Supplementary files

to „Improving Group Lasso for high-dimensional categorical data"

1 Notations

In the main paper we consider the Group Lasso in the form:

$$\arg\min_{\beta} \quad \ell(\beta) + \lambda \sum_{k=1}^{r} ||W_k \beta_k||,$$

where $\ell(\beta) = \frac{1}{n}\sum_{i=1}^{n}[(x_{i.}^T\beta)^2/2 - y_i x_{i.}^T\beta]$ and W_k is a diagonal matrix with $(W_k)_{jj} = ||x_{j,k}||/\sqrt{n}$. In this supplement we use the equivalent notation

$$\arg\min_{\beta} \quad \tilde{\ell}(\beta) + \tilde{\lambda} \sum_{k=1}^{r} ||\tilde{W}_k \beta_k||, \tag{1}$$

where $\tilde{\ell}(\beta) = n\ell(\beta), \tilde{\lambda} = \sqrt{n}\lambda, \tilde{W}_k = \sqrt{n}W_k$. The new notation is more convenient in calculations. **For simplicity, we drop all „tilde" signs in (1) in this supplement.** We hope that it will not lead to confusions. Similar changes have to be done in the information criterion in the step (2b) of the PDMR algorithm.

In (1) we consider fixed weights $||x_{j,k}||$ in the Group Lasso, but in our theoretical investigation they can be arbitrary. What is more, using our results we show that the choice as in the main paper is optimal.

Let $W_1 = diag(w_{0,1}, w_{1,1}, \ldots, w_{p_1,1})$ and $W_k = diag(w_{1,k}, \ldots, w_{p_k,k}), k = 2, \ldots, r$ be diagonal nonrandom matrices with positive entries. Besides, $W = diag(W_1, \ldots, W_r)$ is a $p \times p$ diagonal matrix with matrices W_k on the diagonal.

In the following we consider $k \in \{1, \ldots, r\}$ and for $k = 2, \ldots, r$ we have $j \in \{1, \ldots, p_k\}$, while for $k = 1$ we have $j \in \{0, \ldots, p_1\}$. Let $x_{j,k}$ be a column of X corresponding to the j-th level ofthe k-th factor. The additional notations are $x_M = \max_{j,k} ||x_{j,k}||, x_m = \min_{j,k} ||x_{j,k}||, x_W = \max_{j,k} ||x_{j,k}||/w_{j,k}$.

It is easy to see that $\dot{\ell}(\beta) = \sum_{i=1}^{n}(x_{i.}^T\beta - y_i)x_{i.}$, where $\dot{\ell}$ denotes a derivative of ℓ. Besides, $\dot{\ell}(\mathring{\beta}) = -X^T\varepsilon$ for $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Next, for $k = 1, \ldots, r$ partial derivatives of $\ell(\beta)$ corresponding to coordinates of β_k are denoted by $\dot{\ell}_k(\beta)$.

2 Auxiliary results

We consider the PDMR algorithm with arbitrary diagonal matrices W_k in Group Lasso. The default setting from the main paper will be justified in Proposition 1.

First, we generalize a characteristic of linear models with continuous predictors, which quantifies the degree of separation between the model $M_{\mathring{\beta}}$ and other models [24]. Let $S = \{1 \leq k \leq r : \mathring{\beta}_k \neq 0\}$ and $\bar{S} = \{1, \ldots, r\} \setminus S$. Notice that S need not coincide with $M_{\mathring{\beta}}$. For $a \in (0, 1)$ and a diagonal matrix W we define a cone

$$\mathcal{C}_{a,W} = \{v \in \mathbb{R}^p : \sum_{k \in \bar{S}} ||W_k v_k|| \leq \sum_{k \in S} ||W_k v_k|| + a|Wv|_1\}. \tag{2}$$

A Cone Invertibility Factor (CIF) is defined as

$$\zeta_{a,W} = \inf_{0 \neq \nu \in \mathcal{C}_{a,W}} \frac{|W^{-1}X^T X v|_\infty}{|\nu|_\infty}. \tag{3}$$

In the case that matrix $X^T X$ is orthogonal one can easily find a lower bound on (3). For instance, for the default choice of weights W_k (i.e. $(W_k)_{jj} = ||x_{j,k}||$) we have $\zeta_{a,W} \geq x_m$ for all $a \in (0, 1)$, where we recall that $x_m = \min_{j,k} ||x_{j,k}||$ is the square root of the minimal number of observations per level.

In the case $n > p$ one usually uses the minimal eigenvalue of the matrix $X^T X$ to express the strength of correlations between predictors. Obviously, in the high-dimensional scenario this value is zero. Therefore, CIF can be viewed as a useful analog of the minimal eigenvalue for the case $p > n$. In comparison to more popular restricted eigenvalues [3] or compatibility constants [22], CIF enables sharper ℓ_∞ estimation error bounds [24], [9], [26]. We explain precisely this fact in Section 4 of this supplement. Finally, if all predictors are continuous, then (2) and (3) are the same as the cone and CIF in [24].

2.1 Estimation consistency of Group Lasso

Now we establish an upper bound on an estimation error of the Group Lasso, which can be applied to the high-dimensional scenario $p >> n$. Similar results can be found in the literature, for instance in [14, Theorem 4.5], [23, Theorem 2.2], [12, Theorem 5.1], [5, Theorem 8.1] or [4, Theorem III.6]. The main difference between those results and ours is that we measure the estimation error in the l_∞ norm, which is all we need in partition selection, while in those papers there is a mixture of l_2 and l_1 (or l_∞) norms. Thus, relying on those papers we would need more restrictive assumptions in our results. It is especially true for an orthogonal design. At the end of this subsection we compare our results to [12, Theorem 5.1].

Lemma 1. *Suppose that assumptions (1), (2) from the main paper are satisfied and* $a \in (0,1)$. *Then*

$$\mathbb{P}_{\mathring{\beta}}\left(|\hat{\beta}-\mathring{\beta}|_\infty > (1+a)\lambda\zeta_{a,W}^{-1}\right) \leq 2p\exp\left(-\frac{a^2\lambda^2}{2\sigma^2 x_W^2}\right).$$

Thus, if $\lambda^2 = 2a^{-2}\sigma^2 x_W^2 \log(2p/\alpha)$ *for some* $\alpha \in (0,1)$, *then with probability at least* $1-\alpha$

$$|\hat{\beta}-\mathring{\beta}|_\infty^2/\sigma^2 \leq 2(1+a)^2 a^{-2} x_W^2 \zeta_{a,W}^{-2}\log(2p/\alpha). \tag{4}$$

Proof. For $k=1,\ldots,r$ using KKT for the Group Lasso estimator $\hat{\beta}$, we have that $W_k^{-1}\dot{\ell}_k(\hat{\beta}) = -\lambda W_k\hat{\beta}_k/||W_k\hat{\beta}_k||$ for $\hat{\beta}_k \neq 0$ and $||W_k^{-1}\dot{\ell}_k(\hat{\beta})|| \leq \lambda$ for $\hat{\beta}_k = 0$. Therefore, we obtain that $|W^{-1}\dot{\ell}(\hat{\beta})|_\infty = \max_k |W_k^{-1}\dot{\ell}_k(\hat{\beta})|_\infty \leq \lambda$.

Recall that $\dot{\ell}(\mathring{\beta}) = -X^T\varepsilon$ and suppose that we are on an event $\mathcal{A} = \{|W^{-1}\dot{\ell}(\mathring{\beta})|_\infty \leq a\lambda\}$. First, we prove that $v := \hat{\beta}-\mathring{\beta} \in \mathcal{C}_{a,W}$. Using differentiability of ℓ and Taylor's expanssion we have $v^T\left[\dot{\ell}(\hat{\beta})-\dot{\ell}(\mathring{\beta})\right] = v^T\nabla^2\ell(\tilde{\beta})v$ for some $\tilde{\beta}$ between $\hat{\beta}$ and $\mathring{\beta}$. Obviously, this expression is nonnegative, because ℓ is convex. Moreover, $v_k = \hat{\beta}_k$ for $k \in \bar{S}$, so we also obtain

$$\begin{aligned} v^T\left[\dot{\ell}(\hat{\beta})-\dot{\ell}(\mathring{\beta})\right] &= \sum_{k=1}^r v_k^T\dot{\ell}_k(\hat{\beta}) - \sum_{k=1}^r v_k^T\dot{\ell}_k(\mathring{\beta}) \\ &= \sum_{k\in\bar{S}}\hat{\beta}_k^T\dot{\ell}_k(\hat{\beta}) + \sum_{k\in S} v_k^T\dot{\ell}_k(\hat{\beta}) - \sum_{k=1}^r v_k^T\dot{\ell}_k(\mathring{\beta}). \end{aligned} \tag{5}$$

Consider the first term in (5). Using KKT, it equals

$$\sum_{k\in\bar{S},\hat{\beta}_k\neq 0}\hat{\beta}_k^T\dot{\ell}_k(\hat{\beta}) = -\lambda\sum_{k\in\bar{S},\hat{\beta}_k\neq 0}||W_k\hat{\beta}_k|| = -\lambda\sum_{k\in\bar{S}}||W_k v_k||.$$

Similarly, we bound the second term in (5) by

$$\sum_{k\in S}||W_k v_k||\,||W_k^{-1}\dot{\ell}_k(\hat{\beta})|| \leq \lambda\sum_{k\in S}||W_k v_k||.$$

The last term in (5) can be bounded using the fact that we are on the event $\mathcal{A}$

$$\sum_{k=1}^r |W_k v_k|_1 |W_k^{-1}\dot{\ell}_k(\mathring{\beta})|_\infty \leq a\lambda\sum_{k=1}^r |W_k v_k|_1.$$

Joining the above facts we get that $v \in \mathcal{C}_{a,W}$. Therefore, from the definition (3) we have

$$\begin{aligned}\zeta_{a,W}|\hat{\beta}-\mathring{\beta}|_\infty &\leq \max_{1\leq k\leq r}|W_k^{-1}\dot{\ell}_k(\hat{\beta})-W_k^{-1}\dot{\ell}_k(\mathring{\beta})|_\infty \\ &\leq \max_{1\leq k\leq r}|W_k^{-1}\dot{\ell}_k(\hat{\beta})|_\infty+\max_{1\leq k\leq r}|W_k^{-1}\dot{\ell}_k(\mathring{\beta})|_\infty.\end{aligned}$$

Using again KKT and the fact, that we are on $\mathcal{A}$, we get $|\hat{\beta}-\mathring{\beta}|_\infty \leq (1+a)\lambda\zeta_{a,W}^{-1}$. Now we calculate probability of the event $\mathcal{A}$. To do it, we use the following exponential inequality for independent subgaussian variables $\varepsilon_i, i=1,\ldots,n$: for each $b>0$ and $v\in\mathbb{R}^n$ we have $P(\varepsilon^T v/||v|| > b) \leq \exp\left(-b^2/(2\sigma^2)\right)$. Using union bounds and the definition of x_W, we obtain

$$\begin{aligned}P_{\mathring{\beta}}(\mathcal{A}^c) &\leq \sum_{k,j} P\left(|x_{j,k}^T\varepsilon|/w_{j,k} > a\lambda\right) \leq 2\sum_{j,k}\exp\left(-\frac{a^2\lambda^2 w_{j,k}^2}{2\sigma^2||x_{j,k}||^2}\right) \\ &\leq 2p\exp\left(-\frac{a^2\lambda^2}{2\sigma^2 x_W^2}\right),\end{aligned}$$

where we consider $k\in\{1,\ldots,r\}$ and for $k=2,\ldots,r$ we have $j\in\{1,\ldots,p_k\}$, while for $k=1$ we have $j\in\{0,\ldots,p_1\}$.

The proof of the second claim is straightforward. □

The upper bound on the estimation error in Lemma 1 depends on the choice of a weight matrix W. So, to find optimal weights we should minimize $x_W^2\zeta_{a,W}^{-2}$. Solving this problem in the general case is difficult, so we restrict to the simplified version of the problem in the next result.

Proposition 1. *If X^TX is orthogonal and weights are of the form $w_{j,k}=||x_{j,k}||^q$ for $q\in\mathbb{R}$. Then for each $a\in(0,1)$ we have $x_W^2\zeta_{a,W}^{-2} \leq x_m^{-2}(x_M/x_m)^{\max(0,|2q-3|-1)} =: f(q)$ and $\arg\min_q f(q) = [1,2]$.*

Proof. For a linear model with an orthogonal design we have $|W^{-1}X^TXv|_\infty/|v|_\infty \geq \min_{j,k}||x_{j,k}||^{2-q}$ for all $v\in\mathbb{R}^p$. So, we can easily bound from above $\zeta_{a,W}^{-2}$ by x_m^{2q-4}, when $q\leq 2$ and x_M^{2q-4}, when $q>2$. The rest of the proof follows from the fact that x_W^2 equals x_M^{2-2q}, when $q\leq 1$ and x_m^{2-2q}, when $q>1$. □

Thus, for an orthogonal design with the optimal weights (i.e. $q\in[1,2]$) the upper bound in (4) behaves like $x_m^{-2}\log p$. Consider a *balanced design*, i.e. there are n/p_k observations on every level of k-factor. Then $x_m^{-2}=\max_k p_k/n$ and the upper bound on the estimation error of Group Lasso behaves like $\sqrt{\max_k p_k\ \log p/n}$. The assumption that a design is orthogonal is quite restrictive. The much more

common case, especially for $p >> n$, is an *almost orthogonal* design, i.e. when ${x_{j_1,k_1}}^T x_{j_2,k_2} = o(x_m^2)$ for $(j_1,k_1) \neq (j_2,k_2)$. In such a case weights $w_{j,k} = ||x_{j,k}||^q$ for $q \in [1,2]$ can be treated as *almost optimal*.

In the original paper on the Group Lasso [25] two choices of weights are proposed. The first one, called "obvious", gives a penalty of the form $\lambda \sum_k ||\beta_k||$. In the second one, called "preferred" they have a penalty $\lambda \sum_k \sqrt{p_k}||\beta_k||$. The latter choice is more widely used in the literature [14], [23], [5]. Now we compare these choices of weights to those obtained in Proposition 1. Notice that columns of X are normalized in [25], which is not done in our paper. So, we start with writing their penalty in our setting. We do it under a balanced design (it is defined in the previous paragraph). Their first choice gives a penalty $\lambda\sqrt{n}\sum_k p_k^{-1/2}||\beta_k||$, while the second one gives $\lambda\sqrt{n}\sum_k ||\beta_k||$. On the other hand, by Proposition 1 for $q = 1$ we obtain $\lambda\sqrt{n}\sum_k p_k^{-1/2}||\beta_k||$, while for $q = 2$ we have $\lambda n \sum_k p_k^{-1}||\beta_k||$. Therefore, our optimal choice for $q = 1$ coincides with the "obvious" choice in [25]. However, the "preferred" choice in [25] leads to sub-optimal results. Obviously, Proposition 1 deals with an orthogonal design, so our result is rather a starting point of the thorough analysis on weights optimality.

Finally, notice that for an orthogonal and balanced design [12, Theorem 5.1] bounds the estimation error of Group Lasso by $x_m^{-2}(\max_k p_k + \log r)$, which is greater than $x_m^{-2}\log p$ in Lemma 1.

2.2 Partition selection of PDMR

In this section we state the main theoretical result concerning our algorithm. First, we need to define the Kullback-Leibler (K-L) distance between the true model $M_{\mathring{\beta}}$ and its submodels. The precise definition of a submodel and its cardinality is given in Section 3 of this supplement. Roughly speaking, model M is a submodel of $M_{\mathring{\beta}}$ ($M \subsetneq M_{\mathring{\beta}}$), if M contains at least one additional merging of levels comparing to $M_{\mathring{\beta}}$.

Let M be a submodel of $M_{\mathring{\beta}}$ and $k = |M_{\mathring{\beta}}| - |M|$. Denote

$$\delta_k = ||X\mathring{\beta} - X_M\beta_M^*||^2,$$

where $\beta_M^* = \arg\min_{\beta_M} ||X\mathring{\beta} - X_M\beta_M||^2$. A scaled K-L distance between $M_{\mathring{\beta}}$ and its submodels is

$$\delta_{M_{\mathring{\beta}}} = \min_{k=1,\ldots,|M_{\mathring{\beta}}|-1} \ \min_{M: M\subsetneq M_{\mathring{\beta}}, |M_{\mathring{\beta}}|-|M|=k} \frac{\delta_k}{k}. \tag{6}$$

Different variants of the K-L distance have been used in the consistency analysis of selection algorithms [16, Section 3.1], but $\delta_{M_{\mathring{\beta}}}$ defined in (6) seems to lead to optimal results [18, Theorem 1].

In the next theorem we establish properties of the PDMR algorithm in partition selection. We consider the default setting of weights in Group Lasso (i.e. $(W_k)_{jj} = ||x_{j,k}||$), so (3) simplifies to ζ_a.

Theorem 1. *Suppose that assumptions (1) and (2) from the main paper are satisfied and there exists* $0 < a < 1$ *such that*

$$2a^{-2}\sigma^2 \log(|M_{\mathring{\beta}}|^2/(2\log 2)) \leq \lambda^2 < \frac{\min(\Delta^2\zeta_a^2, 4\delta_{M_{\mathring{\beta}}})}{16(1+a)^2}. \tag{7}$$

Then

$$P(\hat{M}_{PDMR} \subsetneq M_{\mathring{\beta}}) \leq (2p + |M_{\mathring{\beta}}|^2) \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right). \tag{8}$$

The simplified version of the above result is given in Theorem 1 in the main paper. The differences are:
(i) the case that $|M_{\mathring{\beta}}|^2 \leq p$ is considered there, so $\log(|M_{\mathring{\beta}}|^2/(2\log 2))$ in (7) is replaced by $\log p$ and $|M_{\mathring{\beta}}|^2$ in (8) is replaced by p. Obviously, if $|M_{\mathring{\beta}}|^2 > p$, then $\log(|M_{\mathring{\beta}}|^2/(2\log 2))$ can be replaced by slightly larger $2\log p$,
(ii) the sample size n does not appear explicitly in (7) and (8), because we use different notation comparing to the main paper (see the beginning of section „Notations" in this supplement),
(iii) here the identifiability number κ is explicitly stated in (7).

Proof. We will establish two inequalities

$$P(M_{\mathring{\beta}} \notin \mathcal{M}) \leq 2p \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right) \tag{9}$$

and

$$P(M_{\mathring{\beta}} \in \mathcal{M}, \hat{M}_{PDMR} \subsetneq M_{\mathring{\beta}}) \leq (2\log 2)^{-1}|M_{\mathring{\beta}}|^2 \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right). \tag{10}$$

We start with (9). From Lemma 1 we know that

$$P(|\hat{\beta} - \mathring{\beta}|_\infty \leq (1+a)\lambda\zeta_a^{-1}) \geq 1 - 2p\exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right).$$

Now we fix the k-th predictor and take indexes j_1, j_2 such that $\mathring{\beta}_{j_1,k} = \mathring{\beta}_{j_2,k}$, i.e. they correspond to the same cluster in $M_{\mathring{\beta}}$. Let $R = (1+a)\lambda\zeta_a^{-1}$. We obtain

$$|\hat{\beta}_{j_1,k} - \hat{\beta}_{j_2,k}| \leq |\hat{\beta}_{j_1,k} - \mathring{\beta}_{j_1,k}| + |\hat{\beta}_{j_2,k} - \mathring{\beta}_{j_2,k}| \leq 2R. \tag{11}$$

On the other hand, if j_1, j_2 are such that $\mathring{\beta}_{j_1,k} \neq \mathring{\beta}_{j_2,k}$, then

$$|\hat{\beta}_{j_1,k} - \hat{\beta}_{j_2,k}| \geq |\mathring{\beta}_{j_1,k} - \mathring{\beta}_{j_2,k}| - |\hat{\beta}_{j_1,k} - \mathring{\beta}_{j_1,k}| - |\hat{\beta}_{j_2,k} - \mathring{\beta}_{j_2,k}| \geq \Delta - 2R > 2R,$$

because $\Delta > R$ by assumption the (7). Therefore, there is a separation between entries of a dissimilarity matrix D_k in the clustering step of PDMR. Namely, entries corresponding to indexes from the same true cluster are smaller than those corresponding to distinct true clusters. The first step of complete linkage clustering uses the dissimilarity matrix D_k, then in consecutive steps this matrix is updated as follows: a distance between two clusters A and B is defined as $\max_{a\in A, b\in B} |\hat{\beta}_{a,k} - \hat{\beta}_{b,k}|$. Therefore, in some step of clustering we obtain true partitioning of levels of the k-th factor and all cutting heights in h_k to that step are not larger than $2R$, while subsequent coefficients of h_k are larger than $2R$. Clearly, threshold $2R$ does not depend on k. Therefore, this separation property is also satisfied after taking all cutting heights together and sorting increasingly. Thus, the true model is contained in the nested family $\mathcal{M}$ with high probability.

Now, we consider (10). Notice that

$$\begin{aligned} & P(M_{\mathring{\beta}} \in \mathcal{M}, \hat{M}_{PDMR} \subset M_{\mathring{\beta}}) \\ \leq P & \left(\exists_{M: L_M \subsetneq L_{M_{\mathring{\beta}}}} \quad \ell(\hat{\beta}_M) + \lambda^2 |M|/2 < \ell(\hat{\beta}_{M_{\mathring{\beta}}}) + \lambda^2 |M_{\mathring{\beta}}|/2\right) \end{aligned} \tag{12}$$

and we recall that $\hat{\beta}_M$ is a minimum loss estimator over $\mathbb{R}^p$ with constraints determined by the model M. Technical details of this constrained minimization is given in Section 3 of these supplementary materials. Denote $k = |M_{\mathring{\beta}}| - |M|$. We can calculate that $\delta_k = ||(I - H_M)X\mathring{\beta}||^2$ and

$$\ell(\hat{\beta}_M) = \delta_k/2 + \varepsilon^T (I - H_M) X\mathring{\beta} + \varepsilon^T (\mathbb{I} - H_M)\varepsilon/2 - y^T y/2,$$

in particular $\ell(\hat{\beta}_{M_{\mathring{\beta}}}) = \varepsilon^T (\mathbb{I} - H_{M_{\mathring{\beta}}})\varepsilon/2 - y^T y/2$. Since $H_{M_{\mathring{\beta}}} - H_M$ is a projection matrix, we have

$$\ell(\hat{\beta}_M) - \ell(\hat{\beta}_{M_{\mathring{\beta}}}) \geq \delta_k/2 + \varepsilon^T (I - H_M) X\mathring{\beta}$$

and we can bound the rhs of (12) by

$$P\left(\exists_{M: L_M \subsetneq L_{M_{\mathring{\beta}}}} \quad -2\varepsilon^T (I - H_M) X\mathring{\beta} \geq \delta_k - k\lambda^2\right).$$

Clearly, we have $\delta_k \geq k\delta_{M_{\mathring{\beta}}}$, so above probability is bounded by

$$P\left(\exists_{M: L_M \subsetneq L_{M_{\mathring{\beta}}}} \quad \frac{-\varepsilon^T (I - H_M) X\mathring{\beta}}{\sqrt{\delta_k}} \geq \sqrt{k\delta_{M_{\mathring{\beta}}}} \left(1 - \frac{\lambda^2}{\delta_{M_{\mathring{\beta}}}}\right)/2\right). \tag{13}$$

To estimate (13) we use union bounds with the exponential inequality for subgaussian random variables (see the proof of Lemma 1). Thus, we bound (13) from above by

$$\sum_{k=1}^{|M_{\mathring{\beta}}|-1} N_k \exp\left(-\frac{k\delta_{M_{\mathring{\beta}}}}{8\sigma^2}\left(1-\frac{\lambda^2}{\delta_{M_{\mathring{\beta}}}}\right)^2\right), \tag{14}$$

where N_k is a number of models M such that $L_M \subsetneq L_{M_{\mathring{\beta}}}$ and $|M| = |M_{\mathring{\beta}}| - k$. Notice that the value in (14) is the largest, if $M_{\mathring{\beta}}$ consists of one factor on $|M_{\mathring{\beta}}|$ levels, which we assume in the following. In this case $N_k = \left\{ {|M_{\mathring{\beta}}| \atop |M_{\mathring{\beta}}|-k} \right\}$, where $\left\{ {r \atop s} \right\}$ is a Stirling number of the second kind, i.e. a number of ways to partition a set of r objects into s non-empty subsets.

From the assumption $\lambda^2 \leq \delta_{M_{\mathring{\beta}}}/(2+2a)^2$ we obtain

$$\lambda^2/\delta_{M_{\mathring{\beta}}} \leq f_1(a), \qquad \text{where} \quad f_1(a) = 1 + 2a^2 - \sqrt{(1+2a^2)^2 - 1}, \tag{15}$$

which gives

$$\frac{4a^2\lambda^2}{\delta_{M_{\mathring{\beta}}}} \leq \left(1 - \frac{\lambda^2}{\delta_{M_{\mathring{\beta}}}}\right)^2. \tag{16}$$

Therefore, we estimate (14) by

$$\begin{aligned} &\sum_{k=1}^{|M_{\mathring{\beta}}|-1} \left\{ {|M_{\mathring{\beta}}| \atop |M_{\mathring{\beta}}| - k} \right\} \exp\left(-\frac{ka^2\lambda^2}{2\sigma^2}\right) \\ = \;& \exp\left(-|M_{\mathring{\beta}}|\frac{a^2\lambda^2}{2\sigma^2}\right) \sum_{k=1}^{|M_{\mathring{\beta}}|} \left\{ {|M_{\mathring{\beta}}| \atop k} \right\} \exp\left(\frac{ka^2\lambda^2}{2\sigma^2}\right) - 1. \end{aligned} \tag{17}$$

The sum in (17) is called a Touchard polynomial. Its value is closely related to moments of Poisson random variables (see Lemma 2 given below). Therefore, (17) can be estimated by Lemma 3 (given below) as

$$\exp\left[|M_{\mathring{\beta}}|^2 \exp\left(-\frac{a^2\lambda^2}{2\sigma^2}\right)/2\right] - 1.$$

Using the inequality $\exp(c) - 1 \leq \log(2)^{-1} c$ for $0 \leq c \leq \log(2)$, we finish the proof. □

Lemma 2 ([15], Proposition 3.3.2)**.** *For every $n \geq 0$ and $x > 0$, one has that $\mathbb{E}[K(x)]^n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} x^k$, where $K(x)$ is a Poisson random variable with parameter x.*

Lemma 3 ([1], Theorem 1)**.** *Under assumptions of Lemma 2 we have*

$$\mathbb{E}[K(x)/x]^n \leq \left(\frac{n/x}{\log(1+n/x)}\right)^n \leq \exp(n^2/(2x)).$$

3 Models and constrained minimization

For simplicity, we consider the case that all predictors are factors. But the extension to the general case is straighforward.

We recall that each model M is defined as a sequence $M = (P_1, P_2, \ldots, P_r)$, where P_k is some partition of a set of levels of the k-th factor, i.e. $\{0, 1, \ldots, p_k\}$. We will show that every model M corresponds to a linear space

$$L_M = \{\beta \in \mathbb{R}^p : A_{0,M}\beta = 0\}, \tag{18}$$

where a matrix $A_{0,M}$ is defined in the following subsection.

3.1 Matrix $A_{0,M}$

Suppose that $P_k = C_{1,k} \cup C_{2,k} \cup C_{j_k,k}$, so j_k is a number of clusters that the set $\{0, 1, \ldots, p_k\}$ is divided. In further considerations we fix the ordering between clusters. We also suppose that reference levels (i.e. zero levels) belong to $C_{1,k}$ for each k. Let $s_{j,k}$ be the smallest element in $C_{j,k}$. In particular, $s_{1,k} = 0$.

Fix $\beta \in \mathbb{R}^p$. We recall that $\beta = (\beta_1^T, \beta_2^T, \ldots, \beta_r^T)^T$, where $\beta_1 = (\beta_{0,1}, \beta_{1,1}, \ldots, \beta_{p_1,1})^T \in \mathbb{R}^{p_1+1}$ and $\beta_k = (\beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \ldots, \beta_{p_k,k})^T \in \mathbb{R}^{p_k}$ for $k = 2, \ldots, r$. Now we change the ordering between coordinates of β according to the ordering defined by a model M in the following way:

$$\beta = (\underbrace{\beta_{s_{1,1},1}, \beta_{s_{2,1},1}, \ldots, \beta_{s_{j_1,1},1}}_{group\ 1}, \underbrace{\beta_{s_{2,2},2}, \ldots, \beta_{s_{j_2,2},2}}_{group\ 2}, \ldots, \underbrace{\beta_{s_{2,r},2}, \ldots, \beta_{s_{j_r,r},r}}_{group\ r}, \quad \text{remaining coefficients}). \tag{19}$$

In other words, levels of the first factor are partitioned as $\{0, 1, \ldots, p_1\} = C_{1,1} \cup C_{2,1} \cup C_{j_1,1}$ and the smallest numbers in these clusters are $s_{1,1}, s_{2,1}, \ldots, s_{j_1,1}$, respectively. So, *group 1* consists of the corresponding coefficients of $\beta_1 \in \mathbb{R}^{p_1+1}$. Next, levels of the second factor are partitioned as $\{0, 1, \ldots, p_2\} = C_{1,2} \cup C_{2,2} \cup C_{j_2,2}$ and the smallest numbers in these clusters are $s_{1,2}, s_{2,2}, \ldots, s_{j_2,2}$, respectively. So, *group 2* consists of the corresponding coefficients of $\beta_2 \in \mathbb{R}^{p_2}$. In particular, *group 2* does not contain $\beta_{s_{1,2},2}$, because $s_{1,2}$ corresponds to a cluster, which contains a reference level of the second factor and we do not include coefficients corresponding to reference levels in vector β (cf. Section 2 in the main part of the paper, the only exception is a reference level of the first factor). The same

proceeding relates to the following factors. At the end, we write all coefficients, which were not used before. They are called *remaining coefficients* in (19).

To make this new ordering more transparent we consider the example that we have two factors: the first one with 8 levels and the second one with 7 levels. So, $p_1 = 7, p_2 = 6$ and $p = 14$. Let $M = (P_1, P_2)$ be as follows: $P_1 = \{0, 2, 6\} \cup \{3, 4, 5\} \cup \{1, 7\}$, $P_2 = \{0, 4\} \cup \{1, 2, 6\} \cup \{3, 5\}$. Then

$$\beta = (\beta_{0,1}, \beta_{3,1}, \beta_{1,1}, \beta_{1,2}, \beta_{3,2}, \beta_{2,1}, \beta_{6,1}, \beta_{4,1}, \beta_{5,1}, \beta_{7,1}, \beta_{4,2}, \beta_{2,2}, \beta_{6,2}, \beta_{5,2}).$$

Let m be a number of clusters indicated by the model M, which do not contain reference levels plus one, i.e. $m = j_1 + (j_2 - 1) + \ldots + (j_k - 1)$. The matrix A_{0M} is a $(p - m) \times p$ matrix of a form $(B_M, \mathbb{I}_{p-m})$ for a $(p - m) \times m$ matrix B_M and the identity matrix $\mathbb{I}_{p-m}$, where the matrix B_M is constructed as follows: first, we define a connection between columns of B_M and the first m coordinates of (19) as follows: the first column of B_M corresponds to $\beta_{s_{1,1},1}$, the second one corresponds to $\beta_{s_{2,1},1}$ etc. Analogously, columns of $\mathbb{I}_{p-m}$ correspond to the last $p - m$ coordinates of (19) (i.e. those called *remaining coefficients*). Now we find any 1 in the matrix $\mathbb{I}_{p-m}$, say it is in a column t^* and a row t^* of the matrix $\mathbb{I}_{p-m}$. This column corresponds to some coordinate in (19), say β_{j^*,k^*}. It means that this column corresponds to the j^*-th level of the k^*-th factor. Now we check to which cluster this level belongs to. Then we find the smallest element in this cluster, say r^*. If $r^* \neq 0$ or $r^* = 0$ but $k^* = 1$, then we take the matrix B_M and write -1 in its column corresponding to coordinate β_{r^*,k^*} and the row t^*. The remaining entries in B_M are filled in by zeroes.

In the above example we have $m = 5$ and

$$A_{0,M} = \begin{array}{c} \begin{array}{ccccc|ccccccccc} \beta_{0,1} & \beta_{3,1} & \beta_{1,1} & \beta_{1,2} & \beta_{3,2} & \beta_{2,1} & \beta_{6,1} & \beta_{4,1} & \beta_{5,1} & \beta_{7,1} & \beta_{4,2} & \beta_{2,2} & \beta_{6,2} & \beta_{5,2} \end{array} \\ \left(\begin{array}{ccccc|ccccccccc} -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}.$$

Therefore, the space L_M defined in (18) consists of those vectors β, which determines the same partitions of factors' levels (and the same as given by M). These partitions are also coded by $A_{0,M}$.

3.2 Constrained minimization

In the paper we have many places where we consider $\hat{\beta}_M$, *which is a minimum loss estimator over* $\mathbb{R}^p$ *with constraints determined by the model* M. Now we can precisely define this minimization as $\arg\min_{\beta \in L_M} \ell(\beta)$. In this subsection we show that this constrained minimization can be replaced by the unconstrained one.

Let $A_{1,M} = (\mathbb{I}_m, \mathbf{0}_{m\times(p-m)})$ be a $(m \times p)$-complement of $A_{0,M}$ to a invertible matrix A_M, that is:

$$A_M = \left[\frac{A_{1,M}}{A_{0,M}}\right] \quad \text{and} \quad A_M^{-1} = \left[A_M^1 | A_M^0\right] = \left[\begin{smallmatrix} I_m \\ -B_M \end{smallmatrix} \middle| \begin{smallmatrix} \mathbf{0}_{m\times(p-m)} \\ \mathbb{I}_{p-m} \end{smallmatrix}\right],$$

where A_M^{-1} is calculated using the Schur complement.

Let β_M be an arbitrary element of L_M and $\xi_M = A_{1,M}\beta_M$, then $\beta_M = A_M^1 \xi_M$. Indeed, we have

$$\beta_M = A_M^{-1} A_M \beta_M = A_M^{-1} \left[\frac{A_{1,M}\beta_M}{A_{0,M}\beta_M}\right] = \left[A_M^1 | A_M^0\right] \left[\frac{\xi_M}{\mathbf{0}_{p-m}}\right] = A_M^1 \xi_M.$$

Therefore, L_M in (18) can be equivalently expressed as

$$L_M = \{A_M^1 \xi : \xi \in \mathbb{R}^m\}.$$

Therefore, L_M can be viewed as a linear space spanned by columns of A_M^1. The dimension of the space L_M is called a size of the model M and denoted by $|M|$. Clearly, we have $|M| = m$, so $|M|$ is a number of different non-reference levels (again with an exception for the first factor) indicated by the model M.

Fix the model M. We change the ordering of columns in X according to the ordering induced by M, as in (19). Then a matrix $Z_M = XA_M^1$ is simply the matrix X with appropriate columns deleted or added to each other according to partitions in the model $M = (P_1, \ldots, P_r)$. We also have $X\beta_M = Z_M \xi_M$. We assume that the considered models are sufficiently sparse, which means that $r(Z_M) = |M| \leq \bar{m}$, where $\bar{m} < \min(n, p)$ and $r(Z_M)$ is the rank of Z_M.

Therefore, constrained minimization $\arg\min_{\beta \in L_M} \ell(\beta)$ can be replaced by the unconstrained one as follows: we compute an ordinary least squares estimator with a design matrix Z_M, i.e. $\hat{\xi}_M = (Z_M^T Z_M)^{-1} Z_M^T y$. Then we calculate $\hat{\beta}_M = A_M^1 \hat{\xi}_M$.

Finally, we also need the following notation of a projection matrix $H_M = Z_M (Z_M^T Z_M)^{-1} Z_M^T$.

3.3 Submodels

Now we can easily define submodels, namely M_1 is a submodel of M_2, if $L_{M_1} \subset L_{M_2}$. Roughly speaking, partitions induced by M_1 and M_2 are the same ($L_{M_1} = L_{M_2}$) or partitions induced by M_1 contains at least one additional merging of levels comparing to those induced by M_2 ($L_{M_1} \subsetneq L_{M_2}$). Finally, a model determined by a vector β is such M that $\beta \in L_M$ and M has the smallest size among all models with this property. It is denoted M_β.

4 Cone invertibility factor (CIF)

Consider linear model (1) from the main paper with numerical predictors only. Let T be the set of indices corresponding to the support of the true vector $\mathring{\beta}$ and $T' = \{1, \ldots, p\} \setminus T$. Let β_T and $\beta_{T'}$ be the restrictions of the vector $\theta \in \mathbb{R}^p$ to the indices from T and T', respectively. Now, for $a \in (0, 1)$ we consider a cone

$$C_a = \{\theta \in \mathbb{R}^p : |\theta_{T'}|_1 \leq \frac{1+a}{1-a}|\theta_T|_1\} \ .$$

In the case when $p >> n$ three different characteristics measuring the potential for consistent estimation of the model parameters have been introduced:
- the restricted eigenvalue [3]:

$$RE_a = \inf_{0 \neq \theta \in C_a} \frac{\theta^T X^T X \theta}{|\theta|_2^2} \ ,$$

- the compatibility factor [5]:

$$K_a = \inf_{0 \neq \theta \in C_a} \frac{|T| \theta^T X^T X \theta}{|\theta_T|_1^2} \ ,$$

- the cone invertibility factor (CIF, [24]): for $q \geq 1$

$$\zeta_{a,q} = \inf_{0 \neq \theta \in C_a} \frac{|T|^{1/q} |X^T X \theta|_\infty}{|\theta|_q} \ .$$

Relations between the above quantities are discussed, for instance, in van de Geer and Bühlmann [22], Ye and Zhang [24], Huang et al. [8]. Moreover, notice that $\zeta_{a,\infty}$ is the same as (3), if we omit weight matrix W.

We use CIF, since this factor allows for a sharp formulation of convergency results for all l_q norms with $q \geq 1$. Indeed, the following estimation bounds are

established for Lasso with numerical predictors (see Huang et al. [8, Section 3]): with probability close to one

$$|\hat{\beta} - \mathring{\beta}|_1 \leq \frac{2(1+a)|T|\lambda}{(1-a)K_a} =: R_a^1 \tag{20}$$

$$|\hat{\beta} - \mathring{\beta}|_2 \leq \frac{(1+a)|T|^{1/2}\lambda}{RE_a} =: R_a^2 \tag{21}$$

$$|\hat{\beta} - \mathring{\beta}|_q \leq \frac{(1+a)|T|^{1/q}\lambda}{\zeta_{a,q}} =: R_{a,q}^3 \tag{22}$$

Such estimation bounds are the main tool to prove selection consistency of modifications of Lasso such as Thresholded Lasso, Adaptive Lasso or algorithms with nonconvex penalties (SCAD, MCP). Indeed, these inequalities are used to prove *separability* of Lasso, i.e. for each $j \in T$ and $k \notin T$ we have $|\hat{\beta}_j| \geq |\hat{\beta}_k|$. To get it one has to assume additionally that $\mathring{\beta}_{min} = \min_{j \in T} |\mathring{\beta}_j|$ is bounded from below by twice the right-hand side of (20), (21) or (22). The latter condition means that the signal has to be large enough. Obviously, one wants this condition to be as weak as possible. Below we show that the right-hand side of (20), (21), (22) is the smallest for CIF with $q = \infty$.

Clearly, $R_{a,q}^3$ is the smallest for $q = \infty$. Besides, for each $\beta \in \mathcal{C}_a$ we have $|\beta|_1 \leq 2|\beta_T|_1/(1-a)$ and $|\beta_T|_1^2 \leq |T||\beta|_2^2$. These two facts imply that $K_a \geq RE_a$ and $\sqrt{RE_a K_a} \leq 2\zeta_{a,2}/(1-a)$. Therefore, we obtain

$$R_{a,\infty}^3 \leq R_{a,2}^3 \leq \frac{2(1+a)|T|^{1/2}\lambda}{(1-a)\sqrt{RE_a K_a}} \leq \frac{2(1+a)|T|^{1/2}\lambda}{(1-a)RE_a} \leq 2R_a^2/(1-a).$$

Taking a not to close to one (for instance, $a = 0.5$) we obtain that (22) with $q = \infty$ is not larger than (21) with respect to the constant. However, it is possible that $K_a >> RE_a$ [22], which means that $R_{a,\infty}^3$ might be significantly smaller than R_a^2.

Finally, for $\beta \in \mathcal{C}_a$ we have $|\beta|_\infty \leq (1+a)|\beta_T|_1/(1-a)$, which gives $K_a \leq 2(1+a)|T|\zeta_{a,\infty}/(1-a)^2$. Consequenlty, $R_{a,\infty}^3 \leq (1+a)R_a^1/(1-a)$, so $R_{a,\infty}^3$ is at most R_a^1, if one takes a close to zero. Again, we can show the example that $R_{a,\infty}^3$ is significanlty larger than R_a^1. Consider the orthonormal case $X^T X = \mathbb{I}$. Then $\zeta_{a,\infty} = 1$ and $K_a = \inf_{0 \neq \theta \in C_a} \frac{|T||\theta|_2^2}{|\theta_T|_1^2} \geq 1$. On the other hand, K_a is smaller than $\frac{|T||\beta|_2^2}{|\beta_T|_1^2}$ for any $\beta \in \mathcal{C}_a$. Take a vector $d \in \mathbb{R}^p$ such that d has ones on the set T and zeroes elsewhere. Clearly, $d \in \mathcal{C}_a$ for any a. Therefore, $K_a \leq \frac{|T||d|_2^2}{|d_T|_1^2} = 1$, so we have $K_a = 1$. Consequently, $R_a^1 = \frac{2(1+a)|T|\lambda}{(1-a)}$, so $R_a^1 = \frac{2|T|}{1-a} R_{a,\infty}^3 > |T| R_{a,\infty}^3$. Therefore, R_a^1 is significantly larger than $R_{a,\infty}^3$, if $|T|$ can tend to infinity.

5 Description of real data sets and additional results of experiments

We investigate five real data sets: the first two with binary responses and the next three with continuous responses:
- the Adult data set [11] contains data from the 1994 US census. It contains 32,561 observations in a file `adult.data` and 16,281 observations in a file `adult.test`. The response represents whether the individual's income is higher than 50,000 USD per year or not. We preprocessed the data as in [20], i.e. we combined two files together, removed 4 variables representing either irrelevant (*fnlwgt*) or redundant (*education-num*) features or with values for the most part equal to zero (*capital-gain* and *capital-loss*) and then removed the observations with missing values. Preprocessing resulted in 45,222 observations with 2 continuous and 8 categorical variables with $p = 93$;
- the Promoter data set [7], [21] contains E. Coli genetic sequences of length 57. The response indicates whether the region represents a gene promoter. We removed the *name* variable and further worked with a data set consisting of 106 observations with 57 categorical variables, each with 4 levels representing 4 nucleotides, thus with $p = 172$.

Adult and Promoter data sets are available at the UCI Machine Learning Repository [6],
- the Airbnb data set reports rental price for a number of hosts offering rental in the Airbnb service and is available from *insideairbnb.com.* The host is characterised by a number of features like the neighbourhood, number of rooms, is it kids friendly, the length of rental history, reviews etc. We follow [19] in using the same dated version of the data set and in preprocessing the data as in [17], i.e. we compute numeric sentiment for reviews or otherwise transform features into numbers and normalize them (including the log transformation of the rental price), with the following exceptions: (1) we retain the *host_id*, *street* and *neighbourhood* columns as categorical variables, with 39393, 311 and 204 levels, respectively and (2) we skip the feature selection step, since ability of considered methods to screen predictors is one of the things we want to test in this paper. This preprocessing procedure resulted in 49,976 observations with 765 variables (out of which 3 are categorical) with $p = 40668$;
- the Insurance data set [10] contains data describing attributes of life insurance applicants. The response is an 8-level ordinal variable measuring insurance risk of the applicant, which we treat as a continuous response. We preprocessed the data as in [20], i.e. we removed the irrelevant *id* variable and 13 variables with missing values. Preprocessing resulted in 59,381 observations with 5 continuous and 108

Table 1: Execution time of methods.

Dataset	PDMR	DMR	SCOPE-8	SCOPE-32
Setting 1, snr=3,	11.67	15.25	335.6	332.03
Setting 2, snr=3,	11.57	15.56	371.91	353.68
Setting 3, snr=3,	11.47	15.11	352.84	370.06
Setting 4, snr=3,	11.48	15.21	361.14	326.89
Setting 5, snr=3,	12.71	16.81	504.25	764.15
Setting 6, snr=3,	12.17	16.49	712.94	1253.64
Setting 1, snr=4,	11.54	15.26	352.8	320.09
Setting 2, snr=4,	11.5	15.36	406.26	349.75
Setting 3, snr=4,	11.3	15.15	391.03	364.08
Setting 4, snr=4,	11.18	15.01	379.97	312.04
Setting 5, snr=4,	12.53	16.46	505.87	791.67
Setting 6, snr=4,	12.51	16.19	726.14	1255.18

categorical variables with $p = 823$;
- the Antigua data set [2] contains data concerning maize fertilizer experiments on the Island of Antigua and is available at the `R` package `DAAG` [13]. The response measures harvest. We removed the irrelevant *id* variable and one observation with a clearly outlying value of *ears* variable. We treat *plot* as a categorical variable and further worked with a data set consisting of 287 observations with 1 continuous and 4 categorical variables with $p = 58$.

Finally, in Table 1 we state execution times of procedures. In the main paper only results for SNR=3 are included. Here we see that results for SNR=4 are similar.

References

[1] Ahle, T. D. (2022). Sharp and simple bounds for the raw moments of the binomial and poisson distributions. *Statistics & Probability Letters*, 182:109–306.

[2] Andrews, D. F. and Herzberg, A. M. (1985). *Data. A Collection of Problems from Many Fields for the Student and Research Worker*. Springer, New York.

[3] Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732.

[4] Blazere, M., Loubes, J.-M., and Gamboa, F. (2014). Oracle inequalities for a

group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory*, 60:2303–2318.

[5] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data*. Springer, New York.

[6] Dua, D. and Graff, C. (2017). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences.

[7] Harley, C. and Reynolds, R. (1987). Analysis of e. coli promoter sequences. *Nucleic Acids Research*, 15:2343–2361.

[8] Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the Cox model. *Annals of Statistics*, 41:1142–1165.

[9] Huang, J. and Zhang, C. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13:1839–1864.

[10] Kaggle (2015). Prudential life insurance assessment. https://www.kaggle.com/c/prudential-life-insurance-assessment/data.

[11] Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of KDD*, pages 202–207.

[12] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39:2164–2204.

[13] Maindonald, J. and Braun, W. (2010). *Data Analysis and Graphics Using R*. Cambridge University Press.

[14] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633.

[15] Peccati, G. and Taqqu, M. (2011). *Wiener Chaos: Moments, Cumulants and Diagrams: A survey with Computer Implementation*. Springer.

[16] Pokarowski, P. and Mielniczuk, J. (2015). Combined ℓ_0 and ℓ_1 penalized least squares for linear model selection. *Journal of Machine Learning Research*, 16:961–992.

[17] Rezazadeh Kalehbasti, P., Nikolenko, L., and Rezaei, H. (2021). Airbnb price prediction using machine learning and sentiment analysis. In *Proceedings of CD-MAKE 2021*, pages 173–184.

[18] Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65:807–832.

[19] Simchoni, G. and Rosset, S. (2021). Using random effects to account for high-cardinality categorical features and repeated measures in deep neural networks. In *Proceedings of NIPS*, volume 34, pages 25111–25122.

[20] Stokell, B. G., Shah, R. D., and Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:579–611.

[21] Towell, G., Shavlik, J., and Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based artificial neural networks. In *Proceedings of AAAI*.

[22] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

[23] Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16:1369–1384.

[24] Ye, F. and Zhang, C. (2010). Rate minimaxity of the Lasso and Dantzig Selector for the l_q loss in l_r balls. *Journal of Machine Learning Research*, 11:3519–3540.

[25] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67.

[26] Zhang, C. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27:576–593.