INFERENCE FOR HIGH-DIMENSIONAL MODEL AVERAGING ESTIMATORS

Lise Léonard, Eugen Pircalabelu and Rainer von Sachs

UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences

Supplementary Material

Contents

S1 Proofs **2**

S2 Additional simulations **10**

S2.1 Simulation settings for the toy example 10

S2.2 Accuracy metrics 10

S2.3 Noise level estimation 12

S2.4 Sensitivity analysis for M 15

S2.5 Averaging competitors 17

S2.6 When the signal is decreasing with the sample size 22

S2.7 When the precision matrix is no longer sparse 24

S2.8 Best of both worlds 27

S1 Proofs

Proof of Proposition 1. 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 $\hat{\beta}_1^{\mathrm{MA}} \coloneqq \hat{\beta}^{\mathrm{MA}}(\hat{\mathbf{w}}_1)$ and $\hat{\beta}_2^{\mathrm{MA}} \coloneqq \hat{\beta}^{\mathrm{MA}}(\hat{\mathbf{w}}_2)$ the estimators associated. By contradiction, we assume that $\mathbf{X}\hat{\beta}_1^{\mathrm{MA}} \neq \mathbf{X}\hat{\beta}_2^{\mathrm{MA}}$ 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:

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

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 $\mathbf{X}\hat{\beta}_1^{\text{MA}}$ and $\mathbf{X}\hat{\beta}_2^{\text{MA}}$ are equal. Then

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

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

Proof of Proposition 2. Using the decomposition in Equation (2.4) of the main manuscript, we obtain

$$\begin{aligned}\left|\left|\hat{\beta}^{\text{MA}}(\hat{\mathbf{w}}) - \beta_0\right|\right|_\infty &= \left|\left|\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon + \sum_{m=1}^{M}\hat{w}_m\left(I_p - \hat{\Omega}\hat{\Sigma}\right)\left(\hat{\beta}_m^{\text{Lasso}} - \beta_0\right)\right|\right|_\infty \\ &\leq \left|\left|\frac{1}{n}\hat{\Omega}\mathbf{X}^T\epsilon\right|\right|_\infty + \left|\left|\sum_{m=1}^{M}\hat{w}_m\frac{\Delta_m}{\sqrt{n}}\right|\right|_\infty . \qquad \text{(S1.1)}\end{aligned}$$

Under Assumption 2, using the same argument as in Theorem 1, as $\hat{w}_m \in [0,1]$ for each $m \in \{1,\ldots,M\}$ 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 (S1.1), 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 $||\hat{\Omega}||_{\infty,\infty} \coloneqq \max_{j=1,\ldots,p}||\hat{\Omega}_j||_1$ and using further Lemma 5.3 from van de Geer et al. (2014), we have that $\max_{j=1,\ldots,p}||\hat{\Omega}_j||_1 = O_{\mathbb{P}}(\max_{j=1,\ldots,p}\sqrt{s_j})$. As such, on the event ξ_∞

(introduced in Section 4 of the main manuscript),

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

since $\max_{j=1,\ldots,p} \sqrt{s_j} = o(\sqrt{n/\log p})$ by Assumption 3. □

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

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

and as the weights sum to 1

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

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

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

By Assumptions 2 and 3, we have that for each m, $||\hat{\beta}_m^{\mathrm{Lasso}} - \beta_0||_1 = O_{\mathbb{P}}(s_0\lambda_m)$ due to Bühlmann and van de Geer (2011, Theorem 6.1). This implies that

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

$$\leq O_{\mathbb{P}}\left(\sqrt{\log p}\, s_0\, K\sqrt{\frac{\log p}{n}}\right),$$

where we have used the assumption that $\sup_{m\leq M_n} \lambda_m < K\sqrt{(\log p)/n}$ and Assumptions 2 and 3 for the control 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. □

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 $L_n\{\hat{\beta}^{\mathrm{MA}}(\mathbf{w})\}$ and finally, we bound the remainder term, $R\{\hat{\beta}^{\mathrm{MA}}(\mathbf{w})\}$.

Step 1: Using Equation (2.4) and with $\hat{\mu}^{\mathrm{MA}}(\mathbf{w}) = \mathbf{X}\hat{\beta}^{\mathrm{MA}}(\mathbf{w})$ and $\mu_0 = \mathbf{X}\beta_0$, we begin by re-writing $C_n(\mathbf{w})$.

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

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

$$\tilde{C}_n(\mathbf{w}) = L_n\left\{\hat{\beta}^{\mathrm{MA}}(\mathbf{w})\right\} \underbrace{-\frac{2}{n}\epsilon^T\mathbf{X}\left(I_p - \hat{\Omega}\hat{\Sigma}\right)\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\mathrm{Lasso}} - \beta_0\right) + \frac{2\hat{\sigma}^2}{n}\sum_{m=1}^{M} w_m \hat{s}_m}_{=R\left\{\hat{\beta}^{\mathrm{MA}}(\mathbf{w})\right\}}. \tag{S1.2}$$

The last two terms in (S1.2) correspond to the "bias terms" if we regard $\tilde{C}_n(\mathbf{w})$ as an estimator for the loss $L_n\{\hat{\beta}^{\mathrm{MA}}(\mathbf{w})\}$.

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

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

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

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

We develop next the rate of convergence of each term separately. Using the triangular and the Cauchy–Schwarz inequality $\left|u^T v\right| \leq ||u||_2||v||_2$ for all $u, v \in \mathbb{R}^p$, we have

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

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

where $||A||_2$ for a $p \times p$ matrix A denotes the spectral norm, $||A||_2 \coloneqq \sqrt{\Lambda_{\max}(A^T A)}$. The inequality on the last line comes from the basic properties of norms where the Euclidean and the spectral norms are bounded by the sup norm.

The control on $||\Omega||_2$ is implied by Assumption 2. Indeed, the maximum eigenvalue of the precision matrix is $1/\Lambda_{\min}(\Sigma)$ which is $O(1)$ by Assumption 2, and for this positive definite matrix, $||\Omega||_2 = \sqrt{\Lambda_{\max}(\Omega^T \Omega)} = \Lambda_{\max}(\Omega) = 1/\Lambda_{\min}(\Sigma)$ for all $j \in \{1, \ldots, p\}$. We thus have that $||\Omega||_2 = O(1)$.

We know that $||\hat{\Omega}^T \hat{\Sigma} \hat{\Omega} - \Omega||_\infty = O_{\mathbb{P}}(\max_{j=1,\ldots,p} \sqrt{s_j} \sqrt{(\log p)/n})$ by Lemma 5.4 in van de Geer et al. (2014) and so on the event ξ_∞, under Assumption 3, I_1 is of order

$$
I_1 = O_{\mathbb{P}} \left(p^2 \left(\frac{\log p}{n} \right)^{3/2} \max_{j=1,\ldots,p} \sqrt{s_j} + p \frac{\log p}{n} \right) = O_{\mathbb{P}} \left(p^2 \frac{\log p}{n} \right).
$$

For I_2, using again the Cauchy–Schwarz inequality, the triangular inequality and the bounds on the Euclidean and spectral norms, we have

$$
|I_2| = \left| \left\{ \frac{2\epsilon^T \mathbf{X}}{n} \left(\hat{\Omega}^T \hat{\Sigma} - I_p \right) + \frac{2\epsilon^T \mathbf{X}}{n} \right\} \left(I_p - \hat{\Omega} \hat{\Sigma} \right) \left\{ \sum_{m=1}^{M} w_m \left(\hat{\beta}_m^{\text{Lasso}} - \beta_0 \right) \right\} \right|
$$

$$
\begin{aligned}
&\leq \left|\left|2\left(\hat{\Sigma}\hat{\Omega}-I_p\right)\frac{\mathbf{X}^T\epsilon}{n}+\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\left(I_p-\hat{\Omega}\hat{\Sigma}\right)\left\{\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}}-\beta_0\right)\right\}\right|\right|_2 \\
&\leq \left|\left|2\left(\hat{\Sigma}\hat{\Omega}-I_p\right)\frac{\mathbf{X}^T\epsilon}{n}+\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2 \sqrt{p}\left|\left|\left(I_p-\hat{\Omega}\hat{\Sigma}\right)\left\{\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}}-\beta_0\right)\right\}\right|\right|_\infty \\
&\leq \left|\left|2\left(\hat{\Sigma}\hat{\Omega}-I_p\right)\frac{\mathbf{X}^T\epsilon}{n}+\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2 \sqrt{p}\left|\left|I_p-\hat{\Omega}\hat{\Sigma}\right|\right|_\infty \left(\sum_{m=1}^{M} w_m\left|\left|\hat{\beta}_m^{\text{Lasso}}-\beta_0\right|\right|_1\right) \\
&\leq \sqrt{p}\left(2\left|\left|\hat{\Sigma}\hat{\Omega}-I_p\right|\right|_2\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2+\left|\left|\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_2\right)\left|\left|I_p-\hat{\Omega}\hat{\Sigma}\right|\right|_\infty \left(\sum_{m=1}^{M} w_m\left|\left|\hat{\beta}_m^{\text{Lasso}}-\beta_0\right|\right|_1\right) \\
&\leq \sqrt{p}\left(2p^{3/2}\left|\left|\hat{\Sigma}\hat{\Omega}-I_p\right|\right|_\infty\left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty+\sqrt{p}\left|\left|\frac{2\mathbf{X}^T\epsilon}{n}\right|\right|_\infty\right)\left|\left|I_p-\hat{\Omega}\hat{\Sigma}\right|\right|_\infty \left(\sum_{m=1}^{M} w_m\left|\left|\hat{\beta}_m^{\text{Lasso}}-\beta_0\right|\right|_1\right).
\end{aligned}
$$

On the event ξ_∞, using Assumption 3, the term I_2 is of order:

$$
|I_2| = O_{\mathbb{P}}\left(p^2 s_0\left(\frac{\log p}{n}\right)^2+p\left(\frac{\log p}{n}\right)^{3/2}\right) = O_{\mathbb{P}}\left(p^2\frac{\log p}{n^{3/2}}+p\left(\frac{\log p}{n}\right)^{3/2}\right) = O_{\mathbb{P}}\left(p^2\frac{\log p}{n^{3/2}}\right).
$$

For I_3 we have

$$
\begin{aligned}
I_3 &\leq \frac{1}{n}\left|\left|\mathbf{X}\right|\right|_2^2\left|\left|\left(I_p-\hat{\Omega}\hat{\Sigma}\right)\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}}-\beta_0\right)\right|\right|_2^2 \\
&\leq \frac{1}{n}\Lambda_{\max}(\Sigma)\left|\left|\mathbf{Z}\right|\right|_2^2 p\left|\left|\sum_{m=1}^{M} w_m\frac{\Delta_m}{\sqrt{n}}\right|\right|_\infty^2,
\end{aligned}
$$

where, $\mathbf{Z}$ is a $n \times p$ matrix of i.i.d random vector $Z_i \sim \mathcal{N}(0, I_p)$ for $i = 1, \ldots, n$. By Theorem 4.4.5 in Vershynin (2018), we have that, with high probability, $||\mathbf{Z}||_2 \leq C(\sqrt{n}+\sqrt{p})$ for $C > 0$ a constant. As in Theorem 1, $\Delta_m := \sqrt{n}(I_p-\hat{\Omega}\hat{\Sigma})(\hat{\beta}_m^{\text{Lasso}}-\beta_0)$. Thus, under Assumption 3, I_3 is of order

$$
I_3 = O_{\mathbb{P}}\left(p\Lambda_{\max}(\Sigma)\left(1+\sqrt{\frac{p}{n}}\right)^2 s_0\frac{\log p}{n}\right) = O_{\mathbb{P}}\left(\frac{p^2}{n^{3/2}}\Lambda_{\max}(\Sigma)\right).
$$

Step 3: Bound for the remainder term, $R\{\hat{\beta}^{\text{MA}}(\mathbf{w})\}$.

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 inequality, that

$$\begin{aligned}\left|\frac{2}{n}\epsilon^T \mathbf{X}\left(I_p - \hat{\Omega}\hat{\Sigma}\right)\left\{\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}} - \beta_0\right)\right\}\right| &\leq \left|\left|2\left(I_p - \hat{\Sigma}\hat{\Omega}^T\right)\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left|\left|\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}} - \beta_0\right)\right|\right|_2 \\ &\leq 2\left|\left|I_p - \hat{\Sigma}\hat{\Omega}^T\right|\right|_2 \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_2 \left(\sum_{m=1}^{M} w_m \left|\left|\hat{\beta}_m^{\text{Lasso}} - \beta_0\right|\right|_2\right) \\ &\leq 2p^{3/2}\left|\left|I_p - \hat{\Sigma}\hat{\Omega}^T\right|\right|_\infty \left|\left|\frac{\mathbf{X}^T\epsilon}{n}\right|\right|_\infty \left(\sum_{m=1}^{M} w_m \left|\left|\hat{\beta}_m^{\text{Lasso}} - \beta_0\right|\right|_2\right).\end{aligned}$$

From Bühlmann and van de Geer (2011, Chapter 6), we know that $||\hat{\beta}_m^{\text{Lasso}} - \beta_0||_2 = O_{\mathbb{P}}\left(\sqrt{s_0(\log p)/n}\right)$ for all m and as $||I_p - \hat{\Sigma}\hat{\Omega}^T||_\infty = ||I_p - \hat{\Omega}\hat{\Sigma}||_\infty$,

$$\left|\frac{2}{n}\epsilon^T \mathbf{X}\left(I_p - \hat{\Omega}\hat{\Sigma}\right)\left\{\sum_{m=1}^{M} w_m\left(\hat{\beta}_m^{\text{Lasso}} - \beta_0\right)\right\}\right| = O_{\mathbb{P}}\left(p^{3/2}\left(\frac{\log p}{n}\right)^{3/2}\sqrt{s_0}\right) = O_{\mathbb{P}}\left(p^{3/2}\frac{\log p}{n^{5/4}}\right).$$

To conclude, on the event ξ_∞, the loss is of order

$$L_n\left\{\hat{\beta}^{\text{MA}}(\mathbf{w})\right\} = O_{\mathbb{P}}\left(p^2\frac{\log p}{n} + p^2\frac{\log p}{n^{3/2}} + \frac{p^2}{n^{3/2}}\Lambda_{\max}(\Sigma)\right) = O_{\mathbb{P}}\left(p^2\frac{\log p}{n}\right),$$

and the bias term in $\tilde{C}_n(\mathbf{w})$ is of order

$$R\left\{\hat{\beta}^{\text{MA}}(\mathbf{w})\right\} = O_{\mathbb{P}}\left(p^{3/2}\frac{\log p}{n^{5/4}}\right).$$

As the order of $L_n\{\hat{\beta}^{\text{MA}}(\mathbf{w})\}$ 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 $p^2(\log p)/n = o(1)$, which implies that the design must be low-dimensional.

□

S2 Additional simulations

S2.1 Simulation settings for the toy example

The squared prediction error loss is averaged over $R = 1000$ replicates for $n = 100$ and $p = 200$. The sequence of parameters is $\beta_0 = (2, \ldots, 2, 0, \ldots, 0)^T$ with $s_0 = 6$ non-zero entries, the design matrix is Gaussian with a Toeplitz covariance structure and the noise level is $\sigma^2 = 1.5$.

S2.2 Accuracy metrics

Table 1 shows accuracy metrics such as the average bias and MSE for active and non-active coefficients for different sample sizes and noise levels. The settings for the simulations are the same as in Section 5 from the main manuscript. The empirical

Table 1: Average bias and MSE for active and non-active variables over $R = 1000$ replications for different sample sizes and noise levels; the values are expressed in unit 10^{-2} for ease of readability.

σ_0^2	n	MA-d				Debiased Lasso				Lasso				MA Lasso			
		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
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	7.86	2.16	0.04	0.02
	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	5.83	1.05	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	4.76	0.68	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	3.92	0.43	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	3.36	0.29	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	3.03	0.23	<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	11.08	4.32	0.06	0.04
	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	8.22	2.09	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	6.68	1.36	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	5.50	0.85	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	4.70	0.58	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	4.20	0.45	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	15.66	8.63	0.09	0.09
	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	11.59	4.18	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	9.45	2.72	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	7.75	1.70	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	6.61	1.15	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	5.89	0.89	0.01	<0.01

quantities are calculated as follows:

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

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

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

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.

S2.3 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 an MLE version with a factor of $(1/n)$

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

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

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 of the main manuscript 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 2 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 3 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

Table 2: 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 $\hat{\sigma}^2_{\text{cv-mle}}$			
	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

obtained with $\hat{\sigma}^2_{\text{cv}}$ 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.

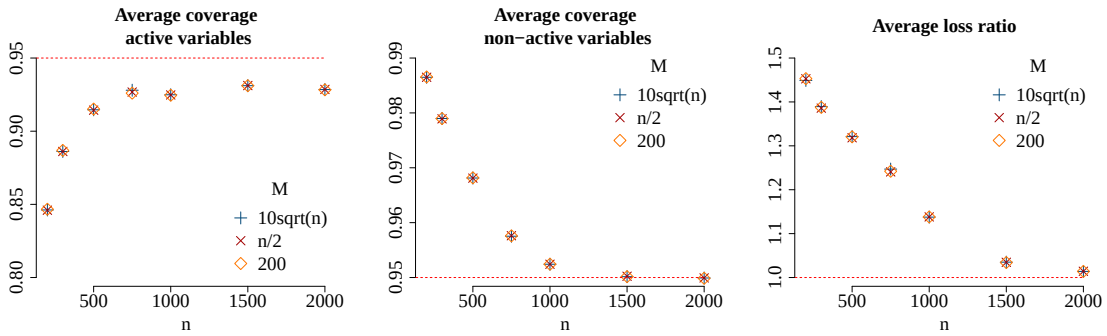
Table 3: Prediction loss ratio (LR) for the MA-d estimator constructed with different estimators for σ_0^2.

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 $\hat{\sigma}^2_{\text{cv}-\text{mle}}$
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

S2.4 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

Figure 1: Average active coverage (left), non-active coverage (middle) and loss ratio (right) over $R = 1000$ replications for the MA-d estimator constructed with $M = 10\sqrt{n}$, $M = n/2$ and $M = 200$.



$M = 200$, as used in the previous simulations, in Section 5 of the main text.

Figure 1 shows the average coverage over $R = 1000$ replications for the active and non-active variables and the average loss ratio (5.2) defined in the main manuscript. 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 increases with the sample size. In practice, we recommend a moderate value for M such as $10\sqrt{n}$ on a sufficiently large grid $[\lambda_{\min}, \lambda_{\max}]$ to avoid long computing times and an excessive use of resources.

S2.5 Averaging competitors

In this section, we compare the simulation results from Section 5 of the main paper with those obtained for five additional model averaging competitors. These are: a averaging method based on the Lasso path; a naive MA estimator based on the debiased Lasso; an MA Ridge estimator, and two ensemble estimators that average the solutions of Lasso, Ridge and debiased Lasso estimator. We also included the Ridge estimator where the penalization parameter was obtained by 10-fold cross-validation. These results can be directly compared with those in Table 2 of the main paper and the accuracy measures in Table 1.

The considered MA estimators are:

- Model averaging Lasso with naive weights (MA-Lasso-naive) is the mean of all the Lasso solutions along the regularization path, each with a weight of $1/M$,

$$\hat{\beta}^{\mathrm{MA-Lasso-naive}} := \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m^{\mathrm{Lasso}}.$$

- Model averaging debiased Lasso with naive weights (MA-d-naive) is the mean of all the debiased Lasso solutions associated with each value in the regularization path, each with a weight of $1/M$,

$$\hat{\beta}^{\mathrm{MA-d-naive}} := \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m^{\mathrm{d}}.$$

- Model averaging Ridge with naive weights (MA-Ridge-naive) is the mean of all the Ridge solutions associated with each value on the regularization path of the Ridge problem, each with a weight of $1/M$,

$$\hat{\beta}^{\text{MA}-\text{Ridge}-\text{naive}} := \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{\text{Ridge}}_{\lambda_m}.$$

The regularization path for the Lasso-based methods and the debiased Lasso is identical. In particular, the cross-validation solution is one of the estimators included the weighted mean of MA-d. Conversely, the Ridge estimator has its own regularization path.

The next two averaging competitors are a combination of three different estimators, a Lasso solution, a Ridge solution and a debiased Lasso solution.

- The first ensemble estimator is an average of the optimal cross-validation solutions from the Lasso, the Ridge, and the debiased Lasso problems. The weights are obtained using an optimization problem inspired by the one used in the paper,

$$\begin{aligned}
\hat{\beta}^{\text{MA}-\text{cv}} :=& \hat{w}_1\hat{\beta}^{\text{Lasso}}_{\text{cv}} + \hat{w}_2\hat{\beta}^{\text{d}}_{\text{cv}} + \hat{w}_3\hat{\beta}^{\text{Ridge}}_{\text{cv}},\\
\hat{\mathbf{w}} :=& \arg\min_{\mathbf{w}\in\mathcal{H}} C(\mathbf{w}, \hat{\beta}^{\text{Lasso}}_{\text{cv}}, \hat{\beta}^{\text{d}}_{\text{cv}}, \hat{\beta}^{\text{Ridge}}_{\text{cv}})\\
=& \arg\min_{\mathbf{w}\in\mathcal{H}} \frac{1}{n}\left|\left|Y - \mathbf{X}\left(w_1\hat{\beta}^{\text{Lasso}}_{\text{cv}} + w_2\hat{\beta}^{\text{d}}_{\text{cv}} + w_3\hat{\beta}^{\text{Ridge}}_{\text{cv}}\right)\right|\right|_2^2 + \frac{2\hat{\sigma}^2}{n}\left(w_1||\hat{\beta}^{\text{Lasso}}_{\text{cv}}||_0 + w_2||\hat{\beta}^{\text{d}}_{\text{cv}}||_0 + w_3||\hat{\beta}^{\text{Ridge}}_{\text{cv}}||_0\right),
\end{aligned} \tag{S2.1}$$

where $||\cdot||_0$ is the ℓ_0 norm that counts the number of non-zero entries of a vector and $\hat{\beta}_{\text{cv}}$ denotes the cross-validation solution of the estimator.

- The second ensemble estimator is an average of the three MA-naive estimators described just above,

$$\begin{aligned} \hat{\beta}^{\text{MA}-\text{naive}} &:= \hat{w}_1 \hat{\beta}^{\text{MA}-\text{Lasso}-\text{naive}} + \hat{w}_2 \hat{\beta}^{\text{MA}-\text{d}-\text{naive}} + \hat{w}_3 \hat{\beta}^{\text{MA}-\text{Ridge}-\text{naive}} \\ \hat{\mathbf{w}} &:= \arg\min_{\mathbf{w} \in \mathcal{H}} C\left(\mathbf{w}, \hat{\beta}^{\text{MA}-\text{Lasso}-\text{naive}}, \hat{\beta}^{\text{MA}-\text{d}-\text{naive}}, \hat{\beta}^{\text{MA}-\text{Ridge}-\text{naive}}\right). \end{aligned} \tag{S2.2}$$

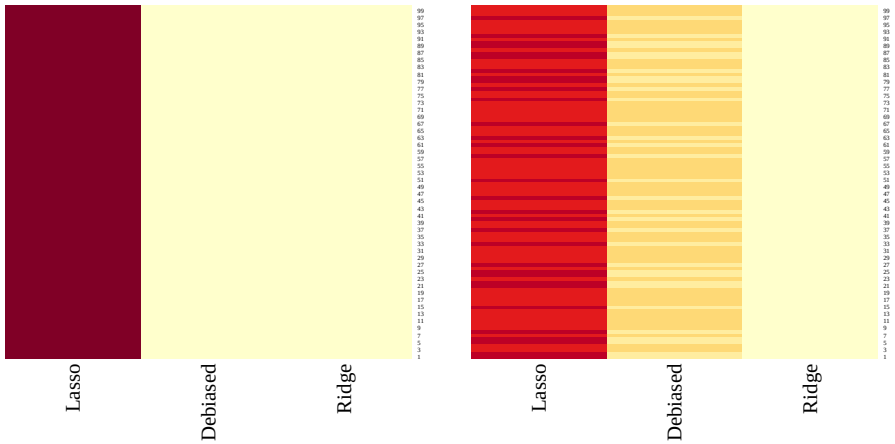


Figure 2: Heatmap of the weights vector for the ensemble estimators (S2.1) and (S2.2), respectively left and right, for 100 replications of the same design with $n = 500, p = 1000$, red corresponds to one and light yellow corresponds to zero.

Figure 2 displays the weight vectors for estimators (S2.1) and (S2.2) on the first 100 replications of the design with $n = 500$, $p = 1000$ and $\sigma^2 = 1.5$. We observe that all the weights of the ensemble MA-cv estimator (S2.1), are on the Lasso-cv solution. Indeed, this solution results in a small squared prediction loss with few non-zero coefficients. For the MA-naive estimator (S2.2), the Lasso solution has weights between 0.80 and 1, the remaining weights are placed on the debiased Lasso

solution. It should also be noted that the properties of such ensemble methods with respect to performing inference are intricate, and that no statistical guarantees are currently available.

Table 4 shows the accuracy measures and the prediction loss ratios for MA-d and the new competitors, averaged over 1000 replications.

In terms of accuracy metrics, the MA-d estimator has a smaller average MSE than the MA-d-naive estimator and a slightly smaller average bias. Except for the MA-Ridge estimator, the MA methods with naive weights appear to be highly variable, as can be seen with the MSE. The MA-Ridge estimator performs slightly better than the Ridge in terms of accuracy but not for the prediction loss. For ensemble methods, since the weights are mostly on Lasso estimates, the performance is similar to that of the Lasso estimator presented in Table 1.

σ^2	n	MA-d		MA-d-naive		MA-Lasso-naive		Ridge		MA-Ridge-naive		MA-cv		MA-naive	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0.75	100	11.59	23.75	12.57	31.68	10.02	18.06	6.83	12.67	5.93	12.02	11.40	22.45	10.58	20.09
	200	5.91	12.11	6.28	15.81	5.13	9.22	3.54	6.39	3.06	6.02	5.82	11.49	5.38	10.17
	300	4.59	9.46	4.84	12.37	4.01	7.19	2.79	5.01	2.40	4.69	4.55	9.01	4.19	7.93
	500	2.78	5.73	2.91	7.39	2.42	4.35	1.70	3.02	1.46	2.82	2.75	5.45	2.53	4.77
	750	1.89	3.84	2.02	4.91	1.62	2.91	1.16	2.01	0.99	1.88	1.84	3.65	1.72	3.18
	1000	1.44	2.89	1.56	3.67	1.22	2.19	0.89	1.51	0.75	1.41	1.39	2.75	1.30	2.39
1.5	100	11.55	24.18	12.55	32.38	10.00	18.06	6.84	12.69	5.93	12.02	11.30	22.16	10.53	20.03
	200	5.89	12.31	6.28	16.21	5.13	9.22	3.54	6.39	3.06	6.02	5.78	11.37	5.37	10.14
	300	4.58	9.59	4.83	12.66	4.01	7.19	2.79	5.01	2.41	4.69	4.53	8.92	4.18	7.90
	500	2.78	5.82	2.91	7.57	2.42	4.35	1.70	3.02	1.46	2.82	2.74	5.41	2.53	4.75
	750	1.89	3.92	2.03	5.04	1.62	2.91	1.17	2.02	0.99	1.88	1.84	3.63	1.71	3.17
	1000	1.45	2.95	1.57	3.77	1.22	2.19	0.89	1.51	0.75	1.41	1.38	2.73	1.30	2.38
3	100	11.51	25.16	12.52	33.72	9.97	18.09	6.85	12.71	5.93	12.03	11.15	21.81	10.45	19.97
	200	5.88	12.78	6.28	16.95	5.12	9.23	3.54	6.41	3.06	6.02	5.73	11.20	5.34	10.09
	300	4.57	9.89	4.82	13.20	4.00	7.19	2.79	5.02	2.41	4.69	4.49	8.81	4.16	7.87
	500	2.77	5.99	2.91	7.92	2.42	4.35	1.70	3.03	1.46	2.82	2.72	5.35	2.52	4.73
	750	1.90	4.04	2.04	5.29	1.62	2.91	1.17	2.02	0.99	1.88	1.83	3.59	1.70	3.16
	1000	1.46	3.05	1.58	3.96	1.22	2.19	0.89	1.52	0.75	1.41	1.38	2.71	1.29	2.36
		Prediction loss ratio													
0.75	100	1.08		21.82		2.61		16.88		32.17		0.11		0.55	
	200	1.01		15.91		1.86		12.42		24.02		0.05		0.41	
	300	1.00		15.13		1.78		11.26		22.67		0.03		0.37	
	500	1.00		11.74		1.39		8.93		18.09		0.01		0.30	
	750	1.00		9.71		1.17		7.58		15.41		0.01		0.26	
	1000	1.00		8.71		1.06		6.86		13.99		0.01		0.24	
1.5	100	1.23		15.13		1.69		10.69		20.40		0.14		0.48	
	200	1.08		11.85		1.26		8.31		16.08		0.06		0.37	
	300	1.02		11.44		1.23		7.72		15.56		0.04		0.33	
	500	1.00		8.95		0.96		6.11		12.40		0.02		0.26	
	750	1.00		7.29		0.79		5.07		10.33		0.01		0.22	
	1000	1.00		6.36		0.69		4.45		9.07		0.01		0.19	
3	100	1.41		10.35		1.04		6.33		12.09		0.17		0.43	
	200	1.27		8.86		0.81		5.23		10.16		0.08		0.34	
	300	1.13		8.81		0.82		5.08		10.26		0.05		0.31	
	500	1.02		7.22		0.66		4.15		8.43		0.03		0.25	
	750	1.00		6.00		0.54		3.47		7.08		0.02		0.21	
	1000	1.00		5.26		0.47		3.03		6.19		0.01		0.18	

Table 4: Bias, mean squared error (MSE) and prediction loss ratio on $R = 1000$ replications. The values of the bias and the MSE were expressed in unit 10^{-2} for readability.

In the second part of Table 4, we observe that the prediction loss ratio of the MA-d estimator is much smaller than that of MA-d-naive estimator. The ensemble estimators (S2.1) and (S2.2) outperform the MA-d estimator in terms of prediction loss. This is likely due to the fact that most of the weights are assigned to the Lasso solutions. However, there are no statistical guarantees for inference with such ensemble estimators.

S2.6 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 data from 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 $\hat{\sigma}^2_{\text{scaled}}$ 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.

The results are shown in Table 5 and we observe that the prediction loss ratio

Table 5: 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.

	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

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 targets a better prediction risk.

S2.7 When the precision matrix is no longer sparse

Assumption 3 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.

Table 6: 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).

	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

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 $\hat{\sigma}^2_{\text{scaled}}$, 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.

From Table 6, 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.

S2.8 Best of both worlds

The proposed estimator is designed to facilitate inference for high-dimensional models with an optimal prediction loss. Notably, it outperforms the debiased Lasso estimator in terms of prediction loss, while maintaining an asymptotically Gaussian distribution. Moreover, in some cases, the MA-d estimator can provide better prediction loss performance even relative to the Lasso and better coverage performance relative to the debiased Lasso. Here, we describe one such design.

We consider model (2.1) with $p = 50$ covariates and a Toeplitz covariance matrix Σ with $\rho = 0.5$. The noise level σ is set to 1, and there are 35 non-zero coefficients, each set at 3. The sample size varies from 100 to 1500. We compare the MA-d estimator with the Lasso and the debiased Lasso, for which the penalization parameter was chosen using 10 fold cross-validation.

Figure 3 shows the average loss ratio, as defined in Equation (5.2), and the average coverage for active and non-active coefficients, based on 1000 replications of the design. Regarding the prediction loss, we observe that the MA-d estimator initially has a slightly higher prediction loss for small sample sizes, but this decreases quickly, and outperforms the Lasso loss as soon as the sample size reaches 200.

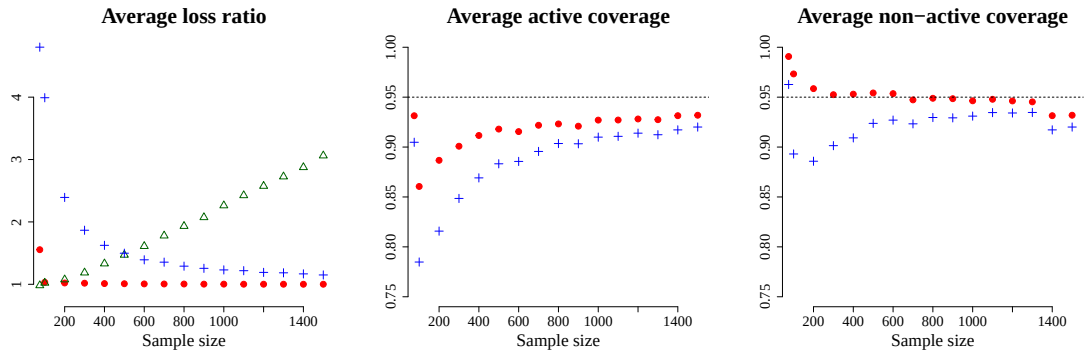


Figure 3: Average loss ratio, average active coverage and average non-active coverage on $R = 1000$ replications for the MA-d estimator (red circle), the Lasso (green triangle) and the debiased Lasso (blue cross). The penalization parameter was chosen by 10-fold cross-validation for the Lasso and the debiased Lasso estimators, and the nominal coverage level is set at .95.

Regarding the empirical coverage, the MA-d estimator is more accurate than the debiased Lasso for both active and non-active coverage, over all sample sizes.

References

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

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

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.