using the proposed random projection matrix, but rather only a subset by using a randomized screening step (similar to Mukhopadhyay and Dunson 2020). In this step, predictors are selected with a probability proportional to their effect on the response conditional on the other predictors. To account for the correlation among the predictors, we rely on the same HOLP estimator used in our proposed random projection as a screening coefficient, which can be efficiently computed and has strong theoretical screening properties.

The proposed method builds an ensemble of linear models using the screened and projected covariates to reduce the variance caused by these two randomness sources (also recommended by Thanei et al. 2017; Guhaniyogi and Dunson 2015). The averaged coefficients over all of the ensemble models can then be used for prediction or interpretation purposes. The method allows for zero coefficients, as some variables may not be included in the models after the probabilistic screening step. To further encourage sparsity in our framework, we introduce a thresholding parameter. This parameter allows us to set any estimated coefficient values in the ensemble smaller than the threshold (in absolute value) to zero before averaging. The number of models in the ensemble and the thresholding parameter can be chosen by cross-validation.

In a broad simulation study with six different covariance structures and three different levels of sparsity, we benchmark this new approach against an extensive collection of existing (sparse and dense) methods and show that it provides the best performance when averaging ranks of prediction ability over all scenarios. While the proposed method is outperformed by sparse techniques like adaptive LASSO or elastic net in sparse settings, it still delivers better predictions than dense methods in these sparse settings, and is among the best methods in all medium to dense settings, making it a suitable choice when the true sparsity level of the problem is unknown. We also show that the version with cross-validation is competitive in terms of ranking the variables according to their impact on the response.

The paper is organized as follows. Section 2 introduces the methodology. An extensive simulation study is presented in Section 3. Section 4 illustrates the proposed method on two real-world datasets, and Section 5 concludes.

2. Methods

In Section 2.1, we introduce variable screening to reduce the dimensionality of predictors and motivate the use of HOLP over other alternatives for this purpose.

In Section 2.2, we propose a random projection tailored to dimension reduction for linear regression and give a theoretical bound on the performance gain in the expected prediction error over using a data-agnostic, conventional random projection. We also show in a simulation example that when estimating the required coefficients using the HOLP estimator, we stay well within the bound and still obtain better predictions than a conventional random projection. Finally, we discuss how to combine these two concepts in Section 2.3 and propose our algorithm in Section 2.4.

The following notation is used throughout the paper. For any integer $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, \dots, n\}$, $I_n \in \mathbb{R}^{n\times n}$ is the n-dimensional identity matrix and $1_n \in \mathbb{R}^n$ is an n-dimensional vector of ones. From model (1), we let $X \in \mathbb{R}^{n\times p}$ be the matrix of predictors with rows $\{x_i \in [n]\}$ and $y = (y_1, \dots, y_n)' \in \mathbb{R}^n$ the response vector.

2.1. Variable screening

The general idea of variable screening is to select a (small) subset of variables, based on some marginal utility measure for the predictors, and disregard the rest for further analysis. In this work, we are interested in screening coefficients that can accommodate correlated predictors, while also being almost proportional to the true regression coefficients. The latter property will prove useful when employing the same screening coefficient in the data-driven random projection matrix proposed in Section 2.3.

Fan and Lv (2008) propose to use the vector of marginal empirical correlations between the response and each predictor for variable screening to select a smaller number of variables (less than n) for subsequent analysis (sure independence screening, SIS). They show *screening consistency* of the estimator for exponential growth of p under the condition that marginal correlations for the important variables must be bounded away from zero. This condition rules out practically possible scenarios where an important variable is marginally uncorrelated to the response. Wang and Leng (2016) relax this assumption and propose the HOLP estimator for screening. Assuming $\text{rank}(XX') = n$ and therefore $p > n$

$$\hat{\beta}_{\text{HOLP}} = X'(XX')^{-1}y = \lim_{\lambda\to 0} \underbrace{X'(\lambda I_n + XX')^{-1}y}_{\text{Ridge estimator}}. \tag{2}$$

which is also the minimum norm solution to $X\beta = y$ (see Lemma 2). It is notable that the HOLP estimator is the limit of the Ridge estimator (Hoerl and Kennard 1970) in the case $p > n$ when $\lambda \to 0$ (see Lemma 1 for the derivation of this alternative form of the Ridge estimator for the case $p > n$). Letting $\lambda \to 0$ is also in line with results in Kobak et al. (2020), who show that the optimal Ridge penalty for minimal mean-squared prediction error can be zero or negative for real-world high-dimensional data because low-variance directions in the predictors can already provide an implicit Ridge regularization. This motivates choosing the absolute values of the tuning-free coefficient vector $\hat{\beta}_{\text{HOLP}}$ for variable screening. Wang et al. (2015) and Wang and Leng (2016) show that HOLP satisfies stronger theoretical screening properties under weaker conditions than SIS.

We now turn to investigating the practical performance of these screening coefficients in a simulation example. The simulation study in Wang and Leng (2016) focuses on correctly selecting a sparse true model, while we are also interested in the HOLP estimator being almost proportional to the true regression coefficients β for later application in the random projection. Therefore, we simulate data from the following setting (similar to the ones employed in the simulation study in Section 3.1).

Example 1. We generate data from (1) with multivariate normal predictors $x_i \sim N(0, \Sigma)$ and normal errors $\varepsilon_i \sim N(0, \sigma^2)$, where we choose $n = 200, p = 2000, \mu = 1$, and $\Sigma = \rho 1_p 1_p' + (1 - \rho) I_p$ has a compound symmetry structure with $\rho = 0.5$ and eigenvalues $\lambda_1 = 1 - \rho + p\rho, \lambda_j = 1 - \rho, j = 2, \ldots, p$. The first $a = 100$ entries of β are uniformly drawn from $\pm\{1, 2, 3\}$ and the rest are zero. The error variance σ^2 is chosen such that the signal-to-noise ratio is $\rho_{\text{snr}} = \beta' \Sigma \beta / \sigma^2 = 10$.

We compare variable screening based on the marginal correlations used in SIS, HOLP, Ridge with penalty $\lambda = \sqrt{n} + \sqrt{p}$ proposed in Wang and Leng (2016) and Ridge with λ chosen by 10-fold cross-validation. Figures 8a and 8b in the Appendix show that HOLP and Ridge with penalty $\lambda = \sqrt{n} + \sqrt{p}$ better separate the active and non-active predictors and achieve better results for precision, recall, true sign recovery and correlation to the true coefficient compared to Ridge with cross-validated penalty and correlation-based screening.

2.2. Random projection

Random projection works by generating a random matrix $\Phi \in \mathbb{R}^{m \times p}$ with $m \ll p$ and transforming the predictors as $z_i = \Phi x_i \in \mathbb{R}^m$ for further analysis. When applied to linear regression, a random projection should ideally preserve predictive power and ensure that $\beta \in \text{span}(\Phi')$, allowing recovery of the true coefficients after reduction. To achieve this, we propose a novel random projection matrix tailored to the regression problem, based on the *sparse embedding matrix* of Clarkson and Woodruff (2013). This projection satisfies the *JL* property and is constructed as follows:

Definition 1. Let $h : [p] \to [m]$ be a random map such that each $j \in [p]$ is assigned a random goal dimension: $h(j) = h_j \overset{i.i.d.}{\sim} \text{Unif}([m])$. Let $B \in \mathbb{R}^{m \times p}$ be a binary matrix with $B_{h_j, j} = 1$ for all $j \in [p]$ and all other entries zero, assuming $\text{rank}(B) = m$. Let $D \in \mathbb{R}^{p \times p}$ be a diagonal matrix with default entries $d_j \sim \text{Unif}(\{-1, 1\})$, independent of h. Then, we call $\Phi = BD$ a CW random projection (CW RP).

Each variable j is mapped to a uniformly random goal dimension h_j with random sign, assuming that each goal dimension $k \in [m]$ is reached by h for some variable $j \in [p]$, which leads to $\text{rank}(B) = m$ (otherwise, the dimension is discarded and m is reduced by one). The sparsity of the CW RP ensures computational efficiency, while its structure makes it analytically tractable. Specifically, we propose to adapt its diagonal elements to ensure: (i) *sign consistency*: variables in the same goal dimension do not have conflicting signs that would cancel out their respective contributions to the response, and, (ii) *coefficient recovery*: $\beta \in \text{span}(\Phi')$, i.e., the true coefficients $\beta \in \mathbb{R}^p$ can be recovered by the reduced predictors $z_i = \Phi x_i$ when modeling the responses as their linear combination $y_i \approx z_i' \gamma = x_i' \Phi' \gamma, \gamma \in \mathbb{R}^m$.

In Lemma 3, we show that for a CW RP Φ with general diagonal entries $d_j \in \mathbb{R}$, the projection of the coefficients β onto the row span of Φ, in matrix form $\tilde{\beta} = P_\Phi \beta = \Phi' (\Phi \Phi')^{-1} \Phi \beta$, is explicitly given by

$$\tilde{\beta}_j = d_j \cdot \frac{\sum_{k:h_k=h_j} d_k \beta_k}{\sum_{k:h_k=h_j} d_k^2}.$$

To ensure $\tilde{\beta} = \beta$ and thus $\beta \in \text{span}(\Phi')$, we propose setting $d_j = c \cdot \beta_j$ for some constant $c \in \mathbb{R}$. Note that for diagonal entries $d_j \in \mathbb{R}$, we ensure $\text{rank}(\Phi) = m$ by assuming that each $i \in [m]$ has at least one mapped variable $j \in h^{-1}(i) = \{k \in [p] : h(k) = i\}$ with $d_j \neq 0$. If not, we set $d_{j_i} = \text{Unif}(\{-1, 1\}) \cdot \min_{j:d_j \neq 0} |d_j|$ where $j_i = \min(h^{-1}(i))$.

The following theorem shows that we can improve the mean square prediction error when using diagonal elements proportional to β rather than random signs.

Theorem 1. *Assume we have data $(y_i, x_i), i = 1, \ldots, n$, from the model (1) with $\mu = 0$, where $x_i \overset{i.i.d.}{\sim} N(0, \Sigma)$ with $0 < \Sigma \in \mathbb{R}^{p \times p}, p > n$, and we want to predict a new observation from the same distribution $\tilde{y} = \tilde{x}'\beta + \tilde{\varepsilon}$ independent from the given data. For a smaller dimension $m < n - 1$, let $\Phi_{rs} = BD_{rs} \in \mathbb{R}^{m \times p}$ be the CW RP with random sign diagonal entries and $\Phi_{pt} = BD_{pt} \in \mathbb{R}^{m \times p}$ the CW RP with diagonal entries $d_j^{pt} = c\beta_j$ for some constant $c > 0$ proportional to the true β.*

Let $Z_{rs} = X\Phi'_{rs} \in \mathbb{R}^{n \times m}$ and $Z_{pt} = X\Phi'_{pt} \in \mathbb{R}^{n \times m}$ be the reduced predictor matrices and $\hat{y}_{rs} = (\Phi_{rs}\tilde{x})'(Z'_{rs}Z_{rs})^{-1}Z'_{rs}y$ and $\hat{y}_{pt} = (\Phi_{pt}\tilde{x})'(Z'_{pt}Z_{pt})^{-1}Z'_{pt}y$ the corresponding least-squares predictions. Then,

$$\mathbb{E}[(\tilde{y} - \hat{y}_{rs})^2] - \mathbb{E}[(\tilde{y} - \hat{y}_{pt})^2] \geq C_{Th1} > 0, \tag{3}$$

$$C_{Th1} = \|\beta\|^2 \left[\lambda_p \left(1 - \frac{2m}{p}\right)\right] + \frac{a}{p-1} m \lambda_p \tau^2 \left(1 - \frac{m+1}{p-1} + \mathcal{O}(p^{-2})\right), \tag{4}$$

where $\mathcal{A} = \{j \in [p] : \beta_j \neq 0\}$ is the active index set, $a = |\mathcal{A}|$ is the number of active variables, $\tau = \min_{j:\beta_j \neq 0} |\beta_j|$ is the smallest non-zero absolute coefficient and $\lambda_p > 0$ is the smallest eigenvalue of Σ.

The proof can be found in Appendix B.

Remark 1.

- This theorem shows that when using the random projection from Definition 1 for least-squares regression, the expected squared prediction error is much smaller when using diagonal elements proportional to the variables' true effect on the response as opposed to the conventional random sign, and gives an explicit conservative lower bound on how much smaller it has to be at least.
- In practice, the true β is unknown, but in Section 2.1 we saw that $\hat{\beta}_{\text{HOLP}}$ asymptotically recovers the true sign and order of magnitude with high probability, and has high correlation to the true β, meaning it is 'almost' proportional to the true β. So we propose to use $\hat{\beta}_{\text{HOLP}}$ as diagonal elements of our projection. See Remark 2 in Appendix B for a short note on the implications of the error bound, the relaxation of distributional assumptions, and the full-rank adaption of the diagonal elements.
- Note that this bound is non-asymptotic and valid for any allowed m, n, p, a (up to the quadratic order in p), and it does not depend on the signal-to-noise ratio ρ_{snr} or the noise level σ^2, because they have the same average effect on the error for both random projections.

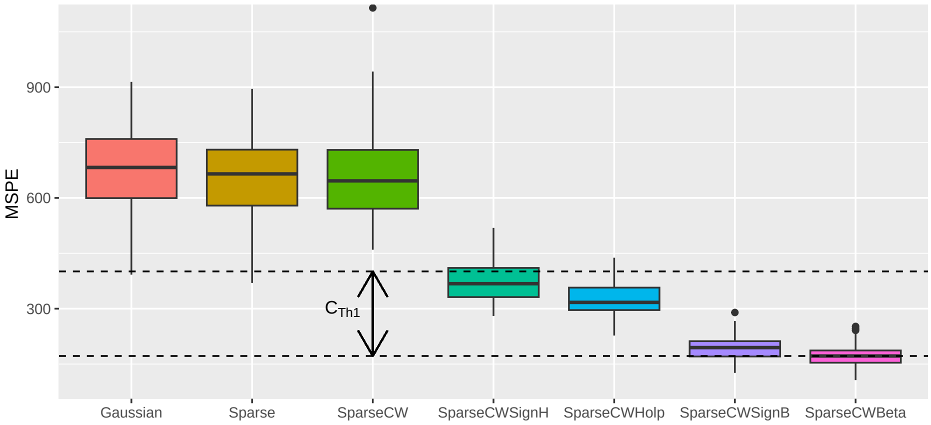
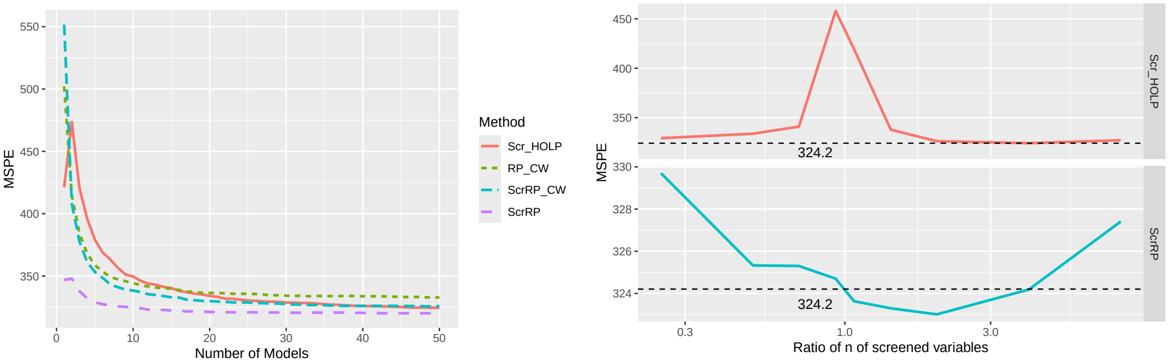


Figure 1: MSPE of different conventional projections, the proposed projection using the HOLP coefficient (SparseCWHolp) or its signs (SparseCWSignH), and the oracle projections using the true β (SparseCWBeta) or its signs (SparseCWSignB). Example 1 setup with 100 replications is used.

In what follows, we want to verify the above considerations and the obtained bound by evaluating the prediction performance of different projections in a small simulation example, using again the setting from Example 1. When $\Phi \in \mathbb{R}^{m\times p}$ is the selected random projection matrix, we fit an ordinary least-squares model to the responses y_i on the reduced predictors $z_i = \Phi x_i$ to obtain predictions for $n_{test} = 100$ new predictor observations. These predictions are evaluated by the mean squared prediction error MSPE. We set the reduced dimension to the true number of active variables $m = a = 100$ and compare Φ with Gaussian i.i.d. $N(0,1)$ entries, the sparse construction $\Phi_{ij} = \pm 1/\sqrt{\psi}$ with probability $\psi/2$ and zero otherwise with $\psi = 1/3$ (Achlioptas 2003), and the following three versions from our Definition 1: SparseCW with standard random sign diagonal elements, SparseCWSignH with $d_j = \text{sign}(\hat{\beta}_{\text{HOLP},j})$ and SparseCWHolp with $d_j = \hat{\beta}_{\text{HOLP},j}$. Additionally, we look at two oracles SparseCWSignB from Definition 1 with $d_j = \text{sign}(\beta_j)$ and SparseCWBeta with $d_j = \beta_j$ with the full-rank adaptions proposed in Theorem 1. Figure 1 shows the prediction performance of these different projections for 100 replications. We also plot the theoretical lower bound C_{Th1} from Theorem 1 from the best oracle to SparseCW with random signs and see that the observed difference is even higher. The conventional random projections stay well above this bound, while our proposed random projections using the HOLP-coefficient manage to stay within the bound of the oracle's performance, with random projection using the sign-information (SparseCWSignH) instead of the coefficients performing only slightly worse than SparseCWHolp.

2.3. Combination of screening and random projection

Similar to Mukhopadhyay and Dunson (2020), we employ a two-step approach where probabilistic variable screening is performed before projecting the screened variables to a random dimension using the random projection matrix in Section 2.2 to avoid noise accumulation from using too many unimportant predictors in the random projection. These steps are repeated several times to build an ensemble of linear models in order to reduce the inherent randomness. We set the number of screened variables to a



(a) (b)

Figure 2: Left: Average MSPE for ensembles using HOLP screening (Src_HOLP), CW random projections (RP_CW), HOLP with conventional random projections (ScrRP_CW), and HOLP with CW random projections (ScrRP) across different numbers of models. Right: Average effect of number of screened variables on MSPE, comparing screening alone to screening plus random projection before linear regression. Example 1 setup with 100 replications is used.

fixed multiple of the sample size $c \cdot n$ (independent of p), draw the variables without replacement with probabilities proportional to their utility based on the HOLP-estimator $\hat{p}_j \propto |\hat{\beta}_{\mathrm{HOLP},j}|$, and use goal dimensions $m \sim \mathrm{Unif}(\{\log(p), \dots, n/2\})$ to increase estimation performance of the linear regression in the reduced model. This is in contrast to the approach in Mukhopadhyay and Dunson (2020), who use probabilities proportional to the marginal correlation, do not control the number of variables selected in the screening directly, and use a slightly larger goal dimension $m \sim \mathrm{Unif}(\{2\log(p), \dots, 3n/4\})$.

In the remainder of the section, we examine the effects of the number of models in the ensemble for different combinations of variable screening and random projection steps as well as the impact of the number of screened variables through a simulation exercise using the data setting from Example 1.

Number of models in ensemble

Figure 2a shows the effect of the number of models used on the average prediction performance over 100 replications and compares the following four methods: screening to $n/2$ variables based on $\hat{\beta}_{\mathrm{HOLP}}$ (Scr_HOLP), random projections with SparseCW matrix (RP_CW), and first screening with $\hat{\beta}_{\mathrm{HOLP}}$ to $2n$ variables and then using the conventional SparseCW random projection (ScrRP_CW) or our proposed SparseCWHolp random projection (ScrRP). When we use just one model, the screening methods deterministically select the variables with highest marginal importance $|\hat{\beta}_{\mathrm{HOLP},j}|$, $j = 1, \dots, p$, otherwise they are drawn without replacement with initial probabilities $\hat{p}_j \propto |\hat{\beta}_{\mathrm{HOLP},j}|$, as previously mentioned. We can see that the combination of screening and the proposed random projection yields the best performance, and that the gain of using more models diminishes at around 20 models already for this method.

Number of screened variables

In Figure 2b, we look at the effect of the number of screened variables $c \cdot n$ on prediction performance, where we compare just screening (Scr_HOLP) to the combination of screening with random projection (ScrRP) as above for a fixed number of models $M = 20$, and we show averages over 100 replications. For just screening, we use the HOLP estimator from Section 2.1 as the subsequent regression method when $c \geq 1$, and in case the system is close to degeneracy we add a small ridge penalty $\lambda = 0.01$ to the OLS estimate. We see that the screening still performs badly for c close to 1 because the sample covariance of the selected predictors is close to singularity. For small and large ratios, it achieves better prediction performance. When combining the screening with the random projection, c does not have such a big impact, and we can achieve lower prediction errors where the best results are achieved for $2 \leq c \leq 4$.

So far, every variable selected once in the screening step will have a contribution to the final regression coefficient, so when we choose a smaller number of models and ratio c, there will be fewer variables in the model. To achieve additional sparsity, we use a thresholding step to actively set less important contributions to 0 before averaging.

2.4. Sparse projected averaged regression (SPAR)

The considerations of the previous sections lead us to propose the following Algorithm 1 for high-dimensional regression where $p > n$.

In Step 3, for a vector $y \in \mathbb{R}^n$ and an index set $I \subset [n]$, $\bar{I}$ denotes the complement, $y_I \in \mathbb{R}^{|I|}$ is the subvector with entries $\{y_i : i \in I\}$; for a matrix $B \in \mathbb{R}^{n \times m}$, $B_{I.} \in \mathbb{R}^{|I| \times m}$ denotes the submatrix with rows $\{B_{i.} : i \in I\}$ (similarly for a subset of columns).

The standardization in Step 1 stabilizes computation and makes the estimated regression coefficients comparable. In the calculation of $\hat{\beta}_{\text{HOLP}}$ with standardized X in Step 2, XX' will have rank $n-1$, so we use $U\Lambda^* U'$ instead of $(XX')^{-1}$, where $U \in \mathbb{R}^{n \times n}$ has the eigenvectors of XX' in its columns, Λ is a diagonal matrix with the corresponding eigenvalues λ_i, and Λ^* is diagonal with entries $1/\lambda_i$ for $i = 1, \ldots, n-1$ and the entry (n, n) is set to zero.

In Step 3.1, the choice of $2n$ is guided by the analysis of prediction performance in Section 2.3.2. However, if the ratio of p/n is exceptionally small (< 5) or large (> 100), a smaller or larger multiple of n can be used for better visual representation of the estimated coefficients. The thresholding Step 4 introduces additional sparsity to the models in the ensemble, where the threshold-level can be selected via cross-validation. Note that when using $\lambda = 0$ and no screening (i.e., all variables are included in the random projection, $I^k = [p]$), then each marginal model will return the same $\hat{\gamma}^k = 1_m, \hat{\beta}^k = \hat{\beta}_{\text{HOLP}}$, since $Z_k \hat{\gamma}^k = X\hat{\beta}_{\text{HOLP}} = y$ achieves zero training loss. The proposed method with screening will yield different coefficients compared to HOLP, but there will remain some similarities.

The simple average in Step 5 can be replaced by a weighted average, where the weights are chosen based on AIC (Burnham and Anderson 2004), (leave-out-one or cross-validation) prediction error, true posterior model weights in a Bayesian approach, or dynamic model weights in time series modeling (Gruber and Kastner 2023). However, across all our efforts, the simple average across all models turned out to yield the

Algorithm 1

```
Step 1. Standardize inputs $X : n \times p$ and $y : n \times 1$;
Step 2. Calculate $\hat{\beta}_{\mathrm{HOLP}} = X'(XX')^{-1}y$ using the standardized inputs;
Step 3. For $k = 1, \dots, M$:
    3.1. Draw $2n$ predictors out of $p$ with probabilities $\hat{p}_j \propto |\hat{\beta}_{\mathrm{HOLP},j}|$ without replacement sequentially yielding the screening index set $I^k = \{j_1^k, \dots, j_{2n}^k\} \subset [p]$; if $p < 2n$ set $I^k = [p]$;
    3.2. Project selected variables to dimension $m_k \sim \mathrm{Unif}\{\log(p), \dots, n/2\}$ using $\Phi_k : m_k \times 2n$ from Definition 1 with diagonal elements $d_i = \hat{\beta}_{\mathrm{HOLP},j_i^k}$ to obtain reduced predictors $Z_k = X_{.I^k}\Phi_k' \in \mathbb{R}^{n \times m_k}$;
    3.3. Fit OLS of $y$ against $Z_k$ to obtain $\hat{\gamma}^k = (Z_k'Z_k)^{-1}Z_k'y$ and $\hat{\beta}^k$, where $\hat{\beta}^k_{I^k} = \Phi_k'\hat{\gamma}^k$ and $\hat{\beta}^k_{\bar{I}^k} = 0$;
Step 4. For a given threshold $\lambda > 0$, set all entries $\hat{\beta}_j^k$ with $|\hat{\beta}_j^k| < \lambda$ to 0 for all $j, k$;
Step 5. Combine via simple average $\hat{\beta} = \sum_{k=1}^{M} \hat{\beta}^k / M$;
Step 6. Choose $M$ and $\lambda$ via 10-fold cross-validation (CV) by repeating Steps 1–5 using the original index sets $I^k$ and projections $\Phi_k$) for each fold; evaluate prediction power by MSE on the withheld fold and choose $(M_{\mathrm{best}}, \lambda_{\mathrm{best}}) = \mathrm{argmin}_{M,\lambda}\widehat{\mathrm{MSE}}(M, \lambda)$;
Step 7. Output the estimated coefficients and predictions for the chosen $M$ and $\lambda$.
```

best predictions for the investigated settings. Similar observations have already been reported in the literature as the forecast combination puzzle (Claeskens et al. 2016).

The number of marginal models M can also be chosen via cross-validation (after specifying a grid of values). However, in Figure 2a, we observed that the effect of M decreases after a certain value, so it would be possible to fix it in the analysis. Note that higher values of M will lead to more variables being employed in the ensemble.

Finally, we note that the input data is used both to compute the screening coefficient in Step 2 and to estimate the ensemble models, raising concerns about potential overfitting. However, additional simulations show that a data-splitting approach – where one subset is used to estimate the screening coefficient and the remainder for estimating the marginal models – does not improve performance. This indicates that the combined data usage does not lead to overfitting in our settings.

3. Simulation Study

3.1. Data generation

We assume the linear model in Equation (1). The covariance matrix Σ of the predictors and the coefficient vector $\beta \in \mathbb{R}^{p\times p}$ will change depending on the simulation setting. The intercept is set to $\mu = 1$ and the error variance σ^2 is chosen such that the signal-to-noise ratio $\rho_{snr} = \beta'\Sigma\beta/\sigma^2 = 10$. We choose $p = 2000$ as a high number of variables and consider the following different simulation settings for Σ: (1) Independent predictors, (2) compound symmetry structure with common covariance $\rho = 0.5$, (3) autoregressive structure with $\rho = 0.9$, (4) a block-diagonal group structure of the previous three choices, (5) a lower-dimensional factor structure, and (6) an extreme correlation setting, where any active predictor has less marginal correlation to the response than other predictors. See Section C in the Appendix for more details on these choices of Σ.

We vary the number of active predictors a between a *sparse* $a = 2\log(p)$, *medium* $a = n/2 + 2\log(p)$ and *dense* $a = p/4$ choice (rounded to closest integer). For settings (1) to (5), the positions of the non-zero entries in β are chosen uniform random (without replacement) in $[p]$ and these entries are independently set as $(-1)^u(4\log(n)/\sqrt{n}+|z|)$, where u is drawn from a Bernoulli distribution with probability of success parameter $p = 0.4$ and z is a standard normal variable. This choice was taken from Fan and Lv (2008), such that the coefficients are bounded away from 0 and vary in sign and magnitude. In setting 6, we choose the first a predictors to be active with $\beta_j = j$ for $j = 1, \ldots, a$ and $\beta_k = 0$ for $k > a$.

For each setting, we generate $n = 200$ observations and evaluate the performance on $n_{\text{test}} = 1000$ further test observations. For setting 4, we also consider $p = 500, 10000$, $n = 100, 400$ as well as $\rho_{\text{snr}} = 1, 5$, and each setting is repeated $n_{\text{rep}} = 100$ times.

3.2. Error measures

We evaluate prediction performance on $n_{test} = 1000$ independent observations via *relative mean squared prediction error* rMSPE $= \sum_{i=1}^{n_{test}}(\hat{y}_i^{test} - y_i^{test})^2 \Big/ \sum_{i=1}^{n_{test}}(y_i^{test} - \bar{y})^2$, which is also used and motivated in Silin and Fan (2022). This measure gives an interpretable performance measure relative to the naive estimator $\hat{\beta} = 0$, which has been shown to achieve a small mean squared error in some high-dimensional settings, and we want to achieve rMSPE < 1 as small as possible.

To evaluate how well the methods are able to rank the variables, we employ the absolute value of the estimated coefficient vector to compute the partial area under the receiver operating characteristic curve (pAUC) (similar to Wang et al. 2019). In the computation of pAUC we limit the number of false positives to $n/2$, which also allows for a fairer comparison between sparse and dense methods than AUC. In all presentations, we rescale pAUC to the interval [0,1] for better interpretation.

Finally, we evaluate the sparse methods and screening-based methods for variable selection using precision (proportion of active predictors among identified active predictors) and recall (proportion of correctly identified active predictors among active predictors).

3.3. Competitors

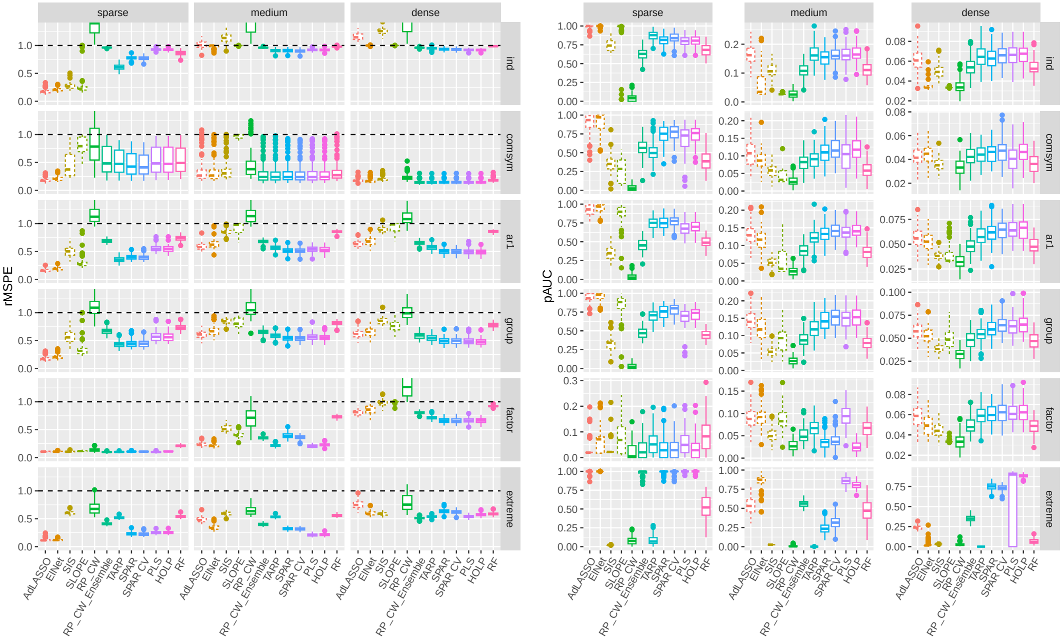
We compare the following list of methods:

- AdLASSO using 10-fold CV (Zou 2006);
- Elastic Net with $\alpha = 3/4$ using 10-fold CV (Zou and Hastie 2005);
- Sorted L-one penalized estimation (SLOPE) (Figueiredo and Nowak 2014; Bogdan et al. 2015);
- SIS (Fan and Lv 2008, screening method);
- Projected linear regression using one draw of a Sparse CW RP matrix with dimension randomly drawn as in Step 3.3 of the SPAR algorithm (RP_CW);
- An ensemble of $M = 100$ models of the projected linear regression with a Sparse CW RP matrix (RP_CW_Ensemble);
- TARP (Mukhopadhyay and Dunson 2020, targeted random projection method), which employs screening based on marginal correlations and the conventional random projection of Achlioptas (2003);
- SPAR with fixed $\lambda = 0, M = 20$;
- SPAR CV with $M \leq 100$ and λ both chosen by cross-validation;
- PLS;
- HOLP, where the coefficients in the linear regression model are computed using Equation (2);
- Random forests (Breiman 2001, RF).

RF is the only non-linear method, but similarly to TARP and SPAR, it relies on sampling covariates for inclusion in smaller models to build an ensemble. However, RF uses uniform sampling, while TARP and SPAR perform probabilistic screening, aiming at giving higher probabilities to more important covariates.

We also performed principal component regression (PCR) and a linear model with Ridge penalty chosen by 10-fold CV, but we omitted the results for a more compact overview. PCR performed similarly to PLS in prediction, while Ridge performed worse than PLS in most settings. Moreover, we also replaced the Sparse CW matrix with Gaussian and sparse conventional random projection, but did not report the results as their performance was very similar to that of the CW RP matrix in all settings. Finally, we employed a linear model with a LASSO penalty, but observed it did not outperform adaptive LASSO or elastic net, so we omitted the results from this section. Adaptive LASSO, elastic net, SLOPE, and SIS can be considered sparse methods and will be marked by dotted boxes in the figures.

All methods were implemented in R (R Core Team 2024) using the packages **glmnet** (Friedman et al. 2010, AdLASSO and ElNet), **SIS** (Saldana and Feng 2018), **pls** (Liland et al. 2022), **randomForest** with `mtry` parameter tuned by Out-of-Bag error (Liaw and Wiener 2002), **SLOPE** (Larsson et al. 2025, with `alpha` estimated by up to 10 iterations of Algorithm 5 in Bogdan et al. 2015) and the source code available online on `https://github.com/david-dunson/TARP` for TARP. Our proposed method is implemented in the R package **spar** available on GitHub (`https://github.com/RomanParzer/SPAR`).



(a)

(b)

Figure 3: Relative MSPE (a) and partial AUC (b) of competing methods for different covariance and active predictor settings ($n_{\text{rep}} = 100$, $n = 200, p = 2000, \rho_{\text{snr}} = 10$). Sparse methods are marked by dotted boxes.

3.4. Results

First, we look at the prediction results of the competing methods for the six different covariance settings and sparse, medium and dense active predictor settings with fixed $n = 200, p = 2000, \rho_{\text{snr}} = 10$ in Figure 3a. We see that the overall performance depends heavily on the covariance setting, and the signal-to-noise ratio alone does not quantify the difficulty of a regression problem. In the 'independent' covariance setting with many active predictors, all methods barely outperform the naive estimator $\hat{\beta} = 0$ with an rMSPE close to one, while in other covariance settings, the errors are much lower. In general, we see that the sparse methods, especially AdLASSO and ElNet, perform well in sparse settings, but not in settings with more active variables. On the other hand, the PLS method and HOLP perform well in all dense settings, but less so in sparse settings. Except in some sparse settings, the SPAR method provides competitive results. Note that Algorithm 5 in Larsson et al. (2025) can fail to estimate `alpha` for SLOPE when it selects more than n variables, e.g., in the extreme correlation setting.

To assess the overall performance, for each scenario and each repetition, we rank the methods from best (= 1) to worst (= 12) in terms of their relative MSPE. Table 1 shows the average of these ranks (and its standard error). The proposed SPAR CV method has the best average rank, followed by SPAR and PLS. SPAR and SPAR CV provide a good prediction performance all-around, showing that it is a viable option, especially in cases where it is not clear how sparse the problem is in practice. We also observe that in terms of prediction ability, SPAR's performance is close to SPAR CV's. Depending on the application context, the additional computational cost of the cross-validation can be avoided with minimal loss in prediction power.

Table 1: Mean and standard error of the rank (best to worst) based on rMSPE and pAUC across all settings for $n_{\text{rep}} = 100$. The cells with the best 3 values are highlighted.

Method	rMSPE		pAUC	
AdLASSO	5.185	(0.076)	**4.146**	(0.057)
ElNet	5.511	(0.068)	5.381	(0.078)
SIS	9.002	(0.063)	9.934	(0.041)
SLOPE	8.596	(0.072)	6.734	(0.09)
RP_CW	11.584	(0.02)	11.227	(0.039)
RP_CW_Ensemble	6.885	(0.062)	7.872	(0.057)
TARP	4.544	(0.041)	6.156	(0.064)
SPAR	**4.291**	(0.047)	4.983	(0.057)
SPAR CV	**3.746**	(0.048)	**4.133**	(0.054)
PLS	**4.304**	(0.066)	4.458	(0.061)
HOLP	4.983	(0.065)	**4.427**	(0.056)
RF	9.369	(0.039)	8.548	(0.052)

Figure 3b provides information on how well the methods rank the variables as measured by pAUC. We observe that the results again highly depend on the investigated setting. The sparse methods achieve a high pAUC in most sparse settings, while SPAR CV performs well in almost all other settings. In Table 1, we see that SPAR CV, followed by AdLASSO and HOLP, perform best when ranking the methods based on pAUC.

In the Appendix, we present the results of the sparse methods and the methods including a screening step in terms of variable selection by looking at precision (Figure 9) and recall (Figure 10). We observe that SPAR achieves a high recall but lower precision, showing that the number of employed variables is rather high compared to the truly active predictors. The same can be observed for TARP. The sparse methods, on the other hand, achieve better precision than a dense method, which would select all variables, and also a high recall in most sparse settings, but not in more dense settings.

Next, we take a closer look at the 'group' covariance setting with medium active variables and look at the effect of changing p, n or ρ_{snr}. Figure 11 in the Appendix shows that all methods achieve increasingly better performance when p is decreasing and that a similar effect can be seen for increasing n and the signal-to-noise ratio ρ_{snr}, where both versions of SPAR are always among the best methods for prediction. RP_CW was left out of these figures because of its worse performance compared to the ensemble version and all other methods.

Finally, Figure 4 shows the average computing times in the 'group' covariance setting. For $p = 10000$, SIS takes the least time to compute, followed by HOLP and SPAR without cross-validation. One pays a price in terms of computing time when employing the cross-validation in SPAR CV. However, the increase in needed computation time for growing p is lower than for most other methods, e.g., RF or PLS. It might be surprising that the random projection ensemble takes longer to compute than other methods consisting of more steps, but this is due to the large input dimension of the random projection matrices compared to SPAR and TARP, which first employ a screening step.

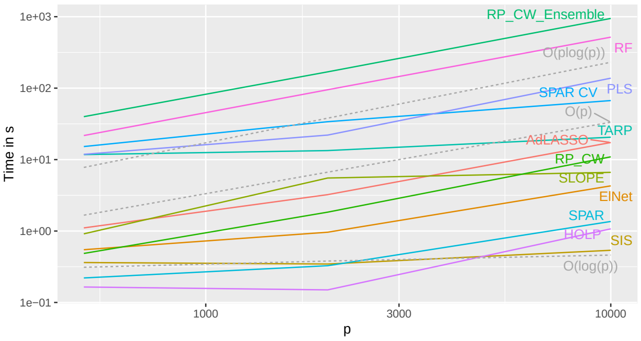


Figure 4: Average computing time in seconds for 'group' covariance over $n_{\text{rep}} = 100$ replications of each active variable setting for increasing p and fixed $n = 200$, $\rho_{\text{snr}} = 10$.

4. Data Applications

In this section, we apply SPAR and its competitors to two real-world high-dimensional regression problems. For both applications, the data is randomly split into a training set of size $3n/4$ and a test set of size $n/4$ (rounded), with this process being repeated 100 times.

4.1. Rat eye gene expression

In this example, we use the data from Scheetz et al. (2006)[1], which measured expression levels of 31,042 (non-control) gene probes on collected tissues from eyes of $n = 120$ rats. Similarly to Huang et al. (2006), we are interested in modeling the relation of all other genes to a specific gene TRIM32, which has been related to Bardet-Biedl syndrome. Since only a few genes are expected to be linked to the given gene, this can be interpreted as a sparse high-dimensional regression problem (Huang et al. 2006). As in Huang et al. (2006) and Scheetz et al. (2006), we only analyze genes expressed in the eye with sufficient variation. A gene is expressed if its maximum observed value is higher than the first quartile of all expression values of all genes, and has sufficient variation if it exhibits a coefficient of variation of at least two. This filtering yields $p = 22{,}905$ genes to be used in the analysis. A subset of this dataset with $p = 200$ genes is available in the R package **flare** (Li et al. 2022), where all but three genes are also contained in our filtered version. The selection process in Li et al. (2022) is not described in more detail, but all 200 genes have a higher marginal correlation to TRIM32 than 75% of all available genes.

Figure 5 shows the prediction performance for these two versions of the dataset, where SPAR CV and HOLP perform best on the bigger dataset. For the small dataset, SLOPE, TARP, and the ensemble of CW random projections achieve the best results. Both SPAR methods improve their performance when the number of variables increases from 200 to 22905, showing that the method is able to make use of additional information, while the sparse methods yield worse predictions on the bigger dataset. Note that the computation for PLS and PCR failed on the larger dataset, so we show Ridge here instead.

[1] The dataset is publicly available in the Gene Expression Omnibus repository `www.ncbi.nlm.nih.gov/geo` (GEO assession id: GSE5680)

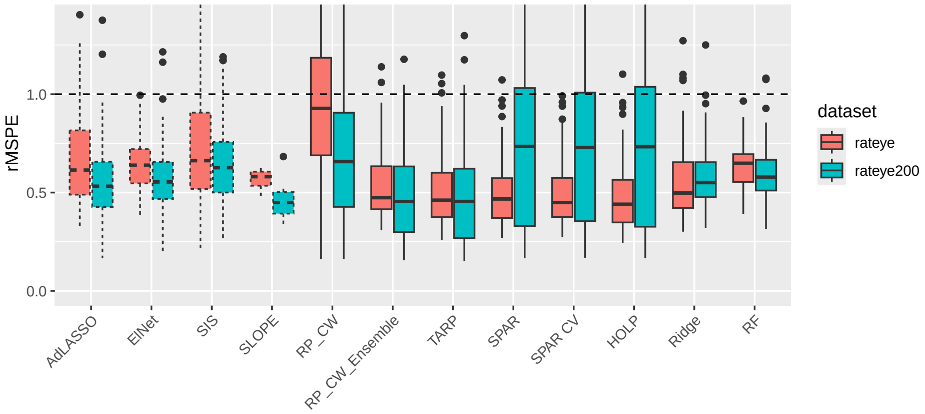


Figure 5: Relative MSPE on the **rat eye gene expression** datasets for 100 random train/test splits. Sparse methods are marked by dotted boxes.

Table 2: Median number of active predictors for all methods on data applications across $n_{\text{rep}} = 100$ random train/test splits.

	AdLASSO	ElNet	SIS	SLOPE	TARP	SPAR	SPAR CV
rateye	44.0	24.5	5.0	41.0	16,490.0	3,199.5	8,737.5
rateye200	10.0	18.5	4.0	36.0	200.0	199.0	181.0
face	19.5	113.5	5.5	309.5	3,626.5	3,506.0	3,746.5

Table 2 shows the median number of active variables of the competing methods on these data applications. SPAR with $M = 20$ and $\lambda = 0$ as well as SPAR CV (where M and λ are selected by cross-validation) reduce the predictor space, but we can see that the number of used variables is much larger than for the sparse methods. Note that SPAR CV can be less sparse than SPAR, even if it performs additional thresholding, in cases where the selected number of models used in the ensemble is larger. The fact that the sparse methods do not achieve the best performance on these datasets raises the question whether this problem is actually sparse, as in our simulated sparse settings, the sparse methods always performed better than the rest. To investigate whether this is due to the observed covariance structure of the predictors rather than the sparsity in β, we perform a small simulation exercise where we generate synthetic data with the observed predictors. Results are presented in Appendix Section D. We show that, for such a covariance structure, the SPAR method performs well even in the sparse setting, leaving the question of the true sparsity in this problem open.

4.2. Face images

The dataset originates from Tenenbaum et al. (2000)[2] and was also studied, among others, in Guhaniyogi and Dunson (2016). It consists of $n = 698$ black and white face images of size $p = 64 \times 64 = 4{,}096$ and the faces' horizontal looking direction angle as the response. The bottom-left plot in Figure 6 illustrates one such instance with the corresponding angle. For each train/test split, we exclude pixels close to the edges and

[2] Data can be found at https://web.archive.org/web/20160913051505/http://isomap.stanford.edu/datasets.html

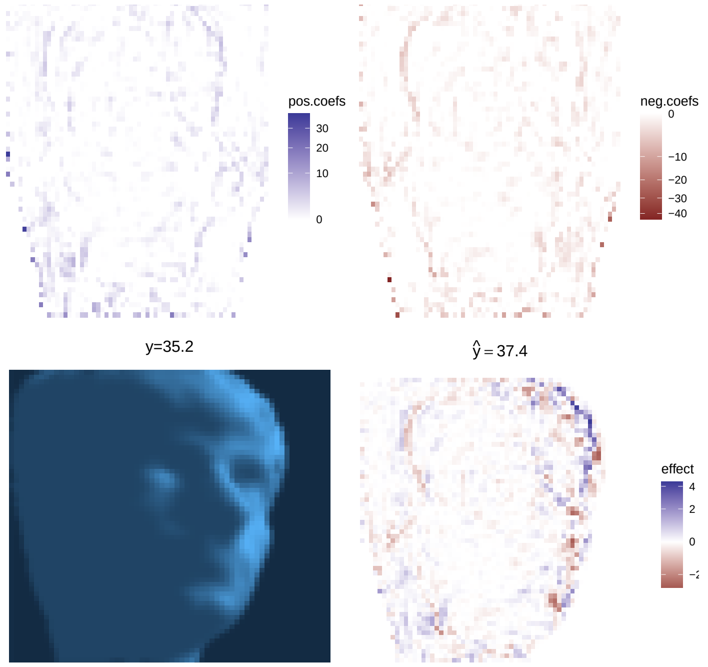


Figure 6: Top: positive (left) and negative (right) estimated coefficients of SPAR CV. Bottom: One new instance (left) and the pixel contributions to its prediction (right).

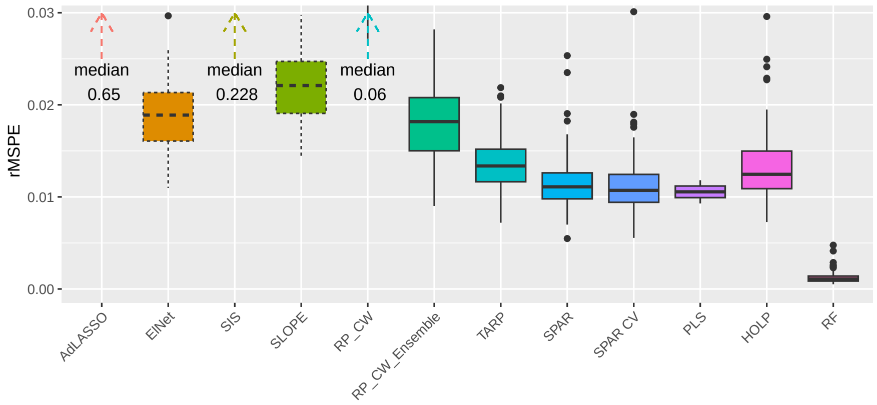


Figure 7: Relative MSPE on the **face angle** dataset for 100 random train/test splits. Sparse methods are marked by dotted boxes.

corners, which are constant on the training set. We expect that many pixels together carry relevant information, making this a rather dense regression problem.

Figure 7 shows the prediction performance results for this dataset. Here, RF yields the lowest prediction error, followed by SPAR, SPAR CV, and PLS. AdLASSO and SIS perform substantially worse than the other methods, as their number of estimated active predictors seems to be way too low; see Table 2. As a non-parametric method, RF is able to estimate a non-linear relationship to the response. Actually, this data example was previously used for illustrating non-linear compressed methods in (low-dimensional) manifold regression in Guhaniyogi and Dunson (2016). When replicating the preprocessing in their paper, SPAR CV achieved an average MSPE of 0.0142 with average bootstrap standard error of 0.0043, while the best non-linear method mentioned in Guhaniyogi and Dunson (2016) achieved 0.06 with standard error 0.009, showing that the proposed method with the linear model assumption is a feasible option for modeling this data, but there might still be some non-linearity indicated by RF's good performance.

One advantage of the linear methods over RF is interpretability. For this dataset, we now illustrate the estimated regression coefficients and their contribution to a new prediction for SPAR CV. We apply our method once on the full dataset except for two test images, thus $n = 696$. The top of Figure 6 shows the positive (left) and negative (right) estimated regression coefficients of the pixels. It yields almost symmetrical images, which is sensible, and highlights the contours of the nose and forehead. For the prediction of a new face image, we can define the contribution of each pixel as the pixel's coefficient multiplied by the corresponding grey-scale value of the new instance. In the bottom-right panel, we visualize these contributions for the test instance on the bottom left. The sum of all these contributions (plus a 'hidden' intercept) yields the prediction of $\hat{y} = 37.4$ for the true angle $y = 35.2$.

5. Summary and Conclusions

This paper introduced a new data-informed random projection aimed at dimension reduction for linear regression, which uses the HOLP estimator (Wang and Leng 2016). We motivated this projection matrix by a theoretical result, where we show how much better we can expect the prediction error to be in a projection that uses the true β coefficients compared to a conventional random projection.

Around this new random projection, we built the SPAR ensemble method with a data-driven threshold selection using cross-validation. In an extensive simulation study, we compare SPAR to different methods employing random projections and other sparse and dense methods. We show that the proposed method achieves the best all-around prediction and variable ranking performance on average across the scenarios. This makes SPAR a viable option, especially in cases where it is unclear how sparse the problem is in practice, because most other methods only perform well in sparse settings but underperform in dense settings or vice versa. We also noticed that SPAR's prediction performance is comparable to the cross-validated SPAR CV. However, SPAR CV shows better performance in terms of variable ranking. In applications where fast prediction is the main task, SPAR without cross-validation can be employed without notable loss in prediction ability.

This methodology can be extended to non-linear (or robust) regression by employing non-linear (or robust) methods, such as generalized linear models or Gaussian processes, in the marginal models instead of OLS. Future work also includes extensions for (multi-) classification or multivariate regression tasks.

Computational Details

All the code to reproduce the results in this paper can be found in the GitHub repository `https://github.com/RomanParzer/SPAR_Paper_Figures_Code`.

Acknowledgments

Roman Parzer and Laura Vana-Gür acknowledge funding from the Austrian Science Fund (FWF) for the project "High-dimensional statistical learning: New methods to advance economic and sustainability policies" (ZK 35).

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, ISSN: 0022-0000, DOI: 10.1016/s0022-0000(03)00025-4. Special Issue on PODS 2001.

Ahfock, D. C., Astle, W. J., and Richardson, S. (2021). Statistical properties of sketching algorithms. *Biometrika*, 108(2):283–297, DOI: 10.1093/biomet/asaa062.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103, DOI: 10.1214/15-AOAS842.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32, ISSN: 1573-0565, DOI: 10.1023/a:1010933404324.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, DOI: 10.1177/0049124104268644.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, DOI: 10.1016/j.ijforecast.2015.12.005.

Clarkson, K. L. and Woodruff, D. P. (2013). Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 81–90. DOI: 10.1145/2488608.2488620.

Cook, R. D. and Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics*, 47:884–908, DOI: 10.1214/18-aos1681.

Cribari-Neto, F., Garcia, N. L., and Vasconcellos, K. L. P. (2000). A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2), DOI: 10.12660/bre.v20n22000.2760.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(5):849–911, DOI: 10.1111/j.1467-9868.2008.00674.x.

Figueiredo, M. A. T. and Nowak, R. D. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. *arXiv preprint arXiv:1409.4005*, DOI: 10.48550/arXiv.1409.4005.

Fornell, C. and Cha, J. (1994). Partial least squares. In Bagozzi, R. P., editor, *Advanced Methods of Marketing Research*, volume 407, pages 52–78.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, DOI: 10.18637/jss.v033.i01.

Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J., and Sohler, C. (2015). Random projections for Bayesian regression. *Statistics and Computing*, 27(1):79–101, DOI: 10.1007/s11222-015-9608-z.

Gruber, L. and Kastner, G. (2023). Forecasting macroeconomic data with Bayesian VARs: Sparse or dense? It depends! DOI: 10.48550/arxiv.2206.04902.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, DOI: 10.1080/01621459.2014.969425.

Guhaniyogi, R. and Dunson, D. B. (2016). Compressed Gaussian process for manifold regression. *Journal of Machine Learning Research*, 17(69):1–26, https://jmlr.org/papers/v17/14-230.html.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, ISSN: 00401706, DOI: 10.2307/1267351.

Huang, J., Ma, S., and Zhang, C.-H. (2006). Adaptive Lasso for sparse high-dimensional regression. *Statistica Sinica*, 18, DOI: https://www3.stat.sinica.edu.tw/sstest/j18n4/j18n420/j18n420.html.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, DOI: 10.1090/conm/026/737400.

Kobak, D., Lomond, J., and Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *The Journal of Machine Learning Research*, 21(1), DOI: 10.5555/3455716.3455885.

Larsson, J., Wallin, J., Bogdan, M., van den Berg, E., Sabatti, C., Candes, E., Patterson, E., Su, W., Kała, J., Grzesiak, K., and Burdukiewicz, M. (2025). ***SLOPE**: Sorted L1 Penalized Estimation*, https://CRAN.R-project.org/package=SLOPE. R package version 0.5.2.

Li, X., Zhao, T., Wang, L., Yuan, X., and Liu, H. (2022). ***flare**: Family of Lasso Regression*, https://CRAN.R-project.org/package=flare. R package version 1.7.0.1.

Liaw, A. and Wiener, M. (2002). Classification and regression by **randomForest**. *R News*, 2(3):18–22, DOI: 10.32614/cran.package.randomforest.

Liland, K. H., Mevik, B.-H., and Wehrens, R. (2022). ***pls**: Partial Least Squares and Principal Component Regression*, https://CRAN.R-project.org/package=pls. R package version 2.8-1.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22, Red Hook, NY. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2009/file/01882513d5fa7c329e940dda99b12147-Paper.pdf.

Mukhopadhyay, M. and Dunson, D. B. (2020). Targeted random projection for prediction from high-dimensional features. *Journal of the American Statistical Association*, 115(532):1998–2010, DOI: 10.1080/01621459.2019.1677240.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Saldana, D. F. and Feng, Y. (2018). **SIS**: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25, DOI: 10.18637/jss.v083.i02.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America*, 103(39):14429–14434, DOI: 10.1073/pnas.0602562103.

Silin, I. and Fan, J. (2022). Canonical thresholding for nonsparse high-dimensional linear regression. *The Annals of Statistics*, 50(1):460 – 486, DOI: 10.1214/21-aos2116.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, DOI: 10.1126/science.290.5500.2319.

Thanei, G.-A., Heinze, C., and Meinshausen, N. (2017). Random projections for large-scale regression. In *Big and Complex Data Analysis: Methodologies and Applications*, pages 51–68. DOI: 10.1007/978-3-319-41573-4_3.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, ISSN: 00359246, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Wang, F., Mukherjee, S., Richardson, S., and Hill, S. M. (2019). High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking. *Statistics and Computing*, 30(3):697–719, DOI: 10.1007/s11222-019-09914-9.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, DOI: 10.1198/jasa.2008.tm08516.

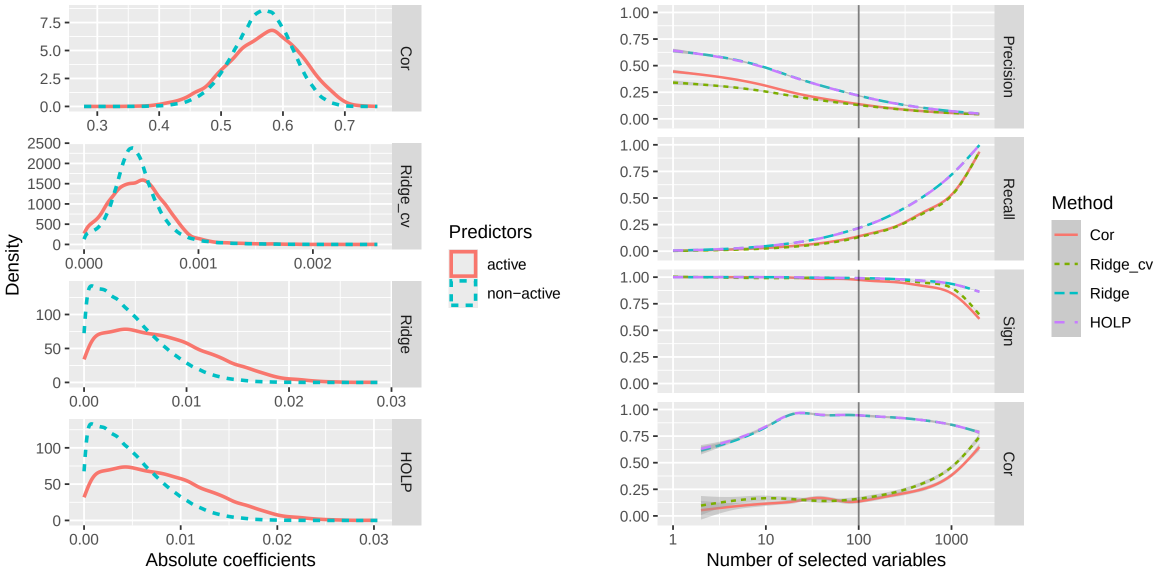
Wang, X. and Leng, C. (2016). High-dimensional ordinary least-squares projection for screening variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78:589–611, DOI: 10.1111/rssb.12127.

Wang, X., Leng, C., and Dunson, D. B. (2015). On the consistency theory of high dimensional variable screening. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2431–2439, Cambridge, MA, USA. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2015/file/540ae6b0f6ac6e155062f3dd4f0b2b01-Paper.pdf

Zhou, S., Wasserman, L., and Lafferty, J. D. (2007). Compressed regression. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, Red Hook, NY. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2007/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, DOI: 10.1198/016214506000000735.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, DOI: 10.1111/j.1467-9868.2005.00503.x.



(a) Density estimates of absolute coefficients

(b) Screening measures

Figure 8: Comparison of screening based on marginal correlations, HOLP, Ridge with $\lambda = \sqrt{n} + \sqrt{p}$ and Ridge with cross-validated λ in the setting in Example 1. (a) shows density estimates of absolute estimated coefficients for active and non-active predictors over $n_{\text{rep}} = 100$ repetitions. (b) shows precision, recall, sign recovery, and correlation of estimates to the true coefficients averaged over 100 replications, where the vertical line indicates the true number of active variables.

A. Simulation Study for Screening

Similarly to Example 1 of Section 2.1, we compare the selection of variables based on marginal correlations, HOLP, Ridge with proposed penalty $\lambda = \sqrt{n} + \sqrt{p}$, and Ridge with λ chosen by 10 fold cross-validation.

Figure 8a shows density estimates of the absolute coefficients estimated by these four methods for truly active and non-active variables for 100 replicated draws of the data. In Figure 8b, we evaluate the selection process of the four methods when selecting the k variables having the highest absolute estimated coefficients and let k vary on the x-axis. We show the precision and recall of this selection, as well as the ratio of correct signs for truly active predictors included in the selection and the correlation of the corresponding true coefficients to the estimates averaged over the 100 replications. We see that HOLP and Ridge with penalty $\lambda = \sqrt{n} + \sqrt{p}$ better separate the active and non-active predictors and achieve better results for precision, recall, true sign recovery and correlation to the true coefficient compared to Ridge with cross-validated penalty and correlation-based screening. In Figure 8a, we see that the absolute coefficients of cross-validated Ridge are much smaller than HOLP and Ridge with $\lambda = \sqrt{n} + \sqrt{p}$, meaning the λ suggested by cross-validation is much higher. In comparison, the choice $\lambda = \sqrt{n} + \sqrt{p}$ even leads to quite similar results as HOLP, which can be interpreted as Ridge with $\lambda = 0$.

B. Lemmas and Proof of Theorem 1

This section states and proves Lemmas 1, 2 and 3 mentioned in Section 2, and gives a detailed proof of Theorem 1 and Lemma 4 needed in the proof.

Lemma 1. *Let $X \in \mathbb{R}^{n\times p}$ be a fixed matrix and $y \in \mathbb{R}^n$ a vector. Then, the Ridge estimator for $\lambda > 0$ has the following alternative form suitable for the $p \gg n$ case.*

$$\hat{\beta}_\lambda := (X'X + \lambda I_p)^{-1}X'y = X'(\lambda I_n + XX')^{-1}y. \tag{5}$$

Proof. Using the Woodbury matrix inversion formula

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

where A, U, C and V are conformable matrices, we have for any penalty $\lambda > 0$

$$\begin{aligned}
\hat{\beta}_\lambda &:= (X'X + \lambda I_p)^{-1}X'y \\
&= \frac{1}{\lambda}\left(I_p - \frac{1}{\lambda}\cdot X'\left(I_n + \frac{1}{\lambda}\cdot XX'\right)^{-1}X\right)X'y \\
&= \frac{1}{\lambda}X'y - \frac{1}{\lambda}X'(\lambda I_n + XX')^{-1}XX'y \pm \frac{1}{\lambda}X'(\lambda I_n + XX')^{-1}\lambda y \\
&= \frac{1}{\lambda}X'y - \frac{1}{\lambda}X'\underbrace{(\lambda I_n + XX')^{-1}(XX' + \lambda I_n)}_{=I_n}y + \frac{1}{\lambda}X'(\lambda I_n + XX')^{-1}\lambda y \\
&= X'(\lambda I_n + XX')^{-1}y.
\end{aligned}$$

□

Lemma 2. *Let $X \in \mathbb{R}^{n\times p}$ be a fixed matrix with rank$(XX') = n$ (implying $p > n$) and $y \in \mathbb{R}^n$ a vector. Then, the minimum norm least-squares solution* $\operatorname{argmin}_{\beta\in\mathbb{R}^p, s.t. X\beta = y}\|\beta\|$ *is uniquely given by $\hat{\beta} = X'(XX')^{-1}y$.*

Proof. Obviously, $\hat{\beta} = X'(XX')^{-1}y$ satisfies $X\hat{\beta} = y$. For any $\tilde{\beta} \in \mathbb{R}^p$ with $X\tilde{\beta} = y$ we have

$$\begin{aligned}
\|\tilde{\beta}\|^2 = \|\hat{\beta} + \tilde{\beta} - \hat{\beta}\|^2 = \|\hat{\beta}\|^2 + \|\tilde{\beta} - \hat{\beta}\|^2 + 2\cdot\hat{\beta}'(\tilde{\beta} - \hat{\beta}) = \\
= \|\hat{\beta}\|^2 + \underbrace{\|\tilde{\beta} - \hat{\beta}\|^2}_{\geq 0} + 2\cdot y'(XX')^{-1}\underbrace{X(\tilde{\beta} - \hat{\beta})}_{=0} \geq \|\hat{\beta}\|^2,
\end{aligned}$$

with equality if and only if $\tilde{\beta} = \hat{\beta}$. □

Lemma 3. *Let $\Phi \in \mathbb{R}^{m\times p}$ be a CW random projection from Definition 1 with general diagonal elements $d_j \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$. Then, the projected vector $\tilde{\beta} = P_\Phi\beta$ for the orthogonal projection $P_\Phi = \Phi'(\Phi\Phi')^{-1}\Phi$ onto the row span of Φ is given by*

$$\tilde{\beta}_j = d_j \cdot \frac{\sum_{k:h_k=h_j} d_k\beta_k}{\sum_{k:h_k=h_j} d_k^2}. \tag{6}$$

Proof. We can split the projection in

$$P_\Phi \beta = \Phi'(\Phi\Phi')^{-1}\Phi\beta = D(B'(\Phi\Phi')^{-1}B)(D\beta).$$

The matrix $\Phi\Phi' = BD^2B' \in \mathbb{R}^{m\times m}$ is diagonal with entries $\{\sum_{l:h_l=i} d_l^2 : i \in [m]\}$, because each variable is only mapped to one goal dimension. Then, for $j, k \in [p]$ we have

$$(B'(\Phi\Phi')^{-1}B)_{jk} = \begin{cases} 0 & h_j \neq h_k \\ 1/(\sum_{l:h_l=h_j} d_l^2) & h_j = h_k \end{cases}.$$

Putting it together, we get

$$\tilde{\beta}_j = d_j \cdot \sum_{k=1}^{p} I\{h_k = h_j\} \cdot \frac{d_k\beta_k}{\sum_{l:h_l=h_j} d_l^2} = d_j \cdot \frac{\sum_{k:h_k=h_j} d_k\beta_k}{\sum_{k:h_k=h_j} d_k^2}.$$

□

Lemma 4. *Let $h : [p] \to [m]$ be a random map such that for each $j \in [p] : h(j) = h_j \overset{i.i.d.}{\sim}$ Unif$([m])$, and let $\mathcal{A} \subset [p]$ be a subset of indices with $a = |\mathcal{A}| > 1$. Then,*

$$\mathbb{E}\left[\frac{|\mathcal{A} \cap h^{-1}(h_j) \setminus \{j\}|}{|h^{-1}(h_j)|}\right] = \frac{a - \mathbf{I}\{j \in \mathcal{A}\}}{p-1} \cdot \left(1 - \frac{m}{p}\left(1 - \left(\frac{m-1}{m}\right)^p\right)\right), \tag{7}$$

$$\mathbb{E}\left[\frac{|\mathcal{A} \cap h^{-1}(h_j) \setminus \{j\}|}{|h^{-1}(h_j)|^2}\right] = m\frac{a - \mathbf{I}\{j \in \mathcal{A}\}}{p-1} \cdot \left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right), \tag{8}$$

where $h^{-1}(k) = \{j \in [p] : h(j) = k\}$ is the (random) preimage set for $k \in [m]$.

Proof. The first random variable $|\mathcal{A} \cap h^{-1}(h_j) \setminus \{j\}|/|h^{-1}(h_j)|$ (random in h) has the distribution of $X_1/(1 + X_1 + X_2)$, where $X_1 \sim \text{Binom}(a_j, 1/m), a_j = a - \mathbf{I}\{j \in \mathcal{A}\}$ corresponding to the active variables (except j) and $X_2 \sim \text{Binom}(p - 1 - a_j, 1/m)$ independent of X_1 corresponding to the inactive variables.

Note that for any $x_1, x_2 \in \mathbb{N}$ $x_1/(1+x_1+x_2) = \int_0^1 x_1 s^{x_1+x_2} ds$ and, by Fubini's theorem, we can interchange the integral and expectation to obtain

$$\mathbb{E}\left[\frac{X_1}{1 + X_1 + X_2}\right] = \int_0^1 \mathbb{E}[X_1 s^{X_1}]\mathbb{E}[s^{X_2}]ds.$$

By using the moment-generating function of a binomial variable and the dominated convergence theorem to interchange the derivative and the expectation, we get

$$\begin{aligned}
\mathbb{E}[s^{X_2}] &= \left(\frac{m-1}{m} + \frac{1}{m}s\right)^{p-1-a_j}, \\
\mathbb{E}[(X_1+1)s^{X_1}] &= \frac{\partial}{\partial s}\mathbb{E}[s^{X_1+1}] = \frac{\partial}{\partial s} s\left(\frac{m-1}{m} + \frac{1}{m}s\right)^{a_j} \\
&= \left(\frac{m-1}{m} + \frac{1}{m}s\right)^{a_j} + s\frac{a_j}{m}\left(\frac{m-1}{m} + \frac{1}{m}s\right)^{a_j-1}, \\
\implies \mathbb{E}[X_1 s^{X_1}] &= \mathbb{E}[(X_1+1)s^{X_1}] - \mathbb{E}[s^{X_1}] = s\frac{a_j}{m}\left(\frac{m-1}{m} + \frac{1}{m}s\right)^{a_j-1}.
\end{aligned}$$

Putting the results together and using partial integration, we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{X_1}{1+X_1+X_2}\right] &= \int_0^1 s\frac{a_j}{m}\left(\frac{m-1}{m}+\frac{1}{m}s\right)^{a_j-1}\left(\frac{m-1}{m}+\frac{1}{m}s\right)^{p-1-a_j} ds \\ &= \frac{a_j}{p-1}\cdot\left(1-\frac{m}{p}\left(1-\left(\frac{m-1}{m}\right)^p\right)\right).\end{aligned}$$

Similarly, the second random variable $|\mathcal{A}\cap h^{-1}(h_j)\setminus\{j\}|/|h^{-1}(h_j)|^2$ has the distribution of $X_1/(1+X_1+X_2)^2$. We will use a similar approach to Cribari-Neto et al. (2000) to obtain a fourth-order approximation.

By use of the Gamma function and similar arguments to the first case, we can write

$$\frac{x_1}{(1+x_1+x_2)^2} = \int_0^\infty x_1 t e^{-(1+x_1+x_2)t} dt$$

for any $x_1, x_2 \in \mathbb{N}$, and

$$\mathbb{E}\left[\frac{X_1}{(1+X_1+X_2)^2}\right] = \int_0^\infty te^{-t}\mathbb{E}[X_1e^{-X_1t}]\mathbb{E}[e^{-X_2t}]dt. \tag{9}$$

By use of the moment-generating functions we get

$$\begin{aligned}\mathbb{E}[e^{-X_2t}] &= \left(\frac{m-1}{m}+\frac{1}{m}e^{-t}\right)^{p-1-a_j}, \\ \mathbb{E}[X_1e^{-X_1t}] &= \mathbb{E}\left[\frac{\partial}{\partial t}\left(-e^{-X_1t}\right)\right] = -\frac{\partial}{\partial t}\mathbb{E}[e^{-X_1t}] = -\frac{\partial}{\partial t}\left(\frac{m-1}{m}+\frac{1}{m}e^{-t}\right)^{a_j} \\ &= a_j\left(\frac{m-1}{m}+\frac{1}{m}e^{-t}\right)^{a_j-1}\frac{1}{m}e^{-t}.\end{aligned}$$

Plugging this into (9) and using the variable substitution $e^{-r} = (m-1)/m + (1/m)e^{-t}$ and the definition $g(r) = -\log[m\{e^{-r} - (m-1)/m\}]me^{-r}$ yields

$$\begin{aligned}&\mathbb{E}\left[\frac{X_1}{(1+X_1+X_2)^2}\right] \\ &= a_j\int_0^\infty \frac{1}{m}te^{-2t}\left(\frac{m-1}{m}+\frac{1}{m}e^{-t}\right)^{p-2} dt \\ &= a_j\int_0^{-\log[\{(m-1)/m\}]} -\log\left(m\left(e^{-r}-\frac{m-1}{m}\right)\right)m\left(e^{-r}-\frac{m-1}{m}\right)e^{-(p-1)r}dr \\ &= a_j\int_0^{-\log[\{(m-1)/m\}]}\left(1-\frac{m-1}{m}e^r\right)g(r)e^{-(p-1)r}dr. \end{aligned}\tag{10}$$

From Cribari-Neto et al. (2000), we use the facts that for $\delta < \min(1, -\log\{(m-1)/m\})$

$$g(r) = m^2r\left[1+\frac{m-3}{2}r+\mathcal{O}(r^2)\right], \tag{11}$$

$$\int_0^\delta r^k e^{-(p-1)r}dr = \frac{\Gamma(k+1)}{(p-1)^{k+1}} + \mathcal{O}(e^{-(p-1)\delta}), \tag{12}$$

$$\int_\delta^{-\log\{(m-1)/m\}} g(r)e^{-(p-1)r}dr = \mathcal{O}(e^{-(p-1)\delta}). \tag{13}$$

We split the integral in (10) in two parts $(0,\delta)$ and $(\delta, -\log\{(m-1)/m\})$. On $r > \delta$ we can use $(1 - \{(m-1)/m\}e^r) \leq 1$ and (13) to obtain

$$\int_\delta^{-\log\{(m-1)/m\}} \left(1 - \frac{m-1}{m}e^r\right) g(r)e^{-(p-1)r}dr = \mathcal{O}(e^{-(p-1)\delta}).$$

On $r < \delta$ we use the Taylor expansion $e^r = 1 + r + \mathcal{O}(r^2)$ and Equations (11) and (12) to get

$$\begin{aligned}
&\int_0^\delta \left(1 - \frac{m-1}{m}e^r\right) g(r)e^{-(p-1)r}dr \\
&= m^2 \int_0^\delta \left(r\frac{1}{m} + r^2\left(-\frac{m-1}{m} + \frac{m-3}{2m}\right) + \mathcal{O}(r^3)\right)e^{-(p-1)r}dr \\
&= m\left[\frac{1}{(p-1)^2} + \frac{2(-(m-1) + (m-3)/2)}{(p-1)^3} + \mathcal{O}(p^{-4})\right],
\end{aligned}$$

where any exponential decay (e.g., from (12))t is omitted when a polynomial decay is present. Together, we obtain from (10)

$$\begin{aligned}
\mathbb{E}\left[\frac{X_1}{(1+X_1+X_2)^2}\right] &= a_j\left[\int_0^\delta \left(1 - \frac{m-1}{m}e^r\right) g(r)e^{-(p-1)r}dr\right. \\
&\quad \left. + \int_\delta^{-\log\{(m-1)/m\}} \left(1 - \frac{m-1}{m}e^r\right) g(r)e^{-(p-1)r}dr\right] \\
&= a_j m\left[\frac{1}{(p-1)^2} + \frac{2(-(m-1) + (m-3)/2)}{(p-1)^3} + \mathcal{O}(p^{-4})\right] \\
&= m\frac{a_j}{p-1} \cdot \left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right).
\end{aligned}$$

□

Proof of Theorem 1. For a general CW projection $\Phi = BD$, reduced predictors $Z = X\Phi'$, and a prediction $\hat{y} = (\Phi\tilde{x})'(Z'Z)^{-1}Z'y = (\Phi\tilde{x})'(Z'Z)^{-1}Z'X\beta + (\Phi\tilde{x})'(Z'Z)^{-1}Z'\varepsilon$ we get the expected squared error (w.r.t $\tilde{x}, \tilde{\varepsilon}$, and ε given X and Φ)

$$\begin{aligned}
&\mathbb{E}[(\tilde{y} - \hat{y})^2|X,\Phi] && (14)\\
&= \mathbb{E}[(\tilde{x}'(I_p - \Phi'(Z'Z)^{-1}Z'X)\beta + \tilde{\varepsilon} - \tilde{x}'\Phi'(Z'Z)^{-1}Z'\varepsilon)^2|X,\Phi] = && (15)\\
&= \mathbb{E}[\beta'(I_p - X'X\Phi'(\Phi X'X\Phi')^{-1}\Phi)\tilde{x}\tilde{x}'(I_p - \underbrace{\Phi'(\Phi X'X\Phi')^{-1}\Phi X'X}_{:=P})\beta && (16)\\
&\quad + \tilde{\varepsilon}^2 + \varepsilon'X\Phi'(\Phi X'X\Phi')^{-1}\Phi\tilde{x}\tilde{x}'\Phi'(\Phi X'X\Phi')^{-1}\Phi X'\varepsilon|X,\Phi] = && (17)\\
&= \beta'(I_p - P)'\Sigma(I_p - P)\beta + \sigma^2 && (18)\\
&\quad + \mathbb{E}[\varepsilon'X\Phi'(\Phi X'X\Phi')^{-1}\Phi\Sigma\Phi'(\Phi X'X\Phi')^{-1}\Phi X'\varepsilon|X,\Phi], && (19)
\end{aligned}$$

where we used that the mixed terms have expectation 0. The third term has conditional expectation given Φ

$$\begin{aligned}
&\mathbb{E}[\varepsilon'X\Phi'(\Phi X'X\Phi')^{-1}\Phi\Sigma\Phi'(\Phi X'X\Phi')^{-1}\Phi X'\varepsilon|\Phi] \\
&= \mathbb{E}[\mathrm{tr}\Big((\Phi X'X\Phi')^{-1}\Phi\Sigma\Phi'(\Phi X'X\Phi')^{-1}\Phi X'\varepsilon\varepsilon'X\Phi'\Big)|\Phi] \\
&= \sigma^2 \cdot \mathrm{tr}\Big(\mathbb{E}[(\Phi X'X\Phi')^{-1}|\Phi]\Phi\Sigma\Phi'\Big),
\end{aligned}$$

where we used the facts that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for matrices A, B of suitable dimensions, $\mathbb{E}[\varepsilon\varepsilon'] = \sigma^2 \cdot I_n$ and ε is independent of X and Φ. For fixed Φ, the matrix $X\Phi'$ has a centered matrix normal distribution with among-row covariance I_n and among-column covariance $\Phi\Sigma\Phi' \in \mathbb{R}^{m\times m}$. Therefore, $\Phi X'X\Phi'$ has a Wishart distribution with scale matrix $\Phi\Sigma\Phi' \in \mathbb{R}^{m\times m}$ and n degrees of freedom, and $(\Phi X'X\Phi')^{-1}$ has an Inverse-Wishart distribution resulting in the expectation $\mathbb{E}[(\Phi X'X\Phi')^{-1}|\Phi] = (\Phi\Sigma\Phi')^{-1}/(n - m - 1)$ and, continuing above calculations, we obtain

$$\mathbb{E}[\varepsilon' X\Phi'(\Phi X'X\Phi')^{-1}\Phi\Sigma\Phi'(\Phi X'X\Phi')^{-1}\Phi X'\varepsilon] = \sigma^2 \cdot \frac{m}{n-m-1}.$$

Since the expectations of the second and third term in (18) and (19) do not depend on Φ or the respective diagonal elements, they will cancel when computing the difference in (3), and we only need to consider the first term $\beta'(I_p - P)'\Sigma(I_p - P)\beta = (\beta - P\beta)'\Sigma(\beta - P\beta)$. The plan is to find an upper bound on its expectation when using diagonal elements proportional to the true coefficient and a lower bound when using random signs as the diagonal elements.

Lower bound for random signs: Let $\lambda_1 \geq \cdots \geq \lambda_p > 0$ be the ordered eigenvalues of Σ and $P_X^{\mathrm{rs}} = \Phi_{\mathrm{rs}}'(\Phi_{\mathrm{rs}} X'X\Phi_{\mathrm{rs}}')^{-1}\Phi_{\mathrm{rs}} X'X$. Then,

$$\mathbb{E}[(\beta - P_X^{\mathrm{rs}}\beta)'\Sigma(\beta - P_X^{\mathrm{rs}}\beta)] \geq \lambda_p \cdot \mathbb{E}[\|\beta - P_X^{\mathrm{rs}}\beta\|^2]. \tag{20}$$

Let $P_\Phi^{\mathrm{rs}} = \Phi_{\mathrm{rs}}'(\Phi_{\mathrm{rs}}\Phi_{\mathrm{rs}}')^{-1}\Phi_{\mathrm{rs}}$ and $\tilde{\beta}^{\mathrm{rs}} = P_\Phi^{\mathrm{rs}}\beta$ be the orthogonal projection. Then, we have

$$\|\beta - P_X^{\mathrm{rs}}\beta\|^2 = \|\beta - \tilde{\beta}^{\mathrm{rs}}\|^2 + \underbrace{\|\tilde{\beta}^{\mathrm{rs}} - P_X^{\mathrm{rs}}\beta\|^2}_{\geq 0} \geq \|\beta - \tilde{\beta}^{\mathrm{rs}}\|^2,$$

because $\tilde{\beta}^{\mathrm{rs}} - P_X^{\mathrm{rs}}\beta \in \mathrm{span}(\Phi_{\mathrm{rs}}')$ and $\beta - \tilde{\beta}^{\mathrm{rs}} \perp \mathrm{span}(\Phi_{\mathrm{rs}}')$.

Using the explicit form of $\tilde{\beta}^{\mathrm{rs}}$ from Lemma 3 and independence of the map h and diagonal elements $d_j \overset{i.i.d.}{\sim} \mathrm{Unif}(\{-1, 1\})$, we get

$$\mathbb{E}[\tilde{\beta}_j^{\mathrm{rs}}] = \mathbb{E}\left[d_j \cdot \frac{\sum_{k:h_k=h_j} d_k\beta_k}{|h^{-1}(h_j)|}\right] = \beta_j \cdot \mathbb{E}\left[\frac{1}{|h^{-1}(h_j)|}\right]. \tag{21}$$

Since we always have $j \in h^{-1}(h_j)$ and the other goal dimensions are independently drawn uniformly at random, the cardinality of this set has distribution $|h^{-1}(h_j)| \sim 1 + \mathrm{Binom}(p-1, 1/m)$. Cribari-Neto et al. (2000) showed that the inverse moments are then given by

$$\begin{aligned}\mathbb{E}\left[\frac{1}{|h^{-1}(h_j)|}\right] &= \frac{m}{p}\left(1 - \left(\frac{m-1}{m}\right)^p\right), \\ \mathbb{E}\left[\frac{1}{|h^{-1}(h_j)|^2}\right] &= \frac{m^2}{(p-1)^2} + \frac{(m-3)m^2}{(p-1)^3} + \mathcal{O}(p^{-4}).\end{aligned}$$

Plugging this into (21) yields

$$\begin{aligned}
\beta_j \mathbb{E}[\tilde{\beta}_j^{\mathrm{rs}}] &= \beta_j^2 \cdot \frac{m}{p}\left(1 - \left(\frac{m-1}{m}\right)^p\right) \leq \beta_j^2 \cdot \frac{m}{p}, \\
\mathbb{E}[(\tilde{\beta}_j^{\mathrm{rs}})^2 | h] &= \mathbb{E}\left[\frac{\sum_{k:h_k=h_j} \sum_{l:h_l=h_j} d_k d_l d_j^2 \beta_k \beta_l}{|h^{-1}(h_j)|^2} | h\right] \\
&= \frac{\sum_{k:h_k=h_j} \beta_k^2}{|h^{-1}(h_j)|^2} \geq \tau^2 \frac{|\mathcal{A} \cap h^{-1}(h_j)|}{|h^{-1}(h_j)|^2},
\end{aligned}$$

where $\tau = \min_{j:\beta_j \neq 0} |\beta_j|$. Using Lemma 4 we get for $\beta_j \neq 0$ (or $j \in \mathcal{A}$)

$$\begin{aligned}
\mathbb{E}[(\tilde{\beta}_j^{\mathrm{rs}})^2] \geq \tau^2 \mathbb{E}\left[\frac{|\mathcal{A} \cap h^{-1}(h_j)|}{|h^{-1}(h_j)|^2}\right] &= \tau^2 \mathbb{E}\left[\frac{1 + |\mathcal{A} \cap h^{-1}(h_j) \setminus \{j\}|}{|h^{-1}(h_j)|^2}\right] \\
&= \tau^2 \Bigg[\frac{m^2}{(p-1)^2} + \frac{(m-3)m^2}{(p-1)^3} \mathcal{O}(p^{-4}) \\
&\quad + m\frac{a-1}{p-1} \cdot \left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right)\Bigg] \\
&\geq \tau^2 \left[m\frac{a}{p-1} \cdot \left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right)\right]
\end{aligned}$$

and, for $\beta_j = 0$ (or $j \notin \mathcal{A}$)

$$\begin{aligned}
\mathbb{E}[(\tilde{\beta}_j^{\mathrm{rs}})^2] \geq \tau^2 \mathbb{E}\left[\frac{|\mathcal{A} \cap h^{-1}(h_j)|}{|h^{-1}(h_j)|^2}\right] &= \tau^2 \mathbb{E}\left[\frac{|\mathcal{A} \cap h^{-1}(h_j) \setminus \{j\}|}{|h^{-1}(h_j)|^2}\right] \\
&= \tau^2 \left[m\frac{a}{p-1} \cdot \left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right)\right].
\end{aligned}$$

Now we can find a lower bound on the expected squared norm as

$$\begin{aligned}
\mathbb{E}[\|\beta - \tilde{\beta}^{\mathrm{rs}}\|^2] = \mathbb{E}\left[\sum_{j=1}^{p} \left(\beta_j - \tilde{\beta}_j^{\mathrm{rs}}\right)^2\right] &= \sum_{j=1}^{p} \beta_j^2 - 2\beta_j \mathbb{E}[\tilde{\beta}_j^{\mathrm{rs}}] + \mathbb{E}[(\tilde{\beta}_j^{\mathrm{rs}})^2] \qquad (22) \\
&\geq \|\beta\|^2 \cdot \left(1 - \frac{2m}{p}\right) + \tau^2 ma\left(\frac{1}{p-1} - \frac{m+1}{(p-1)^2} + \mathcal{O}(p^{-3})\right). \qquad (23)
\end{aligned}$$

Upper bound for true coefficient: The additional assumption on the diagonal elements proportional to the true coefficient ensures that Φ_{pt} has full row rank. From Lemma 3, we see that $\tilde{\beta}^{\mathrm{pt}} = P_{\Phi}^{\mathrm{pt}} \beta$ for $P_{\Phi}^{\mathrm{pt}} = \Phi'_{\mathrm{pt}}(\Phi_{\mathrm{pt}}\Phi'_{\mathrm{pt}})^{-1}\Phi_{\mathrm{pt}}$ still equals

$$\tilde{\beta}_j^{\mathrm{pt}} = \begin{cases} c\beta_j \cdot \{\sum_{k:h_k=h_j} (c\beta_k)\beta_k / \sum_{k:h_k=h_j} c^2\beta_k^2\} = \beta_j & \beta_j \neq 0 \\ 0 \cdot \{\sum_{k:h_k=h_j} (c\beta_k)\beta_k / \sum_{k:h_k=h_j} c^2\beta_k^2\} = 0 & \beta_j = 0, \exists k \in h^{-1}(h_j) : \beta_k \neq 0 \\ d_j \cdot \{\sum_{k:h_k=h_j} d_k \overbrace{\beta_k}^{=0} / \sum_{k:h_k=h_j} d_k^2\} = 0 & \beta_j = 0, \forall k \in h^{-1}(h_j) : \beta_k = 0 \end{cases},$$

the true coefficient β in every case, implying $\beta = P_{\Phi}^{\mathrm{pt}}\beta \in \mathrm{span}(\Phi'_{\mathrm{pt}})$. As a short remark, here we see that the choice of diagonal elements $\{d_k : k \in h^{-1}(h_j)\}$ in the third case has no influence on the projection, as long as at least one is non-zero.

Similarly to before, we need to bound the expectation of $(\beta - P_X^{\text{pt}}\beta)'\Sigma(\beta - P_X^{\text{pt}}\beta)$, where $P_X^{\text{pt}} = \Phi_{\text{pt}}'(\Phi_{\text{pt}}X'X\Phi_{\text{pt}}')^{-1}\Phi_{\text{pt}}X'X$. Since $\beta \in \text{span}(\Phi_{\text{pt}}')$, we have $\beta = P_X^{\text{pt}}\beta$ and, therefore,

$$\mathbb{E}[(\beta - P_X^{\text{pt}}\beta)'\Sigma(\beta - P_X^{\text{pt}}\beta)] = 0. \tag{24}$$

Finally, we can put the results together to obtain

$$\begin{aligned}\mathbb{E}[(\tilde{y} - \hat{y}_{\text{rs}})^2] - \mathbb{E}[(\tilde{y} - \hat{y}_{\text{pt}})^2] &= \mathbb{E}[(\beta - P_{\text{rs}}\beta)'\Sigma(\beta - P_{\text{rs}}\beta)] - \mathbb{E}[(\beta - P_{\text{pt}}\beta)'\Sigma(\beta - P_{\text{pt}}\beta)] \\ &\geq \|\beta\|^2\lambda_p\left(1 - \frac{2m}{p}\right) + \frac{a}{p-1}m\lambda_p\tau^2\left(1 - \frac{m+1}{p-1} + \mathcal{O}(p^{-2})\right).\end{aligned}$$

□

Remark 2.

- When using diagonal elements just almost proportional to the true β, we can obtain the upper bound

$$\mathbb{E}[(\beta - P_X^{\text{pt}}\beta)'\Sigma(\beta - P_X^{\text{pt}}\beta)] \leq \lambda_1 \cdot \mathbb{E}\left[\|\beta - \tilde{\beta}^{\text{pt}}\|^2 \cdot \left(1 + \|P_X^{\text{pt}}\|^2\right)\right], \tag{25}$$

  where $\|P_X^{\text{pt}}\|$ is the spectral norm induced by the Euclidean norm growing bigger when $X'X$ is further away from the identity. As long as $\|\beta - \tilde{\beta}^{\text{pt}}\|^2$ is small enough such that this upper bound remains smaller than the obtained lower bound for random sign diagonal elements, we still have a theoretical guarantee for an average gain in prediction performance.

- We assumed $\mathbb{E}[y_i] = 0, \mathbb{E}[x_i] = 0$ for notational convenience in the proof. With a general center $\mathbb{E}[x_i] = \mu_x$ and intercept $\mu \neq 0$ as in (1), we can just use the centered X and y and the proof will work in a similar way for the same bound, but also needs to consider the estimation of the intercept $\hat{\mu} = \bar{y} - (\Phi\bar{x})'(Z'Z)^{-1}Z'y$ for both $Z = Z_{\text{rs}}, Z_{\text{pt}}$ and $\Phi = \Phi_{\text{rs}}, \Phi_{\text{pt}}$.

- The assumption of multivariate normal distribution for the predictors allows us to explicitly calculate $\mathbb{E}[(\Phi X'X\Phi')^{-1}|\Phi]\Phi\Sigma\Phi'$ from the Inverse-Wishart-distribution, but we could also allow any distribution, for which this expression does not depend on the choice of Φ.

- In the proof, we can see that the concrete adaption of diagonal elements to retain $\text{rank}(\Phi_{\text{pt}}) = m$ after Definition 1 is irrelevant, as long as there is at least one non-zero d_j with $j \in h^{-1}(i)$ for each $i \in [m]$. Our proposed adaption aims at adding minimal noise when we can not choose the diagonal elements exactly proportional to the true β (e.g., when we only use the sign information), while keeping Φ_{rs} not just full rank but also well-conditioned.

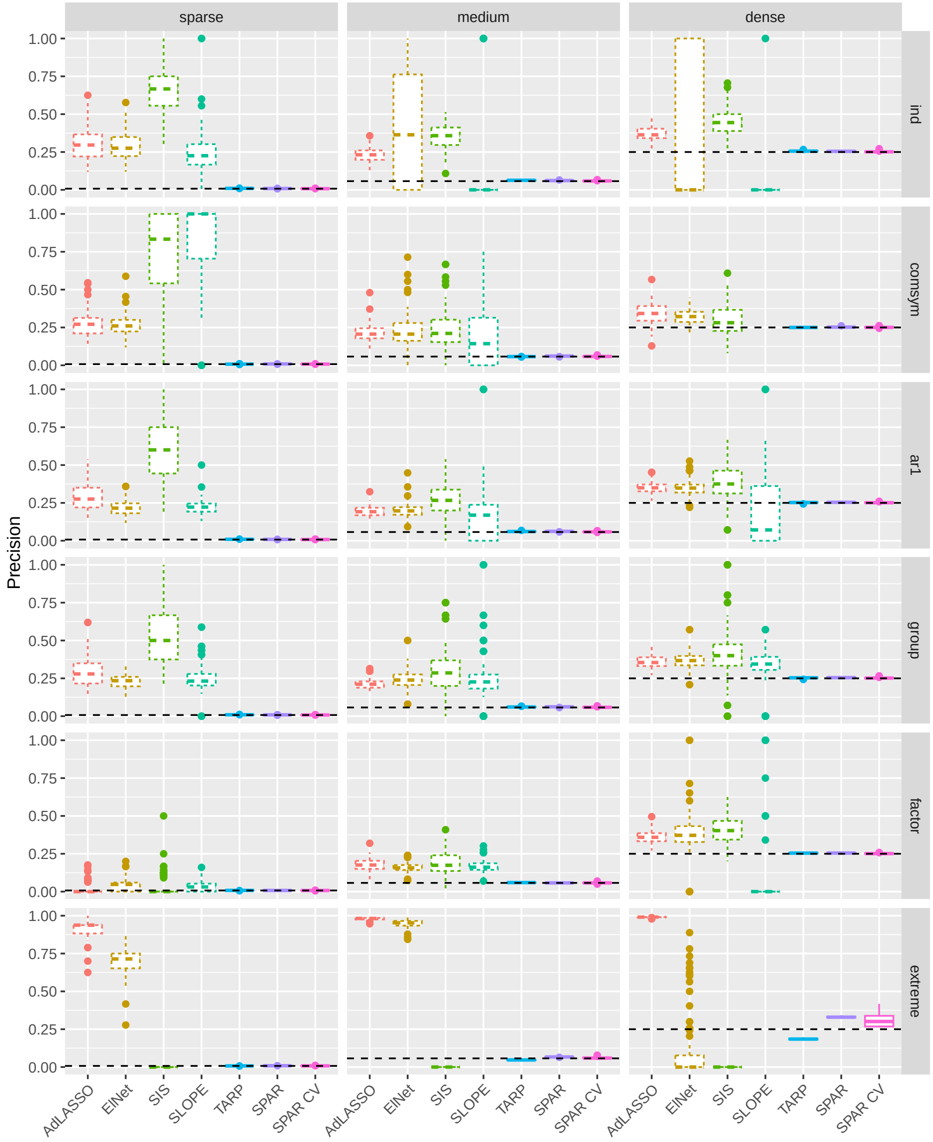


Figure 9: Precision of competing methods which perform any type of variable selection for different covariance and active predictor settings for $n_{\text{rep}} = 100$ replications ($n = 200, p = 2000, \rho_{\text{snr}} = 10$). Sparse methods are marked by dotted boxes.

C. Additional Details for Simulation Study

The following choices were considered for the covariance matrix Σ in Section 3.

1. *Independent predictors*: $\Sigma = I_p$.
2. *Compound symmetry structure*: $\Sigma = \rho 1_p 1_p' + (1-\rho) I_p$, where we set $\rho = 0.5$.
3. *Autoregressive structure*: The (i,j)-th entry is given by $\Sigma_{ij} = \rho^{|i-j|}$ and we choose $\rho = 0.9$. This structure is appropriate if there is a natural order among the predictors and two predictors with larger distances are less correlated, e.g., when they give measurements over time.

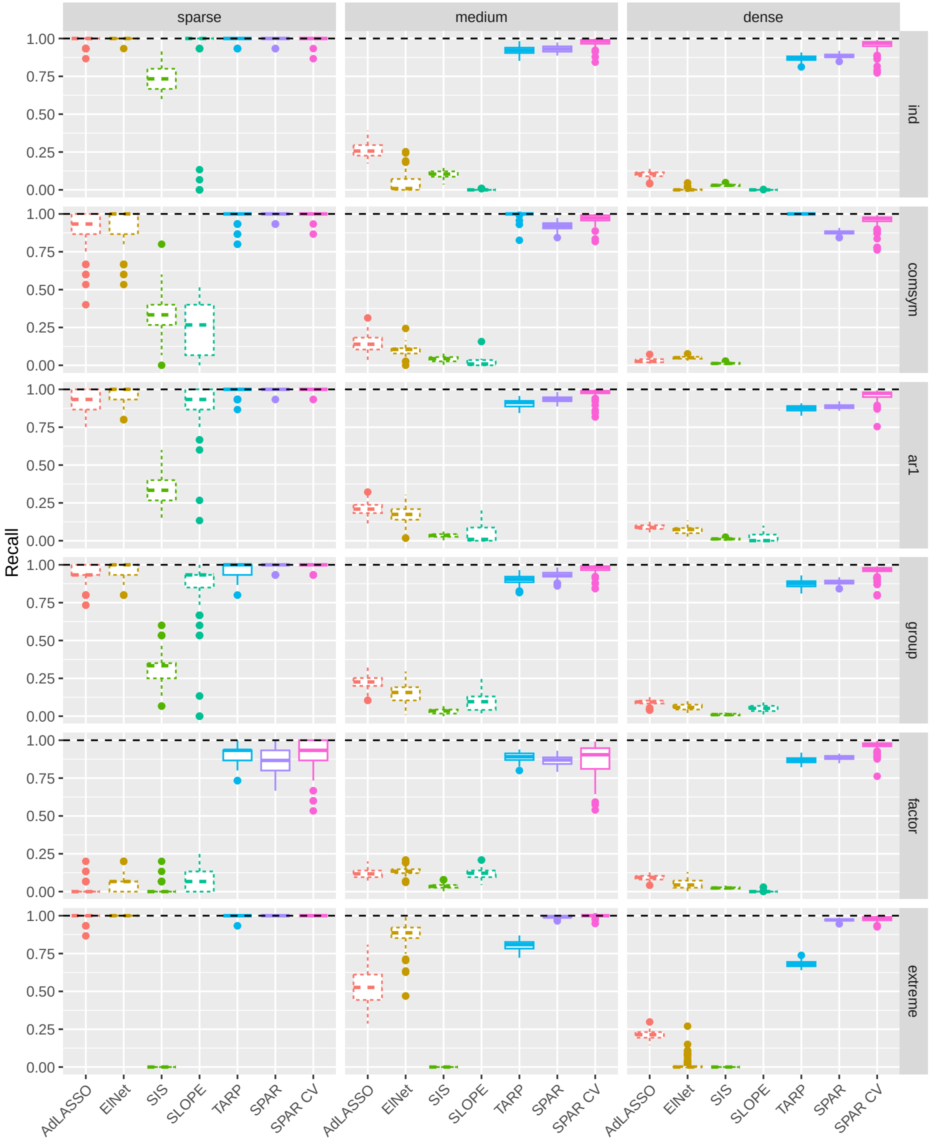


Figure 10: Recall of competing methods which perform any type of variable selection for different covariance and active predictor settings for $n_{\text{rep}} = 100$ replications ($n = 200, p = 2000, \rho_{\text{snr}} = 10$). Sparse methods are marked by dotted boxes.

4. *Group structure*: Similarly to scheme II in Mukhopadhyay and Dunson (2020), Σ follows a block-diagonal structure with blocks of 100 predictors each, where the first half of the blocks has the compound structure from setting 2 and the second half has the AR structure from setting 3. Only the very last block has identity structure corresponding to independent predictors within that block, and the predictors between different blocks are independent.

5. *Factor model*: Inspired by model 4.1.4. in Wang and Leng (2016), we first generate a $p \times k$ factor matrix F with $k = a$ and iid standard normal entries, and then set $\Sigma = FF' + 0.01 \cdot I_p$. Here, dimension reduction of the predictors will be useful, because most of the information lies within the k-dimensional subspace defined by F.

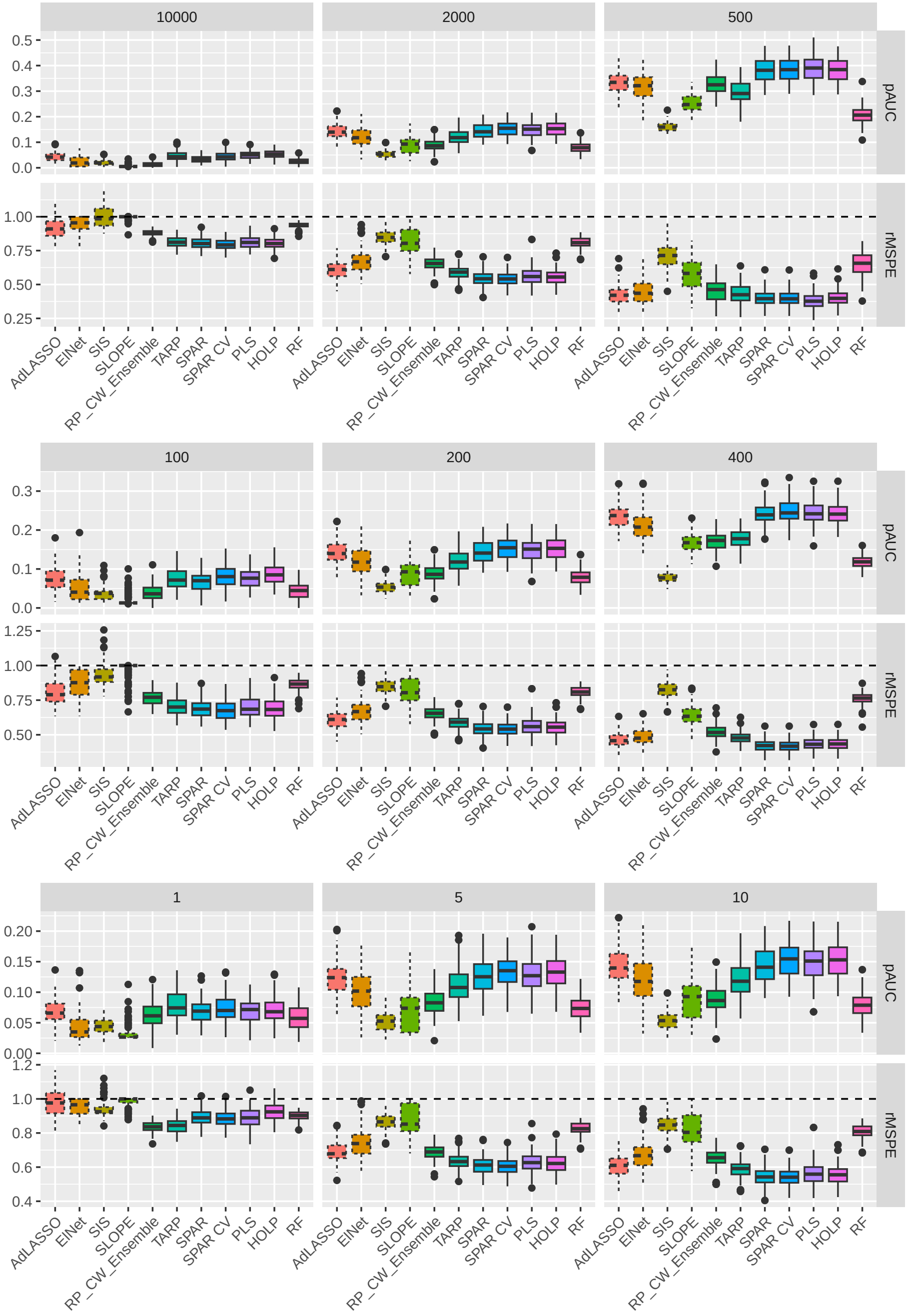


Figure 11: Performance measures of the competing methods, where the sparse methods are marked by dotted boxes, for 'group' covariance setting, medium setting for the active variables and $p = 500, 2000, 10000$ (top panel, $n = 200, \rho_{\text{snr}} = 10$), $n = 100, 200, 400$ (middle panel, $p = 2000, \rho_{\text{snr}} = 10$) and $\rho_{\text{snr}} = 1, 5, 10$ (bottom panel, $p = 2000, n = 200$) for $n_{\text{rep}} = 100$ replications.

6. *Extreme correlation*: This setting is designed such that methods relying on marginal correlations have difficulty in finding any true active predictor. Similarly to example 4 in Wang (2009), we create each predictor variable x_i the following way. For $i = 1, \dots, n$, let $z_{ij} \sim N(0, 1)$ be iid standard normal variables for $j = 1, \dots, p$ and $w_{ij} \sim N(0, 1)$ iid standard normal variables for $j = 1, \dots, a$ independent of

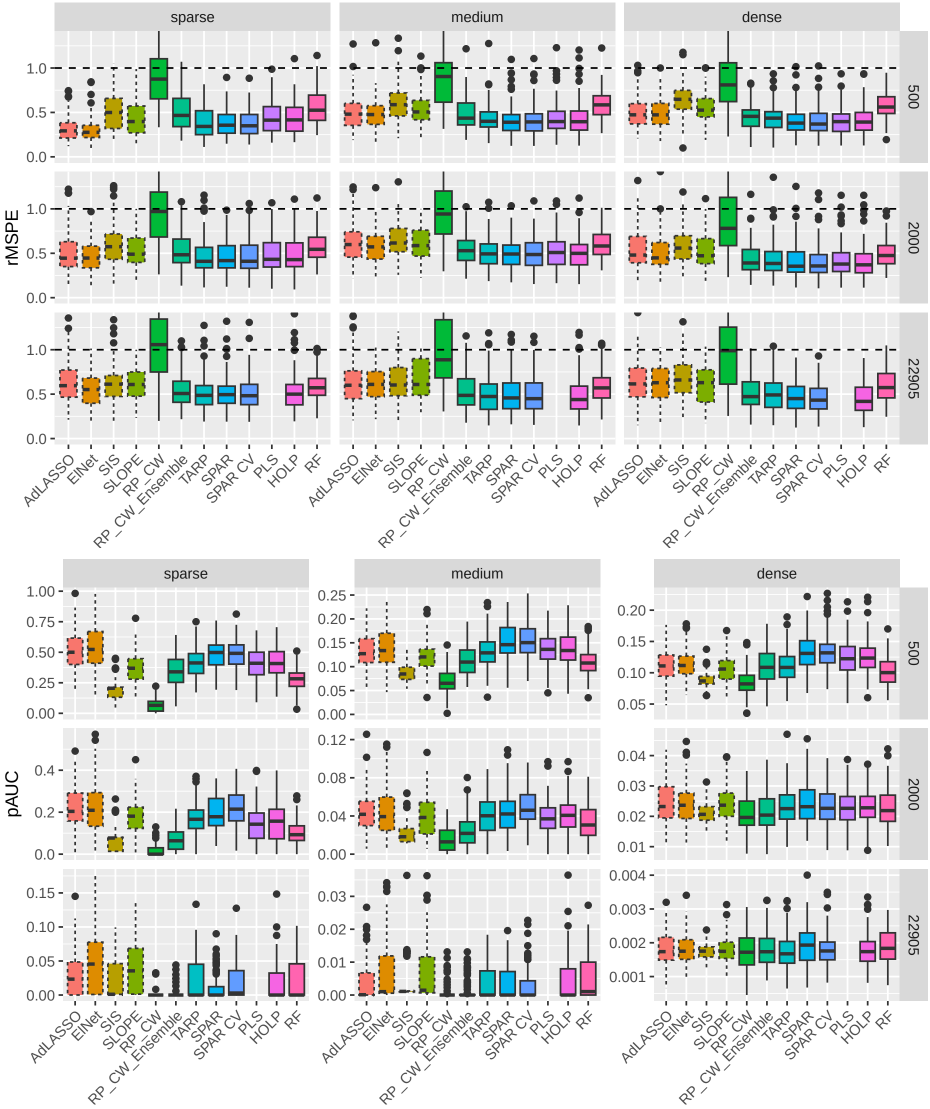


Figure 12: Relative MSPE and pAUC of the competing methods for different active predictor settings and different number of variables $p = 500, 2000, 22905$ over $n_{\text{rep}} = 100$ synthetic datasets using the observed genes in the gene expression dataset. Sparse methods are marked by dotted boxes.

the z_{ij}s. We then set

$$x_{ij} = \begin{cases} (z_{ij} + w_{ij})/\sqrt{2} & j \leq a \\ (z_{ij} + \sum_{k=1}^{a} z_{ik})/\sqrt{a+1} & j > a \end{cases}.$$

The marginal correlation of any active predictor $x_j, j \leq a$ to the response is way smaller than that of any unimportant predictor $x_k, k > a$. The exact ratio between them is $(j/a) \cdot 2^{-3/2} \cdot (a+1)^{-1/2} < 1$ for $j = 1, \ldots, a$.

Next, we include the additional Figures 9,10 and 11 for the simulation results mentioned and explained in Section 3.4.

D. Simulation Study with Synthetic Rat Eye Data

We employ a simulation setting similar to the one in Section 3 where we use the first $p = 500, 2000, 22905$ genes from our filtered gene expression data as predictors, construct sparse, medium, and dense coefficient vectors β as in Section 3.1, and generate a synthetic response with mean $\mu = 1$ from the predictors with noise level chosen such that the signal-to-noise ratio based on the empirical predictor covariance is 10.

Figure 12 shows the relative MSPE and the partial AUC for sparse, medium, and dense settings and the different values of p over 100 replications. We observe that SPAR performs well in all settings, even in the sparse ones, especially for the case $p = 22{,}905$. In contrast to the simulation scenarios in Section 3, the sparse methods do not perform clearly better in the sparse settings. This highlights the strong impact of the correlation structure on the performance of the methods.

Affiliation:

Roman Parzer
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria
E-mail: romanparzer1@gmail.com
GitHub: https://github.com/RomanParzer?tab=repositories

Peter Filzmoser
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria
E-mail: peter.filzmoser@tuwien.ac.at

Laura Vana-Gür
Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria
E-mail: laura.vana.guer@tuwien.ac.at

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
May 2025, Volume V, Issue V *Submitted:* 2025-01-09
doi:10.52933/jdssv.v5i5.138 *Accepted:* 2025-04-15