Online Supplementary Material for the Paper "An Alternating Method for Cardinality-Constrained Optimization: A Computational Study for the Best Subset Selection and Sparse Portfolio Problems"

Carina Moreira Costa, Dennis Kreber, Martin Schmidt

1. Omitted Proofs and Remarks

1.1. Proof of Proposition 5.

Proof. If $e^\top \bar{x}_S = 1$ we have proven the result because $\tilde{w}_S = \bar{x}_S$ must hold for $\tilde{w}$ to be an optimal solution. Hence, we only have to consider the case $e^\top \bar{x}_S < 1$ because $e^\top \bar{x}_S > 1$ is infeasible. First, we show that $\tilde{w}_i \geq \bar{x}_i$ holds for all $i \in S$.

Note that if $e^\top \tilde{w} = e^\top \tilde{w}_S = 1$ and $e^\top \bar{x}_S < 1$ holds, then there must be at least one index $i \in S$ with $\tilde{w}_i > \bar{x}_i$. By contradiction, assume that there also exists an index $j \in S$ with $\tilde{w}_j < \bar{x}_j$. Next, we define $\gamma := \min\{\tilde{w}_i - \bar{x}_i, \bar{x}_j - \tilde{w}_j\} > 0$ and consider $\hat{w} \in \mathbb{R}^n$ such that

$$\hat{w}_l := \begin{cases} \tilde{w}_i - \gamma, & \text{if } l = i, \\ \tilde{w}_j + \gamma, & \text{if } l = j, \\ \tilde{w}_l, & \text{otherwise}, \end{cases} \tag{1}$$

holds. Clearly, $e^\top \hat{w} = 1$ and $\operatorname{supp}(\hat{w}) = S$ hold. Thus, $\hat{w}$ is feasible for (7). Furthermore, it holds that $\hat{w}_i \geq \bar{x}_i$ and $\hat{w}_j \leq \bar{x}_j$. From this and (1), it follows

$$\begin{aligned} \|\hat{w} - \bar{x}\|_1 = \sum_{l=1}^n |\hat{w}_l - \bar{x}_l| &= \sum_{\substack{l=1 \\ l \notin \{i,j\}}}^n |\hat{w}_l - \bar{x}_l| + |\hat{w}_i - \bar{x}_i| + |\hat{w}_j - \bar{x}_j| \\ &= \sum_{\substack{l=1 \\ l \notin \{i,j\}}}^n |\hat{w}_l - \bar{x}_l| + \hat{w}_i - \bar{x}_i + \bar{x}_j - \hat{w}_j \\ &= \sum_{\substack{l=1 \\ l \notin \{i,j\}}}^n |\tilde{w}_l - \bar{x}_l| + \tilde{w}_i - \gamma - \bar{x}_i + \bar{x}_j - \tilde{w}_j - \gamma \\ &= \|\tilde{w} - \bar{x}\|_1 - 2\gamma < \|\tilde{w} - \bar{x}\|_1, \end{aligned}$$

which is a contradiction to $\tilde{w}$ being an optimal solution of (7). Thus, $\tilde{w}_i \geq \bar{x}_i$ holds for all $i \in S$.

Consider now $w^* \in \mathbb{R}^n$ with

$$w_l^* := \begin{cases} \bar{x}_l/(e^\top \bar{x}_S), & \text{if } l \in S, \\ 0, & \text{otherwise}. \end{cases}$$

2010 *Mathematics Subject Classification.* 90Cxx, 90C11, 90C90, 90-08.

Key words and phrases. Cardinality Constraints, Alternating Direction Methods, Penalty Methods, Best Subset Selection, Portfolio Optimization.

Clearly, w^* is feasible for (7) and $w_i^* \geq \bar{x}_i$ holds for all $i \in S$. We are now ready to show that w^* is also an optimal solution of (7):

$$\begin{aligned}\|\tilde{w} - \bar{x}\|_1 &= \sum_{l \notin S} |\bar{x}_l| + \sum_{l \in S} \tilde{w}_l - \bar{x}_l = \sum_{l \notin S} |\bar{x}_l| + 1 - e^\top \bar{x}_s \\ &= \sum_{l \notin S} |\bar{x}_l| + \sum_{l \in S} w_l^* - \bar{x}_l = \|w^* - \bar{x}\|_1.\end{aligned}$$ □

1.2. **Proof of Proposition 6.**

Proof. Let $S^* = \{i_1, \dots, i_k\}$ and $w^* \in \mathbb{R}^n$ with

$$w_l^* := \begin{cases} \bar{x}_l/(e^\top \bar{x}_{S^*}), & \text{if } l \in S^*, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, assume that there is an optimal solution $\tilde{w}$ of (7) with $\tilde{S} = \operatorname{supp}(\tilde{w})$. Due to Proposition 5, we can assume without loss of generality that $\tilde{w}$ has the form

$$\tilde{w}_l := \begin{cases} \bar{x}_l/(e^\top \bar{x}_{\tilde{S}}), & \text{if } l \in \tilde{S}, \\ 0, & \text{otherwise,} \end{cases}$$

and, hence, that $\tilde{w} \geq \bar{x}$ holds. Then, it follows that

$$\begin{aligned}\|\tilde{w} - \bar{x}\|_1 &= \sum_{l \notin \tilde{S}} \bar{x}_l + \sum_{l \in \tilde{S}} \tilde{w}_l - \bar{x}_l = \sum_{l \notin \tilde{S}} \bar{x}_l + 1 - \sum_{l \in \tilde{S}} \bar{x}_l \geq \sum_{l \notin S^*} \bar{x}_l + 1 - \sum_{l \in S^*} \bar{x}_l \\ &= \sum_{l \notin S^*} \bar{x}_l + \sum_{l \in S^*} w_l^* - \bar{x}_l = \|w^* - \bar{x}\|_1.\end{aligned}$$

Thus, w^* is an optimal solution of (7). □

1.3. **Proof of Proposition 7.**

Proof. We need to prove two properties: feasibility and optimality of $\delta^{t,l+1}$. To prove feasibility, we need to check whether $\delta^{t,l+1} \in \bar{V}$. This is obvious since there are at most k non-zero entries in $\delta^{t,l+1}$ and all the others $p-k$ entries are zero. Now, we prove that $\delta^{t,l+1}$ is optimal. Suppose that δ in Problem (11) is not restricted to the set $\bar{V}$, i.e., the problem is an unconstrained problem. Then, the minimum objective value that one can obtain is zero and this is reached by setting $\delta^{t,l+1} = \gamma^{t,l+1}$. However, if δ is restricted to the set $\bar{V}$ and $\gamma^{t,l+1}$ is not the zero vector, at most k entries of δ are allowed to be non-zero and then one needs to choose the k out of p that minimize the sum of the absolute values of the 1 norm. Clearly, the best choice are the k largest entries of $\gamma^{t,l+1}$, because then the k corresponding terms in the sum are zero and the other $p-k$ are the smallest entries of $\gamma^{t,l+1}$. This concludes the proof. □

1.4. **Remark on The Robustness of Our Heuristic.** The PADM can be understood as cutting off the smallest coefficients in each iteration to construct a sparse solution. This idea is similar to the trimmed Lasso [1]. Trimmed Lasso is a generalization of Lasso. For the latter, the coefficients are penalized by an ℓ_1 norm term, while for the trimmed Lasso, the regularization is given by

$$T_k(\beta) = \min_{\delta : |\operatorname{supp}(\delta)| \leq k} \|\beta - \delta\|_1,$$

i.e., the trimmed Lasso is the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \|X\beta - y\|_2^2 + \mu T_k(\beta). \tag{2}$$

Interestingly, the two components of the trimmed Lasso, i.e., the least-squares part and the regularization, are the two subproblems that we identified to be the directions for the PADM. In other words, the result of an inner loop of the PADM can be considered a heuristic solution to the trimmed Lasso. The connection of the PADM theory and the theory behind the trimmed Lasso reveals an interesting relation. In [1] it is shown that for some sufficiently large μ, Problem (2) is equivalent to the best subset selection problem. This insight can as well be derived from Theorem 4. Moreover, the optimization problem (2) can be reformulated as a robust optimization problem. The authors in [1] prove that problem

$$\min_{\beta \in \mathbb{R}^p} \min_{\substack{I \subseteq [p] \\ |I| = p-k}} \max_{\Delta \in \mathcal{L}_I^\mu} \|(X + \Delta)\beta - y\|_2^2 \tag{3}$$

with

$$\mathcal{L}_I^\mu := \{\Delta \in \mathbb{R}^{n \times p} : \|\Delta_i\|_2 \le \mu \text{ for all } i \in [p] \text{ for } \Delta_i = 0 \text{ for all } i \notin I\},$$

is equivalent to (2). Hence, the inner loop solution of the PADM can be considered a solution to (3)—i.e., a robustification of the usual least-squares problem. Therefore, each outer iteration of the PADM increases the robustification severity. Thus, even though we are only guaranteed to obtain a partial minimum, the effects of the underlying robustification can lead to results which are still very good from a statistical point of view.

References

[1] D. Bertsimas, M. S. Copenhaver, and R. Mazumder. *The Trimmed Lasso: Sparsity and Robustness.* 2017. arXiv: 1708.04527.

(C. Moreira Costa, D. Kreber, M. Schmidt) Trier University, Department of Mathematics, Universitätsring 15, 54296 Trier, Germany

Email address: costa@uni-trier.de

Email address: dnis.kk@gmail.com

Email address: martin.schmidt@uni-trier.de