Econometric Theory, **0**, 2024, 1–30.
doi:10.1017/S0266466624000070

BOUNDED SUPPORT IN LINEAR RANDOM COEFFICIENT MODELS: IDENTIFICATION AND VARIABLE SELECTION

PHILIPP HERMANN
Philipps-Universität Marburg

HAJO HOLZMANN
Philipps-Universität Marburg

We consider linear random coefficient regression models, where the regressors are allowed to have a finite support. First, we investigate identification, and show that the means and the variances and covariances of the random coefficients are identified from the first two conditional moments of the response given the covariates if the support of the covariates, excluding the intercept, contains a Cartesian product with at least three points in each coordinate. We also discuss identification of higher-order mixed moments, as well as partial identification in the presence of a binary regressor. Next, we show the variable selection consistency of the adaptive LASSO for the variances and covariances of the random coefficients in finite and moderately high dimensions. This implies that the estimated covariance matrix will actually be positive semidefinite and hence a valid covariance matrix, in contrast to the estimate arising from a simple least squares fit. We illustrate the proposed method in a simulation study.

1. INTRODUCTION

In various statistical analyses in fields such as medicine and economics, there is a large extent of individual heterogeneity in the effect of observed covariates, which is routinely modeled by random coefficients—also called random effects—models. For example, in contemporary microeconomic data sets with many observations and potentially a large number of explanatory variables, unobserved heterogeneity plays an important role (Lewbel, 2005). An important issue then is to select those coefficients which actually are random if there is a large set of potential variables which might have individual-specific effects. To this end, in this paper, we shall consider the following random coefficients regression model:

$$Y = B_0 + \mathbf{W}^\top \mathbf{B}, \tag{1.1}$$

Address correspondence to Hajo Holzmann, Department of Mathematics and Computer Science, Philipps-Universität Marburg, Marburg, Germany, holzmann@mathematik.uni-marburg.de.

where $\mathbf{B}, \mathbf{W} \in \mathbb{R}^{p-1}$ are independent random vectors, B_0 is a random variable, and $\mathbf{W} = (W_1, \ldots, W_{p-1})^\top$ represents the random regressors.

Model (1.1), which is related to random effects models from the literature on biostatistics (Schelldorfer, Bühlmann, and van de Geer, 2011), was introduced by Hildreth and Houck (1968) and Swamy (1970). They assumed that $B_0, \ldots, B_{p-1}$ are independent, and focused on estimating their means and variances by least squares in two stages. Arellano and Bonhomme (2012) studied a panel—version of the random coefficient model. Beran and Hall (1992) initiated the nonparametric analysis of the distribution of the random coefficients. For $p = 2$, Beran, Feuerverger, and Hall (1996) used Fourier methods to construct an estimator of the joint density of $(B_0, B_1)^\top$. Their method was taken up again by Hoderlein, Klemelä, and Mammen (2010), who put it into the form of a more conventional kernel estimator and generalized it to arbitrary dimension p. Masten (2018, Lem. 2) provides necessary and sufficient conditions for identification of the overall distribution of $(B_0, \mathbf{B})$ in terms of moments. Further related literature includes Ichimura and Thompson (1998) and Gautier and Kitamura (2013), who analyze a binary choice version of the model, Lewbel and Pendakur (2017) who study a generalization of (1.1) in which the products $B_1 W_1, \ldots, B_{p-1} W_{p-1}$ are related to Y by some arbitrary (possibly nonlinear) unknown function, as well as Hoderlein, Holzmann, and Meister (2017), Breunig and Hoderlein (2018), Dunker et al. (2019), and Holzmann and Meister (2020). Recently, Gaillac and Gautier (2022) studied nonparametric identification and adaptive estimation in a random coefficient regression model, where covariates have bounded but continuous variation.

The above nonparametric approaches which target the full density of the random coefficients require a large or at least, as in Gaillac and Gautier (2022), continuous support of the covariates, which is often an unrealistic assumption in applications. In this paper, we shall focus on situations in which the covariates have bounded and in particular finite support. In this latter setting, there is little hope to identify and estimate the density of the random coefficients nonparametrically. Therefore, we shall focus on the first and second moments, which are arguably of most interest in applications. Variable selection techniques for means, variances, and covariances of the random coefficients then allow to determine which variables have an effect on average (nonzero mean of the coefficient), which variables have heterogeneous effects (nonzero variances) and for which covariates the effects are correlated. In particular, we shall argue that it is important not to focus exclusively on the variances of the random coefficients, but to take the full variance–covariance matrix into account. Further, estimating the first and second moments of the random coefficients then allows to predict the first and second moments of the response Y conditional on the covariates. Finally, normality of the random coefficients is a common parametric assumption, under which their distribution is fully determined by means, variances, and covariances.

Model (1.1) is related to random effects models from the literature on biostatistics (Schelldorfer et al., 2011). These are studied in a longitudinal framework, and the goal is then to estimate the fixed effects by using a quasi-likelihood approach and to predict the random effects. Papers which study these models in a high-dimensional setting are, among others, Schelldorfer et al. (2011) and Li, Cai, and Li (2021).

The paper is organized as follows: In Section 2, we clarify under which assumptions on the support of the covariates, first and second moments of the random coefficients are identified. It turns out that identification holds if the support of the covariate vector contains a Cartesian product with at least three support points for each covariate. Conversely, identification generally fails if one covariate only has two support points. In Section 3, we turn to estimation and in particular to variable selection with a focus on the variances and covariances in model (1.1). We use the adaptive LASSO originally introduced in Zou (2006), which may achieve variable selection consistency without additional restrictive assumptions such as the irrepresentable assumption required for the ordinary LASSO, and show the variable selection consistency in fixed and moderately high dimensions. The technical issues are to deal with the residuals when estimating centered second moments of the random coefficients as well as with the heteroscedasticity of the model. Section 4 contains some numerical illustrations. Proofs of the main results are given in Section 5, while some further auxiliary results are deferred to the Supplementary Material.

We shall use the following notation: For an $n \times p$ matrix $\mathbb{X}$ and a subset $S \subseteq \{1,\dots,p\}$ of the index set, $\mathbb{X}_S$ denotes the $n \times |S|$ matrix containing those columns of $\mathbb{X}$ with indices in S. A similar notation is $\boldsymbol{v}_S$ for a vector $\boldsymbol{v} \in \mathbb{R}^p$. $\|\mathbb{X}\|_{\mathrm{M},2}$ denotes the operator norm of $\mathbb{X}$ for the Euclidean norm, and $\|\mathbb{X}\|_{\mathrm{F}}$ the Frobenius norm, that is the Euclidean norm of the vectorization of $\mathbb{X}$.

2. IDENTIFICATION OF FIRST AND SECOND MOMENTS

In model (1.1), we also write $\mathbf{A} = (B_0, \mathbf{B}^\top)^\top \in \mathbb{R}^p$, so that $\mathbf{A} = (A_1,\dots,A_p)^\top$ and $\mathbf{W}$ are independent. We assume that the first and second moments of the random coefficients $\mathbf{A}$ exist and set

$$\mu^* := \mathbb{E}[\mathbf{A}] \quad \in \mathbb{R}^p, \qquad \text{and} \qquad \Sigma^* := \mathbb{C}\mathrm{ov}(\mathbf{A}) = \mathbb{E}\big[(\mathbf{A}-\mu^*)(\mathbf{A}-\mu^*)^\top\big] \quad \in \mathbb{R}^{p\times p}. \tag{2.1}$$

In this section, we consider the conditions for identification and partial identification of the moments μ^* and Σ^* in terms of the support of the covariates $\mathbf{W}$.

While one may argue that means and variances are of main applied interest, the joint variation of the random coefficients as described by the covariances and the correlations is also relevant. Further, we shall see that excluding covariances from the analysis and falsely assuming a diagonal covariance matrix a priori can lead to wrong conclusions about the (non-)randomness of the coefficients. The proofs of the results in this section are collected in Section 5.1.

To illustrate, first consider the case of a single regressor, resulting in the model

$$Y = B_0 + W_1 B_1\,. \tag{2.2}$$

For W_1 supported on $\{0,1\}$, we have the following result.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

PROPOSITION 2.1. *Suppose that in model (2.2), the random variable $W_1 \in \{0,1\}$ is binary, and denote the identified standard deviations by*

$$s_1 = \sqrt{\mathbb{V}\mathrm{ar}(B_0)}, \qquad s_2 = \sqrt{\mathbb{V}\mathrm{ar}(B_0+B_1)}.$$

Then each value

$$\sqrt{\mathbb{V}\mathrm{ar}(B_1)} \in \big[|s_1-s_2|, s_1+s_2\big]$$

is consistent with s_1 and s_2, provided the correlation $\rho = \mathbb{C}\mathrm{or}(B_0,B_1)$ is chosen for $\sqrt{\mathbb{V}\mathrm{ar}(B_1)} > 0$ as

$$\rho = \frac{s_2^2 - s_1^2 - \mathbb{V}\mathrm{ar}(B_1)}{2 s_1 \sqrt{\mathbb{V}\mathrm{ar}(B_1)}} \in \begin{cases} [-1,1], & \text{if } s_2 > s_1, \\ \big[-1, -\sqrt{s_1^2 - s_2^2}/s_1\big], & \text{if } s_1 \geq s_2. \end{cases} \tag{2.3}$$

Thus, to conclude from $\mathbb{V}\mathrm{ar}(B_0) = \mathbb{V}\mathrm{ar}(B_0+B_1)$ that $\mathbb{V}\mathrm{ar}(B_1) = 0$ fully relies on the assumption of a diagonal covariance matrix, without this assumption, B_1 can well be random. On the other hand, the following proposition shows that three distinct support points of W_1 are enough to identify the means $\mathbb{E}[B_j]$, the variances $\mathbb{V}\mathrm{ar}(B_j)$, $j = 0,1$, and the covariance $\mathbb{C}\mathrm{ov}(B_0,B_1)$. From Proposition 2.1 and not surprisingly, two support points are insufficient for this purpose.

PROPOSITION 2.2. *In model (2.2), if W_1 has $n+1$ support points and $\mathbb{E}\big[|B_0|^n\big], \mathbb{E}\big[|B_1|^n\big] < \infty$, then all mixed moments $\mathbb{E}\big[B_0^j B_1^k\big]$, $j,k \geq 0$, $j+k \leq n$, are identified.*

2.1. Identification of the Covariance Matrix

Now, let us turn to the identification of μ^* and Σ^* in (2.1) in general dimensions. To this end, consider the half-vectorization of symmetric matrices of dimension $p \times p$,

$$\mathrm{vec}(M) = \big(M_{11}, \ldots, M_{pp}, M_{12}, \ldots, M_{1p}, M_{23}, \ldots, M_{2p}, \ldots, M_{(p-1)p}\big)^\top \in \mathbb{R}^{\frac{p(p+1)}{2}} \tag{2.4}$$

for $M \in \mathbb{R}^{p\times p}$ with $M^\top = M$, and set

$$\sigma^* := \mathrm{vec}(\Sigma^*).$$

Note that the first p entries of σ^* are the variances and the remaining entries are the covariances. In model (1.1), we have that

$$\mathbb{V}\mathrm{ar}\big(Y \,\big|\, \mathbf{W} = \mathbf{w}\big) = (1, \mathbf{w}^\top)\, \Sigma^* \,(1, \mathbf{w}^\top)^\top, \tag{2.5}$$

so that the quadratic form in Σ^* is identified over $(1,\mathbf{w})$ with $\mathbf{w}$ ranging over the support of $\mathbf{W}$. Note that (2.5) can be written in vectorized form as

$$\begin{aligned} \mathbb{V}\mathrm{ar}\big(Y \,\big|\, \mathbf{W} = \mathbf{w}\big) &= \big(1, (\mathbf{w}^2)^\top, 2\mathbf{w}^\top, 2w_1w_2, \ldots, 2w_1w_{p-1}, 2w_2w_3, \ldots, 2w_{p-2}w_{p-1}\big)\,\sigma^* \\ &= \mathrm{v}\big((1,\mathbf{w}^\top)^\top\big)^\top \sigma^*, \end{aligned} \tag{2.6}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

where we recall that $\sigma^* = \mathrm{vec}(\Sigma^*)$, and the vector transformation v is defined by

$$\mathrm{v}(\mathbf{x}) = \left(x_1^2, \ldots, x_p^2, 2x_1x_2, \ldots, 2x_1x_p, 2x_2x_3, \ldots, 2x_2x_p, \ldots, 2x_{p-1}x_p\right)^\top \in \mathbb{R}^{\frac{p(p+1)}{2}}, \quad \mathbf{x} \in \mathbb{R}^p. \tag{2.7}$$

Based on (2.6), we can establish linear equations for the $p(p+1)/2$ entries of Σ^*, respectively, σ^*. With the above notation, we may state the following basic result.

THEOREM 2.3. *In model (1.1), a sufficient condition for identification of the mean vector μ^* and the covariance matrix Σ^* is the existence of $p(p+1)/2$ points $\mathbf{w_1}, \ldots, \mathbf{w_{p(p+1)/2}} \in \mathbb{R}^{p-1}$ in the support of $\mathbf{W}$, for which the matrix*

$$S = \left[\mathrm{v}\Big((1, \mathbf{w_1}^\top)^\top\Big), \ldots, \mathrm{v}\Big((1, \mathbf{w}_{\mathbf{p(p+1)/2}}^\top)^\top\Big)\right]^\top \tag{2.8}$$

of dimension $p(p+1)/2 \times p(p+1)/2$ is of full rank. This condition is also necessary for identification in the subset of full-rank covariance matrices.

The theorem remains valid if one can show that for $m \geq p(p+1)/2$ support points, the resulting matrix S_m has full rank $p(p+1)/2$.

In the following example, we show that the condition of the previous theorem can never be satisfied if one of the regressors only has two support points.

Example 2.1. Suppose that W_1 has only two support points a and b and that the joint support of $\mathbf{W}$ is finite. Then the matrix S_m, where m is the total number of support points, has rank at most $p(p+1)/2-1$. Thus, from Theorem 2.3, full-rank covariance matrices Σ^* are not identified. Indeed, the matrix $S_m^\top$ contains the submatrix

$$\begin{bmatrix} 1 & \ldots & 1 & 1 & \ldots & 1 \\ a^2 & \ldots & a^2 & b^2 & \ldots & b^2 \\ 2a & \ldots & 2a & 2b & \ldots & 2b \end{bmatrix} \in \mathbb{R}^{3\times m}.$$

Evidently, this matrix is of column rank at most 2, since there are only two distinct columns. Thus, its row rank is also at most two, which implies that the corresponding three columns in S_m are linearly dependent.

In contrast, if each covariate has at least three support points and the joint support contains the corresponding Cartesian product, then we retain identification of Σ^*.

THEOREM 2.4. *Consider model (1.1). Suppose that the support of $\mathbf{W} = (W_1, \ldots, W_{p-1})^\top$ contains the Cartesian product of three points in each coordinate. Then there exist $p(p+1)/2$ support points such that the matrix S in (2.8) has full rank $p(p+1)/2$ and consequently, the means and (co-)variances of the random coefficients $\mathbf{A}$ are identified. Conversely, if there is a W_j having only two support points, then in the full-rank covariance matrices identification fails.*

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

2.2. Partial Identification

What can be said about the covariance matrix of the random coefficients if there are binary regressors? Assume a single binary regressor Z, and additional regressors $\mathbf{W} \in \mathbb{R}^{p-2}$ (slightly modifying the notation in this section) for which the support contains a Cartesian product with at least three points in each coordinate. Our model is then written as

$$Y = B_0 + Z B_1 + \mathbf{W}^\top \mathbf{B_2}.$$

The set of covariance matrices of $\mathbf{A} = (B_0, B_1, \mathbf{B_2}^\top)^\top \in \mathbb{R}^p$ consistent with the conditional second moments of Y is

$$\begin{aligned} \mathcal{S} := \Big\{ \Sigma \in \mathbb{R}^{p\times p} \,\Big|\, & \Sigma \text{ positive semi-definite and} \\ & (1, z, \mathbf{w}^\top)\, \Sigma\, (1, z, \mathbf{w}^\top)^\top = \mathbb{V}\mathrm{ar}(Y \,|\, Z = z, \mathbf{W} = \mathbf{w}) \ \forall\ (z, \mathbf{w}) \in \operatorname{supp}(Z, \mathbf{W}) \Big\}. \end{aligned} \tag{2.9}$$

Suppose that the support of $(Z, \mathbf{W}^\top)^\top \in \mathbb{R}^{p-1}$ has a product structure. From Theorem 2.4, using $Z = 0$ and $Z = 1$, we identify the covariance matrices

$$\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big) \quad \text{and} \quad \mathbb{C}\mathrm{ov}\big((B_0 + B_1, \mathbf{B_2}^\top)^\top\big),$$

or equivalently

$$\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big), \qquad \mathbb{C}\mathrm{ov}\big(B_1; \mathbf{B_2}\big), \qquad \mathbb{V}\mathrm{ar}(B_0 + B_1). \tag{2.10}$$

Here, for random vectors $\mathbf{C}$ and $\mathbf{D}$, $\mathbb{C}\mathrm{ov}(\mathbf{C})$ is the covariance matrix of $\mathbf{C}$, whereas $\mathbb{C}\mathrm{ov}(\mathbf{C}; \mathbf{D})$ contains the cross-covariances of $\mathbf{C}$ and $\mathbf{D}$.

Sharp bounds for $\mathbb{V}\mathrm{ar}(B_1)$ are given by

$$\inf_{\Sigma \in \mathcal{S}} \Sigma_{22} \le \mathbb{V}\mathrm{ar}(B_1) \le \sup_{\Sigma \in \mathcal{S}} \Sigma_{22}, \tag{2.11}$$

where the set $\mathcal{S}$ in (2.9) is characterized by the restrictions given by the identified parts (2.10) of the matrix Σ^*. These bounds can be obtained numerically by semi-definite programming. An interesting particular question is the potential randomness of B_1, which is addressed in the following proposition, which relies on the identified quantities in (2.10).

Proposition 2.5. *Suppose that the support of $(Z, \mathbf{W}^\top)^\top$ has a product structure, and that only Z is binary.*

1. *If $\mathbb{V}\mathrm{ar}(B_0) \neq \mathbb{V}\mathrm{ar}(B_0 + B_1)$, or if $\mathbb{C}\mathrm{ov}(B_1; \mathbf{B_2})$ is not the zero vector, then $\mathbb{V}\mathrm{ar}(B_1) > 0$.*
2. *Conversely, suppose that $\mathbb{V}\mathrm{ar}(B_0) = \mathbb{V}\mathrm{ar}(B_0 + B_1)$ and that $\mathbb{C}\mathrm{ov}(B_1; \mathbf{B_2}) = \mathbf{0}_{p-2}$.*
   (a) *If $\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)$ is degenerate, and its kernel contains a vector with nonzero first coordinate, then necessarily $\mathbb{V}\mathrm{ar}(B_1) = 0$.*
   (b) *On the other hand, if $\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)$ has full rank, then the upper bound in (2.11) for $\mathbb{V}\mathrm{ar}(B_1)$ is strictly positive.*

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

2.3. Identification of Higher-Order Moments

The kth-order mixed moments of the random vector $\mathbf{A}$, $k \in \mathbb{N}$, are given by

$$m(k_1,\dots,k_p) = \mathbb{E}\big[A_1^{k_1}\dots A_p^{k_p}\big], \qquad k_j \in \mathbb{N}_0,\ k_1+\dots+k_p = k,$$

of which there are $\binom{p+k-1}{k}$ many. Information on the mixed moments in the linear random coefficient model $Y = A_1 + A_2 W_1 + \dots + A_p W_{p-1}$ comes from the identified conditional kth moments of Y given $\mathbf{W}$,

$$\mathbb{E}[Y^k|\mathbf{W}=\mathbf{w}] = \mathbb{E}\big[\big((1,\mathbf{w}^\top)\,\mathbf{A}\big)^k\big]. \tag{2.12}$$

These can be represented as an inner product of $\binom{p+k-1}{k}$-dimensional vectors, one consisting of the mixed moments $m(k_1,\dots,k_p)$, the other with corresponding entry

$$\binom{k}{k_1\dots k_p} w_1^{k_2}\cdot\ldots\cdot w_{p-1}^{k_p}, \tag{2.13}$$

where $\mathbf{w} = (w_1,\dots,w_{p-1})$. Hence, we have analogously to the result in Theorem 2.3 that if there are $\binom{p+k-1}{k}$ support points $\mathbf{w}_j$ of $\mathbf{W}$ such that if we form the matrix with rows as in (2.13) for the coordinates of the $\mathbf{w}_j$, the resulting quadratic matrix has full rank, then the kth -order mixed moments of $\mathbf{A}$ are identified.

While we were not able to obtain a sufficient condition along the lines of Theorem 2.4, we have the following result which guarantees identification.

THEOREM 2.6. *If in model (1.1), the support of* $\mathbf{W} = (W_1,\dots,W_{p-1})^\top$ *contains* p *points* $\mathbf{w}_1,\dots,\mathbf{w}_p$ *in general position, for which for each* $j \in \{1,\dots,k\}$ *and* $i_1,\dots,i_j \in \{1,\dots,p\}$, *the vector* $(\mathbf{w}_{i_1}+\dots+\mathbf{w}_{i_j})/j$ *is also in the support of* $\mathbf{W}$. *Then the mixed moments of* $\mathbf{A}$ *up to order* k *are identified.*

3. SIGN-CONSISTENCY OF THE ADAPTIVE LASSO ESTIMATOR

In this section, we derive the asymptotic variable selection properties of the adaptive LASSO in the linear random coefficient regression model (1.1), where we focus on estimating and selecting the variances and covariances of the random coefficients. First, in Section 3.1, we consider an asymptotic regime with a fixed number p of regressors, before turning to the moderately high-dimensional setting in which $p \to \infty$ but at a slower rate than the sample size n.

The adaptive LASSO and its variable selection properties, originally introduced in Zou (2006), have already been investigated intensively in the literature. For example, Zou and Zhang (2009) consider the adaptive LASSO and an adaptive version of the elastic net in moderately high dimensions, while Huang, Ma, and Zhang (2008) investigate the high-dimensional situation with strong assumptions on the first-stage estimator, and Wagener and Dette (2013) extend their approach to a heteroscedastic framework. Here, our contributions mainly are to deal with the residuals when estimating centered second moments of the random coefficients,

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

and to extend the analysis of Zou and Zhang (2009) to our setting with random coefficients.

We observe independent random vectors $(Y_1, \mathbf{W_1}^\top)^\top, \dots, (Y_n, \mathbf{W_n}^\top)^\top$ distributed according to the random coefficient regression model (1.1), and write

$$Y_i = B_{i,0} + \mathbf{W_i}^\top \mathbf{B_i} = \mathbf{X_i}^\top \mathbf{A_i}, \quad i = 1, \dots, n,$$

where $\mathbf{X_i} = (1, \mathbf{W_i}^\top)^\top \in \mathbb{R}^p$ with $\mathbf{W_i} \sim \mathbf{W}$ and $\mathbf{A_i} = (B_{i,0}, \mathbf{B_i}^\top)^\top \sim \mathbf{A}$ are independent random vectors. Here, $\mathbf{X_i} = (X_{i,1}, \dots, X_{i,p})^\top$ represents the observed covariates and $\mathbf{A_i} = (A_{i,1}, \dots, A_{i,p})^\top$ the unobserved individual regression coefficients.

In the following, we denote by

$$S_\sigma := \operatorname{supp}(\sigma^*) = \Big\{ k \in \{1, \dots, p(p+1)/2\} \,\big|\, \sigma_k^* \neq 0 \Big\}, \qquad s_\sigma := |S_\sigma|,$$

the support of the half-vectorization σ^* of the covariance matrix Σ^*. $S_\sigma^c := \{1, \dots, p(p+1)/2\} \setminus S_\sigma$ will denote the relative complement of this set. For an estimator $\widehat{\mu}_n$ of μ^*, we define the regression residuals $\widetilde{Y}_i := Y_i - \mathbf{X_i}^\top \widehat{\mu}_n$, and write the squared residuals as

$$Y_i^\sigma := \widetilde{Y}_i^2 = \mathbf{X_i}^\top \big(D_i - \Sigma^* + E_n + F_{n,i}\big) \mathbf{X_i},$$

where we set

$$D_i := \big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top, \qquad E_n := \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top, \tag{3.1}$$

$$F_{n,i} := \big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mathbf{A_i} - \mu^*\big)^\top. \tag{3.2}$$

Applying the half-vectorization vec for symmetric matrices in (2.4) and the corresponding vector transformation v in (2.7), we obtain in vector-matrix form

$$\mathbb{Y}_n^\sigma = \mathbb{X}_n^\sigma \sigma^* + \varepsilon_n^\sigma = \mathbb{X}_{n,S_\sigma}^\sigma \sigma_{S_\sigma}^* + \varepsilon_n^\sigma,$$

where

$$\mathbb{Y}_n^\sigma := \Big(\big(Y_1 - \mathbf{X_1}^\top \widehat{\mu}_n\big)^2, \dots, \big(Y_n - \mathbf{X_n}^\top \widehat{\mu}_n\big)^2 \Big)^\top, \qquad \mathbb{X}_n^\sigma := \Big[\mathrm{v}\big(\mathbf{X_1}\big), \dots, \mathrm{v}\big(\mathbf{X_n}\big) \Big]^\top,$$

$$\varepsilon_n^\sigma := \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \operatorname{vec}\big(D_1 - \Sigma^* + E_n + F_{n,1}\big), \dots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \operatorname{vec}\big(D_n - \Sigma^* + E_n + F_{n,n}\big) \Big)^\top. \tag{3.3}$$

Then the adaptive LASSO estimator with regularization parameter $\lambda_n^\sigma > 0$ is given by

$$\widehat{\sigma}_n^{\mathrm{AL}} \in \rho_{\sigma,n,\lambda_n^\sigma}^{\mathrm{AL}} := \underset{\beta \in \mathbb{R}^{p(p+1)/2}}{\arg\min} \left(\frac{1}{n} \big\| \mathbb{Y}_n^\sigma - \mathbb{X}_n^\sigma \beta \big\|_2^2 + 2\lambda_n^\sigma \sum_{k=1}^{p(p+1)/2} \frac{|\beta_k|}{\big|\widehat{\sigma}_{n,k}^{\mathrm{init}}\big|} \right), \tag{3.4}$$

where $\widehat{\sigma}_n^{\mathrm{init}} \in \mathbb{R}^{p(p+1)/2}$ is an initial estimator of σ^*. Note that if $\widehat{\sigma}_{n,k}^{\mathrm{init}} = 0$, we require $\beta_k = 0$.

3.1. Asymptotics for Fixed Dimension *p*

The proofs of the results in this section are deferred to Section 6 of the Supplementary Material.

Assumption 1 (Fixed p). We assume that $\left(\mathbf{X_i}^\top, \mathbf{A_i}^\top\right)^\top$, $i = 1, \dots, n$, are identically distributed, and that:

(A1) the random coefficients $\mathbf{A}$ have finite forth moments,
(A2) the covariates $\mathbf{X} = (1, \mathbf{W}^\top)^\top$ (or rather $\mathbf{W}$) have finite eighth moments,
(A3) the symmetric matrix

$$\mathrm{C}^\sigma := \mathbb{E}\left[\mathrm{v}\big(\mathbf{X}\big)\,\mathrm{v}\big(\mathbf{X}\big)^\top\right],$$

which contains the fourth moments of the covariates, is positive definite.

In the following proposition, we show that the critical third part of the assumptions follows from our identification results in Section 2.

PROPOSITION 3.1. *Under the assumption of Theorem 2.4, that the support of the covariate vector* $\mathbf{W}$ *contains a Cartesian product with three points in each coordinate, Assumption 1, (A3), is satisfied, that is,* C^σ *is positive definite.*

To formulate an asymptotic result on variable selection consistency and asymptotic normality in fixed dimensions, set

$$\mathrm{B}^\sigma := \mathbb{E}\left[\left(\mathrm{v}\big(\mathbf{X}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{X}\big)\right)\mathrm{v}\big(\mathbf{X}\big)\,\mathrm{v}\big(\mathbf{X}\big)^\top\right], \tag{3.5}$$

where

$$\Psi^* := \Big[\mathrm{vec}\big(\mathcal{M}^{11}\big), \dots, \mathrm{vec}\big(\mathcal{M}^{pp}\big), \mathrm{vec}\big(\mathcal{M}^{12}\big), \dots, \mathrm{vec}\big(\mathcal{M}^{1p}\big),$$
$$\mathrm{vec}\big(\mathcal{M}^{23}\big), \dots, \mathrm{vec}\big(\mathcal{M}^{2p}\big), \dots, \mathrm{vec}\big(\mathcal{M}^{(p-1)p}\big)\Big]^\top$$

with $\mathcal{M}^{kl} \in \mathbb{R}^{p\times p}$ and

$$\big(\mathcal{M}^{kl}\big)_{uv} := \mathbb{C}\mathrm{ov}\Big(\big(A_k - \mu_k^*\big)\big(A_l - \mu_l^*\big), \big(A_u - \mu_u^*\big)\big(A_v - \mu_v^*\big)\Big). \tag{3.6}$$

THEOREM 3.2 (Variable selection and asymptotic normality for fixed p). *Suppose that the estimator* $\widehat{\mu}_n$ *of* μ^* *used in the residuals* $\widetilde{Y}_i$ *is* $\sqrt{n}$*-consistent, that is,* $\sqrt{n}\left(\widehat{\mu}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}(1)$.

Further, let Assumption 1 be satisfied, and assume that for the initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ *in the adaptive LASSO* $\widehat{\sigma}_n^{\mathrm{AL}}$ *in (3.4), we also have that* $\sqrt{n}\left(\widehat{\sigma}_n^{\mathrm{init}} - \sigma^*\right) = \mathcal{O}_{\mathbb{P}}(1)$. *If the regularization parameter is chosen as* $\lambda_n^\sigma \to 0$, $\sqrt{n}\lambda_n^\sigma \to 0$ *and* $n\lambda_n^\sigma \to \infty$, *then it follows that* $\widehat{\sigma}_n^{\mathrm{AL}}$ *is sign-consistent,*

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}_n^{\mathrm{AL}}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) \to 1, \tag{3.7}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

and satisfies

$$\sqrt{n}\left(\widehat{\sigma}^{\mathrm{AL}}_{n,S_\sigma} - \sigma^*_{S_\sigma}\right) \xrightarrow{d} \mathcal{N}_{s_\sigma}\left(\mathbf{0}_{s_\sigma}, \left(\mathrm{C}^\sigma_{S_\sigma S_\sigma}\right)^{-1} \mathrm{B}^\sigma_{S_\sigma S_\sigma} \left(\mathrm{C}^\sigma_{S_\sigma S_\sigma}\right)^{-1}\right). \tag{3.8}$$

We defer the proof of the theorem to the Section 6 of the Supplementary Material.

Remark 3.3 (Guaranteeing a positive semi-definite matrix). Consider the positive semi-definite cone

$$\mathbb{S}_p^+ := \left\{M \in \mathbb{R}^{p\times p} \mid M \text{ is symmetric and positive semi-definite}\right\} \quad \subset \mathbb{R}^{p\times p},$$

and its image under the vectorization operator

$$\mathbb{V}_p^+ := \left\{\mathrm{vec}(M) \mid M \in \mathbb{S}_p^+\right\} \quad \subset \mathbb{R}^{\frac{p(p+1)}{2}}.$$

It would be of interest to directly restrict the estimate of σ^* to $\mathbb{V}_p^+$, resulting in

$$\widehat{\sigma}^{\mathrm{AL}}_{n,\mathrm{pos}} \in \underset{\beta\in\mathbb{V}_p^+}{\arg\min} \left(\frac{1}{n}\left\|\mathbb{Y}^\sigma_n - \mathbb{X}^\sigma_n \beta\right\|_2^2 + 2\lambda^\sigma_n \sum_{k=1}^{p(p+1)/2} \frac{|\beta_k|}{\left|\widehat{\sigma}^{\mathrm{init}}_{n,k}\right|}\right), \tag{3.9}$$

an actual covariance matrix. Computationally, this estimate is feasible in principle by using methods from semidefinite programming as discussed, for example, in Vandenberghe and Boyd (1996), or by reparameterizing positive semidefinite matrices in terms of Cholesky factors and maximizing over these Cholesky factors. However, technically, it is hard to extend the primal-dual witness approach underlying the proof of Theorem 3.2 to this setting. Indeed, the primal-dual witness approach amounts to showing that a vector with the correct sparsity pattern asymptotically satisfies the necessary and sufficient KKT—conditions for a minimizer of (3.4). However, these KKT conditions become intractable for the semidefinite problem in (3.9).

Remark 3.4 (Post selection inference). The main result in Theorem 3.2 is the sign consistency (3.7), as application of asymptotic normality (3.8) suffers from issues of post selection inference (see, e.g., Leeb and Pötscher, 2003).

Fortunately, we have the following result, in which some coefficients are nonrandom, while those which actually are random have a non-singular covariance matrix.

COROLLARY 3.5. *Under the conditions of Theorem 3.2, suppose that the covariance matrix of the random coefficients in (2.1) has the form*

$$\Sigma^* = \begin{bmatrix} \Sigma_1^* & \mathbf{0}_{d\times(p-d)} \\ \mathbf{0}_{(p-d)\times d} & \mathbf{0}_{(p-d)\times(p-d)} \end{bmatrix}$$

for a positive definite $d\times d$-matrix Σ_1^*. Then $\mathbb{P}(\widehat{\sigma}^{\mathrm{AL}}_{n,pos} = \widehat{\sigma}^{\mathrm{AL}}_n) \to 1$, $n\to\infty$.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

This follows from Theorem 3.2 since the blocks of zeros in Σ^* are estimated as zero with probability tending to one, and the estimate for Σ_1^* will be positive definite asymptotically with full probability, since the positive definite matrices are open in $\mathbb{R}^{d\times d}$. Hence, the unconstrained estimator $\widehat{\sigma}_n^{\mathrm{AL}}$ will correspond with probability tending to 1 to a positive semi-definite matrix, which proves the corollary. Note that the corresponding statement would not be true for the ordinary least squares estimator.

3.2. Diverging Number *p* of Parameters

Again, we shall focus on the covariance matrix, for a discussion of estimating the means, see the Section 8 of the Supplementary Material. Recall C^σ and B^σ which are given in (A3) and (3.5).

Assumption 2 (Growing p). We assume that $\left(\mathbf{X}_\mathbf{i}^\top, \mathbf{A}_\mathbf{i}^\top\right)^\top$, $i = 1, \ldots, n$, are identically distributed, and that:

(A4) the random coefficients $\mathbf{A}$ have finite fourth moments,
(A5) the vector transformation $\mathrm{v}(\mathbf{X})$ of the covariates $\mathbf{X}$ is sub-Gaussian after centering,
(A6) $c_{\mathrm{C}^\sigma,\mathrm{l}} \le \lambda_{\min}\left(\mathrm{C}^\sigma\right) \le \lambda_{\max}\left(\mathrm{C}^\sigma\right) \le c_{\mathrm{C}^\sigma,\mathrm{u}}$ for some positive constants $0 < c_{\mathrm{C}^\sigma,\mathrm{l}} \le c_{\mathrm{C}^\sigma,\mathrm{u}} < \infty$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal eigenvalues of a symmetric matrix A,
(A7) $\lambda_{\max}\left(\mathrm{B}^\sigma\right) \le c_{\mathrm{B}^\sigma,\mathrm{u}}$ for some positive constant $c_{\mathrm{B}^\sigma,\mathrm{u}} > 0$,
(A8) $\lim_{n\to\infty} p^4/n = 0$.

The proof of the following result is provided in Section 5.2.

THEOREM 3.6 (Variable selection for diverging p). *Suppose that the estimator $\widehat{\mu}_n$ of μ^* used in the residuals $\widetilde{Y}_i$ is $\sqrt{n/p}$-consistent, that is, $\sqrt{n/p}\,\|\widehat{\mu}_n - \mu^*\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. Further, let Assumption 2 be satisfied, and assume that for the initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ in the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ in (3.4), we have also $\sqrt{n}/p\,\left\|\widehat{\sigma}_n^{\mathrm{init}} - \sigma^*\right\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. Moreover, if the regularization parameter is chosen as $\lambda_n^\sigma \to 0$,*

$$\sqrt{s_\sigma\, n}\,\lambda_n^\sigma/(\sigma_{\min}^* p) \to 0, \quad p/(\sigma_{\min}^*\sqrt{n}) \to 0, \quad n\lambda_n^\sigma/p^2 \to \infty$$

with $\sigma_{\min}^* := \min_{k\in S_\sigma} |\sigma_k^*|$, then it follows that $\widehat{\sigma}_n^{\mathrm{AL}}$ is sign-consistent,

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}_n^{\mathrm{AL}}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) \to 1\,. \tag{3.10}$$

Remark 3.7. Additional technical issues in the proof of Theorem 3.6, as compared to the analysis in Zou (2006), are to deal with the residuals when estimating centered second moments of the random coefficients as well as with the heteroscedasticity of the model. Let us also point out that under the assumptions of the theorem, the least squares estimator satisfies the requirements made on the initial estimator.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

Remark 3.8. For fixed p (and S_σ), we obtain the same conditions for the choice of the regularization parameter as in Theorem 3.2. Moreover, if only S_σ is fixed, but the number of coefficients grows, the first condition on the regularization parameter in Theorem 3.6 simplifies to $\sqrt{n}\lambda_n^\sigma/p \to 0$ and the second one is satisfied by (A8).

Remark 3.9. Assumption (A5) is satisfied for bounded covariates which we mainly focus on in this paper. If we merely assume a sub-Gaussian distribution for the regressor vector $\mathbf{X}$ instead of its vector transformation $\mathrm{v}(\mathbf{X})$, we would require a result for the rate of concentration of the sample fourth moment matrix of sub-Gaussian random vectors in the spectral norm.

Remark 3.10. Assumption (A8) can be relaxed to $\lim_{n\to\infty} p^2/n = 0$, which is the minimal condition so that the assumptions of Theorem 3.6 can be satisfied, if the centered coefficients $\mathbf{A} - \mu^*$ are sub-Gaussian as well and $\lim_{n\to\infty} n\exp(-C_p p) = 0$ holds for some positive constant $C_p > 0$. See Remark 5.9 after the proof of Lemma 5.8.

Remark 3.11 (Elastic net). Our results in Theorem 3.6 should extend to the adaptive elastic net estimator, see Zou and Zhang (2009) for an analysis of the adaptive elastic net in moderately high dimensions. The asymptotic properties should be similar to those of the adaptive LASSO, but its numerical performance may be better since the covariates in the design matrix $\mathbb{X}_n^\sigma$ may be highly correlated.

4. SIMULATIONS

In this section, we investigate numerically the performance of the adaptive LASSO with respect to variable selection of the variances and covariances of the random coefficients in two settings. Moreover, we consider various combinations for the sample size n and the number p of coefficients to study the performance for growing p.

We consider the linear random coefficient regression model (1.1) where the first four coefficients $\left(B_0, B_1, B_2, B_3\right)^\top \sim \mathcal{N}_4\left(\mu_1^*, \Sigma_1^*\right)$ are normally distributed with mean vector $\mu_1^* = \left(40, 15, 0, -10\right)^\top$ and covariance matrix

$$\Sigma_1^* = \begin{bmatrix} 10 & 15.65 & -5.20 & 0 \\ 15.65 & 50 & 0 & 12.65 \\ -5.20 & 0 & 30 & -12.25 \\ 0 & 12.65 & -12.25 & 20 \end{bmatrix}.$$

The exact correlations of the coefficients are $\rho_{01} = \mathbb{C}\mathrm{or}(B_0, B_1) = 0.7$, $\rho_{02} = -0.3$, $\rho_{13} = 0.4$, $\rho_{23} = -0.5$ and evidently $\rho_{03} = \rho_{12} = 0$. Furthermore, we set the fifth coefficient B_4 equal to 20 and add deterministic zeros for the remaining $p-5$ coefficients in model (1.1). Hence, we obtain in total the mean vector

$$\mu^* = \left(\left(\mu_1^*\right)^\top, 20, \mathbf{0}_{(p-5)}^\top\right)^\top$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

and the covariance matrix

$$\Sigma^* = \begin{bmatrix} \Sigma_1^* & \mathbf{0}_{4\times(p-4)} \\ \mathbf{0}_{(p-4)\times 4} & \mathbf{0}_{(p-4)\times(p-4)} \end{bmatrix}$$

(which equals the setting in Corollary 3.5) for the random coefficient vector **A**. Obviously, the number s_σ of nonzero elements in the half-vectorization σ^* of the covariance matrix Σ^* is always equal to 8 for each $p \geq 5$. Moreover, the covariates $W_1, \ldots, W_{p-1}$ in model (1.1) are assumed to be independent and identically uniform distributed on the interval $[-1,1]$ ($\mathcal{U}[-1,1]$) or on the set $\{-1,0,1\}$ ($\mathcal{U}\{-1,0,1\}$).

In our numerical study, we simulate n pairs $(Y_1, \mathbf{W}_\mathbf{1}^\top)^\top, \ldots, (Y_n, \mathbf{W}_\mathbf{n}^\top)^\top$ of data according to one of the above specified models and use them for variable selection of the second central moments of the random coefficients. For that purpose, we apply the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$, which is given in (3.4), with the ordinary LASSO estimator as well as the least squares estimator as initial estimators $\widehat{\sigma}_n^{\mathrm{init}}$. To determine the residuals of the first stage mean regression, we use the ordinary least squares estimator $\widehat{\mu}_n^{\mathrm{LS}}$. The adaptive LASSO is computed in our simulation by using the function `glmnet` of the eponymous package. Note that the intercept of the regression model is not penalized by this function, which means that the variance of the random intercept B_0 is not penalized in our setting. This is plausible since the coefficient B_0 includes the deterministic intercept as well as a random error which is not affected by the covariates.

In each of the following scenarios, we perform a Monte Carlo simulation with $m = 10{,}000$ iterations to illustrate the sign-consistency of the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ for various sample sizes, numbers of coefficients and supports for the regressors. Its regularization parameter λ is always chosen such that the sign-recovery rate is as high as possible. For this purpose, we use $1{,}000$ independent repetitions in each scenario, run through a grid for λ in each data set and determine the regularization parameters with a correct number of degrees of freedom.

The average percentage of correct sign-recoveries are displayed in the subsequent Figure 1 for $n = 5{,}000$ and Figure 4 for $n = 10{,}000$ for both the least squares estimator as well as the ordinary LASSO as initial estimators, and for both choices of covariates. As the LASSO as initial estimator leads to much better selection performance, we concentrate on it in the following, where we consider in more detail, the number of false positives and false negatives.

(a) *Findings for sample size* $n = 5{,}000$.
Let us discuss the findings from Figures 1–3. Evidently, for both kinds of regressors, the sign-recovery rate decreases if the number of coefficients increases. Note that the number of parameters which are estimated grows quadratically with the number p of random coefficients since the half-vectorization σ^* of the covariance matrix Σ^* has dimension $p(p+1)/2$. In particular, if we consider $p = 60$ coefficients in our model, we obtain $1{,}830$ variances and covariances. Hence, the results look quite satisfying, however, if the support of the regressors consists only of the three points $\{-1,0,1\}$, the sign-recovery rate is somewhat lower and decreases also slightly faster, as seen

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

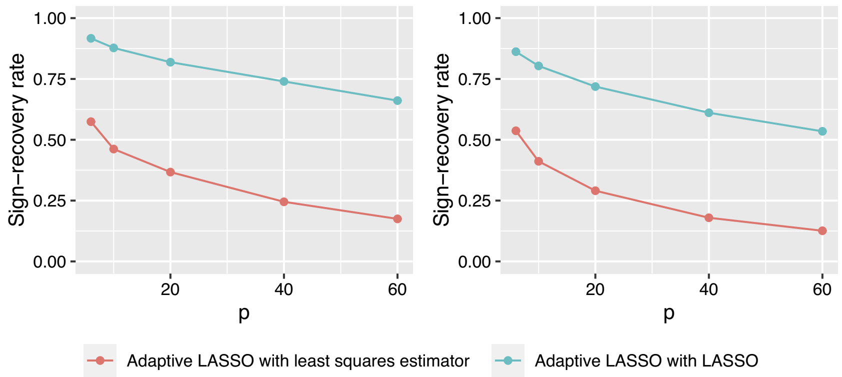

FIGURE 1. Left chart shows the sign-recovery rate for $\mathcal{U}[-1,1]$ distributed regressors, right one for $\mathcal{U}\{-1,0,1\}$ distributed regressors. The sample size is always $n = 5,000$.

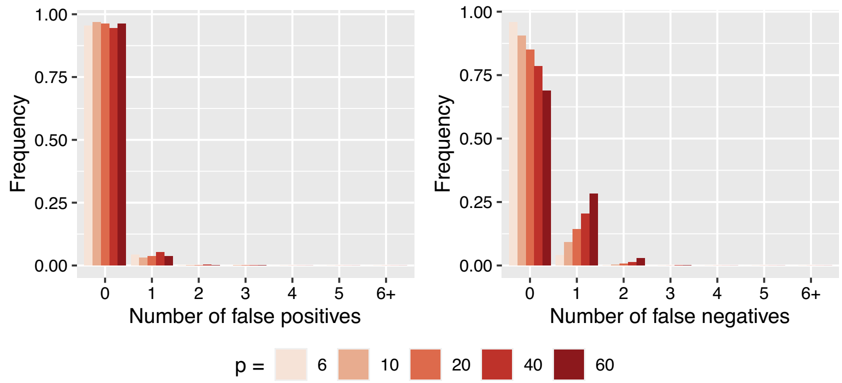

FIGURE 2. Relative frequency of false positives and false negatives for adaptive LASSO with LASSO as initial estimator, $\mathcal{U}[-1,1]$ distributed regressors and sample size $n = 5,000$.

in Figure 1. Second, there are rarely false positives, so that discoveries actually correspond to signals. The error in the sign recovery mainly stems from false negatives, of which there are rarely more than one, as seen in Figures 2 and 3.

(b) *Sample size* $n = 10,000$.

In this setting, the sign-recovery rate is in all scenarios much higher than in the first one with $n = 5,000$. See Figures 5 and 6.

For comparison, we also present a figure for the sign-recovery rate for the mean in Figure 7. Here, even with the simple least squares estimator as initial estimator, the sign-recovery rate is already very high for sample size $n = 5,000$.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

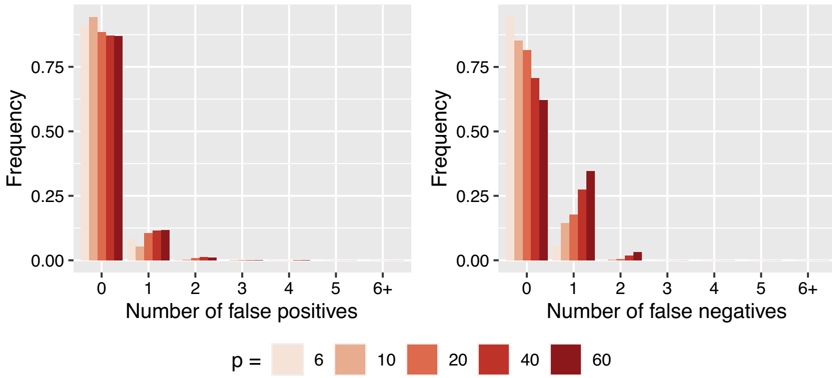

FIGURE 3. Relative frequency of false positives and false negatives for adaptive LASSO with LASSO as initial estimator, $\mathcal{U}\{-1,0,1\}$ distributed regressors and sample size $n = 5,000$.

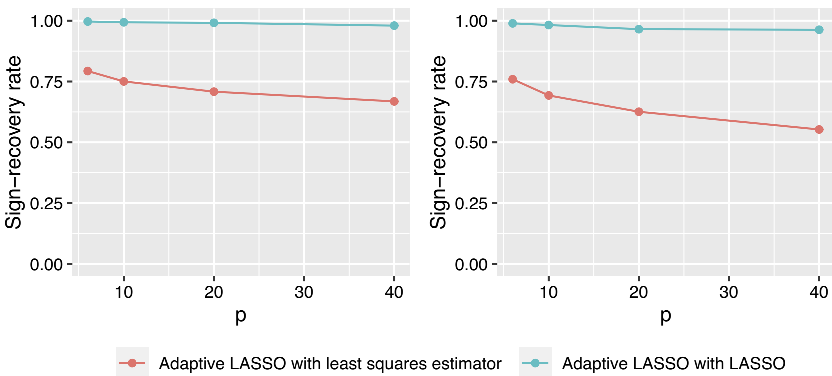

FIGURE 4. Left chart shows the sign-recovery rate for $\mathcal{U}[-1,1]$ distributed regressors, right one for $\mathcal{U}\{-1,0,1\}$ distributed regressors. The sample size is always $n = 10,000$.

5. PROOFS OF THE MAIN RESULTS

5.1. Proofs for Section 2

5.1.1. *Proofs of Propositions 2.1 and 2.2.*

Proof of Proposition 2.1. Set $u = \sqrt{\mathbb{V}\mathrm{ar}(B_1)}$. From $s_2^2 = s_1^2 + u^2 + 2\rho s_1 u$ and $|\rho| \le 1$, we obtain the inequalities

$$(u - s_1)^2 \le s_2^2 \le (u + s_1)^2.$$

By equating $s_2^2 = (u + s_1)^2$, we obtain the solutions $\pm s_2 - s_1$ for u, which yields $u \ge s_2 - s_1$ if $s_2 > s_1$. If $s_2 \le s_1$, we obviously have only the bound $u \ge 0$. Equating

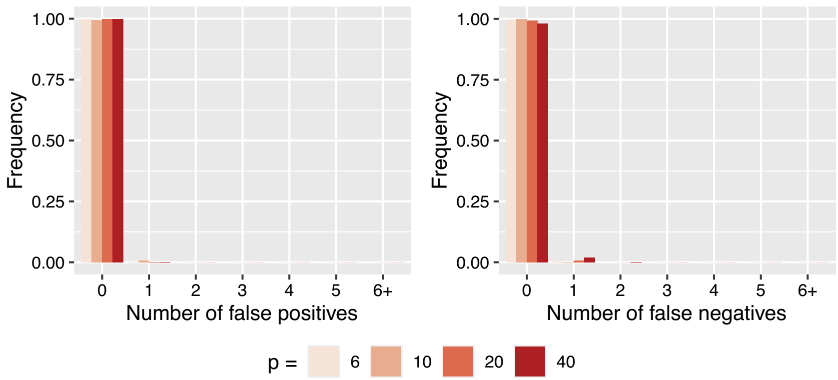

FIGURE 5. Relative frequency of false positives and false negatives for adaptive LASSO with LASSO as initial estimator, $\mathcal{U}[-1,1]$ distributed regressors and sample size $n = 10{,}000$.

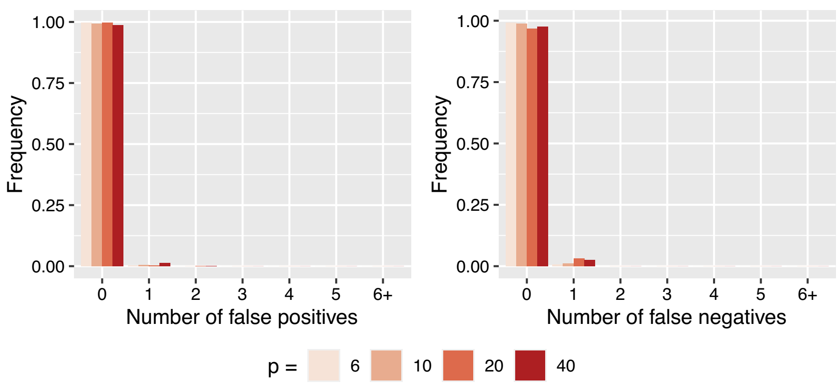

FIGURE 6. Relative frequency of false positives and false negatives for adaptive LASSO with LASSO as initial estimator, $\mathcal{U}\{-1,0,1\}$ distributed regressors and sample size $n = 10{,}000$.

$s_2^2 = (u - s_1)^2$ gives the solutions $\pm s_2 + s_1$ for u, which yields the bounds

$$u \in \big[|s_1 - s_2|, s_1 + s_2\big]$$

for the standard deviation $u = \sqrt{\mathbb{V}\mathrm{ar}(B_1)}$. Solving the equation at the beginning for the correlation gives $\rho = (s_2^2 - s_1^2 - u^2)/(2s_1 u)$, which ranges over the whole interval $[-1,1]$ if $s_2 > s_1$. If $s_1 \geq s_2$, the correlation must be negative, and maximizing the above expression for ρ over u yields $u = \sqrt{s_1^2 - s_2^2}$, and finally the upper bound in (2.3). □

Proof of Proposition 2.2. It is enough to show that all mixed moments of order n are identified from $n+1$ support points, the claim then follows by induction.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

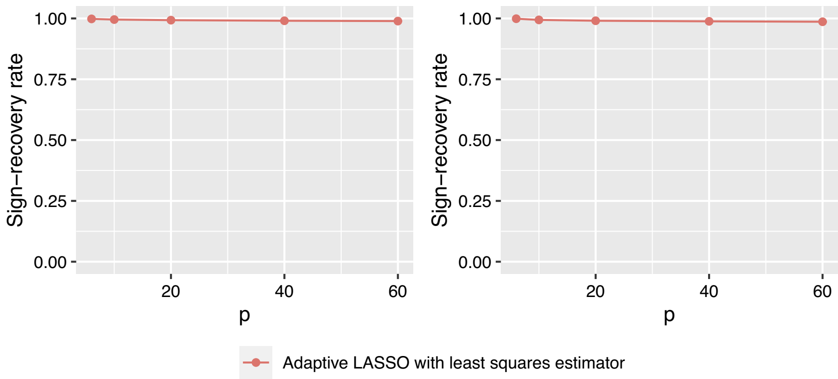

FIGURE 7. Left chart shows the sign-recovery rate for $\mathcal{U}[-1,1]$ distributed regressors, right one for $\mathcal{U}\{-1,0,1\}$ distributed regressors. The sample size is always $n = 5,000$.

By model (2.2), we obtain

$$\mathbb{E}\big[Y^n \,\big|\, W_1 = w\big] = \mathbb{E}\big[(B_0 + wB_1)^n\big] = \sum_{k=0}^{n} \binom{n}{k} w^k\, \mathbb{E}\big[B_0^{n-k} B_1^k\big].$$

If W has distinct support points $w_1, \ldots, w_{n+1}$, we obtain a linear system for the moments $\mathbb{E}\big[B_0^{n-k} B_1^k\big]$, $k = 0, \ldots, n$. Its design matrix satisfies

$$\begin{aligned}\det\left(\binom{n}{k-1} w_j^{k-1}\right)_{j,k\in\{1,\ldots,n+1\}} &= \prod_{l=0}^{n} \binom{n}{l} \det\left(w_j^{k-1}\right)_{j,k\in\{1,\ldots,n+1\}} \\ &= \prod_{l=0}^{n} \binom{n}{l} \prod_{1\le j<k\le n+1} (w_k - w_j) \neq 0,\end{aligned}$$

so that the solution is unique. In the last equation, we used the determinant of the Vandermonde matrix. □

5.1.2. *Proof of Theorem 2.3.* The proof needs some preparations. Recall that points $\mathbf{w_1}, \ldots, \mathbf{w_d} \in \mathbb{R}^{d-1}$ are said to be in general position if $\sum_{k=1}^{d} \alpha_k \mathbf{w_k} = \mathbf{0}_{d-1}$ for $\alpha_k \in \mathbb{R}$, $\sum_{k=1}^{d} \alpha_k = 0$, implies that $\alpha_1 = \cdots = \alpha_d = 0$. The following result is well-known.

LEMMA 5.1. *Points $\mathbf{w_1}, \ldots, \mathbf{w_d} \in \mathbb{R}^{d-1}$ are in general position if and only if one of the following conditions holds.*

(1). $\mathbf{w_2} - \mathbf{w_1}, \ldots, \mathbf{w_d} - \mathbf{w_1}$ *are linearly independent.*
(2). *For each $j \in \{1, \ldots, d\}$, the point $\mathbf{w_j}$ is not contained in $\left\{\sum_{k=1,k\neq j}^{d} \alpha_k \mathbf{w_k} \mid \sum_{k=1,k\neq j}^{d} \alpha_k = 1\right\}$, the hyperplane generated by $\mathbf{w_k}, k \neq j$.*

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

LEMMA 5.2. *If the support of* $\mathbf{W}$ *contains p points* $\mathbf{w_1}, \dots, \mathbf{w_p} \in \mathbb{R}^{p-1}$ *in general position, then the means* $\mu^* = \mathbb{E}[\mathbf{A}]$ *are identified.*

Proof of Lemma 5.2. The design matrix of the linear system $\mathbb{E}[Y \mid \mathbf{W} = \mathbf{w_j}] = \mathbb{E}[B_0] + \mathbf{w_j}^\top \mathbb{E}[\mathbf{B}]$, $j = 1, \dots, p$, has the same rank as the matrix

$$\begin{bmatrix} 1 & \mathbf{w_1}^\top \\ 0 & \mathbf{w_2}^\top - \mathbf{w_1}^\top \\ \vdots & \vdots \\ 0 & \mathbf{w_p}^\top - \mathbf{w_1}^\top \end{bmatrix},$$

which is invertible by Lemma 5.1. □

Proof of Theorem 2.3. Suppose that S is of full rank. Since S contains the matrix

$$\begin{bmatrix} 1 & 2\mathbf{w_1}^\top \\ \vdots & \vdots \\ 1 & 2\mathbf{w}_{\mathbf{p(p+1)/2}}^\top \end{bmatrix} \in \mathbb{R}^{\frac{p(p+1)}{2} \times p}$$

as a submatrix, in order for S to have full rank, it is necessary that this submatrix has rank p. This implies that there are p points among the support points $\mathbf{w_1}, \dots, \mathbf{w_{p(p+1)/2}}$ in general position, thus identifying the means by Lemma 5.2. Then, the linear system which determines $\mathbb{V}\mathrm{ar}(Y \mid \mathbf{W} = \mathbf{w_j})$ in terms of the entries of Σ^* has full-rank design matrix S, see (2.6), thus identifying Σ^* from the conditional variances.

Conversely, let $m = p(p+1)/2$. Suppose that the condition is not satisfied, then all support points $\mathbf{w}$ of $\mathbf{W}$ are such that the vectors $\mathrm{v}\big((1, \mathbf{w}^\top)^\top\big)$ are contained in an $(m-1)$-dimensional linear subspace V of $\mathbb{R}^m$. The $p \times p$-dimensional positive semi-definite matrices form a convex cone with interior consisting of positive definite matrices in the space of all $p \times p$-dimensional symmetric matrices. The image under the map vec is thus a convex cone $\mathcal{C} \subset \mathbb{R}^m$ with non-empty interior in $\mathbb{R}^m$.

Let $\mathbf{z}$ be a unit vector orthogonal to V, and let Z be the $p \times p$-dimensional symmetric matrix for which $\mathrm{vec}(Z) = \mathbf{z}$. Since the positive definite matrices are open in the space of all $p \times p$-dimensional symmetric matrices, given a positive definite matrix Σ^*, for small $\epsilon > 0$ the matrix $\Sigma_1 = \Sigma^* + \epsilon Z$ will still be positive definite, and hence a covariance matrix. Moreover, it is $\mathrm{vec}(\Sigma_1) = \mathrm{vec}(\Sigma^*) + \epsilon\, \mathrm{vec}(Z) = \mathrm{vec}(\Sigma^*) + \epsilon\, \mathbf{z}$ and $(1, \mathbf{w}^\top) Z (1, \mathbf{w}^\top)^\top = \mathrm{v}\big((1, \mathbf{w}^\top)^\top\big)^\top \mathbf{z} = 0$ for $\mathbf{w}$ in the support of $\mathbf{W}$ by construction. Hence, the conditional variances $(1, \mathbf{w}^\top)\, \Sigma^* \,(1, \mathbf{w}^\top)^\top$ and $(1, \mathbf{w}^\top)\, \Sigma_1 \,(1, \mathbf{w}^\top)^\top$ will be the same over the support of $\mathbf{W}$. Thus, for normally distributed $\mathbf{A} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma^*)$ or $\mathbf{A} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma_1)$, the conditional normal distributions of $Y \mid \mathbf{W} = \mathbf{w}$ will coincide, showing nonidentifiability. □

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

5.1.3. *Proof of Theorem 2.4.* For the proof of the theorem, we require the following lemma.

LEMMA 5.3. *Suppose that the support of* $\mathbf{W}$ *in (1.1) contains points satisfying the following properties.*

(1). *The* p *points* $\mathbf{w_1}, \ldots, \mathbf{w_p} \in \mathbb{R}^{p-1}$ *are in general position.*
(2). *For each* $j \in \{1, \ldots, p\}$*, there exist points* $\mathbf{w_{j,1}}, \ldots, \mathbf{w_{j,p-1}} \in \mathbb{R}^{p-1}$*, possibly equal to those in 1, such that:*
   - $\mathbf{w_j}, \mathbf{w_{j,1}}, \ldots, \mathbf{w_{j,p-1}}$ *are in general position,*
   - *for each* $j \in \{1, \ldots, p\}$*,* $k \in \{1, \ldots, p-1\}$*, there is a* $\mathbf{z_{j,k}} \in \mathbb{R}^{p-1}$ *for which* $\mathbf{w_j}, \mathbf{w_{j,k}}, \mathbf{z_{j,k}}$ *are all distinct but generate only a one-dimensional affine space, that is, are all contained in a line.*

Then the design matrix S *in (2.8) formed from all the points* $\mathbf{w_j}, \mathbf{w_{j,k}}, \mathbf{z_{j,k}}$ *has full rank* $p(p+1)/2$ *and hence, the mean vector* μ^* *and the covariance matrix* Σ^* *of the random coefficients* $\mathbf{A}$ *are identified.*

The minimal number of support points required in this lemma is $p + p(p-1)/2 = p(p+1)/2$, which corresponds to the number of free parameters in Σ^*. For the proof of Lemma 5.3, we first need the following two preliminary lemmas.

LEMMA 5.4. *Suppose that* Σ *is a* $p \times p$*-dimensional symmetric matrix and* $\mathbf{v_1}, \ldots, \mathbf{v_p} \in \mathbb{R}^p$ *is a known basis of* $\mathbb{R}^p$*. If* $\mathbf{v} \in \mathbb{R}^p$ *and* $\mathbf{v}^\top \Sigma\, \mathbf{v_j}$*,* $1 \le j \le p$*, is identified, then* $\mathbf{v}^\top \Sigma\, \mathbf{u}$ *is identified for any vector* $\mathbf{u} \in \mathbb{R}^p$*. In particular,* Σ *is identified from the values* $\mathbf{v_j}^\top \Sigma\, \mathbf{v_k}$*,* $1 \le j \le k \le p$*.*

Proof of Lemma 5.4. Given $\mathbf{u} \in \mathbb{R}^p$, we may write $\mathbf{u} = \sum_{j=1}^p \lambda_j \mathbf{v_j}$ with $\lambda_1, \ldots, \lambda_p \in \mathbb{R}$. Then

$$\mathbf{v}^\top \Sigma\, \mathbf{u} = \sum_{j=1}^{p} \lambda_j \mathbf{v}^\top \Sigma\, \mathbf{v_j},$$

showing the first claim. For the second, let $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ denote the kth unit vector in $\mathbb{R}^p$. By assumption, one may write $e_k = \sum_{j=1}^p \lambda_{k,j} \mathbf{v_j}$, where $\lambda_{k,j} \in \mathbb{R}$ and $1 \le k \le p$. Then

$$\Sigma_{kl} = e_k^\top \Sigma\, e_l = \sum_{j_1, j_2 = 1}^{p} \lambda_{k,j_1} \lambda_{l,j_2} \mathbf{v_{j_1}}^\top \Sigma\, \mathbf{v_{j_2}}.$$

The result follows from the assumptions and the symmetry of Σ. □

LEMMA 5.5. *Let* $\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3} \in \mathbb{R}^p$ *be such that each pair is linearly independent, but all three are linearly dependent, so that* $\mathbf{v_3} = \lambda_1 \mathbf{v_1} + \lambda_2 \mathbf{v_2}$*, where* $\lambda_1, \lambda_2 \neq 0$*.*

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

Then for a $p\times p$-dimensional symmetric matrix Σ, it holds that

$$\mathbf{v_1}^\top \Sigma\, \mathbf{v_2} = \frac{1}{2\lambda_1\lambda_2}\Big(\mathbf{v_3}^\top \Sigma\, \mathbf{v_3} - \lambda_1^2\, \mathbf{v_1}^\top \Sigma\, \mathbf{v_1} - \lambda_2^2\, \mathbf{v_2}^\top \Sigma\, \mathbf{v_2}\Big).$$

Proof of Lemma 5.5. Plug in the expression for $\mathbf{v_3}$ and compute the right side of the equation. □

Proof of Lemma 5.3. By Lemma 5.2 and the first assumption, the means μ^* are identified. Hence, we obtain the equations (2.5) or equivalently (2.6) with $\mathbf{w}$ ranging over the support points mentioned in the statement of the lemma. To show that the design matrix S in (2.8) has full rank $p(p+1)/2$, it suffices to show that from these equations one can uniquely solve for σ^*. To this end, from the second assumption, for $j\in\{1,\dots,p\}$ and $k\in\{1,\dots,p-1\}$, letting $\mathbf{v_1}=(1,\mathbf{w_j}^\top)^\top$, $\mathbf{v_2}=(1,\mathbf{w}_{\mathbf{j,k}}^\top)^\top$, and $\mathbf{v_3}=(1,\mathbf{z}_{\mathbf{j,k}}^\top)^\top$ in Lemma 5.5, we identify $(1,\mathbf{w_j}^\top)\,\Sigma^*\,(1,\mathbf{w}_{\mathbf{j,k}}^\top)^\top$. Since $(1,\mathbf{w_j}^\top)\,\Sigma^*\,(1,\mathbf{w_j}^\top)^\top$ are also identified, from the first part in Lemma 5.4, we identify $(1,\mathbf{w_j}^\top)\,\Sigma^*\,(1,\mathbf{w_l}^\top)^\top$, $j,l\in\{1,\dots,p\}$. Hence, from the second part of that lemma and the first assumption Σ^* itself is identified. □

Proof of Theorem 2.4. For the sufficiency, suppose that the support of W_j contains $\{w_{j,k},\ k=1,2,3\}$, $j=1,\dots,p-1$. We apply Lemma 5.3 with:

- $\mathbf{w_j}=(w_{1,1},\dots,w_{j-1,1},w_{j,2},w_{j+1,1},\dots,w_{p-1,1})^\top$, $j=1,\dots,p-1$, and $\mathbf{w_p}=(w_{1,1},\dots,w_{p-1,1})^\top$,
- for $j\in\{1,\dots,p-1\}$, let $\mathbf{w_{j,k}}$, $k\in\{1,\dots,p-1\}$, $k\neq j$, enumerate the points having kth coordinate $w_{k,2}$ and jth coordinate $w_{j,2}$, otherwise coordinates $w_{i,1}$, the corresponding $\mathbf{z_{j,k}}$ having kth coordinate $w_{k,3}$, jth coordinate $w_{j,2}$, otherwise coordinates $w_{i,1}$. Furthermore, let $\mathbf{w_{j,j}}=\mathbf{w_p}$, and let $\mathbf{z_{j,j}}$ have jth coordinate $w_{j,3}$, otherwise $w_{i,1}$,
- let $\mathbf{w_{p,k}}=\mathbf{w_k}$, $k\in\{1,\dots,p-1\}$, and $\mathbf{z_{p,k}}=(w_{1,1},\dots,w_{k-1,1},w_{k,3},w_{k+1,1},\dots,w_{p-1,1})^\top$.

The requirements of the lemma are then easily checked by applying Lemma 5.1(1). The necessity of at least three support points in each coordinate, if Σ^* has full rank, is clear from Example 2.1. □

5.1.4. *Proofs for Section 2.2.*

Proof of Proposition 2.5. The claims in (1) are clear. For (2), in both cases, from $\mathbb{V}\mathrm{ar}(B_0)=\mathbb{V}\mathrm{ar}(B_0+B_1)$, we get that

$$\mathbb{V}\mathrm{ar}(B_1) = -2\,\mathbb{C}\mathrm{ov}(B_0,B_1).$$

Since the covariance matrix $\mathbb{C}\mathrm{ov}\big((B_0,B_1,\mathbf{B_2}^\top)^\top\big)$ is positive semi-definite, setting $s=\mathbb{V}\mathrm{ar}(B_1)$, we hence require

$$s\,(z^2 - z v_1) + \mathbf{v}^\top \mathbb{C}\mathrm{ov}\big((B_0,\mathbf{B_2}^\top)^\top\big)\,\mathbf{v} \geq 0 \tag{5.1}$$

for any $z\in\mathbb{R}$, $\mathbf{v}=(v_1,\dots,v_{p-1})^\top\in\mathbb{R}^{p-1}$.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

(a) Choose $\mathbf{v} \in \mathbb{R}^{p-1}$ in the kernel of $\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)$ for which $v_1 > 0$. If we assume $s > 0$, then for $0 < z < v_1$ the form in (5.1) would be negative. Hence $s = 0$ in this case.

(b) If $\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)$ has full rank, then for the minimal eigenvalue $\lambda_{\min} > 0$ of $\mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)$, we have that

$$\mathbf{v}^\top \mathbb{C}\mathrm{ov}\big((B_0, \mathbf{B_2}^\top)^\top\big)\, \mathbf{v} \geq \lambda_{\min} \, \|v\|_2^2 .$$

Therefore, the form in (5.1) is positive definite for $0 \leq s \leq 4\,\lambda_{\min}$. □

Proof of Theorem 2.6. By (2.12), the symmetric multilinear form

$$u(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \mathbb{E}\big[(\mathbf{A}^\top \mathbf{v}_1) \cdot \ldots \cdot (\mathbf{A}^\top \mathbf{v}_k)\big], \qquad \mathbf{v}_j \in \mathbb{R}^p, \quad j = 1, \ldots, k,$$

is identified over the diagonal

$$\tilde{u}(\mathbf{v}) = u(\mathbf{v}, \ldots, \mathbf{v})$$

for $\mathbf{v}^\top = (1, \mathbf{w}^\top)$ with $\mathbf{w}$ in the support of $\mathbf{W}$. We shall show that the symmetric multilinear form u is identified. Then, inserting unit vectors $(0, \ldots, 0, 1, 0, \ldots, 0)$ yields the kth-order mixed moments.

By multilinearity, it suffices to show that u is identified over a basis of $\mathbb{R}^p$, that is, there exists a basis $\mathbf{v_1}, \ldots, \mathbf{v_p}$ of $\mathbb{R}^p$ such that $u(\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_k})$ is identified for all choices $i_j \in \{1, \ldots, p\}$. To this end, we use the polarization formula for symmetric multilinear forms (Thomas, 2014, Formula (7)), which we write as

$$u(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \sum_{\{i_1, \ldots i_j\} \subseteq \{1, \ldots, k\}} j^k \, \tilde{u}\big((\mathbf{v}_{i_1} + \cdots + \mathbf{v}_{i_j})/j\big). \tag{5.2}$$

Now for $\mathbf{w}_1, \ldots, \mathbf{w}_p$ as in the assumption of the theorem, the vectors $(1, \mathbf{w}_i^\top)$, $i = 1, \ldots, p$ are linearly independent by the proof of Lemma 5.2, and for $j \in \{1, \ldots, k\}$ and $i_1, \ldots, i_j \in \{1, \ldots, p\}$, we have

$$\big((1, \mathbf{w}_{i_1}^\top) + \cdots + (1, \mathbf{w}_{i_j}^\top)\big)/j = \big(1, (\mathbf{w}_{i_1} + \cdots + \mathbf{w}_{i_j})^\top / j\big)$$

with $(\mathbf{w}_{i_1} + \cdots + \mathbf{w}_{i_j})^\top / j$ in the support of $\mathbf{W}$. Hence, the terms on the right of (5.2) are identified for k (not necessarily distinct) vectors $(1, \mathbf{w}_i^\top)$, thus also the form u. □

5.2. Proofs for Section 3.2

5.2.1. *Proof of Theorem 3.6.* Consider the following decomposition of the error term (3.3):

$$\varepsilon_n^\sigma = \delta_n + \zeta_n + \xi_n \tag{5.3}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

with

$$\delta_n := \Big(\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(D_1-\Sigma^*),\ldots,\mathrm{v}(\mathbf{X_n})^\top \mathrm{vec}(D_n-\Sigma^*)\Big)^\top, \tag{5.4}$$

$$\zeta_n := \Big(\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(E_n),\ldots,\mathrm{v}(\mathbf{X_n})^\top \mathrm{vec}(E_n)\Big)^\top,$$

$$\xi_n := \Big(\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(F_{n,1}),\ldots,\mathrm{v}(\mathbf{X_n})^\top \mathrm{vec}(F_{n,n})\Big)^\top.$$

The matrices $D_1,\ldots,D_n,E_n,F_{n,1},\ldots,F_{n,n}$ are defined in (3.1) and (3.2). For the proof of Theorem 3.6, we need the following auxiliary lemmas.

LEMMA 5.6. *Set* $Z_n^{\sigma,1} = \frac{1}{n}\big(\mathbb{X}_n^\sigma\big)^\top \delta_n$, *then* $\|Z_n^{\sigma,1}\|_2 = \mathcal{O}_{\mathbb{P}}\big(p/\sqrt{n}\big)$.

Proof of Lemma 5.6. It is

$$\mathbb{E}\Big[\big\|Z_n^{\sigma,1}\big\|_2^2\Big] = \frac{1}{n^2}\,\mathbb{E}\Big[\delta_n^\top \mathbb{X}_n^\sigma\,(\mathbb{X}_n^\sigma)^\top \delta_n\Big] = \frac{1}{n^2}\,\mathbb{E}\Big[\mathrm{trace}\big((\mathbb{X}_n^\sigma)^\top \delta_n\,\delta_n^\top \mathbb{X}_n^\sigma\big)\Big]$$
$$= \frac{1}{n^2}\,\mathbb{E}\Big[\mathrm{trace}\big((\mathbb{X}_n^\sigma)^\top \mathbb{E}\big[\delta_n\,\delta_n^\top \,\big|\, \mathbb{X}_n^\sigma\big]\mathbb{X}_n^\sigma\big)\Big] = \frac{1}{n}\,\mathrm{trace}\bigg(\mathbb{E}\Big[\frac{1}{n}\,(\mathbb{X}_n^\sigma)^\top \Omega_n^\sigma\,\mathbb{X}_n^\sigma\Big]\bigg),$$

where $\Omega_n^\sigma = \mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big)$ is a diagonal matrix with entries $\mathrm{v}(\mathbf{X_1})^\top \Psi^*\,\mathrm{v}(\mathbf{X_1}),\ldots,$ $\mathrm{v}(\mathbf{X_n})^\top \Psi^*\,\mathrm{v}(\mathbf{X_n})$ (see Lemma 6.2 in the Supplementary Material). It is obvious that

$$\mathbb{E}\Big[\frac{1}{n}\,(\mathbb{X}_n^\sigma)^\top \Omega_n^\sigma\,\mathbb{X}_n^\sigma\Big] = \mathrm{B}^\sigma,$$

and hence, we obtain by Assumption (A7), the estimate

$$\mathbb{E}\Big[\big\|Z_n^{\sigma,1}\big\|_2^2\Big] = \frac{\mathrm{trace}\big(\mathrm{B}^\sigma\big)}{n} \le \frac{\lambda_{\max}\big(\mathrm{B}^\sigma\big)\,p(p+1)}{2n} \le \frac{c_{\mathrm{B}^\sigma,\mathrm{u}}\,p(p+1)}{2n}.$$

Markov's inequality implies the assertion. □

LEMMA 5.7. *Set* $Z_n^{\sigma,2} = \frac{1}{n}\big(\mathbb{X}_n^\sigma\big)^\top \zeta_n$, *then* $\|Z_n^{\sigma,2}\|_2 = \mathcal{O}_{\mathbb{P}}\,(p/n)$.

Proof of Lemma 5.7. It is

$$\big\|Z_n^{\sigma,2}\big\|_2 = \bigg\|\frac{1}{n}\sum_{i=1}^n \big(e_i^\top \zeta_n\big)\mathrm{v}\big(\mathbf{X_i}\big)\bigg\|_2 = \bigg\|\frac{1}{n}\sum_{i=1}^n \Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(E_n\big)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\bigg\|_2$$
$$\le \bigg\|\frac{1}{n}\sum_{i=1}^n \mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top\bigg\|_{\mathrm{M},2}\,\big\|\mathrm{vec}\big(E_n\big)\big\|_2.$$

We multiply each entry of E_n, which is not on the diagonal, with $\sqrt{2}$ and denote the resulting matrix by $\widetilde{E}_n$. Then it is clear that $\big\|\mathrm{vec}\big(E_n\big)\big\|_2 \le \big\|\mathrm{vec}\big(\widetilde{E}_n\big)\big\|_2$ and $\big\|\mathrm{vec}\big(\widetilde{E}_n\big)\big\|_2 = \|E_n\|_{\mathrm{F}}$. Moreover, recall that $E_n = \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top$ is a

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

rank-one matrix; and hence $\|E_n\|_{\mathrm{F}} \leq \|\widehat{\mu}_n - \mu^*\|_2^2$. Hence, we obtain

$$\frac{n}{p}\left\|Z_n^{\sigma,2}\right\|_2 \leq \left\|\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right\|_{\mathrm{M},2} \frac{n}{p}\left\|\widehat{\mu}_n - \mu^*\right\|_2^2 = \mathcal{O}_{\mathbb{P}}(1)\,\mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1)$$

since $\sqrt{n/p}\,\|\widehat{\mu}_n - \mu^*\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. □

LEMMA 5.8. *Set* $Z_n^{\sigma,3} = \frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \xi_n$, *then* $\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(p^{3/2}/n^{5/8} + p^2/n^{3/4}\right)$. *In particular, we obtain by Assumption (A8), the convergence* $\sqrt{n}/p\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(p^{1/2}/n^{1/8} + p/n^{1/4}\right) = \mathrm{o}_{\mathbb{P}}(1)$.

Proof of Lemma 5.8. It is

$$\begin{aligned}\left\|Z_n^{\sigma,3}\right\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^n \left(e_i^\top \xi_n\right)\mathrm{v}\left(\mathbf{X_i}\right)\right\|_2 = 2\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\mathbf{X_i}^\top\left(\mu^* - \widehat{\mu}_n\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\right\|_2 \\ &\leq 2\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top\right\|_{\mathrm{M},2}\left\|\mu^* - \widehat{\mu}_n\right\|_2, \end{aligned} \tag{5.5}$$

and we have again $\sqrt{n/p}\,\|\widehat{\mu}_n - \mu^*\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. Moreover, let

$$\mathcal{T}_n(\tau_n) = \bigcap_{i=1}^n \left\{\left\|\mathbf{A_i} - \mu^*\right\|_2 \leq \tau_n\right\}$$

with $\tau_n > 0$, then we obtain

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top\right\|_{\mathrm{M},2} \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)}\right\|_{\mathrm{M},2} + \left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n^c(\tau_n)}\right\|_{\mathrm{M},2} \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^n \Big(\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)} - \mathbb{E}\Big[\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big)\right\|_{\mathrm{M},2} \\ &+ \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\right\|_{\mathrm{M},2} + \left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n^c(\tau_n)}\right\|_{\mathrm{M},2}. \end{aligned} \tag{5.6}$$

For the first term of the sum, we get

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^n \Big(\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)} - \mathbb{E}\Big[\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)\mathbf{X_i}^\top \mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big)\right\|_{\mathrm{M},2} \\ &\quad= \sup_{\substack{u_1 \in \mathbb{R}^{p(p+1)/2},\, u_2 \in \mathbb{R}^p, \\ \|u_1\|_2, \|u_2\|_2 \leq 1}} \frac{1}{n}\sum_{i=1}^n \Big(u_1^\top \mathrm{v}\left(\mathbf{X_i}\right)\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\mathbf{X_i}^\top u_2\Big)\mathbb{1}_{\mathcal{T}_n(\tau_n)} \\ &\qquad\qquad - \mathbb{E}\Big[u_1^\top \mathrm{v}\left(\mathbf{X_i}\right)\Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\mathbf{X_i}^\top u_2\Big)\mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big). \end{aligned} \tag{5.7}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

For the second factors in brackets, we obtain by the definition of the half-vectorization vec in (2.4) and the vector transformation v in (2.7), the equation

$$\begin{aligned}\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\mathbf{X_i}^\top u_2 &= \mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)u_2^\top\mathbf{X_i} = \frac{1}{2}\mathbf{X_i}^\top\Big((\mathbf{A_i}-\mu^*)u_2^\top + u_2(\mathbf{A_i}-\mu^*)^\top\Big)\mathbf{X_i}\\ &= \frac{1}{2}\mathrm{v}(\mathbf{X_i})^\top \mathrm{vec}\Big((\mathbf{A_i}-\mu^*)u_2^\top + u_2(\mathbf{A_i}-\mu^*)^\top\Big).\end{aligned}$$

For the half-vectorization we can argue analogously as in Lemma 5.7 and bound its Euclidean norm by

$$\begin{aligned}\frac{1}{2}\left\|\mathrm{vec}\Big((\mathbf{A_i}-\mu^*)u_2^\top + u_2(\mathbf{A_i}-\mu^*)^\top\Big)\right\|_2 &\le \frac{1}{2}\left\|(\mathbf{A_i}-\mu^*)u_2^\top + u_2(\mathbf{A_i}-\mu^*)^\top\right\|_F\\ &\le \left\|(\mathbf{A_i}-\mu^*)u_2^\top\right\|_F \le \|\mathbf{A_i}-\mu^*\|_2\,\|u_2\|_2\,.\end{aligned}$$

Suppose that $\mathrm{v}(\mathbf{X})$ is sub-Gaussian with variance proxy $\tau^2_{\mathrm{v}(\mathbf{X})}$, then conditionally on the coefficients $\mathbf{A_1},\dots,\mathbf{A_n}$, which are independent of the regressors $\mathbf{X_1},\dots,\mathbf{X_n}$, we obtain in (5.7) for each u_1,u_2, a sum of centered and independent products consisting of two sub-Gaussian random variables with variance proxies $\tau^2_{\mathrm{v}(\mathbf{X})}\|u_1\|_2^2 \le \tau^2_{\mathrm{v}(\mathbf{X})}$ and $\tau^2_{\mathrm{v}(\mathbf{X})}\|\mathbf{A_i}-\mu^*\|_2^2\,\|u_2\|_2^2 \le \tau^2_{\mathrm{v}(\mathbf{X})}\,\tau_n^2$. Hence, in particular, the products are sub-exponential with parameter bounded by $C_1\,\tau^2_{\mathrm{v}(\mathbf{X})}\,\tau_n$ for a universal constant $C_1>0$ (see Vershynin (2018, Lem. 2.7.7)). Following the covering argument and applying the tail bound of sub-exponential random variables as in Wainwright (2019, Thm. 6.5) leads to

$$\begin{aligned}\mathbb{P}\Bigg(\Bigg\|\frac{1}{n}\sum_{i=1}^n\Big(\big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\big)\mathrm{v}(\mathbf{X_i})\mathbf{X_i}^\top\mathbb{1}_{\mathcal{T}_n(\tau_n)}\\ -\mathbb{E}\Big[\big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\big)\mathrm{v}(\mathbf{X_i})\mathbf{X_i}^\top\mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big)\Bigg\|_{M,2} \ge C_2\,\tau_n\frac{p}{\sqrt{n}}\Bigg) \le C_3\exp\big(-C_4p^2\big)\end{aligned}$$

for universal constants $C_2,C_3,C_4>0$. Furthermore, we obtain for the third term in the sum in (5.6), the estimate

$$\mathbb{P}\Bigg(\Bigg\|\frac{1}{n}\sum_{i=1}^n\big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\big)\mathrm{v}(\mathbf{X_i})\mathbf{X_i}^\top\mathbb{1}_{\mathcal{T}_n^c(\tau_n)}\Bigg\|_{M,2}\ge t\Bigg) \le \mathbb{P}\big(\mathcal{T}_n^c(\tau_n)\big) \le C_5\frac{p^2n}{\tau_n^4}$$

for $t>0$, since

$$\begin{aligned}\mathbb{P}\big(\mathcal{T}_n^c(\tau_n)\big) &= \mathbb{P}\Bigg(\bigcup_{i=1}^n\Big\{\|\mathbf{A_i}-\mu^*\|_2>\tau_n\Big\}\Bigg) \le \sum_{i=1}^n\mathbb{P}\Big(\|\mathbf{A_i}-\mu^*\|_2>\tau_n\Big)\\ &= n\,\mathbb{P}\Big(\|\mathbf{A}-\mu^*\|_2>\tau_n\Big) \le \frac{n\,\mathbb{E}\big[\|\mathbf{A}-\mu^*\|_2^4\big]}{\tau_n^4}\\ &= \frac{n}{\tau_n^4}\sum_{k,l=1}^p\mathbb{E}\Big[\big(A_k-\mu_k^*\big)^2\big(A_l-\mu_l^*\big)^2\Big] \le C_5\frac{p^2n}{\tau_n^4} \qquad \textbf{(5.8)}\end{aligned}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

holds for a positive constant $C_5 > 0$ by Assumption (A4). Moreover, we obtain

$$\mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\,\mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big] = \mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\,\big(-\mathbb{1}_{\mathcal{T}_n^c(\tau_n)}\big)\Big]$$

because

$$\mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\Big] = \mathbb{E}\Big[\Big(\mathbf{X_i}^\top\mathbb{E}\big[\mathbf{A_i}-\mu^*\,\big|\,\mathbf{X_i}\big]\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\Big] = \mathbf{0}_{\frac{p(p+1)}{2}\times p}$$

is satisfied by the independence of $\mathbf{X_i}$ and $\mathbf{A_i}$. Hence, the Cauchy–Schwarz inequality implies for the second term in (5.6), the estimate

$$\begin{aligned}
&\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\,\mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big\|_{\mathrm{M},2}\\
&\quad= \sup_{\substack{u_1\in\mathbb{R}^{p(p+1)/2},\,u_2\in\mathbb{R}^p,\\ \|u_1\|_2,\|u_2\|_2\le 1}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)u_1^\top\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top u_2\,\big(-\mathbb{1}_{\mathcal{T}_n^c(\tau_n)}\big)\Big]\\
&\quad\le \sup_{\substack{u_1\in\mathbb{R}^{p(p+1)/2},\,u_2\in\mathbb{R}^p,\\ \|u_1\|_2,\|u_2\|_2\le 1}} \left(\mathbb{E}\Big[\Big(\mathbf{X}^\top(\mathbf{A}-\mu^*)\Big)^2\Big(u_1^\top\mathrm{v}\big(\mathbf{X}\big)\Big)^2\big(\mathbf{X}^\top u_2\big)^2\Big]\,\mathbb{P}\big(\mathcal{T}_n^c(\tau_n)\big)\right)^{\frac{1}{2}}.
\end{aligned}$$

Further,

$$\begin{aligned}
&\sup_{\substack{u_1\in\mathbb{R}^{p(p+1)/2},\,u_2\in\mathbb{R}^p,\\ \|u_1\|_2,\|u_2\|_2\le 1}} \mathbb{E}\Big[\Big(\mathbf{X}^\top(\mathbf{A}-\mu^*)\Big)^2\Big(u_1^\top\mathrm{v}\big(\mathbf{X}\big)\Big)^2\big(\mathbf{X}^\top u_2\big)^2\Big]^{\frac{1}{2}}\\
&\qquad\le \sup_{\substack{u_1\in\mathbb{R}^{p(p+1)/2},\,u_2\in\mathbb{R}^p,\\ \|u_1\|_2,\|u_2\|_2\le 1}} \mathbb{E}\Big[\Big(\mathbf{X}^\top(\mathbf{A}-\mu^*)\Big)^4\Big]^{\frac{1}{4}}\,\mathbb{E}\Big[\Big(u_1^\top\mathrm{v}\big(\mathbf{X}\big)\Big)^8\Big]^{\frac{1}{8}}\,\mathbb{E}\Big[\big(\mathbf{X}^\top u_2\big)^8\Big]^{\frac{1}{8}}\\
&\qquad\le C_6\,\mathbb{E}\Big[\mathbb{E}\Big[\Big(\mathbf{X}^\top(\mathbf{A}-\mu^*)\Big)^4\,\Big|\,\mathbf{A}\Big]\Big]^{\frac{1}{4}} \le C_7\,\mathbb{E}\Big[\big\|\mathbf{A}-\mu^*\big\|_2^4\Big]^{\frac{1}{4}} \le C_8\sqrt{p},
\end{aligned}$$

where $C_6, C_7, C_8 > 0$ are positive constants, since $\mathrm{v}\big(\mathbf{X}\big)$ and $\mathbf{X}$ are sub-Gaussian and hence their moments exist (see Wainwright (2019, Thm. 2.6)). This implies together with (5.8), the upper bound

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\,\mathbb{1}_{\mathcal{T}_n(\tau_n)}\Big]\Big\|_{\mathrm{M},2} \le C_8\sqrt{p}\Big(\mathbb{P}\big(\mathcal{T}_n^c(\tau_n)\big)\Big)^{\frac{1}{2}} \le C_8\sqrt{C_5}\,\frac{p^{3/2}\sqrt{n}}{\tau_n^2}.$$

So all in all collecting, the terms leads to

$$\begin{aligned}
&\mathbb{P}\left(\Big\|\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{X_i}^\top(\mathbf{A_i}-\mu^*)\Big)\mathrm{v}\big(\mathbf{X_i}\big)\,\mathbf{X_i}^\top\Big\|_{\mathrm{M},2} \ge 2\,C_2\,\frac{\tau_n p}{\sqrt{n}} + 2\,C_8\sqrt{C_5}\,\frac{p^{3/2}\sqrt{n}}{\tau_n^2}\right)\\
&\qquad\qquad \le C_5\,\frac{p^2 n}{\tau_n^4} + C_3\exp\big(-C_4\,p^2\big).
\end{aligned}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

If $p^2 n/\tau_n^4 \to 0$ is satisfied, we obtain by (5.5), the rate

$$\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{\tau_n p^{3/2}}{n} + \frac{p^2}{\tau_n^2}\right).$$

Let $\tau_n = n^{3/8}$, then $p^2 n/\tau_n^4 = p^2/\sqrt{n} \to 0$ by Assumption (A8), and

$$\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{p^{3/2}}{n^{5/8}} + \frac{p^2}{n^{3/4}}\right).$$

□

Remark 5.9. Suppose that the vector $\mathbf{A} - \mu^*$ is sub-Gaussian with variance proxy $\tau_{\mathbf{A}}^2$, then we can use in the proof of Lemma 5.8, the estimate

$$\mathbb{P}\Big(\left\|\mathbf{A} - \mu^*\right\|_2 > \tau_n\Big) = \Big(\sup_{v \in \mathbb{R}^p,\, \|v\|_2 \le 1} v^\top(\mathbf{A} - \mu^*) > \tau_n\Big) \le 6^p \exp\left(-\frac{\tau_n^2}{8\,\tau_{\mathbf{A}}^2}\right),$$

(see Rigollet and Hütter (2019, Thm. 1.19)). Let $\tau_n = \sqrt{(C_p + \log(6))8\,\tau_{\mathbf{A}}^2 p}$ with $C_p > 0$, then

$$n\mathbb{P}\Big(\left\|\mathbf{A} - \mu^*\right\|_2 > \tau_n\Big) \le n \exp\left(-\frac{(C_p + \log(6))8\,\tau_{\mathbf{A}}^2 p}{8\,\tau_{\mathbf{A}}^2} + p\log(6)\right) = n \exp\big(-C_p p\big) \to 0.$$

Hence $\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(p^2/n\right)$ and, in particular, $\sqrt{n}/p\left\|Z_n^{\sigma,3}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(p/\sqrt{n}\right) = o_{\mathbb{P}}(1)$.

Proof of Theorem 3.6. We shall use the primal-dual witness characterization of the adaptive LASSO in Lemma 7.1 in the Section 7 of the Supplementary Material, to prove the sign-consistency (3.10). We obtain by Assumption (A5) and Wainwright (2019, Thm. 6.5) that

$$\left\|\frac{1}{n}(\mathbb{X}_n^\sigma)^\top \mathbb{X}_n^\sigma - \mathrm{C}^\sigma\right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{p(p+1)/n}\right) = \mathcal{O}_{\mathbb{P}}\left(p/\sqrt{n}\right),$$

which implies together with the Assumptions (A6) and (A8), the invertibility of the Gram matrix for large n, and hence by Loh and Wainwright (2017, Lem. 11), we get also

$$\left\|\left(\frac{1}{n}(\mathbb{X}_n^\sigma)^\top \mathbb{X}_n^\sigma\right)^{-1} - \left(\mathrm{C}^\sigma\right)^{-1}\right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}}\left(p/\sqrt{n}\right).$$

Furthermore, basic properties of the ℓ_2 operator norm and Assumption (A6) lead to

$$\begin{aligned}
&\left\|\left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma \left(\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1}\right\|_{\mathrm{M},2} \\
&\quad = \left\|\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1}\right\|_{\mathrm{M},2}
\end{aligned}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

$$\leq \left(\left\| \mathrm{C}^{\sigma}_{S^c_\sigma S_\sigma} \right\|_{\mathrm{M},2} + \left\| \frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} - \mathrm{C}^{\sigma}_{S^c_\sigma S_\sigma} \right\|_{\mathrm{M},2} \right) \cdot \left\| \left(\frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} - \left(\mathrm{C}^{\sigma}_{S_\sigma S_\sigma}\right)^{-1} \right\|_{\mathrm{M},2} + \left\| \frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} - \mathrm{C}^{\sigma}_{S^c_\sigma S_\sigma} \right\|_{\mathrm{M},2} \left\| \left(\mathrm{C}^{\sigma}_{S_\sigma S_\sigma}\right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}}\left(p/\sqrt{n}\right).$$

In particular, this implies

$$\left\| \left(\frac{1}{n} (\mathbb{X}^{\sigma}_n)^\top \mathbb{X}^{\sigma}_n \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}}(1), \qquad \left\| \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}}(1). \tag{5.9}$$

Moreover, let $\widehat{\sigma}^{\text{init}}_{n,\min} := \min_{k \in S_\sigma} |\widehat{\sigma}^{\text{init}}_{n,k}|$, then

$$\left| \frac{\widehat{\sigma}^{\text{init}}_{n,\min} - \sigma^*_{\min}}{\sigma^*_{\min}} \right| \leq \frac{1}{\sigma^*_{\min}} \left\| \widehat{\sigma}^{\text{init}}_n - \sigma^* \right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{p}{\sigma^*_{\min}\sqrt{n}} \right) = \mathrm{o}_{\mathbb{P}}(1)$$

since $\sqrt{n}/p \left\| \widehat{\sigma}^{\text{init}}_n - \sigma^* \right\|_2 = \mathcal{O}_{\mathbb{P}}(1)$ and $p/(\sigma^*_{\min}\sqrt{n}) \to 0$. This implies

$$\left(1 + \frac{\widehat{\sigma}^{\text{init}}_{n,\min} - \sigma^*_{\min}}{\sigma^*_{\min}} \right)^{-1} = \mathcal{O}_{\mathbb{P}}(1)$$

(see van der Vaart (1998, Section 2.2)). Hence, we obtain

$$\frac{\sqrt{n}}{p} \left\| \lambda^{\sigma}_n \left(\frac{1}{|\widehat{\sigma}^{\text{init}}_{n,S_\sigma}|} \odot \operatorname{sign}\left(\sigma^*_{S_\sigma}\right) \right) \right\|_2 \leq \frac{\sqrt{n}\,\lambda^{\sigma}_n}{p} \left\| \frac{1}{|\widehat{\sigma}^{\text{init}}_{n,S_\sigma}|} \right\|_2 \leq \frac{\sqrt{s_\sigma\, n}\,\lambda^{\sigma}_n}{p} \left\| \frac{1}{|\widehat{\sigma}^{\text{init}}_{n,S_\sigma}|} \right\|_\infty = \frac{\sqrt{s_\sigma\, n}\,\lambda^{\sigma}_n}{p} \left(\widehat{\sigma}^{\text{init}}_{n,\min}\right)^{-1} = \mathrm{o}_{\mathbb{P}}(1) \tag{5.10}$$

since $\sqrt{s_\sigma\, n}\,\lambda^{\sigma}_n / (\sigma^*_{\min} p) \to 0$ by assumption. It follows that

$$\begin{aligned}
&\frac{\sqrt{n}}{p} \left\| \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \left(\lambda^{\sigma}_n \left(\frac{1}{|\widehat{\sigma}^{\text{init}}_{n,S_\sigma}|} \odot \operatorname{sign}\left(\sigma^*_{S_\sigma}\right) \right) \right) + \frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathrm{P}_{\mathbb{X}^{\sigma}_{n,S_\sigma}} \varepsilon^{\sigma}_n \right\|_2 \\
&\leq \left\| \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \right\|_{\mathrm{M},2} \frac{\sqrt{n}}{p} \left\| \lambda^{\sigma}_n \left(\frac{1}{|\widehat{\sigma}^{\text{init}}_{n,S_\sigma}|} \odot \operatorname{sign}\left(\sigma^*_{S_\sigma}\right) \right) \right\|_2 \\
&\quad + \frac{\sqrt{n}}{p} \left\| \frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \varepsilon^{\sigma}_n \right\|_2 + \left\| \left(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \right\|_{\mathrm{M},2} \frac{\sqrt{n}}{p} \left\| \frac{1}{n} \left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^\top \varepsilon^{\sigma}_n \right\|_2 \\
&= \mathcal{O}_{\mathbb{P}}(1)\,\mathrm{o}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1),
\end{aligned} \tag{5.11}$$

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

by Lemmas 5.6–5.8 and (5.9), where

$$\mathrm{P}_{\mathbb{X}^{\sigma}_{\mathrm{n},S_\sigma}} = \mathrm{I}_n - \mathbb{X}^{\sigma}_{n,S_\sigma}\left(\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{-1}\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}.$$

Furthermore, it is

$$\frac{\left|\widehat{\sigma}^{\,\mathrm{init}}_{n,k}\right|}{\lambda^{\sigma}_n} \le \frac{\left\|\widehat{\sigma}^{\,\mathrm{init}}_{n,S^c_\sigma}\right\|_2}{\lambda^{\sigma}_n} = \frac{\left\|\widehat{\sigma}^{\,\mathrm{init}}_{n,S^c_\sigma} - \sigma^*_{S^c_\sigma}\right\|_2}{\lambda^{\sigma}_n} \le \frac{\left\|\widehat{\sigma}^{\,\mathrm{init}}_{n} - \sigma^*\right\|_2}{\lambda^{\sigma}_n} = \frac{\sqrt{n}/p\left\|\widehat{\sigma}^{\,\mathrm{init}}_{n} - \sigma^*\right\|_2}{(\sqrt{n}/p)\,\lambda^{\sigma}_n}$$

for all $k \in S^c$. The condition $\sqrt{n}/p\left\|\widehat{\sigma}^{\,\mathrm{init}}_{n} - \sigma^*\right\|_2 = \mathcal{O}_{\mathbb{P}}(1)$ together with $n\lambda^{\sigma}_n/p^2 \to \infty$ implies the convergence

$$\frac{\left|\widehat{\sigma}^{\,\mathrm{init}}_{n,k}\right|}{(\sqrt{n}/p)\,\lambda^{\sigma}_n} = \frac{1}{n\lambda^{\sigma}_n/p^2}\,\mathcal{O}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1).$$

Hence, it follows by (5.11) that the first condition (7.1) of Lemma 7.1 in the Supplementary Material is satisfied with high probability for a sufficient large sample size n. Furthermore, let

$$\widetilde{\sigma}_{n,S_\sigma} = \sigma^*_{S_\sigma} + \left(\frac{1}{n}\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{-1}\left(\frac{1}{n}\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}\varepsilon^{\sigma}_n - \lambda^{\sigma}_n\left(\frac{1}{|\widehat{\sigma}^{\,\mathrm{init}}_{n,S_\sigma}|}\odot \mathrm{sign}\big(\sigma^*_{S_\sigma}\big)\right)\right).$$

Then, we obtain

$$\begin{aligned}\frac{\sqrt{n}}{p}\left\|\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\right\|_2 &\le \left\|\left(\frac{1}{n}\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{-1}\right\|_{\mathrm{M},2}\left(\frac{\sqrt{n}}{p}\left\|\frac{1}{n}\left(\mathbb{X}^{\sigma}_{n,S_\sigma}\right)^{\top}\varepsilon^{\sigma}_n\right\|_2\right.\\ &\qquad\left. + \frac{\sqrt{n}}{p}\left\|\lambda^{\sigma}_n\left(\frac{1}{|\widehat{\sigma}^{\,\mathrm{init}}_{n,S_\sigma}|}\odot \mathrm{sign}\big(\sigma^*_{S_\sigma}\big)\right)\right\|_2\right)\\ &= \mathcal{O}_{\mathbb{P}}(1)\big(\mathcal{O}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1)\big) = \mathcal{O}_{\mathbb{P}}(1)\end{aligned}$$

by (5.9), (5.10) and Lemmas 5.6–5.8. In particular, this implies

$$\left\|\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(p/\sqrt{n}\right) = \mathrm{o}_{\mathbb{P}}(1)$$

by Assumption (A8), and hence the second condition, $\mathrm{sign}\big(\widetilde{\sigma}_{n,S_\sigma}\big) = \mathrm{sign}\big(\sigma^*_{S_\sigma}\big)$, of Lemma 7.1 in the Supplementary Material is also satisfied with high probability for large sample sizes n. Sign-consistency of the adaptive LASSO and $\widehat{\sigma}^{\,\mathrm{AL}}_{n,S_\sigma} = \widetilde{\sigma}_{n,S_\sigma}$ is the consequence. □

SUPPLEMENTARY MATERIAL

Hermann, P. and Holzmann, H. (2024): Supplement to "Bounded Support in Linear Random Coefficient Models: Identification and Variable Selection," Econometric Theory Supplementary Material. To view, please visit https://doi.org/10.1017/S0266466624000070.

https://doi.org/10.1017/S0266466624000070 Published online by Cambridge University Press

REFERENCES

Arellano, M., & Bonhomme, S. (2012). Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies*, 79(3), 987–1020.

Beran, R., Feuerverger, A., & Hall, P. (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Annals of Statistics*, 24(6), 2569–2592.

Beran, R., & Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. *Annals of Statistics*, 20(4), 1970–1984.

Breunig, C., & Hoderlein, S. (2018). Specification testing in random coefficient models. *Quantitative Economics*, 9(3), 1371–1417.

Dunker, F., Eckle, K., Proksch, K., & Schmidt-Hieber, J. (2019). Tests for qualitative features in the random coefficients model. *Electronic Journal of Statistics*, 13(2), 2257–2306.

Gaillac, C., & Gautier, E. (2022). Adaptive estimation in the linear random coefficients model when regressors have limited variation. *Bernoulli*, 28, 504–524.

Gautier, E., & Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2), 581–607.

Hildreth, C., & Houck, J. P. (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association*, 63(322), 584–595.

Hoderlein, S., Holzmann, H., & Meister, A. (2017). The triangular model with random coefficients. *Journal of Econometrics*, 201(1), 144–169.

Hoderlein, S., Klemelä, J., & Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26, 804–837.

Holzmann, H., & Meister, A. (2020). Rate-optimal nonparametric estimation for random coefficient regression models. *Bernoulli*, 26(4), 2790–2814.

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18, 1603–1618.

Ichimura, H., & Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2), 269–295.

Leeb, H., & Pötscher, B. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations *Econometric Theory*, 19(1), 100–142.

Lewbel, A. (2005). Modeling heterogeneity. *Working Papers in Economics*, Boston College, 402.

Lewbel, A., & Pendakur, K. (2017). Unobserved preference heterogeneity in demand using generalized random coefficients. *Journal of Political Economy*, 125(4), 1100–1148.

Li, S., Cai, T. T., & Li, H. (2022). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association*, 117, 1835–1846.

Loh, P.-L., & Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, 45(6), 2455–2482.

Masten, M. A. (2018). Random coefficients on endogenous variables in simultaneous equations models. *Review of Economic Studies*, 85(2), 1193–1250.

Rigollet, P., & Hütter, J.-C. (2019). High dimensional statistics. *Lecture notes for course 18S997*, MIT OpenCourseWare, Cambridge.

Schelldorfer, J., Bühlmann, P., & van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2), 197–214.

Swamy, P. (1970). Efficient inference in a random coefficients model. *Econometrica*, 38, 311–324.

Thomas, E. G. (2014). A polarization identity for multilinear maps. *Indagationes Mathematicae*, 25(3), 468–474.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vandenberghe, L., & Boyd, S. (1996). Semidefinite programming. *SIAM Review*, 38(1), 49–95.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wagener, J., & Dette, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic

models. *Mathematical Methods of Statistics*, 22(2), 137–154.
Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4), 1733.