Data-Driven Random Projection and Screening for High-Dimensional Generalized Linear Models

Roman Parzer Peter Filzmoser Laura Vana-Gür

Institute of Statistics and Mathematical Methods in Economics
TU Wien
Vienna, Austria
romanparzer1@gmail.com

October 1, 2024

Abstract

We address the challenge of correlated predictors in high-dimensional GLMs, where regression coefficients range from sparse to dense, by proposing a data-driven random projection method. This is particularly relevant for applications where the number of predictors is (much) larger than the number of observations and the underlying structure – whether sparse or dense – is unknown. We achieve this by using ridge-type estimates for variable screening and random projection to incorporate information about the response-predictor relationship when performing dimensionality reduction. We demonstrate that a ridge estimator with a small penalty is effective for random projection and screening, but the penalty value must be carefully selected. Unlike in linear regression, where penalties approaching zero work well, this approach leads to overfitting in non-Gaussian families. Instead, we recommend a data-driven method for penalty selection. This data-driven random projection improves prediction performance over conventional random projections, even surpassing benchmarks like elastic net. Furthermore, an ensemble of multiple such random projections combined with probabilistic variable screening delivers the best aggregated results in prediction and variable ranking across varying sparsity levels in simulations at a rather low computational cost. Finally, three applications with count and binary responses demonstrate the method's advantages in interpretability and prediction accuracy.

Keywords Generalized Linear Models · High-Dimensional Data · Predictive Modeling · Random Projection · Screening

1 Introduction

High-dimensional data in a regression context, where the number of variables exceeds the number of observations (i.e., $p > n$ or even $p \gg n$), has become increasingly common across various applications, posing substantial computational and statistical challenges, particularly when dealing with discrete responses. In such cases, predictors are often correlated and the sparsity of the true model is uncertain. Moreover, interpretability is increasingly becoming a model requirement in a variety of fields. This calls for computationally efficient approaches that enable both accurate predictions and interpretable relationships between the predictors and the response.

The generalized linear model (GLM) extends the linear model to both continuous and discrete responses while maintaining interpretability. In high-dimensional settings, GLMs are typically regularized [e.g., Tibshirani, 1996, Fan and Li, 2001, Zou and Hastie, 2005]. Alternatively or complementarily, the dimensionality of the feature space can be reduced to a moderate size while learning and inference is performed in this reduced predictor space. One fast way to achieve this is *variable screening*, i.e., selecting a subset of the predictors based on their utility. Methods for screening often rely on parametric [such as the maximum likelihood estimates of univariate GLMs, Fan et al., 2009, Fan and Song, 2010] or nonparametric [e.g., Fan et al., 2011, Mai and Zou, 2013, 2015, Ke, 2023] measures but typically ig-

nore predictor correlations. Fan et al. [2009] suggest an iterative procedure to address this issue, while Wang and Leng [2016] propose screening variables in linear regression using the high dimensional ordinary least squares projection (HOLP), a ridge-type estimator with a closed form solution when the penalty converges to zero. However, screening approaches based on ridge-type estimators are still rare in the context of GLMs.

An approach similar in scope to variable screening is random projection (RP), which reduces the dimensionality of the feature space by linearly projecting the features onto a lower dimensional space, rather than employing a reduced set of the original features. Conventional random projections contain iid entries from a suitable distribution and are oblivious to the data to be used in the regression. Such projections have been used in a classification setting in e.g., Cannings and Samworth [2017], while Guhaniyogi and Dunson [2015], Mukhopadhyay and Dunson [2020] focus on linear regression. On the other hand, Ryder et al. [2019] propose a data-informed random projection using an asymmetric transformation of the predictor matrix without using information of the response. Parzer et al. [2024] propose a data-driven projection for linear regression which incorporates information from the estimated HOLP coefficients, i.e., about both the predictors and the response.

In this paper, we leverage the computational advantages of variable screening and random projection and introduce a data-driven random projection method for GLMs that accounts for the relationship between predictors and the response, while addressing the potentially complex correlation structure among the predictors. For this purpose we propose a ridge-type estimator which can be integrated into a sparse random projection matrix and can also be employed for screening the variables prior to projection, as it performs well in preserving the true relationship between predictors and the response.

A key aspect of the proposed ridge-type estimator is the selection of the penalty term. We extend the HOLP estimator to GLMs with canonical link, deriving a closed form solution, and explore if it retains the same benefits for random projection and screening in GLMs as in linear regression. We find that for non-Gaussian families, a ridge estimator with zero penalty can overfit, making penalty selection a non-trivial task. While small penalty values reduce bias, a data-driven approach to choosing the penalty works best. Specifically, we propose selecting the smallest penalty value for which the deviance ratio in the fit stays under a certain threshold (e.g., 0.8 for non-Gaussian families and 0.999 for Gaussian). Simulations show that using these ridge estimates in the sparse RP matrix outperforms conventional RP techniques.

Given the randomness in the RP matrix, variability can be reduced by building ensembles of multiple RPs. For example, Guhaniyogi and Dunson 2015 propose repeatedly sampling RPs of different size and estimating an ensemble of linear regressions on the reduced predictors, while Cannings and Samworth 2017 generate various RPs for each classifier in an ensemble and pick the best one based on a appropriate loss function. Additionally, ensembles of multiple RPs and variable screening steps have also been proposed achieve better predictive performance in linear regression [Mukhopadhyay and Dunson, 2020, Parzer et al., 2024].

In a similar fashion we extend the variable screening and random projection procedure by building an ensemble of GLMs and averaging them to form the final model, adapting the algorithm in Parzer et al. [2024] to GLMs. An extensive simulation study reveals that this ensemble improves predictive performance, ranks predictors effectively, and is computationally efficient, particularly with increasing predictor dimensions. It consistently performs well against state-of-the-art approaches such as penalized regression, random forests, and support vector machines (SVMs) across a range of sparsity settings and yields the best overall performance when aggregated across all scenarios. This broad applicability makes the method versatile for high-dimensional regression with correlated predictors, especially when the sparsity of the underlying data generating process is unknown, and capable to computationally handle datasets with small n and a (very) large number of predictors.

The integration of the GLM framework with probabilistic screening improves model interpretability, as the coefficients in the marginal models can be extracted and their reliability and relevance can be assessed. The GLM framework also offers modeling flexibility, facilitating seamless comparison across different family-link combinations.

The paper is organized as follows: Section 2 introduces the GLM model, the proposed data-informed random projection and variable screening in GLMs as well as the ensemble algorithm. In Section 3, extensive simulations motivating and comparing the method with state-of-the-art approaches are performed. Applications are presented in Section 4. Section 5 concludes the paper.

2   Method

This section begins by presenting the GLM model class, followed by an introduction to the dimension reduction tools random projection and variable screening. Then, we propose a novel coefficient estimator useful for both these

concepts, and state, how this estimator can be used to extend the algorithm in Parzer et al. [2024] to GLMs. Throughout the section, we use the notation $[n] = \{1, \dots, n\}$ for any $n \in \mathbb{N}$.

2.1 Generalized linear models

We assume to observe high-dimensional data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}, \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ with $p \gg n$ from a GLM with the responses having conditional densities from a (reproductive) exponential dispersion family of the form

$$f(y_i|\theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \tag{1}$$

where θ_i is the natural parameter, $a(.) > 0$ and $c(.)$ are specific real-valued functions determining different families, ϕ is a dispersion parameter, and $b(.)$ is the log-partition function normalizing the density to integrate to one. If ϕ is known, we obtain densities in the natural exponential family for our responses. It can be shown that $b(.)$ is twice differentiable and convex with $\mathbb{E}[y_i|\theta_i, \phi] = b'(\theta_i)$ and $\text{Var}(y_i|\theta_i, \phi) = a(\phi)b''(\theta_i) > 0$, if the responses have (positive) second moments [see e.g. McCullagh and Nelder, 1989, Section 2.2.2].

The responses are related to the p-dimensional predictors through the conditional mean, i.e., the conditional mean of y_i given $\boldsymbol{x}_i$ depends on a linear combination of the predictors through a (invertible) link function $g(.)$

$$g(\mathbb{E}[y_i|\boldsymbol{x}_i]) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} =: \eta_i, \tag{2}$$

where $\beta_0 \in \mathbb{R}$ is the intercept and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients. Equations (1) and (2) give the functional relation $\theta_i = \theta_i(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}_i) = (b')^{-1}(g^{-1}(\eta_i))$ between θ_i and η_i. For each family, $g := (b')^{-1}$ is the canonical link function, such that $\theta_i = \eta_i$. The full log-likelihood of the regression parameter $\boldsymbol{\beta}$ given the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is

$$\ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(\theta_i(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}_i))}{a(\phi)} + c(y_i, \phi),$$

but for maximization with respect to $\boldsymbol{\beta}$, it suffices to use

$$\tilde{\ell}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i\theta_i(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}_i) - b(\theta_i(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}_i)),$$

and treat ϕ as a nuisance parameter. In our general high-dimensional setting $p > n$, the predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with rows $\boldsymbol{x}_i$ can be assumed to have full $\text{rank}(\boldsymbol{X}) = n$, so there is typically a whole (affine) subspace of $\boldsymbol{\beta}$s yielding the same η_is and we can not hope to find a unique maximizer $\hat{\boldsymbol{\beta}}$. In order to reduce the dimension of the problem, we resort to two techniques namely, random projection and variable screening.

2.2 Random projection and variable screening

Random projection can be used as a dimension-reduction tool for high-dimensional regression by creating a random matrix $\Phi \in \mathbb{R}^{m \times p}$ with $m \ll p$ and using the reduced predictors $\boldsymbol{z}_i = \Phi\boldsymbol{x}_i \in \mathbb{R}^m$ in a regression model. When using this method for GLMs, we would like the predictors to still have most of the predictive power and that the true regression coefficients $\boldsymbol{\beta}$ are close to the row span of Φ, such that they can be approximately recovered by the reduced predictors. For this purpose, we propose to employ the following random projection.

Definition 2.1. *Let $h : [p] \to [m]$ be a random map such that for each $j \in [p] : h(j) = h_j \overset{iid}{\sim} \text{Unif}([m])$. Let $\boldsymbol{B} \in \mathbb{R}^{m \times p}$ be a binary matrix with $B_{h_j, j} = 1$ for all $j \in [p]$ and remaining entries 0, where we assume $\text{rank}(\boldsymbol{B}) = m$. Let $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ be a diagonal matrix with entries $d_j \in \mathbb{R} \setminus \{0\}, j \in [p]$. Then we call $\boldsymbol{\Phi} = \boldsymbol{BD}$ a CW random projection.*

When using random sign diagonal elements $d_j \sim \text{Unif}(\{-1, 1\})$ independent of h, we obtain a *sparse embedding matrix* $\boldsymbol{\Phi} \in \mathbb{R}^{m \times p}$, $m \ll p$ from Clarkson and Woodruff [2013]. Aside from being sparse and computationally efficient, this random projection also exhibits the property $\boldsymbol{d} = (d_1, \dots, d_p)' \in \text{span}(\boldsymbol{\Phi'})$. Thus, by choosing $d_j \propto \beta_j, j = 1, \dots, p$ instead of random sign diagonal elements, we can reach our goal of combining the variables to reduced predictors with strong predictive power which are able to recover the true regression coefficient. In Theorem 1 of Parzer et al. [2024], it is shown that this approach significantly reduces the expected squared error of future predictions in the linear regression setting. Below, we will propose a new estimator for a general family to use as diagonal elements.

Variable screening   aims at selecting a small subset of variables based on some marginal utility measure, and using the ones with the highest utility for further analysis. This procedure can complement the RP approach and further reduce the dimensionality of the problem, by first screening for the important variables and then performing the random projection.

A seminal contribution in variable screening is the sure independence screening (SIS) of Fan and Lv [2008], who proposed to use the absolute marginal correlation of the predictors to the response in linear regression, which was later extended to GLMs in Fan et al. [2009], Fan and Song [2010] by employing the maximum likelihood coefficient estimates of univariate GLMs instead of the correlation coefficient. However, in the presence of predictor correlation, screening based on a conditional utility measure (i.e., conditional on all the other variables in the model) is to be preferred to the unconditional one. To tackle this issue, Fan and Lv [2008] propose iterative SIS, which involves iteratively applying SIS and penalized regression to select a small set of variables, computing the residuals of a model fitted with the selected variables and using these residuals residuals as a response variable to continue finding relevant variables. This procedure was later extended to more general model classes in Fan et al. [2009].

In general, in a GLM each variable's utility given all the other variables amounts to $|\beta_j|$, so another approach is to find a screening coefficient which capable of detecting the correct order of magnitudes of the regression coefficients (but not necessarily their signs). We note that an estimator which performs well for the purpose of random projection (as discussed above) would also be a good candidate as a screening coefficient. In the next section, we propose such an estimator for GLMs.

2.3   Proposed estimator for random projection and screening

In general, a ridge-type estimator

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \min_{\beta_0\in\mathbb{R}} \left\{ -\tilde{\ell}(\beta_0,\boldsymbol{\beta}) + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2 \right\}, \quad \lambda > 0 \tag{3}$$

promises to be a sensible choice, both for screening and for inclusion in the CW random projection, because it considers all variables in the model and is non-sparse. For linear regression, the HOLP estimator [Wang and Leng, 2016] provides a screening utility measure considering all variables' effects simultaneously and has strong theoretical properties (see Section 2.1 in Parzer et al. [2024] for a discussion). It is explicitly given by

$$\hat{\boldsymbol{\beta}}_{\text{HOLP}} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{y} = \lim_{\lambda\to 0}\left(\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2\right\}\right), \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the predictor matrix and $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector. Here, a model without an intercept β_0 is assumed, which can be justified by centering $\boldsymbol{X}$ and $\boldsymbol{y}$. Parzer et al. [2024] used this HOLP coefficient (4) as the diagonal elements in a CW random projection.

Motivated by (4) and Kobak et al. [2020], who showed that the optimal ridge-penalty in linear regression can be negative due to implicit regularization from high-dimensional predictors, we first investigate whether $\lim_{\lambda\to 0}\hat{\boldsymbol{\beta}}_\lambda$ is also an appropriate choice in GLMs. For logistic regression (binomial family with logit link), it is known that the estimator (3), scaled by its norm, converges to a hard-margin support vector machine (SVM) coefficient for $\lambda \to 0$ (Theorem 3 in Rosset et al. [2004], see Section 3.6.1 in Hastie et al. [2015] for a discussion). More generally, the following Theorem shows an explicit form of $\lim_{\lambda\to 0}\hat{\boldsymbol{\beta}}_\lambda$ for families with canonical link function such that $g(y_i) \in \mathbb{R}$ for all $i \in [n]$.

Theorem 2.2. *For a family with canonical link function satisfying $g(y_i) \in \mathbb{R}$ for all $i \in [n]$, for a full rank predictor matrix rank$(\boldsymbol{X}\boldsymbol{X}') = n$, and in a model without an intercept β_0, we obtain*

$$\lim_{\lambda\to 0}\hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}g(\boldsymbol{y}). \tag{5}$$

The proof can be found in the Appendix A. The intercept-free assumption can be avoided by appropriate centering of $\boldsymbol{X}$ and $g(\boldsymbol{y})$. For practical usage, this exact limit has a few drawbacks. For example, when centering by the exact sample mean, $\boldsymbol{X}$ would not have full rank n anymore. A workaround could be employing a generalized inverse or using different location estimators (like median or trimmed mean) for the purpose of centering the variables. In the cases where $g(y_i) \notin \mathbb{R}$, one could approximate $g(y_i)$ by using a continuity correction, but there is no guarantee that this approximation works well. Also, Theorem 2.2 only covers the canonical link for each family.

As an alternative that can cover all cases, we propose to approximate $\lim_{\lambda\to 0}\hat{\boldsymbol{\beta}}_\lambda$ by a ridge estimator with a small fixed $\lambda_{\min} > 0$ and to use $\hat{\boldsymbol{\beta}}_{\lambda_{\min}}$ from (3) for variable screening and as the diagonal elements in the data-informed random projection.

Furthermore, in order to understand whether more penalization (i.e., through higher values of λ) is needed for other families as compared to the Gaussian case, we investigate alternative strategies to choosing the penalty value. In a simulation example in Section 3.3, we investigate the choice of this $\lambda_{\min}$ for different families. It shows, that for non-Gaussian families, it is beneficial to use a higher $\lambda_{\min}$ to avoid a saturated fit. We also show that the resulting estimator allows good recovery of sign and magnitude of the true non-zero coefficients, while also performing well in terms of prediction when employed in the RP matrix.

2.4 SPAR algorithm for GLMs

Employing one single data-driven random projection with the proposed estimator and then estimating a GLM on the reduced predictors can lead to high variability due to randomness. We address this by adapting the Sparse Projected Averaged Regression (SPAR) algorithm from Parzer et al. [2024] to GLMs, which builds an ensemble of GLMs in the following way: i) randomly sampling predictors for inclusion in the random projection based on the proposed screening coefficient, ii) projecting the sampled variables to a randomly chosen lower dimension using the proposed random projection, iii) estimating penalized GLMs with the reduced predictors and iv) averaging them to form the final model. The adapted algorithm is given below, where $*$ mark changes compared to the linear regression formulation:

1.* choose family with corresponding log-likelihood $\ell(.)$ and link and standardize covariate inputs $\boldsymbol{X} : n \times p$

2.* calculate $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_{\lambda_{\min}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \min_{\beta_0 \in \mathbb{R}} \left\{ -\tilde{\ell}(\beta_0, \boldsymbol{\beta}) + \frac{\lambda_{\min}}{2} \sum_{j=1}^p \beta_j^2 \right\}$, see below for choice of $\lambda_{\min} > 0$

3. For $k = 1, \ldots, M$:

     3.1. draw $2n$ predictors with probabilities $p_j \propto |\hat{\alpha}_j|$ yielding screening index set $\boldsymbol{I}_k = \{j_1^k, \ldots, j_{2n}^k\} \subset [p]$

     3.2. project remaining variables to dimension $m_k \sim \mathrm{Unif}\{\log(p), \ldots, n/2\}$ using $\boldsymbol{\Phi}_k : m_k \times 2n$ from Definition 2.1 with diagonal elements $d_i = \hat{\alpha}_{j_i^k}$ to obtain reduced predictors $\boldsymbol{Z}_k = \boldsymbol{X}_{\cdot I_k} \boldsymbol{\Phi}_k' \in \mathbb{R}^{n \times m_k}$

     3.3.* fit a GLM of $\boldsymbol{y}$ against $\boldsymbol{Z}_k$ (with small L_2-penalty) to obtain estimated coefficients $\boldsymbol{\gamma}^k \in \mathbb{R}^{m_k}$ and $\hat{\boldsymbol{\beta}}^k$, where $\hat{\boldsymbol{\beta}}_{I_k}^k = \boldsymbol{\Phi}_k' \boldsymbol{\gamma}^k$ and $\hat{\boldsymbol{\beta}}_{\bar{I}_k}^k = 0$.

4. for a given threshold $\nu > 0$, set all entries $\hat{\beta}_j^k$ with $|\hat{\beta}_j^k| < \nu$ to 0 for all j, k

5. combine via simple average on link-level $\hat{\boldsymbol{\beta}} = \sum_{k=1}^M \hat{\boldsymbol{\beta}}^k / M$ or on response level $\hat{\boldsymbol{y}} = \sum_{k=1}^M \hat{\boldsymbol{y}}^k / M$

6.* *optional:* choose M and ν via 10-fold cross-validation by repeating steps 2 to 6 (with fixed index sets $\boldsymbol{I}_k$ and projections $\boldsymbol{\Phi}_k$) for each fold and evaluating the performance by model deviance (Dev) on the withheld fold; and choose

$$(M_{\mathrm{best}}, \nu_{\mathrm{best}}) = \arg\min_{M, \nu} \mathrm{Dev}(M, \nu)$$

7. output the estimated coefficients and predictions for the chosen M and ν

We propose the following strategy for choosing the penalty $\lambda_{\min}$. Along a decreasing path of λs (e.g., equally spaced on the logarithmic scale), the fitted model will get closer and closer to a saturated fit and a deviance ratio of one. We propose to use the smallest λ, where the deviance ratio does not exceed to a certain threshold. It turns out that this threshold can be set to a value close to one (e.g., 0.999) for the Gaussian family but to a lower value (e.g., 0.8) for other families, as we will show in subsequent simulations.

In fitting the marginal models in step 3.3, we also obtain intercept estimates, which can also be averaged and translated back to the original predictor scale to give an overall estimate $\hat{\beta}_0$ of the intercept β_0.

This algorithm allows for several variations. For instance, different measures can be used in the cross-validation, such as mean squared error instead of the family-dependent deviance. If stricter thresholding is desired, (M, ν) can be chosen by the one-standard-error rule, which yields the sparsest $\hat{\boldsymbol{\beta}}$ within one standard error of the score of the best parameters. While averaging in step 6 can also be done on response-level, simulations showed no performance gain in any setting, and averaging at the linear predictor-level is more interpretable, as a single final coefficient estimate $\hat{\boldsymbol{\beta}}$, as well as a distribution of each $\hat{\beta}_j$ over the marginal models can be reported.

3 Simulation study

In a first simulation study we compare how well the estimator (3) recovers the true active $\boldsymbol{\beta}$ across different penalty choices, and evaluate the predictive performance of the data-driven random projection with these estimators. We then compare SPAR algorithm's predictive performance and variable ranking ability against various benchmarks in a comprehensive simulation study.

3.1 Setup

We generate data from Equation (1) for five family-link combinations with $n = 200$ and use additionally $n_{\text{test}} = 200$ observations as a test sample. The p-dimensional predictors are simulated as $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where we investigate different structures for $\boldsymbol{\Sigma}$: *identity*, *compound* with $\Sigma_{ij} = 0.5$ if $i \neq j$ and 1 otherwise, *autocorrelated* with $\Sigma_{ij} = 0.9^{|i-j|}$, and a *block* structure (blocks of size 100 where half of the blocks have a compound, the remaining blocks except the last have an autocorrelated structure, plus independent predictors in the last block.

We consider $p = 500, 2000, 10000$ and three sparsity settings for $\boldsymbol{\beta}$: sparse ($a = [2\log(p)]$), medium ($a = [2\log(p) + n/2]$) and dense ($a = p/4$), where a is the number of non-zero entries in $\boldsymbol{\beta}$. These entries are independently set as $(-1)^u (4\log(n)/\sqrt{n} + |z|)$ at uniformly random positions, where $u \sim \text{Bernoulli}(0.4)$ and $z \sim N(0, 1)$ [Fan and Lv, 2008]. Finally, for each family-link we rescale $\boldsymbol{\beta}$ to control the signal strength by $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} = c$, where $c = 100, 1000, 0.25, 10, 0.125$ for binomial-logit, binomial-complementary log log (cloglog), Poisson-log, Gaussian-identity and Gaussian-log, respectively. The intercept β_0 is set such that $\sum_{i=1}^{n} \mathbb{E}[y_i|\boldsymbol{x}_i]/n = 0.5, 0.7, 10, 1, 10$ for the respective family-links. These values are chosen to ensure that the problem is not too easy to solve but also that most methods can explain some of the deviance.

3.2 Measures

Prediction performance is assessed on the independent test samples $\{(\boldsymbol{x}_{n+i}, y_{n+i}) : i \in [n_{\text{test}}]\}$ by several measures: *mean squared prediction error* and its relative version

$$\text{MSPE}(\hat{\boldsymbol{y}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_{n+i} - \hat{y}_{n+i})^2, \quad \text{rMSPE}(\hat{\boldsymbol{y}}) = n_{\text{test}} \text{MSPE}(\hat{\boldsymbol{y}}) / \sum_{i=1}^{n_{\text{test}}} (y_{n+i} - \bar{y})^2$$

where $\bar{y}$ is the mean of the responses in the training sample (for the binomial family, MSPE corresponds to the Brier score). We also compute *mean squared link estimation error* to assess the linear predictor accuracy

$$\text{MSLE}(\hat{\boldsymbol{\beta}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\eta_i - \hat{\eta}_i)^2, \quad \text{rMSLE}(\hat{\boldsymbol{\beta}}) = n_{\text{test}} \frac{\text{MSLE}(\hat{\boldsymbol{\beta}})}{\sum_{i=1}^{n_{\text{test}}} (\hat{\eta}_i)^2},$$

where $\eta_i = \beta_0 + \boldsymbol{x}'_{n+i}\boldsymbol{\beta}$ and $\hat{\eta}_i = \hat{\beta}_0 + \boldsymbol{x}'_{n+i}\hat{\boldsymbol{\beta}}$. For the binomial family we also consider the area under the receiver operating characteristic curve (AUC) of the predicted probabilities $\hat{p}_{n+i} = 1/(1 + \exp(-\hat{\eta}_i))$ to the binary responses.

Furthermore, we assess variable ranking using the *partial AUC* (pAUC) which considers whether a variable is truly active and the absolute value of their estimated coefficient. To allow fair comparison between sparse methods and methods delivering a dense coefficient vector, we limit false positive to $n/2$ [see also Parzer et al., 2024].

3.3 Simulations for screening and random projection

Recovery of true active $\boldsymbol{\beta}$ We investigate whether the ridge estimator in (3) is appropriate for the purpose of screening and data-informed random projection in the sense that it is able to recover the true active coefficients well by considering the correlation to the true active $\boldsymbol{\beta}$. We present results for $n = 200, p = 2000$, medium sparsity and $\boldsymbol{\Sigma}$ block-diagonal. Figure 1 shows the distribution of the correlation coefficient of the true active coefficients to different screening coefficients over 100 replications. We consider the ridge (L2) estimator from (3) with the following choices for λ: chosen by cross-validation based on the deviance criterion (L2_cv); fixed at 10 (L2_10), 1 (L2_1), 0.1 (L2_01), 0.01 (L2_001); the estimates for the penalty converging to zero (L2_limit0). For binomial-logit we use the exact limit for $\lambda \to 0$, which is a hard-margin linear support-vector-machine (SVM) coefficient (estimated with R package **e1071** with cost set to 10^{10}; Meyer et al. 2023). For the other families we use Equation 5 and impute any zeros present in the response variable by a small positive value for the poisson family. Furthermore, we employ a data-driven approach to choosing λ as the smallest value for which the fraction of (null) deviance explained does not exceed 80% (L2_dev08), 95% (L2_dev095) and 99.9% (L2_dev0999). Finally, we examine the screening coefficient in Fan and Song [2010], where each β_j is estimated by the slope of a marginal GLM of the corresponding predictor $\{x_{ij}, i \in [n]\}$ with an intercept (marGLM).

We observe that L2_limit0 does not deliver the best results for all investigated family links (we omit the results for binomial-cloglog as they are similar to binomial-logit). L2_cv also underperforms, and the marginal GLM coefficients are the least effective at recovering the true coefficients. Fixing λ to predetermined values is also not appropriate for all family links, since the strength of the penalty varies between families. Table 2 in Appendix shows the average resulting λ for each family using cross-validation and the different deviance cut-offs. For binomial, differences among the ridge estimators are minor, but for Poisson the performance declines as the deviance ratio threshold increases or
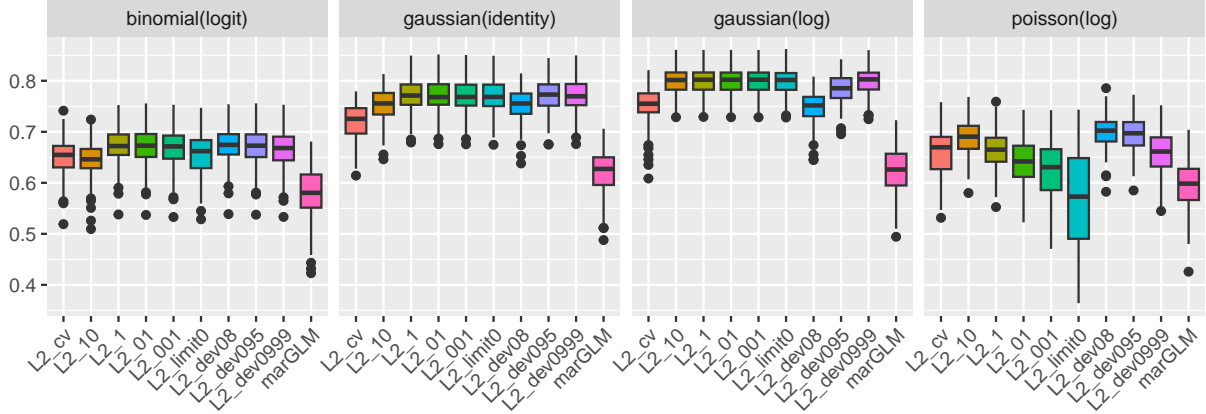
Figure 1: Correlation of true active coefficients to different screening estimators for 100 replications ($n = 200, p = 2000$, medium sparsity and block-diagonal Σ).

λ becomes too value. While the cross-validated penalty seems to be too large for the purpose of screening, using a 0.8 deviance ratio for non-Gaussian families and 0.999 for Gaussian responses yields good results. The corresponding estimates are therefore well-suited as diagonal elements approximately proportional to the true β in our data-informed random projection from Definition 2.1. In this comparison, we also computed the pAUC of these coefficients, and the ratio of true active variables within the highest $3a$ absolute estimated values, but omit these results as they are similar to those based on correlation.

Data-informed random projection Next, we investigate the predictive performance of a model where the predictors are first projected onto a $m = n/4 = 50$ lower dimensional space using the proposed data-informed random projection with different screening coefficients, namely L2_cv, L2_limit0, L2_dev08, L2_dev095 and L2_dev0999 from the previous section. Additionally, we also show the oracle performance of our proposal using the true β in the random projection (True_Beta). We also consider models where conventional RPs (i.e., Gaussian with iid standard normal entries and SparseCW from Definition 2.1 with random sign diagonal elements) are used. Furthermore, we estimate the models with adaptive LASSO [Zou, 2006] and elastic net ($\alpha = 3/4$) with penalty chosen by CV based on deviance as performance benchmarks.

In Figure 2 we present the rMSLE and the prediction error ($1-$AUC for binomial, rMSPE for Gaussian and Poisson) for $n_{\text{test}} = 200$ over 100 repetitions and see that the proposed data-informed projection with ridge estimates in the diagonal generally increases the performance with respect to both metrics over the conventional RPs (Gaussian and SparseCW), reaching a lower link estimation error and a prediction error. The differences between the ridge estimators are less obvious, but generally, we see that the performance of the estimators in terms of screening also translates to the prediction power, namely that L2_dev0999 or L2_limit0 deliver the best results for the Gaussian family with both identity and log link, while L2_dev08 achieves the best prediction performance for the other families.

For this example, it can be seen that this adaption of the diagonal of the projection matrix suffices to reach a better performance than high-dimensional regression benchmarks adaptive LASSO and elastic net, but there is a noticeable gap to the oracle performance.

3.4 Benchmark simulations for SPAR

We consider $n = 200, p = 2000$ in sparse, medium and dense setting for the five family-link combinations and a block structure for Σ. We report here results for the canonical links. In the supplementary materials, we report additional results, which include different covariance structures for $p = 2000$ predictors for the medium setting, as well as different values of p for the medium setting with block-covariance.

We compare the following methods. Assuming the GLM, we use LASSO, elastic net ($\alpha = 3/4$), ridge, adaptive LASSO (using package **glmnet** Tay et al. 2023) and sure independence screening [SIS; Fan and Song, 2010]. Then, as general regression benchmarks, we use random forest (RF implemented R-package **randomForest** with `mtry` parameter tuned by CV; Liaw and Wiener 2002) and support vector machines (R package **e1071** with cost and kernel – linear or radial – tuned by CV; Meyer et al. 2023). For these two methods, we do not report results for link estimation, since they do not (necessarily) estimate a linear predictor. For variable ranking, we use the reported importance
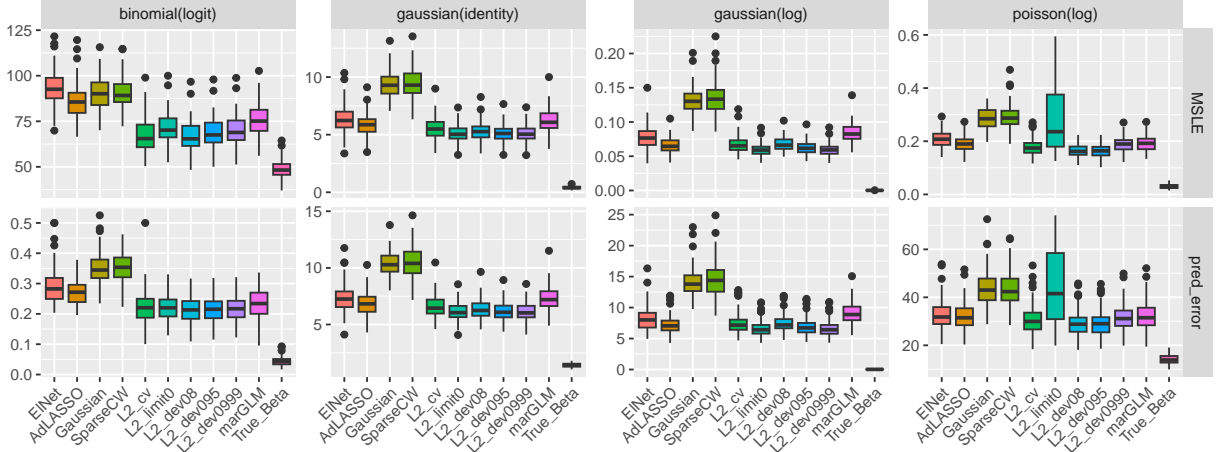
Figure 2: Link estimation error and prediction error (1−AUC for Binomial, MSPE otherwise) of Adaptive LASSO, Elastic Net, Gaussian and CW random projections, and our proposed projection with different ridge estimators, for 100 replications ($n = 200, p = 2000$, medium sparsity and block-diagonal $\boldsymbol{\Sigma}$).

measure (i.e., (weighted) mean of the individual trees' decrease in Gini-index or MSE produced by each variable) for RF and the inner product of the estimated coefficients and the support vectors for SVM.

Finally, as a set of methods using random projections, we use an ensemble of 50 models with the conventional RP from Definition 2.1 with random sign entries without any screening and random goal dimensions as in step 3.2 of the SPAR-algorithm in Section 2.4 (RP_CW_Ensemble), Targeted Random Projections (TARP), which is an adaptation of Mukhopadhyay and Dunson [2020] to GLMs where we perform screening based on marginal GLM coefficients and use the conventional RP of Achlioptas [2003] with $\psi = 1/6$ for an ensemble of 20 models, and our proposed SPAR algorithm from Section 2.4, once without cross-validation with fixed 20 models and λ chosen from a grid of values based on the best model deviance achieved on the training set, and once with cross-validation for a grid of λ values and up to 50 models.

Figure 3 shows prediction and link estimation performance for 100 replications of $n = 200, p = 2000$ in the block co-variance setting. The results for LASSO and SIS are omitted in this figure to provide a more comprehensive overview since they are always outperformed by AdLASSO and ElNet, respectively. Generally, SPAR and SPAR-CV are outperformed by AdLASSO and elastic net in terms of prediction and link estimation only in the sparse settings, while they are among the best performing methods in all other settings. Especially the performance in link estimation for the logistic regression is remarkable. Figure 9 in the Supplementary materials shows the prediction error for the five family-link combinations across the different covariance settings for $\boldsymbol{\Sigma}$ for fixed medium sparsity. The results for autocorrelated and block covariance are rather similar, while all methods perform better for the compound covariance and worse for the identity covariance, which could be due to the amount of information contained in the covariance structure. As final inspection of prediction performance, Figure 10 in the Appendix shows prediction errors for increasing p in the block covariance and medium sparsity. All methods lose some performance when the dimension increases, but adaptive LASSO and elastic net seem to suffer the most. Figure 4 shows that SPAR and SPAR-CV both perform well in terms of variable ranking as measured by pAUC, where only SVM yields a similar performance across all settings. For the ensemble methods, we here compute the pAUC of the final averaged $\hat{\boldsymbol{\beta}}$. As expected, AdLASSO and ElNet perform best in the sparse settings.

To assess the performance across all investigated settings (including those in the supplementary materials), we rank the methods from best (1) to worst (11) for each replication and setting. Figure 5 shows the average ranks (with 99% confidence intervals) across all investigated settings. Aside from SVM in variable ranking, we find that SPAR-CV and SPAR achieve the best average ranks for all of the three measures. Using the Friedman and the post-hoc Nemenyi tests for multiple comparisons [Hollander et al., 2013], we can also report that SPAR and SPAR-CV are significantly better than all other methods for prediction and link estimation, and, together with SVM, they are also significantly better for variable ranking. Even if SPAR and SPAR-CV are not best in every scenario, they perform well across-the-board, making them especially suitable when the degree of sparsity is unknown.

Finally, Figure 6 shows computing time for three increasing values of p for the binomial family for the medium sparsity setting with block covariance. Most methods, except RF, SVM, and SIS, inherit computational efficiency from **glmnet**, which uses a fast C++ implementation for canonical links. SPAR is the second fastest for larger p (after SIS) for the
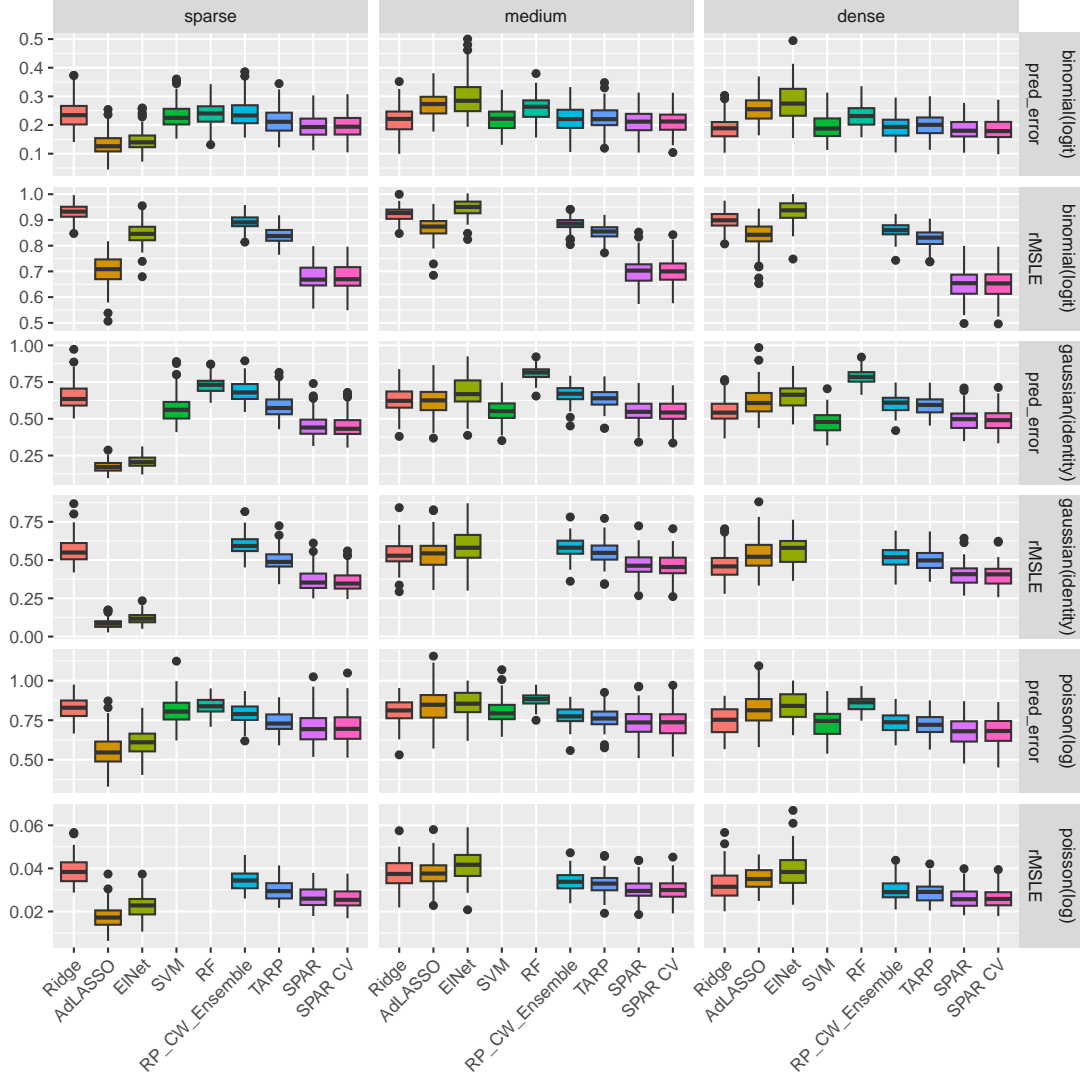
Figure 3: Comparison of prediction performance ($1-$AUC for binomial, rMSPE otherwise) and link estimation (rMSLE, not reported for SVM and RF) over 100 repetitions ($n = 200$, $p = 2000$, medium sparsity and block $\boldsymbol{\Sigma}$).

non-canonical family-link, with its time mostly spent fitting the $M = 20$ marginal models and thus hardly affected by p. We note that the ensemble of SparseCW random projections (RP_CW_Ensemble) is slower than SPAR due to projecting all p variables, resulting in larger projection matrices. TARP is slower than SPAR likely due to using less sparse projections. SPAR-CV, while slower than most methods for small p, scales efficiently with increasing p. The computing time for other family-link combinations follows similar patterns based on whether the link is canonical, so those plots are omitted for brevity.

4 Data applications

We illustrate the proposed method on one high-dimensional dataset with a count response and two datasets with binary response. Table 1 shows the average prediction performance metrics for all analyzed datasets over 100 random training/test splits of ratio three to one. For SPAR-CV, we use deviance as cross-validation measure for all datasets. The other methods have been tuned as described in Section 3.4.
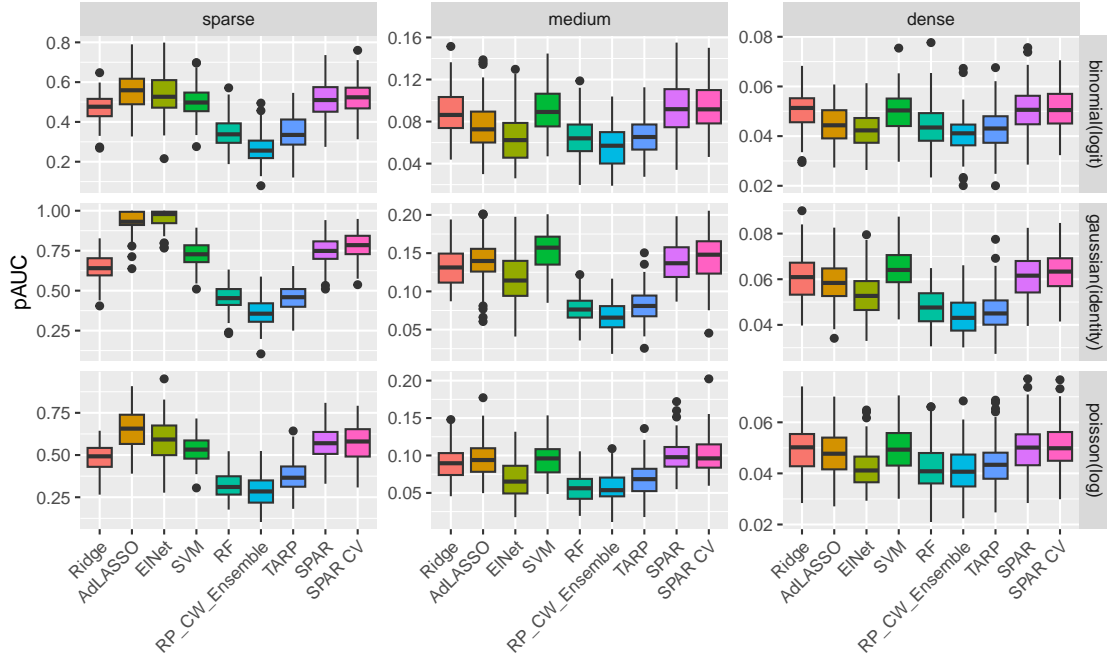
9

Figure 4: Comparison of variable ranking (pAUC rescaled to [0, 1] for better comparison) over 100 repetitions for $n = 200$, $p = 2000$, medium sparsity and block Σ.
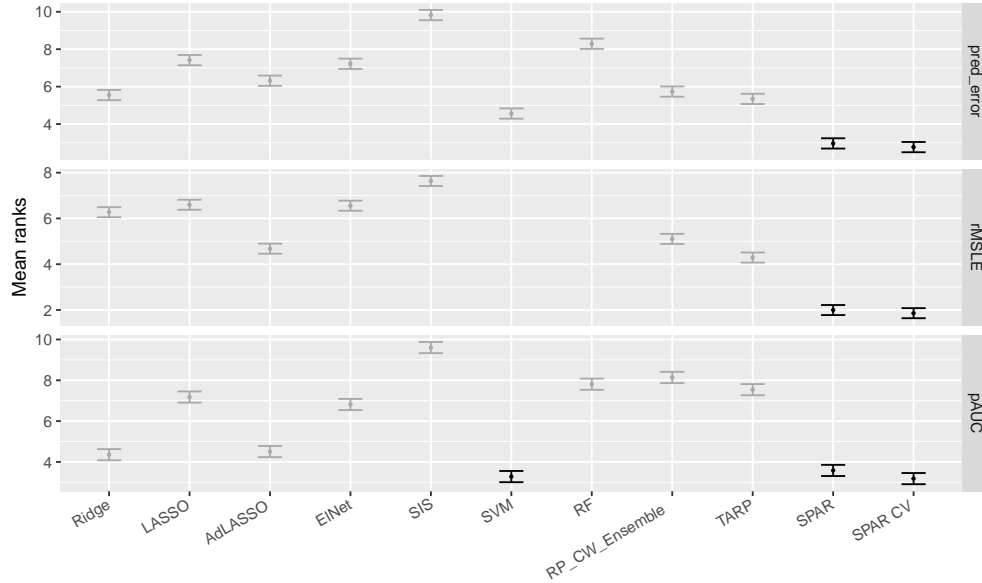

Figure 5: Mean ranks with 99% confidence intervals for prediction error, link estimation (rMSLE), and variable ranking (pAUC) across all investigated settings and $n_{\text{rep}} = 100$ replications. Methods not significantly worse than the best method are colored black.
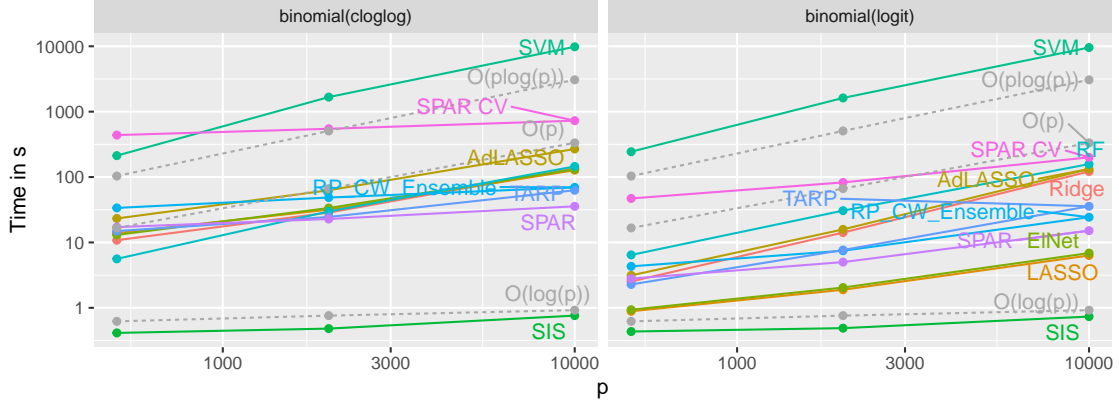
Figure 6: Comparison of average computing time for increasing p, $n = 200$, medium sparsity and block Σ, for the binomial family.

Table 1: Mean prediction metrics (with standard errors) on datasets over 100 three-to-one training/test splits, the best three methods for each dataset and metric are marked in bold.

Method	FTIR spectra		Darwin				DLBCL				DLBCL_extended			
	rMSPE		AUC		rMSPE		AUC		rMSPE		AUC		rMSPE	
Ridge	**0.059**	(0.004)	0.915	(0.004)	0.542	(0.007)	**0.995**	(0.001)	0.282	(0.01)	0.812	(0.009)	0.856	(0.008)
LASSO	0.079	(0.006)	0.914	(0.004)	0.509	(0.01)	0.963	(0.005)	0.456	(0.02)	**0.942**	(0.007)	**0.601**	(0.016)
AdLASSO	0.588	(0.058)	0.782	(0.006)	0.805	(0.007)	0.875	(0.009)	0.676	(0.023)	0.838	(0.011)	0.727	(0.02)
ElNet	0.078	(0.006)	0.921	(0.003)	0.491	(0.008)	0.979	(0.004)	0.346	(0.017)	**0.964**	(0.005)	**0.514**	(0.017)
SIS	0.424	(0.051)	0.898	(0.004)	0.600	(0.007)	0.902	(0.009)	0.678	(0.021)	0.902	(0.009)	0.678	(0.021)
SVM	0.855	(0.009)	**0.954**	(0.003)	**0.355**	(0.011)	0.992	(0.002)	**0.165**	(0.012)	0.885	(0.013)	0.648	(0.021)
RF	0.126	(0.009)	**0.957**	(0.003)	0.444	(0.006)	0.960	(0.004)	0.544	(0.01)	0.912	(0.007)	0.679	(0.011)
RP_CW_Ensemble	0.066	(0.005)	0.927	(0.003)	0.450	(0.009)	0.988	(0.003)	0.403	(0.013)	0.815	(0.011)	0.835	(0.016)
TARP	0.067	(0.007)	0.927	(0.003)	0.447	(0.01)	0.983	(0.003)	0.359	(0.012)	**0.939**	(0.005)	**0.515**	(0.017)
SPAR	**0.036**	(0.004)	0.935	(0.003)	**0.439**	(0.013)	**0.994**	(0.002)	**0.179**	(0.013)	0.921	(0.007)	0.627	(0.022)
SPAR CV	**0.033**	(0.003)	0.935	(0.003)	**0.437**	(0.013)	**0.994**	(0.001)	**0.183**	(0.012)	0.932	(0.006)	0.603	(0.023)

4.1 FTIR spectra

Fourier transform infrared (FTIR) spectroscopy is commonly used in tribology to analyze changes in oil samples during use. The dataset, introduced in Pfeiffer et al. [2022], consists of $n = 34$ artificially altered automotive engine oil samples. Two types of artificial alteration were done in a laboratory by heating the oil and exposing it to dried air, simulating real-life degradation. For each of the $n = 34$ samples, the difference FTIR spectra, i.e., the difference between the absorbances of the fresh oil and those of the degraded oils at $p = 1814$ wavenumbers, and alteration duration (in hours) were recorded along with the type of alteration.

Table 1 shows the relative MSPE for various methods. SPAR(-CV) was estimated using both a (quasi) Poisson family and a Gaussian family, both with log links. We only present results for the Gaussian model, as it gave the best predictions. Moreover, SPAR(-CV) performed best overall, followed by ridge regression.

Figure 7 (top panel) displays the difference spectrum for one sample, highlighting intervals with high or total absorption. These regions, typically non-informative due to hydrocarbon properties, are often pre-processed or discarded from analysis. The bottom panel shows the standardized coefficients estimated by SPAR-CV across $M = 50$ marginal models, where the coefficients for each variable are sorted by their absolute values and displayed vertically as a color gradient. Even without pre-processing, non-informative variables rarely appear in the models and have low coefficients when they do, demonstrating the reliability of the SPAR method. The distribution of coefficients indicates which wavenumbers correlate positively or negatively with longer alteration durations.

4.2 Darwin Alzheimer dataset

The dataset, introduced in Cilia et al. [2022], contains a binary response for Alzheimer's disease (AD) together with $p = 450$ extracted variables from 25 handwriting tests (18 features per task) for 89 AD patients and 85 healthy people ($n = 174$) and can be downloaded from the UC Irvine Machine Learning Repository. The dataset has been reported to contain outliers, so before proceeding with the analysis, we screened for multivariate outliers and imputed them
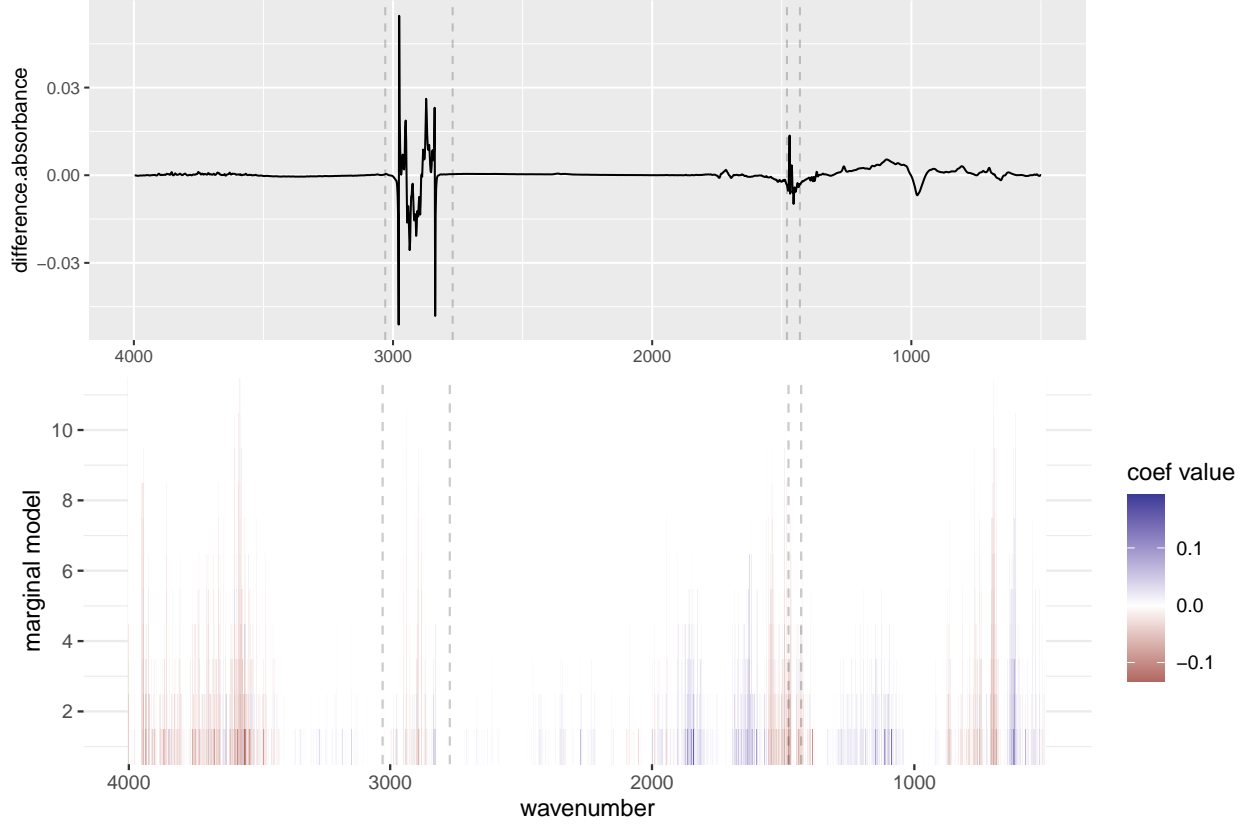
Figure 7: Difference FTIR spectrum for one oil sample in the tribology dataset (top) and coefficients of $p = 1814$ wavenumbers estimated by SPAR-CV in each of $M = 50$ marginal models (bottom). Marked intervals represent non-informative variables.

using the *detect deviating cells* algorithm of Rousseeuw and Bossche [2018] implemented in R package **cellWise** [Raymaekers and Rousseeuw, 2023].

Table 1 presents the area under the ROC curve and the relative MSPE (Brier score) as prediction metrics for the binary classification task. For all methods based on GLMs (i.e., except SVMs and RFs), results are shown for the binomial family with logit link, as it provided the best performance among all investigated link functions. SPAR and SPAR-CV performed similarly and were outperformed in AUC by both SVM and RF, and in rMSPE by SVM alone. This suggests SPAR is a viable option for modeling this dataset, while offering lower computational cost.

Figure 8 displays the estimated standardized coefficients for the $p = 450$ variables grouped by feature, across all marginal models. Feature blocks generally show either a positive or negative impact on the probability of AD across all 25 tasks (indicated by blocks of blue or red lines). For example, the probability of AD increases with the total time spent on a task (total_time). The number of pendowns (num_of_pendown, the number of times the pen hits the paper) is positively associated with the AD likelihood for the first few tasks, but negatively correlated for the remaining tasks, with a strong negative association observed for the task "Copy the fields of a postal order", which requires many pendowns when copying the individual fields. This may suggest that a lower number of pendowns in this task indicates failure to complete the task and thus a higher likelihood of AD.

4.3 Diffuse large B-cell lymphoma (DLBCL)

The *Diffuse large B-cell lymphoma (DLBCL)* microarray dataset has been introduced in Shipp et al. [2002] and is available at the OpenML platform. It contains expression data for $p = 5469$ genes of $n = 77$ patients diagnosed with two different types of lymphomas: DLBCL (58 cases) and follicular lymphoma (FL, 19 cases), making this a rather *imbalanced* dataset. Given that we do not know the true degree of sparsity in any of the employed datasets, we artificially extend the DLBCL dataset by randomly permuting the rows of the covariate matrix and appending
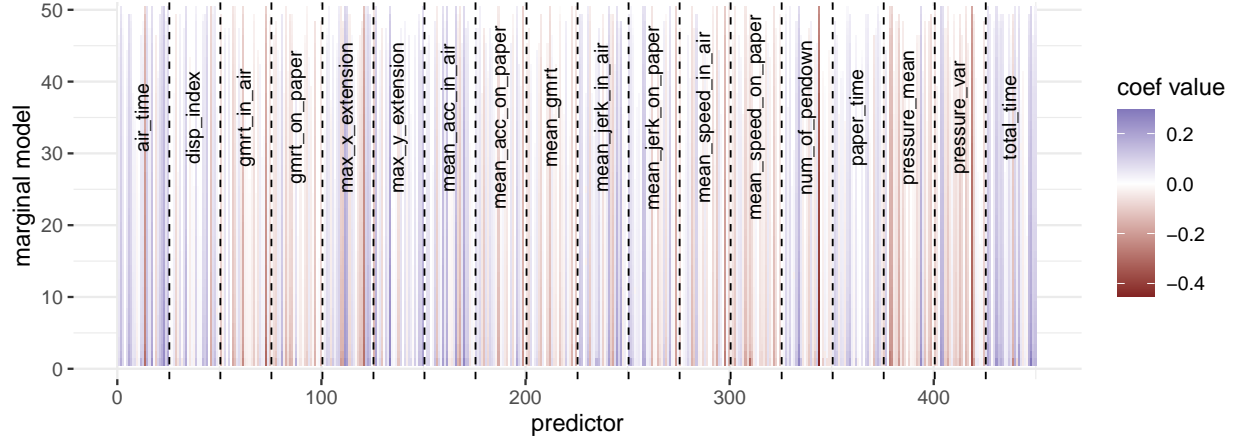
Figure 8: Estimated coefficients for the $p = 450$ variables in the Darwin dataset, for each of the $M = 50$ marginal models in the SPAR-CV algorithm.

the permuted covariates to the existing dataset twice. This results in an *extended DLBCL dataset* with $n = 77$ and $p = 5469 \times 3 = 16407$.

Again, Table 1 presents the area under the ROC curve and the relative MSPE (Brier score) as prediction metrics and for all methods based on GLMs results are shown for the binomial family with logit. SPAR and SPAR-CV perform very well on the original dataset, being outperformed only by SVM in terms of rMSPE and by ridge in terms of AUC. For the extended dataset, the best-performing methods in terms of AUC are elastic net followed by LASSO and TARP. SPAR(-CV) performs satisfactorily, being outperformed only by elastic net and TARP in terms of rMSPE. The superior performance of the sparse methods is not unexpected and is in line with the simulation results, as the degree of sparsity in this dataset has been artificially increased through the permutation of the variables.

5   Conclusion

In this paper, we propose a novel data-driven random projection method to be employed in high-dimensional GLMs, which efficiently reduces the dimensionality of the problem while preserving essential information between the response and the (possibly correlated) predictors. We achieve this by using ridge-type estimates of the regression coefficients to construct the random projection matrix which should accurately recover the true regression coefficients. These coefficients can also be employed for variable screening, which can be used prior to random projection to further reduce the dimensionality of the problem.

A critical aspect of the proposed method is the selection of the penalty term for the ridge-type estimator. The penalty should generally be small to avoid over-regularization. However, determining the optimal size of the penalty has proven to be a non-trivial task. For linear regression, a ridge estimator with penalty converging to zero has shown good properties [Wang and Leng, 2016, Parzer et al., 2024]. In this paper, we derive the analytical formula for such an estimator in GLMs with canonical links and find that this estimator leads to lower predictive performance for non-Gaussian families, likely due to overfitting. More generally, there is no *one-size-fits-all* penalty value for all families. Instead, we advocate for a data-driven approach by decreasing the penalty value as long as the improvement in a goodness-of-fit criterion (e.g., deviance) exceeds a certain threshold (e.g., 0.8 for non-Gaussian families and 0.999 for Gaussian) seems to be the best strategy.

Through extensive simulations, we show that integrating multiple probabilistic variable screening and projection steps into an ensemble of medium-sized GLMs improves prediction accuracy and variable ranking, without too much computational cost. To implement this method, we adapt the SPAR algorithm from Parzer et al. [2024], ensuring that it is tailored to the specific requirements of high-dimensional GLMs and show that the method achieves top overall performance aggregated over all investigated scenarios, making it a valid choice when the true degree of sparsity is not known in practice. At the cost of higher computation time, which still scales well with p, the method can, to some degree, benefit from cross-validation, most notably in terms of ranking the variables based on their relevance.

Finally, three real data sets illustrate the interpretability and superior prediction performance of the proposed approach. A potential extension includes adapting the method to multivariate GLMs (e.g., multinomial) and multivariate responses (e.g., multivariate linear regression). A key extension in this direction would be designing a data-driven RP

that can preserve the multivariate structure in the data while also being straightforward and fast to compute. Additionally, ways of incorporating non-linearities in the RP could be explored.

Supplementary materials

The code to produce all simulations and data applications can be found in `https://github.com/RomanParzer/SPAR_GLM_Paper_Code`, and the R-package performing the SPAR-algorithm can be found on `https://github.com/RomanParzer/SPAR`.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

Appendix A   Proof of Theorem 2.2

To prove Theorem 2.2, we will need the following Lemma.

Lemma A.1. *Let* $f_\lambda(\boldsymbol{\beta}) = -\tilde{\ell}(\boldsymbol{\beta}) + \frac{\lambda}{2}\sum_{j=1}^p \beta_j^2$ *and* $\boldsymbol{\beta}_\lambda := argmin_{\boldsymbol{\beta}\in\mathbb{R}^p} f_\lambda(\boldsymbol{\beta})$. *Then,* $\lambda\boldsymbol{\beta}_\lambda \to \mathbf{0}$ *for* $\lambda \to 0$.

Proof of Lemma A.1. Each summand $f(\theta_i) = y_i\theta_i - b(\theta_i)$ of $\tilde{\ell}(.)$ is a concave function of θ_i and has a unique maximum at $\hat{\theta}_i = (b')^{-1}(y_i)$ if $(b')^{-1}(y_i) \in \mathbb{R}$, or is unbounded otherwise (as is the case for a Bernoulli variable with logit-link, where $(b')^{-1}(y_i) = \log(y_i/(1-y_i))$ and $y_i \in \{0, 1\}$). Note that for the family with the canonical link and defined values $g(y_i)$, we have $\theta(\boldsymbol{\beta}, \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and, therefore, f_λ is strictly convex and has a unique minimum. We will even show the stronger result, that $\lambda\|\boldsymbol{\beta}_\lambda\|^2 \to 0$. If that were not the case, there exists some $\varepsilon > 0$ such that $\forall\Lambda \exists\lambda < \Lambda$ such that $\lambda\|\boldsymbol{\beta}_\lambda\|^2 > \varepsilon$. Let $\Lambda := \varepsilon/\|\hat{\boldsymbol{\beta}}\|^2$, where $\hat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}g(\boldsymbol{y})$. According to Section 2.1, any $\boldsymbol{\beta}$ maximizes $\tilde{\ell}$, if and only if $\boldsymbol{x}_i'\boldsymbol{\beta} = g(y_i)$ for all $i \in [n]$ or $\boldsymbol{X}\boldsymbol{\beta} = g(\boldsymbol{y})$ in matrix notation, since we use the canonical link $g = (b')^{-1}$. It is easy to see that $\hat{\boldsymbol{\beta}}$ satisfies this condition, and has the smallest L_2-norm among all such $\boldsymbol{\beta}$ (see e.g. proof of Lemma 3 in Parzer et al. [2024]). For the λ corresponding to Λ, we then get

$$f_\lambda(\hat{\boldsymbol{\beta}}) - f_\lambda(\boldsymbol{\beta}_\lambda) = \underbrace{\tilde{\ell}(\boldsymbol{\beta}_\lambda) - \tilde{\ell}(\hat{\boldsymbol{\beta}})}_{\leq 0} + \underbrace{\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}\|^2}_{<\Lambda/2} - \underbrace{\frac{\lambda}{2}\|\boldsymbol{\beta}_\lambda\|^2}_{>\varepsilon/2}$$

$$< \frac{\Lambda}{2}\|\hat{\boldsymbol{\beta}}\|^2 - \frac{\varepsilon}{2} = 0,$$

which contradicts the definition of $\boldsymbol{\beta}_\lambda$. $\qquad\qquad\square$

Proof of Theorem 2.2. We need to show that $\boldsymbol{\beta}_\lambda \to \hat{\boldsymbol{\beta}}$ for $\lambda \to 0$, where $\hat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}g(\boldsymbol{y})$. Let $\boldsymbol{P}\boldsymbol{\beta}_\lambda = \hat{\boldsymbol{\beta}} + (\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X})\boldsymbol{\beta}_\lambda$ be the orthogonal projection of $\boldsymbol{\beta}_\lambda$ to the (affine) subspace $\{\boldsymbol{X}\boldsymbol{\beta} = g(\boldsymbol{y})\} = \hat{\boldsymbol{\beta}} + \ker(\boldsymbol{X})$. Then

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|^2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{P}\boldsymbol{\beta}_\lambda\|^2 + \|\boldsymbol{P}\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_\lambda\|^2 =$$
$$= \|(\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X})\boldsymbol{\beta}_\lambda\|^2 + \|\underbrace{\hat{\boldsymbol{\beta}} - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{\beta}_\lambda}_{=\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}(g(\boldsymbol{y}) - \boldsymbol{X}\boldsymbol{\beta}_\lambda)}\|^2.$$

For the canonical link, f_λ is explicitly given by $f_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^n b(\boldsymbol{x}_i'\boldsymbol{\beta}) - y_i\boldsymbol{x}_i'\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|^2/2$ with the following first order optimality conditions at $\boldsymbol{\beta}_\lambda$.

$$0 = \frac{\partial f_\lambda}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_\lambda) = \boldsymbol{X}'(b'(\boldsymbol{X}\boldsymbol{\beta}_\lambda) - \boldsymbol{y}) + \lambda\boldsymbol{\beta}_\lambda.$$

Figure 9: Prediction error for different covariance structures with medium setting for $p = 2000$

This implies, that $\boldsymbol{\beta}_\lambda = -\frac{1}{\lambda}\boldsymbol{X}'(b'(\boldsymbol{X}\boldsymbol{\beta}_\lambda) - \boldsymbol{y}) \in \text{span}(\boldsymbol{X}')$ and, therefore, $(\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X})\boldsymbol{\beta}_\lambda = 0$. Rewriting the optimality conditions and using Lemma A.1 also show that

$$\boldsymbol{X}'(b'(\boldsymbol{X}\boldsymbol{\beta}_\lambda) - \boldsymbol{y}) = -\lambda\boldsymbol{\beta}_\lambda \to \boldsymbol{0} \text{ for } \lambda \to 0.$$

Since $\text{rank}(\boldsymbol{X}) = n$, this is only possible if $b'(\boldsymbol{X}\boldsymbol{\beta}_\lambda) \to \boldsymbol{y}$ for $\lambda \to 0$, and, since the canonical link g is continuous, this implies $g(b'(\boldsymbol{X}\boldsymbol{\beta}_\lambda)) = \boldsymbol{X}\boldsymbol{\beta}_\lambda \to g(\boldsymbol{y})$. Therefore, the second term also vanishes for $\lambda \to 0$ and $\boldsymbol{\beta}_\lambda \to \hat{\boldsymbol{\beta}}$. $\qquad\square$

Appendix B Additional tables and figures for simulations

Method	binomial(logit)		gaussian(identity)		gaussian(log)		poisson(log)	
L2_cv	9.993	(1.399)	28.435	(0.839)	1458.825	(140.805)	580.668	(45.346)
L2_dev08	0.677	(0.008)	10.065	(0.182)	1448.049	(32.237)	77.657	(1.487)
L2_dev095	0.094	(0.001)	2.695	(0.038)	345.963	(6.554)	16.759	(0.305)
L2_dev0999	0.001	(0)	0.246	(0.002)	23.346	(0.316)	0.791	(0.042)

Table 2: Average chosen λ for different estimators over 100 replications ($n = 200, p = 2000$, block-diagonal $\boldsymbol{\Sigma}$ and medium sparsity).

References

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. ISSN 01621459. URL http://www.jstor.org/stable/3085904.
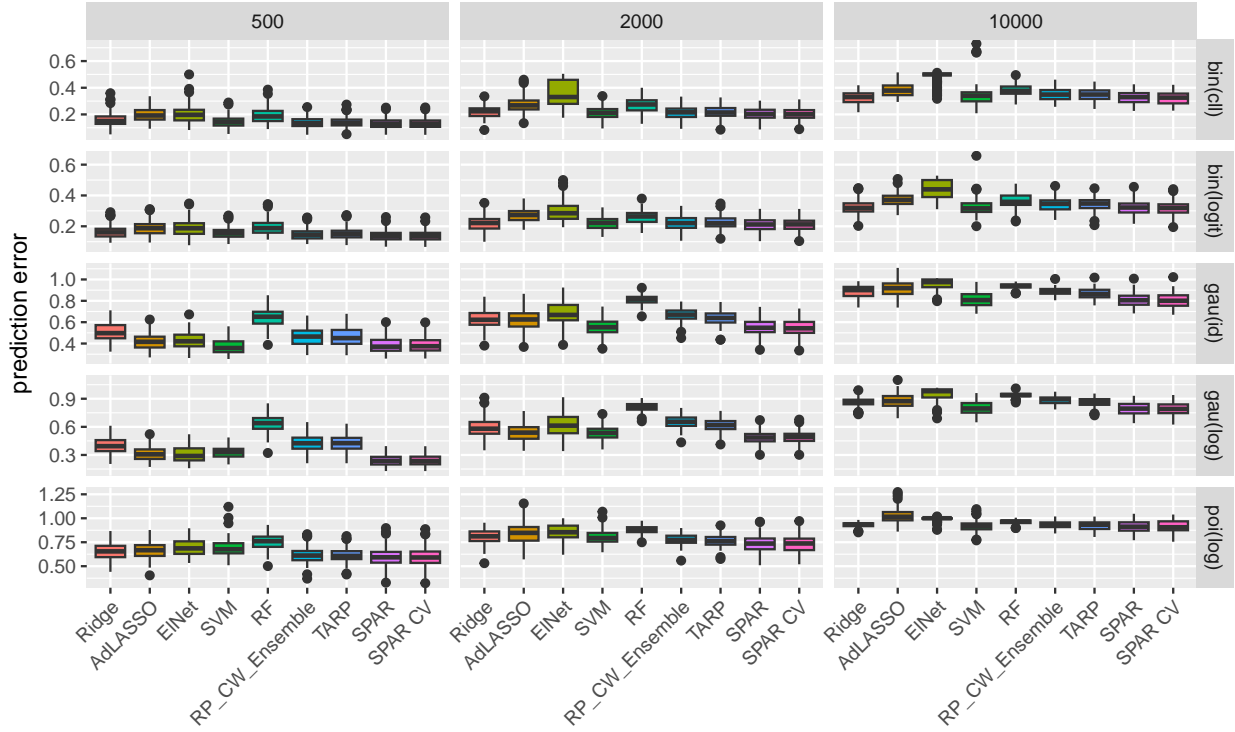
Figure 10: Prediction error for increasing p with medium setting and block-diagonal covariance

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi:0.1111/j.1467-9868.2005.00503.x.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038, 2009. URL https://jmlr.csail.mit.edu/papers/v10/fan09a.html.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567 – 3604, 2010. doi:10.1214/10-AOS798.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. doi:10.1198/jasa.2011.tm09779.

Qing Mai and Hui Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234, 2013. doi:10.1093/biomet/ass062.

Qing Mai and Hui Zou. The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471 – 1497, 2015. doi:10.1214/14-AOS1303.

Chenlu Ke. Sufficient variable screening with high-dimensional controls. *Electronic Journal of Statistics*, 17(2):2139 – 2179, 2023. doi:10.1214/23-EJS2150.

Xiangyu Wang and Chenlei Leng. High-dimensional ordinary least-squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:589–611, 06 2016. doi:10.1111/rssb.12127.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):959–1035, 2017. doi:10.1111/rssb.12228.

Rajarshi Guhaniyogi and David B. Dunson. Bayesian Compressed Regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015. doi:10.1080/01621459.2014.969425.

Minerva Mukhopadhyay and David B. Dunson. Targeted Random Projection for Prediction From High-Dimensional Features. *Journal of the American Statistical Association*, 115(532):1998–2010, 2020. doi:10.1080/01621459.2019.1677240.

Nick Ryder, Zohar S. Karnin, and Edo Liberty. Asymmetric random projections. *arXiv*, abs/1906.09489, 2019. doi:10.48550/arXiv.1906.09489.

Roman Parzer, Peter Filzmoser, and Laura Vana-Gür. Sparse data-driven random projection in regression for high-dimensional data. *arXiv*, 2024. doi:10.48550/arXiv.2312.00130.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606. URL `https://books.google.at/books?id=h9kFH2_FfBkC`.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 81–90, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi:10.1145/2488608.2488620.

Jianqing Fan and Jinchi Lv. Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 10 2008. ISSN 1369-7412. doi:10.1111/j.1467-9868.2008.00674.x. URL `https://doi.org/10.1111/j.1467-9868.2008.00674.x`.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.*, 21(1), 1 2020. ISSN 1532-4435.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 08 2004. URL `https://www.jmlr.org/papers/v5/rosset04a.html`.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 1498712169. doi:10.1201/b18401.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2023. R package version 1.7-13.

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

J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31, 2023. doi:10.18637/jss.v106.i01.

Andy Liaw and Matthew Wiener. Classification and regression by **randomForest**. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. ISSN 0022-0000. doi:10.1016/S0022-0000(03)00025-4. Special Issue on PODS 2001.

M. Hollander, D.A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118553299. URL `https://books.google.at/books?id=Y5s3AgAAQBAJ`.

Pia Pfeiffer, Bettina Ronai, Georg Vorlaufer, Nicole Dörr, and Peter Filzmoser. Weighted lasso variable selection for the analysis of ftir spectra applied to the prediction of engine oil degradation. *Chemometrics and Intelligent Laboratory Systems*, 228:104617, 2022. ISSN 0169-7439. doi:10.1016/j.chemolab.2022.104617.

Nicole D. Cilia, Giuseppe De Gregorio, Claudio De Stefano, Francesco Fontanella, Angelo Marcelli, and Antonio Parziale. Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111:104822, 2022. ISSN 0952-1976. doi:10.1016/j.engappai.2022.104822.

Peter J Rousseeuw and Wannes Van Den Bossche. Detecting deviating data cells. *Technometrics*, 60(2):135–145, 2018. doi:10.1080/00401706.2017.1340909.

Jakob Raymaekers and Peter Rousseeuw. *cellWise: Analyzing Data with Cellwise Outliers*, 2023. R package version 2.5.3.

Margaret A. Shipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Ricardo C.T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002. doi:10.1038/nm0102-68.