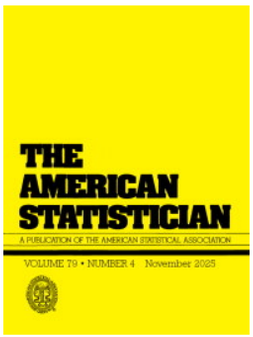
Taylor & Francis
Taylor & Francis Group



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

Bad Estimation, Good Prediction: The Lasso in Dense Regimes

Andrea Bratsberg, Magne Thoresen & Jelle J. Goeman

To cite this article: Andrea Bratsberg, Magne Thoresen & Jelle J. Goeman (21 Nov 2025): Bad Estimation, Good Prediction: The Lasso in Dense Regimes, The American Statistician, DOI: 10.1080/00031305.2025.2569464

To link to this article: https://doi.org/10.1080/00031305.2025.2569464

THE AMERICAN STATISTICIAN
2025, VOL. 00, NO. 0, 1–9: Statistical Practice
https://doi.org/10.1080/00031305.2025.2569464

Bad Estimation, Good Prediction: The Lasso in Dense Regimes

Andrea Bratsberg[a], Magne Thoresen[a], and Jelle J. Goeman[b]

[a]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway; [b]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

ABSTRACT

For high-dimensional omics data, sparsity-inducing regularization methods such as the Lasso are widely used and often yield strong predictive performance, even in settings when the assumption of sparsity is likely violated. We demonstrate that under a specific dense model, namely the high-dimensional joint latent variable model, the Lasso produces sparse prediction rules with favorable prediction error bounds, even when the underlying regression coefficient vector is not sparse at all. We further argue that this model better represents many types of omics data than sparse linear regression models. We prove that the prediction bound under this model in fact decreases with increasing number of predictors, and confirm this through simulation examples. These results highlight the need for caution when interpreting sparse prediction rules, as strong prediction accuracy of a sparse prediction rule may not imply underlying biological significance of the individual predictors.

ARTICLE HISTORY
Received January 2025
Accepted September 2025

KEYWORDS
High-dimensional omics; Latent variable models; Multicollinearity; Sparsity; Eigenvalue decay

1. Introduction

Technological developments in the last few decades have improved our ability to measure vast amounts of data across various research fields, including omics, econometrics, and finance. In omics research (transcriptomics, in particular), microarrays and next-generation sequencing has made it possible to quickly and cheaply measure gene expression from thousands of genes for a number of subjects, simultaneously. If properly analyzed, these data could improve our understanding of the human genome, which in turn can potentially enhance current research on personal treatment or precision medicine. A common objective in the development of new biomarkers, for example, is to identify subsets of genes or molecular networks that are predictive of a phenotype, such as disease status or prognosis (Segal, Dahlquist, and Conklin 2003).

To this end, sparse methods for high-dimensional regression models have become immensely popular due to their ability to produce parsimonious and interpretable solutions, even when the number of predictors far exceeds the number of independent samples. These methods rely on the assumption that only a small subset of predictors are truly relevant for the outcome of interest, while the rest are noise, and aim at selecting these nonzero effects (i.e., variable selection). One of the most widely used variable selection methods is the Lasso (Tibshirani 1996), which favors solutions with small ℓ_1-norm, resulting in a sparse estimated regression coefficient vector. Lasso and other sparsity-inducing methods often show good prediction performance across a wide range of settings, even when their underlying assumptions are likely violated (Greenshtein and Ritov 2004; Zhao and Simon 2010; Waldron et al. 2011; Ogutu, Schulz-Streeck, and Piepho 2012; Abraham et al. 2013; Wang et al. 2019; Ajana et al. 2019). Prediction is fundamentally an easier task than estimation, and it is well known that a good predictor does not necessarily provide an accurate estimate of the underlying model. Still, we believe there is a need to clarify this more explicitly for high-dimensional omics data, as sparse solutions are in practice often interpreted, either explicitly or implicitly, as being biologically meaningful. We argue in the following that the assumption of sparsity is often unrealistic for omics data and that researchers should be cautious when interpreting a good sparse prediction rule. Similar questions have been raised in other fields, such as econometrics (Giannone, Lenza, and Primiceri 2021). The aim of this article is to first propose the joint latent variable model as a plausible model for the data-generating mechanism for omics data, and then demonstrate that sparse methods, in particular the Lasso, can still achieve strong prediction accuracy under this non-sparse model.

1.1. The Sparsity Assumption for Omics Data

First, omics data, such as gene expression measurements, are typically highly correlated. Gene expression data are the results of the combinatorial effects from many highly connected biological pathways (Sun and Zhao 2004), resulting in potentially long-range correlations and strong local correlations. Additionally, the measured variables are inherently noisy, due to for example

CONTACT Andrea Bratsberg a.m.bratsberg@medisin.uio.no Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, P.O.Box 1122 Blindern, 0317 Oslo, Norway.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

technical noise or biological variation, and even the noise may be correlated (Segura-Lepe, Keun, and Ebbels 2019). Such highly correlated designs may contradict the assumption of a sparse underlying multiple regression model. If, for example, there is only one variable in a group of correlated variables that is directly related to the outcome, the remaining variables may partially explain its residual variation, leading to a dense representation in the model. Thus, if many or all variables are highly correlated, then the resulting (multiple) regression coefficient will not be sparse (Klein et al. 2014, chap. 15). Empirical evidence also appears to support this issue; Ein-Dor et al. (2005) showed that for a single breast cancer survival dataset, there are several different sets of 70 genes that are equally predictive of survival, undermining the notion of an underlying unique smaller subset of "driving genes". Some genomic data may coincide with the assumption of sparsity, such as single nucleotide polymorphisms (SNPs) (Goeman and Böhringer 2020). However, this is rather the exception than the rule.

Second, sparse methods such as the Lasso rely on fairly strong assumptions to achieve model selection consistency, that is, the ability to recover the true sparse model when one exists. For example, the design matrix needs to fulfill the so-called irrepresentability condition (Zhao and Yu 2006), which essentially says that the non-influential (or "non-active") variables cannot depend too strongly on the influential ("active") variables. When the number of variables exceeds the number of independent samples, the columns of the design matrix are necessarily linear combinations of each other. Thus, when the number of variables is extremely large compared to the number of independent samples, this condition is generally difficult to satisfy. In particular, Wang et al. (2019) demonstrated that for three types of real omics data (gene expression, methylation and miRNA) and three cancer types, the irrepresentability condition is often violated even when the true model is sparse. Another key assumption for most sparse methods is that the truly influential variables must have large effects on the outcome of interest, while the remaining variables have negligible effects (Van De Geer 2018). This assumption is also often violated for genomic data, particularly for complex diseases, where correlations between each variable and the outcome seem to be much more distributed across the whole genome (Ein-Dor et al. 2005; Boyle, Li, and Pritchard 2017; Goeman and Böhringer 2020). Even when there is strong evidence for the existence of "core genes" that directly affect a disease, and that can in principle be interpreted, these are affected by an abundance of tiny influences from other genes (Boyle, Li, and Pritchard 2017). Consequently, such "core genes" may explain only a small part of the risk of a disease, even though they have a significant role in its development (Ball 2023; Boyle, Li, and Pritchard 2017).

Given the noisy and highly correlated nature of omics data, the assumption that a unique subset of variables drives the outcome is often unrealistic. Even in settings where this assumption is plausible, the stringent conditions for sparse methods to select this set are rarely met in real omics data. While there exist many extensions to the Lasso, such as Elastic net (Zou and Hastie 2005) that deal with the issue of collinearity by forcing a grouping structure on the coefficient estimates, the fundamental issue of sparsity still remains. Sparse methods are not stable, whereas stable solutions are not sparse (Xu, Caramanis, and Mannor 2011). Efforts to stabilize the solution by aggregating the results from multiple models (ensemble methods) still lead to similar conclusions in high-dimensional omics settings; different subsets of variables with almost no overlap give comparable prediction performance (Pes, Dessì, and Angioni 2017).

1.2. A Dense Model

To better capture the complex interactions between genes and their effects phenotypes (e.g., disease risk), Boyle, Li, and Pritchard (2017, p. 1184) proposed that a more realistic model is "that disease risk is largely driven by genes with no direct relevance to disease and is propagated through regulatory networks to a much smaller number of core genes with direct effects". Motivated by this perspective, we suggest the investigation of a joint latent variable model to represent the dependencies among observed variables and their relationship with the outcome (Goeman 2006, chap. 6). The latent variable model is a common approach to characterize dependency structures among features (Fan, Lou, and Yu 2024), and it allows for the inclusion of random errors in the observed features such as measurement error, which is very common in omics data. These models assume that both the observed predictors (e.g., gene expression) and the outcome are driven by a common set of unobserved latent variables. These may be the shared metabolic pathways or regulatory mechanisms (Carvalho et al. 2008; Leek 2011). In such models, any linear predictor based on the high-dimensional set of observed predictors is far from sparse. In this article, we illustrate, both through simulations and theoretical results, that Lasso can still yield accurate predictions even when the underlying model is not sparse, but rather, a (dense) latent variable model. In particular, we show that the prediction error decreases as the number of predictors diverges. Sparse solutions, such as those produced by Lasso, should thus be viewed as practical prediction tools rather than representations of the underlying biological reality.

1.3. Illustrative Example

To illustrate our point, we revisit a simple (low-dimensional) example from the original Elastic net article (Zou and Hastie 2005). The Elastic net penalizes both large ℓ_1-norm and ℓ_2-norm solutions. This way, it decorrelates the variables (due to the ℓ_2 penalty) and can be seen as a stabilizing version of the Lasso. In the example, they consider two latent variables Z_1 and Z_2 which are independent $\mathcal{U}(0, 20)$. The response y is generated as $\mathcal{N}(\boldsymbol{\beta}^T\mathbf{z}, 1)$, where $\mathbf{z} = (Z_1, Z_2)^T$ and $\boldsymbol{\beta} = (1, 0.1)^T$. The observed variables are generated as $\mathbf{x} = A^T\mathbf{z} + \mathbf{e}$, where $\mathbf{e} \in \mathcal{N}(0, \frac{1}{16}I)$ and

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 \end{pmatrix}.$$

Thus, the first three variables form a group whose underlying factor is Z_1, and the remaining variables form a second group whose underlying factor is Z_2. The within-group correlations are almost 1 and the between-group correlations are almost 0. From Figure 5 in Zou and Hastie (2005), it is clear that the Elastic net can be thought of as a stabilized version of the Lasso, with the latter showing a highly unstable variable selection.

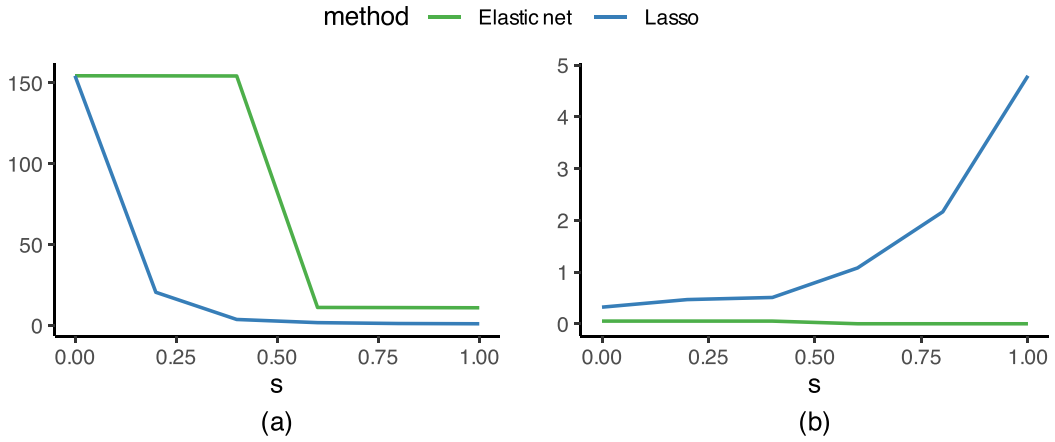


Figure 1. (a) Out-of-sample prediction error $\|y_* - \mathbf{x}_*^T\hat{\boldsymbol{\gamma}}\|_2^2$, and (b) estimation error $\|\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}\|_2^2$ for Lasso versus Elastic net.

But what about the prediction error? The best linear unbiased predictor of y using only $\mathbf{x}$ is $\mathbf{x}^T\boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}_0 = (A^TA + \frac{1}{16}I)^{-1}A^T\boldsymbol{\beta}$. We may think of this as the "true" underlying regression coefficient. We now rerun the exact same example and instead record the out-of-sample prediction error $\|y_* - \mathbf{x}_*^T\hat{\boldsymbol{\gamma}}\|_2^2$ for a new data point $(\mathbf{x}_*^T, y_*)$, together with the ℓ_2-estimation error $\|\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}\|_2^2$, where $\hat{\boldsymbol{\gamma}}$ is either the Lasso or Elastic net estimate. We repeat the simulation 100 times and report the average, as a function of $s = \|\hat{\boldsymbol{\gamma}}\|_1/\|\hat{\boldsymbol{\gamma}}_n\|_1$, where $\hat{\boldsymbol{\gamma}}_n$ is the full ordinary least squares estimate.

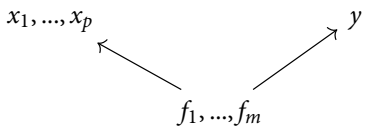
The display to the right in Figure 1 shows that the Elastic net gives a much more correct estimate of the underlying regression coefficient in terms of the ℓ_2-estimation error. However, when we consider the prediction error (left display), Lasso clearly outperforms the Elastic net. By focusing on the results at, for example, $s = 0.75$, we can draw two key conclusions: (a) a good sparse estimator (the Elastic net) is not necessarily a good sparse predictor, and (b) a good sparse predictor (the Lasso) is not necessarily a good sparse estimator.

1.4. Outline of the Article

The rest of this work is organized as follows. In Section 2 we formally define the high-dimensional joint latent variable model, allowing the number of observed variables p to approach infinity. In Section 3, we build on the theory for Lasso prediction for correlated designs and prove that, under the latent variable model, the in-sample prediction error exhibits favorable bounds for increasing p. Section 4 presents simulation examples that explore how the prediction (and estimation) errors are influenced by p and the correlation structure of the random error in the observed variables.

2. Joint Latent Variable Model

Assume we observe p predictors $\mathbf{x} = (x_1, \ldots, x_p)^T$ and response y, and that both are driven by a set of latent variables $\mathbf{f} = (f_1, \ldots, f_m)^T$:



and that $m \ll p$. The variables $f_1, \ldots, f_m$ can for example be shared biological pathways, and we assume that $\mathbf{f}$ has finite mean and covariance matrix. In addition, the predictor variables are subject to (possibly correlated) random error, which consists of uncontrollable sources of variability, including sampling and technical measurement error. Specifically, we assume the following model:

$$\begin{aligned} y &= \boldsymbol{\beta}^T\mathbf{f} + \epsilon \\ \mathbf{x} &= A^T\mathbf{f} + \mathbf{e}, \end{aligned} \tag{1}$$

where we have implicitly assumed without loss of generality that the marginal means of y and $\mathbf{x}$ are zero. The parameters $\boldsymbol{\beta}$ and A are an m-vector and an $m \times p$ matrix of loadings. We assume that ϵ and $\mathbf{e}$ are uncorrelated and that they have mean zero and variance-covariance σ^2 and Ψ, respectively. The joint vector $\mathbf{z} = (y, \mathbf{x}^T)^T$ then has mean zero and covariance matrix

$$\Sigma_z = \begin{pmatrix} \boldsymbol{\beta}^T\boldsymbol{\beta} + \sigma^2 & \boldsymbol{\beta}^TA \\ A^T\boldsymbol{\beta} & A^TA + \Psi \end{pmatrix}.$$

If we assume normality of $\mathbf{z}$, we have that

$$\mathrm{E}[y|\mathbf{x}] = \boldsymbol{\gamma}_0^T\mathbf{x}, \quad \text{where } \boldsymbol{\gamma}_0 = (A^TA + \Psi)^{-1}A^T\boldsymbol{\beta}.$$

However, if normality is not assumed, $\boldsymbol{\gamma}_0^T\mathbf{x}$ is still the best linear unbiased prediction of y given $\mathbf{x}$. We may therefore predict the mean of y from $\mathbf{x}$ by

$$\mathrm{E}[y|\mathbf{x}] = \boldsymbol{\gamma}_0^T\mathbf{x}.$$

Under this high-dimensional latent variable model, $\boldsymbol{\gamma}_0$ is potentially very dense in the sense that the number of nonzero entries may be as large as p. The term $A^T\boldsymbol{\beta}$ captures the direct correlation between the predictors and the outcome y and the nonzero elements may correspond to the group of "core genes"

(see Section 1.2), while it is zero for the rest of the predictors. Note that the model presented here can sometimes be referred to as a factor model (Fan et al. 2021; Fan, Lou, and Yu 2024), which is a special case of latent variable models. However, the latter are more general as they include other types like latent class and structural equation models. In addition, traditional factor models typically assume that Ψ is diagonal.

To allow for the possibility of diverging p, we need some additional assumptions to make the model well-defined. Assume that when p grows, new columns are added to A, and rows and columns are added to Ψ, and impose the following restrictions on the matrices, similarly to that of Goeman (2006, chap. 6):

A.1 There are constants $0 < k \leq K < \infty$ such that all eigenvalues of Ψ are between k and K for all p.

A.2 The limit $\lim_{p\to\infty} \frac{1}{p} AA^T \in \mathbb{R}^{m\times m}$ exists and is of full rank m.

We may note a few implications of A.1 and A.2. First, they neatly separate the covariance matrix into a structural part, A^TA, and a noise part Ψ. Thus, $\mathbf{x}$ can be considered coming from a spiked covariance model (Johnstone 2001), that is that the ordered eigenvalues $\omega_1, \ldots, \omega_p$ of $\mathrm{E}[\mathbf{x}\mathbf{x}^T] = A^TA + \Psi$ are distributed as follows:

$$\omega_1 > \omega_2 > \cdots > \omega_m \gg \omega_{m+1} \geq \cdots \geq \omega_p > 0,$$

where the "non-spiked" eigenvalues $\omega_{m+1}, \ldots, \omega_p$ are bounded due to A.1, and the spiked eigenvalues $\omega_1, \ldots, \omega_m = \mathcal{O}(p)$ due to A.2. Second, A.2 ensures that there is a non-vanishing proportion of nonzero entries in A as the dimension p grows, which is referred to as the *pervasiveness* assumption in the latent factor model literature (Fan, Liao, and Mincheva 2013). This means that when the number of variables p is very large, we have abundant information about the underlying factor $\mathbf{f}$ in $\mathbf{x}$.

While the number of nonzero entries of $\boldsymbol{\gamma}_0$ grows with p, the conditions A.1 and A.2 in fact ensure that $\|\boldsymbol{\gamma}_0\|_1$ is approximately sparse in one specific way; the ℓ_1-norm stays bounded for $p \to \infty$, as shown in the following proposition, with proof given in the Appendix. Note, however, that $\|\boldsymbol{\gamma}_0\|_1$ may be large if $K \gg k$.

Proposition 1. Under A.1 and A.2,

$$\|\boldsymbol{\gamma}_0\|_1 = \mathcal{O}(1). \tag{2}$$

This proposition indicates that while the assumption of sparsity (in the ℓ_0 sense) implies a small ℓ_1-norm, the converse is not necessarily true. The joint latent variable model (1) is thus an example of a situation with small ℓ_1-norm but no sparsity.

As argued in Section 1.2, this latent variable model is suitable for modeling the types of data that have become common in omics research. We now turn to show that under the latent variable model, a sparse Lasso solution gives favorable prediction bounds, even when $p \to \infty$, if the regularization parameter is chosen accordingly.

3. Lasso Prediction Under the Joint Latent Variable Model

Suppose we have n observations from the model (1), giving us the design matrix $X \in \mathbb{R}^{n\times p}$ and the vector of observed outcomes $\mathbf{y} \in \mathbb{R}^n$. The Lasso estimator Tibshirani (1996) is defined as

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\|\mathbf{y} - X\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1 \right\}, \tag{3}$$

where $\lambda \geq 0$ is a regularization parameter that penalizes large values of $\|\boldsymbol{\gamma}\|_1$ and typically forces many of the entries of $\boldsymbol{\gamma}$ to be exactly zero.

For large p, any sparse approximation $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}_0 = (A^TA + \Psi)^{-1}A^T\boldsymbol{\beta}$ may quickly become very different from $\boldsymbol{\gamma}_0$ in terms of the estimation error $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2$ since the number of nonzero entries of $\boldsymbol{\gamma}_0$ grows linearly with p. However, for prediction accuracy, it is sufficient that $\|\boldsymbol{\gamma}_0\|_1$ remains bounded, as shown in Theorem 1. Of course, a more accurate estimate of $\boldsymbol{\gamma}_0$ can be obtained by plugging in estimates of A, $\boldsymbol{\beta}$ and Ψ. A formal estimation of these model parameters would require additional identifiability conditions on the model (1). However, in this work, we are interested in the prediction performance of a Lasso solution under a joint variable model, not the estimation of the factor loadings themselves. We refer to Goeman (2006) and references therein for a more detailed treatment of statistical inference under the current framework.

We use the following measure of prediction performance:

$$\mathrm{MSE}(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\|X\boldsymbol{\gamma}_0 - X\hat{\boldsymbol{\gamma}}\|_2^2 = (\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}})^T\hat{\Sigma}(\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}})$$

where $\hat{\Sigma} = X^TX/n$. Note that under the normality assumption for $(y, \mathbf{x}^T)$, $\mathrm{MSE}(\hat{\boldsymbol{\gamma}})$ differs from the prediction error for fixed designs, $\frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\gamma}}\|_2^2$, by an amount $\sigma_{|x}^2$, defined as

$$\sigma_{|x}^2 := \boldsymbol{\beta}^T\left(I - (A\Psi^{-1}A^T + I)A\Psi^{-1}A^T\right)\boldsymbol{\beta} + \sigma^2.$$

When p is large, the identity matrix becomes negligible compared to $A\Psi^{-1}A^T$, and $\sigma_{|x}^2 \approx \sigma^2$ (Goeman 2006). Hence, $\mathrm{MSE}(\hat{\boldsymbol{\gamma}})$ adequately reflects the excess prediction risk for the fixed design case.

The simplest bound for Lasso prediction error, valid for an arbitrary design matrix X, is given by

$$\mathrm{MSE}(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\|X\boldsymbol{\gamma}_0 - X\hat{\boldsymbol{\gamma}}\|_2^2 \leq \frac{2}{n}\lambda\|\boldsymbol{\gamma}_0\|_1 \tag{4}$$

on the set $\mathcal{T} = \{\sup_{\boldsymbol{\gamma}} 2\sigma\|\epsilon^TX\boldsymbol{\gamma}\|/\|\boldsymbol{\gamma}\|_1 \leq \lambda\}$ (Hebiri and Lederer 2012, eq. (2)). By Proposition 1, this bound may be favorable and it holds for any value of λ. However, some λ make the probability of being on the set $\mathcal{T}$ higher than others. Usually, λ is set to be proportional to $\sigma\sqrt{n\log(p)}$. However, this choice of λ is not optimal if the columns of X are highly correlated, as in the case of model (1) when p is large; the optimal choice of λ is typically smaller for highly correlated designs than for orthogonal or weakly correlated designs (Hebiri and Lederer 2012; Van de Geer and Lederer 2013). Still, for nearly collinear designs, Dalalyan, Hebiri, and Lederer (2017) shows that even the universal tuning parameter $\lambda \propto \sigma\sqrt{n\log(p)}$ can yield favorable prediction bounds for the Lasso.

Under the joint latent variable model (1), a faster prediction error bound than (4) is achievable for a properly chosen penalty parameter λ. To establish this, we use results from Van de Geer and Lederer (2013), which addresses highly correlated designs. They consider a design matrix to be highly correlated if the rate

at which the eigenvalues of $\hat{\Sigma} := X^T X/n$ decay is sufficiently fast. The structural assumption of spectral decay has recently gained more attention for its agreement with real-life examples; see Silin and Fan (2022) and references therein. For the model (1), condition A.2 (pervasiveness assumption) implies that the m spiked eigenvalues $\omega_1, \ldots, \omega_m$ grow linearly with p, while the non-spiked eigenvalues $\omega_{m+1}, \ldots, \omega_p$ will increase in number but remain bounded (due to Condition A.1). Consequently, the ordered set of empirical eigenvalues $\tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_p$ will indeed decay fast when p grows. The rate of this decay in the limit $p \to \infty$ depends on the ratio p/n and the magnitude of the leading population eigenvalues $\omega_1, \ldots, \omega_m$ (Wang and Fan 2017; Cai, Han, and Pan 2020). See Baik and Silverstein (2006) for the (almost sure) relationship between $\tilde{\omega}_i$ and ω_i for the case where $\omega_{m+1}, \ldots, \omega_p = 1$ and $p, n \to \infty$. By adapting Lemma 6.1 and Corollary 5.2 from Van de Geer and Lederer (2013) to our setting, we derive an improved bound, as stated in Theorem 1. The proof is given in the Appendix.

Theorem 1. Assume that A.1 and A.2 hold and $p \geq n$. Then, there exists positive constants C and c_0 such that for any $t > 0$, $p \geq c_0$, and

$$\lambda = \left[C\sigma^2 \left(\frac{(m + c_0 p^{-1})^{1/2}}{\sqrt{2} - 1} + t \right) \right]^{4/3} \|\boldsymbol{\gamma}_0\|_1^{-1/3}$$

the bound

$$\frac{1}{n}\|X(\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}})\|_2^2 \leq \frac{21}{2n} \left[C\sigma^2 \left(\frac{(m + c_0 p^{-1})^{1/2}}{\sqrt{2} - 1} + t \right) \right]^{4/3} \|\boldsymbol{\gamma}_0\|_1^{2/3} \quad (5)$$

holds with probability at least $1 - \exp[-t^2]$.

Remark 1. We did not attempt to optimize the constants in (5) and the bound may not be tight.

Remark 2. The constant c_0 is largely determined by Ψ and its eigenvalues (see the proof of Theorem 1 for details); in particular by the quantity

$$\frac{1}{\tilde{\omega}_1} \sum_{j=m+1}^{n} \tilde{\omega}_j,$$

which we will refer to as the "partial effective rank." This is related to the idea of effective rank found in the literature on high-dimensional sample covariance estimators (Koltchinskii and Lounici 2017).

Under the latent variable model, the prediction bound (5) in fact decreases with increasing p, and increases with increasing m. This is in sharp contrast to the prediction error bounds for the high-dimensional linear model under the sparsity assumption, in which increasing the number of (noise) variables can only hurt the prediction accuracy (Flynn, Hurvich, and Simonoff 2017). However, our results align well with the conclusions drawn in Greenshtein and Ritov (2004), that increasing the number of variables p does not hurt the prediction accuracy of a sparse Lasso solution, and that "Occam's razor does not seem relevant for prediction." Furthermore, consistent with the findings for highly correlated designs, the optimal λ is much smaller than in traditional results. This is intuitive under the current (dense) model, as a large p implies that many variables contribute useful information for predicting y, and the model should therefore not be overly penalized.

In the subsequent sections, we complement these theoretical insights through numerical studies, illustrating how the prediction accuracy is affected by the dimensionality and the correlation structure of X.

4. Numerical Studies

For data generation, we set the number of latent variables to $m = 3$. We let the number of nonzero entries of each row of A_j, $j \in \{1, \ldots, 3\}$ be to $0.2p$, that is, they are pervasive factors. The entries of A are fixed, but they are generated from $\mathcal{U}(-1, 1)$. The latent variables $\mathbf{f} \sim \mathcal{N}(0, 1)$ and the noise $\epsilon \sim \mathcal{N}(0, 1)$. We fix $n = 100$, but vary the number of predictors $p \in \{5, \ldots, 10000\}$. We consider four settings of Ψ:

1. Ψ_1 is the $p \times p$ identity matrix,
2. $\Psi_2 = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, where a random subset of size $p/2$ are considered common variants (low variance) and the other $p/2$ are rare (high variance). The low variances are generated from $\mathcal{U}(0.01, 0.1)$ and the high variances from $\mathcal{U}(0.5, 2)$,
3. Ψ_3 has block-diagonal structure, and Toeplitz correlation structure within blocks. The correlation coefficient within blocks is set to 0.8 and the between-blocks correlation is 0.1. In addition, the eigenvalues of Ψ are forced to be between 0.1 and 3. The number of blocks is set to $\lceil \sqrt{p} \rceil$.
4. Ψ_4 is a random non-diagonal matrix with eigenvalues between 0.1 and 3 for all p.

Simulation settings 1 and 2 correspond to uncorrelated error, while settings 3 and 4 have correlated error. To evaluate the out-of-sample prediction error, we consider not only $\text{MSE}(\hat{\boldsymbol{\gamma}})$, but also

$$\text{PE}(\hat{\boldsymbol{\gamma}}) = \text{E}[(\mathbf{x}^T\boldsymbol{\gamma}_0 - \mathbf{x}^T\hat{\boldsymbol{\gamma}})^2] = (\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}})^T \Sigma (\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}),$$

where $\Sigma = A^T A + \Psi$. Similar to $\text{MSE}(\hat{\boldsymbol{\gamma}})$ in the fixed design setting, $\text{PE}(\hat{\boldsymbol{\gamma}})$ reflects the excess prediction risk in the random design case, particularly for large p. We include this additional measure of prediction error because, in high dimensions, the matrices $\hat{\Sigma}$ and Σ may differ considerably, and it is thus of interest to see how this is reflected in the prediction performance of the Lasso. Since $\boldsymbol{\gamma}_0$ also changes with p, we report the *relative* MSE and PE, that is, $\text{MSE}(\hat{\boldsymbol{\gamma}})/\text{MSE}(0)$ and $\text{PE}(\hat{\boldsymbol{\gamma}})/\text{PE}(0)$.

Even when $p > n$, the Lasso cannot select more than a total of n variables. Let $\hat{\boldsymbol{\gamma}}_n$ be the Lasso estimate for the maximum number of selected variables. For simpler comparisons, we report the results for $s \in [0, 1]$, where $s = \|\hat{\boldsymbol{\gamma}}\|_1 / \|\hat{\boldsymbol{\gamma}}_n\|_1$ instead of λ, but we note that there exists a λ for any s that gives the same solution and that larger values of s correspond to smaller λ. We also consider the standardized estimation error $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 / \|\boldsymbol{\gamma}_0\|_2^2$. We repeated the simulation 100 times and report the average.

Figure 2 shows the results for the case $\Psi = \Psi_3$. Note that we do not include the results for $p = n = 100$ as this often results in

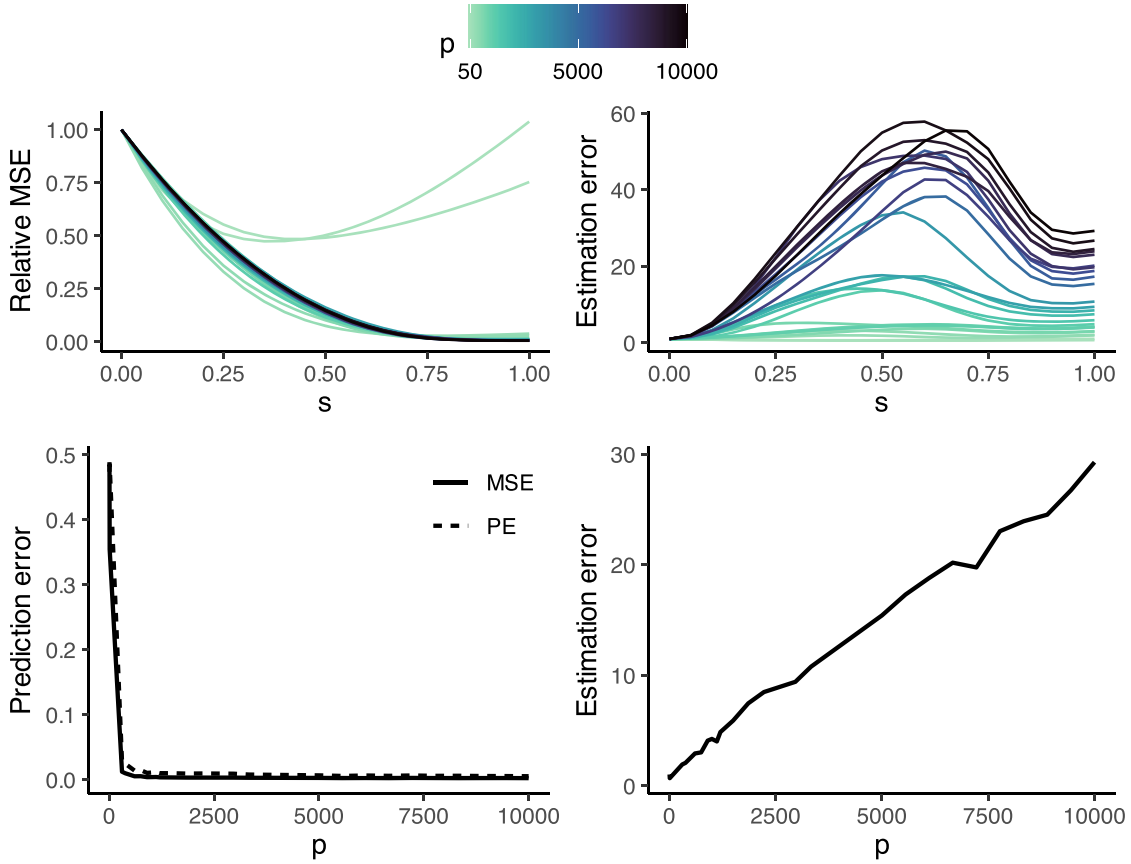


Figure 2. Top display: the relative MSE (left) and standardized ℓ_2 estimation error (right) for $s \in \{0, \ldots, 1\}$ for different number of variables p. Bottom display: the relative MSE and relative PE at the optimal choice of s (left), and the ℓ_2 estimation error for $s = 0.5$ (right). $\Psi = \Psi_3$.

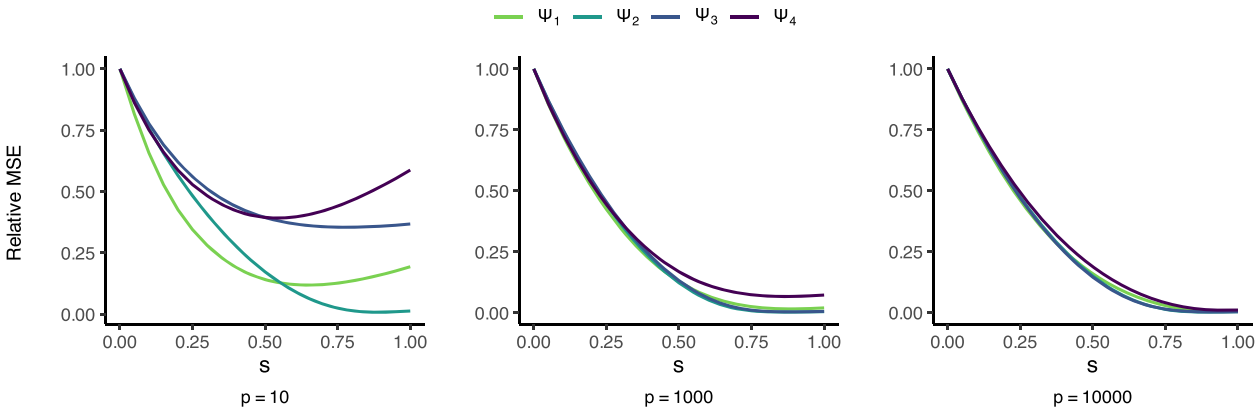


Figure 3. The relative MSE for $p \in \{10, 1000, 10{,}000\}$ for the four different examples of Ψ. The structure of Ψ becomes less influential when the number of variables p is large.

a much worse prediction performance that will overshadow the other curves. This worsening of performance is associated with the "double descent" phenomenon, which has recently gained much attention in the machine learning literature (Belkin et al. 2019).

From Figure 2, it is clear that the prediction error of the Lasso indeed decreases for increasing number of variables p, while the estimation error naturally increases. Similar figures for the other cases can be found in Web Appendix A. The decay of the relative MSE as a function of p (bottom left in Figure 2) is also nicely captured by the in-sample prediction bound (5). The difference between the in-sample and out-of-sample prediction error in this case is minimal, showing that the predictions are stable with respect to new observations.

Figure 3 displays the relative MSE for different settings of Ψ, across three values of p. The merging of the four curves as p increases suggests that the structure of Ψ becomes less influential as p increases, which is due to the corresponding strengthening of the signal (under A.1 and A.2).

Figure 4 shows the prediction performance at the optimal value of $s \in [0, 1]$ as a function of $p > n$ for $\Psi_1, \ldots, \Psi_4$. As expected, the out-of-sample prediction error (PE) is consistently larger than the in-sample prediction error (MSE) across all structures of Ψ. However, PE shows a similar decay as p

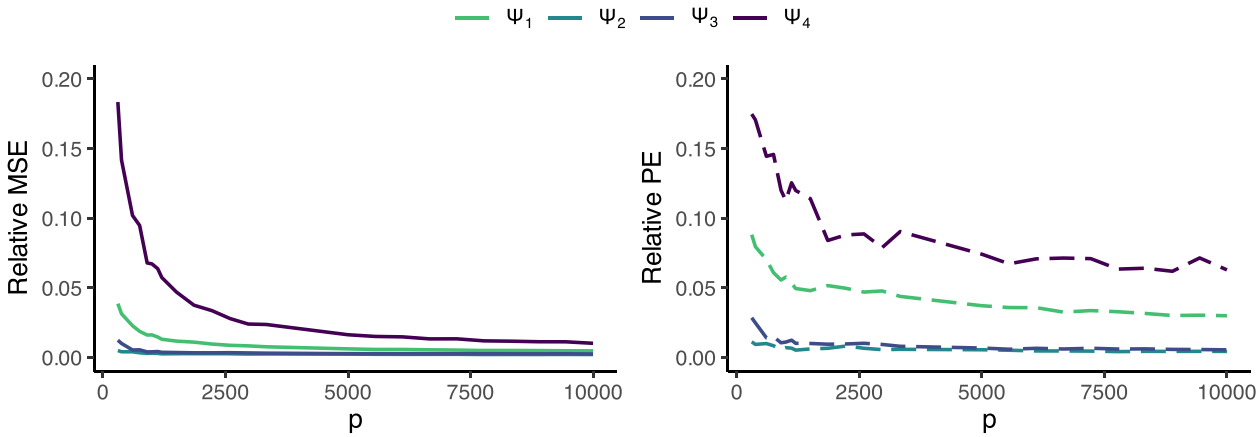


Figure 4. The relative MSE (left) and PE (right) at the optimal choice of s for the different choices of Ψ, as a function of $p > n$.

increases. Furthermore, it is clear that both MSE and PE are worse for Ψ_1 and Ψ_4 compared to Ψ_2 and Ψ_3. This can indeed be related to the constant c_0 in (5), as the partial effective ranks of the resulting design matrices corresponding to Ψ_1 and Ψ_4 are much larger than for Ψ_2 and Ψ_3 (see Figure 4 in Web Appendix A). However, the constant c_0 determines only an upper bound and does not completely account for the ranking of the different Ψ structures, which depends not only on the eigenvalues of Ψ but also on its eigenvectors.

As additional exploration (Web Appendix B), we show that under the latent variable model, there exist multiple disjoint subset that are have comparable predictive accuracy. This phenomenon is further illustrated using the real breast cancer dataset of Van De Vijver et al. (2002).

5. Discussion

The aim of this work has been to highlight the importance of caution when interpreting sparse prediction rules, particularly in the context of high-dimensional omics data. To this end, we examined Lasso's prediction performance when data are generated from a latent variable model, which we have presented as a plausible and dense alternative to the usual sparsity assumption made when analyzing omics datasets. Our theoretical results rely on a rapid decay of the eigenvalues of the sample covariance matrix $\hat{\Sigma}$. This condition is not difficult to verify in practice; one can simply examine the eigenspectrum of the data. The latent variable model is thus an example of a data-generating process that produces such eigenvalue decay whenever p is large, and under which the Lasso yields good prediction results.

While it is well understood that a strong predictive rule does not necessarily imply an accurate representation of the underlying truth, the use of sparse methods are often motivated by the fact that they are interpretable. For example, when genes, SNPs, or proteins are identified as potential novel biomarkers using sparse methods, they are frequently reported as "important variables" and implicitly regarded as potential targets for future treatment or intervention. It is, however, important to distinguish between two different goals: to find disease-related "master" genes, or to construct a prognostic tool.

Nonetheless, an accurate sparse prediction rule can be extremely useful. For instance, in prognostic tests, a sparse biomarker panel, comprising only a few gene expressions or proteins, reduces the need for extensive measurements (e.g., the whole genome) for each patient (Rahnenführer et al. 2023). Sparse predictors are thus economical and efficient, and makes the results more actionable for both clinicians and researchers (Klein et al. 2014, chap. 15). While a sparse prediction rule in the $p \gg n$ setting is not stable, it does not mean that the selected variables are not useful, as they are indeed very useful in terms of prediction; they are unstable as a prediction rule, but stable as a prediction. Furthermore, as $p \gg n$ and the true model is dense, there is no hope of estimating the regression coefficient $\boldsymbol{\gamma}_0$ correctly. Hence, we might as well use a sparse predictor. This is the "bet on sparsity" principle, as coined by Hastie, Tibshirani, and Friedman (2009).

In this work, we have focused on the Lasso as it is the most widely used and theoretically studied sparse regression method. However, we believe that the main conclusions also apply to other sparse methods, including those that use non-convex penalties, such as SCAD (Fan and Li 2001) and MCP (Zhang 2010). While these methods require less strict assumptions for achieving model selection consistency by reducing the bias induced by ℓ_1 penalty, this does not resolve the fundamental issue of interpretation when the true model is dense. The main take-away is that sparse solutions in the setting of high-dimensional omics data should be treated with caution, serving as a powerful tool for prediction rather than a definitive list of biologically important variables.

It is not surprising that the Lasso yields good prediction results in the present setting. As p increases, so does the information about the underlying factors. As $p \gg n$, many columns of X are near linear combinations of each other, and Lasso tends to select the optimal linear combinations in terms of MSE. An interesting further study would be to examine the variable selection properties of the Lasso in the limit $p \to \infty$, as the selected variables by the Lasso would perhaps resemble the directions chosen by partial least squares.

In this work, we consider the setting where p diverges while m remains fixed. When m is very large, the prediction error bound (5) becomes correspondingly large. However, a large m implies

that the problem is not sparse in any sense of the word, and it cannot be expected that the Lasso would be an appropriate method. In certain situations, it is reasonable to assume that m also grows with p, albeit at a slow rate. We believe that our theoretical results can be extended to this case, but this is beyond the scope of this current work.

Our results confirm the virtue of regularization techniques when working with high-dimensional data to avoid overfitting and serve as an illustration of the conclusions drawn in Greenshtein and Ritov (2004); under the present model, there is no harm in introducing many more explanatory variables than observations as long as some constraint (on the ℓ_1-norm) is placed on the solution.

Appendix A. Appendix

Proof of Proposition 1. First note that $\boldsymbol{\gamma}_0 = (A^TA + \Psi)^{-1}A^T\boldsymbol{\beta} = \Psi^{-1}A^T(A\Psi^{-1}A^T + I)^{-1}\boldsymbol{\beta}$ due to the Woodbury identity. We then have that

$$\begin{aligned}\|\boldsymbol{\gamma}_0\|_1 &= \|\Psi^{-1}A^T(A\Psi^{-1}A^T + I)^{-1}\boldsymbol{\beta}\|_1 \\ &\underset{(i)}{\leq} \sqrt{p}\|\Psi^{-1}A^T(A\Psi^{-1}A^T + I)^{-1}\boldsymbol{\beta}\|_2 \\ &\leq \sqrt{p}\|\Psi^{-1}\|_2\|A^T\|_2\|(A\Psi^{-1}A^T + I)^{-1}\|_2\|\beta\|_2,\end{aligned}$$

where (i) is because $\|x\|_1 \leq \sqrt{d}\|x\|_2$ for any d-vector x. Now, $\|\Psi^{-1}\|_2 = 1/k$ due to A.1, $\|A^T\|_2 = \mathcal{O}(\sqrt{p})$ due to A.2 and $\|(A\Psi^{-1}A^T + I)^{-1}\|_2 = \mathcal{O}(K/p)$ and $\|\beta\|_2 = \mathcal{O}(1)$. Hence,

$$\|\boldsymbol{\gamma}_0\|_1 \leq \frac{\sqrt{p}}{k}\mathcal{O}(\sqrt{p})\mathcal{O}(Kp^{-1}) = \mathcal{O}(1).$$

□

Proof of Theorem 1. We follow the lines of the proof of Lemma 6.1 of Van de Geer and Lederer (2013) to show that Corollary 5.2 of Van de Geer and Lederer (2013) holds in our setting. Here, they assume that $\|X_j\|_2^2/n \leq 1$ for each column X_j, $j \in \{1,\ldots,p\}$ of X. To fulfill this assumption in our case that allows for $p \to \infty$, we will consider a scaled version of X, $X' = \frac{1}{\sqrt{\tilde{\omega}_1}}X$, where $\sqrt{\tilde{\omega}_1}$ is the maximum singular value of X. Then, for any p, $\|X'_j\|_2^2/n \leq 1$. Let $\hat{\Sigma} = U\Omega U^T$ be the eigendecomposition of $\hat{\Sigma}$, where $U^TU = UU^T = I$ and $\Omega = \text{diag}(\tilde{\omega}_1,\ldots,\tilde{\omega}_p)$, and since $p \geq n$, the last $n-p$ eigenvalues are equal to zero. The spiked sample eigenvalues $\tilde{\omega}_1,\ldots,\tilde{\omega}_m = \mathcal{O}(p)$, while the remaining $n-m$ sample eigenvalues are bounded for any p (Cai, Han, and Pan 2020). Denote by $\mathbf{u}_i$ the ith column of U and note that $\sum_{i=1}^p(\mathbf{u}_i^T\boldsymbol{\gamma})^2 = \boldsymbol{\gamma}^TUU^T\boldsymbol{\gamma} = \|\boldsymbol{\gamma}\|_2^2 \leq \|\boldsymbol{\gamma}\|_1 = 1$. Let $\mathcal{F} := \{X'\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_1 = 1\}$ and consider any $f \in \mathcal{F}$ and note that since $\|\boldsymbol{\gamma}\|_1 = 1$, $\|f\|_n \leq 1$, and

$$\begin{aligned}\|f\|_n^2 &= \frac{1}{n}\sum_{i=1}^n f_i = \frac{1}{\tilde{\omega}_1}\boldsymbol{\gamma}^T\hat{\Sigma}\boldsymbol{\gamma} = \sum_{i=1}^p\frac{\tilde{\omega}_i}{\tilde{\omega}_1}(\mathbf{u}_i^T\boldsymbol{\gamma})^2 \\ &= \sum_{i=1}^m\frac{\tilde{\omega}_i}{\tilde{\omega}_1}(\mathbf{u}_i^T\boldsymbol{\gamma})^2 + \sum_{i=m+1}^n\frac{\tilde{\omega}_i}{\tilde{\omega}_1}(\mathbf{u}_i^T\boldsymbol{\gamma})^2 \\ &\leq \frac{\tilde{\omega}_1}{\tilde{\omega}_1}\sum_{i=1}^m(\mathbf{u}_i^T\boldsymbol{\gamma})^2 + \frac{\tilde{\omega}_{m+1}}{\tilde{\omega}_1}\sum_{i=m+1}^n(\mathbf{u}_i^T\boldsymbol{\gamma})^2 \leq 1 + \frac{c_0}{p}, \quad (6)\end{aligned}$$

for some positive constant c_0, where the last inequality is due to the fact that $\tilde{\omega}_{m+1}/\tilde{\omega}_1 = \mathcal{O}(p^{-1})$ and $\sum_{i=1}^p(\mathbf{u}_i^T\boldsymbol{\gamma})^2 \leq 1$.

Define $f_m := P(f)$ as the projection operator that maps f to a m-dimensional vector so that $\|f_m\|_n^2 = \sum_{i=1}^m\frac{\tilde{\omega}_i}{\tilde{\omega}_1}(\mathbf{u}_i^T\boldsymbol{\gamma})^2$, and $\mathcal{F}_m := \{P(f) : f \in \mathcal{F}\}$. Let $f_{m^c} = f - f_m$ be the part of f orthogonal to f_m. Then,

$$\begin{aligned}\|f\|_n^2 = \|f_m\|_n^2 + \|f_{m^c}\|_n^2 &\underset{(i)}{\leq} \|f_m\|_n^2 + \frac{c_0}{p} \\ \implies \|f_{m^c}\|_n = \|f - f_m\|_n &\leq \sqrt{\frac{c_0}{p}},\end{aligned}$$

where (i) is due to (6).

Define $\delta := \sqrt{c_0p^{-1}}$. Since $p \geq c_0$, $0 < \delta \leq 1$. Let $\{x_1,\ldots,x_N\} \subset \mathbb{R}^m$ be the minimal δ-covering of $\mathcal{F}_m$ with corresponding cover number $N := N(\delta, \mathcal{F}_m, \|\cdot\|_n)$, that is, the minimal number of balls of radius δ with respect to $\|\cdot\|_n$ required to cover the set. Note that since $\|f_m\|_n \leq 1$, $\mathcal{F}_m$ is an m-ball with radius 1, whose δ-covering number is upper bounded by $(3/\delta)^m$ (Bühlmann and van de Geer 2011, Lemma 14.27).

Now, for each $f_m \in \mathcal{F}_m$, $\|f_m - x_i\|_n \leq \delta$ for some x_i, $i \in \{1,\ldots,N\}$ (by definition of δ-covering). Let $\{\tilde{x}_i,\ldots,\tilde{x}_N\} \in \mathbb{R}^n$ be the set consisting of each x_i appended with $n-m$ zeros. Thus, for all $f \in \mathcal{F}$,

$$\|f - \tilde{x}_i\|_n \leq \|f - f_m\|_n + \|f_m - x_i\|_n \leq 2\delta.$$

Hence, $\{\tilde{x}_i,\ldots,\tilde{x}_N\}$ is a 2δ-covering number of $\mathcal{F}$ (not necessarily minimal), and we get that

$$\begin{aligned}N(2\delta, \mathcal{F}, \|\cdot\|_n) \leq N(\delta, \mathcal{F}_m, \|\cdot\|_n) &\leq \left(\frac{3}{\delta}\right)^m \\ \implies N(\delta, \mathcal{F}, \|\cdot\|_n) &\leq \left(\frac{6}{\delta}\right)^m \\ \implies \log(1 + 2N(\delta, \mathcal{F}, \|\cdot\|) &\leq \left(\frac{m + c_0p^{-1}}{\delta}\right),\end{aligned}$$

where $\delta = \sqrt{c_0p^{-1}}$. Thus, Corollary 5.2 Van de Geer and Lederer (2013) holds with $A = m + c_0p^{-1}$ and $\alpha = 1/2$. This means that the bound (5) holds with probability at least $1 - \exp[-t^2](1 + 2B)$, where

$$B := \left(\exp\left[\frac{m + c_0p^{-1}}{4(\sqrt{2}-1)^2} - 1\right]\right)^{-1}.$$

B becomes extremely small even for small m and c_0, so we set $B \approx 0$. This completes the proof. □

Supplementary Materials

Web Appendix A contains additional simulation results. Web Appendix B contains explorations of the impact of sequential removal of variables on prediction.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement and Disclosure of AI

The breast cancer dataset used in the real data example (Appendix C.2) is openly available in the Bioconductor package 'breastCancerNKI'. Minor language corrections were supported by Generative AI tools, in particular, ChatGPT (OpenAI, GPT-4) and Gemini 2.5 (Google). These tools were used to improve grammar and clarity during revision. All intellectual content is the work of the authors.

ORCID

Andrea Bratsberg http://orcid.org/0000-0002-6784-5687

References

Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013), "Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease," *Genetic Epidemiology*, 37, 184–195. [1]

Ajana, S., Acar, N., Bretillon, L., Hejblum, B. P., Jacqmin-Gadda, H., and Delcourt, C. (2019), "Benefits of Dimension Reduction in Penalized Regression Methods for High-Dimensional Grouped Data: A Case Study in Low Sample Size," *Bioinformatics*, 35, 3628–3634. [1]

Baik, J., and Silverstein, J. W. (2006), "Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models," *Journal of Multivariate Analysis*, 97, 1382–1408. [5]

Ball, P. (2023), *How Life Works: A User's Guide to the New Biology*, Chicago, IL: University of Chicago Press. [2]

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019), "Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off," *Proceedings of the National Academy of Sciences*, 116, 15849–15854. [6]

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017), "An Expanded View of Complex Traits: From Polygenic to Omnigenic," *Cell*, 169, 1177–1186. [2]

Bühlmann, P., and van de Geer, S. (2011), "Probability and Moment Inequalities," in *Statistics for High-Dimensional Data: Methods, Theory and Applications*, pp. 481–538, Berlin, Heidelberg: Springer. [8]

Cai, T. T., Han, X., and Pan, G. (2020), "Limiting Laws for Divergent Spiked Eigenvalues and Largest Nonspiked Eigenvalue of Sample Covariance Matrices," *The Annals of Statistics*, 48, 1255–1280. [5,8]

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [2]

Dalalyan, A. S., Hebiri, M., and Lederer, J. (2017), "On the Prediction Performance of the Lasso," *Bernoulli*, 23, 552–581. [4]

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005), "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?" *Bioinformatics*, 21, 171–178. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [7]

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [4]

Fan, J., Lou, Z., and Yu, M. (2024), "Are Latent Factor Regression and Sparse Regression Adequate?" *Journal of the American Statistical Association*, 119, 1076–1088. [2,4]

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021), "Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning," *Statistical Science*, 36, 303–327. [4]

Flynn, C. J., Hurvich, C. M., and Simonoff, J. S. (2017), "On the Sensitivity of the Lasso to the Number of Predictor Variables," *Statistical Science*, 32, 88–105. [5]

Giannone, D., Lenza, M., and Primiceri, G. E. (2021), "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437. [1]

Goeman, J. J. (2006), "Statistical Methods for Microarray Data," Ph. D. thesis, Universiteit Leiden. [2,4]

Goeman, J. J., and Böhringer, S. (2020), "Comments On: Hierarchical Inference for Genome-Wide Association Studies by Jelle J. Goeman and Stefan Böhringer," *Computational Statistics*, 35, 41–45. [2]

Greenshtein, E., and Ritov, Y. (2004), "Persistence in High-Dimensional Linear Predictor Selection and the Virtue of Overparametrization," *Bernoulli*, 10, 971–988. [1,5,8]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.), New York: Springer. [7]

Hebiri, M., and Lederer, J. (2012), "How Correlations Influence Lasso Prediction," *IEEE Transactions on Information Theory*, 59, 1846–1854. [4]

Johnstone, I. M. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of Statistics*, 29, 295–327. [4]

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2014), *Handbook of Survival Analysis*, Boca Raton, FL: CRC Press. [2,7]

Koltchinskii, V., and Lounici, K. (2017), "Normal Approximation and Concentration of Spectral Projectors of Sample Covariance," *The Annals of Statistics*, 45, 121–157. [5]

Leek, J. T. (2011), "Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data," *Biometrics*, 67, 344–352. [2]

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012), "Genomic Selection Using Regularized Linear Regression Models: Ridge Regression, Lasso, Elastic Net and their Extensions," in *BMC Proceedings* (Vol. 6), pp. 1–6, Springer. [1]

Pes, B., Dessì, N., and Angioni, M. (2017), "Exploiting the Ensemble Paradigm for Stable Feature Selection: A Case Study on High-Dimensional Genomic Data," *Information Fusion*, 35, 132–147. [2]

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A.-L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., et al. (2023), "Statistical Analysis of High-Dimensional Biomedical Data: A Gentle Introduction to Analytical Goals, Common Approaches and Challenges," *BMC Medicine*, 21, 182. [7]

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression Approaches for Microarray Data Analysis," *Journal of Computational Biology*, 10, 961–980. [1]

Segura-Lepe, M. P., Keun, H. C., and Ebbels, T. M. (2019), "Predictive Modelling Using Pathway Scores: Robustness and Significance of Pathway Collections," *BMC Bioinformatics*, 20, 1–11. [2]

Silin, I., and Fan, J. (2022), "Canonical Thresholding for Non-Sparse High-Dimensional Linear Regression," *Annals of Statistics*, 50, 460–486. [5]

Sun, N., and Zhao, H. (2004), "Genomic Approaches in Dissecting Complex Biological Pathways," *Pharmacogenomics*, 5, 163–179. [1]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1,4]

Van De Geer, S. (2018), "On Tight Bounds for the Lasso," *Journal of Machine Learning Research*, 19, 1–48. [2]

Van de Geer, S., and Lederer, J. (2013), "The Lasso, Correlated Design, and Improved Oracle Inequalities," in *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner* (Vol. 9), pp. 303–317, Institute of Mathematical Statistics. [4,5,8]

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *New England Journal of Medicine*, 347, 1999–2009. [7]

Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C., and Jurisica, I. (2011), "Optimized Application of Penalized Regression Methods to Diverse Genomic Data," *Bioinformatics*, 27, 3399–3406. [1]

Wang, H., Lengerich, B. J., Aragam, B., and Xing, E. P. (2019), "Precision Lasso: Accounting for Correlations and Linear Dependencies in High-Dimensional Genomic Data," *Bioinformatics*, 35, 1181–1187. [1,2]

Wang, W., and Fan, J. (2017), "Asymptotics of Empirical Eigenstructure for High Dimensional Spiked Covariance," *Annals of Statistics*, 45, 1342–1374. [5]

Xu, H., Caramanis, C., and Mannor, S. (2011), "Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 187–193. [2]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [7]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [2]

Zhao, Y., and Simon, R. (2010), "Development and Validation of Predictive Indices for a Continuous Outcome Using Gene Expression Profiles," *Cancer Informatics*, 9, CIN–S3805. [1]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [2]