Open Journal of Mathematical Optimization

Gabriel Ponte, Marcia Fampa, Jon Lee & Luze Xu

Good and Fast Row-Sparse ah-Symmetric Reflexive Generalized Inverses

Volume 6 (2025), article no. 12 (24 pages)

https://doi.org/10.5802/ojmo.48

Article submitted on November 19, 2024, revised on July 2, 2025,
accepted on November 20, 2025.

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

Good and Fast Row-Sparse ah-Symmetric Reflexive Generalized Inverses

Gabriel Ponte
University of Michigan, Ann Arbor, USA
gabponte@umich.edu

Marcia Fampa
Federal University of Rio de Janeiro, Brazil
fampa@cos.ufrj.br

Jon Lee
University of Michigan, Ann Arbor, USA
jonxlee@umich.edu

Luze Xu
University of California, Davis, USA
lzxu@ucdavis.edu

Abstract

We present several algorithms aimed at constructing sparse and structured sparse (row-sparse) generalized inverses, with application to the efficient computation of least-squares solutions, for inconsistent systems of linear equations, in the setting of multiple right-hand sides and a rank-deficient constraint matrix. Leveraging our earlier formulations to minimize the 1- and 2,1-norms of generalized inverses that satisfy important properties of the Moore–Penrose pseudoinverse, we develop efficient and scalable ADMM algorithms to address these norm-minimization problems and to limit the number of nonzero rows in the solution. We establish a 2,1-norm approximation result for a local-search procedure that was originally designed for 1-norm minimization, and we compare the ADMM algorithms with the local-search procedure and with general-purpose optimization solvers.

Digital Object Identifier 10.5802/ojmo.48

Keywords Moore–Penrose properties, generalized inverse, sparse optimization, norm minimization, least squares, local search, convex optimization, ADMM.

Acknowledgments G. Ponte was supported in part by CNPq GM-GD scholarship 161501/2022-2. M. Fampa was supported in part by CNPq grant 307167/2022-4. J. Lee was supported in part by AFOSR grant FA9550-22-1-0172.

1 Introduction

The M-P (Moore–Penrose) pseudoinverse is an important object in numerical linear algebra. It has many uses; in particular, and relevant to our study, it can be used to compute least-squares solutions in the context of linear statistical models. Given a real singular-value decomposition $A = U\Sigma V^\intercal$ of A, the M-P pseudoinverse of A is $A^\dagger := V\Sigma^\dagger U^\intercal$, where the diagonal matrix $\Sigma^\dagger$ has the shape of the transpose of the diagonal matrix Σ, and is calculated from Σ by taking reciprocals of the nonzero elements of Σ, and is otherwise 0 (see [12, Section 2.5.3], for example). The following well-known theorem characterizes the M-P pseudoinverse.

▶ **Theorem 1** (see [19, Theorem 1]). *For $A \in \mathbb{R}^{m\times n}$, the following equations have a unique solution $H \in \mathbb{R}^{n\times m}$.*

$$\text{(generalized inverse)} \qquad AHA = A \tag{P1}$$
$$\text{(reflexive)} \qquad HAH = H \tag{P2}$$
$$\text{(ah-symmetric)} \qquad (AH)^\intercal = AH \tag{P3}$$
$$\text{(ha-symmetric)} \qquad (HA)^\intercal = HA \tag{P4}$$

It is easy to check that the M-P pseudoinverse of A, that is $A^\dagger := V\Sigma^\dagger U^\intercal$, satisfies these equations, and so it is the unique solution.

As is common (see [1, Introduction, Section 2]), we say that a *generalized inverse* of A is any H satisfying (P1), and a generalized inverse of A is *reflexive* if it additionally satisfies (P2). Following [24, p. 1723], we say that a generalized inverse H of A is *ah-symmetric* (resp. *ha-symmetric*) if AH (resp. HA) is symmetric. It is clear that under the mapping $(A, H) \to (A^\intercal, H^\intercal)$, properties (P1) and (P2) are invariant, while properties (P3) and (P4) are exchanged; so, although we will concentrate on ah-symmetric reflexive generalized inverses, much of what we work out directly transfers to the case of ha-symmetric reflexive generalized inverses.

We are interested in least-squares problems $\min\{\|A\theta - b\|_2 : \theta \in \mathbb{R}^n\}$, related to fitting a model of $\beta = \alpha^\intercal\theta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma)$. We regard α as a vector of n independent variables, and β as the (random) response. We view the rows of $[A, b]$ as m independent realizations of $[\alpha^\intercal, \beta]$. We are particularly interested in the situation where $m > n \gg \mathrm{rank}(A)$. A very well-known solution of the least-squares problem is $\widehat{\theta} := A^\dagger b$, which fits the model $\widehat{\beta} = \alpha^\intercal\widehat{\theta}$. This solution has a wonderful property — it quickly gives us the solution for *every* response vector b, and moreover, it is linear in b. This all begs the question, are there other matrices H (besides $A^\dagger$) that also have this attractive property? It is well known that the answer is yes: if H is any ah-symmetric generalized inverse, then $\widehat{\theta} := Hb$ solves $\min\{\|A\theta - b\|_2 : \theta \in \mathbb{R}^n\}$ (see [3, Theorem 6.2.4] or [11, Corollary 4.4]). Because a given matrix A does not typically possess a unique ah-symmetric generalized inverse, there is room to optimize various criteria.

One important criterion for H is *sparsity*. Even if a given matrix is sparse, its M-P pseudoinverse can be completely dense, often leading to a high computational burden in its applications involving many response vectors b, especially when we are dealing with a large matrix A. A sparse H leads to efficiency in calculating $\widehat{\theta} := Hb$. We can seek such a matrix H with a minimal number of nonzero elements, by applying the standard minimization of its (vector) 1-norm $\|H\|_1 := \sum_{ij} |H_{ij}|$, as a surrogate for the nonconvex "0-norm". Besides aiming to induce sparsity, minimization of $\|H\|_1$, or even a different norm, has the important effect of keeping the elements of H under control, which reduces numerical errors in the calculation of $\widehat{\theta} := Hb$.

We are further interested in *structured sparsity* for H; specifically, such an H with a *limited number of nonzero rows*. In evaluating $\widehat{\theta} := Hb$, potentially for many b, it is easy to do this very efficiently when H has a limited number of nonzero rows, taking advantage of sparsity without having to use elaborate sparse data structures. Specifically, if $\widehat{H} := H[S, \cdot\,]$ contains all of the nonzeros of H, then $\widehat{\theta}_{\overline{S}} = 0$, and $\widehat{\theta}_S = \widehat{H}b$. Moreover, for such an H, the nonzero rows align with a corresponding limited number of columns (i.e., features) of the data matrix A, and so $\widehat{\theta}$ is sparse, and the linear model $\widehat{\beta} = \alpha^\intercal\widehat{\theta}$ is more *explainable*.

A specific example where we want to evaluate Hb for many right-hand sides b is in an ADMM approach to the least-absolute-deviations problem $\min\{\|A\theta - b\|_1 : \theta \in \mathbb{R}^n\}$, which aims at a robust fit, under the assumption that $m > n \gg \mathrm{rank}(A)$; see [2, Section 6.1] (where they assume that A has full column rank; but that assumption is not needed even in their approach, if they simply replace $(A^\intercal A)^{-1}A^\intercal$ with $A^\dagger$).

There are different ways to work with the criterion of wanting H to have a limited number of rows, i.e. "row-sparsity". For example, (i) we can try to induce it by minimizing an appropriate norm, in this case the 2,1-norm $\|H\|_{2,1} := \sum_i \|H_{i\cdot}\|_2$, or (ii) we can impose this structure and then further seek to (locally) minimize some other criterion, e.g., the (vector) 1-norm or the 2,1-norm. In both approaches, we gain an important additional benefit (mentioned above); by keeping some norm of H under control, we can expect to gain some numerical stability in the calculation of Hb.

With the same level of sparsity, structured sparsity should always be preferred to unstructured sparsity. Additionally, algorithms aimed at inducing structured (resp., unstructured) sparsity may or may not actually achieve a high level of structured (resp., unstructured) sparsity. Finally, different algorithms with different (but related) goals may of course have significantly different running times. Therefore, as we will do in Section 5, we compare the behavior of different algorithms having different (but related) goals, on all relevant measures.

Another important criterion is the *rank* of H. While the sparsity or row-sparsity of H can be viewed as a kind of simplicity for H, such a viewpoint is basis dependent. If we make an invertible linear transformation of the columns of A, that is $A \to AT$, for an invertible and possibly dense $T_{n\times n}$, then with an ah-symmetric generalized inverse H of A, we have $\widehat{\theta} = T^{-1}Hb$. But this matrix $T^{-1}H$ loses the sparsity of H. So we are also interested in a criterion for simplicity of H that is not basis dependent. The natural such criterion is the *rank* of H, which has been employed in many contexts, in tandem with sparsity criteria (such as sparse PCA (see [26]); sparse Gaussian mixture models (see [15], and the references therein); low-rank/sparsity matrix decomposition (see [4]); low-rank graphical models (see [17])), to capture a different kind of simplicity/explainability than is captured by sparsity.

In many contexts, rank minimization is naturally induced by minimizing the nuclear norm (i.e., the 1-norm of the vector of singular values). But in our context, we have an easier path. It is well known that: (i) if H is a generalized inverse of A, then $\mathrm{rank}(H) \geq \mathrm{rank}(A)$, and (ii) a generalized inverse H of A is reflexive if and only if $\text{rank}(H) = \text{rank}(A)$ (see [21, Theorem 3.14]). So the minimum possible rank for a generalized inverse of A is $\text{rank}(A)$, and we can achieve minimization of $\text{rank}(H)$ by simply enforcing (P2), that is, that H be a reflexive generalized inverse (e.g., $A^\dagger$ is a reflexive generalized inverse, so it has minimum rank). Now, it appears that (P2) is *nonlinear* in H (i.e., it is a generic system of quadratic equations) typically defining a nonconvex region, making enforcing it (in the context of *exact global optimization*) very difficult. But, it is easy to check and well known that (P2) becomes linear under properties (P1) and (P3). Specifically, (P1) and (P3) imply that $AH = AA^\dagger$, and so (P2) becomes $HAA^\dagger = H$.

In summary, our overarching goal is to efficiently calculate a sparse or row-sparse reflexive ah-symmetric generalized inverse H, with a controlled value for some norm, in the setting of $m > n \gg \text{rank}(A)$. We develop several methods for achieving this goal. We note that because we have two objectives (sparsity and low norm), there is naturally a trade-off to look at.

Notation. $\|\cdot\|_p$ denotes vector p-norm ($p \geq 1$), $\|\cdot\|_0$ denotes vector "0-norm". For matrices, $\text{tr}(\cdot)$ denotes trace, $\det(\cdot)$ denotes determinant, $\text{rank}(\cdot)$ denotes rank, $\|\cdot\|_F$ denotes the Frobenius norm, $\mathcal{R}(\cdot)$ denotes the column space. $A[S,T]$ is the submatrix of A having row indices S and column indices T, with $S = \cdot$ (resp. $T = \cdot$) indicating all rows (resp. columns); further, $A_{i\cdot} := A[\{i\},\cdot]$, $A_{\cdot j} := A[\cdot,\{j\}]$, and $A_{ij} := A[\{i\},\{j\}]$. For $A \in \mathbb{R}^{m\times n}$, $\|A\|_{p,q} := \|(\|A_{1\cdot}\|_p, \ldots, \|A_{m\cdot}\|_p)\|_q$ ($p, q \geq 0$). $\mathbf{e}_i$ denotes the i-th standard unit vector. I denotes an identity matrix (sometimes with a subscript indicating its order). For $A \in \mathbb{R}^{m\times n}$, for the *compact singular-value decomposition* (compact SVD), we write $A = U\Sigma V^\top$, where $r := \text{rank}(A)$, $U_{m\times r}$ and $V_{n\times r}$ are real orthogonal matrices, $\Sigma_{r\times r}$ is a diagonal matrix, with positive diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, known as the singular values of A. For the *(full) singular-value decomposition* (SVD), we instead have $U_{m\times m}$, $V_{n\times n}$, with nonnegative diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}}$. We denote by $\text{argmin}\{\cdot\}$ *any* solution of the associated minimization problem.

Literature. For basic information on generalized inverses, we refer to [21, 1, 3]. [7] (and then [5, 6]) introduced the idea of seeking to induce sparsity in H by linear optimization; specifically, they minimize the (vector) 1-norm over the set of left inverses (i.e., $HA = I_n$) or right inverses (i.e., $AH = I_m$). [5] also considered the idea of inducing row-sparsity by minimizing the 2,1-norm over the same sets of matrices. Additionally, [5] presents ADMM algorithms for these minimization problems, using specific projections for the cases where the matrices are full row- or column-rank. It is important to realize that the left (resp., right) inverse case applies only to full column (resp., row) rank matrices A. In particular, every left inverse H is a reflexive ha-symmetric generalized inverse, and so for such an H, $\widehat{\theta} := Hb$ solves the least-squares problem (for all b) *only* when $H = A^\dagger$. In fact, we are motivated by the case in which A is neither full row rank nor full column rank, and so the left/right inverse approach does not apply at all in our setting.

In [11], we proposed the idea of finding various types of sparse generalized inverses by imposing combinations of the M-P properties in the context of straight-forward linear-optimization formulations, minimizing the (vector) 1-norm of H. But we did not propose any practical idea for controlling the rank of solutions. Moreover, the approaches proposed did not scale well.

In [8], we introduced the idea of imposing a block structure on H, which implicitly enforces (P1) and (P2), and then carrying out a combinatorial local search, giving an approximation guarantee on (vector) 1-norm minimization for reflexive generalized inverses. That work does not directly related to our present setting because it ignores property (P3). Additionally, at the time, the empirical quality (1-norm) of the solutions for that method on large instances was unknown, because we could not solve the large linear-optimization formulations that would give us lower bounds.

In [9], we proposed different algorithmic approaches that start from a (vector) 1-norm minimizing ah-symmetric generalized inverse, and gradually decrease its rank, by iteratively imposing the reflexive property. The algorithms iterate until the ah-symmetric generalized inverse has the least possible rank producing intermediate solutions during the iterations, trading off low 1-norm against low rank. The best approach investigated was a cutting-plane method that solves linear-optimization problems at each iteration, although it is capable of constructing interesting intermediate solutions (trading off rank and 1-norm), it does not scale well, as the solution of many dense linear-optimization problems is required. Only square matrices with size 50 and rank 25 were considered in the numerical experiments.

In [24], we significantly extended the work in [8], in particular to approximate (vector) 1-norm minimizing ah-symmetric reflexive generalized inverses, using a column-block construction, which implicitly enforces (P1), (P2), and (P3), and maintains structured row-sparsity during a local search. Comparisons between solutions of the local-search procedure and the minimal 1-norm of ah-symmetric generalized inverses were presented for matrices of size up to 120×60 and rank 30. It was also mentioned that, within a time limit of two hours, the solutions of the linear-optimization problems to minimize the 1-norm of generalized and ah-symmetric generalized inverses were obtained for only one out of five given matrices of size 200×100 and rank 50. The results indicated that, in practice, the solutions of the linear-optimization problems were not easy for large matrices. [23] is a companion work that in particular analyzes the motivating cases for which $\operatorname{rank}(A) \in \{1, 2\}$; while the case of rank one is trivial, the complexity of the case of rank two reveals that there is no simple solution in general.

In [10], we carried out a detailed computational study of local-search procedures based on the results of [8] and [24]. An experimental analysis of the procedures was performed, but comparisons between the local search solutions and the minimum 1-norm of generalized and ah-symmetric generalized inverses were only presented for square matrices of size up to 100 and rank 50, again because of the intractability of the large dense linear-optimization problems.

In [20], in addition to considering 1-norm minimization with linear optimization to induce (unstructured) sparsity of generalized inverses, which was first advanced by [11], we also considered 2,1-norm minimization with second-order-cone optimization to induce row-sparsity. Furthermore, [20] showed how to make the ideas of linear and second-order-cone optimization much more efficient/scalable. A key idea exploited in [20] is as follows. Let $A \in \mathbb{R}^{m\times n}$ with rank r and $A =: U\Sigma V^\intercal$ be the (full) SVD of A, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices ($U^\intercal U = I_m, V^\intercal V = I_n$), and $\Sigma \in \mathbb{R}^{m\times n}$ with

$$\Sigma =: \begin{bmatrix} \underset{r\times r}{D} & \underset{r\times(n-r)}{0} \\ \underset{(m-r)\times r}{0} & \underset{(m-r)\times(n-r)}{0} \end{bmatrix},$$

with D being a diagonal matrix with rank r. Let $H \in \mathbb{R}^{n\times m}$ and $\Gamma := V^\intercal H U$, where

$$\Gamma =: \begin{bmatrix} \underset{r\times r}{X} & \underset{r\times(m-r)}{Y} \\ \underset{(n-r)\times r}{Z} & \underset{(n-r)\times(m-r)}{W} \end{bmatrix},$$

then $H = I_n H I_m = (VV^\intercal)H(UU^\intercal) = V\Gamma U^\intercal$.

▶ **Lemma 2** (see [1], p. 208, Ex. 14 (no proof); see [20, Lemmas 2-5] (with proof))**.**

- (P1) *is equivalent to* $X = D^{-1}$.
- *If* (P1) *is satisfied, then* (P2) *is equivalent to* $ZDY = W$.
- *If* (P1) *is satisfied, then* (P3) *is equivalent to* $Y = 0$.
- *If* (P1) *is satisfied, then* (P4) *is equivalent to* $Z = 0$.

So, for the purpose of enforcing any subset of the M-P properties on a generalized inverse H, except the case of (P2) alone, we simply set appropriate blocks of Γ to zeros.

Consider the following natural convex-optimization problems, aimed at inducing sparsity and row-sparsity for ah-symmetric reflexive generalized inverses, respectively.

$$\min_{H\in\mathbb{R}^{n\times m}} \{\|H\|_1 \ : \ \text{(P1)}, \text{(P2)}, \text{(P3)}\} \tag{P_{123}^1}$$

and

$$\min_{H\in\mathbb{R}^{n\times m}} \{\|H\|_{2,1} \ : \ \text{(P1)}\}\,. \tag{$P_1^{2,1}$}$$

It is important to note that due to [20, Corollary 9], optimal solutions of ($P_1^{2,1}$) automatically satisfy (P2) and (P3) (i.e., they are reflexive and ah-symmetric), which we exploit in establishing Theorem 3 below.

Let

$$V := \begin{bmatrix} \underset{n\times r}{V_1} & \underset{n\times(n-r)}{V_2} \end{bmatrix},\ U := \begin{bmatrix} \underset{m\times r}{U_1} & \underset{m\times(m-r)}{U_2} \end{bmatrix}.$$

We remark that U_1 and V_2 are full column rank matrices and $U_1^\intercal U_1 = I$ and $V_2^\intercal V_2 = I$, then $U_1^\dagger = (U_1^\intercal U_1)^{-1}U_1^\intercal = U_1^\intercal$ and $V_2^\dagger = (V_2^\intercal V_2)^{-1}V_2^\intercal = V_2^\intercal$.

From Lemma 2, we can deduce the following result.

► **Theorem 3** ([20, Section 5]). *(P_{123}^1) and $(P_1^{2,1})$, respectively, can be efficiently reformulated as*

$$z(\mathcal{P}_{123}^1) := \min_{Z\in\mathbb{R}^{(n-r)\times r}} \left\|V_1D^{-1}U_1^\top + V_2ZU_1^\top\right\|_1 \,, \tag{$\mathcal{P}_{123}^1$}$$

$$z(\mathcal{P}_1^{2,1}) := \min_{Z\in\mathbb{R}^{(n-r)\times r}} \left\|V_1D^{-1} + V_2Z\right\|_{2,1} \,. \tag{$\mathcal{P}_1^{2,1}$}$$

The more compact formulations presented in Theorem 3 were exploited in the numerical experiments performed in [20] to increase the size of matrices for which the solutions of (P_{123}^1) and $(P_1^{2,1})$ were known. Within a time limit of 5 hours, the solutions of $(\mathcal{P}_{123}^1)$ and $(\mathcal{P}_1^{2,1})$ were obtained for instances with (m,n,r) equal to $(280,140,60)$ and $(2000,1000,500)$, respectively. These results allowed us to finally evaluate the empirical quality of the proposed local-search solutions on larger instances.

Concerning the contribution of our present work, we take a further step in efficiently obtaining solutions to (P_{123}^1) and $(P_1^{2,1})$. We leverage Theorem 3 once again and develop very efficient/scalable ADMM (Alternating Direction Method of Multipliers) algorithms to solve them. Furthermore, we develop two ADMM algorithms aiming at imposing structured sparsity by limiting the number of nonzero rows for an ah-symmetric reflexive generalized inverse. For the first, we do *not* consider the minimization of a norm of the nonzero submatrix of H, leading to a fast ADMM algorithm that can take advantage of the efficient solutions presented in the literature for the subproblems solved. For the second, we minimize the 2,1-norm of H, in addition to limiting its number of nonzero rows. We provide a closed-form solution for the nonconvex subproblem solved, deriving an efficient ADMM algorithm for this problem as well. Although there is no guarantee of the convergence of these two ADMMs to solutions of the nonconvex problems addressed, we demonstrate their efficacy through our numerical results. We show that: (i) our ADMM approaches are much more scalable than direct methods aimed at $(\mathcal{P}_{123}^1)$ and $(\mathcal{P}_1^{2,1})$ (using general-purpose software like `Gurobi` and `MOSEK`), and (ii) our ADMM methods provide a good complement to our local-search methods. They not only provide a means to better evaluate the solutions computed with the local-search methods, but can in fact construct solutions of comparable quality with respect to sparsity and row-sparsity in less time. Finally, we also demonstrate in this work that the efficient minimum (vector) 1-norm r-approximating local search of [24], for an ah-symmetric generalized inverse, in fact gives a factor-r approximation for 2,1-norm minimization as well, which aims directly at inducing row sparsity.

Organization. In Section 2, we develop the ADMM algorithms aimed at inducing unstructured and structured sparsity in our setting. In Section 3, we give an approximation guarantee on 2,1-norm minimization for ah-symmetric generalized inverses, for the minimum (vector) 1-norm r-approximating local search of [24]. For very low r, we do not need to settle for approximation, and so we make a detailed analysis of the cases of rank one and two in Appendix A. In Section 4, we develop the two ADMM algorithms aiming at imposing structured sparsity by limiting the number of nonzero rows for an ah-symmetric reflexive generalized inverse. In Section 5, we present the results of computational experiments, demonstrating the favorable performance of our new methods. In Section 6, we indicate some next steps for this line of research.

2 Inducing sparsity and row sparsity with ADMM

It is natural to consider specialized algorithms for attacking problems like (P_{123}^1) and $(P_1^{2,1})$, seeking fast convergence to near optima. It is even more enticing to seek such methods for their compact forms, $(\mathcal{P}_{123}^1)$ and $(\mathcal{P}_1^{2,1})$. In these compact forms, we have unconstrained minimization problems in a single variable. Motivated by for example [2, Section 6.1], we can introduce a second variable and linear linking constraints to seek to develop efficient ADMM algorithms.

2.1 ADMM for 1-norm minimization

We seek to develop an ADMM algorithm for $(\mathcal{P}_{123}^1)$. Initially, by introducing a variable $E\in\mathbb{R}^{n\times m}$, we rewrite $(\mathcal{P}_{123}^1)$ as

$$\min\left\{\|E\|_1 \;:\; E = V_1D^{-1}U_1^\top + V_2ZU_1^\top\right\}. \tag{1}$$

The augmented Lagrangian function associated to (1) is

$$\begin{aligned}\mathcal{L}_\rho(Z,E,\Lambda) &:= \|E\|_1 + \langle\Theta, V_1D^{-1}U_1^\top + V_2ZU_1^\top - E\rangle + \frac{\rho}{2}\left\|V_1D^{-1}U_1^\top + V_2ZU_1^\top - E\right\|_F^2\\ &= \|E\|_1 + \frac{\rho}{2}\left\|V_1D^{-1}U_1^\top + V_2ZU_1^\top - E + \Lambda\right\|_F^2 - \frac{\rho}{2}\|\Lambda\|_F^2 \,,\end{aligned}$$

where $\rho > 0$ is the penalty parameter, $\Theta \in \mathbb{R}^{n\times m}$ is the Lagrangian multiplier, and Λ is the scaled Lagrangian multiplier, that is $\Lambda := \Theta/\rho$. We will apply the ADMM method to $(\mathcal{P}^1_{123})$, by iteratively solving, for $k = 0, 1, \ldots,$

$$Z^{k+1} := \underset{Z}{\operatorname{argmin}} \ \mathcal{L}_\rho(Z, E^k, \Lambda^k), \tag{2}$$

$$E^{k+1} := \underset{E}{\operatorname{argmin}} \ \mathcal{L}_\rho(Z^{k+1}, E, \Lambda^k), \tag{3}$$

$$\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} U_1^\intercal + V_2 Z^{k+1} U_1^\intercal - E^{k+1}.$$

Next, we detail how to solve the subproblems above.

Update Z. To update Z, we consider subproblem (2), more specifically,

$$Z^{k+1} := \underset{Z}{\operatorname{argmin}}\{\|J - V_2 Z U_1^\intercal\|_F^2\}, \tag{4}$$

where $J := E^k - V_1 D^{-1} U_1^\intercal - \Lambda^k$. We can easily verify that the solution of (4) is given by $Z^{k+1} = V_2^\intercal J U_1$.

Update E. To update E, we consider subproblem (3), more specifically,

$$E^{k+1} := \underset{E}{\operatorname{argmin}}\{\|E\|_1 + \frac{\rho}{2}\,\|E - Y\|_F^2\}, \tag{5}$$

where $Y := V_1 D^{-1} U_1^\intercal + V_2 Z^{k+1} U_1^\intercal + \Lambda^k$.

▶ **Proposition 4** (see [2, Section 4.4.3]). *The solution of (5) is given by $E^{k+1}_{ij} = S_{1/\rho}(Y_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where the soft thresholding operator S is defined as*

$$S_\kappa(a) := \begin{cases} a - \kappa, & a > \kappa; \\ 0, & |a| \le \kappa; \\ a + \kappa, & a < -\kappa. \end{cases}$$

Initialization of the variables. We need to initialize Λ and E. First, we set $\Lambda^0 := \widehat{\Theta}/\rho$, where $\widehat{\Theta} := \frac{1}{\|V_1 U_1^\intercal\|_\infty} V_1 U_1^\intercal$. Our goal was to select the Lagrangian multiplier $\widehat{\Theta}$ as an easily computable feasible solution to the dual problem of (1), given by

$$\max_\Theta\{\operatorname{tr}(D^{-1} V_1^\intercal \Theta U_1) : V_2^\intercal \Theta U_1 = 0,\ \|\Theta\|_\infty \le 1\}.$$

Due to the orthogonality of V, we can easily verify the feasibility of $\widehat{\Theta}$. Furthermore, it gives a positive objective value and has maximum infinity norm; in contrast to the zero matrix which is also feasible.

Then, from Lemma 2, we note that the ah-symmetric reflexive generalized inverses of A can be written as $V_1 D^{-1} U_1^\intercal + V_2 Z U_1^\intercal$, and the M-P pseudoinverse $A^\dagger$ can be written as $V_1 D^{-1} U_1^\intercal$. We recall that $A^\dagger$ is the generalized inverse of A with minimum Frobenius norm. Then, aiming to obtain $Z^1 = 0$ when solving (4) in the first iteration of the algorithm, and consequently starting the algorithm with a Frobenius norm minimizing ah-symmetric reflexive generalized inverse, we set $E^0 := V_1 D^{-1} U_1^\intercal + \Lambda^0$.

Stopping criterion. We consider a stopping criterion from [2, Section 3.3.1], and select an absolute tolerance ϵ^{abs} and a relative tolerance ϵ^{rel}. The algorithm stops at iteration k if

$$\|r^k\|_F \le \epsilon^{\text{abs}}\sqrt{nm} + \epsilon^{\text{rel}} \max\left\{\|E^k\|_F\,, \|V_2 Z^k U_1^\intercal\|_F\,, \|V_1 D^{-1} U_1^\intercal\|_F\right\}, \tag{6}$$

$$\|s^k\|_F \le \epsilon^{\text{abs}}\sqrt{(n-r)r} + \epsilon^{\text{rel}}\rho\|V_2^\intercal \Lambda^k U_1\|_F\,, \tag{7}$$

where $r^k := V_1 D^{-1} U_1^\intercal + V_2 Z^k U_1^\intercal - E^k$ is the primal residual, and $s^k := \rho V_2^\intercal (E^k - E^{k-1}) U_1$ is the dual residual.

In Algorithm 1, we present the ADMM algorithm for $(\mathcal{P}^1_{123})$. We observe that in steps 7 to 9, the elements of E^{k+1} can be computed in parallel.

Input: $A \in \mathbb{R}^{m\times n}$, $\Lambda^0 \in \mathbb{R}^{n\times m}$, $E^0 \in \mathbb{R}^{n\times m}$, $\rho > 0$.
Output: $H \in \mathbb{R}^{n\times m}$.
$U, \Sigma, V := \texttt{svd}(A)$, $k := 0$;
Get U_1, V_1, V_2, D^{-1} from U, Σ, V;
while *not converged* **do**
  $J := E^k - V_1 D^{-1} U_1^\intercal - \Lambda^k$;
  $Z^{k+1} := V_2^\intercal J U_1$;
  $Y := V_1 D^{-1} U_1^\intercal + V_2 Z^{k+1} U_1^\intercal + \Lambda^k$;
  for $i = 1, \ldots, n$ **do**
    for $j = 1, \ldots, m$ **do**
      $E_{ij}^{k+1} := S_{1/\rho}(Y_{ij})$; (see Proposition 4)
  $\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} U_1^\intercal + V_2 Z^{k+1} U_1^\intercal - E^{k+1}$;
  $k := k + 1$;
$H := V_1 D^{-1} U_1^\intercal + V_2 Z^k U_1^\intercal$;

Algorithm 1: ADMM for $(\mathcal{P}_{123}^1)$ (ADMM$_1$)

2.2 ADMM for 2,1-norm minimization

Next, we apply ADMM for solving $(\mathcal{P}_1^{2,1})$. Initially, by introducing a variable $E \in \mathbb{R}^{n\times r}$, we rewrite $(\mathcal{P}_1^{2,1})$ as

$$\min \left\{ \|E\|_{2,1} \ : \ E = V_1 D^{-1} + V_2 Z \right\}. \tag{8}$$

► Remark 5. In Theorem 3, we use the orthogonality of U_1 to replace $\|V_1 D^{-1} U_1^\intercal + V_2 Z U_1^\intercal\|_{2,1}$ with $\|V_1 D^{-1} + V_2 Z\|_{2,1}$ in the objective function of $(\mathcal{P}_1^{2,1})$. This leads to a very helpful reduction in the dimension of the variable E in the ADMM algorithm presented in this section, compared to what we presented in the previous section.

The augmented Lagrangian function associated to (8) is

$$\mathcal{L}_\rho(Z, E, \Lambda) := \|E\|_{2,1} + \frac{\rho}{2} \left\|V_1 D^{-1} + V_2 Z - E + \Lambda\right\|_F^2 - \frac{\rho}{2} \|\Lambda\|_F^2 \,,$$

where $\rho > 0$ is the penalty parameter and $\Lambda \in \mathbb{R}^{n\times r}$ is the scaled Lagrangian multiplier. We will apply the ADMM method to $(\mathcal{P}_1^{2,1})$, by iteratively solving, for $k = 0, 1, \ldots$,

$$Z^{k+1} := \operatorname*{argmin}_Z \ \mathcal{L}_\rho(Z, E^k, \Lambda^k), \tag{9}$$

$$E^{k+1} := \operatorname*{argmin}_E \ \mathcal{L}_\rho(Z^{k+1}, E, \Lambda^k), \tag{10}$$

$$\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} + V_2 Z^{k+1} - E^{k+1}.$$

Next, we detail how to solve the subproblems above.

Update of Z. To update Z, we consider subproblem (9), that is

$$Z^{k+1} := \operatorname*{argmin}_Z \|J - V_2 Z\|_F^2 \,, \tag{11}$$

where $J := E^k - V_1 D^{-1} - \Lambda^k$. Similarly to (4), we can verify that $Z^{k+1} = V_2^\intercal J$ is an optimal solution to (11).

Update of E. To update E, we consider subproblem (10), that is

$$E^{k+1} := \operatorname*{argmin}_E \{\|E\|_{2,1} + \frac{\rho}{2} \|E - Y\|_F^2\}, \tag{12}$$

where $Y := V_1 D^{-1} + V_2 Z^{k+1} + \Lambda^k$.

► **Proposition 6** (see [25, Proposition 1]). *The solution of* (12) *is given by*

$$E_{i\cdot}^{k+1} := \begin{cases} \frac{\|Y_{i\cdot}\|_2 - 1/\rho}{\|Y_{i\cdot}\|_2} Y_{i\cdot} \,, & \text{if } 1/\rho < \|Y_{i\cdot}\|_2 \,; \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Initialization of the variables. To initialize the variables, we apply the same ideas discussed in Section 2.1. Considering now the dual problem of (8),

$$\max_{\Theta}\{\operatorname{tr}(D^{-1}V_1^{\mathsf{T}}\Theta) : V_2^{\mathsf{T}}\Theta = 0, \|\Theta_{i\cdot}\|_2 \leq 1, i = 1,\ldots,n\},$$

we set $\widehat{\Theta} := V_1/\kappa$, where $\kappa := \max_{i=1,\ldots,n}\|V_1[i,\cdot]\|_2$, $\Lambda^0 := \widehat{\Theta}/\rho$, and $E^0 := V_1D^{-1} + \Lambda^0$.

Stopping criterion. We adopt the same stopping criterion described in Section 2.1. With the primal and dual residuals at iteration k given respectively by $r^k := V_1D^{-1} + V_2Z^k - E^k$ and $s^k := \rho V_2^{\mathsf{T}}(E^k - E^{k-1})$, the criterion becomes

$$\|r^k\|_F \leq \epsilon^{\text{abs}}\sqrt{nr} + \epsilon^{\text{rel}}\max\left\{\|E^k\|_F\,, \|V_2Z^k\|_F\,, \|V_1D^{-1}\|_F\right\}, \tag{14}$$

$$\|s^k\|_F \leq \epsilon^{\text{abs}}\sqrt{(n-r)r} + \epsilon^{\text{rel}}\rho\|V_2^{\mathsf{T}}\Lambda^k\|_F\,. \tag{15}$$

In Algorithm 2, we present the ADMM algorithm for $(\mathcal{P}_1^{2,1})$.

Input: $A \in \mathbb{R}^{m\times n}$, $\Lambda^0 \in \mathbb{R}^{n\times r}$, $E^0 \in \mathbb{R}^{n\times r}$, $\rho > 0$.
Output: $H \in \mathbb{R}^{n\times m}$.
$U, \Sigma, V := \texttt{svd}(A)$, $k := 0$;
Get U_1, V_1, V_2, D^{-1} from U, Σ, V;
while *not converged* **do**
  $J := E^k - V_1D^{-1} - \Lambda^k$;
  $Z^{k+1} := V_2^{\mathsf{T}}J$;
  $Y := V_1D^{-1} + V_2Z^{k+1} + \Lambda^k$;
  for $i = 1,\ldots,n$ **do**
    if $\|Y_{i\cdot}\|_2 > 1/\rho$ **then** $E_{i\cdot}^{k+1} := \frac{\|Y_{i\cdot}\|_2 - 1/\rho}{\|Y_{i\cdot}\|_2}Y_{i\cdot}$ **; (see Proposition 6)**
    else $E_{i\cdot}^{k+1} := 0$**;**
  $\Lambda^{k+1} := \Lambda^k + V_1D^{-1} + V_2Z^{k+1} - E^{k+1}$;
  $k := k + 1$;
$H := V_1D^{-1}U_1^{\mathsf{T}} + V_2Z^kU_1^{\mathsf{T}}$;

Algorithm 2: ADMM for $(\mathcal{P}_1^{2,1})$ $(\text{ADMM}_{2,1})$

3 Imposing row sparsity with column-block solutions and local search

[24, Section 3] developed a local-search based approximation algorithm to efficiently calculate an approximate (vector) 1-norm minimizing ah-symmetric generalized inverse with $r := \operatorname{rank}(A)$ nonzero rows. It is very natural to investigate how well this approximation algorithm does for other norms, in particular the 2,1-norm, which we emphasize in our present work. Here we show that the same matrix calculated by this approximation algorithm is also a good approximate 2,1-norm minimizing ah-symmetric generalized inverse. We note that our approximation factor for the 2,1-norm is much better than what would be obtained by simply applying norm inequalities to the result for the 1-norm (see Remark 12). Our local search works with "column-block solutions", as presented in the following result.

▶ **Theorem 7** ([24, Theorem 3.3])**.** *For $A \in \mathbb{R}^{m\times n}$, let $r := \operatorname{rank}(A)$. For any T, an ordered subset of r elements from $\{1,\ldots,n\}$, let $\hat{A} := A[\cdot, T]$ be the $m \times r$ submatrix of A formed by columns T. If $\operatorname{rank}(\hat{A}) = r$, let $\hat{H} := \hat{A}^{\dagger} = (\hat{A}^{\mathsf{T}}\hat{A})^{-1}\hat{A}^{\mathsf{T}}$. Then the $n \times m$ matrix H with all rows equal to zero, except rows T, which are given by $\hat{H}$, is an ah-symmetric reflexive generalized inverse of A.*

In Appendix A, we consider the cases in which $\operatorname{rank}(A) \in \{1, 2\}$. As was done for the 1-norm in [23], we use the result in Theorem 7 to address the exact 2,1-norm minimization of ah-symmetric reflexive generalized inverses for these simpler cases. After observing the difficulty of the exact minimization already for rank 2, we consider an approximation algorithm for general matrices of constant rank r, as discussed next.

▶ **Definition 8** ([24, Definition 3.5])**.** *For $A \in \mathbb{R}^{m\times n}$, let $r := \operatorname{rank}(A)$, and S be an ordered subset of r elements from $\{1,\ldots,m\}$ such that $A[S,\cdot]$ has linearly independent rows. For T an ordered subset of r elements from $\{1,\ldots,n\}$, and fixed $\epsilon \geq 0$, if $|\det(A[S,T])|$ cannot be increased by a factor of more than $1+\epsilon$ by swapping an element of T with one from its complement, we say that $A[S,T]$ is a $(1+\epsilon)$-local maximizer for the absolute determinant on the set of $r \times r$ nonsingular submatrices of $A[S,\cdot]$.*

▶ Remark 9. [24, Theorem 3.9] showed that a local optimum satisfying Definition 8 can be calculated in polynomial time (when A is rational), for any fixed $\epsilon > 0$.

▶ **Theorem 10** ([24, Theorem 3.7])**.** *For $A \in \mathbb{R}^{m\times n}$, let $r := \operatorname{rank}(A)$, and let S be an ordered subset of r elements from $\{1,\ldots,m\}$ such that $A[S,\cdot\,]$ has linearly independent rows. Choose $\epsilon \geq 0$, and let $\widetilde{A} := A[S,T]$ be a $(1+\epsilon)$-local maximizer for the absolute determinant on the set of $r\times r$ nonsingular submatrices of $A[S,\cdot\,]$. Then the $n\times m$ matrix H constructed by Theorem 7 over $\widehat{A} := A[\,\cdot\,,T]$, is an ah-symmetric reflexive generalized inverse of A satisfying $\|H\|_1 \leq r(1+\epsilon)\|H^1_{opt}\|_1$, where H^1_{opt} is a (vector) 1-norm minimizing ah-symmetric reflexive generalized inverse of A.*

Next, we will show that the constructed H of Theorem 10 has 2,1-norm within a factor of r of the 2,1-norm minimizing solution as well.

▶ **Theorem 11.** *For $A \in \mathbb{R}^{m\times n}$, let $r := \operatorname{rank}(A)$, and let S be an ordered subset of r elements from $\{1,\ldots,m\}$ such that $A[S,\cdot\,]$ has linearly independent rows. Choose $\epsilon \geq 0$, and let $\widetilde{A} := A[S,T]$ be a $(1+\epsilon)$-local maximizer for the absolute determinant on the set of $r\times r$ nonsingular submatrices of $A[S,\cdot\,]$. Then the $n\times m$ matrix H constructed by Theorem 7 over $\widehat{A} := A[\,\cdot\,,T]$, is an ah-symmetric reflexive generalized inverse of A satisfying $\|H\|_{2,1} \leq r(1+\epsilon)\|H^{2,1}_{opt}\|_{2,1}$, where $H^{2,1}_{opt}$ is a 2,1-norm minimizing ah-symmetric reflexive generalized inverse of A.*

We note that $H^{2,1}_{opt}$ defined in Theorem 11 is also a 2,1-norm minimizing generalized inverse of A because as stated in Section 1, optimal solutions of $(P^{2,1}_1)$ automatically satisfy (P2) and (P3) (i.e., they are reflexive and ah-symmetric).

▶ Remark 12. Theorem 11 is stronger than the direct consequence of Theorem 10 that can be obtained via norm inequalities. Specifically, for $x \in \mathbb{R}^m$, we have $\|x\|_2 \leq \|x\|_1 \leq \sqrt{m}\|x\|_2$ [12, Eq. 2.2.5]. Then, for $H \in \mathbb{R}^{n\times m}$, we have $\|H\|_{2,1} = \sum_{i=1}^n \|H_{i\cdot}\|_2 \leq \sum_{i=1}^n \|H_{i\cdot}\|_1 = \|H\|_1$, and $\|H\|_1 = \sum_{i=1}^n \|H_{i\cdot}\|_1 \leq \sqrt{m}\sum_{i=1}^n \|H_{i\cdot}\|_2 = \sqrt{m}\|H\|_{2,1}$. Then, for $H \in \mathbb{R}^{n\times m}$ constructed as in Theorem 10 we have $\|H\|_{2,1} \leq \|H\|_1 \leq r(1+\epsilon)\|H^1_{opt}\|_1 \leq r(1+\epsilon)\|H^{2,1}_{opt}\|_1 \leq r\sqrt{m}(1+\epsilon)\|H^{2,1}_{opt}\|_{2,1}$. That dependence on m is concerning, in the context of least-squares applications, where m can be huge.

To prove Theorem 11, we will work with the convex-optimization formulation for a 2,1-norm minimizing generalized inverse,

$$\min\left\{\|H\|_{2,1} \;:\; AHA = A\right\}, \tag{$P^{2,1}_1$}$$

and its dual

$$\max\left\{\operatorname{tr}(A^\intercal W) \;:\; \|(A^\intercal W A^\intercal)_{i\cdot}\|_2 \leq 1,\ i=1,\ldots,n\right\}. \tag{$D^{2,1}_1$}$$

Next, we present some technical results that will be used to prove our main result.

▶ **Lemma 13.** *Let T be an ordered subset of r elements from $\{1,\ldots,n\}$ and $\widehat{A} := A[\,\cdot\,,T]$ be the $m\times r$ submatrix of an $m\times n$ matrix A formed by columns T, and $\operatorname{rank}(\widehat{A}) = r$. There exists an $r\times r$ matrix E with $\|E_{i\cdot}\widehat{A}^\intercal\|_2 = 1$ and $E_{ii} = \|\widehat{A}^\dagger_{i\cdot}\|_2$ for $i \in \{1,\ldots,r\}$.*

Proof. Suppose that $\widehat{A} = U\Sigma V^\intercal$ is the compact singular value decomposition of $\widehat{A}$, where $U \in \mathbb{R}^{m\times r}$, $\Sigma, V \in \mathbb{R}^{r\times r}$, and Σ is diagonal. Let $\widehat{V} = \Sigma^{-1}V^\intercal = [z_1,\ldots,z_r]$ ($\Sigma^{-1}V^\intercal \mathbf{e}_i = z_i$), and $E_{ij} = \frac{z_i^\intercal z_j}{\|z_i\|_2}$ ($E = M\widehat{V}^\intercal\widehat{V} = MV\Sigma^{-2}V^\intercal$, where M is a diagonal matrix with $M_{ii} = \frac{1}{\|z_i\|_2}$). We know that $\widehat{A}^\dagger = V\Sigma^{-1}U^\intercal$. Then $E_{ii} = \frac{z_i^\intercal z_i}{\|z_i\|_2} = \|z_i\|_2 = \|\Sigma^{-1}V^\intercal\mathbf{e}_i\|_2 = \|U\Sigma^{-1}V^\intercal\mathbf{e}_i\|_2 = \|\widehat{A}^\dagger_{i\cdot}\|_2$, where the second last equality follows from orthogonality of U. Also, as $\widehat{A}^\intercal = V\Sigma U^\intercal$, we have $E_{i\cdot}\widehat{A}^\intercal = \frac{1}{\|z_i\|_2}\mathbf{e}_i^\intercal V\Sigma^{-2}V^\intercal V\Sigma U^\intercal = \frac{1}{\|z_i\|_2}\mathbf{e}_i^\intercal V\Sigma^{-1}U^\intercal = \frac{1}{\|z_i\|_2}\widehat{A}^\dagger_{i\cdot}$. Because $\|z_i\|_2 = \|\widehat{A}^\dagger_{i\cdot}\|_2$, we have $\|E_{i\cdot}\widehat{A}^\intercal\|_2 = 1$. ◀

▶ **Lemma 14.** *Let T be an ordered subset of r elements from $\{1,\ldots,n\}$ and $\widehat{A} := A[\,\cdot\,,T]$ be the $m\times r$ submatrix of an $m\times n$ matrix A formed by columns T, and $\operatorname{rank}(\widehat{A}) = r$. Let E be an $r\times r$ matrix such that $\|E_{i\cdot}\widehat{A}^\intercal\|_2 = 1$ and $E_{ii} = \|\widehat{A}^\dagger_{i\cdot}\|_2$, for $i \in \{1,\ldots,r\}$ (which exists by Lemma 13). There exists an $m\times n$ matrix W such that $\widehat{A}^\intercal W A^\intercal = E\widehat{A}^\intercal$ and $\operatorname{tr}(A^\intercal W) = \|\widehat{A}^\dagger\|_{2,1}$.*

Proof. Let S be an ordered subset of r elements from $\{1,\ldots,m\}$ such that $\widetilde{A} := \widehat{A}[S,\cdot\,]$ is a nonsingular $r\times r$ submatrix of $\widehat{A}$ formed by rows S. Let $\widehat{W}$ be an $r\times r$ matrix and W be an $m\times n$ matrix with all elements equal to zero, except the ones in rows S and columns T which are given by the respective elements in $\widehat{W}$. Choose $\widehat{W} := \widetilde{A}^{-\intercal}E$, then we have $\widehat{A}^\intercal W A^\intercal = \widetilde{A}^\intercal\widehat{W}\widehat{A}^\intercal = E\widehat{A}^\intercal$ and $\operatorname{tr}(A^\intercal W) = \operatorname{tr}(\widetilde{A}^\intercal\widehat{W}) = \operatorname{tr}(E) = \|\widehat{A}^\dagger\|_{2,1}$. ◀

We now proceed to prove Theorem 11.

Proof. We will construct a dual-feasible solution with objective value $\frac{1}{r(1+\epsilon)}\|H\|_{2,1}$. By weak duality, we will then have $\frac{1}{r(1+\epsilon)}\|H\|_{2,1} \leq \|H_{opt}\|_{2,1}$.

By Lemma 14, we can always choose W such that $\widehat{A}^\intercal W A^\intercal = E\widehat{A}^\intercal$ and $\operatorname{tr}(A^\intercal W) = \operatorname{tr}(E) = \|\widehat{A}^\dagger\|_{2,1} = \|H\|_{2,1}$, where E is any given $r \times r$ matrix such that $\|E_{i\cdot}\widehat{A}^\intercal\|_2 = 1$ and $E_{ii} = \|\widehat{A}^\dagger_{i\cdot}\|_2$, for $i \in \{1, \ldots, r\}$.

So it is sufficient to demonstrate that $\|(A^\intercal W A^\intercal)_{i\cdot}\|_2 \leq r(1+\epsilon)$ for $i = 1, \ldots, n$ (then $\frac{1}{r(1+\epsilon)}W$ is dual feasible), and $\operatorname{tr}\left(A^\intercal\left(\frac{1}{r(1+\epsilon)}W\right)\right) = \frac{1}{r(1+\epsilon)}\|H\|_{2,1}$.

First, it is clear that $\|(\widehat{A}^\intercal W A^\intercal)_{i\cdot}\|_2 = \|E_{i\cdot}\widehat{A}^\intercal\|_2 = 1 \leq r(1+\epsilon)$, for $i \in T$. Now, consider any column $\widehat{b}$ of $A[\cdot, N \setminus T]$. Because $\operatorname{rank}(\widehat{A}) = r$, we have that $\widehat{b} = \widehat{A}\beta$, for some $\beta \in \mathbb{R}^r$, which implies that $\widetilde{b} = \widetilde{A}\beta$, where $\widetilde{b} := \widehat{b}[S]$. By Cramer's rule, where $\widetilde{A}_i(\widetilde{b})$ is $\widetilde{A}$ with column i replaced by $\widetilde{b}$, we have $|\beta_i| = \frac{|\det(\tilde{A}_i(\tilde{b}))|}{|\det(\tilde{A})|} \leq 1+\epsilon$, because $\widetilde{A}$ is a $(1+\epsilon)$-local maximizer for the absolute determinant of $A[S,\cdot]$. Therefore

$$\begin{aligned}
\|\widehat{b}^\intercal W A^\intercal\|_2 &= \|\beta^\intercal \widehat{A}^\intercal W A^\intercal\|_2 = \|\beta^\intercal E \widehat{A}^\intercal\|_2 = \|\sum_{i=1}^{r} \beta_i \cdot E_{i\cdot}\widehat{A}^\intercal\|_2 \\
&\leq \sum_{i=1}^{r} \|\beta_i \cdot E_{i\cdot}\widehat{A}^\intercal\|_2 = \sum_{i=1}^{r} |\beta_i| \cdot \|E_{i\cdot}\widehat{A}^\intercal\|_2 \leq r(1+\epsilon),
\end{aligned}$$

where the first inequality comes from the triangle inequality. ◀

► Remark 15. We wish to emphasize that even though the local search of Definition 8 does not directly consider either the (vector) 1-norm or the 2,1-norm in its operation, Theorems 10 and 11 still provide approximation guarantees for both. Furthermore, we showed in [24, Example A.1 (3)], that the approximation ratio established in Theorem 10 is essentially the best possible for the 1-norm. Using the same example, we can also show that the approximation ratio established in Theorem 11 for the 2,1-norm is also essentially best possible.

As emphasized in Remark 15, although our goal is to minimize the 2,1-norm over ah-symmetric reflexive generalized inverses, the local-search procedure presented above uses instead, the absolute determinant of $r \times r$ nonsingular submatrices of A as a criterion for improving the constructed solution. The advantage of the criterion used is twofold: it leads to the approximation result shown in Theorem 11 and to a very efficient implementation based on rank-1 update of the determinant.

Nevertheless, a natural further investigation concerns how the results compare to the solution obtained by the local search modified to use as a criterion for improving the solution, the actual 2,1-norm of H. More specifically, the criterion for improving the column-block solution H constructed by Theorem 7, is modified to be the decrease in its 2,1-norm, or equivalently, the decrease in the 2,1-norm of the M-P pseudoinverse of the $m \times r$ full column-rank submatrix of A being considered. To evaluate how much the 2,1-norm of the M-P pseudoinverse of the submatrix changes when each column of $A[\cdot, T]$ is replaced by a given column γ of $A[\cdot, N \setminus T]$, we use the next result.

► **Proposition 16.** *Let $A := (a_1, \ldots, a_j, \ldots, a_r) \in \mathbb{R}^{m \times r}$ with $\operatorname{rank}(A) = r$ and $\gamma \in \mathcal{R}(A)$. Let $\bar{A}$ be the matrix obtained by replacing the j^{th} column of A by γ, and $v = (v_1, \ldots, v_j, \ldots, v_r)^\intercal := A^\dagger\gamma$. If $v_j \neq 0$, let $\bar{v} := \frac{1}{v_j}(-v_1, \ldots, -v_{j-1}, 1, -v_{j+1}, \ldots, -v_r)^\intercal$. Then*

$$\|\bar{A}^\dagger\|_{2,1} = |\bar{v}_j| \cdot \|A^\dagger_{j\cdot}\|_2 + \sum_{\substack{i=1\\ i\neq j}}^{r} \left(\|A^\dagger_{i\cdot}\|_2^2 + 2\bar{v}_i A^\dagger_{i\cdot}(A^\dagger_{j\cdot})^\intercal + \bar{v}_i^2\|A^\dagger_{j\cdot}\|_2^2\right)^{1/2}.$$

Proof. Because $\gamma \in \mathcal{R}(A)$, we have $Av = AA^\dagger\gamma = \gamma$. Then, $\bar{A} = A\Theta$, where $\Theta := (e_1, \ldots, e_{j-1}, v, e_{j+1}, \ldots, e_r)$ is non-singular, and we can verify that $\operatorname{rank}(\bar{A}) = \operatorname{rank}(A) = r$ and that

$$\bar{A}\bar{A}^\dagger = \bar{A}(\bar{A}^\intercal\bar{A})^{-1}\bar{A}^\intercal = A\Theta(\Theta^\intercal A^\intercal A\Theta)^{-1}\Theta^\intercal A^\intercal = A(A^\intercal A)^{-1}A^\intercal = AA^\dagger.$$

Then, we have

$$\bar{A}\bar{A}^\dagger = A\Theta\bar{A}^\dagger \Rightarrow AA^\dagger = A\Theta\bar{A}^\dagger \Rightarrow A^\dagger AA^\dagger = A^\dagger A\Theta\bar{A}^\dagger \Rightarrow A^\dagger = \Theta\bar{A}^\dagger,$$

where the last implication holds because A has full column rank.

Then, $\bar{A}^{\dagger} = \Theta^{-1}A^{\dagger}$, with $\Theta^{-1} = (e_1, \ldots, e_{j-1},\ \bar{v},\ e_{j+1}, \ldots, e_r)$, or equivalently, $\bar{A}^{\dagger}_{i\cdot} = A^{\dagger}_{i\cdot} + \bar{v}_i A^{\dagger}_{j\cdot}$, for $i \neq j$, and $\bar{A}^{\dagger}_{j\cdot} = \bar{v}_j A^{\dagger}_{j\cdot}$. So, we finally have

$$\|\bar{A}^{\dagger}_{i\cdot}\|_2 = \left((A^{\dagger}_{i\cdot} + \bar{v}_i A^{\dagger}_{j\cdot})(A^{\dagger}_{i\cdot} + \bar{v}_i A^{\dagger}_{j\cdot})^{\mathsf{T}}\right)^{1/2} = \left(\|A^{\dagger}_{i\cdot}\|_2^2 + 2\bar{v}_i A^{\dagger}_{i\cdot}(A^{\dagger}_{j\cdot})^{\mathsf{T}} + \bar{v}_i^2\|A^{\dagger}_{j\cdot}\|_2^2\right)^{1/2},$$

for $i \neq j$, and $\|\bar{A}^{\dagger}_{j\cdot}\|_2 = |\bar{v}_j| \cdot \|A^{\dagger}_{j\cdot}\|_2$. The result follows. ◀

▶ Remark 17. Computing $A^{\dagger}_{i\cdot}(A^{\dagger}_{j\cdot})^{\mathsf{T}}$ for a given j and all $i \in \{1, \ldots, r\}$ to update the 2,1-norm of the M-P pseudoinverse, as described in Proposition 16, at every iteration of our local-search procedure, is still quite time consuming. To address this, we note that $\bar{A}^{\dagger}_{i\cdot} = A^{\dagger}_{i\cdot} + \bar{v}_i A^{\dagger}_{j\cdot}$, for $i \neq j$, and $\bar{A}^{\dagger}_{j\cdot} = \bar{v}_j A^{\dagger}_{j\cdot}$. Therefore, defining $W := A^{\dagger}(A^{\dagger})^{\mathsf{T}}$ we note that it is possible to compute $\bar{A}^{\dagger}(\bar{A}^{\dagger})^{\mathsf{T}}$ by the following.

$$\left(\bar{A}^{\dagger}(\bar{A}^{\dagger})^{\mathsf{T}}\right)_{i\ell} = \begin{cases} W_{i\ell} + \bar{v}_i\bar{v}_\ell\|A^{\dagger}_{j\cdot}\|_2^2 + W_{ij}\bar{v}_\ell + W_{j\ell}\bar{v}_i, & i \neq j,\ \ell \neq j; \\ \bar{v}_j\left(W_{i\ell} + \|A^{\dagger}_{j\cdot}\|_2^2\bar{v}_i\right), & i \neq j,\ \ell = j; \\ \bar{v}_j^2\|A^{\dagger}_{j\cdot}\|_2^2, & i = j,\ \ell = j. \end{cases}$$

The update above drastically improves upon computing the $A^{\dagger}_{i\cdot}(A^{\dagger}_{j\cdot})^{\mathsf{T}}$ from scratch.

4 Targeting row sparsity with ADMM

It is a common approach to induce row-sparsity for a matrix by minimizing its 2,1-norm, in order to take advantage of the convexity of the problem addressed (see [25, 16, 18, 13] and [2, Section 6.4.2], for example). However, we can observe from our numerical experiments, that the ah-symmetric reflexive generalized inverse of a given matrix $A \in \mathbb{R}^{m\times n}$, obtained with this approach, has in general significantly more nonzero rows than the solution of the local-search procedure described in Section 3, for which the number of nonzero rows, or equivalently, the 2,0-norm, is $r := \operatorname{rank}(A)$, the least possible number. On the other side, we observe that the solution obtained by the local search has in general much larger 2,1-norm than a 2,1-norm minimizing generalized inverse. Aiming at matrices with both nice features of having small 2,1- and 2,0-norms, we now seek an ah-symmetric reflexive generalized inverse with 2,1-norm smaller than the local-search solution and 2,0-norm smaller than the 2,1-norm minimizing solution. Inspired by ideas in [14], we will present ADMM algorithms to compute an ah-symmetric reflexive generalized inverses for A, with the 2,0-norm limited to $\gamma := \omega r + (1-\omega)\|H^{2,1}_{opt}\|_{2,0}$, where $0 < \omega < 1$, and $H^{2,1}_{opt}$ an optimal solution of $(\mathcal{P}^{2,1}_1)$. More specifically, we will present ADMM algorithms for *nonconvex* problems of obtaining a solution $\bar{Z}$ of $\|V_1D^{-1} + V_2Z\|_{2,0} \leq \gamma$. Initially, aiming at a more efficient application of the ADMM method, we do not consider the minimization of the 2,1-norm, so the problem addressed is a nonconvex feasibility problem. Then, we investigate the impact on the solutions by minimizing the 2,1-norm subject to the same nonconvex feasible set. Our goal is to compare through numerical experiments, the solutions obtained by the ADMM algorithms proposed next, with each other and also to the local-search solutions and the solutions of $(\mathcal{P}^{2,1}_1)$.

We wish to note that the ADMM approach that we develop, for handling an inequality-constrained minimization problem in one variable by introducing a second variable, an indicator function on the second variable, and linear linking constraints is a known scheme (at a high level); see [2, introductory passages of Section 5, and Sections 5.2, 6.2 and 9.1], for example.

4.1 ADMM for limited 2,0-norm

By introducing a variable $E \in \mathbb{R}^{n\times r}$, we reformulate the feasibility problem of obtaining a solution $\bar{Z}$ of $\|V_1D^{-1} + V_2Z\|_{2,0} \leq \gamma$ as

$$\min\{\mathcal{I}_{\mathcal{M}}(E)\ :\ E = V_1D^{-1} + V_2Z\}, \tag{16}$$

where the indicator function $\mathcal{I}_{\mathcal{M}}(\,\cdot\,)$ is defined by

$$\mathcal{I}_{\mathcal{M}}(X) := \begin{cases} 0, & X \in \mathcal{M}; \\ +\infty, & X \notin \mathcal{M}. \end{cases}$$

for the set $\mathcal{M} := \{X \in \mathbb{R}^{n \times r} \ : \ \|X\|_{2,0} \leq \gamma\}$.

The augmented Lagrangian function associated to (16) is

$$\mathcal{L}_\rho(Z, E, \Lambda) := \mathcal{I}_\mathcal{M}(E) + \frac{\rho}{2}\left\|V_1 D^{-1} + V_2 Z - E + \Lambda\right\|_F^2 - \frac{\rho}{2}\left\|\Lambda\right\|_F^2 ,$$

where $\rho > 0$ is the penalty parameter and $\Lambda \in \mathbb{R}^{n \times r}$ is the scaled Lagrangian multiplier. We will apply the ADMM method to (16), by iteratively solving, for $k = 0, 1, \ldots,$

$$Z^{k+1} := \underset{Z}{\operatorname{argmin}} \ \mathcal{L}_\rho(Z, E^k, \Lambda^k), \tag{17}$$

$$E^{k+1} := \underset{E}{\operatorname{argmin}} \ \mathcal{L}_\rho(Z^{k+1}, E, \Lambda^k), \tag{18}$$

$$\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} + V_2 Z^{k+1} - E^{k+1}.$$

Subproblem (17) is exactly the same as (9). Next, we detail how to solve (18).

Update of E.

$$E^{k+1} := \underset{E}{\operatorname{argmin}}\{\mathcal{I}_\mathcal{M}(E) + \frac{\rho}{2}\left\|E - Y\right\|_F^2\},$$

where $Y := V_1 D^{-1} + V_2 Z^{k+1} + \Lambda^k$. Although $\mathcal{M}$ is a nonconvex set, the exact solution for the above subproblem can be efficiently computed (see [2, Section 9.1]). It is given by the projection of Y onto $\mathcal{M}$, which we will denote in the following by $\Pi_\mathcal{M}(Y)$. $\Pi_\mathcal{M}(Y)$ keeps the rows of Y with the γ largest 2-norms and zeros out the other rows.

Initialization. To initialize the variables, we follow the approach described in Section 2.1 and Section 2.2. But because of the nonconvexity of (16), instead of considering its dual problem to initialize Λ, we simply set $\Lambda^0 := 0$. As in Section 2.2, we set $E^0 := V_1 D^{-1} + \Lambda^0$.

Stopping criteria Considering the nonconvexity of our feasibility problem, formulated as (16), we stop the ADMM when a feasible solution is found, i.e., when $\|V_1 D^{-1} + V_2 Z^k\|_{2,0} \leq \gamma$, (or when the time limit is reached).

Pseudocode. In Algorithm 3, we present the ADMM algorithm for (16) .

Input: $A \in \mathbb{R}^{m \times n}$, $\Lambda^0 \in \mathbb{R}^{n \times r}$, $E^0 \in \mathbb{R}^{n \times r}$, $0 < \omega < 1$, $r := \operatorname{rank}(A)$, $\|H_{opt}^{2,1}\|_{2,0}$.
Output: $H \in \mathbb{R}^{n \times m}$.
$U, \Sigma, V := \mathtt{svd}(A)$, $k := 0$;
Get U_1, V_1, V_2, D^{-1} from U, Σ, V;
$\mathcal{M} := \{X \in \mathbb{R}^{n \times r} \ : \ \|X\|_{2,0} \leq \omega r + (1-\omega)\|H_{opt}^{2,1}\|_{2,0}\}$;
while *not converged* **do**
  $J := E^k - V_1 D^{-1} - \Lambda^k$;
  $Z^{k+1} := V_2^\intercal J$;
  $Y := V_1 D^{-1} + V_2 Z^{k+1} + \Lambda^k$;
  $E^{k+1} := \Pi_\mathcal{M}(Y)$;
  $\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} + V_2 Z^{k+1} - E^{k+1}$;
  $k := k + 1$;
$H := V_1 D^{-1} U_1^\intercal + V_2 Z^k U_1^\intercal$;

Algorithm 3: ADMM for (16) ($\text{ADMM}_{2,0}$)

4.2 ADMM for 2,1-norm minimization subject to limited 2,0-norm

Next, we introduce variable $E \in \mathbb{R}^{n \times r}$ in $(\mathcal{P}_1^{2,1})$ and restrict it to solutions with 2,0-norm limited to γ, reformulating it as

$$\min_{E,Z}\left\{\|E\|_{2,1} \ : \ E = V_1 D^{-1} + V_2 Z, \ \|E\|_{2,0} \leq \gamma\right\}. \tag{19}$$

Then, we reformulate (19) as

$$\min_{E,Z}\left\{\|E\|_{2,1} + \mathcal{I}_\mathcal{M}(E) \ : \ E = V_1 D^{-1} + V_2 Z\right\}. \tag{20}$$

The augmented Lagrangian function associated to (20) is

$$\mathcal{L}_\rho(Z,E,\Lambda) := \|E\|_{2,1} + \mathcal{I}_{\mathcal{M}}(E) + \frac{\rho}{2}\left\|V_1D^{-1} + V_2Z - E + \Lambda\right\|_F^2 - \frac{\rho}{2}\|\Lambda\|_F^2,$$

where $\rho > 0$ is the penalty parameter and $\Lambda \in \mathbb{R}^{n\times r}$ is the scaled Lagrangian multiplier. We will apply the ADMM method to (20), by iteratively solving, for $k = 0, 1, \ldots,$

$$Z^{k+1} := \underset{Z}{\operatorname{argmin}}\ \mathcal{L}_\rho(Z, E^k, \Lambda^k), \tag{21}$$

$$E^{k+1} := \underset{E}{\operatorname{argmin}}\ \mathcal{L}_\rho(Z^{k+1}, E, \Lambda^k), \tag{22}$$

$$\Lambda^{k+1} := \Lambda^k + V_1D^{-1} + V_2Z^{k+1} - E^{k+1}.$$

Subproblem (21) is exactly the same as (9). Next, we detail how to solve (22).

Update of E. To update E, we consider subproblem (22), that is

$$\begin{aligned} E^{k+1} &:= \underset{E\in\mathbb{R}^{n\times r}}{\operatorname{argmin}}\left\{\|E\|_{2,1} + \mathcal{I}_{\mathcal{M}}(E) + \frac{\rho}{2}\left\|V_1D^{-1} + V_2Z^{k+1} - E + \Lambda^k\right\|_F^2\right\} \\ &= \underset{E\in\mathcal{M}}{\operatorname{argmin}}\left\{\|E\|_{2,1} + \frac{\rho}{2}\|E - Y\|_F^2\right\}, \end{aligned} \tag{23}$$

where $Y := V_1D^{-1} + V_2Z^{k+1} + \Lambda^k$.

▶ **Theorem 18.** *Let $\widetilde{E}^{k+1} \in \mathbb{R}^{n\times r}$ be the optimal solution to the unconstrained version of subproblem (23), that is,*

$$\widetilde{E}^{k+1} := \operatorname{argmin}\{\|E\|_{2,1} + \frac{\rho}{2}\|E - Y\|_F^2\}. \tag{24}$$

Let

$$g(i) := \frac{\rho}{2}\|Y_{i\cdot}\|_2^2 - \|\widetilde{E}_{i\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|\widetilde{E}_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2, \tag{25}$$

for $i \in N := \{1,\ldots,n\}$. Let ϕ be the permutation of the indices in N such that $g(\phi_1) \ge g(\phi_2) \ge \cdots \ge g(\phi_n)$. Then, an optimal solution $E^{k+1} \in \mathbb{R}^{n\times r}$ of (23) is given by

$$E_{\phi_i\cdot}^{k+1} := \begin{cases} \widetilde{E}_{\phi_i\cdot}^{k+1}, & \textit{if } i \le \gamma; \\ 0, & \textit{otherwise.} \end{cases} \tag{26}$$

Proof. From (13), we see that

(i) if $1/\rho < \|Y_{i\cdot}\|_2$, then $\widetilde{E}_{i\cdot}^{k+1} \neq 0$, and we can verify that $g(i) = \frac{\rho}{2}\left(\|Y_{i\cdot}\| - \frac{1}{\rho}\right)^2 > 0$,

(ii) if $1/\rho \ge \|Y_{i\cdot}\|_2$, then $\widetilde{E}_{i\cdot}^{k+1} = 0$, and we can verify that $g(i) = 0$.

From (i) and (ii), we see that $E^{k+1} = \widetilde{E}^{k+1}$ if $\|\widetilde{E}^{k+1}\|_{2,0} \le \gamma$, and the statement of the theorem trivially follows. Therefore, in the following, we assume that $\|\widetilde{E}^{k+1}\|_{2,0} > \gamma$. Note that in this case we have that $\|\widetilde{E}_{\phi_i\cdot}^{k+1}\|_2 > 0$, for all $i \le \gamma$, so

$$\|E_{\phi_i\cdot}^{k+1}\|_2 > 0, \text{ for all } i \le \gamma \text{ and } \|E^{k+1}\|_{2,0} = \gamma. \tag{27}$$

Next, we will demonstrate that there is no feasible solution to (23) with a better objective value than that of E^{k+1}. Let us suppose, by contradiction, that there exists a solution $X \in \mathbb{R}^{n\times r}$ with $\|X\|_{2,0} \le \gamma$, such that

$$\begin{aligned} &\|X\|_{2,1} + \frac{\rho}{2}\|X - Y\|_F^2 < \|E^{k+1}\|_{2,1} + \frac{\rho}{2}\|E^{k+1} - Y\|_F^2 \\ &\iff \sum_{i\in N}\|X_{i\cdot}\|_2 + \frac{\rho}{2}\|X_{i\cdot} - Y_{i\cdot}\|_2^2 < \sum_{i\in N}\|E_{i\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2. \end{aligned} \tag{28}$$

In this case, there exists $\hat{\jmath} \in N$, such that

$$\|X_{\hat{\jmath}\cdot}\|_2 + \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 < \|E_{\hat{\jmath}\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{\hat{\jmath}\cdot}^{k+1} - Y_{\hat{\jmath}\cdot}\|_2^2. \tag{29}$$

Because $\widetilde{E}^{k+1}$ is optimal and X is feasible for (24), we must have

$$\|X_{\hat{\jmath}\cdot}\|_2 + \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 \geq \|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1} - Y_{\hat{\jmath}\cdot}\|_2^2 \,, \tag{30}$$

otherwise we would obtain a better solution than $\widetilde{E}^{k+1}$ for (24), by replacing its $\widehat{\jmath}$-th row with $X_{\hat{\jmath}\cdot}$.

From (29) and (30), we can see that $E_{\hat{\jmath}\cdot}^{k+1} \neq \widetilde{E}_{\hat{\jmath}\cdot}^{k+1}$. Thus, from (26), we have $E_{\hat{\jmath}\cdot}^{k+1} = 0$, and (29) reduces to

$$\|X_{\hat{\jmath}\cdot}\|_2 + \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 < \frac{\rho}{2}\|Y_{\hat{\jmath}\cdot}\|_2^2 \,.$$

Note that we cannot have $X_{\hat{\jmath}\cdot} = 0$ in the inequality above, hence we have $\|X_{\hat{\jmath}\cdot}\|_2 > 0$ and $\|E_{\hat{\jmath}\cdot}^{k+1}\|_2 = 0$.

Now, we recall that $\|E^{k+1}\|_{2,0} = \gamma$ and $\|X\|_{2,0} \leq \gamma$, therefore for each $\widehat{\jmath}$ that satisfies (29), there must exist a distinct $\widehat{\ell} \in N \setminus \{\widehat{\jmath}\}$ such that $\|X_{\hat{\ell}\cdot}\|_2 = 0$ and $\|E_{\hat{\ell}\cdot}^{k+1}\|_2 > 0$. Moreover, because $\|E_{\hat{\ell}\cdot}^{k+1}\|_2 > 0$ and $\|E_{\hat{\jmath}\cdot}^{k+1}\|_2 = 0$, we see from (26) and (27) that $\widehat{\ell} \in \{\phi_1, \phi_2, \ldots, \phi_\gamma\}$ and $\widehat{\jmath} \in \{\phi_{\gamma+1}, \phi_{\gamma+2}, \ldots, \phi_n\}$, so $g(\widehat{\ell}) \geq g(\widehat{\jmath})$, that is

$$\frac{\rho}{2}\|Y_{\hat{\ell}\cdot}\|_2^2 - \|\widetilde{E}_{\hat{\ell}\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|\widetilde{E}_{\hat{\ell}\cdot}^{k+1} - Y_{\hat{\ell}\cdot}\|_2^2 \geq \frac{\rho}{2}\|Y_{\hat{\jmath}\cdot}\|_2^2 - \|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1} - Y_{\hat{\jmath}\cdot}\|_2^2 \,. \tag{31}$$

Also, from (30), we can see that

$$\frac{\rho}{2}\|Y_{\hat{\jmath}\cdot}\|_2^2 - \|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|\widetilde{E}_{\hat{\jmath}\cdot}^{k+1} - Y_{\hat{\jmath}\cdot}\|_2^2 \geq \frac{\rho}{2}\|Y_{\hat{\jmath}\cdot}\|_2^2 - \|X_{\hat{\jmath}\cdot}\|_2 - \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 \,. \tag{32}$$

From (31) and (32), we have that

$$\frac{\rho}{2}\|Y_{\hat{\ell}\cdot}\|_2^2 - \|\widetilde{E}_{\hat{\ell}\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|\widetilde{E}_{\hat{\ell}\cdot}^{k+1} - Y_{\hat{\ell}\cdot}\|_2^2 \geq \frac{\rho}{2}\|Y_{\hat{\jmath}\cdot}\|_2^2 - \|X_{\hat{\jmath}\cdot}\|_2 - \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 \,.$$

We recall that $X_{\hat{\ell}\cdot} = 0$ and $E_{\hat{\jmath}\cdot}^{k+1} = 0$. Moreover, $\|E_{\hat{\ell}\cdot}^{k+1}\|_2 > 0$, thus $E_{\hat{\ell}\cdot}^{k+1} = \widetilde{E}_{\hat{\ell}\cdot}^{k+1}$. Then, the last inequality is equivalent to

$$\begin{aligned}
&\|X_{\hat{\ell}\cdot}\|_2 + \frac{\rho}{2}\|X_{\hat{\ell}\cdot} - Y_{\hat{\ell}\cdot}\|_2^2 - \|E_{\hat{\ell}\cdot}^{k+1}\|_2 - \frac{\rho}{2}\|E_{\hat{\ell}\cdot}^{k+1} - Y_{\hat{\ell}\cdot}\|_2^2 \\
&\quad \geq \|E_{\hat{\jmath}\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{\hat{\jmath}\cdot}^{k+1} - Y_{\hat{\jmath}\cdot}\|_2^2 - \|X_{\hat{\jmath}\cdot}\|_2 - \frac{\rho}{2}\|X_{\hat{\jmath}\cdot} - Y_{\hat{\jmath}\cdot}\|_2^2 \,,
\end{aligned}$$

which we rewrite as

$$\sum_{i \in \{\hat{\jmath}, \hat{\ell}\}} \|E_{i\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2 - \|X_{i\cdot}\|_2 - \frac{\rho}{2}\|X_{i\cdot} - Y_{i\cdot}\|_2^2 \leq 0. \tag{33}$$

Taking into account the above, we see that for each $\widehat{\jmath}$ that satisfies (29), there is a distinct $\widehat{\ell} := \widehat{\ell}(\widehat{\jmath})$, such that (33) holds for $(\widehat{\jmath}, \widehat{\ell}(\widehat{\jmath}))$. Let $\widehat{N}$ be the set of all indices $\widehat{\jmath} \in N$ that satisfy (29) and the corresponding indices $\widehat{\ell}(\widehat{\jmath})$. Then we have

$$\begin{aligned}
&\sum_{i \in N} \|E_{i\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2 - \sum_{i \in N} \left(\|X_{i\cdot}\|_2 + \frac{\rho}{2}\|X_{i\cdot} - Y_{i\cdot}\|_2^2\right) \\
&\quad = \sum_{i \in \hat{N}} \|E_{i\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2 - \|X_{i\cdot}\|_2 - \frac{\rho}{2}\|X_{i\cdot} - Y_{i\cdot}\|_2^2 \\
&\qquad + \sum_{i \in N \setminus \hat{N}} \|E_{i\cdot}^{k+1}\|_2 + \frac{\rho}{2}\|E_{i\cdot}^{k+1} - Y_{i\cdot}\|_2^2 - \|X_{i\cdot}\|_2 - \frac{\rho}{2}\|X_{i\cdot} - Y_{i\cdot}\|_2^2 \\
&\quad \leq 0 \,,
\end{aligned}$$

which contradicts (28), showing that there is no better solution than E^{k+1} to (23). ◀

▶ Remark 19. It is possible to verify that the result of Theorem 18 still holds if subproblem (23) is replaced by the more general problem

$$E^{k+1} := \operatorname*{argmin}_{E \in \mathcal{M}} \sum_{i=1}^{n} f(E_{i\cdot}) + \frac{\rho}{2}\|E - Y\|_F^2 \,,$$

where $f : \mathbb{R}^r \to \mathbb{R}$ is nonnegative with $f(\delta) = 0$ if and only if $\delta = 0$. For example, we could have the 2,1-norm in the objective function of subproblem (23) replaced by the 1-norm or the square of the Frobenius norm.

▶ **Corollary 20.** *Let τ be the permutation of the indices in $N := \{1,\ldots,n\}$, such that $\|Y_{\tau_1\cdot}\|_2 \geq \|Y_{\tau_2\cdot}\|_2 \geq \cdots \geq \|Y_{\tau_n\cdot}\|_2$. Then an optimal solution of (23) is given by*

$$E^{k+1}_{\tau_i\cdot} := \begin{cases} \frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2} Y_{\tau_i\cdot}\,, & \text{if } 1/\rho < \|Y_{\tau_i\cdot}\|_2 \text{ and } i \leq \gamma\,; \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

Proof. Let us consider $\widetilde{E}^{k+1}$ as the optimal solution of the unconstrained version of (23) defined in Proposition 6, that is

$$\widetilde{E}^{k+1}_{\tau_i\cdot} := \begin{cases} \frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2} Y_{\tau_i\cdot}\,, & \text{if } 1/\rho < \|Y_{\tau_i\cdot}\|_2\,; \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

Considering the notation and statement of Theorem 18, it suffices to prove that $g(\tau_1) \geq g(\tau_2) \geq \cdots \geq g(\tau_n)$. For all i, such that $1/\rho < \|Y_{\tau_i\cdot}\|_2$, we have

$$\begin{aligned} g(\tau_i) &= \frac{\rho}{2}\|Y_{\tau_i\cdot}\|_2^2 - \left\|\left(\frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2}\right) Y_{\tau_i\cdot}\right\|_2 - \frac{\rho}{2}\left\|\left(\frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2}\right) Y_{\tau_i\cdot} - Y_{\tau_i\cdot}\right\|_2^2 \\ &= \frac{\rho}{2}\|Y_{\tau_i\cdot}\|_2^2 - \left(\frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2}\right)\|Y_{\tau_i\cdot}\|_2 - \frac{\rho}{2}\left(1 - \frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2}\right)^2 \|Y_{\tau_i\cdot}\|_2^2 \\ &= \frac{\rho}{2}\|Y_{\tau_i\cdot}\|_2^2 - \|Y_{\tau_i\cdot}\|_2 + \frac{1}{\rho} - \frac{1}{2\rho} = \left(\sqrt{\frac{\rho}{2}}\|Y_{\tau_i\cdot}\|_2 - \sqrt{\frac{1}{2\rho}}\right)^2. \end{aligned}$$

We note that, because $1/\rho < \|Y_{\tau_i\cdot}\|_2$, we have $\sqrt{\frac{\rho}{2}}\|Y_{\tau_i\cdot}\|_2 - \sqrt{\frac{1}{2\rho}} > 0$. Then, for any pair of indices j,ℓ with $j<\ell$, $1/\rho < \|Y_{\tau_j\cdot}\|_2$, and $1/\rho < \|Y_{\tau_\ell\cdot}\|_2$, as we have $\|Y_{\tau_j\cdot}\|_2 \geq \|Y_{\tau_\ell\cdot}\|_2$, we also have $g(\tau_j) \geq g(\tau_\ell) > 0$. Moreover, for all i, such that $1/\rho \geq \|Y_{\tau_i\cdot}\|_2$, we have $g(\tau_i) = 0$, confirming that $g(\tau_1) \geq g(\tau_2) \geq \cdots \geq g(\tau_n)$. Finally, from the ordering $\|Y_{\tau_1\cdot}\|_2 \geq \|Y_{\tau_2\cdot}\|_2 \geq \cdots \geq \|Y_{\tau_n\cdot}\|_2$, it is clear that $\|\widetilde{E}^{k+1}_{\tau_1\cdot}\|_2 \geq \|\widetilde{E}^{k+1}_{\tau_2\cdot}\|_2 \geq \cdots \geq \|\widetilde{E}^{k+1}_{\tau_n\cdot}\|_2$, completing the proof. ◀

Initialization of the variables. We initialize the variables with the solution obtained by the ADMM for 2,1-norm minimization described in Section 2.2.

Stopping criteria. We adopt the stopping criterion described in Section 2.2, additionally requiring that $\|H\|_{2,0} \leq \gamma$.

Pseudocode. In Algorithm 4, we present the ADMM algorithm for (20) .

Input: $A \in \mathbb{R}^{m\times n}$, $\Lambda^0 \in \mathbb{R}^{n\times r}$, $E^0 \in \mathbb{R}^{n\times r}$, $\rho > 0$, $0 < \omega < 1$, $r := \operatorname{rank}(A)$, $\|H^{2,1}_{opt}\|_{2,0}$.
Output: $H \in \mathbb{R}^{n\times m}$.
$U, \Sigma, V :=$ `svd`(A), $k := 0$;
Get U_1, V_1, V_2, D^{-1} from U, Σ, V;
$\mathcal{M} := \{X \in \mathbb{R}^{n\times r} \;:\; \|X\|_{2,0} \leq \gamma := \omega r + (1-\omega)\|H^{2,1}_{opt}\|_{2,0}\}$;
while *not converged* **do**
  $J := E^k - V_1 D^{-1} - \Lambda^k$;
  $Z^{k+1} := V_2^{\intercal} J$;
  $Y := V_1 D^{-1} + V_2 Z^{k+1} + \Lambda^k$;
  $\tau :=$ permutation of indices in $\{1,\ldots,n\}$, such that $\|Y_{\tau_1\cdot}\|_2 \geq \|Y_{\tau_2\cdot}\|_2 \geq \cdots \geq \|Y_{\tau_n\cdot}\|_2$;
  for $i = 1,\ldots,n$ **do**
    if $\|Y_{\tau_i\cdot}\|_2 > 1/\rho$ & $i \leq \gamma$ **then** $E^{k+1}_{\tau_i\cdot} := \frac{\|Y_{\tau_i\cdot}\|_2 - 1/\rho}{\|Y_{\tau_i\cdot}\|_2} Y_{\tau_i\cdot}$**; (see Cor. 20)**
    else $E^{k+1}_{\tau_i\cdot} := 0$**;**
  $\Lambda^{k+1} := \Lambda^k + V_1 D^{-1} + V_2 Z^{k+1} - E^{k+1}$;
  $k := k+1$;
$H := V_1 D^{-1} U_1^{\intercal} + V_2 Z^k U_1^{\intercal}$;

Algorithm 4: ADMM for (20) ($\text{ADMM}_{2,1/0}$)

5 Numerical experiments

We constructed test instances of varying sizes for our numerical experiments using the MATLAB function `sprand` to randomly generate $m \times n$ matrices A with rank r, as described in [10, Section 2.1]. The instances considered in [10] were too small for our experiments, but we considered the results of this previous work by selecting the formulations for our norm minimization problems as those in which `Gurobi` and `MOSEK` performed best, namely $(\mathcal{P}^1_{123})$ and $(\mathcal{P}^{2,1}_1)$. We divided our instances into two categories related to m with $n := 0.5m$, $r := 0.25m$; small instances with $m := 100, 200, \ldots, 500$ (S1, S2,..., S5) and large instances with $m := 1000, 2000, \ldots, 5000$ (L1, L2,..., L5).

We ran our experiments on 'zebratoo', a 32-core machine (running Windows Server 2022 Standard): two Intel Xeon Gold 6444Y processors running at 3.60GHz, with 16 cores each, and 128 GB of memory. We consider 10^{-5} as the tolerance to distinguish nonzero elements. The symbol '*' in column 'Time' of our tables, indicates that the problem was not solved to optimality because the time limit was reached, and the symbol '☠' indicates that we ran out of memory. We set a time limit of 2 hours for solving each instance. We coded our algorithms in `Julia` v.1.10.0.

In our first numerical experiment, we compare the solutions of ADMM_1 for $(\mathcal{P}^1_{123})$ with the solutions of `Gurobi`, and the solutions of $\text{ADMM}_{2,1}$ for $(\mathcal{P}^{2,1}_1)$ with the solutions of `MOSEK`. We used `Gurobi` v.11.0.0 to solve $(\mathcal{P}^1_{123})$ as a linear-optimization problem:

$$\min_{\substack{F\in\mathbb{R}^{n\times m},\\ Z\in\mathbb{R}^{(n-r)\times r}}} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} F_{ij} \,:\, F - V_2 Z U_1^{\mathsf{T}} \geq V_1 D^{-1} U_1^{\mathsf{T}}\,,\ F + V_2 Z U_1^{\mathsf{T}} \geq -V_1 D^{-1} U_1^{\mathsf{T}} \right\},$$

and we used `MOSEK` v.10.1.21 to solve $(\mathcal{P}^{2,1}_1)$ a second-order-cone optimization problem:

$$\min_{\substack{t\in\mathbb{R}^{n},\\ Z\in\mathbb{R}^{(n-r)\times r}}} \left\{ \sum_{i=1}^{n} t_i \ :\ \left\| \mathbf{e}_i^{\mathsf{T}} V \begin{bmatrix} D^{-1} \\ Z \end{bmatrix} \right\|_2 \leq t_i\,, \quad i = 1,\ldots,n \right\}.$$

Our first goal is to compare solutions and running times of the ADMM algorithms to state-of-the-art general-purpose convex-optimization solvers. For this comparison, in addition to running the ADMM algorithms with the stopping criteria described in Section 2.1 and Section 2.2, where the residual tolerances depend not only on the parameters ϵ^{abs} and ϵ^{rel}, but also on the values of the variables that are dynamically updated throughout the iterations (see (6–7), (14–15)), we also run the ADMM algorithms with a fixed tolerance ϵ for the Frobenius norms of the primal and dual residuals. With this fixed tolerance, we have a more fair comparison with the solvers that also work with a fixed tolerance. In the following, we refer to this second version of ADMM_1 and $\text{ADMM}_{2,1}$ with a fixed tolerance ϵ, respectively, by ADMM^{ϵ}_1 and $\text{ADMM}^{\epsilon}_{2,1}$.

Our second goal is to compare the solutions and running times of ADMM_1 (and ADMM^{ϵ}_1) to $\text{ADMM}_{2,1}$ (and $\text{ADMM}^{\epsilon}_{2,1}$). With these comparisons, we verify whether the different norms used to induce sparsity and row sparsity are effective and how much we actually lose in sparsity and gain in row sparsity when applying $\text{ADMM}_{2,1}$ compared to ADMM_1.

In Table 1, we show from the first to the last column, the instance, the method adopted, the different norms of the solutions and the (elapsed) running time (in seconds). In the second column, the labels identifying the methods are: `Gurobi`, ADMM_1, and ADMM^{ϵ}_1 applied to solve $(\mathcal{P}^1_{123})$; `MOSEK`, $\text{ADMM}_{2,1}$, and $\text{ADMM}^{\epsilon}_{2,1}$ applied to solve $(\mathcal{P}^{2,1}_1)$.

The following parameters were used.

- `Gurobi`: optimality and feasibility tolerances of 10^{-4};
- ADMM_1 : $\epsilon^{\text{abs}} = \epsilon^{\text{rel}} := 10^{-4}$, $\rho := 3$;
- ADMM^{ϵ}_1 : $\epsilon := 10^{-4}$, $\rho := 3$;
- `MOSEK`: optimality and feasibility tolerances of 10^{-5};
- $\text{ADMM}_{2,1}$: $\epsilon^{\text{abs}} = \epsilon^{\text{rel}} := 10^{-7}$, $\rho := 1$.
- $\text{ADMM}^{\epsilon}_{2,1}$: $\epsilon := 10^{-5}$, $\rho := 1$;

We note that we initially tried to select the values of ϵ^{abs}, ϵ^{rel} and ϵ, all equal to 10^{-4}. However, when we ran our experiments with `MOSEK` using tolerance 10^{-4}, we observed solutions with much higher 2,0-norms than the solutions of $\text{ADMM}^{\epsilon}_{2,1}$ (probably because `MOSEK` is using interior-point methods). To get a fair comparison between `MOSEK` and $\text{ADMM}^{\epsilon}_{2,1}$ with comparable solutions in terms of row-sparsity, we selected the smaller tolerance of 10^{-5} for both. Finally, the small tolerance of 10^{-7} was selected for $\text{ADMM}_{2,1}$ to highlight the fast convergence of this algorithm even for small tolerances, which was not the case for ADMM_1 .

Table 1 Comparison between ADMM and solvers

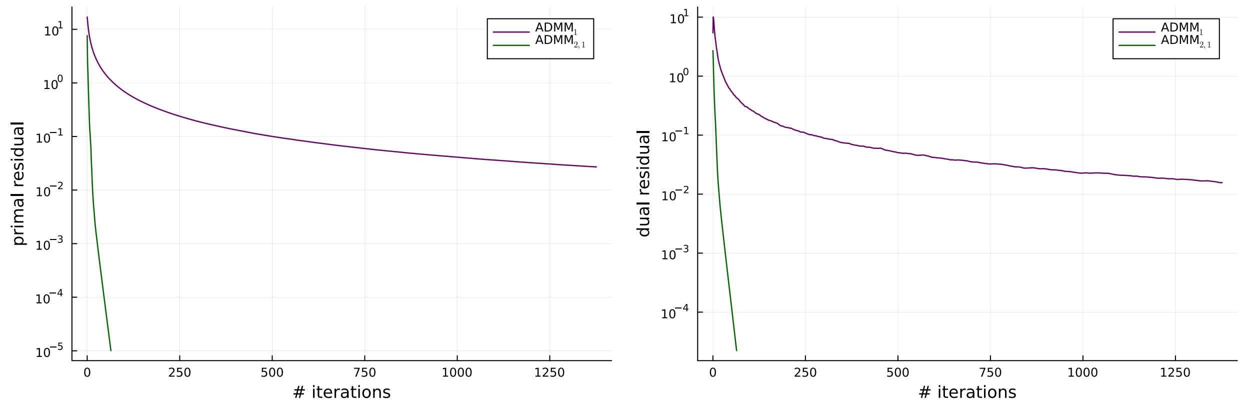
Inst.	Method	$\|H\|_1$	$\|H\|_0$	$\|H\|_{2,1}$	$\|H\|_{2,0}$	Time (sec)
S1	`Gurobi`	194.28	3734	39.14	44	6.06
	ADMM_1^ϵ	194.28	3733	39.14	44	1.67
	ADMM_1	194.29	4152	39.13	50	0.32
	`MOSEK`	211.60	3824	36.10	42	1.43
	$\text{ADMM}_{2,1}^\epsilon$	211.60	3819	36.10	39	0.08
	$\text{ADMM}_{2,1}$	211.60	3819	36.10	39	0.08
S2	`Gurobi`	539.70	14943	79.91	90	119.21
	ADMM_1^ϵ	539.71	14927	79.91	90	71.58
	ADMM_1	539.79	16804	79.89	99	3.38
	`MOSEK`	595.24	16222	73.41	84	0.38
	$\text{ADMM}_{2,1}^\epsilon$	595.24	16219	73.41	83	0.01
	$\text{ADMM}_{2,1}$	595.24	16219	73.41	83	0.02
S3	`Gurobi`	868.19	33756	119.40	133	2295.54
	ADMM_1^ϵ	868.19	33702	119.40	134	309.21
	ADMM_1	868.42	38962	119.43	150	7.48
	`MOSEK`	972.46	34723	109.98	118	29.96
	$\text{ADMM}_{2,1}^\epsilon$	972.47	34729	109.98	117	0.24
	$\text{ADMM}_{2,1}$	972.47	34729	109.98	117	0.33
S4	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	1339.90	64174	163.73	194	870.73
	ADMM_1	1340.32	73519	163.74	200	11.81
	`MOSEK`	1504.61	67338	150.07	193	2.33
	$\text{ADMM}_{2,1}^\epsilon$	1504.64	66956	150.07	171	0.26
	$\text{ADMM}_{2,1}$	1504.64	66956	150.07	171	0.33
S5	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	1870.64	100758	208.31	245	1030.52
	ADMM_1	1871.28	115194	208.31	250	16.20
	`MOSEK`	2102.71	106021	190.28	237	5.48
	$\text{ADMM}_{2,1}^\epsilon$	2102.77	105312	190.28	217	0.31
	$\text{ADMM}_{2,1}$	2102.77	105312	190.28	217	0.48
L1	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	4886.76	409821	425.60	492	3178.79
	ADMM_1	4890.11	471964	425.67	500	47.44
	`MOSEK`	5637.29	437653	384.94	480	40.90
	$\text{ADMM}_{2,1}^\epsilon$	5637.67	435820	384.94	444	1.03
	$\text{ADMM}_{2,1}$	5637.67	435820	384.94	444	0.58
L2	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	11781.79	1622951	853.10	987	*
	ADMM_1	11799.06	1901119	853.46	1000	141.72
	`MOSEK`	14098.24	1757540	765.78	978	443.30
	$\text{ADMM}_{2,1}^\epsilon$	14099.58	1746738	765.78	891	2.14
	$\text{ADMM}_{2,1}$	14099.58	1746741	765.78	891	2.26
L3	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	20459.62	3672370	1306.86	1488	*
	ADMM_1	20495.70	4286145	1307.31	1500	292.84
	`MOSEK`	-	-	-	-	☠
	$\text{ADMM}_{2,1}^\epsilon$	24904.76	4014677	1160.70	1367	4.25
	$\text{ADMM}_{2,1}$	24904.76	4014676	1160.70	1367	4.05
L4	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	30029.31	6474511	1738.97	1995	*
	ADMM_1	30089.82	7595298	1739.63	2000	475.33
	`MOSEK`	-	-	-	-	☠
	$\text{ADMM}_{2,1}^\epsilon$	36984.26	7075074	1547.08	1813	7.65
	$\text{ADMM}_{2,1}$	36984.26	7075080	1547.08	1813	7.23
L5	`Gurobi`	-	-	-	-	*
	ADMM_1^ϵ	40716.32	10285553	2191.30	2498	*
	ADMM_1	40805.04	11876100	2192.73	2498	656.24
	`MOSEK`	-	-	-	-	☠
	$\text{ADMM}_{2,1}^\epsilon$	50516.23	11223830	1947.08	2297	14.40
	$\text{ADMM}_{2,1}$	50516.22	11223836	1947.08	2297	12.91

From the results in Table 1, we see that with the application of `Gurobi` to our linear-optimization model for $(\mathcal{P}^1_{123})$, we can only solve the three smallest instances. ADMM_1^ϵ converges for all small instances and for one large instance as well. When using the looser dynamic stopping criterion, we have convergence of ADMM_1 for all instances tested. ADMM_1^ϵ and `Gurobi` converge to very similar solutions when both converge, but ADMM_1^ϵ converges much faster, showing the advantage of the application of the ADMM algorithm over our application of

Gurobi to construct 1-norm minimizing ah-symmetric reflexive generalized inverses. ADMM_1 , with the dynamic stopping criterion, converges much faster than ADMM_1^ϵ for all tested instances and, comparing the solutions of both algorithms, we see only an increase of at most 0.1% in the 1-norm of the solutions obtained by ADMM_1 , although the 0-norms increase by about 10%.

Analyzing now the statistics for $\mathrm{ADMM}_{2,1}^\epsilon$ and $\mathrm{ADMM}_{2,1}$, we see that both algorithms converge for all instances tested to solutions with the same 2,1-norms, and both take less than 15 seconds on the largest instance. The ADMM for the 2,1-norm is robust and quickly converges to high-precision solutions. MOSEK did not solve the three largest instances due to lack of memory, and for the other instances, MOSEK took much longer than the ADMM algorithms and obtained solutions with the same 2,1-norms. The results in Table 1 show high superiority of the ADMM algorithm proposed to construct 2,1-norm minimizing ah-symmetric reflexive generalized inverses, when compared to our application of MOSEK. We note from the 2,0-norms of the solutions, that the minimization of the 2,1-norm is in fact a more effective strategy to induce row sparsity than the minimization of the 1-norm. Moreover, we always obtain fewer nonzero rows with $\mathrm{ADMM}_{2,1}$ than with MOSEK.

Finally, we can observe from the statistics in Table 1, that $\mathrm{ADMM}_{2,1}$ converges much faster than ADMM_1 . In Figure 1, we illustrate the convergence of these algorithms, showing that the number of iterations for $\mathrm{ADMM}_{2,1}$ is also much lower. Although Figure 1 corresponds to instance L1 only, it illustrates the typical behavior in our experiments.



(a) (b)

Figure 1 Convergence: ADMM_1 vs. $\mathrm{ADMM}_{2,1}$; $m, n, r = 1000, 500, 250$ (Instance L1)

We also note from Table 1 that, as expected, ADMM_1^ϵ obtains the sparsest solutions among the four ADMM algorithms tested, but at a high computational cost. Therefore, if our goal is to compute solutions quickly, we should note that although $\mathrm{ADMM}_{2,1}$ aims at row sparsity, it is actually more effective than ADMM_1 at obtaining sparse solutions.

Finally, we note that our goal when setting the parameters of the solvers was to achieve the best possible performance. In the case of Gurobi, this was achieved by letting Gurobi choose the algorithm. We also experimented with selecting specific algorithms, but Gurobi did not perform well. From the Gurobi log for an instance, we can see that it starts with a barrier method and then uses a dual simplex method, but this can change depending on the difficulty of the instance for Gurobi.

For MOSEK, we can see that it recognizes each instance as a conic optimization problem, for which MOSEK employs a primal-dual interior point algorithm. From MOSEK's log for some instances, we can see that it performs many fewer iterations than $\mathrm{ADMM}_{2,1}$ and $\mathrm{ADMM}_{2,1}^\epsilon$, but each iteration of MOSEK takes much longer than the iterations of the ADMM algorithm. We can then conclude that, for the instances considered, solving the linear system of equations to obtain the Newton direction in the primal-dual interior point method iterations is much more computationally expensive than updating the variables in $\mathrm{ADMM}_{2,1}$ with the closed-form formulas presented in Algorithm 2.

In our second numerical experiment, our goal is to verify the efficiency of the ADMM algorithms described in Section 4 ($\mathrm{ADMM}_{2,0}$ and $\mathrm{ADMM}_{2,1/0}$), in computing ah-symmetric reflexive generalized inverses within the target 2,0-norm given by $\omega r + (1-\omega)\|H_{opt}^{2,1}\|_{2,0}$, where $H_{opt}^{2,1}$ is the solution obtained by $\mathrm{ADMM}_{2,1}$. We also aim at observing the trade-off between the norms of the ah-symmetric reflexive generalized inverses obtained by these algorithms while varying ω in the interval $(0, 1)$. We note that for $\omega = 0$, the target 2,0-norm is the 2,0-norm of the 2,1-norm minimizing generalized inverse $H^{2,1}_{opt}$, while for $\omega = 1$, the target 2,0-norm is r, the smallest possible 2,0-norm of an ah-symmetric reflexive generalized inverse. Then, increasing ω in the interval $(0, 1)$, we intend to construct ah-symmetric reflexive generalized inverses with decreasing 2,0-norms, and we would like to see the impact on the 1- and 2,1-norms of the matrices. We are particularly interested in verifying if the 1- and 2,1-norms of the solutions obtained by $\text{ADMM}_{2,0}$ are too large, and how effective $\text{ADMM}_{2,1/0}$ is in decreasing the 2,1-norm with respect to $\text{ADMM}_{2,0}$. We recall that the problems addressed by both algorithms are nonconvex, therefore there is no guarantee of convergence of these ADMM algorithms. The first is a feasibility problem and seeks any solution within the target 2,0-norm, while the second seeks solutions within the target 2,0-norm and minimum 2,1-norm. We run both ADMM algorithms for the large instances, for $\omega = 0.25, 0.50, 0.75, 0.80, 0.90, 0.95$, and compare the solutions obtained with $H^{2,1}_{opt}$ and with the local-search solutions described in Section 3. The first one, denoted in the following by LS, uses the absolute determinant as a criterion to improve a given solution. The linearly-independent rows and columns of A used to construct its initial solution are obtained from the QR factorization of A. The second one, denoted in the following by $\text{LS}_{2,1}$, starts with the solution of LS, and try to improve it using the 2,1-norm as the criterion for improvement. In this case, the result in Proposition 16 is used in the implementation. We recall that both local-search procedures construct ah-symmetric reflexive generalized inverses with 2,0-norms equal to r.

Table 2 $\text{ADMM}_{2,0}$ / $\text{ADMM}_{2,1/0}$ for target 2,0-norm, for various ω

Inst.	Method	$\Vert H\Vert_1$		$\Vert H\Vert_0$		$\Vert H\Vert_{2,1}$		$\Vert H\Vert_{2,0}$	Time (sec)	
L1	$\text{ADMM}_{2,1}$	5637.7		435820		384.9		444	0.6	
	$\omega = 0.25$	6044.9	5796.4	388481	388547	393.5	387.1	395	0.8	26.9
	$\omega = 0.50$	6836.2	6325.5	341730	341640	411.6	398.7	347	1.4	20.2
	$\omega = 0.75$	8956.9	7686.8	293605	293557	483.2	441.9	298	3.2	26.3
	$\omega = 0.80$	9759.6	8691.6	283825	283825	512.1	474.1	288	3.8	36.9
	$\omega = 0.90$	12467.8	11164	265108	265133	613.0	566.2	269	8.1	42.6
	$\omega = 0.95$	15223.1	13143.9	255298	255287	729.5	643.8	259	19.5	33.1
	LS	10797.9		246392		589.9		250	4.4	
	$\text{LS}_{2,1}$	10452.9		246373		570.5		250	9.4	
L2	$\text{ADMM}_{2,1}$	14099.6		1746741		765.8		891	2.3	
	$\omega = 0.25$	15084.8	14576.0	1557984	1558091	781.5	769.2	793	2.8	22.0
	$\omega = 0.50$	17169.9	16068.1	1367070	1366850	807.8	790.0	695	5.4	85.6
	$\omega = 0.75$	23083.7	20739.7	1173280	1175007	932.4	883.6	597	11.6	93.1
	$\omega = 0.80$	26747.9	22149.5	1133992	1137706	1020.3	915.0	578	12.6	83.7
	$\omega = 0.90$	34401.6	30997.8	1061345	1061284	1231.7	1138.1	539	81.8	155.5
	$\omega = 0.95$	51969.9	40647.4	1022152	1022053	1734.5	1412.2	519	273.3	272.9
	LS	32796.0		984545		1272.5		500	15.2	
	$\text{LS}_{2,1}$	31583.0		984526		1232.4		500	66.2	
L3	$\text{ADMM}_{2,1}$	24904.8		4014676		1160.7		1367	4.0	
	$\omega = 0.25$	26922.6	25896.1	3568117	3567608	1184.4	1166.2	1212	6.9	114.5
	$\omega = 0.50$	32515.0	29371.3	3116491	3115482	1249.9	1206.6	1058	15.1	121.1
	$\omega = 0.75$	45489.9	40631.5	2660981	2663715	1482.5	1395.7	904	28.1	183.0
	$\omega = 0.80$	51695.7	45042.0	2568254	2572501	1609.5	1485.9	873	34.8	287.2
	$\omega = 0.90$	70276.9	61176.6	2387213	2389950	2027.8	1826.9	811	69.1	283.6
	$\omega = 0.95$	94905.5	83185.4	2298796	2298668	2580.9	2322.2	780	376.9	589.0
	LS	69338.6		2209609		2131.2		750	26.8	
	$\text{LS}_{2,1}$	62062.7		2208832		1974.1		750	380.4	
L4	$\text{ADMM}_{2,1}$	36984.3		7075080		1547.1		1813	7.2	
	$\omega = 0.25$	40413.6	38780.2	6302160	6301531	1582.1	1556.5	1609	9.2	147.5
	$\omega = 0.50$	49000.8	44611.0	5519866	5517665	1665.2	1612.3	1406	19.6	225.8
	$\omega = 0.75$	69070.6	60934.9	4722393	4727485	1967.3	1839.5	1203	35.5	437.8
	$\omega = 0.80$	77766.4	68502.2	4559968	4567287	2117.3	1964.7	1162	42.6	423.8
	$\omega = 0.90$	112003.5	93549.8	4247329	4250758	2747.0	2417.1	1081	103.0	727.3
	$\omega = 0.95$	154855.3	129358.9	4086725	4090408	3592.4	3125.0	1040	183.5	864.6
	LS	124026.2		3932651		3162.3		1000	93.9	
	$\text{LS}_{2,1}$	111118.3		3932550		2902.5		1000	1335.4	
L5	$\text{ADMM}_{2,1}$	50516.2		11223836		1947.1		2297	12.9	
	$\omega = 0.25$	55039.3	53092.0	9971973	9981248	1987.4	1958.6	2035	10.4	126.2
	$\omega = 0.50$	66928.5	61609.2	8713776	8715794	2091.4	2033.1	1773	24.3	681.0
	$\omega = 0.75$	95752.6	84276.1	7426600	7439856	2478.9	2329.6	1511	48.7	456.6
	$\omega = 0.80$	106328.0	93241.4	7172048	7185451	2636.9	2461.0	1459	61.0	668.3
	$\omega = 0.90$	161475.1	131572.0	6661864	6672091	3563.9	3073.8	1354	136.8	935.5
	$\omega = 0.95$	222013.5	185469.1	6413109	6412843	4653.3	4009.2	1302	332.1	1214.1
	LS	195014.1		6160665		4401.7		1250	192.4	
	$\text{LS}_{2,1}$	168359.4		6159949		3893.4		1250	5256.7	

In Table 2, we have the same information presented in Table 1. In the second column, we identify the methods addressed. For each value of ω, we have in each column, first the information for $\text{ADMM}_{2,0}$, and then for $\text{ADMM}_{2,1/0}$. Both algorithms always obtain solutions with the target 2,0-norm, and therefore, there is only one result for $\Vert H\Vert_{2,0}$ for each ω. Next, we present some important details about these algorithms.

- For $\text{ADMM}_{2,0}$ and $\text{ADMM}_{2,1/0}$, we consider $H_{i\cdot}$ to be a row of all zeros if $\Vert H_{i\cdot}\Vert_2 \leq 10^{-5}$.

- The parameter ρ is not used in the calculations of $\mathrm{ADMM}_{2,0}$, so any $\rho > 0$ could be used to define the augmented Lagrangian function considered in Section 4.1.
- For $\mathrm{ADMM}_{2,1/0}$, we consider $\rho := 10^4$ until we obtain $\|H\|_{2,0} \leq \gamma$ and the Frobenius norm of the primal residual is less than 10^{-4}. Then we update ρ at each iteration if the Frobenius norm of the primal residual remains less than 10^{-4}, by setting $\rho := \rho/\alpha$, where α ranges from 2 to 1, starting at the largest value. The convergence of ADMM for a nonconvex problem is not guaranteed; and the selection of ρ described above was important to obtain convergence of $\mathrm{ADMM}_{2,1/0}$ for all tested instances (see, for example, [2, Section 3.4.1], for a discussion of the update of ρ in the ADMM algorithm). The large value of ρ used to start the algorithm is related to the observed difficulty in satisfying the nonconvex constraint. Large values for ρ leads to a fast convergence of the primal residual to zero, so we were able to satisfy $\|H\|_{2,0} \leq \gamma$. However, with large ρ we have very slow convergence of the Frobenius norm of the dual residual to zero. Therefore, when we have $\|H\|_{2,0} \leq \gamma$ with Frobenius norm of the primal residual less than 10^{-4}, we set $\rho := \rho/\alpha$, if the Frobenius norm of the primal residual remains less than 10^{-4}. Otherwise, we iteratively decrease α and then update ρ, until the Frobenius norm of the primal residual remains less than 10^{-4} with the current ρ or $\alpha = 1$.
- For $\mathrm{ADMM}_{2,1/0}$, in addition to using the solution of $\mathrm{ADMM}_{2,1}$ to initialize the algorithm, we experimented with initializing it with the solution of $\mathrm{ADMM}_{2,0}$, however, in this case the algorithm failed to converge for some instances.
- For $\mathrm{ADMM}_{2,1/0}$, we set the parameters $\epsilon^{\mathrm{abs}} = \epsilon^{\mathrm{rel}} := 10^{-4}$.

From the results presented in Table 2, we can observe that $\mathrm{ADMM}_{2,0}$ computes ah-symmetric reflexive generalized inverses with the target 2,0-norms for all instances and all values of ω in less than 400 seconds. Increasing ω leads to slower convergence (in Figure 2, we illustrate this result, where we see a better behavior of the algorithm for $\omega \leq 0.80$). Despite disregarding the 1- and 2,1-norms in the formulation of problem (16), for

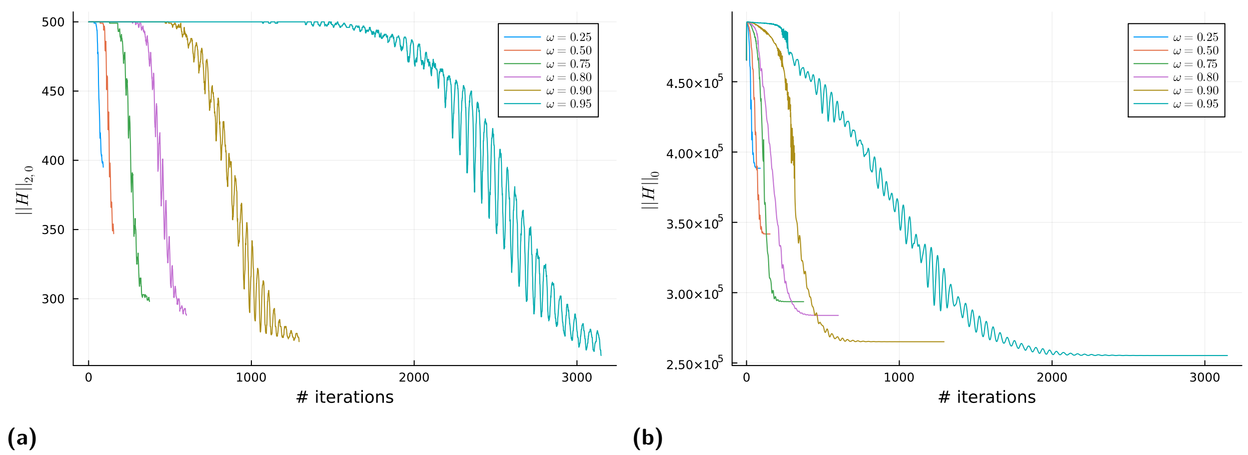


Figure 2 Norms per iteration: ADMM for target 2,0-norm $\omega r + (1-\omega)\|H^{2,1}_{opt}\|_{2,0}$; $m, n, r = 1000, 500, 250$ (Instance L1)

which $\mathrm{ADMM}_{2,0}$ is applied, the trade-off between the 2,0-norm and the 1- and 2,1-norms when ω increases is very clear from the results in Table 2. The 1- and 2,1-norms of the solutions are relatively small, probably because we start $\mathrm{ADMM}_{2,0}$ with a Frobenius-norm minimizing solution (see discussion in Section 2.1). This initialization was, in fact, very important in this case. When using different initializations we observed solutions with large 1- and 2,1- norms. For $\omega = 0.80$, for example, we see that $\mathrm{ADMM}_{2,0}$ is an excellent approach to construct fast structured ah-symmetric reflexive generalized inverses with small 2,1-norms for our test instances, when compared to the 2,1-norm of the local-search procedures, even $\mathrm{LS}_{2,1}$, that can actually obtain solutions with significant smaller 2,1-norm than LS, but at a high computational cost. We note that $\mathrm{ADMM}_{2,0}$, for $\omega = 0.80$, scales even better than the fastest local-search LS, converging in 61 seconds for the largest instance while LS takes 192.4 seconds to converge. The number of nonzero rows in the solutions obtained with $\omega = 0.80$ is always small when compared to $H^{2,1}_{opt}$, and the 1- and 2,1-norms are always small when compared to the local-search solution. Observing now the results for $\mathrm{ADMM}_{2,1/0}$, we see that it is effective in obtaining solutions with smaller 2,1-norm than $\mathrm{ADMM}_{2,0}$, but again, at a much greater computational cost. An important observation from our experiments is that, with our parameter settings, the ADMM algorithms proposed for the two nonconvex problems addressed in Section 4, converge for all tested instances.

Figure 3 is a visual representation of the "L5 block" of Table 2. We are comparing $\mathrm{ADMM}_{2,0}$ and $\mathrm{ADMM}_{2,1/0}$, varying $\omega \in \{0.25, 0.50, 0.75, 0.80, 0.90, 0.95\}$, to obtain a row-sparse solution, for the largest instance that we have considered. Additionally, we compare $\mathrm{ADMM}_{2,1}$ (at the left of the figure), and the local searches LS and $\mathrm{LS}_{2,1}$ (at the right of the figure). In blue, we show the 2,1-norms of the solutions, scaled by 1/2 for convenience. The solid blue line corresponds to $\mathrm{ADMM}_{2,0}$ and the dashed blue line corresponds to $\mathrm{ADMM}_{2,1/0}$. In red, we show the 2,0-norms of the solutions; they are *the same* for $\mathrm{ADMM}_{2,0}$ and $\mathrm{ADMM}_{2,1/0}$. Under the horizontal axis, the first row indicates the algorithm, with the numeric values corresponding to ω for $\mathrm{ADMM}_{2,0}$ and $\mathrm{ADMM}_{2,1/0}$. The second row indicates the running times for $\mathrm{ADMM}_{2,1}$, $\mathrm{ADMM}_{2,0}$ for each ω, LS and $\mathrm{LS}_{2,1}$. The third row indicates the running times for $\mathrm{ADMM}_{2,1/0}$.

There is no overall winner, and the best choice depends on considering the 2,0-norm, the 2,1-norm, and the elapsed time. For modest to moderate values of ω, say $\omega \in \{0.25, 0.50, 0.75, 0.80\}$, we can see a reasonable trade-off in 2,0-norm vs. 2,1-norm, and with a modest but increasing elapsed time. For the larger values of ω, we pay a large penalty in the 2,1-norm for further decrease in the 2,0-norm, and the elapsed time grows as well. Still, some user might prefer the solution of say $\mathrm{ADMM}_{2,1/0}$ at $\omega = 0.95$ to the local search LS, giving a lesser 2,1-norm and only slightly greater 2,0-norm, albeit with a much greater elapsed time. If one is willing to suffer a very long elapsed time, the local search $\mathrm{LS}_{2,1}$ dominates both the 2,1-norm and 2,0-norm of $\mathrm{ADMM}_{2,1/0}$ at $\omega = 0.95$.

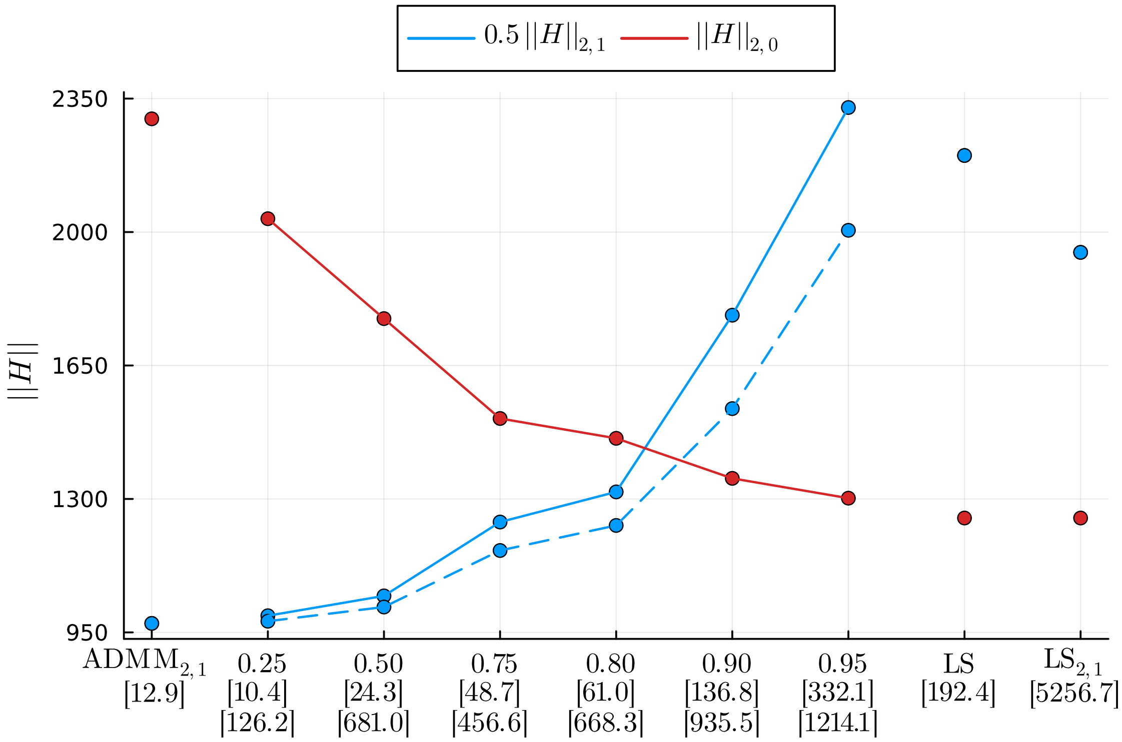


Figure 3 Comparison of our algorithms aimed at row-sparsity on a large instance

6 Conclusions

The results from the numerical experiments performed in this work show that the four ADMM algorithms developed are effective in obtaining sparse and row-sparse ah-symmetric reflexive generalized inverses. It is important to note that only two of the four ADMM algorithms address convex optimization problems, and there is no guarantee of convergence for the other two. Nevertheless, all four of the algorithms obtain good results and always converge. The code and data associated with our experiments can be obtained at `https://github.com/GabrielPonte/admmGinv/`.

If ensuring convergence to the global optimum is an important point, the algorithms ADMM_1 (aimed at inducing sparsity) and $\mathrm{ADMM}_{2,1}$ (aimed at inducing row-sparsity) are well suited, as they address convex optimization problems. Both algorithms converge faster than general-purpose optimization solvers. Comparing both algorithms, the convergence of $\text{ADMM}_{2,1}$ is much faster than ADMM_1 , and it converges in many fewer iterations. $\text{ADMM}_{2,1}$ is robust and quickly converges to high-precision solutions. As expected, the minimization of the 2,1-norm is more effective than the minimization of the 1-norm if we are aiming for low 2,0-norm (i.e., row-sparse solutions). But if we seek low 0-norm (i.e., sparse solutions), we naturally should instead minimize the 1-norm. ADMM_1^ϵ obtains the sparsest solutions, but we can expect a high computational cost to have convergence to such solutions. If we are aiming for sparse solutions computed quickly, then $\text{ADMM}_{2,1}$ and $\text{ADMM}_{2,1}^\epsilon$ are preferred to ADMM_1 .

If we do not require guaranteed convergence to a global optimum, the ADMM algorithms developed for the nonconvex problems addressed, where we limit the number of nonzero rows in the solution, are very effective in obtaining row-sparse solutions. Moreover, interesting solutions can be obtained with these algorithms by varying the number of nonzero rows allowed. Decreasing this number from the 2,0-norm of the 2,1-norm minimizing ah-symmetric reflexive generalized inverse of a given matrix A, until it approaches the minimum number of rows (given by $\text{rank}(A)$), we see an increase in the 1- and 2,1-norms of the solutions. With an appropriate number of nonzero rows allowed, $\text{ADMM}_{2,0}$ becomes an excellent approach to quickly construct structured ah-symmetric reflexive generalized inverses with small 2,1-norms. $\text{ADMM}_{2,0}$ scales even better than our fastest local-search. $\text{ADMM}_{2,1/0}$, is effective in obtaining solutions with smaller 2,1-norm than $\text{ADMM}_{2,0}$, but at a much greater computational cost. If one insists on a solution with minimum 2,0-norm, then the local searches are quite appropriate. In summary, there is no overall winner, and all of our algorithms have their use, as one trades off 2,0-norm, 2,1-norm and elapsed time.

We note that the ADMM algorithms for the 1- and 2,1-norms presented in Section 2.1 and Section 2.2 have their convergence guaranteed to an optimal solution by general convergence results for ADMM algorithms that can be found, for example, in [2, Section 3.2]. The nonconvexity of the feasibility problems addressed in Section 4 precludes the direct use of these results. There are convergence results for ADMM applied to nonconvex problems, but the ones that we are aware of (for example, [22]) do not apply to our situation. Nevertheless, because we were able to get practical convergence for our nonconvex ADMMs, exploring a theoretical reason for this looks to be a promising direction for future research.

ADMM is just one proximal-gradient algorithm, and there are other such algorithms, and relatives as well. We see our work as opening the door for examining whether any other algorithms of this type may further improve the state-of-the-art for the computation of row-sparse ah-symmetric reflexive generalized inverses.

Acknowledgment

The authors gratefully acknowledge Laura Balzano and Ahmad Mousavi for suggesting to us an ADMM approach to structured-sparse generalized inverses, but aimed at an objective minimizing a balance between the nuclear norm and 2,1-norm of H, subject to (P1) and (P3). By exploiting ideas in [20], we instead worked with $(\mathcal{P}_1^{2,1})$.

A Ranks one and two

Following the ideas of [23], we consider the construction of a 2,1-norm minimizing ah-symmetric reflexive generalized inverse based in Theorem 7, for the cases where $\text{rank}(A) \in \{1, 2\}$.

▶ **Theorem 21.** *Let A be an arbitrary $m \times n$, rank-1 matrix. Choose any row $s \in \{1, \ldots, m\}$ (except a zero row), then define $t := \text{argmax}_j\{|A_{sj}| : j = 1, \ldots, n\}$. Let $\widehat{a}$ be column t of A. Then the $n \times m$ matrix H constructed by Theorem 7 over $\widehat{a}$, is an ah-symmetric reflexive generalized inverse of A with minimum 2,1-norm.*

Proof. First, note that $\|\widehat{a}^\dagger\|_2 = \|(\widehat{a}^\intercal\widehat{a})^{-1}\widehat{a}^\intercal\|_2 = \left\|\frac{1}{\|\hat{a}\|_2^2}\widehat{a}^\intercal\right\|_2 = \frac{1}{\|\hat{a}\|_2}$, and construct W from Lemma 14 with $E = \|\widehat{a}^\dagger\|_2$. As we have a rank-1 matrix, W is an $m \times n$ matrix with all elements equal to zero except $W_{st} = \|\widehat{a}^\dagger\|_2 / A_{st}$. Then, $A^\intercal W$ is a matrix with all elements equal to zero, except for column t which is $(A^\intercal W)_{it} = \frac{A_{si}\|\hat{a}^\dagger\|_2}{A_{st}} = \frac{A_{si}}{A_{st}\|\hat{a}\|_2}$, then $(A^\intercal W A^\intercal)_{ij} = \frac{A_{si}A_{jt}}{A_{st}\|\hat{a}\|_2}$, and so $\|(A^\intercal W A^\intercal)_{i\cdot}\|_2 = \frac{|A_{si}|}{|A_{st}|\,\|\hat{a}\|_2}\|\widehat{a}\|_2 = \frac{|A_{si}|}{|A_{st}|}$, for all $i = 1, \ldots, n$, $j = 1, \ldots, m$.

For $i = t$ we have an active dual constraint, and for $i \neq t$ we have $\|(A^\intercal W A^\intercal)_{i\cdot}\|_2 \leq 1$. Then by weak duality, we have that the constructed H is optimal for $(P_1^{2,1})$. ◀

Generally, when $\text{rank}(A) = 2$, we cannot construct a 2,1-norm minimizing ah-symmetric reflexive generalized inverse based on the column-block construction. Even under the condition that A is totally unimodular and $m = r$, we have the example: $A := \left[\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right]$. We have an ah-symmetric reflexive generalized inverse with minimum

2,1-norm $\frac{\sqrt{2}}{2}\left(1+\sqrt{3}\right)$, $H := \frac{1}{6}\begin{bmatrix} 3+\sqrt{3} & \sqrt{3}-3 \\ 3-\sqrt{3} & 3-\sqrt{3} \\ \sqrt{3}-3 & 3+\sqrt{3}\end{bmatrix}$. However, the three ones based on our column block construction have 2,1-norm $1+\sqrt{2}$, 2, $1+\sqrt{2}$, respectively.

Next, we demonstrate that under an efficiently-checkable technical condition, when $\mathrm{rank}(A) = 2$, construction of a 2,1-norm minimizing ah-symmetric reflexive generalized inverse can be based on the column block construction. Let T be an ordered subset of r elements from $\{1,\ldots,n\}$ and $\widehat{A} := A[\,\cdot\,, T]$ be the $m\times r$ submatrix of an $m\times n$ matrix A formed by columns T, and $\mathrm{rank}(\widehat{A}) = r$. Let S be an ordered subset of r elements from $\{1,\ldots,m\}$, such that $\widetilde{A} := A[S,T]$ is a nonsingular matrix.

▶ **Lemma 22.** *Let $k\in\{1,2\}$ and $\sigma(k) := \{1,2\}\setminus\{k\}$. For rank 2, $\|\widehat{A}^\dagger_{k\cdot}\|_2^2 = \frac{\|\hat{A}_{\cdot\sigma(k)}\|_2^2}{\det(\hat{A}^\intercal\hat{A})}$.*

Proof. We have $\widehat{A}^\dagger_{i\cdot} = (\widehat{A}^\intercal\widehat{A})^{-1}_{i\cdot}\widehat{A}$, then $\|\widehat{A}^\dagger_{k\cdot}\|_2^2 = (\widehat{A}^\intercal\widehat{A})^{-1}_{k\cdot}\widehat{A}^\intercal\widehat{A}((\widehat{A}^\intercal\widehat{A})^{-1}_{k\cdot})^\intercal = \mathbf{e}_k^\intercal((\widehat{A}^\intercal\widehat{A})^{-1}_{k\cdot})^\intercal = (\widehat{A}^\intercal\widehat{A})^{-1}_{k\cdot}\mathbf{e}_k = (\widehat{A}^\intercal\widehat{A})^{-1}_{kk} = \frac{\|\hat{A}_{\cdot\sigma(k)}\|_2^2}{\det(\hat{A}^\intercal\hat{A})}$. ◀

▶ **Theorem 23.** *Let A be an arbitrary $m\times n$, rank-2 matrix. For any $j_1, j_2\in\{1,\ldots,n\}$, with $j_1 < j_2$, let $\widehat{A} := [\widehat{a}_{j_1}\ \widehat{a}_{j_2}]$ be the $m\times 2$ submatrix of A formed by columns j_1 and j_2. Suppose that j_1, j_2 are chosen to minimize the 2,1-norm of $\widehat{H} := \widehat{A}^\dagger$ among all $m\times 2$ rank-2 submatrices of A. Every column $\widehat{b}$ of A, can be uniquely written in the basis $\widehat{a}_{j_1}, \widehat{a}_{j_2}$, say $\widehat{b} = \beta_1\widehat{a}_{j_1} + \beta_2\widehat{a}_{j_2}$. Suppose that for each such column $\widehat{b}$ of A we have $|\beta_1| + |\beta_2| \le 1$. Then the $n\times m$ matrix H constructed by Theorem 7 based on $\widehat{A}$, is an ah-symmetric reflexive generalized inverse of A with minimum 2,1-norm.*

Proof. Let $\alpha_1 := \|\widehat{A}^\dagger_{1\cdot}\|_2/\|\widehat{A}_{\cdot 2}\|_2^2$ and $\alpha_2 := \|\widehat{A}^\dagger_{2\cdot}\|_2/\|\widehat{A}_{\cdot 1}\|_2^2$, and construct W from Lemma 14 with

$$E := \begin{bmatrix} \|\widehat{A}^\dagger_{1\cdot}\|_2 & -\alpha_1\widehat{A}^\intercal_{\cdot 1}\widehat{A}_{\cdot 2} \\ -\alpha_2\widehat{A}^\intercal_{\cdot 1}\widehat{A}_{\cdot 2} & \|\widehat{A}^\dagger_{2\cdot}\|_2 \end{bmatrix},$$

so $\widehat{A}^\intercal W A^\intercal = E\widehat{A}^\intercal$, and let $k\in\{1,2\}$ and $\sigma(k) := \{1,2\}\setminus\{k\}$. Then we have

$$\begin{aligned}
\|(\widehat{A}^\intercal W A^\intercal)_{k\cdot}\|_2^2 &= \|E_{k\cdot}\widehat{A}^\intercal\|_2^2 = \frac{\|\widehat{A}^\dagger_{k\cdot}\|_2^2}{\|\widehat{A}_{\cdot\sigma(k)}\|_2^4}\left(\|\widehat{A}_{\cdot k}\|_2^2\|\widehat{A}_{\cdot\sigma(k)}\|_2^4 - (\widehat{A}^\intercal_{\cdot k}\widehat{A}_{\cdot\sigma(k)})^2\|\widehat{A}_{\cdot\sigma(k)}\|_2^2\right) \\
&= \frac{\|\widehat{A}^\dagger_{k\cdot}\|_2^2}{\|\widehat{A}_{\cdot\sigma(k)}\|_2^2}(\|\widehat{A}_{\cdot k}\|_2^2\|\widehat{A}_{\cdot\sigma(k)}\|_2^2 - (\widehat{A}^\intercal_{\cdot k}\widehat{A}_{\cdot\sigma(k)})^2) \\
&= \frac{\|\widehat{A}^\dagger_{k\cdot}\|_2^2}{\|\widehat{A}_{\cdot\sigma(k)}\|_2^2}((\widehat{A}^\intercal\widehat{A})_{kk}(\widehat{A}^\intercal\widehat{A})_{\sigma(k)\sigma(k)} - (\widehat{A}^\intercal\widehat{A})_{k\sigma(k)}(\widehat{A}^\intercal\widehat{A})_{\sigma(k)k}) = \frac{\|\widehat{A}^\dagger_{k\cdot}\|_2^2}{\|\widehat{A}_{\cdot\sigma(k)}\|_2^2}\det(\widehat{A}^\intercal\widehat{A}) = 1,
\end{aligned}$$

where the last equation comes from Lemma 22. As $\widehat{b} = \beta_1\widehat{a}_{j_1} + \beta_2\widehat{a}_{j_2}$, then

$$\begin{aligned}
\|\widehat{b}^\intercal W A^\intercal\|_2 &= \|\beta_1\widehat{a}^\intercal_{j_1}WA^\intercal + \beta_2\widehat{a}^\intercal_{j_2}WA^\intercal\|_2 \le \|\beta_1\widehat{a}^\intercal_{j_1}WA^\intercal\|_2 + \|\beta_2\widehat{a}^\intercal_{j_2}WA^\intercal\|_2 \\
&= |\beta_1|\cdot\|\widehat{a}^\intercal_{j_1}WA^\intercal\|_2 + |\beta_2|\cdot\|\widehat{a}^\intercal_{j_2}WA^\intercal\|_2 = |\beta_1| + |\beta_2|,
\end{aligned}$$

where the inequality comes from the triangle inequality. Then $|\beta_1| + |\beta_2| \le 1 \Rightarrow \|\widehat{b}WA^\intercal\|_2 \le 1$, so by weak duality, we establish that the constructed H is optimal. ◀

References

1 Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, 1974.

2 Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

3 Stephen L. Campbell and Carl D. Meyer. *Generalized Inverses of Linear Transformations*. Society for Industrial and Applied Mathematics, 2009.

4 Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-Sparsity Incoherence for Matrix Decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

5 Ivan Dokmanić and Rémi Gribonval. Beyond Moore–Penrose Part I: generalized inverses that minimize matrix norms, 2017. `https://inria.hal.science/hal-01547183v1/document`.

6 Ivan Dokmanić and Rémi Gribonval. Beyond Moore–Penrose Part II: the sparse pseudoinverse, 2017. `https://hal.inria.fr/hal-01547283/file/pseudo-part2.pdf`.

7 Ivan Dokmanić, Mihailo Kolundžija, and Martin Vetterli. Beyond Moore–Penrose: sparse pseudoinverse. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6526–6530. IEEE, 2013.

8 Marcia Fampa and Jon Lee. On sparse reflexive generalized inverses. *Oper. Res. Lett.*, 46(6):605–610, 2018.

9 Marcia Fampa, Jon Lee, and Gabriel Ponte. Trading off 1-norm and sparsity against rank for linear models using mathematical optimization: 1-norm minimizing partially reflexive ah-symmetric generalized inverses. *Open J. Math. Optim.*, 2: article no. 4 (14 pages), 2021.

10 Marcia Fampa, Jon Lee, Gabriel Ponte, and Luze Xu. Experimental analysis of local searches for sparse reflexive generalized inverses. *J. Glob. Optim.*, 81:1057–1093, 2021.

11 Victor K. Fuentes, Marcia Fampa, and Jon Lee. Sparse pseudoinverses via LP and SDP relaxations of Moore–Penrose. In *CLAIO 2016 (18th Latin-Iberian-American Conference on Operations Research)*, pages 343–350, 2016.

12 Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, 1996.

13 Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. In *UAI'09: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348. AUAI Press, 2009.

14 Zekun Liu and Siwei Yu. Alternating Direction Method of Multipliers Based on $\ell_{2,0}$-Norm for Multiple Measurement Vector Problem. *IEEE Trans. Signal Process.*, 71:3490–3501, 2023.

15 Zhongyuan Lyu and Dong Xia. Optimal estimation and computational limit of low-rank Gaussian mixtures. *Ann. Stat.*, 51(2):646–667, 2023.

16 Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 70(1):53–71, 2008.

17 Karthik Mohan. *Learning structured matrices in high dimensions: Low rank estimation and structured graphical models*. PhD thesis, University of Washington, 2015. `http://hdl.handle.net/1773/33147`.

18 Guillaume R Obozinski, Martin J Wainwright, and Michael Jordan. High-dimensional support union recovery in multivariate regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 1217–1224. Curran Associates, Inc., 2008.

19 Roger Penrose. A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.*, 51:406–413, 1955.

20 Gabriel Ponte, Marcia Fampa, Jon Lee, and Luze Xu. On computing sparse generalized inverses. *Oper. Res. Lett.*, 52: article no. 107058 (6 pages), 2024.

21 Charles A. Rohde. *Contributions to the theory, computation and application of generalized inverses*. PhD thesis, North Carolina State University, Raleigh, 1964. `https://www4.stat.ncsu.edu/~boos/library/mimeo.archive/isms_1964_392.pdf`.

22 Yu Wang, Wotao Yin, and Jinshan Zeng. Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. *J. Sci. Comput.*, 78(1):29–63, 2019.

23 Luze Xu, Marcia Fampa, and Jon Lee. 1-norm minimization and minimum-rank structured sparsity for symmetric and ah-symmetric generalized inverses: rank one and two. `https://arxiv.org/abs/2010.11406`, 2020, to appear in S. Dang, A. Deza, S. Gupta, P. McNicholas, S. Pokutta, M. Sugiyama (eds.) *Data Science and Optimization*, volume 91 of *Fields Institute Communications*, Springer, ISBN 978-3-032-03843-2.

24 Luze Xu, Marcia Fampa, Jon Lee, and Gabriel Ponte. Approximate 1-norm minimization and minimum-rank structured sparsity for various generalized inverses via local search. *SIAM J. Optim.*, 31(3):1722–1747, 2021.

25 Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 68(1):49–67, 2006.

26 Hui Zou and Lingzhou Xue. A Selective Overview of Sparse Principal Component Analysis. *Proc. IEEE*, 106(8):1311–1320, 2018.