HIGH-DIMENSIONAL INFERENCE FOR MODEL AVERAGING ESTIMATORS

Lise Léonard, Eugen Pircalabelu, Rainer von Sachs

LIDAM Discussion Paper ISBA
2025 / 14

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

High-dimensional inference for Model Averaging estimators

Lise Léonard, Eugen Pircalabelu and Rainer von Sachs

June 11, 2025

UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences
Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
lise.leonard@uclouvain.be, eugen.pircalabelu@uclouvain.be
and rainer.vonsachs@uclouvain.be.

Abstract

Selection methods for high-dimensional models are well developed, but they do not take into account the choice of the model, which leads to an underestimated variance. We propose a procedure for high-dimensional model averaging that allows inference even when the number of predictors is greater than the sample size. The proposed estimator is constructed from the debiased Lasso and the weights are chosen to reduce the prediction risk associated with them. We derive the asymptotic distribution of the estimator within a high-dimensional framework and offer guarantees for the minimal loss prediction obtained from the weights. With this, in contrast to existing approaches, our proposed method combines the advantages of model averaging with the possibility of inference based on asymptotic normality. In a simulation study and on a real, high-dimensional dataset, the estimator shows a smaller prediction risk than its competitors.

Keywords: Debiased Lasso, High-Dimensional Inference, Model Averaging, Prediction Risk.

1 Introduction

With the rapid evolution of technology, the amount of data to analyze is growing exponentially, leading to new challenges. In particular, samples with a size much smaller than the number of features are increasingly common in many research fields. For example, in biology, transcriptomic data have thousands or tens of thousands of gene expressions per sample unit. This situation does not occur only in biology, and in the last decades a great deal of work has already been done in the area of high-dimensional statistics. For an introduction to the subject and a survey of the existing literature, see Bühlmann and van de Geer (2011) and Giraud (2014).

In high-dimensional settings, model selection methods are very common and intensely used in practice to reduce the number of parameters. The most famous estimator for regression is the Lasso introduced by Tibshirani (1996), which selects the variables with the largest signal and shrinks the other coefficients to zero. Other model selection methods are based on information criteria such as the AIC (Akaike, 1979), BIC (Schwarz, 1978) or Mallow's Cp (Mallows, 1973). See Burnham and Anderson (2002) for an overview on model selection in statistics. However, one disadvantage of such methods is the loss of information for the unselected variables. Especially when it comes to performing inference, the variance estimation does not take into account the selection step, which can lead to an overestimated precision (Burnham and Anderson, 2002).

An alternative to model section is model averaging (MA) as it consists in aggregating several models instead of selecting just one. The aims are not only to propose an alternative to selection but also to reduce the risk associated with the estimator by incorporating information from different models.

Such averaging strategies were first developed for Bayesian techniques; for a literature review see Hoeting et al. (1999) and Raftery and Zheng (2003). In frequentist theory, MA techniques

have been developed more recently, namely over the last two decades. What is yet missing in the literature is an MA procedure which, in high-dimensional setting, achieves both a low prediction loss and is asymptotically normal. We fill this gap with our proposed procedure.

MA estimators are constructed by a weighted mean of different estimators obtained from different models and most of the time, the method is based on least squares with nested models. The choice of the weights is one important point of research. On this topic, Buckland et al. (1997) proposed weights based on the AIC, Hansen (2007) constructed an estimator based on Mallow's Cp weights, shown further in Hansen (2014) to provide minimal prediction risk for homoscedastic models. Hansen and Racine (2012) proposed a jackknife MA strategy for models with heteroscedastic errors, Chen et al. (2022) studies the consistency of MA weights based on BIC and Zhang et al. (2015) developed a Kullback-Leibler loss model averaging.

In high-dimensional settings, MA is not well developed and most methods still rely on model selection first. However, we can cite the work of Ando and Li (2014), who developed a two-step model averaging procedure using leave-one-out cross-validation, Xie et al. (2021), who extended the methods to models with missing responses and Lan et al. (2018) proposed a sequential model averaging procedure that works even in ultra-high dimensions. Recently, without any theoretical results, Feng and Liu (2020) also proposed an averaging estimator based on the Lasso path with high empirical performance in prediction loss reduction.

Apart from the risk, one is also interested in the asymptotic properties of the MA estimators. For low-dimensional models, Hjort and Claeskens (2003) studied the asymptotic properties of model averaging estimators constructed from likelihood-based models. For the distribution, Pötscher (2006) shows an impossibility result for the estimation of the distribution of a least squares model averaging and Charkhi et al. (2016) derived a non-standard distribution for another averaged estimator. Indeed, when the weights are selected from the data, they are random and even when each estimator is Gaussian, there is no guarantee to obtain a standard distribution for the random weighted mean.

In this paper, we propose a model averaging estimator for high-dimensional regression based on the debiased Lasso estimator, an asymptotically Gaussian estimator in high-dimensional settings (van de Geer et al., 2014). Our proposed averaging is done on the entire Lasso path and the weights are chosen based on an optimization problem that aims to reduce the prediction risk. We show that the proposed estimator is asymptotically Gaussian and that the chosen weight vector leads to the best asymptotic prediction loss among all weight vectors. Thus, the proposed estimator combines asymptotic normality and loss optimality in a high-dimensional setting. Furthermore, to the best of our knowledge, it is the only MA estimator that has been shown to be asymptotically Gaussian and prediction efficient.

The paper is organized as follows: Section 2 introduces the high-dimensional model, the debiased Lasso estimator and it discusses the challenges associated with the choice of the regularization parameter. In Section 3, we present the proposed estimator and the weight construction. Section 4 shows the theoretical properties of the new estimator. Section 5 uses a simulation study to show the practical performance of the estimator, and in Section 6 we show an application on a real dataset. Section 7 concludes with a discussion.

2 Framework

In this section, we briefly present the context in which we work. In particular, we succinctly introduce the high-dimensional model and the debiased Lasso. We then present the challenges associated with this estimator and illustrate the advantages of the proposed method with a simple toy example.

The classical high-dimensional linear regression model we use here is

$$Y = \mathbf{X}\beta_0 + \epsilon, \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ random design matrix, Y is an $n \times 1$ response vector, β_0 is the $p \times 1$ vector of coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is the vector of errors with I_n the $n \times n$ identity matrix. Let $S_0 = \{\beta_{0,j} : \beta_{0,j} \neq 0\}$ be the true active set with cardinality denoted by s_0. In this work, we allow for $p \geq n$ but we assume that $(\log p)/n = o(1)$.

The Lasso is defined as

$$\hat{\beta}_\lambda^{Lasso} := \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} ||Y - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1 \right), \tag{2}$$

where $||v||_2^2 = v^T v$ denotes the squared Euclidean norm and $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm of the coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^T$. The parameter λ is a regularization parameter yet to be determined.

Even if the Lasso estimator is consistent in a high-dimensional framework, it is biased in finite sample setting. In particular, the signal for the active variables is underestimated due to the penalization term (Fan and Li, 2001). To reduce this bias and also to perform inference on all the coefficients, one possibility is to use the debiased Lasso estimator, a desparsified estimator constructed from the Lasso.

The debiased Lasso was introduced by Zhang and Zhang (2014) and further developed in van de Geer et al. (2014) and Javanmard and Montanari (2014). It is derived from the following construction

$$\hat{\beta}_\lambda^d := \hat{\beta}_\lambda^{Lasso} + \frac{1}{n} \hat{\Omega} \mathbf{X}^T \left(Y - \mathbf{X} \hat{\beta}_\lambda^{Lasso} \right), \tag{3}$$

where $\hat{\beta}_\lambda^{Lasso}$ is an initial Lasso estimator with some chosen λ and $\hat{\Omega}$ is a $p \times p$ matrix that plays the role of the inverse of the sample variance-covariance matrix, $\hat{\Sigma} := (1/n)\mathbf{X}^T\mathbf{X}$. Indeed, in high-dimensional settings when $p > n$, $\text{rank}(\mathbf{X}^T\mathbf{X}) < p$ and so the matrix $\hat{\Sigma}$ is not invertible. The motivation for the debiased Lasso estimator comes from the following decomposition

$$\hat{\beta}_\lambda^d = \beta_0 + \frac{1}{n} \hat{\Omega} \mathbf{X}^T \epsilon + \left(I_p - \hat{\Omega}\hat{\Sigma} \right) \left(\hat{\beta}_\lambda^{Lasso} - \beta_0 \right). \tag{4}$$

The intuition is that the debiased Lasso estimator converges to the true parameter if $\hat{\Omega}\hat{\Sigma}$ is close to the identity matrix and if the Lasso estimator converges. Then, the estimator is motivated by the second term on the right-hand side as, conditionally on the design, the term $(1/\sqrt{n})\hat{\Omega}\mathbf{X}^T\epsilon$ is Gaussian. This allows for showing the asymptotic normality of the estimator. Another important motivation for using this estimator is the significant bias reduction it achieves in high-dimensional settings compared to the Lasso, hence its name.

The debiased Lasso estimator depends on the Lasso estimator which itself depends on the regularization parameter λ. This parameter is essential since it determines the weight of the penalization term in the Lasso equation and thus the sparsity of the final Lasso estimator. A small value for λ will result in a Lasso estimator with a small bias but a high variability, which carries over to the debiased Lasso estimator. On the other hand, a high value for λ results in a smaller variability and a larger degree of sparsity for the estimator but with a larger bias. Thus, the choice of this parameter is a trade-off between bias and variability. From a theoretical point of view, a parameter that is large enough and proportional to $\sigma\sqrt{(\log p)/n}$ is recommended to ensure that the mean squared prediction error converges to zero when n grows and p is a

function of n (Lahiri, 2021). However, these theoretical recommendations can not help in the practical choice of the parameter, as the noise level σ is unknown as well as the proportionality constant.

Many methods for selecting the regularization parameter have been proposed and a data-driven choice has prevailed in the literature. The most common methods are based on the minimization of a certain criterion, such as the AIC and the BIC with some improvements as in Chen and Chen (2008). These methods work well for selecting a useful model (Wang et al., 2009), but the theoretical results hold only when $p/n \to 0$ and not for high-dimensional designs (see Homrighausen and McDonald, 2018). Another range of methods is based on resampling techniques such as cross-validation which are risk-consistent in high-dimensional settings (see Homrighausen and McDonald, 2017). Nowadays, there is still no consensus on the method to use even if some methods are particularly popular. The choice of a regularization parameter can be seen as a model selection problem since it dictates which variables are retained in the model. In this paper, we propose an averaging procedure that avoids the choice of λ. Moreover, the proposed method aims to reduce the prediction risk associated with the proposal model averaging estimator.

To provide a motivating example, Table 1 shows the performance of the proposed Model Averaging debiased estimator (MA-d), presented thoroughly in Section 3, for a toy example. We compare MA-d with several commonly used methods for choosing the regularization parameter and we inspect the out-of-sample squared prediction error loss

$$L_n(\hat{\beta}) := \frac{1}{n} \left|\left| \mathbf{X}\beta_0 - \mathbf{X}\hat{\beta} \right|\right|_2^2. \tag{5}$$

The table clearly shows that the different methods for λ selection can provide very different results, and that the proposed procedure, MA-d, that optimally averages the estimates across the entire path of λ values, performs better than any competitor.

Methods	**MA-d**	Debiased Lasso with				
		$\lambda_{\text{CV}-1\text{se}}$	λ_{AIC}	λ_{BIC}	λ_{CV}	λ_{LOO}
L_n	**1.88**	4.51	3.26	3.77	3.03	2.82
std.	**0.38**	1.38	0.89	1.08	0.81	0.82

Table 1: Out-of-sample mean squared prediction error loss and its standard deviation for the MA-d and the debiased Lasso estimator with different values of the penalization parameter. The loss is averaged over 1000 replicates for $n = 100$ and $p = 200$. The sequence of parameters is $\beta_0 = (2, \dots, 2, 0, \dots, 0)^T$ with $s_0 = 6$ non-zero entries, the design matrix is Gaussian with a Toeplitz covariance structure and the noise level is $\sigma^2 = 1.5$. MA-d denotes the proposed method, CV-1se is the 'one standard error' rule (Chen and Yang, 2021), AIC is the λ value that minimizes the AIC criterion, BIC is the value that minimizes the BIC, CV is a choice based on 5-fold cross-validation, LOO stands for a choice by leave-one-out cross-validation.

3 Proposed procedure

For low-dimensional models, the regularization parameter for the Lasso can take any value in $\mathbb{R}^+$. However, there is a point, $\lambda_{\max}$, such that $\forall \lambda > \lambda_{\max}$, $\hat{\beta}_\lambda^{Lasso}$ is estimated at zero for each entry. Note that $\lambda_{\max}$ depends on the design matrix, as well as on the response vector and it can be determined numerically. In the high-dimensional case, there is moreover a minimal positive value, $\lambda_{\min}$, such that the number of estimated active coefficients by $\hat{\beta}_{\lambda_{\min}}^{Lasso}$ is $\min(n, p)$ and problem (2) has no solution for any $\lambda < \lambda_{\min}$ (Tibshirani, 2013). In practical applications, a

choice for the regularization parameter is always taken inside the interval $[\lambda_{\min}, \lambda_{\max}]$ and these bounds are obtained empirically.

Let M be a positive integer independent of p and n and let $\lambda_1, \ldots, \lambda_M$ be a finite sequence of regularization parameters in $[\lambda_{\min}, \lambda_{\max}]$. Therefore, for each λ_m with $m = 1 \ldots, M$, one can obtain a solution for optimization problem (2). This defines a finite sequence of estimated parameters $\hat{\beta}_1^{Lasso}, \ldots, \hat{\beta}_M^{Lasso}$. Using further Equation (3), we construct a sequence of M debiased Lasso estimators $\hat{\beta}_1^d, \ldots, \hat{\beta}_M^d$. Finally, the model averaging estimator is constructed as

$$\hat{\beta}^{MA} := \sum_{m=1}^{M} w_m \hat{\beta}_m^d,$$

where $\hat{\beta}_m^d := \hat{\beta}_{\lambda_m}^d$ is the debiased Lasso estimator for model m and $\mathbf{w} = (w_1, \ldots, w_M)^T$ is a vector of weights satisfying that $\mathbf{w} \in [0,1]^M$ and $\sum_{m=1}^{M} w_m = 1$. Figure 1 summarizes these steps.

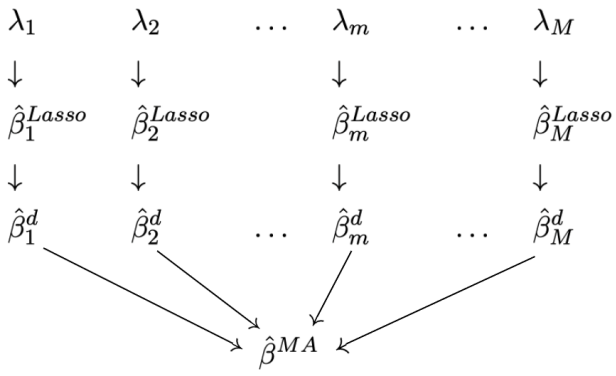

Figure 1: Illustration of the construction of the model averaging estimator: from a grid of penalization parameter values, we estimate first the Lasso, then construct the debiased Lasso and finally the model averaging estimator.

Remark 1. In Equation 3, we use the node-wise method from van de Geer et al. (2014) to obtain the matrix $\hat{\Omega}$ as it is the most used method and as there is a dedicated R package for it (Dezeure et al., 2015). With this method, the matrix $\hat{\Omega}$ is constructed only from $\mathbf{X}$ and is independent of the choice of the penalization parameter. Thus, it can be constructed once and be reused to estimate the debiased Lasso for different values of the penalization parameter.

Remark 2. In practice, a log-scale is used for the construction of the grid of candidate values for λ as the coefficients vary more for small values. On the other hand, the coarseness of the grid empirically seems not to affect the results too much. We explore this point with a simulation study in the Supplementary Material, where we recommend a moderate value such as $M = 10\sqrt{n}$.

3.1 Choice of the weight vector

The choice of the weight vector $\mathbf{w}$ is crucial to ensure good properties of the estimator. Hansen (2014) has shown that for low-dimensional model averaging based on the OLS estimator, the weight that minimizes Mallow's Cp leads to the best prediction risk for the model averaging estimator in the case of homoscedastic errors. The Mallow's Cp criterion is not directly defined

for high-dimensional regression, but following the same idea of minimization of an empirical loss with a penalization term, we use the following quantity to derive the weights:

$$\begin{aligned} \hat{\mathbf{w}} &:= \arg\min_{\mathbf{w}\in\mathcal{H}} C_n(\mathbf{w}), \qquad \mathcal{H} := \left\{ \mathbf{w} : \mathbf{w} \in [0,1]^M; \sum_{m=1}^{M} w_m = 1 \right\}, \\ C_n(\mathbf{w}) &:= \frac{1}{n} \left|\left| Y - \mathbf{X} \left(\sum_{m=1}^{M} w_m \hat{\beta}_m^d \right) \right|\right|_2^2 + \frac{2\hat{\sigma}^2}{n} \sum_{m=1}^{M} w_m \hat{s}_m, \end{aligned} \tag{6}$$

where $\hat{s}_m$ is the number of active coefficients estimated by the Lasso for model m, i.e. $\hat{s}_m = |\hat{S}_m| = |\{\hat{\beta}_{m,j}^{Lasso} : \hat{\beta}_{m,j}^{Lasso} \neq 0\}|$ and $\hat{\sigma}^2$ is an estimator for σ^2. The set $\mathcal{H}$ is the convex set of all weight vectors for which the components are between 0 and 1 and which sum to 1; this set is fixed and does not depend on n, nor on p.

The penalization term in the function $C_n(\cdot)$ is included to avoid overfitting and to favor models that are not too complex. Indeed, for two models with the same squared prediction error, the one with the smaller number of estimated active variables by the Lasso is privileged. The number of estimated active coefficients is an unbiased estimator for the degrees of freedom of the Lasso regression (Zou et al., 2007), so this criterion can be seen as an extension of the Mallows' Cp criterion for high-dimensional models, and so optimization problem (6) is very similar in spirit to the one used in Hansen (2014). Moreover, the $C_n(\cdot)$ function has already shown good empirical performance in Feng and Liu (2020) for model averaging based on the Lasso only, but no statistical guarantees, nor distributional results were provided.

Optimization problem (6) can be re-written as

$$\min_{\mathbf{w}\in\mathcal{H}} \frac{1}{2n} \mathbf{w}^T \bar{E}^T \bar{E} \mathbf{w} + \frac{\hat{\sigma}^2}{n} K^T \mathbf{w}, \tag{7}$$

where $\mathbf{w} = (w_1, \ldots, w_M)^T$ is the $M \times 1$ vector of weights, $K = (\hat{s}_1, \ldots, \hat{s}_M)^T$ is an $M \times 1$ vector and $\bar{E} = (e_1, \ldots, e_M)$ is an $n \times M$ matrix containing the residual vectors of each model, $e_m = Y - \mathbf{X}\hat{\beta}_m^d$, for all $m \in \{1, \ldots, M\}$.

Remark 3. The optimization program in (7) is a quadratic program, so the solution is unique if and only if the matrix $(1/n)\bar{E}^T\bar{E}$ is positive definite. This implies that the matrix $\bar{E}$ needs to be full rank which is not to be expected in practice. In fact, the residuals of model k may be highly correlated with the residuals of model $(k+1)$. However, even if the weights are not necessarily unique, the fitted values *are* unique as shown in Proposition 1. This result implies that a sufficient condition for the uniqueness of the MA-d estimator is to have an invertible design matrix, and this even if the weight vector is not unique.

Proposition 1. *For any fixed sequence $\lambda_1, \lambda_2, \ldots, \lambda_M$, any Y and any matrix $\mathbf{X}$ and for any estimator $\hat{\sigma}^2$ of σ^2, the optimization problem (6) is such that*

1. *there exists either one solution $\hat{\mathbf{w}}$ or an infinite number of solutions,*
2. *the fitted values $\mathbf{X}\hat{\beta}^{MA}(\hat{\mathbf{w}})$ associated to the solutions are unique,*
3. *if $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ are two different solutions, then $\sum_{m=1}^{M} \hat{s}_m \hat{w}_{m,1} = \sum_{m=1}^{M} \hat{s}_m \hat{w}_{m,2}$.*

Proof. 1. The support $\mathcal{H}$ is convex and closed while the function $C_n(\mathbf{w})$ is convex in $\mathbf{w}$. So, there exists at least one solution to the problem. Then if $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ are two different solutions and $\alpha \in (0,1)$, we have that $\alpha\hat{\mathbf{w}}_1 + (1-\alpha)\hat{\mathbf{w}}_2$ is also a solution, since the solution set of a convex problem is convex. This gives us an infinite number of possible solutions.

2. Let $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ be two different solutions to the problem. We denote by $\hat{\beta}_1^{MA}$ and $\hat{\beta}_2^{MA}$ the estimators associated. By contradiction, we assume that $\mathbf{X}\hat{\beta}_1^{MA} \neq \mathbf{X}\hat{\beta}_2^{MA}$ and we denote by c the minimal value of the optimization program, thus $C_n(\hat{\mathbf{w}}_1) = c = C_n(\hat{\mathbf{w}}_2)$.

   Let α be a constant in $(0, 1)$. Then, the vector $\tilde{\mathbf{w}} \coloneqq \alpha\hat{\mathbf{w}}_1 + (1-\alpha)\hat{\mathbf{w}}_2$ is also a solution as argued above. By this, we have:

$$\begin{aligned}
C_n(\tilde{\mathbf{w}}) &= \frac{1}{n}\left|\left|Y - \mathbf{X}\left(\sum_{m=1}^{M}(\alpha\hat{\mathbf{w}}_1 + (1-\alpha)\hat{\mathbf{w}}_2)_m\hat{\beta}_m^d\right)\right|\right|_2^2 + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}(\alpha\hat{\mathbf{w}}_1 + (1-\alpha)\hat{\mathbf{w}}_2)_m\hat{s}_m \\
&= \frac{1}{n}\left|\left|\alpha(Y - \mathbf{X}\hat{\beta}_1^{MA}) + (1-\alpha)(Y - \mathbf{X}\hat{\beta}_2^{MA})\right|\right|_2^2 + \alpha\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{w}_{1,m}\hat{s}_m + (1-\alpha)\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{w}_{2,m}\hat{s}_m \\
&< \alpha\left(\frac{1}{n}\left|\left|Y - \mathbf{X}\hat{\beta}_1^{MA}\right|\right|_2^2 + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{w}_{1,m}\hat{s}_m\right) + (1-\alpha)\left(\frac{1}{n}\left|\left|Y - \mathbf{X}\hat{\beta}_2^{MA}\right|\right|_2^2 + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{w}_{2,m}\hat{s}_m\right) \\
&= \alpha c + (1-\alpha)c \\
&= c,
\end{aligned}$$

   where the strict inequality comes from the strict convexity of the ℓ_2 squared norm after applying the triangle inequality. This is a contradiction with $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ both being optimal solutions. As such, the fitted values must be equal.

3. Let $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ be two different solutions to the problem. As shown earlier, the fitted values $\mathbf{X}\hat{\beta}_1^{MA}$ and $\mathbf{X}\hat{\beta}_2^{MA}$ are equal. Then

$$\frac{1}{n}\left|\left|Y - \mathbf{X}\hat{\beta}_1^{MA}\right|\right|_2^2 = \frac{1}{n}\left|\left|Y - \mathbf{X}\hat{\beta}_2^{MA}\right|\right|_2^2$$

   and since $C_n(\hat{\mathbf{w}}_1) = C_n(\hat{\mathbf{w}}_2)$, we must have that $\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{s}_m\hat{w}_{m,1} = \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}\hat{s}_m\hat{w}_{m,2}$ and thus $\sum_{m=1}^{M}\hat{s}_m\hat{w}_{m,1} = \sum_{m=1}^{M}\hat{s}_m\hat{w}_{m,2}$.

□

4 Theoretical properties

We present here novel theoretical results on the MA-d estimator. The two main results focus on the asymptotic normality of the estimator and the loss reduction. All the rates presented in this paper are valid for an increasing n and for p being a function of n. We also rely on the asymptotic normality of the debiased Lasso, hence the first two assumptions below are from van de Geer et al. (2014). The regression model is defined in Equation (1) and the rows of the design matrix are assumed to be drawn randomly from a $\mathcal{N}(0, \Sigma)$ distribution where 0 is the $p \times 1$ zero vector.

Assumption 0. *The sample size, n, and the number of parameters, p, are allowed to grow simultaneously, satisfying $(\log p)/n = o(1)$.*

Assumption 1. *The rows of $\mathbf{X}$ are i.i.d. realizations of a Gaussian distribution whose p-dimensional inner product matrix Σ has a strictly positive smallest eigenvalue $\Lambda_{\min}(\Sigma)$ satisfying $1/\{\Lambda_{\min}(\Sigma)\} = O(1)$. Furthermore for all p, $\max_{j=1,\ldots,p}\Sigma_{j,j} = O(1)$.*

Assumption 2. *The sparsity index for the coefficient vector from model (1), s_0, satisfies $s_0 = o\left(\sqrt{n}/\log p\right)$ and $\max_{j=1,\dots,p} s_j = o\left(n/\log p\right)$, where $s_j \coloneqq |\{\Omega_{j,k} : j \neq k \text{ and } \Omega_{j,k} \neq 0\}|$ denotes the sparsity in the rows of the precision matrix, $\Omega \coloneqq \Sigma^{-1}$.*

Assumption 3. *The design matrix $\mathbf{X}$ satisfies $||\mathbf{X}||_\infty = O_{\mathbb{P}}(\sqrt{n})$, where $||\cdot||_\infty$ denotes the entry-wise sup norm of the matrix.*

Assumption 1 concerns the variance-covariance population matrix Σ. In particular, it implies the compatibility condition required for Lasso convergence (Bühlmann and van de Geer, 2011, Chap. 6). Assumption 2 concerns the sparsity of the coefficient vector β_0 and of the Ω matrix. As studied in Javanmard and Montanari (2018), the sparsity assumption for the debiased Lasso is stricter than the usual one for Lasso, which is $o\left(\sqrt{n/\log p}\right)$. Assumption 3 controls the boundedness of the entries in the design matrix and it is needed to ensure the convergence of the prediction error in Theorem 2. A sufficient condition for this assumption is that $p = O(n^2)$. Indeed, by Theorem 4.4.5 in Vershynin (2018), we have $||\mathbf{X}||_\infty \leq ||\mathbf{X}||_{spect} = \sqrt{\Lambda_{\max}(\mathbf{X}^T\mathbf{X})} = O_{\mathbb{P}}(p^{1/4})$, where $||\mathbf{X}||_{spect}$ is the spectral norm.

Our first result focuses on the asymptotic normality of the proposal estimator. The result is not trivial, as usually model averaging estimators have an unknown and non-standard distribution even if all the estimators are Gaussian due to the randomness of the weight vector.

Theorem 1. *Under Assumptions 0 – 2, for the linear model with Gaussian error term $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ where $\sigma^2 = O(1)$ and whose $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_M)^T$ denotes the weights selected by (6) with $\hat{\sigma}^2$ as estimator for the noise level, if for every model the penalization parameter satisfies $\lambda_m = O\left(\sqrt{(\log p)/n}\right)$, we have:*

$$\begin{aligned}
\sqrt{n}(\hat{\beta}^{MA} - \beta_0) &= W + \sum_{m=1}^{M} \hat{w}_m \Delta_m, \\
W|\mathbf{X} &\sim \mathcal{N}(0, \sigma^2 \hat{\Omega}\hat{\Sigma}\hat{\Omega}^T), \\
\left|\left|\sum_{m=1}^{M} \hat{w}_m \Delta_m\right|\right|_\infty &= o_{\mathbb{P}}(1),
\end{aligned}$$

where $W \coloneqq \hat{\Omega}\mathbf{X}^T\epsilon/\sqrt{n}$ and $\Delta_m \coloneqq \sqrt{n}(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)$ is a bias term.

Theorem 1 shows that the MA-d estimator can be decomposed in two parts. The first one does not depend on the weights and is asymptotically Gaussian and the second part is a bias term whose sup norm converges in probability to zero. The asymptotic normality holds when n and p both grow.

Proof. By construction of the debiased estimator, we have

$$\sqrt{n}(\hat{\beta}^{MA} - \beta_0) = \sum_{m=1}^{M} \hat{w}_m \sqrt{n}(\hat{\beta}_m^d - \beta_0) = \sum_{m=1}^{M} \hat{w}_m (W + \Delta_m) = W + \sum_{m=1}^{M} \hat{w}_m \Delta_m,$$

where, as in the proof of Theorem 2.2 of van de Geer et al. (2014), $W = \hat{\Omega}\mathbf{X}^T\epsilon/\sqrt{n}$ is Gaussian when we condition on the design matrix $\mathbf{X}$ and does not depend on the penalization parameter.

Under Assumption 1, for the bias term $\Delta_m \coloneqq \sqrt{n}(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)$, we have

$$||\Delta_m||_\infty \leq \sqrt{n}||I_p - \hat{\Omega}\hat{\Sigma}||_\infty ||\hat{\beta}_m^{Lasso} - \beta_0||_1 = O_{\mathbb{P}}\left(s_0 \frac{\log p}{\sqrt{n}}\right),$$

where $||\Delta_m||_\infty = \max_{j=1,\dots,p} |\Delta_{m,j}|$ denotes the sup norm of the vector and where $||I_p - \hat{\Omega}\hat{\Sigma}||_\infty = O_{\mathbb{P}}(\sqrt{(\log p)/n})$ is shown in Theorem 2.2 from van de Geer et al. (2014). Since $\hat{\mathbf{w}} \in [0,1]^M$ for all n and p, and as the number of considered models M is fixed, we also have that

$$\left|\left|\sum_{m=1}^{M} \hat{w}_m \Delta_m\right|\right|_\infty \leq \sum_{m=1}^{M} ||\hat{w}_m \Delta_m||_\infty \leq \sum_{m=1}^{M} ||\Delta_m||_\infty = O_{\mathbb{P}}\left(s_0 \frac{\log p}{\sqrt{n}}\right).$$

Moreover, under Assumption 2, which controls the growth rate of the sparsity index s_0, we have

$$\left|\left|\sum_{m=1}^{M} \hat{w}_m \Delta_m\right|\right|_\infty = O_{\mathbb{P}}\left(s_0 \frac{\log p}{\sqrt{n}}\right) = o_{\mathbb{P}}(1).$$

□

Due to its construction, the decomposition of the MA-d estimator in Theorem 1 shows that crucially, the weights play a role *only* in the second term. This gives us the result of Proposition 2 below which determines the rate of convergence for the estimated coefficients.

For the next results, we need to control the noise level of the statistical problem and more precisely, the term $\mathbf{X}^T\epsilon/n$. For this purpose, we work on the event $\xi_\infty := \left\{\left|\left|\mathbf{X}^T\epsilon/n\right|\right|_\infty \leq \lambda_0\right\}$, with $\lambda_0 = O\left(\sqrt{(\log p)/n}\right)$ and by Lemma 6.2 from Bühlmann and van de Geer (2011), directly applicable to our setting, the probability of the event tends to one for increasing samples sizes.

Proposition 2. *Under Assumptions 0 – 2, for the linear model with Gaussian error term $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ where $\sigma^2 = O(1)$, if for every model $m \in \{1,\dots,M\}$, the penalization parameter satisfies $\lambda_m = O\left(\sqrt{(\log p)/n}\right)$, we have*

$$\left|\left|\hat{\beta}^{MA}(\hat{\mathbf{w}}) - \beta_0\right|\right|_\infty = O_{\mathbb{P}}\left(\max_{j=1,\dots,p} \sqrt{s_j}\sqrt{\frac{\log p}{n}} + s_0\frac{\log p}{n}\right) = o_{\mathbb{P}}(1),$$

where $\hat{\beta}^{MA}(\hat{\mathbf{w}})$ denotes the model averaging estimator with weights selected by (6) with $\hat{\sigma}^2$ as estimator for the noise level.

The proof is provided in the Supplementary Material. The main idea is to decompose the MA-d estimator and derive a bound for each term.

These two results show that, unlike many model averaging estimators, the proposed estimator is asymptotically Gaussian and maintains the same rate of convergence in sup-norm *even after* the averaging step. In addition to these interesting properties, the following theorem shows that the weights obtained by (6) lead to the asymptotically best loss compared to *any* other weights. In particular, the MA-d is no worse than any individual debiased model.

Theorem 2. *Under Assumptions 0 – 3, for the linear model with Gaussian error term $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, let $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_M)^T$ denote the weights selected by (6) with $\hat{\sigma}^2$ as estimator for the noise level. If for every model the penalization parameter satisfies $\lambda_m = O\left(\sqrt{(\log p)/n}\right)$ and if $\hat{\sigma}^2$ satisfies $|(\hat{\sigma}^2/\sigma^2) - 1| = o_{\mathbb{P}}(1)$ we have:*

$$\frac{L_n(\hat{\beta}^{MA}(\hat{\mathbf{w}}))}{\inf_{\mathbf{w}\in\mathcal{H}} L_n(\hat{\beta}^{MA}(\mathbf{w}))} \to 1 \text{ in probability,}$$

where $L_n(\cdot)$ denotes the quadratic loss function defined in Equation (5).

The proof of Theorem 2 can be found in the Supplementary Material, here we just explain briefly the main idea. We begin by showing that if the weights minimize $C_n(\mathbf{w})$ by definition of optimization problem (6), they also minimize a function $\tilde{C}_n(\mathbf{w})$ defined as $\tilde{C}_n(\mathbf{w}) = L_n(\hat{\beta}^{MA}(\mathbf{w})) + R(\hat{\beta}^{MA}(\mathbf{w}))$, where $R(\hat{\beta}^{MA}(\mathbf{w}))$ is a remainder term to be defined in the proof. Afterward, we explore under which conditions these two terms converge to zero, and we show that the convergence rate of $L_n(\hat{\beta}^{MA}(\mathbf{w}))$ is slower than the one of $R(\hat{\beta}^{MA}(\mathbf{w}))$. The closing argument is then: asymptotically choosing the weights by minimizing $C_n(\mathbf{w})$ is equivalent to choosing the weights by minimizing $L_n(\cdot)$.

The next proposition extends the result of Theorem 1 to the case where the number of models, M, increases with n. For this, one needs to control all the regularization parameters in the grid for each n and M. Interestingly, our proposition shows that the growth rate of M does not affect the normality result. The proof can be found in the Supplementary Material.

Proposition 3. *Under Assumption 0 – 2, for the linear model with Gaussian error term $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ where $\sigma^2 = O(1)$, let $M_n = M(n)$ be the number of models and $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_{M_n})^T$ the weights selected by (6) with $\hat{\sigma}^2$ as estimator for the noise level. If for every model and every sample size n, we have $\lambda_m \propto \sqrt{(\log p)/n}$ and $\sup_{m \leq M_n} \lambda_m < K\sqrt{(\log p)/n}$ with $K < \infty$ for all M_n, then*

$$\sqrt{n}(\hat{\beta}^{MA} - \beta_0) = W + \sum_{m=1}^{M_n} \hat{w}_m \Delta_m,$$

$$W|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 \hat{\Omega}\hat{\Sigma}\hat{\Omega}^T),$$

$$\left|\left| \sum_{m=1}^{M_n} \hat{w}_m \Delta_m \right|\right|_\infty = o_{\mathbb{P}}(1),$$

where $W := \hat{\Omega}\mathbf{X}^T\epsilon/\sqrt{n}$ and $\Delta_m := \sqrt{n}(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)$ is a bias term.

5 Simulation

We illustrate the performance of the proposed method with a numerical study. There are two main results we expect to verify with this study: (i) the optimality of the loss as shown in Theorem 2 and (ii) the asymptotic normality of the estimator as shown in Theorem 1.

We work under model (1) where the rows of the design matrix $\mathbf{X}$ are i.i.d. Gaussian, having a correlation matrix Σ that follows a Toeplitz structure, $\Sigma_{ab} = \rho^{|a-b|}$ with $\rho = 0.5$. The vector of coefficients is set to $\beta_0 = (2, \ldots, 2, 0, \ldots, 0)^T$ with the number of non-null entries, the sparsity index s_0, set at $10\, n^{1/4}/\log p$ rounded to the nearest integer. The sample size is $n \in \{100, 200, 300, 500, 750, 1000\}$ and the number of covariates is $p = 2n$, so that each design corresponds to a high-dimensional setting. For the noise level σ_0^2, we allow it to vary in the set $\{0.75, 1.5, 3\}$.

The MA-d estimator is constructed with an estimated value for the noise level in the weights optimization problem (6). Here the estimator $\hat{\sigma}^2_{\text{scaled}}$ from Sun and Zhang (2012) is used. It is a consistent estimator in high-dimensional settings and it is obtained as

$$\left(\hat{\beta}_{\text{scaled}}, \hat{\sigma}_{\text{scaled}}\right) := \arg\min_{\beta,\sigma} \left(\frac{||Y - \mathbf{X}\beta||_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 ||\beta||_1 \right),$$

where $\lambda_0 \approx \sqrt{(2/n)\log p}$ is chosen using the quantile method defined and studied in Sun and Zhang (2013). In practice, we use the `scalreg` R package with the default parameters to obtain the estimator.

The number of models is set to $M = 200$. The competitors are the debiased Lasso, the Ridge, and the Lasso with λ obtained by 10-fold cross-validation. For all lasso-based methods, the regularization parameter is chosen from the same path of M values. For the construction of the inference metrics, $\hat{\sigma}^2_{\text{scaled}}$ is used for all the competitors that require an estimate of the noise level. Additional simulations comparing different estimators for σ^2 are shown in the Supplementary Material, but they all point roughly to similar conclusions so we skip presenting the results here.

Table 2 shows, averaged over $R = 1000$ repetitions, the following loss ratio

$$\text{LR} := \frac{1}{R}\sum_{r=1}^{R}\left(\frac{L_n\left(\hat{\beta}^{(r)}\right)}{\min_{\mathbf{w}\in\mathcal{H}} L_n\left(\hat{\beta}^{MA,(r)}(\mathbf{w})\right)}\right). \tag{8}$$

The denominator is the minimum loss function for model averaging with oracle weights. These weights are the ones that directly minimize $L_n(\hat{\beta}^{MA}(\mathbf{w}))$, which is known in this simulation setting. The numerator is the loss function for the estimator at the rth iteration. This ratio allows us to evaluate if the weights obtained by minimizing (6) give a similar loss as the oracle ones.

		Prediction loss ratio				Inference metrics *(nominal level: .95)*			
σ_0^2	n	MA-d	Debiased Lasso	Ridge	Lasso	MA-d		Debiased Lasso	
						Active	Non-active	Active	Non-active
0.75	100	1.08	2.81	16.88	0.11	.87	.98	.85	.95
	200	1.01	2.19	12.42	0.05	.89	.98	.87	.95
	300	1.00	1.78	11.26	0.03	.90	.98	.88	.95
	500	1.00	1.43	8.93	0.01	.92	.97	.90	.95
	750	1.00	1.24	7.58	0.01	.93	.97	.91	.95
	1000	1.00	1.13	6.86	0.01	.93	.96	.91	.95
1.5	100	1.23	3.18	10.69	0.14	.86	.98	.86	.96
	200	1.08	2.78	8.31	0.06	.88	.98	.89	.96
	300	1.02	2.34	7.72	0.04	.89	.98	.89	.96
	500	1.00	1.91	6.11	0.02	.91	.98	.91	.96
	750	1.00	1.63	5.07	0.01	.92	.98	.92	.95
	1000	1.00	1.45	4.45	0.01	.92	.97	.92	.95
3	100	1.41	3.56	6.33	0.17	.85	.98	.87	.96
	200	1.27	3.43	5.23	0.08	.88	.98	.90	.96
	300	1.13	3.03	5.08	0.05	.88	.98	.90	.96
	500	1.02	2.57	4.15	0.03	.89	.98	.91	.96
	750	1.00	2.23	3.47	0.02	.91	.98	.92	.96
	1000	1.00	1.98	3.03	0.01	.91	.98	.92	.95

Table 2: Prediction loss ratio (LR) averaged over $R = 1000$ repetitions for estimators constructed with a path of $M = 200$ values and average coverage for active and non-active variables.

From Table 2, left-hand-side panel, we conclude that the Lasso is the only method with a ratio below 1, which was expected since it is the only sparse method and the true model is sparse. Among the non-sparse methods, we observe that for each sample size and noise level, the MA-d has the smallest ratio. As the noise level increases, the ratio increases as well, and as the sample size increases, the ratio decreases to unity as expected. The ratio of the competitors

also converges to unity, but at a much slower rate. The fast decay for MA-d shows that in practice the weights obtained by minimizing (6) are very close to the optimal ones, pointing to substantial gains for finite sample analyzes.

Another point of interest is the asymptotic distribution of the estimator. The right-hand side panel in Table 2 shows the average coverage for active and non-active variables. The target coverage is 95%. The only competitor is the debiased Lasso estimator, since it is the only other considered competitor with a known distribution in the high-dimensional setting. The empirical coverage is calculated as

$$\text{Cvg}(\hat{\beta}_j) := \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\left\{\beta_{0,j} \in \widehat{CI}_{1-\alpha}\left(\hat{\beta}_j^{(r)}\right)\right\},$$

$$\text{avgCvg}_a(\hat{\beta}) := \frac{1}{s}\sum_{j=1}^{s}\text{Cvg}(\hat{\beta}_j), \qquad \text{avgCvg}_{na}(\hat{\beta}) := \frac{1}{p-s}\sum_{j=s+1}^{p}\text{Cvg}(\hat{\beta}_j),$$

where $\mathbf{1}\{A\} = 1$ if A takes place, is the indicator function and $\widehat{CI}_{1-\alpha}(\hat{\beta}_j^{(r)})$ is the confidence interval at level $1-\alpha$ for the jth coefficient at iteration r. The confidence interval is computed with the asymptotic variance presented in Theorem 1, which is, for $\hat{\beta}_j$, $\hat{\sigma}^2(\hat{\Omega}\hat{\Sigma}\hat{\Omega})_{j,j}$. For both estimators, the noise level is estimated by $\hat{\sigma}^2_{\text{scaled}}$.

Table 2 shows that the performance of the MA-d and of the debiased Lasso estimator, with respect to the empirical coverage, are comparable. The coverage for both methods is close and gets closer to the target as the sample size increases as expected. The coverage is slightly lower for the active variables and slightly higher for the non-active variables and this phenomenon occurs for both methods. When the noise level is high, the coverage performance of MA-d is slightly worse. Since the standard error of the model averaging estimator does not depend on the weights, the lengths of the confidence intervals of the two methods are the same and as such, their values are not presented here.

We show next in Figure 2 the distribution of the MA-d estimator, for a randomly chosen active and a non-active standardized coefficient for different values of n. The standardization was obtained by

$$\frac{\hat{\beta}_j^{MA} - \beta_{0,j}}{\text{sd}(\hat{\beta}_j^{MA})} = \frac{\hat{\beta}_j^{MA} - \beta_{0,j}}{\sqrt{\hat{\sigma}^2_{\text{scaled}}(\hat{\Omega}\hat{\Sigma}\hat{\Omega})_{j,j}}}.$$

Figure 2 complements well the results for coverage. It illustrates that the distribution for the non-active coefficients has slightly larger variance for small sample size, but it approaches the Gaussian limit as n increases.

Table 3 shows accuracy metrics such as the average bias and MSE for active and non-active coefficients for different sample sizes and noise levels. These empirical quantities are calculated as follows:

$$\text{Bias}(\hat{\beta}_j) := \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_j^{(r)}\right) - \beta_{0,j}, \qquad \text{MSE}(\hat{\beta}_j) := \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_j^{(r)} - \beta_{0,j}\right)^2,$$

$$\text{avgBias}(\hat{\beta})_a := \frac{1}{s}\sum_{j=1}^{s}\left|\text{Bias}(\hat{\beta}_j)\right|, \qquad \text{avgMSE}(\hat{\beta})_a := \frac{1}{s}\sum_{j=1}^{s}\text{MSE}(\hat{\beta}_j),$$

$$\text{avgBias}(\hat{\beta})_{na} := \frac{1}{p-s}\sum_{j=s+1}^{p}\left|\text{Bias}(\hat{\beta}_j)\right|, \qquad \text{avgMSE}(\hat{\beta})_{na} := \frac{1}{p-s}\sum_{j=s+1}^{p}\text{MSE}(\hat{\beta}_j).$$

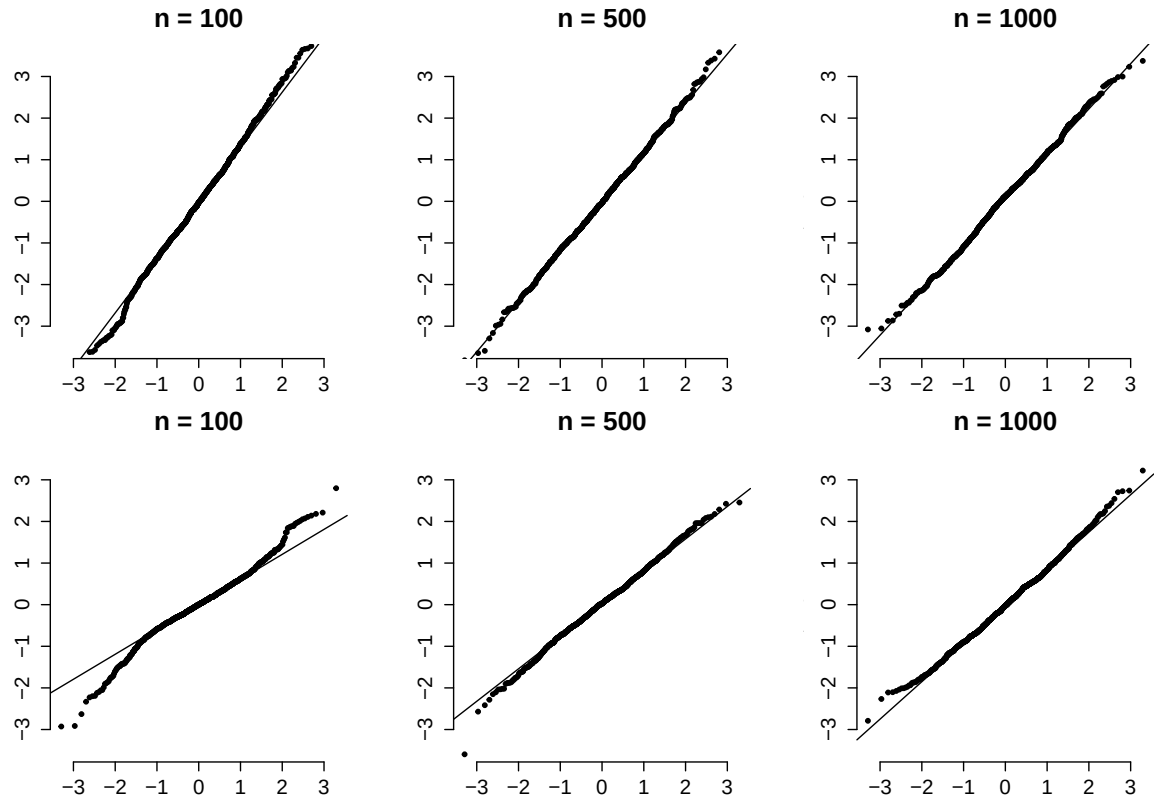

Figure 2: QQplot for an active (top row) and a non-active (bottom row) standardized estimated coefficient by MA-d over 1000 replications.

σ_0^2	n	MA-d				Debiased Lasso				Lasso			
		Active		Non-active		Active		Non-active		Active		Non-active	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0.75	100	1.25	1.64	0.20	0.53	2.37	1.78	0.26	0.93	7.94	2.20	0.05	0.03
	200	0.66	0.74	0.15	0.30	1.94	0.81	0.19	0.48	6.01	1.08	0.02	0.01
	300	0.67	0.46	0.13	0.23	1.62	0.52	0.16	0.33	4.91	0.70	0.01	<0.01
	500	0.61	0.27	0.10	0.16	1.14	0.29	0.11	0.20	4.04	0.44	0.01	<0.01
	750	0.47	0.17	0.09	0.12	0.91	0.19	0.10	0.14	3.49	0.30	<0.01	<0.01
	1000	0.50	0.13	0.08	0.10	0.79	0.14	0.08	0.11	3.13	0.24	<0.01	<0.01
1.5	100	1.90	3.43	0.27	1.02	3.33	3.29	0.36	1.71	11.19	4.41	0.07	0.05
	200	1.00	1.58	0.20	0.53	2.75	1.52	0.27	0.92	8.49	2.16	0.03	0.01
	300	0.77	1.00	0.16	0.38	2.29	0.98	0.22	0.64	6.92	1.40	0.02	<0.01
	500	0.60	0.57	0.13	0.26	1.61	0.56	0.16	0.40	5.71	0.88	0.01	<0.01
	750	0.41	0.36	0.11	0.20	1.27	0.36	0.14	0.27	4.94	0.61	0.01	<0.01
	1000	0.41	0.26	0.10	0.16	1.09	0.27	0.12	0.21	4.42	0.47	<0.01	<0.01
3	100	2.78	6.95	0.38	2.03	4.69	6.33	0.51	3.28	15.82	8.83	0.10	0.11
	200	1.62	3.24	0.28	1.03	3.90	2.95	0.37	1.79	12.01	4.32	0.04	0.02
	300	1.33	2.13	0.21	0.71	3.24	1.91	0.30	1.26	9.79	2.80	0.03	0.01
	500	1.07	1.23	0.17	0.45	2.27	1.10	0.22	0.78	8.07	1.76	0.02	<0.01
	750	0.61	0.77	0.15	0.33	1.79	0.72	0.19	0.54	6.98	1.21	0.01	<0.01
	1000	0.41	0.56	0.13	0.27	1.53	0.54	0.16	0.42	6.24	0.94	0.01	<0.01

Table 3: Average bias and MSE for active and non-active variables over 1000 replications for different sample sizes and noise levels; the values are expressed in unit 10^{-2} for ease of readability.

Both the bias and the MSE decrease with n, but as expected, increase as the noise level increases. The bias on the active variables is slightly lower for the MA-d and both the MA-d, and the debiased Lasso estimator have a much lower bias than the Lasso estimator for the active and non-active variables. The results show empirically that the MA-d estimator retains the bias reduction property of the debiased Lasso.

6 Real data analysis

Bardet-Biedl Syndrome (BBS) is a genetic disease that affects several organs. The main symptoms are obesity and retinal failure, but polydactia and learning disabilities can also occur. The `eyedata` database made available in the R package `flare` (Li et al., 2015) contains the expression of 200 genes for 120 rats affected by BBS, and is derived from Scheetz et al. (2006). In that study, the TRIM32 gene was identified as a disease-causing gene by looking at the correlation with other known BBS genes.

The aim of this section is to identify whether the predictor genes for the TRIM32 gene are BBS genes that could play a role in the disease. To identify genes linked to TRIM32, we perform a regression analysis on the dataset with TRIM32 as response and since there are more covariates than sampling units, the usage of high-dimensional methods is mandatory. On the whole dataset, the Lasso selects 57 coefficients when the regularization parameter is chosen by 10-fold cross-validation and 19 if it is chosen by the '1se' rule. The debiased estimator based on this cross-validation Lasso estimates that 3 variables are significant for a confidence level of 95% and with a Bonferroni correction for multiplicity. The significant genes are genes labelled 10540, 16984 and 17599 if we refer to the notation of the dataset. The MA-d estimator identifies that only 2 coefficients are significant for the same confidence level and multiplicity correction; these are genes 10540 and 17599.

To test the performance of the estimators, we compute the leave-one-out prediction risk (LOO) of the estimator and compare it with the debiased Lasso estimator, the Lasso and the Ridge estimator where the penalization parameter was chosen by 10-fold cross-validation. The noise level is estimated with $\hat{\sigma}_{\text{scaled}}$ when needed. We define

$$\text{LOO}(\hat{\beta}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\hat{\beta}_{-i})^2,$$

where $\hat{\beta}_{-i}$ is the estimator computed from all observations except the ith one.

To obtain a more stable calculation, both the predictors and the response vector have been standardized, and the prediction risk is expressed in terms of these standardized data. Table 4 shows the results of this analysis, and clearly the MA-d estimator performs similarly to the Lasso or the Ridge with a smaller standard deviation. However, the debiased estimator has a very high prediction risk for this example.

	MA-d	Debiased Lasso	Ridge	Lasso
LOO	0.48	3.87	0.40	0.45
std.	0.73	21.52	0.95	1.38

Table 4: Leave-one-out error and its standard deviation on the *eyedata* for the MA-d, Lasso, the debiased Lasso and the Ridge. The penalization parameter was selected by 10-fold cross-validation for the methods that require it.

7 Conclusion

In this paper, we proposed a new model averaging estimator for high-dimensional regression. The estimator is based on the debiased Lasso estimator, with averaging taken along the Lasso solution path, and the weights obtained through optimization of a sample loss. We showed the asymptotic normality of this novel averaging estimator and the optimality of the weights in terms of prediction risk. In practice, the proposed estimator outperformed competitors in terms of prediction loss and the asymptotic normality continues to hold even for the high-dimensional design. Moreover, since the procedure is based on the debiased estimator, our empirical results show that the MA-d estimator still maintains a bias advantage, even after averaging on the λ path.

While in this work we have focused on model averaging applied to the debiased Lasso estimator, it would be interesting to explore whether for other high-dimensional estimators, model averaging can lead to substantial gains in terms of prediction or accuracy. Another open question is the control of the False Discovery Rate (FDR) under model averaging. Linking to the previous idea, it would be interesting to evaluate if, under high-dimensional settings, a model averaging step on the SLOPE estimator (Bogdan et al., 2015) maintains FDR control.

8 Acknowledgement

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

9 Supplementary material for "High-dimensional inference for Model Averaging estimators"

Proofs

Proof of Proposition 2. Using the decomposition in Equation (4), we obtain

$$\begin{aligned}\left|\left|\hat{\beta}^{MA}(\hat{\mathbf{w}})-\beta_0\right|\right|_\infty &= \left|\left|\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon+\sum_{m=1}^{M}\hat{w}_m(I_p-\hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso}-\beta_0)\right|\right|_\infty \\ &\leq \left|\left|\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_\infty+\left|\left|\sum_{m=1}^{M}\hat{w}_m\frac{\Delta_m}{\sqrt{n}}\right|\right|_\infty. \end{aligned} \tag{9}$$

Under Assumption 1, using the same argument as in Theorem 1, as $\hat{w}_m \in [0,1]$ for each $m \in \{1,\dots,M\}$ and M is fixed, the weighted error is bounded at the rate

$$\left|\left|\sum_{m=1}^{M}\hat{w}_m\frac{\Delta_m}{\sqrt{n}}\right|\right|_\infty = O_{\mathbb{P}}\left(s_0\frac{\log p}{n}\right).$$

For the first term in (9), we have

$$\left|\left|\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_\infty \leq ||\hat{\Omega}||_{\infty,\infty}\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty,$$

where $||\hat{\Omega}||_{\infty,\infty} \coloneqq \max_{j=1,\dots,p}||\hat{\Omega}_j||_1$ and using further Lemma 5.3 from van de Geer et al. (2014), we have that $\max_{j=1,\dots,p}||\hat{\Omega}_j||_1 = O_{\mathbb{P}}\left(\max_{j=1,\dots,p}\sqrt{s_j}\right)$. As such, on the event ξ_∞,

$$\left|\left|\hat{\beta}^{MA}(\hat{\mathbf{w}})-\beta_0\right|\right|_\infty = O_{\mathbb{P}}\left(\max_{j=1,\dots,p}\sqrt{s_j}\sqrt{\frac{\log p}{n}}+s_0\frac{\log p}{n}\right)=o_{\mathbb{P}}(1),$$

since $\max_{j=1,\dots,p}\sqrt{s_j} = o\left(\sqrt{n/\log p}\right)$ by Assumption 2. □

Proof of Theorem 2. The proof is structured in three parts. We show first that minimizing $C_n(\mathbf{w})$ is equivalent to minimizing $\tilde{C}_n(\mathbf{w})$. We next derive a bound for $L_n(\hat{\beta}^{MA}(\mathbf{w}))$ and finally, we bound the remainder term, $R(\hat{\beta}^{MA}(\mathbf{w}))$.

Step 1: Using Equation (4) and with $\hat{\mu}^{MA}(\mathbf{w}) = \mathbf{X}\hat{\beta}^{MA}(\mathbf{w})$ and $\mu_0 = \mathbf{X}\beta_0$, we begin by re-writting $C_n(\mathbf{w})$.

$$\begin{aligned}C_n(\mathbf{w}) &= \frac{1}{n}\left|\left|Y-\hat{\mu}^{MA}(\mathbf{w})\right|\right|_2^2+\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}w_m\hat{s}_m \\ &= \frac{1}{n}\left|\left|\mu_0-\hat{\mu}^{MA}(\mathbf{w})\right|\right|_2^2+\frac{1}{n}||\epsilon||_2^2+\frac{2}{n}\epsilon^T\left(\mu_0-\hat{\mu}^{MA}(\mathbf{w})\right)+\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}w_m\hat{s}_m \\ &= L_n(\hat{\beta}^{MA}(\mathbf{w}))+\frac{1}{n}||\epsilon||_2^2+\frac{2}{n}\epsilon^T\mathbf{X}\left(\sum_{m=1}^{M}w_m\left(\beta_0-\hat{\beta}_m^d\right)\right)+\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}w_m\hat{s}_m \\ &= L_n(\hat{\beta}^{MA}(\mathbf{w}))+\frac{1}{n}||\epsilon||_2^2-\frac{2}{n}\epsilon^T\mathbf{X}\left(\sum_{m=1}^{M}w_m\right)\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon-\frac{2}{n}\epsilon^T\mathbf{X}\left(\sum_{m=1}^{M}w_m(I_p-\hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso}-\beta_0)\right)+\frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M}w_m\hat{s}_m\end{aligned}$$

$$= L_n(\hat{\beta}^{MA}(\mathbf{w})) + \frac{1}{n}||\epsilon||_2^2 - \frac{2}{n^2}\epsilon^T\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon - \frac{2}{n}\epsilon^T\mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0) + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M} w_m\hat{s}_m.$$

Isolating the terms that depend on the weights, we have that if the weights minimize $C_n(\mathbf{w})$, they also minimize

$$\tilde{C}_n(\mathbf{w}) = L_n(\hat{\beta}^{MA}(\mathbf{w})) \underbrace{-\frac{2}{n}\epsilon^T\mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0) + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M} w_m\hat{s}_m}_{=R(\hat{\beta}^{MA}(\mathbf{w}))}. \tag{10}$$

The last two terms in (10) correspond to the "bias terms" if we regard $\tilde{C}_n(\mathbf{w})$ as an estimator for the loss $L_n(\hat{\beta}^{MA}(\mathbf{w}))$.

Step 2: We now derive a bound for $L_n(\hat{\beta}^{MA}(\mathbf{w}))$.

First, remark that

$$\begin{aligned}
\hat{\mu}^{MA}(\mathbf{w}) - \mu_0 &= \sum_{m=1}^{M} w_m\mathbf{X}(\hat{\beta}_m^d - \beta_0) \\
&= \sum_{m=1}^{M} w_m\mathbf{X}\left(\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon + (I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)\right) \\
&= \frac{1}{n}\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon + \mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right).
\end{aligned}$$

We can now rewrite $L_n(\hat{\beta}^{MA}(\mathbf{w}))$ as

$$\begin{aligned}
L_n(\hat{\beta}^{MA}(\mathbf{w})) &= \frac{1}{n}\left(\frac{1}{n}\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon + \mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right)^T\left(\frac{1}{n}\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon + \mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right) \\
&= \frac{1}{n^3}\left|\left|\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_2^2 + \frac{2}{n^2}(\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon)^T\mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right) + \frac{1}{n}\left|\left|\mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right|\right|_2^2 \\
&= \underbrace{\frac{1}{n^3}\left|\left|\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_2^2}_{=I_1} + \underbrace{\frac{2}{n}\epsilon^T\mathbf{X}\hat{\Omega}^T\hat{\Sigma}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)}_{=I_2} + \underbrace{\frac{1}{n}\left|\left|\mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^{M} w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right|\right|_2^2}_{=I_3}.
\end{aligned}$$

We develop next the rate of convergence of each term separately.

Using the triangular and the Cauchy–Schwarz inequality $\left|u^Tv\right| \leq ||u||_2||v||_2$ for all $u, v \in \mathbb{R}^p$, we have

$$\begin{aligned}
\frac{1}{n^3}\left|\left|\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_2^2 &= \frac{1}{n^3}\left|\epsilon^T\mathbf{X}\hat{\Omega}^T\mathbf{X}^T\mathbf{X}\hat{\Omega}\mathbf{X}^T\epsilon\right| \\
&= \left|\frac{\epsilon^T\mathbf{X}}{n}\hat{\Omega}^T\hat{\Sigma}\hat{\Omega}\frac{\mathbf{X}^T\epsilon}{n}\right| \\
&= \left|\frac{\epsilon^T\mathbf{X}}{n}\left(\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right)\frac{\mathbf{X}^T\epsilon}{n} + \frac{\epsilon^T\mathbf{X}}{n}\Omega\frac{\mathbf{X}^T\epsilon}{n}\right| \\
&\leq \left|\frac{\epsilon^T\mathbf{X}}{n}\left(\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right)\frac{\mathbf{X}^T\epsilon}{n}\right| + \left|\frac{\epsilon^T\mathbf{X}}{n}\Omega\frac{\mathbf{X}^T\epsilon}{n}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\left(\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right)\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 + \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\Omega\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \\
&\leq \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right|\right|_{\infty,2} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty + \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 ||\Omega||_{\infty,2} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty \\
&\leq \sqrt{p}\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty \left|\left|\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right|\right|_{\infty,2} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty + \sqrt{p}\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty ||\Omega||_{\infty,2} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty,
\end{aligned}
$$

where the inequality on line 6 comes from the consistency of the induced norm $||\cdot||_{\infty,2}$ defined as $||A||_{\infty,2} = \sup_{\mathbf{x}\neq 0} ||A\mathbf{x}||_2/||\mathbf{x}||_\infty$, for a general matrix A. The inequality on line 7 comes from the basic properties of norms where the Euclidean norm is bounded by the sup norm. From Drakakis and Pearlmutter (2009), we have that $||A||_{\infty,2} = \max_{j=1,\dots,p} ||A_j||_2 \leq \sqrt{p}||A||_\infty$ and as such,

$$
I_1 \leq p\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty \left|\left|\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right|\right|_\infty \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty + \sqrt{p}\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty \max_{j=1,\dots,p}\sqrt{s_j}\,||\Omega||_\infty \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty.
$$

We know that $\left|\left|\hat{\Omega}^T\hat{\Sigma}\hat{\Omega} - \Omega\right|\right|_\infty = O_{\mathbb{P}}\left(\max_{j=1,\dots,p}\sqrt{s_j}\sqrt{(\log p)/n}\right)$ by Lemma 5.4 in van de Geer et al. (2014) and so on the event ξ_∞, I_1 is of order

$$
\begin{aligned}
I_1 &= O_{\mathbb{P}}\left(p\frac{\log p}{n}\max_{j=1,\dots,p}\sqrt{s_j}\sqrt{\frac{\log p}{n}} + \sqrt{p}\frac{\log p}{n}\max_{j=1,\dots,p}\sqrt{s_j}||\Omega||_\infty\right) \\
&= O_{\mathbb{P}}\left(p\left(\frac{\log p}{n}\right)^{1.5}\max_{j=1,\dots,p}\sqrt{s_j} + \sqrt{p}\frac{\log p}{n}\max_{j=1,\dots,p}\sqrt{s_j}||\Omega||_\infty\right).
\end{aligned}
$$

For I_2, using again the Cauchy–Schwarz inequality, the triangular inequality, the consistency of the induced norm and the bound on the Euclidean norm, we have that

$$
\begin{aligned}
|I_2| &= \left|\left(\frac{2\epsilon^T\mathbf{X}}{n}(\hat{\Omega}^T\hat{\Sigma} - I_p) + \frac{2\epsilon^T\mathbf{X}}{n}\right)(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^M w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right| \\
&\leq \left|\left|2(\hat{\Sigma}\hat{\Omega} - I_p)\frac{\mathbf{X}^T\epsilon}{n} + \frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|(I_p - \hat{\Omega}\hat{\Sigma})\left(\sum_{m=1}^M w_m(\hat{\beta}_m^{Lasso} - \beta_0)\right)\right|\right|_2 \\
&\leq \left|\left|2(\hat{\Sigma}\hat{\Omega} - I_p)\frac{\mathbf{X}^T\epsilon}{n} + \frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|I_p - \hat{\Omega}\hat{\Sigma}\right|\right|_{\infty,2}\left(\sum_{m=1}^M w_m\left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_\infty\right) \\
&\leq \left(\left|\left|2(\hat{\Sigma}\hat{\Omega} - I_p)\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 + \left|\left|\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2\right)\sqrt{p}\left|\left|I_p - \hat{\Omega}\hat{\Sigma}\right|\right|_\infty\left(\sum_{m=1}^M w_m\left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_\infty\right) \\
&\leq \left(2\left|\left|\hat{\Sigma}\hat{\Omega} - I_p\right|\right|_{\infty,2}\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty + \left|\left|\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2\right)\sqrt{p}\left|\left|I_p - \hat{\Omega}\hat{\Sigma}\right|\right|_\infty\left(\sum_{m=1}^M w_m\left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_\infty\right) \\
&\leq \left(2\sqrt{p}\left|\left|\hat{\Sigma}\hat{\Omega} - I_p\right|\right|_\infty\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty + \sqrt{p}\left|\left|\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_\infty\right)\sqrt{p}\left|\left|I_p - \hat{\Omega}\hat{\Sigma}\right|\right|_\infty\left(\sum_{m=1}^M w_m\left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_\infty\right).
\end{aligned}
$$

For any model $m \in \{1,\dots,M\}$, is it known that $||\hat{\beta}_m^{Lasso} - \beta_0||_\infty = O_{\mathbb{P}}\left(\sqrt{(\log p)/n}\right)$ (Lounici, 2008). On the event ξ_∞, the term I_2 is then of order:

$$
|I_2| = O_{\mathbb{P}}\left(\sqrt{p}\frac{\log p}{n}\max\left\{\sqrt{p}\frac{\log p}{n}, \sqrt{p}\sqrt{\frac{\log p}{n}}\right\}\right) = O_{\mathbb{P}}\left(p\left(\frac{\log p}{n}\right)^{1.5}\right),
$$

since $(\log p)/n = o(1)$.

For I_3 we have

$$I_3 \leq \frac{1}{n} ||\mathbf{X}||^2_{\infty,2} \left|\left|(I_p - \hat{\Omega}\hat{\Sigma}) \sum_{m=1}^{M} w_m (\hat{\beta}_m^{Lasso} - \beta_0)\right|\right|^2_{\infty} \leq ||\mathbf{X}||^2_{\infty} \left|\left|\sum_{m=1}^{M} w_m \frac{\Delta_m}{\sqrt{n}}\right|\right|^2_{\infty}.$$

where, as in Theorem 1, $\Delta_m := \sqrt{n}(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)$. Thus, the term I_3 is of order

$$I_3 = O_{\mathbb{P}}\left(||\mathbf{X}||^2_{\infty} s_0^2 \left(\frac{\log p}{n}\right)^2\right).$$

Step 3: Bound for the remainder term, $R(\hat{\beta}^{MA}(\mathbf{w}))$.
As $\hat{s}_m \leq \min(n, p)$ for all m (Tibshirani, 2013) and as the weights sum to 1,

$$\frac{2\sigma^2}{n} \sum_{m=1}^{M} w_m \hat{s}_m \leq 2\sigma^2 \frac{\max_{m \leq M}(\hat{s}_m)}{n} \leq 2\sigma^2.$$

Thus, as $\hat{\sigma}^2$ is a consistent estimator, we have:

$$\frac{2\hat{\sigma}^2}{n} \sum_{m=1}^{M} w_m \hat{s}_m = \frac{2\left(\hat{\sigma}^2 - \sigma^2\right)}{n} \sum_{m=1}^{M} w_m \hat{s}_m + \frac{2\sigma^2}{n} \sum_{m=1}^{M} w_m \hat{s}_m = O_{\mathbb{P}}(1).$$

We derive next, based on the Cauchy-Schwarz and the triangular inequality, that

$$\begin{aligned}
\left|\frac{2}{n}\epsilon^T \mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma}) \left(\sum_{m=1}^{M} w_m (\hat{\beta}_m^{Lasso} - \beta_0)\right)\right| &\leq \left|\left|2(I_p - \hat{\Sigma}\hat{\Omega}^T)\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\sum_{m=1}^{M} w_m (\hat{\beta}_m^{Lasso} - \beta_0)\right|\right|_2 \\
&\leq 2\left|\left|I_p - \hat{\Sigma}\hat{\Omega}^T\right|\right|_{\infty,2} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_{\infty} \left(\sum_{m=1}^{M} w_m \left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_2\right) \\
&\leq 2\sqrt{p}\left|\left|I_p - \hat{\Sigma}\hat{\Omega}^T\right|\right|_{\infty} \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_{\infty} \left(\sum_{m=1}^{M} w_m \left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_2\right).
\end{aligned}$$

From Bühlmann and van de Geer (2011), we know that $\left|\left|\hat{\beta}_m^{Lasso} - \beta_0\right|\right|_2 = O_{\mathbb{P}}\left(\sqrt{s_0(\log p)/n}\right)$ for all m and as $\left|\left|I_p - \hat{\Sigma}\hat{\Omega}^T\right|\right|_{\infty} = \left|\left|I_p - \hat{\Omega}\hat{\Sigma}\right|\right|_{\infty}$,

$$\left|\frac{2}{n}\epsilon^T \mathbf{X}(I_p - \hat{\Omega}\hat{\Sigma}) \left(\sum_{m=1}^{M} w_m (\hat{\beta}_m^{Lasso} - \beta_0)\right)\right| = O_{\mathbb{P}}\left(\sqrt{p}\left(\frac{\log p}{n}\right)^{1.5}\sqrt{s_0}\right).$$

To conclude the proof, under sparsity Assumption 2 and Assumption 3, on the event ξ_{∞}, the loss is of order

$$L_n(\hat{\beta}^{MA}(\mathbf{w})) = O_{\mathbb{P}}\left(p\frac{\log p}{n} + \sqrt{p}\sqrt{\frac{\log p}{n}}||\Omega||_{\infty} + \frac{||\mathbf{X}||^2_{\infty}}{n}\right) = O_{\mathbb{P}}\left(p\frac{\log p}{n}\right),$$

and since $s_0 = o(\sqrt{n}/\log p)$ by Assumption 2, the bias term in $\tilde{C}_n(\mathbf{w})$ is of order

$$R\left(\hat{\beta}^{MA}(\mathbf{w})\right) = O_{\mathbb{P}}\left(\sqrt{p}\frac{\log p}{n^{5/4}}\right).$$

The control on $||\Omega||_\infty$ is implied by Assumption 1. Indeed, the maximum eigenvalue of the precision matrix is $1/(\Lambda_{\min}(\Sigma))$ which is $O(1)$ by Assumption 1, and for this positive definite matrix, $||\Omega_j||_\infty \leq ||\Omega||_{spect} = \sqrt{\Lambda_{\max}(\Omega^T\Omega)} = \Lambda_{\max}(\Omega) = 1/(\Lambda_{\min}(\Sigma))$ for all $j \in \{1, \ldots, p\}$. We thus have that $||\Omega||_\infty = O(1)$.

As the order of $L_n(\hat{\beta}^{MA}(\mathbf{w}))$ is greater than the order of the bias term by a factor (at least) $\sqrt{p}n^{1/4}$, the loss term is dominant in the choice of weights.

This result suggests that a sufficient condition for convergence of the prediction error is $p(\log p)/n = o(1)$, which implies that the design must be low-dimensional. If Assumption 3 is relaxed, the loss function $L_n(\cdot)$ will only increase, thereby maintaining the convergence of the ratio of the loss functions. □

Proof of Proposition 3. By the same arguments as in the proof of Theorem 1 we have

$$\sqrt{n}(\hat{\beta}^{MA} - \beta_0) = W + \sum_{m=1}^{M_n} \hat{w}_m \Delta_m,$$

and as the weights sum to 1

$$\left|\left|\sum_{m=1}^{M_n} \hat{w}_m \Delta_m\right|\right|_\infty \leq \sum_{m=1}^{M_n} \left(\hat{w}_m \, ||\Delta_m||_\infty\right) \leq \left(\sum_{m=1}^{M_n} \hat{w}_m\right) \sup_{m \leq M_n} ||\Delta_m||_\infty = \sup_{m \leq M_n} ||\Delta_m||_\infty .$$

Let us recall that $\Delta_m = \sqrt{n}(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)$ and that neither $\hat{\Omega}$ nor $\hat{\Sigma}$ depends on m. Thus,

$$\begin{aligned}
\sup_{m \leq M_n} ||\Delta_m||_\infty &= \sup_{m \leq M_n} \left(\sqrt{n}||(I_p - \hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{Lasso} - \beta_0)||_\infty\right) \\
&\leq \sup_{m \leq M_n} \left(\sqrt{n}||I_p - \hat{\Omega}\hat{\Sigma}||_\infty ||\hat{\beta}_m^{Lasso} - \beta_0||_1\right) \\
&= \sqrt{n}||I_p - \hat{\Omega}\hat{\Sigma}||_\infty \sup_{m \leq M_n} ||\hat{\beta}_m^{Lasso} - \beta_0||_1.
\end{aligned}$$

By Assumptions 1 and 2, we have that for each m, $||\hat{\beta}_m^{Lasso} - \beta_0||_1 = O_{\mathbb{P}}(s_0\lambda_m)$ (Bühlmann and van de Geer, 2011, Theorem 6.1). This implies that

$$\begin{aligned}
\sup_{m \leq M_n} ||\Delta_m||_\infty &\leq \sqrt{n}||I_p - \hat{\Omega}\hat{\Sigma}||_\infty \sup_{m \leq M_n} O_{\mathbb{P}}(s_0\lambda_m) \\
&= \sqrt{n}\, O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right) \sup_{m \leq M_n} O_{\mathbb{P}}(s_0\lambda_m) \\
&\leq O_{\mathbb{P}}\left(\sqrt{\log p}\, s_0\, K \sqrt{(\log p)/n}\right)
\end{aligned}$$

where we have used the assumption that $\sup_{m \leq M_n} \lambda_m < K\sqrt{(\log p)/n}$ and Assumptions 1 and 2 for the convergence of $\hat{\Omega}$. As s_0 satisfies $s_0 = o\left(\sqrt{n}/\log p\right)$ and since K is a finite constant, this concludes the proof. □

Additional simulations

Noise level estimation

When the noise level σ_0^2 is unknown, ones needs an estimator in the optimization program for the weights, as well as in the expression of the asymptotic variance of the debiased Lasso coefficients. In this section, we compare the MA-d estimator constructed with different plug-in estimators for σ_0^2 with the estimator constructed with the true value of the noise level.

The first noise level estimator is $\hat{\sigma}^2_{\text{scaled}}$, obtained with the scaled Lasso from Sun and Zhang (2012). This estimator was presented in Section 5 of the main text. Then we construct two estimators from the sum of the squared residuals of the Lasso estimator corresponding to the one with regularization parameter selected by cross-validation. We have two versions of this estimator as we consider the sample variance estimator with a factor of $1/(n - \hat{s}_{m^*})$, as well as a MLE version with a factor of $(1/n)$

$$\hat{\sigma}^2_{\text{cv}} \coloneqq \frac{1}{n - \hat{s}_{m^*}} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \hat{\beta}^{Lasso}_{m^*} \right)^2,$$

$$\hat{\sigma}^2_{\text{cv-mle}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \hat{\beta}^{Lasso}_{m^*} \right)^2,$$

where $\mathbf{X}_i$ is the vector containing all the covariates for individual i, m^* denotes the model selected by cross-validation and $\hat{s}_{m^*}$ is the number of estimated active coefficients in that model.

In low dimensional settings, all the estimators are consistent but the one using the $(1/n)$ factor is biased. In high dimensional settings, only the scaled estimator $\hat{\sigma}^2_{\text{scaled}}$ has been shown to be consistent and this for the following oracle quantity

$$\sigma^2_{\text{oracle}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta_0)^2.$$

The setting for this simulation setup is the same as in Section 5 where the number of covariates is set to $p = 2n$ and for simplicity we only consider the case where $\sigma_0^2 = 1.5$ for generating data. Table 5 shows the average bias and average MSE on active and non-active coefficients, as well as the average coverage and length as defined in the main document. Table 6 shows the ratio of the prediction loss function (LR).

From both tables, we can observe the following behavior: the MA-d seems to be robust to the choice of the noise level estimator, the results for the different methods being close. The bias is the only metric where there is a difference between the methods, but as the sample size increases, the difference quickly disappears. When the noise level is estimated with $\hat{\sigma}^2_{\text{scaled}}$, the coverage is slightly better, but the results of $\hat{\sigma}^2_{\text{cv}}$ are very close. Only $\hat{\sigma}^2_{\text{cv-mle}}$ seems to have a bias significantly higher and a coverage significantly lower when n is small.

n	MA-d with σ_0^2				MA-d with $\hat{\sigma}^2_{\text{scaled}}$				MA-d with $\hat{\sigma}^2_{\text{cv}}$				MA-d with $\hat{\sigma}^2_{\text{cv-mle}}$			
	Accuracy metrics															
	Active		Non-active		Active		Non-active		Active		Non-active		Active		Non-active	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
100	1.92	0.27	3.44	1.02	1.90	0.27	3.43	1.02	1.92	0.27	3.44	1.02	2.07	0.27	3.53	1.00
200	1.02	0.20	1.58	0.53	1.00	0.20	1.58	0.53	1.01	0.20	1.58	0.53	1.06	0.19	1.59	0.53
300	0.77	0.16	1.00	0.38	0.77	0.16	1.00	0.38	0.77	0.16	1.00	0.38	0.77	0.16	1.00	0.38
500	0.60	0.13	0.57	0.26	0.60	0.13	0.57	0.26	0.60	0.13	0.57	0.26	0.60	0.13	0.57	0.26
750	0.41	0.11	0.36	0.20	0.41	0.11	0.36	0.20	0.41	0.11	0.36	0.20	0.41	0.11	0.36	0.20
1000	0.41	0.10	0.26	0.16	0.41	0.10	0.26	0.16	0.41	0.10	0.26	0.16	0.41	0.10	0.26	0.16
	Inference metrics *(nominal level .95)*															
	Active		Non-active		Active		Non-active		Active		Non-active		Active		Non-active	
	Cvg.	Length	Cvg.	Length	Cvg.	Length	Cvg.	Length	Cvg.	Length	Cvg.	Length	Cvg.	Length	Cvg.	Length
100	0.85	0.54	0.98	0.54	0.86	0.54	0.98	0.55	0.84	0.53	0.98	0.53	0.79	0.48	0.96	0.48
200	0.87	0.38	0.98	0.39	0.88	0.39	0.98	0.39	0.87	0.38	0.98	0.39	0.85	0.36	0.97	0.37
300	0.88	0.32	0.98	0.32	0.89	0.32	0.98	0.33	0.88	0.32	0.98	0.32	0.87	0.31	0.98	0.31
500	0.90	0.25	0.98	0.25	0.91	0.25	0.98	0.25	0.90	0.25	0.98	0.25	0.89	0.24	0.98	0.24
750	0.92	0.21	0.98	0.21	0.92	0.21	0.98	0.21	0.92	0.21	0.98	0.21	0.91	0.20	0.97	0.20
1000	0.92	0.18	0.97	0.18	0.92	0.18	0.97	0.18	0.92	0.18	0.97	0.18	0.92	0.18	0.97	0.18

Table 5: Average Bias, MSE, average coverage and length of the 95% confidence interval for active and non-active variables for the MA-d estimator constructed with different estimators for σ_0^2. The Bias and the MSE values are expressed in unit 10^{-2} for ease of readability.

n	MA-d with σ_0^2	MA-d with $\hat{\sigma}^2_{\text{scaled}}$	MA-d with $\hat{\sigma}^2_{\text{cv}}$	MA-d with $\hat{\sigma}^2_{\text{cv-mle}}$
100	1.22	1.23	1.21	1.15
200	1.08	1.08	1.08	1.06
300	1.01	1.02	1.02	1.01
500	1.00	1.00	1.00	1.00
750	1.00	1.00	1.00	1.00
1000	1.00	1.00	1.00	1.00

Table 6: Prediction loss ratio (LR) for the MA-d estimator constructed with different estimators for σ_0^2.

Sensitivity analysis for M

We perform a sensitivity analysis for the choice of M, the number of models, in the optimization program for the model-averaging debiased estimator. We compare the MA-d estimator when M depends on the sample size to assess whether we still have approximate normality, as stated in Proposition 3. In the following simulation, the number of variables is fixed at 500, n varies from 200 (high-dimensional setting) to 2000 (low-dimensional setting). The sparsity is the same as in the previous simulations and the noise level $\sigma_0^2 = 1.5$ is estimated by $\hat{\sigma}^2_{\text{scaled}}$. We compare the MA-d estimator constructed with two different sequences for M, namely $M = 10\sqrt{n}$ rounded to the nearest integer and $M = n/2$, as well as with a fixed value, namely $M = 200$, as used in the previous simulations, in Section 5 of the main text.

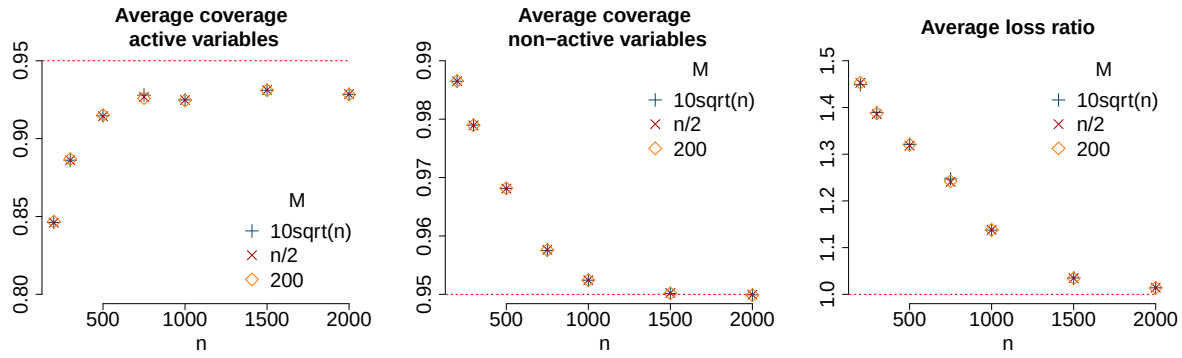

Figure 3: Average active coverage (left), non-active coverage (middle) and loss ratio (right) over 1000 replications for the MA-d estimator constructed with $M = 10\sqrt{n}$, $M = n/2$ and $M = 200$.

Figure 3 shows the average coverage over 1000 replications for the active and non-active variables and the average loss ratio from (8). First, we observe that the values for each sample size are similar for each choice of M. Moreover, for both active and non-active variables, the coverage converges to the target of 0.95 for increasing sample sizes. The loss ratio converges to 1, indicating that the loss of the MA-d estimator reaches optimality.

All in all, from this simulation study, we conclude that the choice of M seems to have a limited impact on the estimation of the parameters and that the MA-d still maintains a competitive performance even when M_n increases with the sample size. In practice, we recommend a moderate value for M as $10\sqrt{n}$ on a sufficiently large grid $[\lambda_{\min}, \lambda_{\max}]$ to avoid long computing times and an excessive use of resources.

When the signal is decreasing with the sample size

An important point is to check whether the estimator is robust to the settings where the signal decreases with the sample size. In this section, we simulate a setting where the signal is proportional to $1/(n^{1/3})$, and we observe that the loss optimality as well as the normality result still hold.

We compare the MA-d estimator with the debiased Lasso, where the regularization parameter has been chosen by 10-fold cross-validation. The noise level is $\sigma_0^2 = 1.5$ and was estimated by $\hat{\sigma}_{\text{scaled}}$ for both estimators. The number of active coefficients is the same as before, i.e. $10n^{1/4}/\log p$ rounded to the nearest integer, and the amplitude of the active coefficients is $20/(n^{1/3})$. This setting implies a fast decay of the signal as the sample size increases, even if the number of active coefficients grows slowly.

	Prediction loss ratio				Inference metrics *(nominal level .95)*			
n	MA-d	Debiased Lasso	Ridge	Lasso	MA-d		Debiased Lasso	
					Active	Non-active	Active	Non-active
100	1.19	3.09	11.86	0.14	.75	.98	.68	.96
200	1.16	3.08	6.80	0.07	.88	.98	.89	.96
300	1.10	2.92	5.45	0.05	.88	.98	.90	.96
500	1.05	2.82	3.61	0.03	.89	.98	.91	.96
750	1.03	2.76	2.60	0.02	.90	.98	.92	.96
1000	1.02	2.68	2.05	0.02	.90	.98	.92	.95

Table 7: Prediction loss ratio averaged over $R = 1000$ repetitions for estimators constructed with a path of $M = 200$ values and average coverage for active and non-active variables (target is .95), where the signal decreases with the sample size.

The results are shown in Table 7 and observe that the prediction loss ratio for the MA-d estimator converges to unity slightly slower than when the signal was fixed as in Table 2 in the main manuscript, but the performance of the proposal estimator is still better than that of the debiased Lasso and the Ridge.

In terms of inference, the MA-d estimator has a slightly lower coverage than the debiased Lasso for the active variables and a slight over-coverage for the non-active variables. This illustrates that the estimator loses more in accuracy than the competing estimator. However, the overall performance of the estimator shows that it is still interesting to consider the MA-d estimator if one wants to obtain a better prediction risk.

When the precision matrix is no more sparse

Assumption 2 requires that the population precision matrix Ω is sparse which was essential to guaranty normality and risk optimality. In this section we explore the case where this assumption is no longer valid.

We compare the performance of the MA-d estimator under two cases with a non-sparse precision matrix: (i) the population variance-covariance matrix, Σ, is positive definite and contains randomly generated entries and (ii) the Σ matrix is equicorrelated as detailed further in the next paragraph.

The simulation setup is similar to the previous studies, the competitors are the debiased Lasso, the Lasso and the Ridge, and each time the regularization parameter was chosen by 10-fold cross-validation. The noise level $\sigma_0^2 = 1.5$ is estimated by $\hat{\sigma}_{\text{scaled}}$, the rows of the design matrix $\mathbf{X}$ are i.i.d. and follow a multivariate Gaussian distribution, $\mathbf{X}_i \sim \mathcal{N}(0, \Sigma)$. In case (i), Σ was generated using the `genPositiveDefMat(·)` function in the `clusterGeneration` R package, and then scaled to produce a correlation matrix. This process involves two steps: first the eigenvalues are randomly generated, and then an orthogonal matrix is randomly generated and used to construct the covariance matrix. In case (ii), $\Sigma_{a,b} = 0.3$ for all $a \neq b$, meaning that its inverse is full. In both cases, the inverse, Ω, is no longer sparse.

	Prediction loss ratio				Inference metrics *(nominal level .95)*			
n	MA-d	Debiased Lasso	Ridge	Lasso	MA-d		Debiased Lasso	
					Active	Non-active	Active	Non-active
Case (i)								
100	1.68	3.57	6.45	0.29	0.88	0.98	0.91	0.98
200	1.59	3.90	5.73	0.16	0.87	0.98	0.92	0.97
300	1.51	3.84	5.87	0.12	0.88	0.98	0.94	0.97
500	1.38	3.73	5.18	0.07	0.88	0.98	0.94	0.96
750	1.25	3.48	4.58	0.04	0.88	0.98	0.95	0.96
1000	1.15	3.28	4.15	0.03	0.87	0.98	0.94	0.96
Case (ii)								
100	1.54	14.04	6.81	0.26	0.78	0.96	0.83	0.95
200	1.70	14.31	6.50	0.15	0.82	0.97	0.88	0.95
300	1.75	19.28	7.35	0.12	0.83	0.97	0.90	0.95
500	1.81	19.79	7.46	0.08	0.85	0.97	0.91	0.95
750	1.83	21.34	7.33	0.06	0.86	0.97	0.93	0.95
1000	1.84	21.20	7.40	0.05	0.86	0.98	0.94	0.95

Table 8: Prediction loss ratio averaged over $R = 1000$ repetitions for estimators constructed with a path of $M = 200$ values and average coverage for active and non-active variables (target = .95), when the precision matrix is not sparse. The variance covariance matrix, Σ, is random for case (i) and equicorrelated for case (ii).

From Table 8, we can see that the MA-d estimator outperforms the debiased Lasso and the Ridge in terms of prediction loss ratio for both cases. This means that even with a non-sparse precision matrix, the proposal estimator achieves better loss performance. However, when the variance-covariance matrix is equicorrelated, the ratio does not converge to 1. For the inference metrics, we observe that in case (i), the average coverage for the active variable is below the nominal level and does not improve as the sample size increases. In case (ii), the active coverage is below the nominal level but increases slightly with the sample size. On the other hand, the debiased Lasso gives more robust results. Empirically, the MA-d estimator seems to be more sensitive to the non-sparsity in the precision matrix for the inferential tasks. Therefore, if the focus is on inference and if the population precision matrix is not sparse, we recommend caution when using the MA-d estimator.

References

Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrika 66*(2), 237–242.

Ando, T. and K.-C. Li (2014). A Model-Averaging Approach for High-Dimensional Regression. *Journal of the American Statistical Association 109*(505), 254–265.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). SLOPE—Adaptive Variable Selection via Convex Optimization. *The Annals of Applied Statistics 9*(3), 1103–1140.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model Selection: An Integral Part of Inference. *Biometrics 53*(2), 603–618.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Berlin, Heidelberg: Springer.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference*. New York, NY: Springer.

Charkhi, A., G. Claeskens, and B. E. Hansen (2016). Minimum Mean Squared Error Model Averaging in Likelihood Models. *Statistica Sinica 26*(2), 809–840.

Chen, J. and Z. Chen (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika 95*(3), 759–771.

Chen, Y. and Y. Yang (2021). The One Standard Error Rule for Model Selection: Does It Work? *Stats 4*(4), 868–892.

Chen, Z., J. Zhang, W. Xu, and Y. Yang (2022). Consistency of Bic Model Averaging. *Statistica Sinica 32*, 635–640.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science 30*(4), 533–558.

Drakakis, K. and B. A. Pearlmutter (2009). On the Calculation of the l2→l1 Induced Matrix Norm. *International Journal of Algebra 3*(5), 231–240.

Fan, J. and R. Li (2001, December). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Feng, Y. and Q. Liu (2020). Nested model averaging on solution path for high-dimensional linear regression. *Stat 9*(1), e317.

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Boca Raton: Chapman and Hall/CRC.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics 5*(3), 495–530.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hjort, N. L. and G. Claeskens (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science 14*(4), 382–401.

Homrighausen, D. and D. J. McDonald (2017). Risk Consistency of Cross-Validation with Lasso-Type Procedures. *Statistica Sinica 27*(3), 1017–1036.

Homrighausen, D. and D. J. McDonald (2018). A study on tuning parameter selection for the high-dimensional lasso. *Journal of Statistical Computation and Simulation 88*(15), 2865–2892.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Lahiri, S. N. (2021). Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. *The Annals of Statistics 49*(2), 820–844.

Lan, W., Y. Ma, J. Zhao, H. Wang, and C.-L. Tsai (2018). Sequential Model Averaging for High Dimensional Linear Regression Models. *Statistica Sinica 28*(1), 449–469.

Li, X., T. Zhao, X. Yuan, and H. Liu (2015). The flare Package for High Dimensional Linear Regression and Precision Matrix Estimation in R. *The Journal of Machine Learning Research 16*(18), 553–557.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics 2*, 90–102.

Mallows, C. L. (1973). Some Comments on CP. *Technometrics 15*(4), 661–675.

Pötscher, B. M. (2006). The Distribution of Model Averaging Estimators and an Impossibility Result regarding Its Estimation. *Lecture Notes-Monograph Series 52*, 113–129.

Raftery, A. E. and Y. Zheng (2003). Discussion: Performance of Bayesian Model Averaging. *Journal of the American Statistical Association 98*(464), 931–938.

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Sun, T. and C.-H. Zhang (2013). Sparse Matrix Inversion with Scaled Lasso. *Journal of Machine Learning Research 14*(106), 3385–3418.

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wang, H., B. Li, and C. Leng (2009). Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 71*(3), 671–683.

Xie, J., X. Yan, and N. Tang (2021). A Model-averaging method for high-dimensional regression with missing responses at random. *Statistica Sinica 31*(2), 1005–1026.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhang, X., G. Zou, and R. J. Carroll (2015). Model Averaging Based on Kullback-Leibler Distance. *Statistica Sinica 25*, 1583–1598.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics 35*(5), 2173–2192.