On sparsity and clustering properties

Xavier Dupuis, Patrick J C Tardivel

On sparsity and clustering properties by OWL penalized estimator

Xavier Dupuis* Patrick J.C. Tardivel*

Abstract

The Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) and the related Ordered Weighted ℓ_1 (OWL) penalized estimator have the particularity to exhibit some null components (sparsity) and some components equal in absolute value (clustering). Recently introduced, the notion of OWL pattern allows to derive theoretical properties on sparsity and clustering by OSCAR and OWL penalized estimator. Specifically, the OWL pattern of a given vector provides: the sign of its components (positive, negative or null), the clusters (indices of components equal in absolute value) and the hierarchy between the clusters. In this article we give some conditions under which OWL estimator recovers the OWL pattern of an unknown parameter of regression coefficients. Finally, numerical experiments illustrate that when some columns of the design are almost equal, OWL estimator outperforms LASSO estimator for recovering the unknown parameter of regression coefficients.

Keywords : OSCAR, OWL estimator, ordered weighted ℓ_1 norm, OWL pattern recovery

1 Introduction

Let us consider the linear regression model

$$Y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and ε is a random noise having iid $\mathcal{N}(0,\sigma^2)$ components. Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008) is a penalized estimator solution of the following optimization problem:

$$\underset{b\in\mathbb{R}^p}{\text{Argmin}}\ \frac{1}{2}\|Y - Xb\|_2^2 + \gamma_1\|b\|_1 + \gamma_2 \sum_{1\leq i<j\leq p} \max\{|b_i|,|b_j|\},\ \text{where } \gamma_1 > 0, \gamma_2 \geq 0.$$

In the acronym OSCAR, the letter C, as "Clustering", refers to the fact that some components of this estimator can be equal in absolute value. This property can be intuitively illustrated by drawing ellipses representing level curves of the function $b \in \mathbb{R}^p \mapsto \|Y - Xb\|_2^2$ together with balls of the penalty term $b \in \mathbb{R}^p \mapsto \gamma_1\|b\|_1 + \gamma_2 \sum_{1\leq i<j\leq p} \max\{|b_i|,|b_j|\}$ (see Figure 2 in Bondell and Reich (2008)). Clustering property is a specificity of OSCAR (and the related OWL estimator) and this property no longer holds when $\gamma_2 = 0$. Indeed in this case, the penalty term is $\gamma_1\|.\|_1$ thus OSCAR coincides with the Least Absolute Shrinkage and Selection Operator (LASSO) (Chen and Donoho, 1994; Tibshirani, 1996). LASSO, as well as OSCAR, exhibits some null components (Tibshirani, 2013; Osborne et al., 2000; Schneider and Tardivel, 2020) but LASSO does not exhibit clusters. Note that contrarily to fused LASSO (Tibshirani et al., 2005), a cluster for OSCAR does not have, in broad generality, adjacent components.

*Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000 Dijon, France. Xavier.Dupuis@u-bourgogne.fr Patrick.Tardivel@u-bourgogne.fr

Since the seminal article of Bondell and Reich (2008), OSCAR estimator have been extended by the penalized estimator: Pairwise Absolute Clustering and Sparsity (PACS) Sharma et al. (2013). Other extensions of OSCAR are derived in Zeng and Figueiredo (2014); Negrinho and Martins (2014); Bogdan et al. (2015). In these articles OSCAR is generalized as a penalized estimator, based on the Ordered Weighted ℓ_1 norm (OWL norm), solution of the following optimization problem:

$$\underset{b\in\mathbb{R}^p}{\text{Argmin}}\ \frac{1}{2}\|Y-Xb\|_2^2+\gamma\sum_{i=1}^{p}w_i|b|_{\downarrow i}\ \text{where}\ \gamma>0, w_1>0\ \text{and}\ w_1\geq\cdots\geq w_p\geq 0. \tag{1}$$

For OSCAR, $w_1,\ldots,w_p$ is an arithmetic sequence but sequences which are not arithmetic are also relevant. For instance, for the Sorted L One Penalized Estimation (SLOPE) (Bogdan et al., 2015) an appropriate sequence is $w_i=\sigma z(1-\frac{i\alpha}{2p}), i=1,\ldots,p$ where $z(\cdot)$ represents a quantile of the $\mathcal{N}(0,1)$ distribution. This last sequence, related to the Benjamini-Hochberg multiple testing procedure, allows to derive a procedure based on SLOPE controlling the false discovery rate when the design X is orthogonal (Bogdan et al., 2015). From now on, we call OWL estimator an estimator solution of the problem (1)

1.1 OWL pattern

The notion of OWL pattern first introduced by Schneider and Tardivel (2020) allows to describe the structure induced of an OWL estimator by extracting from a given vector:

1. the sign of the components (positive, negative or null),
2. the clusters (indices of components equal in absolute value),
3. the hierarchy between the clusters.

More precisely an OWL pattern is a vector in $\mathbb{Z}^p$ coding points 1), 2) and 3) as defined hereafter.

Definition 1 (OWL pattern (Schneider and Tardivel, 2020))**.** *We say that a vector $m=(m_1,\ldots,m_p)\in\mathbb{Z}^p$ is an OWL pattern, if either $m=0$, or if $\{j\in\{1,\ldots,p\}:|m_j|=l\}\neq\varnothing$ for all $l\in\{1,\ldots,\|m\|_\infty\}$. We denote the set of all OWL patterns of dimension p by $\mathcal{P}_p^{\text{owl}}$. Moreover, for any $x\in\mathbb{R}^p$, we denote by $\mathrm{p}_{\text{owl}}(x)$ the unique OWL pattern for which the following statements hold:*

1) $\text{sign}(\mathrm{p}_{\text{owl}}(x))=\text{sign}(x)$ *(sign preservation),*

2) $|x_i|=|x_j|\implies|\mathrm{p}_{\text{owl}}(x)_i|=|\mathrm{p}_{\text{owl}}(x)_j|$ *(clusters preservation),*

3) $|x_i|>|x_j|\implies|\mathrm{p}_{\text{owl}}(x)_i|>|\mathrm{p}_{\text{owl}}(x)_j|$ *(hierarchy preservation).*

Example 1. *For* $x=(4.7,-4.7,0,1.8,4.7,-1.8)$ *we have* $\mathrm{p}_{\text{owl}}(x)=(2,-2,0,1,2,-1)$. *For* $z=(1.2,-2.3,3.5,1.2,2.3,-3.5)$ *we have* $\mathrm{p}_{\text{owl}}(z)=(1,-2,3,1,2,-3)$.

Hereafter we remind an important proposition related to the notion of OWL pattern and to sub-differential of the OWL norm.

Proposition 1 (Theorem 4 in Schneider and Tardivel (2020))**.** *Let $w=(w_1,\ldots,w_p)$ where $w_1>\cdots>w_p>0$ and $x,z\in\mathbb{R}^p$. We have $\mathrm{p}_{\text{owl}}(x)=\mathrm{p}_{\text{owl}}(z)$ if and only if $\partial_{\|.\|_w}(x)=\partial_{\|.\|_w}(z)$.*

From now on, we always assume that $w_1>\cdots>w_p>0$ so that the OWL pattern characterizes the sub-differential of the OWL norm. Given $X\in\mathbb{R}^{n\times p}$, $Y\in\mathbb{R}^n$ and $\beta\in\mathbb{R}^p$, we aim at recovering $\mathrm{p}_{\text{owl}}(\beta)$ with a solution $\widehat{\beta}$ of (1). Pattern recovery by OWL estimator refers to the event $\mathrm{p}_{\text{owl}}(\widehat{\beta})=\mathrm{p}_{\text{owl}}(\beta)$ and accessibility, defined hereafter, is a key concept related to this event (a similar notion of accessibility by LASSO is given in Sepehri and Harris (2017)).

Definition 2 (Accessible OWL pattern)**.** *Let $X \in \mathbb{R}^{n\times p}$, $w = (w_1, \dots, w_p)$. We say that $m \in \mathcal{P}_p^{\mathrm{owl}}$ is an accessible OWL pattern if one may pick $Y \in \mathbb{R}^n$, $\gamma > 0$ and $\widehat{\beta}$ solution of (1) for which $\mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m$.*

Accessible OWL patterns have been characterized analytically and geometrically in Theorem 5 in Schneider and Tardivel (2020). Specifically, one may pick $Y \in \mathbb{R}^n$, $\gamma > 0$ and $\widehat{\beta}$ solution of (1) for which $\mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m$ if and only if one of the following statements holds:

Analytical characterization: for all $z \in \mathbb{R}^p$ such that $Xz = Xm$ we have $\|z\|_w \geq \|m\|_w$.

Geometrical characterization: the vector space $\mathrm{row}(X) := \{X'z : z \in \mathbb{R}^n\}$ intersects the subdifferential $\partial_{\|.\|_w}(m)$.

It follows from the definition that when $\mathrm{p}_{\mathrm{owl}}(\beta)$ is not accessible then OWL estimator cannot recover the pattern of β and thus the event $\mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = \mathrm{p}_{\mathrm{owl}}(\beta)$ has a null probability. However, accessibility of $\mathrm{p}_{\mathrm{owl}}(\beta)$ does not mean that the probability of pattern recovery is large. In this article we introduce a stronger condition on β than accessibility called noiseless pattern recovery. Actually, Theorem 3 proves that i) this condition is necessary for a probability of pattern recovery larger than $1/2$ and ii) under this condition, the probability that OWL estimator recovers $\mathrm{p}_{\mathrm{owl}}(\beta)$ can be arbitrarily close to 1 as soon as gaps between unique absolute values of β are large.

In the following example, when the design matrix X is orthogonal, one illustrates that OWL estimator may have some null components and some clusters.

Example 2. *When $X'X = I$, one may also explicitly compute the OWL estimator $\widehat{\beta}$. Indeed $\widehat{\beta}$ is the image of $\widehat{\beta}^{\mathrm{ols}} = X'Y$ by the the proximal operator of the OWL norm and this operator has a closed form formula (Bogdan et al., 2015; Tardivel et al., 2020; Dupuis and Tardivel, 2022). This explicit expression gives an analytical way to learn that OWL estimator is sparse and has clusters. The orthogonal projection of the ordinary least squares estimator $\widehat{\beta}^{\mathrm{ols}}$ onto the signed permutahedron $P^{\pm}(w) := \mathrm{conv}\{\sigma_1 w_{\pi(1)}, \dots, \sigma_p w_{\pi(p)} : \sigma_1, \dots, \sigma_p \in \{-1, 1\}, \pi \in \mathcal{S}_p\}$ is equal to $\widehat{\beta}^{\mathrm{ols}} - \widehat{\beta}$ (see Lemma 2 in Minami (2020) or Proposition 3 in Schneider and Tardivel (2020)). This result illustrated on Figure 1 provides a geometrical way to learn that OWL estimator is sparse and has clusters (a similar figure is reported in Tardivel et al. (2020); Skalski et al. (2022) or, for LASSO, in Ewald and Schneider (2020)).*

The particular setup where X is orthogonal is a case study to illustrate pattern recovery properties by OWL estimator (Skalski et al., 2022). However, our article is not restricted to the orthogonal case and we do not consider any restriction on the design matrix X. Theorem 1 gives a characterization of pattern recovery by OWL estimator. This characterization provides some properties on the probability of pattern recovery by OWL estimator. The structure of the article is given hereafter:

Section 2: The main concepts to study pattern recovery by OWL estimator are introduced.

Section 3: Theorem 1 provides a characterization of pattern recovery by OWL estimator.

Section 4: Theorem 2 provides a characterization of pattern recovery by OWL estimator in the noiseless case (when $\varepsilon = 0$).

Section 5: Theorem 3 provides a sharp upper bound for the probability of pattern recovery by OWL estimator in the noisy case (when $\varepsilon \neq 0$).

Section 6: A testing procedure for the null hypothesis "all regression coefficients are equal in absolute value" is derived from OWL estimator. A numerical comparison between LASSO and OWL estimator is also performed.

All the proofs are postponed to the Appendix.

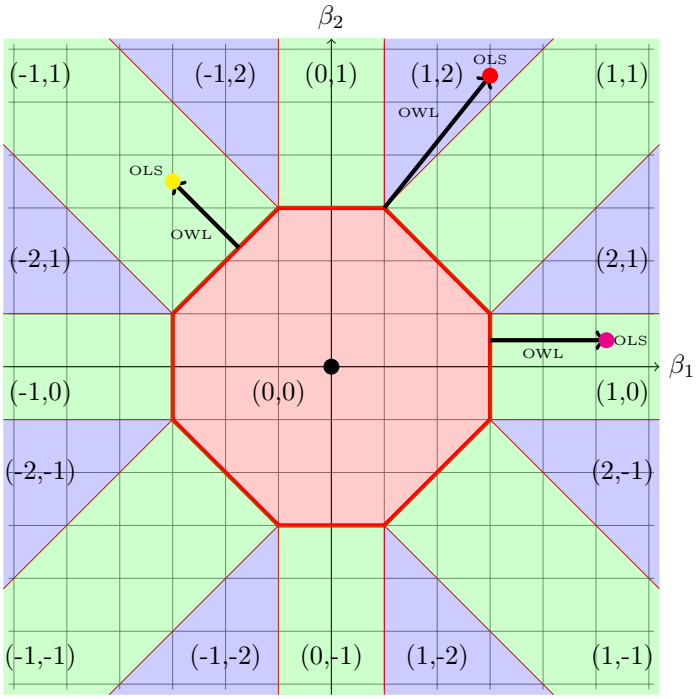

Figure 1: This figure represents the OWL estimator $\widehat{\beta}$ (illustrated by black arrows) depending on the localization of $\widehat{\beta}^{\text{ols}}$ in the particular case where $X'X = I$, $w = (3, 1)$ and $\gamma = 1$. When $\widehat{\beta}^{\text{ols}}$ is the pink point located on the area labelled by $(1, 0)$ then the first component of $\widehat{\beta}$ is positive and the second is null. When $\widehat{\beta}^{\text{ols}}$ is the yellow point located on the area labelled by $(-1, 1)$ then both components of $\widehat{\beta}$ are equal in absolute value; the first component is negative and the second is positive. When $\widehat{\beta}^{\text{ols}}$ is the red point located on the area labelled by $(1, 2)$ then both components of $\widehat{\beta}$ are positive and the first component is smaller than the second. The red polytope is the signed permutahedron $P^{\pm}(w)$ and labels $\mathcal{P}_2^{\text{owl}} = \{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1), \pm(2,1), \pm(2,-1), \pm(1,2), \pm(1,-2)\}$ associated to areas of this figure correspond to OWL patterns in $\mathbb{R}^2$.

2 Notions related to clustering properties by OWL estimator

2.1 Pattern matrix U_m.

Definition 3 (Pattern matrix)**.** *Let $m \neq 0$ be an OWL pattern in $\mathbb{R}^p$ with $k = \|m\|_\infty$ non-null clusters. The* pattern matrix $U_m \in \mathbb{R}^{p \times k}$ *is defined as follows:*

$$\forall i \in \{1, \ldots, p\}, \forall j \in \{1, \ldots, k\}, (U_m)_{ij} = \text{sign}(m_i)\mathbf{1}_{(|m_i|=k+1-j)}.$$

Example 3. *Let $p = 6$, $m = (3, -1, 2, 2, -3, 0)$ and $|m|_\downarrow = (3, 3, 2, 2, 1, 0)$. Then*

$$U_m = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad U_{|m|_\downarrow} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Denote by $\mathbb{R}^{k+}$ the cone $\{x \in \mathbb{R}^k \mid x_1 > \ldots > x_k > 0\}$. It is clear by Definition 3 that we have

$$\mathrm{p}_{\text{owl}}(x) = m \iff x \in U_m \mathbb{R}^{k+}. \tag{2}$$

2.2 Clustered matrix $\widetilde{X}_m$ and clustered parameter $\widetilde{w}_m$

Definition 4. *Let $X \in \mathbb{R}^{n\times p}$, $w = (w_1, \ldots, w_p)$ and $m \in \mathcal{P}_p^{\mathrm{owl}}$. The* clustered matrix *is defined by $\widetilde{X}_m = XU_m$. The* clustered parameter *is defined by $\widetilde{w}_m = (U_{|m|_\downarrow})'w$.*

Thus an OWL pattern m leads naturally to reduce the dimension of the design matrix X as follows:

- A null component $m_i = 0$ leads to discard the column X_i from the design matrix X.
- A cluster $K \subset \{1, \ldots, p\}$ of m (indices of elements of m equal in absolute value) leads to substitute the family of columns $(X_i)_{i\in K}$ by a single column defined by the signed sum: $\sum_{i\in K} \mathrm{sign}(m_i)X_i$.

Example 4. *Let X, m and w as follows:*

$$X = \begin{pmatrix} 1 & -1 & 2 & 3 \\ -2 & 1 & 0 & 2 \\ 1 & 0 & -2 & 3 \end{pmatrix}, m = (1, 2, -2, 0) \text{ and } w = (w_1, w_2, w_3, w_4).$$

Then the clustered matrix and the clustered parameter are given hereafter:

$$\widetilde{X}_m = \begin{pmatrix} -3 & 1 \\ 1 & -2 \\ 2 & 1 \end{pmatrix} \text{ and } \widetilde{w}_m = \begin{pmatrix} w_1 + w_2 \\ w_3 \end{pmatrix}.$$

2.3 Dual OWL norm, signed permutahedron $P^{\pm}(w)$ and sub-differential

The OWL norm is defined as follows:

$$\forall x \in \mathbb{R}^p, \|x\|_w = \sum_{i=1}^{p} w_i |x|_{\downarrow i},$$

where $|x|_{\downarrow 1} \geq \cdots \geq |x|_{\downarrow p}$ are the sorted components of x with respect to the absolute value. The dual OWL norm has an explicit expression given in Zeng and Figueiredo (2014); Negrinho and Martins (2014); Bogdan et al. (2015) and reminded hereafter:

$$\forall x \in \mathbb{R}^p, \|x\|_w^* = \max\left\{ \frac{|x|_{\downarrow 1}}{w_1}, \frac{\sum_{i=1}^{2} |x|_{\downarrow i}}{\sum_{i=1}^{2} w_i}, \ldots, \frac{\sum_{i=1}^{p} |x|_{\downarrow i}}{\sum_{i=1}^{p} w_i} \right\}.$$

The unit ball of the dual OWL norm, also called signed permutahedron Schneider and Tardivel (2020); Negrinho and Martins (2014); Godland and Kabluchko (2020), can be written as a V-polytope as follows:

$$P^{\pm}(w) := \mathrm{conv}\{(\sigma_1 w_{\pi(1)}, \ldots, \sigma_p w_{\pi(p)})' : \sigma_1, \ldots, \sigma_p \in \{-1, 1\}, \pi \in \mathcal{S}_p\}.$$

Geometrical descriptions of the sub-differential of the OWL norm at $x \in \mathbb{R}^p$ have been given in the particular case where $x_1 \geq \cdots \geq x_p \geq 0$. In this setup, the sub-differential is a cartesian product of permutahedra when $x_p > 0$ or a cartesian product of permutahedra with a signed permutahedron when $x_p = 0$ (Tardivel et al., 2020; Dupuis and Tardivel, 2022; Schneider and Tardivel, 2020). Proposition 2 provides an analytic descriptions of the sub-differential of the OWL norm.

Proposition 2. *Let $x \in \mathbb{R}^p$ and $m = \mathrm{p}_{\mathrm{owl}}(x)$ then, we have the following equalities:*

$$\begin{aligned} \partial_{\|.\|_w}(x) &= \{v \in \mathbb{R}^p : \|z\|_w \geq \|x\|_w + v'(z - x)\ \forall z \in \mathbb{R}^p\}, \\ &= \{v \in \mathbb{R}^p : \|v\|_w^* \leq 1 \text{ and } v'x = \|x\|_w\}, \\ &= \{v \in \mathbb{R}^p : \|v\|_w^* \leq 1 \text{ and } U_m'v = \widetilde{w}_m\}. \end{aligned}$$

When $x \neq 0$, the smallest affine space containing $\partial_{\|.\|_w}(x)$ is given by

$$\mathrm{aff}(\partial_{\|.\|_w}(x)) = \{v \in \mathbb{R}^p : U_m'v = \widetilde{w}_m\}. \tag{3}$$

2.4 Characterization of OWL estimator

OWL estimator is a minimizer of the following optimization problem:

$$S_{X,\gamma\|.\|_w}(Y) := \underset{b\in\mathbb{R}^p}{\operatorname{Argmin}} \frac{1}{2}\|Y - Xb\|_2^2 + \gamma\sum_{i=1}^{p} w_i|b|_{\downarrow i} \text{ where } \gamma > 0. \tag{4}$$

In this article we do not assume that $S_{X,\gamma\|.\|_w}(Y)$ contains a unique element and potentially $S_{X,\gamma\|.\|_w}(Y)$ can be a non-trivial compact and convex set. Note however that cases in which $S_{X,\gamma\|.\|_w}(Y)$ is not a singleton are pathological. Indeed, the set of matrices $X \in \mathbb{R}^{n\times p}$ for which there exists a $Y \in \mathbb{R}^n$ where $S_{X,\gamma\|.\|_w}(Y)$ is not a singleton is negligible for the Lebesgue measure (Schneider and Tardivel, 2020). Clearly, the OWL estimator satisfies the following characterization:

$$\widehat{\beta} \in S_{X,\gamma\|.\|_w}(Y) \Leftrightarrow \frac{1}{\gamma}X'(Y - X\widehat{\beta}) \in \partial_{\|.\|_w}(\widehat{\beta}).$$

3 Characterization of pattern recovery by OWL estimator

Given $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $w = (w_1, \ldots, w_p)$, $\gamma > 0$, $m \in \mathcal{P}_p^{\text{owl}}$, where $m \neq 0$ and $\widehat{\beta} \in S_{X,\gamma\|.\|_w}(y)$ we want to characterize the equality $\mathrm{p}_{\text{owl}}(\widehat{\beta}) = m$. Such a characterization will be useful to provide a necessary and sufficient condition for pattern recovery by OWL estimator in the noiseless case in Section 4, as well as an upper bound on the probability of pattern recovery by OWL estimator in the noisy case in Section 5. Let us begin with the following statements:

Cone condition: The fitted value $X\widehat{\beta}$ lies in $\operatorname{cone}(\widetilde{X}_m) := \{\widetilde{X}_m\alpha : \alpha \in \mathbb{R}^{k+}\}$.

Indeed, $\widetilde{X}_m = XU_m$ and $\widehat{\beta} = U_m z$ for some $z \in \mathbb{R}^{k+}$.

Sub-differential condition: $\frac{1}{\gamma}X'(y - X\widehat{\beta}) \in \partial_{\|.\|_w}(\widehat{\beta}) = \partial_{\|.\|_w}(m)$.

It follows from the characterization of an OWL estimator and Proposition 1.

Based on the facts above, we derive in Theorem 1 a characterization of pattern recovery by OWL estimator. From now on A^+ denotes the Moore-Penrose pseudo-inverse of a matrix A.

Theorem 1. *Let $y \in \mathbb{R}^p$, $\gamma > 0$, $w = (w_1, \ldots, w_p)$, $m \neq 0 \in \mathcal{P}_p^{\text{owl}}$ be an OWL pattern having $k := \|m\|_\infty$ clusters, $k \geq 1$. Let U_m and $\widetilde{X}_m := XU_m$ be, respectively, the corresponding pattern matrix and clustered matrix. Let $\operatorname{cone}(\widetilde{X}_m) = \{\widetilde{X}_m\alpha : \alpha \in \mathbb{R}^{k+}\}$ and $\widetilde{P}_m := (\widetilde{X}'_m)^+\widetilde{X}'_m = \widetilde{X}_m\widetilde{X}_m^+$ be the projector onto $\operatorname{col}(\widetilde{X}_m)$. Then, the following conditions:*

$$\begin{aligned}
&\begin{cases} \xi := \widetilde{P}_m y - \gamma(\widetilde{X}'_m)^+\widetilde{w}_m \in \operatorname{cone}(\widetilde{X}_m) & \text{(cone condition)} \\ \zeta := X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma}X'(I - \widetilde{P}_m)y \in \partial_{\|.\|_w}(m) & \text{(sub-differential condition)} \end{cases} \\
\Leftrightarrow\ &\begin{cases} \xi \in \operatorname{cone}(\widetilde{X}_m) & \text{(cone condition)} \\ \|\zeta\|_w^* \leq 1 \text{ and } \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m & \text{(sub-differential condition)} \end{cases}
\end{aligned}$$

are equivalent to the existence of $\widehat{\beta} \in S_{X,\|.\|_w}(y)$ for which $\mathrm{p}_{\text{owl}}(\widehat{\beta}) = m$.

Remark 1. *Clearly when the characterization given in Theorem 1 occurs then, by construction, m is an accessible OWL pattern. The sub-differential condition, $X'[(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma}(I - \widetilde{P}_m)y] \in \partial_{\|.\|_w}(m)$ implying that $\operatorname{row}(X)$ intersects $\partial_{\|.\|_w}(m)$, corroborates this fact.*

4 Pattern recovery by OWL estimator in the noiseless case

The notion of pattern recovery in the noiseless case were recently defined in Tardivel et al. (2021) for a broad class of estimators including LASSO, generalized LASSO, clustered LASSO, OSCAR, CAPS, OWL estimator... Hereafter, we remind the notion of pattern recovery in the noiseless case by OWL.

Definition 5 (Noiseless pattern recovery)**.** *Let $X \in \mathbb{R}^{n\times p}, w = (w_1, \dots, w_p)$ and $\beta \in \mathbb{R}^p$. We say that OWL estimator recovers the pattern of β in the noiseless case ($\varepsilon = 0$) when*

$$\exists \gamma > 0 \ \exists \widehat{\beta} \in S_{X,\gamma\|.\|_w}(X\beta) \text{ such that } \mathrm{p_{owl}}(\widehat{\beta}) = \mathrm{p_{owl}}(\beta). \tag{5}$$

Example 5. *We give two illustrations of noiseless pattern recovery by OWL estimator in the particular case where $w = (4, 2)$, $\beta = (5, 0)$, $\bar{\beta} = (5, 3)$ and $X \in \mathbb{R}^{n\times 2}$ such that*

$$X'X := \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

Figure 2 left (resp. right) illustrates that pattern recovery by OWL estimator does not occur (resp. occurs) for β (resp. for $\bar{\beta}$). Note that, in this setup, clearly the OWL estimator is unique (since $\ker(X) = \{0\}$); we denote by $\widehat{\beta}(\gamma)$ the unique element of $S_{X,\gamma\|.\|_w}(X\beta)$ and the OWL solution path refers to the function $\gamma > 0 \mapsto \widehat{\beta}(\gamma)$.

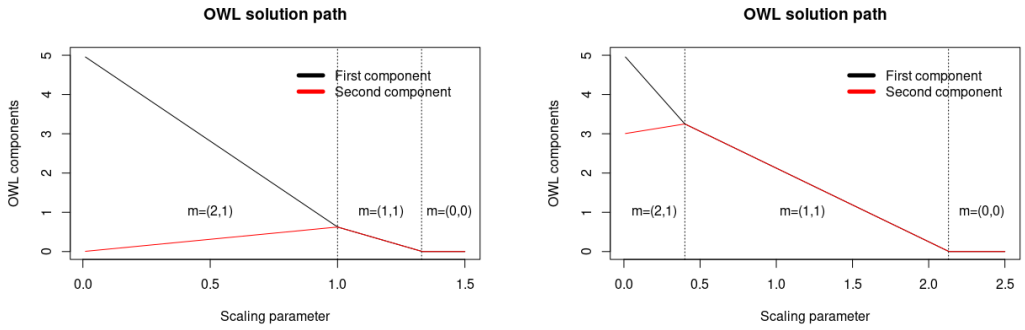

Figure 2: On the left the signal is $\beta = (5, 0)$. Based on this figure one may observe that the pattern of β cannot be recovered by OWL estimator in the noiseless case (despite the fact that $\widehat{\beta}(\gamma)$ tends to β when γ tends to 0). Indeed, for $\gamma \in (0, \gamma_1)$ (where $\gamma_1 \approx 1$) we have $\mathrm{p_{owl}}(\widehat{\beta}(\gamma)) = (2, 1)$; when $\gamma \in [\gamma_1, \gamma_2)$ (where $\gamma_2 \approx 1.33$) we have $\mathrm{p_{owl}}(\widehat{\beta}(\gamma)) = (1, 1)$ and when $\gamma > \gamma_2$ then $\widehat{\beta}(\gamma) = 0$. Consequently, for every $\gamma > 0$ we have $\mathrm{p_{owl}}(\widehat{\beta}(\gamma)) \neq \mathrm{p_{owl}}(\beta) = (1, 0)$. On the right the signal is $\bar{\beta} = (5, 3)$. Based on this figure one may observe that $\mathrm{p_{owl}}(\bar{\beta})$ is recovered by OWL estimator in the noiseless case. Indeed, for $\gamma \in (0, \gamma_1)$ (where $\gamma_1 \approx 0.4$) we have $\mathrm{p_{owl}}(\widehat{\beta}(\gamma)) = (2, 1) = \mathrm{p_{owl}}(\bar{\beta})$.

Theorem 2. *Let $X \in \mathbb{R}^{n\times p}$, $w = (w_1, \dots, w_p)$ and $\beta \in \mathbb{R}^p$ where $\mathrm{p_{owl}}(\beta) = m \neq 0$. A necessary and sufficient condition for noiseless pattern recovery by OWL estimator is $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|^*_w \leq 1$ and $\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m$ (or equivalently $X'(\widetilde{X}'_m)^+\widetilde{w}_m \in \partial_{\|.\|_w}(m)$). Moreover, when this is the case, there exists $\gamma_0 > 0$ such that for all $\gamma \in (0, \gamma_0)$ there exists $\widehat{\beta} \in S_{X,\gamma\|.\|_w}(X\beta)$ for which $\mathrm{p_{owl}}(\widehat{\beta}) = m$.*

Note that when $S_{X,\gamma\|.\|_w}(X\beta)$ has a unique element, denoted by $\widehat{\beta}(\gamma)$, then $\lim_{\gamma\to 0} \mathrm{p_{owl}}(\widehat{\beta}(\gamma)) = \mathrm{p_{owl}}(\beta)$ is equivalent to the pattern recovery of β in noiseless case by OWL estimator.

Remark 2. *By definition, when the pattern recovery in the noiseless case occurs then the accessibility condition occurs for m. The condition $X'(\widetilde{X}'_m)^+\widetilde{w}_m \in \partial_{\|.\|_w}(m)$, implying that $\mathrm{row}(X)$ intersects $\partial_{\|.\|_w}(m)$, corroborates this fact. This fact can be also deduced from Theorem 5 in Vaiter et al. (2015).*

Remark 3. *Theorem 2 for OWL estimator is the analog of Theorem 7.1 in Bühlmann and Van De Geer (2011) dealing with support recovery by LASSO in the noiseless case (see also Theorem 2 in Fuchs (2004)). Actually, the inequality* $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* \leq 1$ *and* $\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m$ *is similar to the irrepresentability condition for LASSO (Fuchs, 2004; Zhao and Yu, 2006; Zou, 2006). This similarity is clear when* $\ker(\widetilde{X}_m) = \{0\}$ *since in this setup,* $\widetilde{X}'_m(\widetilde{X}'_m)^+$ *is the identity matrix,* $X'(\widetilde{X}'_m)^+ = X'\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^{-1}$ *and consequently the noiseless pattern recovery by OWL estimator is equivalent to* $\|X'\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^{-1}\widetilde{w}_m\|_w^* \leq 1$.

Example 6. *Some examples are reported hereafter:*

Example from Figure 2 (left): *We observe on the left picture in Figure 2 that the noiseless pattern recovery does not occur when* $\beta = (5, 0)$. *To corroborate this fact, let us check that* $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* > 1$ *where* $m = \mathrm{p}_{\mathrm{owl}}(\beta) = (1, 0)$. *Let* $X = (X_1|X_2)$ *then* $\widetilde{X}_m = X_1$ *(thus* $\widetilde{X}'_m\widetilde{X}_m = 1$*) and* $\widetilde{w}_m = w_1 = 4$ *therefore*

$$\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = \|X'\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^{-1}\widetilde{w}_m\|_w^* = \|4X'XU_{(1,0)}\|_w^* = \|(4, 2.4)\|_w^* = 6.4/6 > 1.$$

Example from Figure 2 (right): *We observe on the right picture in Figure 2 that the noiseless pattern recovery does not occur when* $\bar{\beta} = (5, 3)$. *To corroborate this fact, let us check that* $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* \leq 1$ *and* $\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m$. *Since* $\mathrm{p}_{\mathrm{owl}}(\bar{\beta}) = (2, 1) = m$, *we have* $\widetilde{X}_m = X$ *and* $\widetilde{w}_m = w$. *In particular* $\ker(\widetilde{X}_m) = \{0\}$ *and* $\widetilde{X}'_m(\widetilde{X}'_m)^+ = I$. *Finally*

$$\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = \|X'X(X'X)^{-1}w\|_w^* = \|w\|_w^* = 1 \leq 1.$$

Orthogonal case: *Let* $X \in \mathbb{R}^{n\times p}$ *such that* $X'X = I$, $w = (w_1, \ldots, w_p)$, $|\beta_1| = \cdots = |\beta| > 0$ *and thus* $\mathrm{p}_{\mathrm{owl}}(\beta) = (\sigma_1, \ldots, \sigma_p) = m$ *for some* $\sigma_1, \ldots, \sigma_p \in \{-1, 1\}^p$. *Then* $\widetilde{w}_m = \sum_{i=1}^p w_i$ *and consequently*

$$\begin{aligned}\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* &= \|X'\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^{-1}\widetilde{w}_m\|_w \\ &= \|X'XU_m\underbrace{(U'_mX'XU_m)^{-1}}_{=1/p}\widetilde{w}_m\|_w^* = \left\|\left(\sigma_1\frac{\sum_{i=1}^p w_i}{p}, \ldots, \sigma_p\frac{\sum_{i=1}^p w_i}{p}\right)\right\|_w^*.\end{aligned}$$

Since $(\sigma_1\sum_{i=1}^p w_i/p, \ldots, \sigma_p\sum_{i=1}^p w_i/p)$ *is the isobarycenter of the sub-differential* $\partial_{\|.\|_w}(\beta) = \mathrm{conv}\{(\sigma_1 w_{\pi(1)}, \ldots, \sigma_p w_{\pi(p)}) : \pi \in \mathcal{S}_p\}$ *(which is a face of the signed permutahedron), we deduce that* $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = 1$. *More generally, in the orthogonal case, for an arbitrary* $\beta \in \mathbb{R}^p$ $X'(\widetilde{X}'_m)^+\widetilde{w}_m$ *is the isobarycenter of the set* $\partial_{\|.\|_w}(\beta)$ *where* $m = \mathrm{p}_{\mathrm{owl}}(\beta)$. *Consequently, the noiseless pattern recovery by OWL estimator occurs for* β.

Strongly correlated covariates: *Let* $X = (X_1|\ldots|X_p) \in \mathbb{R}^{n\times p}$ *and* $w = (w_1, \ldots, w_p)$ *such that:*

- *columns are normalized:* $\|X_1\|_2 = \cdots = \|X_p\|_2 = 1$*;*
- r *columns of* X *are equal (for some* $r \leq p$*) and without loss of generality we set* $X_1 = \cdots = X_r$.
- $|X'_jX_1| < rw_j/\sum_{i=1}^r w_i$ *for all* $j > r$.

A similar setup in which "covariates are strongly correlated" is given in Figueiredo and Nowak (2016) (see Figure 1). Let $\beta \in \mathbb{R}^p$ *having* r *non-null components such that* $\beta_1 = \cdots = \beta_r > 0$ *(thus* m *has* r *non-null components and* $m_1 = \cdots = m_r = 1$*). We have* $\widetilde{X}_m = rX_1$ *and* $\widetilde{w}_m = w_1 + \cdots + w_r$ *and consequently:*

$$\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = \|X'\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^{-1}\widetilde{w}_m\|_w^* \leq \left\|\left(\underbrace{\frac{\sum_{i=1}^r w_i}{r}, \ldots, \frac{\sum_{i=1}^r w_i}{r}}_{\text{r components}}, w_{r+1}, \ldots, w_p\right)\right\|_w^* = 1.$$

Thus the noiseless pattern recovery occurs for β which corroborates Figure 1 in Figueiredo and Nowak (2016).

Pathological example (the null matrix): *We provide very general statements since we do not assume that* $\ker(\widetilde{X}_m) = \{0\}$. *For instance, one may consider the pathological example where X is the zero matrix of dimension $n \times p$ and where $\beta \in \mathbb{R}^p$ with $\beta \neq 0$ (thus $m = \mathrm{p_{owl}}(\beta) \neq 0$). Clearly the pattern of β cannot be recovered in the noiseless case by OWL estimator and this fact is confirmed by Theorem 2. Indeed, in this setup $\widetilde{X}_m$ is the null matrix, thus $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = 0 \leq 1$ but $0 = \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m \neq \widetilde{w}_m$.*

Pathological example (with non-uniqueness): *In some particular cases OWL estimator can recover* $\mathrm{p_{owl}}(\beta) \in \mathcal{P}_p^{\mathrm{owl}}$ *in the noiseless case even if* $\ker(\widetilde{X}_m) \neq \{0\}$. *For instance, let* $X \in \mathbb{R}^{1\times 2}, W \in \mathbb{R}^2$ *and* $\beta \in \mathbb{R}^2$ *such that*

$$X := \begin{pmatrix} 8 & 4 \end{pmatrix}, w := (4,2) \text{ and } \mathrm{p_{owl}}(\beta) := (2,1) = m.$$

Note that $U_m = I_2$ so that $\widetilde{X}_m = X$ (thus clearly $\ker(\widetilde{X}_m) \neq \{0\}$) and $\widetilde{w}_m = w$. We check that $\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* = \|\widetilde{w}_m\|_w^* = 1$ and $\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m$. Indeed, $X'(\widetilde{X}'_m)^+ = \widetilde{X}'_m(\widetilde{X}'_m)^+$ is the projection on $\mathrm{col}(\widetilde{X}'_m)$ and clearly $\widetilde{w}_m \in \mathrm{col}(\widetilde{X}'_m)$. Thus the noiseless pattern recovery occurs namely, for $\gamma > 0$ small enough one may pick $\widehat{\beta} \in S_{X,\gamma\|\cdot\|_w}(X\beta)$ such that $\mathrm{p_{owl}}(\widehat{\beta}) = \mathrm{p_{owl}}(\beta)$. Note that this example is rather pathological since $S_{X,\gamma\|\cdot\|_w}(X\beta)$ is not reduced to a singleton. Indeed for $h \in \ker(X)$ small enough we have both $X(\widehat{\beta}+h) = X\widehat{\beta}$ and $\mathrm{p_{owl}}(\widehat{\beta}+h) = \mathrm{p_{owl}}(\widehat{\beta})$ and thus $\widehat{\beta}+h \in S_{X,\gamma\|\cdot\|_w}(X\beta)$.

5 Pattern recovery in the noisy case

Very recently, Tardivel et al. (2021) have shown that pattern recovery in the noiseless case is a necessary condition for pattern recovery by penalized estimators (including LASSO, generalized LASSO, clustered LASSO, OSCAR, PACS, OWL estimator...) with a probability larger than 1/2 when ε is no longer 0 (see Theorem 2 in Tardivel et al. (2021)). In this section we focus on a sufficient condition for pattern recovery by OWL estimator rather than on necessary conditions. The sub-differential condition, given in Theorem 1, is related to the following Gaussian vector:

$$\zeta_\gamma := X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P})Y = X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P})\varepsilon. \tag{6}$$

This condition naturally leads to introduce the probability $\mathbb{P}_\varepsilon(\|\zeta_\gamma\|_w^* \leq 1)$ as an upper bound on the probability of pattern recovery by OWL estimator. Point 1 in Theorem 3 establishes this upper bound and this result implies that when the noiseless recovery condition does not hold for OWL estimator the probability of pattern recovery is smaller than 1/2. Point 2 shows that the probability of pattern recovery matches with the upper bound when gaps between unique absolute value of β are large enough (*i.e.* when r tends to $+\infty$ in Theorem 3). Based on point 2, one easily observes that OWL estimator can recover the pattern with a probability arbitrarily close to 1 provided that the noiseless recovery by OWL estimator holds. In particular, point 3 shows that the probability of pattern recovery by OWL estimator may tend to 1 when the noiseless recovery condition occurs.

Theorem 3. *Let $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p, \beta \neq 0$ with $\mathrm{p_{owl}}(\beta) = m$, ε has a $\mathcal{N}(0,\sigma^2 I)$ distribution and $w = (w_1, \ldots, w_p)$.*

1. **Upper bound:** *Let $\gamma > 0$ be a fixed scaling parameter and ζ_γ be the Gaussian vector described in (6). Then, we have the following upper bound for the probability of pattern recovery by OWL estimator:*

$$\mathbb{P}_\varepsilon(\exists \widehat{\beta} \in S_{X,\gamma\|\cdot\|_w}(Y) \text{ such that } \mathrm{p_{owl}}(\widehat{\beta}) = m) \leq \begin{cases} \mathbb{P}_\varepsilon\left(\|\zeta_\gamma\|_w^* \leq 1\right) \\ 0 \text{ if } \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m \neq \widetilde{w}_m \end{cases}.$$

2. Sharpness of the upper bound: *To prove the sharpness of the upper bound we consider a sequence of signals* $(\beta^{(r)})_{r\geq 1}$ *with pattern* m*:*

$$\beta^{(r)} = U_m s^{(r)} \quad \text{with } s^{(r)} \in \mathbb{R}^{k+} \text{ and } k = \|m\|_\infty,$$

whose strength is increasing in the following sense:

$$\Delta^{(r)} = \min_{1\leq i<k} (s_i^{(r)} - s_{i+1}^{(r)}) \underset{r\to+\infty}{\longrightarrow} \infty, \text{ with the convention } s_{k+1}^{(r)} = 0.$$

Let $Y^{(r)} = X\beta^{(r)} + \varepsilon$*. The previous upper bound is asymptotically reached when* r *tends to* $+\infty$*:*

$$\lim_{r\to+\infty} \mathbb{P}_\varepsilon \left(\exists \widehat{\beta} \in S_{X,\gamma\|.\|_w}(Y^{(r)}) \text{ such that } \mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m\right) = \begin{cases} \mathbb{P}_\varepsilon \left(\|\zeta_\gamma\|_w^* \leq 1\right) \\ 0 \text{ if } \widetilde{X}'_m(\widetilde{X}'_m)^+ \widetilde{w}_m \neq \widetilde{w}_m \end{cases}.$$

3. Pattern recovery with a probability tending to 1: *Let* $Y^{(r)} = X\beta^{(r)} + \varepsilon$ *as above and in addition* $\lim_{r\to+\infty} \gamma_r = +\infty$ *and* $\lim_{r\to+\infty} \gamma_r/\Delta_r = 0$*. If* $X'(\widetilde{X}'_m)^+ \widetilde{w}_m \in ri(\partial_{\|.\|_w}(m))$*, then*

$$\lim_{r\to+\infty} \mathbb{P}_\varepsilon(\exists \widehat{\beta} \in S_{X,\gamma_r\|.\|_w}(Y) \text{ such that } \mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m) = 1.$$

Remark 4. *Some consequences of Theorem 3 are listed hereafter:*

- *Because the unit ball of the dual OWL norm is convex, when* $\|X'(\widetilde{X}'_m)^+ \widetilde{w}_m\|_w^* > 1$ *then, independently on* $\gamma > 0$*, the probability on pattern recovery is smaller than* $1/2$ *namely*

$$\mathbb{P}_\varepsilon(\exists \widehat{\beta} \in S_{X,\gamma\|.\|_w}(Y) \text{ such that } \mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m) \leq 1/2.$$

  This inequality corroborates Theorem 2 in Tardivel et al. (2021). For LASSO, a similar inequality on the probability of sign recovery is given in Wainwright (2009).

- *When* $\widetilde{P} = \widetilde{X}_m \widetilde{X}_m^+ \neq I$ *and* $X'(\widetilde{X}'_m)^+ \widetilde{w}_m \in \mathrm{ri}(\partial_{\|.\|_w}(m))$ *(thus* $\widetilde{X}'_m(\widetilde{X}'_m)^+ \widetilde{w}_m = \widetilde{w}_m$*) the function*

$$B : \gamma > 0 \mapsto \mathbb{P}_\varepsilon \left(\|\zeta_\gamma\|_w^* \leq 1\right)$$

  is continuous, increasing and satisfies $\lim_{\gamma\to 0} B(\gamma) = 0$ *and* $\lim_{\gamma\to+\infty} B(\gamma) = 1$*. Consequently, one may select* γ *so that the probability of pattern recovery by OWL estimator is upper bounded by* η *for* $\eta < 1$ *arbitrarily close to* 1 *(a similar sharp upper bound for sign recovery by LASSO is given in Tardivel and Bogdan (2022)).*

6 Numerical experiments

6.1 Testing procedure when the design is orthogonal

Based on OWL estimator $\widehat{\beta}$ (which is uniquely defined when X is orthogonal) we would like to test:

$$\mathcal{H}^0 : |\beta_1| = \cdots = |\beta_p| = \|\beta\|_\infty \text{ vs } \mathcal{H}^1 : \exists i \in \{1, \ldots, p\}, \, |\beta_i| < \|\beta\|_\infty.$$

Given a sequence $w_1 > \cdots > w_p > 0$, we reject the null hypothesis when $|\widehat{\beta}(\gamma_\alpha)|_{\downarrow 1} > |\widehat{\beta}(\gamma_\alpha)|_{\downarrow p}$, where $\gamma_\alpha > 0$ is an appropriately chosen scaling parameter allowing to control the type I error at level $\alpha \in (0, 1)$.

Selecting γ_α based on the sharp upper bound: According to Example 6 (item entitled "Orthogonal case"), when X is orthogonal we have $X'(\widetilde{X}'_m)^+ \widetilde{w}_m \in ri(\partial_{\|.\|_w}(\mathrm{p}_{\mathrm{owl}}(\beta)))$ and thus, by Theorem 3

one may select γ_α in order to fix the sharp upper bound $\mathbb{P}_\varepsilon(\|S_{\gamma_\alpha}\|_w^* \leq 1)$ at level $1-\alpha$. Note that under the null hypothesis (and when $\beta \neq 0$) we have $\mathrm{p}_{\mathrm{owl}}(\beta) = (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_p)) \in \{-1,1\}^p$ and one may select $\gamma_\alpha > 0$ independently of $\beta \in \mathbb{R}^p$ for which $|\beta_1| = \cdots = |\beta_p|$. Indeed, $\mathbb{P}(\|\zeta_\gamma\|_w^* \leq 1) = \mathbb{P}(\|D\zeta_\gamma\|_w^* \leq 1)$ where D is the diagonal matrix $\mathrm{diag}(\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_p))$. Consequently, without loss of generality, to select $\gamma_\alpha > 0$, one may consider the particular case where $\mathrm{p}_{\mathrm{owl}}(\beta) = (1, \ldots, 1)$. In Theorem 3, ζ_γ is a Gaussian vector having a $\mathcal{N}(X'(\widetilde{X}'_m)^+\widetilde{w}_m, \frac{\sigma^2}{\gamma^2}X'(I - \widetilde{X}_m\widetilde{X}_m^+)X)$ distribution. In the particular setup where $X'X = I$ and $\mathrm{p}_{\mathrm{owl}}(\beta) = (1, \ldots, 1)$ we have

$$X'(\widetilde{X}'_m)^+\widetilde{w}_m = \left(\frac{1}{p}\sum_{i=1}^p w_i, \ldots, \frac{1}{p}\sum_{i=1}^p w_i\right) \text{ and } X'(I - \widetilde{X}_m\widetilde{X}_m^+)X = \begin{pmatrix} 1-1/p & -1/p & \ldots & -1/p \\ -1/p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1/p \\ -1/p & \ldots & -1/p & 1-1/p \end{pmatrix}.$$

Since the distribution of ζ_γ is given, one may pick $\gamma_\alpha > 0$ for which $\mathbb{P}_\varepsilon(\|S_{\gamma_\alpha}\|_w^* \leq 1) = 1-\alpha$. According to Theorem 3, under the null hypothesis and when $\|\beta\|_\infty$ tends to $+\infty$, we have $\mathbb{P}(\mathrm{p}_{\mathrm{owl}}(\widehat{\beta}(\gamma_\alpha)) \neq \mathrm{p}_{\mathrm{owl}}(\beta)) = \alpha$ and thus the type I error is asymptotically controlled at level α by the above procedure.

Prescription for the sequence $w_1 > \cdots > w_p > 0$: The above procedure can be run with any arbitrary sequence for which $w_1 > \cdots > w_p > 0$. Hereafter we suggest a sequence for this procedure. Let Z be a standard Gaussian vector having a $\mathcal{N}(0, \sigma^2 I_p)$ distribution, let $Z_{\downarrow 1} \geq \cdots \geq Z_{\downarrow p}$ be the components of Z ordered by decreasing values and $|Z|_{\downarrow p}$ be the smallest components of Z in absolute value. For any $i \in \{1, \ldots, p\}$, we set $w_i = \sigma\mathbb{E}(Z_{\downarrow i} - Z_{\downarrow p} + |Z|_{\downarrow p})$. The Cesàro sequence $(\frac{1}{i}\sum_{j=1}^i(|\widehat{\beta}^{\mathrm{ols}}|_{\downarrow j} - \gamma w_j))_{1\leq i\leq p}$ is closely related to the explicit expression of OWL estimator $\widehat{\beta}(\gamma)$ when X is orthogonal (Dupuis and Tardivel, 2022; Tardivel et al., 2020). Intuitively, under the null hypothesis, when γ is larger than 1 the Cesàro sequence tends to be increasing, implying $|\widehat{\beta}(\gamma)|_{\downarrow p} = |\widehat{\beta}(\gamma)|_{\downarrow 1}$. Thus one may control the type I error at prescribed level $\alpha \in (0,1)$ by choosing appropriately a scaling parameter γ slightly larger than 1.

Type I error: When the number of regression coefficients is $p = 100$, the standard error of the noise is $\sigma = 1$ and $\alpha = 0.05$, Figure 3 report the type I error as a function of $\|\beta\|_\infty \in [0, 20]$.

Power: Under the alternative hypothesis, we report in Figure 4 the power (the probability to reject the null hypothesis) of the above procedure in the particular case where

$$|\beta|_{\downarrow 1} = \cdots = |\beta|_{\downarrow k} = c > |\beta|_{\downarrow k+1} = \cdots = |\beta|_{\downarrow p} = 0.8c, \tag{7}$$

as a function of $c > 0$.

Mean squared error of the mean-clustered estimator: When one knows for instance that $\beta_{i_1} = \cdots = \beta_{i_l}$, it is intuitive to estimate the common value of these regression coefficients by averaging component of the ordinary least squares estimator over this cluster as follows: $(\widehat{\beta}^{\mathrm{ols}}_{i_1} + \cdots + \widehat{\beta}^{\mathrm{ols}}_{i_l})/l$. OWL estimator may identify some components of β equal in absolute value leading to introduce the mean-clustered estimator defined hereafter. Let $\widehat{\beta}(\gamma_\alpha)$ be an OWL estimator penalized estimator and $\widehat{m} = \mathrm{p}_{\mathrm{owl}}(\widehat{\beta}(\gamma_\alpha))$ be its pattern; then one defines the mean-clustered estimator as follows:

$$\forall i \in \{1, \ldots, p\}\, \widehat{\beta}^{\mathrm{mc}}_i := \begin{cases} \frac{1}{\#\{j:\widehat{m}_j=0\}}\sum_{j:\widehat{m}_j=0}\widehat{\beta}^{\mathrm{ols}}_j \text{ if } \widehat{m}_i = 0 \\ \frac{\mathrm{sign}(\widehat{m}_i)}{\#\{j:|\widehat{m}_j|=l\}}\sum_{j:|\widehat{m}_j|=l}|\widehat{\beta}^{\mathrm{ols}}_j| \text{ if } |\widehat{m}_i| = l > 0 \end{cases}.$$

In Figure 4 we also report the mean square error of the mean-clustered estimator as a function of $c > 0$. Let us first remind that $\mathbb{E}(\|\widehat{\beta}^{\mathrm{ols}} - \beta\|_2^2) = p$ and the James-Stein estimator, defined by $\widehat{\beta}^{\mathrm{js}} := \widehat{\beta}^{\mathrm{ols}} - (p-2)\widehat{\beta}^{\mathrm{ols}}/\|\widehat{\beta}^{\mathrm{ols}}\|_2^2$, satisfies $\mathbb{E}(\|\widehat{\beta}^{\mathrm{js}} - \beta\|_2^2) \leq \mathbb{E}(\|\widehat{\beta}^{\mathrm{ols}} - \beta\|_2^2) = p$ (James and Stein, 1992).

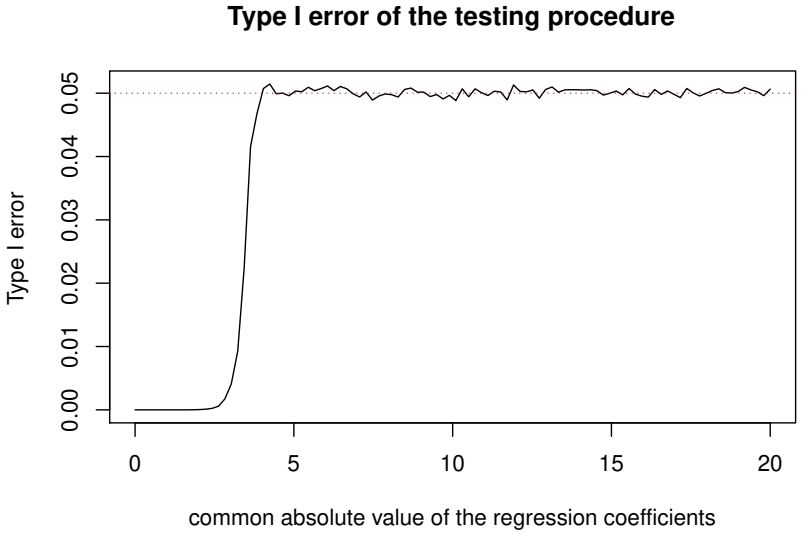

Figure 3: This picture reports the type I error (on the y-axis) of the testing procedure as a function of $\|\beta\|_\infty$ (on the x-axis). Each point from this plot is obtained via 100000 simulations. Theoretically, under the null hypothesis and when $\|\beta\|_\infty$ is infinitely large, the type I error is controlled at level $\alpha = 0.05$. This curve corroborates this fact. Moreover, it seems the type I error is controlled at level $\alpha = 0.05$ for any value of $\|\beta\|_\infty \geq 0$.

Note that the OWL estimator performs poorly for the mean squared error and thus $\mathbb{E}(\|\widehat{\beta}(\gamma_\alpha) - \beta\|_2^2)$ is not reported on Figure 4.

6.2 Noiseless recovery curves and accessibility curves

Numerical experiments where some columns of the design matrix are almost identical were already performed in the literature (Figueiredo and Nowak, 2016). In this particular setup, OWL estimator outperform LASSO for estimating β as illustrated in Figure 1 in Figueiredo and Nowak (2016). Our numerical experiments support these findings.

Hereafter $M_r(a, b)$ denotes a $r \times r$ matrix whose diagonal coefficients are all equal to a and non-diagonal coefficients are all equal to b. For these numerical experiments, we consider the following setting:

1- Distribution of the design: $X = (X_1|\dots|X_p) \in \mathbb{R}^{n\times p}$, where $n = 50$ and $p = 150$, is a Gaussian matrix having iid $\mathcal{N}(0, \Sigma)$ rows and Σ is block diagonal as described hereafter:

$$\Sigma = diag(\underbrace{M_r(1,\rho), \dots, M_r(1,\rho)}_{k \text{ blocks}}, I_{p-kr}).$$

Note that when ρ is close to 1 then, for some $i \in \{0, \dots, k-1\}$, columns $X_{ir+1}, \dots, X_{(i+1)r}$ are almost all equal. In particular we take $\rho = 0.99$ and $r = 10$.

2- OWL pattern: Within a cluster (a set of columns which are almost all equal), we consider that the components of $m := \mathrm{p}_{\mathrm{owl}}(\beta)$ are equal and more precisely we consider the following setting:

$$\underbrace{m_1 = \dots = m_r}_{=k} > \underbrace{m_{r+1} = \dots = m_{2r}}_{=k-1} > \dots > \underbrace{m_{(k-1)r+1} = \dots = m_{kr}}_{=1} > m_{kr+1} = \dots = m_p = 0,$$

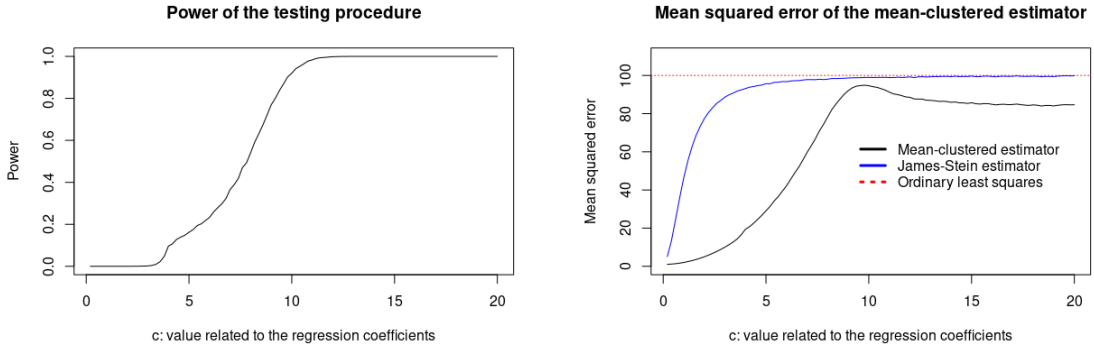

Figure 4: The picture on the left reports the power (on the y-axis) of the testing procedure as a function of the value c given in (7). One may observe that, approximately, the power is 1 when $c > 10$ and the power is smaller than 1 when $c < 10$. The picture on the right reports the mean squared error as a function of c of the mean-clustered estimator, the ordinary least squared estimator and the James-Stein estimator. Note that the mean-clustered estimator outperforms both the ordinary least squares estimator as well as the James-Stein estimator. One may observe that the curve is increasing when $c < 10$ and is slightly decreasing when $c > 10$. Thus, the mean squared error starts to decrease when the power of the procedure is 1, namely when components of the OWL estimator cannot be all equal in absolute value.

3- Sequence w: We take the sequence $w = (w_1, \dots, w_p)$ as in the in the testing procedure when $\sigma = 1$. Namely, for any $i \in \{1, \dots, p\}$, we set $w_i = \mathbb{E}(Z_{\downarrow i} - Z_{\downarrow p} + |Z|_{\downarrow p})$.

For OWL estimator, the noiseless recovery condition as well as the accessibility condition depends on β through $\mathrm{p}_{\mathrm{owl}}(\beta) \in \mathcal{P}_p^{\mathrm{owl}}$. When $\mathrm{p}_{\mathrm{owl}}(\beta) = m$ for some $m \in \mathcal{P}_p^{\mathrm{owl}}$ and when X is a random matrix, the probability that β satisfies the noiseless recovery condition and the probability that β satisfies the accessibility condition are respectively given hereafter:

$$\begin{cases} \mathbb{P}_X(\|X'(\widetilde{X}'_m)^+\widetilde{w}_m\|_w^* \le 1) \\ \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m \end{cases} \quad \text{and } \mathbb{P}_X(\min\{\|\gamma\|_w : X\gamma = Xm\} = \|m\|_w).$$

In practice, in these numerical experiments, the condition $\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m$ (or equivalently, $\widetilde{w}_m \in \mathrm{row}(\widetilde{X}_m)$) always occurs. For LASSO, the noiseless sign recovery condition (more famously known as the irrepresentability condition) as well as the accessibility condition depends on β through $\mathrm{sign}(\beta) \in \{-1, 0, 1\}^p$. Note that when $\mathrm{p}_{\mathrm{owl}}(\beta) = m$ then $\mathrm{sign}(\beta) = \mathrm{sign}(m)$. When X is a random matrix, the probability that β satisfies the noiseless sign recovery condition and the probability that β satisfies the accessibility condition are respectively given hereafter:

$$\mathbb{P}_X(\|X'(X'_I)^+\mathrm{sign}(m_I)\|_\infty \le 1) \text{ and } \mathbb{P}_X(\min\{\|\gamma\|_1 : X\gamma = X\mathrm{sign}(m)\} = \|\mathrm{sign}(m)\|_1),$$

where $I = \mathrm{supp}(m)$ and X_I is the matrix whose columns are $(X_i)_{i \in I}$,

Figure 5 provides these probabilities as a function of the number of clusters k when $\rho = 0.99$.

We consider a linear regression model $Y = X\beta + \varepsilon$ where X is a random matrix as in 1), ε has iid $\mathcal{N}(0, 1)$ entries and β has 40 non-null components as follows:

$$\underbrace{\beta_1 = \cdots = \beta_{10}}_{=80} > \underbrace{\beta_{11} = \cdots = \beta_{20}}_{=60} > \underbrace{\beta_{21} = \cdots = \beta_{30}}_{=40} = \underbrace{\beta_{31} = \cdots = \beta_{40}}_{=20}. \tag{8}$$

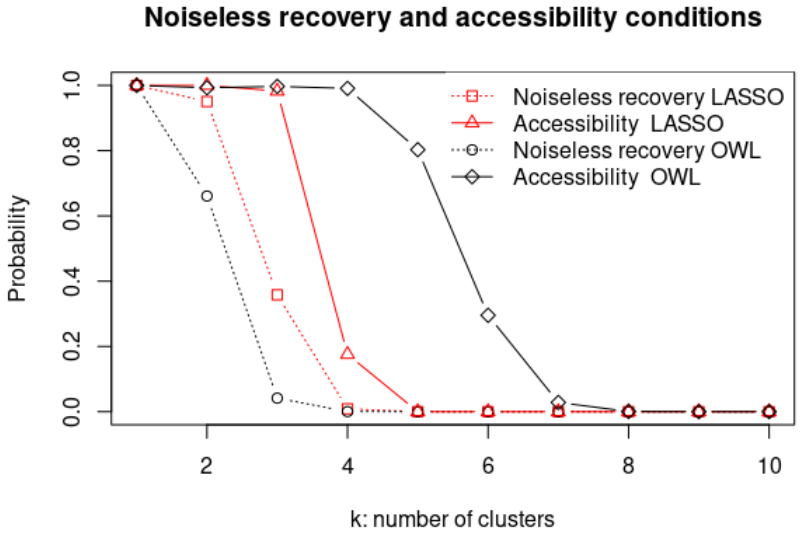

Figure 5: When $X \in \mathbb{R}^{50\times 150}$ is a random matrix described in 1), these curves provide the probability that the noiseless recovery condition occurs and the probability that the accessibility condition occurs as a function of k (the number of clusters) for both LASSO and OWL estimator when $\rho = 0.99$. One may notice that noiseless recovery curves are below accessibility curves; these observations comply with theoretical property: noiseless recovery condition implies accessibility condition. Note that accessibility curves for OWL estimator are above accessibility curves for LASSO. This suggest that, potentially, OWL estimator is a better estimator than LASSO for pattern recovery in the particular setting where some columns are almost equal.

We illustrate, in Figure 6, performance of OWL estimator as well as LASSO for recovering the parameter of regression coefficients β

7 Appendix

7.1 Proof of Proposition 2

Let $w = (w_1, \dots, w_p)$ where $w_1 > \cdots > w_p$. Before starting the proof of Proposition 2, we remind the following facts:

Facts on permutahedron: The permutahedron is the following polytope

$$P(w) := \text{conv}\{(w_{\pi(1)}, \dots, w_{\pi(p)}) : \pi \in \mathcal{S}_p\}.$$

1) The codimension of the permutahedron $P(w)$ is 1 (see for instance Ziegler (2012) page 24).

2) Given $z \in P(w)$ we have $z_1 + \cdots + z_p = w_1 + \cdots + w_p$.

Facts on the OWL norm and on the dual OWL norm: Let $\pi \in \mathcal{S}_p$, $\sigma_1, \dots, \sigma_p \in \{-1, 1\}$ and let us define the following orthogonal transformation:

$$\forall z \in \mathbb{R}^p, \phi(z) = (\sigma_1 z_{\pi(1)}, \dots, \sigma_p z_{\pi(p)}).$$

The following identities are straightforward.

$$\|z\|_w = \|\phi(z)\|_w \quad \forall z \in \mathbb{R}^p \text{ and } \|z\|_w^* = \|\phi(z)\|_w^* \quad \forall z \in \mathbb{R}^p.$$

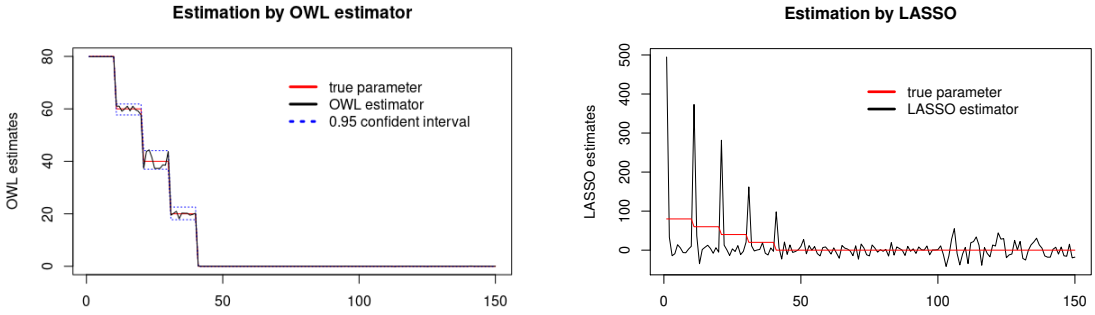

Figure 6: The picture on the left illustrates performance of OWL estimator for recovering β as described in (8). Note that the number of non-null clusters is $k = 4$ and thus according to Figure 5 the noiseless recovery condition does not hold thus OWL estimator cannot recover exactly the pattern of β. However, the accessibility occurs thus OWL estimator may separate clusters (see Tardivel et al. (2021) for additional details). Indeed, for each $i \in \{1, \dots, 150\}$ the probability that $\widehat{\beta}_i$ lies into the blue band is 0.95 thus OWL estimator is a quite accurate estimator of β. The right picture illustrates that LASSO fails to recover β in this setting.

Proof. According to Hiriart-Urruty and Lemaréchal (2004) page 180 we have

$$\partial_{\|.\|_w}(x) := \{v \in \mathbb{R}^p : \|z\|_w \geq \|x\|_w + v'(z - x)\ \forall z \in \mathbb{R}^p\} = \{v \in \mathbb{R}^p : \|v\|_w^* \leq 1 \text{ and } v'x = \|x\|_w\}.$$

The second expression tells us that $\partial_{\|.\|_w}(x)$ is a face of the unit ball for the dual norm. Once we will have established (3), which gives its affine envelope, the last expression of $\partial_{\|.\|_w}(x)$ as the intersection of the ball and its affine envelope will follow. Let us first establish (3) in the particular case where components of x are non-negative and non-increasing. Let $1 \leq k_1 < \cdots < k_l \leq p$ be a subdivision of $\{1, \dots, p\}$ where k_l is the number of non-null components of x and such that

$$x_1 = \cdots = x_{k_1} > x_{k_1+1} = \cdots = x_{k_2} > \cdots > x_{k_{l-1}+1} = \cdots = x_{k_l} > 0.$$

Since components of x are non-increasing, the sub-differential of $\partial_{\|.\|_w}(x)$ is easy to describe as follows (Tardivel et al., 2020; Dupuis and Tardivel, 2022; Schneider and Tardivel, 2020):

$$\partial_{\|.\|_w}(x) = \begin{cases} P(w_1, \dots, w_{k_1}) \times \cdots \times P(w_{k_{l-1}+1}, \dots, w_{k_l}) \text{ if } k_l = p \\ P(w_1, \dots, w_{k_1}) \times \cdots \times P(w_{k_{l-1}+1}, \dots, w_{k_l}) \times P^{\pm}(w_{k_l+1} \dots w_p) \text{ if } k_l < p \end{cases}.$$

Since the codimension of the signed permutahedron is 0 and the codimension of the permutahedron is 1 we have

$$\text{codim}(\partial_{\|.\|_w}(x)) = \text{codim}(P(w_1, \dots, w_{k_1})) + \cdots + \text{codim}(P(w_{k_{l-1}+1}, \dots, w_{k_l})) = l.$$

Moreover when $v \in \partial_{\|.\|_w}(x)$ then, according to the above fact 2), v lies on the following affine space describe hereafter:

$$\begin{cases} v_1 + \cdots + v_{k_1} = w_1 + \cdots + w_{k_1} = (\widetilde{w}_m)_1 \\ \vdots \\ v_{k_{l-1}+1} + \cdots + v_{k_l} = w_{k_{l-1}+1} + \cdots + w_{k_l} = (\widetilde{w}_m)_l \end{cases} \Leftrightarrow U_m' v = \widetilde{w}_m$$

The codimension of the affine space: $\{z \in \mathbb{R}^p : U_m' z = \widetilde{w}_m\}$ is $l = \dim(\partial_{\|.\|_w}(x))$ which concludes the proof when x has non-increasing and non-negative components.

Now, let us prove this proposition for the general case. Let $x \in \mathbb{R}^p$ and $\pi \in \mathcal{S}_p$ such that $|x_{\pi(1)}| \geq \cdots \geq |x_{\pi(p)}|$. Let ϕ be the following orthogonal transformation:

$$\forall z \in \mathbb{R}^p : \phi(z) = (\mathrm{sign}(x_{\pi(1)})z_{\pi(1)}, \ldots, \mathrm{sign}(x_{\pi(p)})z_{\pi(p)}) = (\mathrm{sign}(m_{\pi(1)})z_{\pi(1)}, \ldots, \mathrm{sign}(m_{\pi(p)})z_{\pi(p)}).$$

Note that, by construction, $|x|_\downarrow = \phi(x)$ and clearly $|m|_\downarrow = \phi(m)$. Using facts on the OWL norm and on the dual OWL norm (described above), one may deduce the following equivalences:

$$\begin{aligned} v \in \partial_{\|.\|_w}(x) &\Leftrightarrow \|v\|_w^* \leq 1 \text{ and } v'x = \|x\|_w \\ &\Leftrightarrow \|\phi(v)\|_w^* \leq 1 \text{ and } \phi(v)'\phi(x) = \|\phi(x)\|_w \\ &\Leftrightarrow \phi(v) \in \partial_{\|.\|_w}(|x|_\downarrow). \end{aligned}$$

Thus, according to the beginning of the proof, $\phi(v)$ lies onto the affine space $\{z \in \mathbb{R}^p : U'_{|m|_\downarrow} z = \widetilde{w}_m\}$. Clearly, because $m = \mathrm{p}_{\mathrm{owl}}(x)$ and by definition of π we have $|m_{\pi(1)}| \geq \cdots \geq |m_{\pi(p)}|$. Let $1 \leq k_1 < \cdots < k_l \leq p$ be a subdivision of $\{1, \ldots, p\}$ where k_l is the number of non-null components of m and such that

$$\underbrace{|m_{\pi(1)}| = \cdots = |m_{\pi(k_1)}|}_{=l} > \underbrace{|m_{\pi(k_1+1)}| = \cdots = |m_{\pi(k_2)}|}_{=l-1} > \cdots > \underbrace{|m_{\pi(k_{l-1}+1)}| = \cdots = |m_{\pi(k_l)}|}_{=1} > 0.$$

Consequently, we have the following equivalences:

$$\begin{aligned} & \begin{cases} (\phi(v))_1 + \cdots + (\phi(v))_{k_1} = w_1 + \cdots + w_{k_1} = (\widetilde{w}_m)_1 \\ \vdots \\ (\phi(v))_{k_{l-1}+1} + \cdots + (\phi(v))_{k_l} = w_{k_{l-1}+1} + \cdots + w_{k_l} = (\widetilde{w}_m)_l \end{cases} \\ \Leftrightarrow & \begin{cases} \mathrm{sign}(m_{\pi(1)})v_{\pi(1)} + \cdots + \mathrm{sign}(m_{\pi(k_1)})v_{\pi(k_1)} = (\widetilde{w}_m)_1 \\ \vdots \\ \mathrm{sign}(m_{\pi(k_{l-1}+1)})v_{\pi(k_{l-1}+1)} + \cdots + \mathrm{sign}(m_{\pi(k_l)})v_{\pi(k_l)} = (\widetilde{w}_m)_l \end{cases} \\ \Leftrightarrow & \begin{cases} \sum_{i:|m_i|=l} \mathrm{sign}(m_i)v_i = (\widetilde{w}_m)_1 \\ \vdots \\ \sum_{i:|m_i|=1} \mathrm{sign}(m_i)v_i = (\widetilde{w}_m)_l \end{cases} \\ \Leftrightarrow & \quad U'_m v = \widetilde{w}_m \end{aligned}$$

Finally, $\{z \in \mathbb{R}^p : U'_m z = \widetilde{w}_m\}$ is the smallest affine space containing $\partial_{\|.\|_w}(x)$ since the codimension of $\{z \in \mathbb{R}^p : U'_m z = \widetilde{w}_m\}$ is equal to $l = \mathrm{codim}(\partial_{\|.\|_w}(x)) = \mathrm{codim}(\partial_{\|.\|_w}(|x|_\downarrow))$. □

7.2 Proof of Theorem 1

Proof. First, according to Proposition 2, one may deduce that

$$\begin{aligned} & \zeta := X'(\widetilde{X}'_m)^+ \widetilde{w}_m + \frac{1}{\gamma} X'(I - \widetilde{P}_m)y \in \partial_{\|.\|_w}(m) \\ \Leftrightarrow \quad & \|\zeta\|_w^* \leq 1 \text{ and } \underbrace{U'_m X'(\widetilde{X}'_m)^+ \widetilde{w}_m}_{=\widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m} + \frac{1}{\gamma} \underbrace{U_m X'(I - \widetilde{P}_m)y}_{=0} = \widetilde{w}_m \\ \Leftrightarrow \quad & \|\zeta\|_w^* \leq 1 \text{ and } \widetilde{X}'_m(\widetilde{X}'_m)^+ \widetilde{w}_m = \widetilde{w}_m. \end{aligned}$$

Necessity. Let us assume that there exists $\widehat{\beta} \in S_{X,\gamma\|.\|_w}(y)$ such that $\mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = m$. Consequently, $\widehat{\beta} = U_m z$ for some $z \in \mathbb{R}^{k+}$ and thus $X\widehat{\beta}$ lies in $\mathrm{cone}(\widetilde{X}_m)$. Since $\widehat{\beta}$ is an OWL estimator and

according to Proposition 1 we have $\frac{1}{\gamma}X'(y-X\widehat{\beta}) \in \partial_{\|.\|_w}(\widehat{\beta}) = \partial_{\|.\|_w}(m)$. We want to deduce $X\widehat{\beta}$ from this.
By Proposition 2, multiplying by U_m', we get $\frac{1}{\gamma}\widetilde{X}_m'(y-X\widehat{\beta}) = \widetilde{w}_m$. We multiply this equality by $(\widetilde{X}_m')^+$ and use the fact that $\widetilde{P}_m = (\widetilde{X}_m')^+\widetilde{X}_m'$ is the projector onto $\mathrm{col}(\widetilde{X}_m)$. We have $X\widehat{\beta} \in \mathrm{cone}(\widetilde{X}_m)$ so that $\widetilde{P}_m X\widehat{\beta} = X\widehat{\beta}$. We get

$$(\widetilde{X}_m')^+\widetilde{X}_m' X\widehat{\beta} = \widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m \Rightarrow X\widehat{\beta} = \widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m$$

and the cone condition is proven.
Now, replacing the term $X\widehat{\beta}$ by $\widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m$ in $\frac{1}{\gamma}X'(y-X\widehat{\beta})$ gives the sub-differential condition:

$$\partial_{\|.\|_w}(m) \ni \frac{1}{\gamma}X'(y-X\widehat{\beta}) = \frac{1}{\gamma}X'(y-(\widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m)) = X'(\widetilde{X}_m')^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P}_m)y.$$

Sufficiency. Let us assume that the sub-differential condition and the cone condition occur. Then, by the cone condition, one may pick $z \in \mathbb{R}^{k+}$ for which

$$\widetilde{X}_m z = \widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m.$$

Moreover, by definition of U_m, $\mathrm{p_{owl}}(U_m z) = m$. Let us prove that $U_m z \in S_{X,\gamma\|.\|_W}(y)$. Clearly, the following equalities occur

$$\begin{aligned}\frac{1}{\gamma}X'(y-XU_m z) &= \frac{1}{\gamma}X'(y-(\widetilde{P}_m y - \gamma(\widetilde{X}_m')^+\widetilde{w}_m)) \\ &= X'(\widetilde{X}_m')^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P}_m)y \in \partial_{\|.\|_w}(m) = \partial_{\|.\|_w}(U_m z)\end{aligned}$$

where, at the end, we applied the sub-differential condition. Thus $U_m z \in S_{X,\gamma\|.\|_w}(y)$. □

7.3 Proof of Theorem 2

Proof. Consider the sub-differential condition from Theorem 1 when $y = X\beta$. As $\beta = U_m z$ for some $z \in \mathbb{R}^{k+}$, we have $X\beta = \widetilde{X}_m z \in \mathrm{col}(\widetilde{X}_m)$ so that $(I-\widetilde{P})X\beta = 0$. Thus, in the noiseless case, the sub-differential condition is equivalent to $X'(\widetilde{X}_m')^+\widetilde{w}_m \in \partial_{\|.\|_w}(m)$ or equivalently $\|X'(\widetilde{X}_m')^+\widetilde{w}_m\|_w^* \leq 1$ and $\widetilde{X}_m'(\widetilde{X}_m')^+\widetilde{w}_m = \widetilde{w}_m$ and the first statement follows from Theorem 1.

For the second statement, by Theorem 1, it remains to show that for $y = X\beta$ the cone condition $\widetilde{P}_m X\beta - \gamma(\widetilde{X}_m')^+\widetilde{w}_m \in \mathrm{cone}(\widetilde{X}_m)$ occurs for $\gamma > 0$ small enough. We have $\widetilde{P}_m X\beta = X\beta = \widetilde{X}_m z \in \mathrm{cone}(\widetilde{X}_m)$. Note that $(\widetilde{X}_m')^+\widetilde{w}_m \in \mathrm{col}(\widetilde{X}_m)$. Thus

$$z := \underbrace{\widetilde{P}_m X\beta}_{=\widetilde{X}_m z \in \mathrm{cone}(\widetilde{X}_m)} - \gamma \underbrace{(\widetilde{X}_m')^+\widetilde{w}_m}_{\in \mathrm{col}(\widetilde{X}_m)} \in \mathrm{col}(\widetilde{X}_m)$$

Since $\mathrm{cone}(\widetilde{X}_m)$ is open in $\mathrm{col}(\widetilde{X}_m)$, one may deduce that for γ small enough, we have $z \in \mathrm{cone}(\widetilde{X}_m)$. □

7.4 Proof of Theorem 3

Proof. **1** Note that $(I-\widetilde{P}_m)Y = (I-\widetilde{P}_m)\varepsilon$ (since $(I-\widetilde{P}_m)(X\beta) = 0$). Consequently the sub-differential condition event may be rewritten as follows:

$$\left\{X'(\widetilde{X}_m')^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P}_m)Y \in \partial_{\|.\|_w}(m)\right\} = \left\{X'(\widetilde{X}_m')^+\widetilde{w}_m + \frac{1}{\gamma}X'(I-\widetilde{P}_m)\varepsilon \in \partial_{\|.\|_w}(m)\right\}.$$

Note that the probability of pattern recovery by OWL estimator is smaller than the probability of the sub-differential condition event and consequently

$$\mathbb{P}_\varepsilon(\exists \widehat{\beta} \in S_{X,\gamma\|.\|_w}(Y) \text{ such that } \mathrm{p}_{\mathrm{owl}}(\widehat{\beta}) = \mathrm{p}_{\mathrm{owl}}(\beta)) \leq \begin{cases} \mathbb{P}_\varepsilon(\|\zeta_\gamma\|_w^* \leq 1) \text{ if } \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m = \widetilde{w}_m \\ 0 \text{ if } \widetilde{X}'_m(\widetilde{X}'_m)^+\widetilde{w}_m \neq \widetilde{w}_m \end{cases}.$$

2: According to Theorem 1, pattern recovery by OWL estimator is equivalent to have simultaneously the cone condition and the sub-differential condition. The upper bound coincides with the probability of the sub-differential condition. Thus to prove that this upper bound is sharp, it remains to show that the probability of the cone condition tends to 1 when r tends to $+\infty$. Let us assume that ε is a random vector defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let us consider a particular observation $\varepsilon(\omega)$ for some $\omega \in \Omega$ (thus $Y^{(r)}(\omega) = X\beta^{(r)} + \varepsilon(\omega)$). Since $\widetilde{P} = \widetilde{X}_m\widetilde{X}_m^+$ and since $(\widetilde{X}'_m)^+ = \widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^+$, one may rewrite the cone condition as follows:

$$\begin{aligned} \widetilde{P}_m Y^{(r)}(\omega) - \gamma(\widetilde{X}'_m)^+\widetilde{w}_m &= \widetilde{X}_m s^{(r)} + \widetilde{X}_m\widetilde{X}_m^+\varepsilon(\omega) - \gamma\widetilde{X}_m(\widetilde{X}'_m\widetilde{X}_m)^+\widetilde{w}_m \\ &= \widetilde{X}_m(s^{(r)} + \widetilde{X}_m^+\varepsilon(\omega) - \gamma(\widetilde{X}'_m\widetilde{X}_m)^+\widetilde{w}_m). \end{aligned}$$

Note that by assumption on $\Delta^{(r)}$ we have that:

- the vector $s^{(r)}/\Delta^{(r)}$ is (component-wise) larger or equal than$(k, \dots, 1)$;
- $\lim_{r\to+\infty} \widetilde{X}_m^+\varepsilon(\omega)/\Delta^{(r)} = 0$ and $\lim_{r\to+\infty} \gamma(\widetilde{X}'_m\widetilde{X}_m)^+\widetilde{w}_m/\Delta^{(r)} = 0$.

Consequently, for $r \geq r_0(\omega)$ large enough we have

$$\widetilde{P}_m Y(\omega) - \gamma(\widetilde{X}'_m)^+\widetilde{w}_m \in \mathrm{cone}(\widetilde{X}_m).$$

Since this fact is true for all ω, one may deduce that

$$\lim_{r\to+\infty} \mathbb{P}_\varepsilon(\widetilde{P}_m Y - \gamma(\widetilde{X}'_m)^+\widetilde{w}_m \in \mathrm{cone}(\widetilde{X}_m)) = 1.$$

3: In the proof of 2, we see that the probability of the cone condition tends to 1 when $\lim_{r\to+\infty} \gamma_r = +\infty$ and when $\lim_{r\to+\infty} \gamma_r/\Delta^{(r)} = 0$. Thus it remains to prove that the probability of the sub-differential condition tends to 1 when $X'(\widetilde{X}'_m)^+\widetilde{w}_m \in \mathrm{ri}(\partial_{\|.\|_w}(m))$. Let us point out the following asymptotic result:

$$X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma_r}X'(I - \widetilde{P}_m)\varepsilon \xrightarrow[r\to+\infty]{\mathbb{P}} X'(\widetilde{X}'_m)^+\widetilde{w}_m. \tag{9}$$

Note by Proposition 2 that $X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma_r}X'(I - \widetilde{P}_m)\varepsilon \in \mathrm{aff}(\partial_{\|.\|_w}(m))$. When $X'(\widetilde{X}'_m)^+\widetilde{w}_m \in ri(\partial_{\|.\|_w}(m))$, one may deduce by (9) the following limit:

$$\lim_{r\to+\infty} \mathbb{P}_\varepsilon\left(X'(\widetilde{X}'_m)^+\widetilde{w}_m + \frac{1}{\gamma_r}X'(I - \widetilde{P}_m)\varepsilon \in \partial_{\|.\|_w}(m)\right) = 1.$$

□

Aknowledgements

The Burgundy Institute of Mathematics receives support from the EIPHI Graduate School (contract ANR-17-EURE-0002).

References

Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3): 1103, 2015.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Xavier Dupuis and Patrick JC Tardivel. Proximal operator for the sorted 1 norm: Application to testing procedures based on slope. *Journal of Statistical Planning and Inference*, 2022.

Karl Ewald and Ulrike Schneider. On the distribution, model selection properties and uniqueness of the lasso estimator in low and high dimensions. *Electronic Journal of Statistics*, 14(1):944–969, 2020.

Mario Figueiredo and Robert Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Artificial Intelligence and Statistics*, pages 930–938. PMLR, 2016.

J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE transactions on Information theory*, 50(6):1341–1344, 2004.

Thomas Godland and Zakhar Kabluchko. Projections and angle sums of permutohedra. *arXiv preprint arXiv:2009.04186*, 2020.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

Kentaro Minami. Degrees of freedom in submodular regularization: A computational perspective of stein's unbiased risk estimate. *Journal of Multivariate Analysis*, 175:104546, 2020.

Renato Negrinho and Andre Martins. Orbit regularization. *Advances in neural information processing systems*, 27, 2014.

Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *arXiv preprint arXiv:2004.09106*, 2020.

Amir Sepehri and Naftali Harris. The accessible lasso models. *Statistics*, 51(4):711–721, 2017.

Dhruv B Sharma, Howard D Bondell, and Hao Helen Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340, 2013.

Tomasz Skalski, Piotr Graczyk, Bartosz Kołodziejek, and Maciej Wilczyński. Pattern recovery and signal denoising by slope when the design matrix is orthogonal. *arXiv preprint arXiv:2202.08573*, 2022.

Patrick Tardivel, Tomasz Skalski, Piotr Graczyk, and Ulrike Schneider. The geometry of model recovery by penalized and thresholded estimators. 2021.

Patrick JC Tardivel and Malgorzata Bogdan. On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator, and thresholded basis pursuit denoising. *SCANDINAVIAN JOURNAL OF STATISTICS*, 2022.

Patrick JC Tardivel, Rémi Servien, and Didier Concordet. Simple expressions of the lasso and slope estimators in low-dimension. *Statistics*, 54(2):340–352, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

Samuel Vaiter, Mohammad Golbabaee, Jalal Fadili, and Gabriel Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, 2015.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Xiangrong Zeng and Mário AT Figueiredo. Decreasing weighted sorted ell_1 regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Günter M Ziegler. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.