Appendix. Online Supplement

A. Proofs

Proof of Proposition 1 Consider any point $(\bar{\boldsymbol{x}}, \bar{z}, \bar{t}) \in \mathbb{R}^n \times [0,1] \times \mathbb{R}_+$ with $0 < \bar{z} < 1$. Observe that

$$(\bar{\boldsymbol{x}}, \bar{z}, \bar{t}) = \bar{z}\left(\frac{\bar{\boldsymbol{x}}}{\bar{z}} - \frac{1-\bar{z}}{\bar{z}}\frac{c \cdot \boldsymbol{a}}{\|\boldsymbol{a}\|_2^2}, 1, 0\right) + (1-\bar{z})\left(\frac{c \cdot \boldsymbol{a}}{\|\boldsymbol{a}\|_2^2}, 0, \frac{\bar{t}}{1-\bar{z}}\right),$$

where both $\left(\frac{\bar{\boldsymbol{x}}}{\bar{z}} - \frac{1-\bar{z}}{\bar{z}}\frac{c\cdot\boldsymbol{a}}{\|\boldsymbol{a}\|_2^2}, 1, 0\right) \in Y_c$ and $\left(\frac{c\cdot\boldsymbol{a}}{\|\boldsymbol{a}\|_2^2}, 0, \frac{\bar{t}}{1-\bar{z}}\right) \in Y_c$, and thus $(\bar{\boldsymbol{x}}, \bar{z}, \bar{t}) \in \operatorname{conv}(Y_c)$. Moreover, since $(\bar{\boldsymbol{x}}, 0, \bar{t}) = \lim_{z \to 0^+}(\bar{\boldsymbol{x}}, z, \bar{t})$, we find that $(\bar{\boldsymbol{x}}, \bar{z}, \bar{t}) \in \operatorname{cl}\operatorname{conv}(Y_c)$ even if $z = 0$. □

Proof of Proposition 2 Let

$$T = \left\{(\boldsymbol{x}, z, t) \in \mathbb{R}^n \times [0,1] \times \mathbb{R} :\ t \geq \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \frac{(1-z)\left(\boldsymbol{a}^\top \boldsymbol{x}\right)^2}{1 + z\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2}\right\}.$$

We show next that: • T is convex, • T induces a relaxation of $Y_{0,\boldsymbol{Q}}$, and • optimization of a linear function over set T is equivalent to optimization over $Y_{0,\boldsymbol{Q}}$.

• ***Convexity*** We show convexity of T by establishing it is equivalent to the SDP-representable set given by constraints

$$0 \leq z \leq 1,\ \begin{pmatrix} \boldsymbol{W} & \boldsymbol{x} \\ \boldsymbol{x}^\top & t \end{pmatrix} \succeq 0,\ \boldsymbol{W} = \boldsymbol{Q}^{-1} - (1-z)\frac{\boldsymbol{Q}^{-1}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Q}^{-1}}{1 + \|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\|_2^2}.$$

Note that $\boldsymbol{W} \succ 0$ since for any $\boldsymbol{y} \neq \boldsymbol{0}$,

$$\boldsymbol{y}^\top \boldsymbol{W}\boldsymbol{y} \geq \boldsymbol{y}^\top\boldsymbol{Q}^{-1}\boldsymbol{y} - \frac{\left(\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\boldsymbol{y}\right)^2}{1 + \left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2} \geq \left(\boldsymbol{y}^\top\boldsymbol{Q}^{-1}\boldsymbol{y}\right)\cdot\left(1 - \frac{\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\|_2^2}{1 + \|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\|_2^2}\right) > 0,$$

where the second inequality uses Cauchy-Schwarz inequality and the last one follows from $\boldsymbol{y} \neq \boldsymbol{0}$ and the definition of $\boldsymbol{Q}^{-1} \succ 0$. Since $\boldsymbol{W}$ is invertible, we find by using the Schur complement [3] that $\begin{pmatrix} \boldsymbol{W} & \boldsymbol{x} \\ \boldsymbol{x}^\top & t \end{pmatrix} \succeq 0 \Leftrightarrow t \geq \boldsymbol{x}^\top\boldsymbol{W}^{-1}\boldsymbol{x}$, and using the Sherman Morrison formula [47, 48] we can establish that $\boldsymbol{W}^{-1} = \boldsymbol{Q} + \frac{1-z}{1+z\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\|_2^2}\boldsymbol{a}\boldsymbol{a}^\top$.

• ***Relaxation*** Observe that if $z = 0$ then T reduces to the inequality $t \geq \boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} + (\boldsymbol{a}^\top\boldsymbol{x})^2$, and if $z = 1$ then T reduces to $t \geq \boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}$. This is precisely the disjunction encoded by $Y_{0,\boldsymbol{Q}}$, hence T is indeed a relaxation.

• ***Equivalence*** Now, to prove $T \subseteq \operatorname{cl}\operatorname{conv}\left(Y_{0,\boldsymbol{Q}}\right)$, let us consider the optimization of an arbitrary linear function over the sets $Y_{0,\boldsymbol{Q}}$ and T:

$$\min_{(\boldsymbol{x},z,t)\in Y_{0,\boldsymbol{Q}}} \boldsymbol{\alpha}^\top\boldsymbol{x} + \beta z + \gamma t \tag{17}$$

$$\min_{(\boldsymbol{x},z,t)\in T} \boldsymbol{\alpha}^\top\boldsymbol{x} + \beta z + \gamma t \tag{18}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ and $\gamma \in \mathbb{R}$. Obviously if (18) has an optimal solution $(\boldsymbol{x}^*, z^*, t^*)$ with $z^* \in \{0,1\}$, then it is also an optimal solution for (17). We then show that whenever (18) admits an optimal solution, there exists one with z binary. And if no optimal solution exists, then both problems (17)-(18) are unbounded.

We can assume that $\gamma > 0$ since (18) trivially has a binary solution if $\gamma = 0$ and $\boldsymbol{\alpha} = \boldsymbol{0}$, or both problems are unbounded (for any other combination of parameters with $\gamma \leq 0$). Moreover, by scaling, we can suppose that $\gamma = 1$. Then, assume that (18) has an optimal solution $(\boldsymbol{x}^*, z^*, t^*)$ with $0 < z^* < 1$. The point $(\boldsymbol{x}^*, z^*)$ is an optimal solution of

$$\min_{(\boldsymbol{x},z)\in\mathbb{R}^n\times[0,1]} q(\boldsymbol{x}, z) \tag{19}$$

with

$$q(\boldsymbol{x}, z) = \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta z + \left\| \boldsymbol{Q}^{1/2} \boldsymbol{x} \right\|_2^2 + \frac{(1-z)\,(\boldsymbol{a}^\top \boldsymbol{x})^2}{1 + z \left\| \boldsymbol{Q}^{-1/2} \boldsymbol{a} \right\|_2^2}. \tag{20}$$

Fixing z in (20) and using the first order optimality conditions, we deduce the following expression of an optimal solution $\boldsymbol{x}(z)$ of $\min_{\boldsymbol{x} \in \mathbb{R}^n} q\,(\boldsymbol{x}, z)$:

$$\boldsymbol{x}(z) = -\frac{1}{2} \boldsymbol{Q}^{-1} \boldsymbol{\alpha} + \frac{1-z}{2\left(1 + \left\| \boldsymbol{Q}^{-1/2a} \right\|_2^2\right)} \boldsymbol{Q}^{-1} \boldsymbol{a} \boldsymbol{a}^\top \boldsymbol{Q}^{-1} \boldsymbol{\alpha}. \tag{21}$$

Thus, problem (19) reduces to $\min_{z \in [0,1]} q\,(\boldsymbol{x}(z), z)$. Substituting $\boldsymbol{x}(z)$ by its expression (21) in (20), we obtain that $q\,(\boldsymbol{x}(z), z)$ is a linear function of z. To be more precise, after computations, we get the following expression.

$$q\,(\boldsymbol{x}(z), z) = \beta z - \frac{1}{4} \left\| \boldsymbol{Q}^{-1/2} \boldsymbol{\alpha} \right\|_2^2 + \frac{(\boldsymbol{a}^\top \boldsymbol{Q}^{-1} \boldsymbol{\alpha})^2}{4\left(1 + \left\| \boldsymbol{Q}^{-1/2} \boldsymbol{a} \right\|_2^2\right)} (1-z)\,. \tag{22}$$

Thus, (19) admits an optimal solution with $z \in \{0, 1\}$, concluding the proof. □

Proof of Theorem 1 Observe that

$$\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + (c + w - \boldsymbol{a}^\top \boldsymbol{x})^2 = c^2 + 2c\,(w - \boldsymbol{a}^\top \boldsymbol{x}) + (\boldsymbol{x}^\top \; w)\, \boldsymbol{Q_1} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix} \tag{23}$$

and an expression of cl conv $\left(\hat{Y}_{c,\boldsymbol{Q}}\right)$ easily follows from one of the convex hull of

$$A_{\boldsymbol{Q_1}} = \left\{ (\boldsymbol{x}, z, t, w) \in \mathbb{R}^n \times \{0,1\} \times \mathbb{R}^2 : t \geq (\boldsymbol{x}^\top \; w)\, \boldsymbol{Q_1} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix}, \qquad w(1-z) = 0 \right\}.$$

Let $\delta \in \mathbb{R}_+$ such that $\boldsymbol{Q_{1-\delta}} \succeq \boldsymbol{0}$. The inequality in the description of $A_{\boldsymbol{Q_1}}$ is equivalent to

$$t \geq \delta w^2 + (\boldsymbol{x}^\top \; w)\, \boldsymbol{Q_{1-\delta}} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix}$$

The perspective reformulation [18, 23] applied to this inequality leads to the definition of the convex set

$$A^\delta_{\boldsymbol{Q_1}} = \left\{ (\boldsymbol{x}, z, t, w) \in \mathbb{R}^n \times [0,1] \times \mathbb{R}^2 \times \mathbb{R}_+ : t \geq \delta \frac{w^2}{z} + (\boldsymbol{x}^\top \; w)\, \boldsymbol{Q_{1-\delta}} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix} \right\}.$$

with the convention $0/0 = 0$ and $\epsilon/0 = +\infty$ for any $\epsilon > 0$. One can easily check that $A_{\boldsymbol{Q_1}} \subset A^\delta_{\boldsymbol{Q_1}}$, which implies cl conv $\left(A_{\boldsymbol{Q_1}}\right) \subseteq$ cl $A^\delta_{\boldsymbol{Q_1}}$. Also, if δ_1 and δ_2 belong to the set $\{\delta \in \mathbb{R}_+ : \boldsymbol{Q_{1-\delta}} \succeq \boldsymbol{0}\}$ and satisfy $\delta_1 \leq \delta_2$ then $A^{\delta_2}_{\boldsymbol{Q_1}} \subseteq A^{\delta_1}_{\boldsymbol{Q_1}}$. It follows that the smallest convex set of the form $A^\delta_{\boldsymbol{Q_1}}$ containing cl conv $\left(A_{\boldsymbol{Q_1}}\right)$ is obtained for δ corresponding to an optimal solution of the SDP: $\max\,\{\delta : \boldsymbol{Q_{1-\delta}} \succeq \boldsymbol{0},\ \delta \in \mathbb{R}_+\}$. Since the matrix $\boldsymbol{Q} + \boldsymbol{a}\boldsymbol{a}^\top$ si positive definite, we have

$$\boldsymbol{Q_{1-\delta}} \succeq \boldsymbol{0} \iff \delta \leq 1 - \boldsymbol{a}^\top \left[\boldsymbol{Q} + \boldsymbol{a}\boldsymbol{a}^\top\right]^{-1} \boldsymbol{a} = \frac{1}{1 + \left\| \boldsymbol{Q}^{-1/2} \boldsymbol{a} \right\|_2^2}$$

where the equivalence follows from Schur complement properties and the equation is obtained using the Sherman-Morrison formula. Thus an optimal solution to the above SDP is $\delta^{\max} = \frac{1}{1 + \|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\|^2}$. And substituting δ by $\delta^{\max}$ in the expression

$$t \geq \delta \frac{w^2}{z} + (\boldsymbol{x}^\top \; w)\, \boldsymbol{Q_{1-\delta}} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix},$$

we get

$$t \geq \frac{w^2}{\left(1+\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2\right) z} + \left(\boldsymbol{x}^\top \; w\right) \boldsymbol{Q}_{\boldsymbol{\Delta}} \begin{pmatrix} \boldsymbol{x} \\ w \end{pmatrix} \tag{24}$$

thus leading to cl conv $\left(A_{\boldsymbol{Q_1}}\right) \subseteq T$, where T denotes the set $T = \{(\boldsymbol{x}, z, t, w) \in \mathbb{R}^n \times [0,1] \times \mathbb{R}^2 \colon (24)\}$.

We now prove $T \subseteq$ cl conv $\left(A_{\boldsymbol{Q_1}}\right)$. Consider the optimization of an arbitrary linear function over the sets $A_{\boldsymbol{Q_1}}$ and T:

$$\min_{(\boldsymbol{x},z,t,w)\in A_{\boldsymbol{Q_1}}} \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta z + \gamma t + \zeta w \tag{25}$$

$$\min_{(\boldsymbol{x},z,t,w)\in T} \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta z + \gamma t + \zeta w \tag{26}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $(\beta, \gamma, \zeta) \in \mathbb{R}^3$. Obviously if (26) has an optimal solution $(\boldsymbol{x}^*, z^*, t^*, w^*)$ with $z^* \in \{0, 1\}$, then it is also an optimal solution for (25). We then show that whenever (26) admits an optimal solution, there exists one with z binary. And if no optimal solution exists, then both problems (25)-(26) are unbounded.

If $(\boldsymbol{\alpha}^\top, \gamma, \zeta) = \boldsymbol{0}$, then (26) admits an optimal solution with z binary. If $(\boldsymbol{\alpha}^\top, \gamma, \zeta) \neq \boldsymbol{0}$ and $\gamma \leq 0$, then (25) and (26) are unbounded.

So let us now assume $\gamma > 0$, and with no loss of generality (by scaling) take $\gamma = 1$. Let $(\boldsymbol{x}^*, z^*, t^*, w^*)$ denote an optimal solution of (26). Then, $(\boldsymbol{x}^*, z^*, w^*)$ is an optimal solution of

$$\min_{(\boldsymbol{x},z,w)\in\mathbb{R}^n\times[0,1]\times\mathbb{R}} q\left(\boldsymbol{x}, z, w\right)$$

with

$$q\left(\boldsymbol{x}, z, w\right) = \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta z + \zeta w + \delta^{\max}\frac{w^2}{z} + \left(1-\delta^{\max}\right) w^2 - 2w\boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{x}^\top\left(\boldsymbol{Q} + \boldsymbol{a}\boldsymbol{a}^\top\right)\boldsymbol{x}$$

First order optimality conditions lead to

$$w = \frac{\left(2\boldsymbol{a}^\top x - \zeta\right) z}{2\left(\delta^{\max} + z\left(1-\delta^{\max}\right)\right)}.$$

Substituting w by the right-hand side above in q and then minimizing the resulting expression over $\boldsymbol{x}$, we get that z^* minimizes

$$\overline{q}(z) = -\frac{1}{4}\left(h_2(z)\boldsymbol{a} + \boldsymbol{\alpha}\right)^\top \left[\boldsymbol{Q} + h_1(z)\boldsymbol{a}\boldsymbol{a}^\top\right]^{-1}\left(h_2(z)\boldsymbol{a} + \boldsymbol{\alpha}\right) - \frac{\zeta}{4}h_2(z) + \beta z$$

where $h_1(z) = \frac{\delta^{\max}(1-z)}{\delta^{\max}+z(1-\delta^{\max})} = \frac{1-z}{1+z\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2}$, and $h_2(z) = \frac{\zeta z}{\delta^{\max}+z(1-\delta^{\max})} = \frac{\zeta z\left(1+\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2\right)}{1+z\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2}$. Expanding the expression of $\overline{q}$, we get:

$$-\frac{1}{4}\left[\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{\alpha}\right\|_2^2 + 2\zeta z\boldsymbol{a}^\top\boldsymbol{Q}^{-1}\boldsymbol{\alpha} + \zeta^2 z\left(1+\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2\right) - \frac{(1-z)\left(\boldsymbol{\alpha}^\top\boldsymbol{Q}^{-1}\boldsymbol{a}\right)^2}{1+\left\|\boldsymbol{Q}^{-1/2}\boldsymbol{a}\right\|_2^2}\right] + \beta z$$

which is linear in z, thus implying that the minimization problem of q above admits an optimal solution with z integral.

□

Proof of Proposition 3 From the generalized Schur complement [3], we find that constraint (15b) is equivalent to

$$\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{d}) \succeq 0, \tag{27a}$$

$$\left(\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{d})\right)\left(\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{d})\right)^\dagger \boldsymbol{A} = \boldsymbol{A}, \text{ and} \tag{27b}$$

$$\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{T}^\top\boldsymbol{T} - \boldsymbol{A}^\top\left(\boldsymbol{I} - \mathrm{Diag}(\boldsymbol{d})\right)^\dagger \boldsymbol{A} \succeq 0. \tag{27c}$$

Constraint (27a) is equivalent to $\boldsymbol{d} \leq \boldsymbol{1}$. Constraint (27b) is automatically satisfied if $\boldsymbol{d} < \boldsymbol{1}$, since in that case matrix $(\boldsymbol{I} - \text{Diag}(\boldsymbol{d}))^{\dagger} = (\boldsymbol{I} - \text{Diag}(\boldsymbol{d}))^{-1}$. In general, however, $\Omega = (\boldsymbol{I} - \text{Diag}(\boldsymbol{d}))\,(\boldsymbol{I} - \text{Diag}(\boldsymbol{d}))^{\dagger}$ is the diagonal matrix such that $\Omega_{ii} = \boldsymbol{1}_{\{d_i<1\}}$. Therefore, if $d_i = 1$, then the i-th row of matrix ΩA is a row of 0s, and constraint (27b) cannot be satisfied in that case unless the i-th row of $\boldsymbol{A}$ is also $\boldsymbol{0}$.

Finally, perform a change of variables $u_i = \frac{1}{1-d_i}$, well defined since $d_i < 1$ holds. From constraints $\boldsymbol{d} \geq 0$ we find $\boldsymbol{u} \geq \boldsymbol{1}$. Problem (15) reduces to

$$\max_{\boldsymbol{u}\in\mathbb{R}^m} \sum_{i=1}^{m} \left(\bar{w}_i^2 \left(\frac{1}{\bar{z}_i} - 1 \right) - \bar{w}_i^2 \left(\frac{1}{\bar{z}_i} - 1 \right) \frac{1}{u_i} \right)$$
$$\text{s.t. } \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{T}^\top \boldsymbol{T} - \boldsymbol{A}^\top \text{Diag}(\boldsymbol{u}) \boldsymbol{A} \succeq 0,\ \boldsymbol{u} \geq \boldsymbol{1}.$$

The result then follows by removing terms in the objective not involving $\boldsymbol{u}$. □

B. Implementation details and numerical considerations

We now discuss how we select and tune parameters for the different methods, as well as discuss potential issues if the parameters are poorly chosen.

`big-M` Formulation (11) depends on the parameter M. If a small value is chosen, then the formulation might remove optimal solutions. If the value chosen is too large, then numerical issues can be encountered: for example, the solver might set $z_j = 10^{-5}$ for some j (which is interpreted as 0 due to numerical precision of solvers) but set w_j to a large value, while satisfying constraint $|w_j| \leq M z_j$. In our experiments we set $M = 1{,}000$, and we did not observe any numerical issue in our experiments. Observe that since data is standardized, thus $|y_i| \leq 1$ for all datapoints $i \in \{1, \ldots, m\}$, we intuitively expect that maximum correction $\max_i |w_i|$ to be less than one (note: this intuition is not guaranteed to hold in practice); in other words, the value of M is several times larger than the intuitive bound. This parameter was not tuned.

`conic` Formulation (12) does not involve any parameter, however it requires to have $\lambda > 0$ and may result in incorrect behavior if $\lambda \to 0$. Moreover, based on past experience by the authors, mixed-integer SOCP formulations may result in poor performance or numerical difficulties in large problems. Our experiments satisfy $\lambda \geq 0.01$ and we did not observe any numerical issues.

To handle the intercept (recall the discussion in §4.4), we tested both fixing $x_0 = 0$, or using the intercept $\bar{x}_0$ produced by `ls+l2` as a proxy, and adding the regularization term $\lambda(\bar{x}_0 - x_0)^2$ to the objective (where λ is the same coefficient as the one appearing in term $\lambda\|\boldsymbol{x}\|_2^2$). We did not observe major differences between the two approaches, in terms of quality or solution times. In our experiments with real data we set $x_0 = 0$, and in our experiments with synthetic data we use the value of `ls+l2` as a proxy (note that the synthetic experiment includes the instances where `ls+l2` performs worse).

`conic+` As the most sophisticated formulation, there are several implementation details associated with method `conic+`. First, note that if an optimal solution of problem (16) satisfies $u_i^* = 1$ for some index $i \in [m]$, then $d_i^* = 0$ and formulation (14) does not include term w_i^2/z_i (thus the MIO formulation is not exact, but rather a relaxation). Thus, in our computations, we set a constraint $u_i \geq 1.001$. We did not tune this lower bound, although we noted that simply setting $u_i \geq 1$ does indeed result in incorrect results.

We now discuss the termination criterion of Algorithm 1. Note that at each iteration, at line 5 of the algorithm, a lower bound on the optimal objective value of (7) is computed. Moreover, given the solution to the relaxation $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}, \bar{\boldsymbol{w}})$, we can compute an upper bound on the optimal objective value using a rounding heuristic, by setting $z_j = 1$ for indexes corresponding to the $m - h$ largest values of $\boldsymbol{z}$, and then solving (7) with $\boldsymbol{z}$ fixed. Neither the sequence of lower and upper bounds produced by the algorithm is guaranteed to be monotonic, so we track the best lower (LB) and upper bound (UB) found throughout all previous iterations, and compute an optimality gap at any given iteration as gap $= (UB - LB)/UB$. Finally, we stop the algorithm after 20 iterations (not necessarily consecutive) in which the gap improvement from one iteration to the next is less than 10^{-6}. The parameters $(20, 10^{-6})$ were tuned minimally, based on one synthetic instance and one real instance with the alcohol dataset (and on these datasets we did not observe major differences for different choices of parameters).

In terms of numerical difficulties, in addition to those already mentioned for the `conic` formulation, method `conic+` requires solving several SDPs with low-dimensional cones. Certainly, SDPs are inherently more difficult than quadratic or conic quadratic problems, more sensitive to the input data and more prone to numerical instabilities. For example, we observed that if the raw data $(\boldsymbol{A}, \boldsymbol{y})$ is used without standardization, the SDP solver encounters numerical difficulties in several of the instances. In our experiments, with standardized data, we did encounter numerical issues in a single instance (out of over 400). The intercept is handled similarly to `conic`.

Hyperparameter selection In general, in problems with outliers, it is not possible to use cross-validation to select hyperparameters, as the validation set would also be contaminated with outliers. For the `LTS` method, [42] show that setting $h = \lfloor m/2 \rfloor + \lfloor (n+1)/2 \rfloor$ achieves the optimal breakdown point of $1/2$. However, since this approach involves discarding almost half the datapoints (if n is small), it is not ideal in practice, and the same authors suggest using a smaller proportion of the data (e.g., 20%) as a rule of thumb. Clearly, if the number of outliers is known a priori, it is possible to set h to that quantity – this is the approach we use in our experiments with synthetic data, where we evaluate the statistical performance. Note that for the `huber` method, choosing δ is non-trivial even if the number of outliers is known. In our experiments, in every instance we test $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and select the one that results in best out-of-sample performance (thus the reported results for `huber` are better than what could be achieved in practice).

C. Detailed computational results with real data

Table 4 presents detailed results for each dataset as a function of the budget parameter $m - h$, and Table 5 presents detailed results as a function of the regularization parameter λ. It shows for each dataset and budget/regularization parameter the average time and branch-and-bound nodes used by the solver (time-outs are counted as 600 seconds), as well as the end gaps as reported by the solver (instances solved to optimality count as 0%). We see that formulation `big-M` can solve instances if the parameter $m - h$ is small (and the number of feasible solutions $\binom{m}{m-h}$ is small as well) but struggles in other instances. Formulations `conic` and `conic+` also perform better if the parameter $m - h$ is small (since enumeration is more effective) or if the regularization parameter is large (since the relaxations are stronger). Formulation `conic` is competitive or better than `conic+` in the smaller datasets such as alcohol, but `conic+` is superior overall.

Table 4 Performance of MIO methods as a function of the budget $m-h$. Each row is an average over three instances, with different values of parameter λ. `conic+` encountered numerical difficulties in an instance of radarimage with parameter $m-h=\lfloor 0.4m\rfloor$ and $\lambda=0.2$, which is indicated with a † in the table.

name	$m-h$	`big-M`			`conic`			`conic+`		
		`time`	`nodes`	`gap`	`time`	`nodes`	`gap`	`time`	`nodes`	`gap`
alcohol	$\lfloor 0.1m\rfloor$	2	37,063	0.0%	1	3,880	0.0%	3	5,481	0.0%
	$\lfloor 0.2m\rfloor$	174	2,226,198	0.0%	5	20,999	0.0%	11	34,757	0.0%
	$\lfloor 0.3m\rfloor$	600	9,415,411	7.3%	12	40,484	0.0%	10	25,980	0.0%
	$\lfloor 0.4m\rfloor$	576	8,681,299	7.8%	7	24,367	0.0%	14	40,210	0.0%
education	$\lfloor 0.1m\rfloor$	4	65,793	0.0%	1	2,841	0.0%	2	1,288	0.0%
	$\lfloor 0.2m\rfloor$	600	5,015,602	35.1%	49	73,475	0.0%	13	21,019	0.0%
	$\lfloor 0.3m\rfloor$	600	4,037,385	60.4%	571	601,629	8.9%	143	179,170	0.0%
	$\lfloor 0.4m\rfloor$	600	2,961,573	60.0%	600	772,419	19.6%	256	142,113	4.4%
epilepsy	$\lfloor 0.1m\rfloor$	10	43,399	0.0%	1	1,382	0.0%	3	510	0.0%
	$\lfloor 0.2m\rfloor$	424	4,425,843	1.1%	26	70,898	0.0%	6	9,211	0.0%
	$\lfloor 0.3m\rfloor$	600	5,396,641	29.1%	600	1,705,862	11.1%	249	215,514	1.7%
	$\lfloor 0.4m\rfloor$	600	6,885,941	29.8%	600	1,922,058	18.6%	351	519,278	2.1%
pulpfiber	$\lfloor 0.1m\rfloor$	5	49,645	0.0%	2	4,385	0.0%	3	4,219	0.0%
	$\lfloor 0.2m\rfloor$	600	2,907,040	26.5%	397	565,807	2.3%	414	591,617	0.0%
	$\lfloor 0.3m\rfloor$	600	2,881,002	35.7%	600	839,198	13.0%	600	562,363	11.8%
	$\lfloor 0.4m\rfloor$	600	3,161,658	36.0%	600	348,519	16.7%	600	553,325	12.5%
wagner	$\lfloor 0.1m\rfloor$	292	3,080,940	0.0%	225	464,437	0.0%	28	68,909	0.0%
	$\lfloor 0.2m\rfloor$	600	5,003,290	60.1%	600	873,538	36.9%	428	737,124	12.2%
	$\lfloor 0.3m\rfloor$	600	3,249,964	65.9%	600	553,961	48.1%	600	313,817	21.4%
	$\lfloor 0.4m\rfloor$	600	3,263,027	63.1%	600	730,974	53.0%	600	279,949	29.3%
milk	$\lfloor 0.1m\rfloor$	600	2,166,874	52.6%	600	664,593	21.4%	424	414,886	6.8%
	$\lfloor 0.2m\rfloor$	600	2,334,757	75.6%	600	615,562	46.7%	600	283,008	21.9%
	$\lfloor 0.3m\rfloor$	600	2,342,248	74.7%	600	601,391	51.4%	600	263,689	26.1%
	$\lfloor 0.4m\rfloor$	600	2,457,988	73.4%	600	453,377	52.4%	600	336,605	25.0%
foodstamp	$\lfloor 0.1m\rfloor$	600	3,679,125	79.7%	600	962,111	15.6%	387	652,085	5.4%
	$\lfloor 0.2m\rfloor$	600	3,428,673	94.1%	600	937,462	51.1%	600	628,270	31.0%
	$\lfloor 0.3m\rfloor$	600	1,365,841	95.4%	600	232,394	62.7%	600	260,683	42.2%
	$\lfloor 0.4m\rfloor$	600	1,568,131	96.2%	600	250,902	65.9%	600	260,335	47.0%
radarimage	$\lfloor 0.1m\rfloor$	600	117,206	99.7%	600	18,268	72.7%	600	4,604	30.8%
	$\lfloor 0.2m\rfloor$	600	219,463	99.8%	600	17,128	80.7%	600	5,066	44.1%
	$\lfloor 0.3m\rfloor$	600	258,303	99.8%	600	21,989	87.9%	600	7,681	53.4%
	$\lfloor 0.4m\rfloor$	600	255,181	99.9%	600	19,899	92.1%	†	†	†

Table 5 **Performance of MIO methods as a function of the regularization λ. Each row is an average over four instances, with different values of parameter $m-h$. `conic+` encountered numerical difficulties in an instance of radarimage with parameter $m-h=\lfloor 0.4m \rfloor$ and $\lambda=0.2$, which is indicated with a † in the table.**

name	λ	`big-M`			`conic`			`conic+`		
		`time`	`nodes`	`gap`	`time`	`nodes`	`gap`	`time`	`nodes`	`gap`
alcohol	0.05	331	4,998,999	2.7%	15	49,984	0.0%	18	52,293	0.0%
	0.10	317	5,014,222	4.0%	4	12,738	0.0%	8	22,117	0.0%
	0.20	366	5,256,758	4.7%	1	4,576	0.0%	3	5,422	0.0%
education	0.05	451	2,672,308	32.0%	321	527,878	12.4%	244	208,764	3.3%
	0.10	452	3,013,939	38.5%	311	279,599	7.1%	51	37,059	0.0%
	0.20	451	3,374,019	46.0%	284	280,297	1.8%	15	11,871	0.0%
epilepsy	0.05	456	4,936,641	12.9%	310	733,918	9.9%	303	296,293	2.8%
	0.10	392	3,381,921	17.1%	307	998,623	7.9%	125	206,952	0.0%
	0.20	378	4,245,307	15.0%	304	1,042,610	4.5%	28	55,140	0.0%
pulpfiber	0.05	451	2,151,589	26.2%	451	513,025	11.5%	445	605,184	8.2%
	0.10	451	2,277,393	24.7%	410	432,838	6.9%	388	327,832	6.1%
	0.20	452	2,320,527	22.8%	339	372,570	5.6%	379	346,628	4.0%
wagner	0.05	533	4,175,966	49.8%	547	751,548	40.9%	467	477,785	26.0%
	0.10	524	3,663,863	47.6%	500	593,842	34.7%	453	335,998	16.0%
	0.20	512	3,108,087	44.4%	472	621,794	27.9%	322	236,066	5.1%
milk	0.05	600	2,366,186	70.0%	600	594,747	57.2%	600	276,703	36.4%
	0.10	600	2,322,270	69.8%	600	603,373	44.2%	600	395,896	19.1%
	0.20	600	2,317,945	67.3%	600	553,073	27.5%	467	301,042	4.4%
foodstamp	0.05	600	2,489,055	92.2%	600	579,845	62.9%	600	466,488	46.5%
	0.10	600	2,466,540	90.9%	600	552,682	49.4%	586	518,835	31.0%
	0.20	600	2,575,733	90.9%	600	654,626	34.2%	453	365,707	16.8%
radarimage	0.05	600	220,631	99.8%	600	22,315	92.1%	600	9,388	67.2%
	0.10	600	200,154	99.8%	600	22,358	85.5%	600	4,670	46.1%
	0.20	600	216,832	99.8%	600	13,291	72.4%	†	†	†