OUTLIER DETECTION IN TIME SERIES VIA MIXED-INTEGER CONIC QUADRATIC OPTIMIZATION*

ANDRÉS GÓMEZ†

Abstract. We consider the problem of estimating the true values of a Wiener process given noisy observations corrupted by outliers. In this paper we show how to improve existing mixed-integer quadratic optimization formulations for this problem. Specifically, we convexify the existing formulations via lifting, deriving new mixed-integer conic quadratic reformulations. The proposed reformulations are stronger and substantially faster when used with current mixed-integer optimization solvers. In our experiments, solution times are improved by at least two orders-of-magnitude.

Key words. Outlier detection, mixed-integer optimization, conic quadratic optimization, convexification, lifting

AMS subject classifications. 90C11

1. Introduction.

Analysis of time series data plays a critical role in forecasting and signal processing problems. Therefore, there is an extensive body of literature devoted to this topic [18]. Some well-known techniques to tackle such problems include Kalman filtering [19, 50, 51], ARIMA and exponential smoothing models [17, 20, 24, 37, 72], and regression-based approaches [30, 49, 63, 68, 71]. However, time series may be affected by gross errors, system changes, strikes, natural disasters, or other forms of outliers. The previously mentioned methods are known to be sensitive to outliers [23], with even a single outlier potentially resulting in incorrect statistical conclusions.

In this paper we consider the problem of, given a mix of noisy and anomalous observations of a Wiener process, obtain a *maximum a posteriori* (MAP) estimate of the true values of the process by simultaneously detecting and removing all outliers. The MAP problem we study involves combinatorial decisions (which points to discard). It has typically been tackled by iterative "greedy" procedures [5, 23, 44, 60, 61, 69, 70] that may result in sub-optimal decisions. Alternatively, the MAP problem can be easily formulated as a mixed-integer quadratic optimization (MIQO) problem [77, 78], but the natural convex relaxation of such formulations is weak and branch-and-bound methods may require a prohibitive amount of time to converge.

Contributions and outline. In this paper, we close the gap between fast but heuristic approaches and exact but impractical MIQO approaches. By exploiting the structure imposed by the Wiener process, we formulate the inference problem as a *mixed-integer conic quadratic optimization* (MICQO) that substantially improves the relaxation quality of the natural MIQO formulation [77]. In our computations, the MICQO formulation reduces the number of branch-and-bounds nodes required to proved optimality by at least three orders-of-magnitude (with respect to natural MIQO formulations), resulting in a speed-up of two orders-of-magnitude or more. The exact approach is shown to outperform heuristic methods in correctly identifying outlier observations and estimating the true value of the Wiener process. Moreover, the convex relaxation of the MICQO is strong enough to provide high quality estimates, which can be computed fast by solving a convex optimization problem. The MICQO

*Submitted to the editors on 12/10/2019.

Funding: This paper is based upon work supported by the National Science Foundation under Grant 1930582.

†Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, CA 90089 (gomezand@usc.edu).

formulation is based on the convexification *in the original space of variables* of a rank-three quadratic function with four continuous variables and two indicator variables; we use the lifting theory developed in [62] to obtain the desired ideal formulations.

The remainder of this paper is organized as follows. In §2 we provide the relevant background for the paper, including a literature review, description of the model and MAP estimation approach. In §3 we discuss the natural (but weak) MIQO formulation. In §4 we focus on the convex relaxations of the mixed-integer optimization problems: we discuss why existing techniques are not able to improve upon the natural MIQO formulation, and then present the proposed conic quadratic reformulation (presented in Corollary 4.4 in the paper). In §5 we report computational results. In §6 we give a formal derivation of the proposed convexification, and in §7 we conclude the paper.

Notation. Throughout the paper, we denote vectors in **bold**. Let $N = \{1, \dots, n\}$. Given a vector $\boldsymbol{y} \in \mathbb{R}^N$ and a set $T \subseteq N$, let $\boldsymbol{y_T}$ denote the subvector of $\boldsymbol{y}$ induced by T. Given $a \in \mathbb{R}$, we use the convention that $a^2/0 = 0$ if $a = 0$ and $a^2/0 = \infty$ otherwise. We use $\boldsymbol{e}$ to denote a vectors of ones (the dimension can be inferred from the context). Given two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same dimension, we denote by $\boldsymbol{x} \circ \boldsymbol{y}$ denotes their Hadamard product, i.e., $(\boldsymbol{x} \circ \boldsymbol{y})_i = x_i y_i$. Given a random variable W, we use to notation $W \sim \mathcal{N}(\mu, \sigma^2)$ to denote that W follows a normal distribution with mean μ and variance σ^2.

2. Background.

In this section we give relevant background for the paper.

2.1. Wiener process.

A stochastic process $\{W_t\}_{t \geq 0}$ is a Wiener process [64] (also called a Brownian motion) if: *(i)* $W_0 = 0$, *(ii)* $\{W_t\}_{t \geq 0}$ has stationary independent increments, and *(iii)* $W_t \sim \mathcal{N}(0, t\gamma^2)$ for some $\gamma > 0$. Note that by transforming the process to W_t/γ, we assume without loss of generality that $\gamma = 1$.

Alternatively, $\{W_t\}_{t \geq 0}$ is a Wiener process if and only if it is Gaussian process with $\mathbb{E}[W_t] = 0$, and, for all $0 \leq t_1 \leq t_2$, $\text{cov}(W_{t_1}, W_{t_2}) = t_1$. In other words, given any sequence $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, the random variables $(W_{t_1}, W_{t_2}, \dots, W_{t_n})$ follow a multivariate random variable with mean vector $\mathbf{0}$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} t_1 & t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & t_2 & \dots & t_2 \\ t_1 & t_2 & t_3 & \dots & t_3 \\ \vdots & \vdots & \vdots & \ddots & \\ t_1 & t_2 & t_3 & \dots & t_n \end{pmatrix}. \tag{2.1}$$

The inverse of matrix Σ can be computed in closed form [26] as

$$\Sigma^{-1}_{ij} = \begin{cases} 0 & \text{if } j > i+1 \\ -\frac{1}{t_j - t_i} & \text{if } j = i+1 \\ \frac{1}{t_1} + \frac{1}{t_2 - t_1} & \text{if } i = j = 1 \\ \frac{1}{t_i - t_{i-1}} + \frac{1}{t_{i+1} - t_i} & \text{if } 1 < i = j < n \\ \frac{1}{t_n - t_{n-1}} & \text{if } i = j = n. \\ \Sigma^{-1}_{ji} & \text{if } i > j. \end{cases} \tag{2.2}$$

Note that if times $t_1, \dots, t_n$ are equally spaced, i.e., $t_i = it_1$, then it follows that

$$W_t = W_{t-1} + \xi_t,$$

where $\{\xi_t\}_{t=1}^n$ are independent with $\xi_t \sim \mathcal{N}(0, t_1)$. Thus, in the context of equally spaced observations, $\{W_t\}_{t\geq 0}$ is an autoregressive model of order one, a common class of processes studied in time-series analysis [31].

2.2. Model. Let $\{W_t\}_{t\geq 0}$ be a Wiener process. Observations $y_1, \ldots, y_n$ of process $\{W_t\}_{t\geq 0}$ are obtained for (a finite set of) times $t_1, \ldots, t_n$, but such observations are corrupted by noise and outliers. Specifically, let $S \subsetneq N$ be the set of outliers, let $\bar{S} := N \setminus S$ be the set of noisy observations, and let $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\sigma} \in \mathbb{R}^N_+$ be known vectors of expected values and standard deviations of the noise at each time period, respectively. Then, for $i \in \bar{S}$, the observation y_i corresponds to the true value of the process W at time t_i plus some Gaussian noise, i.e.,

$$y_i = W_{t_i} + \epsilon_i \text{ for } i \in \bar{S}, \tag{2.3}$$

where $\epsilon_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and independent of other errors; in contrast, for $i \in S$, the observations y_i do not follow (2.3) and are in fact be independent of the process W_t. To simplify the notation, we write W_i instead of W_{t_i}. The goal is to identify the set of outliers $S \in \mathcal{Z} \subseteq 2^N$ – where set $\mathcal{Z}$ encodes priors on the set of outliers – and estimate the true values of the process $\{W_t\}_{t\geq 0}$ at times $t_1, \ldots, t_n$.

This problem can be interpreted as an inference problem in a one-dimensional graphical model where the underlying structure is not fully known, see Figure 1. Note that the conditional independence of $\{W_t\}_{0\leq t<i}$ and $\{W_t\}_{t>i}$ given W_i follows from the fact that matrix $\boldsymbol{\Sigma}^{-1}$ given in (2.2) is sparse and tridiagonal. In the absence of outliers ($S = \emptyset$), then the inference problem could be tackled using well-known tools such as Kalman filters, but such methods fail in the presence of outliers [52]. Since outliers may be independent of the actual process $\{W_t\}_{t\geq 0}$ (and outlier data is not assumed to follow a known distribution), observations y_i with $i \in S$ corresponding to outliers should be discarded before inferring the values W_i.

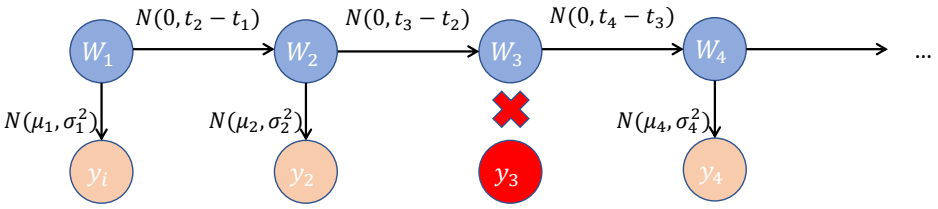


Fig. 1: Schematic representation as a graphical model. A Wiener process (blue) is evolving over time, and observations are obtained. In most cases the observations are noisy (orange, in time indexes 1,2 and 4) but in some cases they are outliers independent of the process (red, at time 3). Which observations are noisy and which are outliers is unknown to the decision maker. The goal is to estimate the values of W_i (blue) given the observations (orange,red).

2.3. Literature review on outlier detection. There has been substantial effort devoted to detecting and removing outliers over the past five decades. In the seminal work of [31], the authors study an autoregressive model without noise, in which there is a single outlier inducing a fixed error of magnitude δ. In other words, $|S| = 1$, $W_{t_i} = y_i$ for $i \in \bar{S}$ and $W_{t_i} = y_i + \delta$ for $i \in S$. They propose to identify the

outlier as the index optimal for the maximum likelihood statistic given by

$$\min_{i \in N} (\boldsymbol{y} - \boldsymbol{\Delta^i})' \boldsymbol{\Sigma^{-1}} (\boldsymbol{y} - \boldsymbol{\Delta^i}), \tag{2.4}$$

where $\boldsymbol{\Delta^i} = (0, \ldots, 0, \delta, 0, \ldots, 0)'$ with the non-zero entry corresponding to the i-th position, and $\boldsymbol{\Sigma^{-1}}$ reduces to (2.2) in the case of a process of order one. Similar models are studied in [2, 23], see also [41, 55].

Methods for detecting and removing *multiple* outliers are often based on the maximum likelihood statistic based on a single outlier. Specifically, [23, 69] propose an iterative procedure identifying one outlier at the time: *(i)* detect a candidate anomalous observation using (2.4); *(ii)* estimate the actual values of the time series $\boldsymbol{y}$ based on the points identified as outliers; and *(iii)* repeat. Several other authors use similar ideas to identify outliers, see [5, 21, 44, 60, 61, 70]. Such procedures, however, may fail if outliers are close or clustered together, causing the diagnostic tests for single outliers to fail. To address this issue, [25] use a procedure similar in spirit to the expectation-maximization algorithm [27]. All discussed methods rely on independent phases for estimating the values of process $\{W_t\}_{t \geq 0}$ (given a tentative set of outliers) and identifying outliers (given tentative estimates of the process), and are thus heuristic in nature. Moreover, the aforementioned methods also implicitly assume that the outliers follow a specific probability distribution, either given by a fixed value as in (2.4), or by a suitable heavy-tail distribution [3]. Naturally, such methods struggle when the distribution is misspecified, or when outliers correspond to gross errors unrelated to the stochastic process.

Robust estimators [65, 67], which call for a *joint* deletion (instead of correction) of spurious observations and estimation of the values of the process, are preferable from a statistical perspective. The class of robust estimators, which include *trimmed least squares* and *least median of squares*, achieve an optimal *breakdown point* [42] (the minimum proportion of anomalous data that could cause the estimator to fail) of 1/2. Unfortunately, robust estimation problems involve combinatorial decisions (which points to discard), are NP-hard in general [10] and difficult to approximate [57]. Classical methods to compute such estimators are either heuristic or rely in complete enumeration [4, 66]. More recently, mixed-integer optimization formulations have been proposed to tackle such problems. In particular, least median of squares estimation and, more generally, least quantile of squares estimation, admit *mixed-integer linear optimization* (MILO) formulations [12, 38]; owing to recent progress in MILO solvers [15], problems with hundreds or thousands of variables can be solved to optimality. In contrast, the trimmed least squares problem is naturally nonlinear, and thus more difficult to solve: Zioutas and Avramidis [77] propose a MIQO formulation, see also [78], but comment that the method is only suitable for small-sample data due to the complexity. In time series analysis, where the sample size (number of time periods) is typically large, existing MIQO methods do not scale.

3. MAP estimation and mixed-integer quadratic optimization. We now discuss the proposed approach to solve the joint estimation and outlier detection problem discussed in §2.2. We propose to simultaneously identify outliers to discard and estimate the value of process W_t by solving a MAP estimation problem, this is, simultaneously find a set of outliers $S \in \mathcal{Z}$ and estimates x of W that result in the most probable outcome of the graphical model depicted in Figure 1. Denote by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be the vector of random variables corresponding to the observations, and let $p(E)$ be the density function of event E happening. The MAP estimate can

be obtained by solving the optimization problem

$$\begin{aligned}
&\max_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} p\left(\boldsymbol{W}=\boldsymbol{x},\boldsymbol{Y_{N\setminus S}}=\boldsymbol{y_{N\setminus S}}\right)\\
&=\max_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} p(\boldsymbol{W}=\boldsymbol{x})p\left(Y_{N\setminus S}=\boldsymbol{y_{N\setminus S}}|\boldsymbol{W}=\boldsymbol{x}\right)\\
&=\max_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} p(\boldsymbol{W}=\boldsymbol{x})\prod_{i\in N\setminus S} p(Y_i=y_i|W_i=x_i)\\
&=\max_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} \frac{e^{-\frac{1}{2}\boldsymbol{x'\Sigma^{-1}x}}}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\prod_{i\in N\setminus S}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(y_i-x_i-\mu_i)^2}{2\sigma_i^2}},
\end{aligned}\tag{3.1}$$

where the second equality follows from the conditional independence Y_i and $\boldsymbol{W}$ given W_i, matrix $\boldsymbol{\Sigma}$ is the covariance matrix of the Wiener process given by (2.1), and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. See also [53] for a similar formula to compute likelihoods corresponding to patches of outliers. Instead of maximizing (3.1), we instead minimize the negative logarithm, resulting in the optimization problem

$$\min_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} \frac{1}{2}\boldsymbol{x'\Sigma^{-1}x}+\sum_{i\in N\setminus S}\frac{1}{2\sigma_i^2}(y_i-x_i-\mu_i)^2+\frac{1}{2}\sum_{i\in N\setminus S}\ln(2\pi\sigma_i^2),\tag{3.2}$$

where we dropped from the optimization the constant term $\frac{1}{2}\ln(2\pi)n+\frac{1}{2}\ln(|\boldsymbol{\Sigma}|)$. Using the closed form expression of $\boldsymbol{\Sigma^{-1}}$ given in (2.2), we find that (3.2) simplifies to

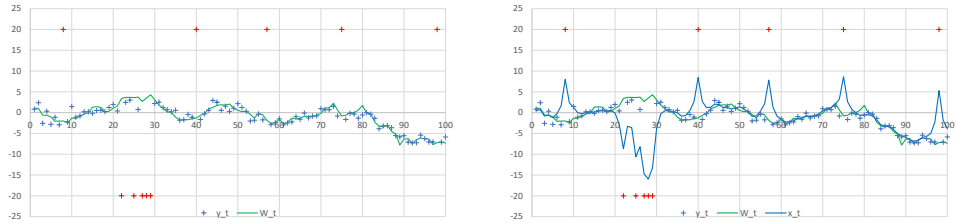
$$\min_{S\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^N} \frac{x_1^2}{2t_1}+\sum_{i=1}^{n-1}\frac{(x_{i+1}-x_i)^2}{2(t_{i+1}-t_i)}+\sum_{i\in N\setminus S}\frac{(y_i-\mu_i-x_i)^2}{2\sigma_i^2}+\sum_{i\in N\setminus S}\frac{\ln(2\pi\sigma_i^2)}{2}.\tag{3.3}$$

Observe that if $\mathcal{Z}=\emptyset$, then (3.3) reduces to a 1D Markov Random Fields problem for which efficient algorithms exist [45, 46, 47]. If $\mathcal{Z}=\{S\subseteq N:|S|=k\}$ for some $k\in\mathbb{Z}_+$, $t_i=i$ for all $i\in N$, $\boldsymbol{\mu}=\boldsymbol{0}$ and $\boldsymbol{\sigma}=\boldsymbol{e}$, then the last term in (3.3) is a constant that can be dropped from the formulation.

Throughout the paper, we will illustrate the formulations using the data described in the following example.

Example. Figure 2 depicts a Wiener process W_t (in green), 100 equally spaced observations $\boldsymbol{y}$ (as crosses) and the MAP estimator that does not account for outliers, $\mathcal{Z}=\emptyset$ (in blue). If there are no outliers, then the MAP estimator is a good estimator of the underlying process. However, if 10 of the observations are actually outliers (red crosses), then the MAP estimator (without removing the outliers) is poor. In this example the noisy observations follow i.i.d. standard normal distributions. The outliers values are equal to ± 20 and are chosen "adversarially" by the author.

Mixed-integer quadratic optimization formulation. We now describe a natural MIQO formulation for problem (3.3). The formulation is a direct adaptation of the MIQO formulation of [77] for the trimmed least squares problem. Define $\boldsymbol{z}\in\{0,1\}^N$ be the indicator vector of S, i.e., $z_i=1$ if and only if $i\in S$. Moreover, let $Z\subseteq\{0,1\}^N$ such that $S\in\mathcal{Z}\Leftrightarrow z\in Z$; we assume that $Z=\left\{\boldsymbol{z}\in\{0,1\}^N:\boldsymbol{Gz}\leq\boldsymbol{b}\right\}$.



(a) No outliers (b) 10 outliers.

Fig. 2: MAP inference of Wiener process with $t_i = i$, $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\sigma} = \boldsymbol{e}$ without accounting for outliers. Blue crosses correspond to noisy observations, red crosses indicate outliers, the green curve indicates the true values of W_t and the blue curve is the MAP estimator. Non-outlier observations are the same in both cases.

Then problem (3.3) can be modeled as the mixed-integer optimization problem

$$\min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}} \frac{x_1^2}{2t_1} + \sum_{i=1}^{n-1} \frac{(x_{i+1} - x_i)^2}{2(t_{i+1} - t_i)} + \sum_{i=1}^{n} \frac{(y_i + v_i - \mu_i - x_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n} \frac{\ln(2\pi\sigma_i^2)}{2} z_i \tag{3.4a}$$

$$\text{s.t. } \boldsymbol{v} \circ (\boldsymbol{e} - \boldsymbol{z}) = \mathbf{0} \tag{3.4b}$$

$$\boldsymbol{G}\boldsymbol{z} \leq \boldsymbol{b} \tag{3.4c}$$

$$\boldsymbol{x} \in \mathbb{R}^N,\ \boldsymbol{z} \in \{0,1\}^N,\ \boldsymbol{v} \in \mathbb{R}^N. \tag{3.4d}$$

Constraints (3.4b) ensure that $z_i = 0 \implies v_i = 0$ for all $i \in N$. Thus, if an outlier is identified at point $i \in N$ then $z_i = 1$ and in optimal solutions of (3.4) we have that $v_i = x_i + \mu_i - y_i$ (independently of the value of x_i), and the error term corresponding to i vanishes – modeling the effect of discarding the value y_i. In contrast, if the datapoint $i \in N$ is not discarded then $z_i = v_i = 0$.

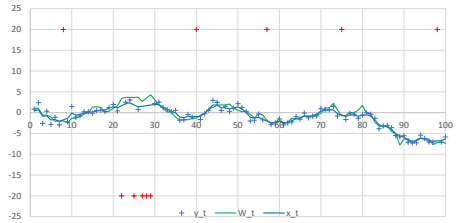
Each quadratic constraint $v_i(1 - z_i) = 0$ in (3.4b) is non-convex, but it can be linearized using big-M constraints. It can be easily shown that there exists an optimal solution of (3.4) satisfying $\min_{j\in N}\{y_j - \mu_j\} \leq x_i \leq \max_{j\in N}\{y_j - \mu_j\}$ for all $i \in N$. Therefore it follows that

$$|v_i| \leq \max_{j\in N}\{y_j - \mu_j\} - \min_{j\in N}\{y_j - \mu_j\} \stackrel{\text{def}}{=} M \quad \forall i \in N,$$

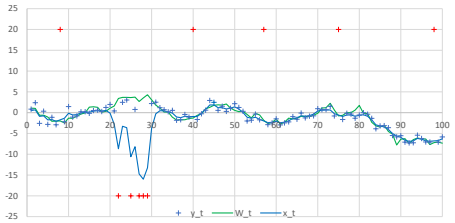
and constraints (3.4b) can be replaced with $-M\boldsymbol{z} \leq \boldsymbol{v} \leq M\boldsymbol{z}$ – note however that the convex relaxation induced by big-M constraints is notoriously weak [54]. Thus, problem (3.4) can be formulated as a MIQO. Perhaps the more common and intuitive specification of Z corresponds to the prior that there is an upper bound k on the number of outliers, which can be modeled via the cardinality constraint

$$Z = \left\{ \boldsymbol{z} \in \{0,1\}^N : \sum_{i\in N} z_i \leq k \right\}.$$

Example. Figure 3 depicts the MAP estimators when a cardinality constraint is imposed. We see that if the true cardinality $k = 10$ is used, then all the outliers are removed and the MAP estimator is a good estimator of the underlying process. If $k = 5$ is used instead, then the five "isolated" outliers are discarded, but the estimator fits the five "clustered" outliers.



(a) $\boldsymbol{e}'\boldsymbol{z} \leq 10$.



(b) $\boldsymbol{e}'\boldsymbol{z} \leq 5$.

Fig. 3: MAP inference of the Wiener process for different cardinalities.

In addition to the simple cardinality constraint, several other priors on the structure of the outliers can be easily incorporated thanks to the modeling power of mixed-integer optimization. Additional modifications of (3.4) can also be envisioned to tackle forecasting and isotonic regression problems, among others. Some of these variants are discussed in Appendix A.

4. Convexification. A direct use of formulation (3.4) with current mixed-integer optimization solvers is sufficient to solve small problems within a reasonable amount of time, but fails in larger instances. To address this limitation, a by now standard approach in both the optimization, statistical and machine learning communities is to derive tight convex relaxations of (3.4). The resulting convex relaxations can then be directly used as a proxy of (3.4) to obtain approximate MAP estimators very efficiently, or can be exploited within a branch-and-bound algorithm to aggressively prune the search tree and prove optimality faster. This section is devoted to studying such relaxations.

4.1. The natural convex relaxation. If $Z = \left\{\boldsymbol{z} \in \{0,1\}^N : \boldsymbol{G}\boldsymbol{z} \leq \boldsymbol{b}\right\}$, then the simplest convex relaxation is obtained simply by relaxing the integrality constraints:

$$\min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}} \ \frac{x_1^2}{2t_1} + \sum_{i=1}^{n-1} \frac{(x_{i+1}-x_i)^2}{2(t_{i+1}-t_i)} + \sum_{i=1}^{n} \frac{(y_i+v_i-\mu_i-x_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n} \frac{\ln(2\pi\sigma_i^2)}{2} z_i \tag{4.1a}$$

$$\text{s.t.} \ -M\boldsymbol{z} \leq \boldsymbol{v} \leq M\boldsymbol{z} \tag{4.1b}$$

$$\boldsymbol{G}\boldsymbol{z} \leq \boldsymbol{b} \tag{4.1c}$$

$$\boldsymbol{x} \in \mathbb{R}^N, \ \boldsymbol{z} \in [0,1]^N, \ \boldsymbol{v} \in \mathbb{R}^N. \tag{4.1d}$$

If the constraints (4.1c) are given by a cardinality constraint $\boldsymbol{e}'\boldsymbol{z} \leq k$, then (4.1) is closely related to the common ℓ_1 norm relaxation $\|\boldsymbol{v}\|_1 \leq Mk$ [28]. For more sophisticated constraints, standard integer programming techniques [73] can be used to improve the formulations. Problem (4.1) is a convex quadratic optimization problem, that can be solved very efficiently in practice. Unfortunately, even for simple constraints (4.1c), the convex relaxation (4.1) is weak and is a poor approximation of (3.4).

Example. Figure 4 depicts the optimal solution of (4.1) with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$. We also use black dots (corresponding to the secondary axis) to represent the optimal values of $\boldsymbol{z}$. Although the optimal values of $\boldsymbol{z}$ are in general low, less than 0.1 in most cases, all points are actually discarded as outliers and the estimator is $\hat{\boldsymbol{x}} \approx \boldsymbol{0}$.

Moreover, the optimal objective value of (4.1) is -9.2 while the optimal objective value of (3.4) is 18.3, for an integrality gap of $\frac{18.3-(-9.2)}{18.3} \times 100 = 150\%$.

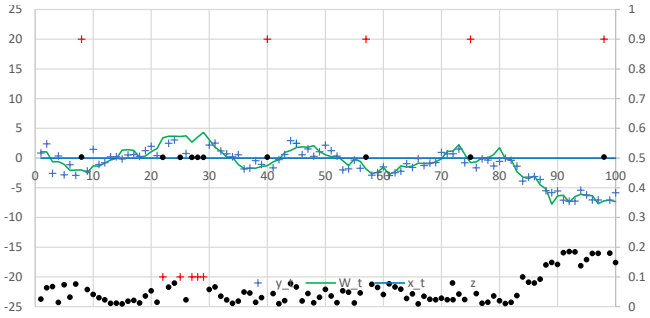


Fig. 4: Solution of the convex relaxation (4.1) with constraint $\boldsymbol{e'z} \leq 10$.

4.2. No separable or rank-one strengthening. Motivated by applications in portfolio optimization [13] and, more recently, sparse regression [11], there has been progress in the past two decades in improving the convex relaxations of MIQO problems [6, 7, **?**, 8, 14, 16, 29, 28, 33, 34, 35, 32, 36, 40, 43, 48, 74, 75, 76]. Most of the existing approaches rely on the perspective reformulation [22, 33], which can be used to derive ideal formulations of *separable* nonlinear functions with indicators: a separable quadratic term $\sum_{i=1}^n (d_i v_i)^2$ in the objective value for some vector $\boldsymbol{d} \in \mathbb{R}^n$ can be replaced with its convexification $\sum_{i=1}^n (d_i v_i)^2 / z_i$. However, the objective function (3.4a) does not have any such terms in $\boldsymbol{v}$, and the perspective reformulation cannot be naturally used.

An alternative is to derive convexifications based on the rank-one quadratic terms $q_i(x_i, v_i) = (y_i + v_i - \mu_i - x_i)^2 = (y_i - \mu_i)^2 + 2(y_i - \mu_i)(v_i - x_i) + (v_i - x_i)^2$. Towards this end, let

$$X_1 = \left\{(x, z, v, s) \in \mathbb{R} \times \{0,1\} \times \mathbb{R} \times \mathbb{R} : (x - v)^2 \leq s,\ v(1 - z) = 0\right\}.$$

Unfortunately, as Proposition 4.1 shows, no strengthening can be derived from the study of X_1, as the closure of its convex hull corresponds to simply relaxing the integrality constraints and dropping the complementary constraints.

PROPOSITION 4.1. *cl conv*$(X_1) = \left\{(x, z, v, s) \in \mathbb{R} \times [0,1] \times \mathbb{R} \times \mathbb{R} : (x - v)^2 \leq s\right\}$.

Proof. Let

$$\bar{X}_1 = \Big\{(w, x, z, v, s) \in \{0,1\} \times \mathbb{R} \times \{0,1\} \times \mathbb{R} \times \mathbb{R} : (x - v)^2 \leq s, \\ v(1 - z) = 0,\ x(1 - w) = 0\Big\}.$$

Atamtürk and Gómez [8] show that

$$\text{cl conv}(\bar{X}_1) = \left\{(w, x, z, v, s) \in [0,1] \times \mathbb{R} \times [0,1] \times \mathbb{R} \times \mathbb{R} : (x - v)^2 \leq s,\ \frac{(x - v)^2}{w + z} \leq s\right\}.$$

By fixing $w = 1$ and noting that the second constraint is redundant, we obtain the result. ∎

Therefore, in order to improve upon the natural convex relaxation (4.1), it is necessary to study sets with more sophisticated quadratic functions.

4.3. Convexification via lifting. As discussed in §4.2, existing convexification schemes cannot be used to strengthen the natural convex relaxation (4.1). In this section we derive a new convex relaxation which exploits additional structure of the objective function of (3.4).

By expanding the quadratic terms

$$(y_i + v_i - \mu_i - x_i)^2 = (y_i - \mu_i)^2 - 2(y_i - \mu_i)(x_i - v_i) + (x_i - v_i)^2$$

and rearranged terms, problem (3.4) can be stated equivalently as

$$\min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}} \frac{1}{2}\sum_{i=2}^{n-2}\left(\frac{1}{2\sigma_i^2}(x_i - v_i)^2 + \frac{(x_{i+1} - x_i)^2}{t_{i+1} - t_i} + \frac{1}{2\sigma_{i+1}^2}(x_{i+1} - v_{i+1})^2\right) \tag{4.2a}$$

$$+ \frac{1}{2}\left(\frac{1}{\sigma_1^2}(x_1 - v_1)^2 + \frac{(x_2 - x_1)^2}{t_2 - t_1} + \frac{1}{2\sigma_2^2}(x_2 - v_2)^2\right) \tag{4.2b}$$

$$+ \frac{1}{2}\left(\frac{1}{2\sigma_{n-1}^2}(x_{n-1} - v_{n-1})^2 + \frac{(x_n - x_{n-1})^2}{t_n - t_{n-1}} + \frac{1}{\sigma_n^2}(x_n - v_n)^2\right) \tag{4.2c}$$

(4.2d)

$$+ \frac{x_1^2}{2t_1} - \sum_{i=1}^{n}\frac{(y_i - \mu_i)(x_i - v_i)}{\sigma_i^2} - \sum_{i=1}^{n}\frac{\ln(2\pi\sigma_i^2)}{2}z_i + \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma_i^2} \tag{4.2e}$$

$$\text{s.t. } -M\boldsymbol{z} \le \boldsymbol{v} \le M\boldsymbol{z} \tag{4.2f}$$

$$\boldsymbol{x} \in \mathbb{R}^N,\ \boldsymbol{z} \in Z,\ \boldsymbol{v} \in \mathbb{R}^N. \tag{4.2g}$$

Since rearranging terms does not impact the relaxation quality, the natural convex relaxation of (4.2) is exactly (4.1). Now, given $a_1, a_2 > 0$, define

$$X = \left\{(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, s) \in \mathbb{R}^2 \times \{0,1\}^2 \times \mathbb{R}^2 \times \mathbb{R} : v_1(1 - z_1) = 0, v_2(1 - z_2) = 0,\right.$$
$$\left.\frac{a_1}{2}(x_1 - v_1)^2 + \frac{(x_1 - x_2)^2}{2} + \frac{a_2}{2}(x_2 - v_2)^2 \le s\right\},$$

and observe that X is the mixed-integer epigraph of the functions in (4.2a)-(4.2c) (after appropriate scaling). The relaxation of (3.4) is then obtained by replacing each term in (4.2a)-(4.2c) by its convexification. Define

$$\bar{a} \stackrel{\text{def}}{=} a_1 a_2 + a_1 + a_2. \tag{4.3}$$

Our main convexification result is stated below.

THEOREM 4.2. *Define functions* $\nu_1, \nu_2 : \mathbb{R}^3 \to \mathbb{R}$ *and* $\bar{\zeta} : \mathbb{R}^2_+ \to \mathbb{R}$ *as*

$$\nu_1(x_1, v_1, v_2) = x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1 - v_2)$$
$$\nu_2(x_2, v_1, v_2) = x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1 - v_2)$$
$$\bar{\zeta}(z_1, z_2) = \min\{1, z_1 + z_2\}.$$

The closure of the convex hull of X *is given by*

$$\textit{cl conv}(X) = \left\{(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, s) \in \mathbb{R}^2 \times [0,1]^2 \times \mathbb{R}^2 \times \mathbb{R} :\right.$$
$$\left.\frac{a_1}{2}\nu_1(x_1, v_1, v_2)^2 + \frac{a_2}{2}\nu_2(x_2, v_1, v_2)^2 + \frac{(\nu_1(x_1, v_1, v_2) - \nu_2(x_2, v_1, v_2))^2}{2} + a_1 a_2 \frac{(v_1 - v_2)^2}{2\bar{a}\bar{\zeta}(z_1, z_2)} \le s\right\}.$$

The proof of Theorem 4.2 is given at the end of the paper, in §6.

Remark 4.3. Set cl conv(X) is conic quadratic representable by adding additional variables $w_1 \stackrel{\text{def}}{=} \nu_1(x_1, v_1, v_2)$, $w_2 \stackrel{\text{def}}{=} \nu_2(x_2, v_1, v_2)$, $\bar{z} \stackrel{\text{def}}{=} \bar{\zeta}(z_1, z_2)$ and $r \stackrel{\text{def}}{=} (v_1 - v_2)^2/\bar{\zeta}(z_1, z_2)$, and using the system of inequalities

$$w_1 = x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1 - v_2) \tag{4.4a}$$

$$w_2 = x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1 - v_2) \tag{4.4b}$$

$$\bar{z} \leq 1,\ \bar{z} \leq z_1 + z_2,\ \bar{z} \geq 0 \tag{4.4c}$$

$$(v_1 - v_2)^2 \leq r\bar{z},\ r \geq 0 \tag{4.4d}$$

$$\frac{a_1}{2}w_1^2 + \frac{a_2}{2}w_2^2 + \frac{(w_1 - w_2)^2}{2} + \frac{a_1 a_2}{2\bar{a}}r \leq s. \tag{4.4e}$$

Constraints (4.4a)-(4.4b) correspond directly to the definition of functions w_1 and w_2, and are linear. Constraints (4.4c) correspond to the linearization of $\min\{1, z_1 + z_2\}$ and are linear as well. The first constraint (4.4d) corresponds to the epigraph of the ratio $(v_1 - v_2)^2/\bar{\zeta}(z_1, z_2)$, and is a rotated cone constraint which can be used with conic quadratic solvers. Finally, constraint (4.4e) is a restatement of the inequality defining cl conv(X), and is convex quadratic.

From Theorem 4.2 and Remark 4.3, we obtain a strong MICQO formulation of (3.4). Define $\lambda_1 = \frac{1}{\sigma_1^2}$, $\lambda_n = 0$ and $\lambda_i = \frac{1}{2\sigma_i^2}$ for $1 < i < n$, and let $L_i = \lambda_i\left(1/\sigma_{i+1}^2 - \lambda_{i+1}\right)(t_{i+1} - t_i) + \lambda_i + 1/\sigma_{i+1}^2 - \lambda_{i+1}$, $i = 1, \ldots, n-1$.Noting that each term in (4.2a) can be written as

$$\frac{1}{2(t_{i+1} - t_i)}\Big(\lambda_i(t_{i+1} - t_i)(x_i - v_i)^2 + (x_{i+1} - x_i)^2 + \left(1/\sigma_{i+1}^2 - \lambda_{i+1}\right)(t_{i+1} - t_i)(x_{i+1} - v_{i+1})^2\Big),$$

and is thus of the form of set X where $a_1 = \lambda_i(t_{i+1} - t_i)$, $a_2 = \left(1/\sigma_{i+1}^2 - \lambda_{i+1}\right)(t_{i+1} - t_i)$ and $\bar{a} = L_i(t_{i+1} - t_i)$. Terms (4.2b)-(4.2c) can be handled identically, and we obtain the strong formulation:

Corollary 4.4. *Formulation*

$$\min \frac{1}{2}\sum_{i=1}^{n-1}\left(\lambda_i w_{i,1}^2 + \frac{(w_{i,1} - w_{i,2})^2}{t_{i+1} - t_i} + \left(\frac{1}{\sigma_{i+1}^2} - \lambda_{i+1}\right)w_{i,2}^2 + \frac{\lambda_i\left(1/\sigma_{i+1}^2 - \lambda_{i+1}\right)}{L_i}r_i\right)$$

$$+ \frac{x_1^2}{2t_1} - \sum_{i=1}^{n}\frac{(y_i - \mu_i)(x_i - v_i)}{\sigma_i^2} - \sum_{i=1}^{n}\frac{\ln(2\pi\sigma_i^2)}{2}z_i + \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma_i^2} \tag{4.5a}$$

(4.5b)

$$s.t.\ w_{i,1} = x_i - v_i + \frac{1/\sigma_{i+1}^2 - \lambda_{i+1}}{L_i}(v_i - v_{i+1}) \qquad i = 1, \ldots, n-1$$

$$w_{i,2} = x_{i+1} - v_{i+1} - \frac{\lambda_i}{L_i}(v_i - v_{i+1}) \qquad i = 1, \ldots, n-1 \tag{4.5c}$$

$$\bar{z}_i \leq 1,\ \bar{z}_i \leq z_i + z_{i+1},\ (v_i - v_{i+1})^2 \leq r_i\bar{z}_i \qquad i = 1, \ldots, n-1 \tag{4.5d}$$

$$-M\boldsymbol{z} \leq \boldsymbol{v} \leq M\boldsymbol{z} \tag{4.5e}$$

$$\boldsymbol{G}\boldsymbol{z} \leq \boldsymbol{b} \tag{4.5f}$$

$$\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \{0,1\}^n,\ \boldsymbol{v} \in \mathbb{R}^n,\ \boldsymbol{w} \in \mathbb{R}^{n\times 2},\ \bar{\boldsymbol{z}} \in \mathbb{R}_+^{n-1},\ \boldsymbol{r} \in \mathbb{R}_+^{n-1}, \tag{4.5g}$$

is a correct formulation for (3.3), and dominates (3.4) in terms of strength of the natural convex relaxations of both formulations.

Example. Figure 5 depicts the optimal solution of the natural convex relaxation of (4.5) with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$. We see that the resulting estimators $\hat{\boldsymbol{x}}$ are a better estimate of W_t than those obtained from (4.1). All indicator variables corresponding to isolated outliers have optimal solutions equal to $z_i = 1$ (while clustered outliers are more difficult to recognize). Moreover, the optimal objective value of the convex relaxation of (4.5) in this case is 15.0, resulting in a much improved integrality gap of $\frac{18.3-15.0}{18.3} \times 100 = 18\%$.

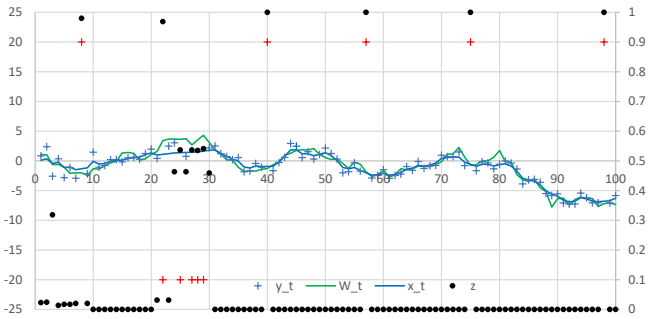


Fig. 5: Solution of the convex relaxation (4.5) with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$.

5. Computations. In this section we report computational experiments comparing the standard formulation (3.4) (`MIQO`) and the proposed stronger formulation (4.5) (`MICQO`). We also benchmark the statistical performance of solving (4.5) against the heuristic methods used in the statistical literature. All tested formulations were coded using AMPL modeling language and the computations were performed in the NEOS Server [1] using Gurobi 8.1.0 with default settings and a 30 minutes time limit. All the raw data, and AMPL model, data and command files used can be found at https://sites.google.com/usc.edu/gomez/data.

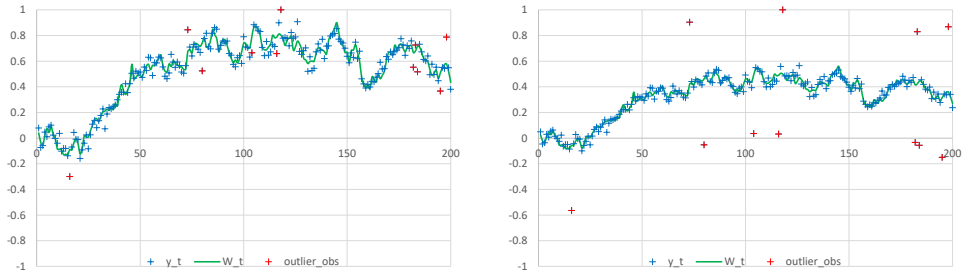
5.1. Instances. The formulations were tested in synthetic instances generated as follows. First, unless stated otherwise, we generate n equally spaced observations of a Wiener process by sampling from a multivariate normal distribution, $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the matrix given in (2.1) with $t_i = i$. Then, given a outlier probability $\tau = 0.1$, we generate observations $\boldsymbol{y}$ as follows. For $i = 1, \ldots, n$, with probability $1 - \tau$, we generate a noisy observations $y_i \sim W_i + \mathcal{N}(0, 1)$, and with probability τ the observation y_i is an outlier. We consider three main classes of outliers:

dev-3 Outliers y_i are generated from a mixture of two normal distributions (with equal weights), each with unit variance and mean equal to $W_i \pm 3$. Thus, on average, outliers are 3 standard deviations away from the mean of a noisy observation.

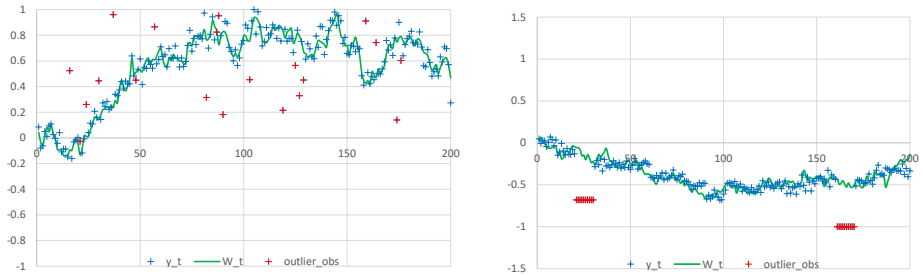
dev-15 Outliers y_i are generated from a mixture of two normal distributions (with equal weights), each with unit variance and mean equal to $W_i \pm 15$. Thus, on average, outliers are 15 standard deviations away from the mean of a noisy observation.

uni Let $\ell = \min_{i \in N} W_i$ and $u = \max_{i \in N} W_i$. Then outliers y_i are generated uniformly in $[\ell, u]$. Observe that unlike **dev-3** and **dev-15**, outliers here are very weakly correlated with $\boldsymbol{W}$.

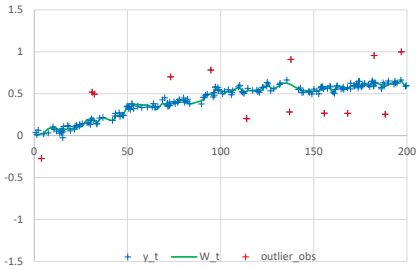
Moreover, we consider two additional variants of **dev-15** instances, resulting in



(a) **dev-3**. (b) **dev-15**.



(c) **uni**. (d) **clu**.



(e) **rti**.

Fig. 6: The five classes of outliers.

more challenging instances (both from an optimization and statistical perspective):

clu Outliers are clustered together in batches of 10. The instance are generated as follows. Let $q = \lfloor n/10 \rfloor$. Then for $0 \leq i \leq q$, either the none of the 10 consecutive datapoints $y_{10i}, y_{10i+1}, \ldots, y_{10i+9}$ are outliers (with probability $1 - \tau$), or all 10 points are outliers (with probability τ). In the later case, y_{10i} is generated as described in **dev-15**, and $y_{10i+j} = y_{10i}$ for $j = 1, \ldots, 9$.

rti The observations are no longer equally spaced. Instead, all times t_i are generated uniformly between 0 and 200.

In all cases, we scale the data so that $\|\boldsymbol{y}\|_\infty = 1$. Figure 6 illustrates all five classes of instances. In our computations, we impose an outlier sparsity constraint $\sum_{i=1}^{n} z_i \leq \tau n$, so that the number of points removed is equal to the expected number

of outliers. Nonetheless, the actual number of outliers may differ substantially from τn in specific instances; we focus on those instances in §5.4. We use $\sigma_i^2 = 1$ in formulations (3.4) and (4.5) (matching the variance of the noise).

5.2. Computational results. First, we test the performance of the formulations for different classes of outliers. Table 1 shows for $n = 200$, different classes of outliers and for both the `MIQO` and `MICQO` formulations (from left to right): the **time** in seconds required to solved the natural convex relaxation; the initial **gap** of the natural convex relaxation, computed as $\frac{\text{obj}_{\text{MIO}} - \text{obj}_{\text{relax}}}{\text{obj}_{\text{MIO}}}$, where obj_{MIO} is the best objective value known for the MIO and $\text{obj}_{\text{relax}}$ is the optimal value of the natural convex relaxation; the **time** required to solve the MIO in seconds; the end **gap** reported by the solver after 30 minutes of branch & bound; the number of branch & bound **nodes** explored; and the number (#) of instances that were solved within the time limit. Each row represents the average over five instances generated with the same parameters. We see that the initial gaps of the `MIQO` are invariably bad, close to 100% in all cases. As a consequence, the branch-and-bound method struggles to prove optimality in most instances, and average end gaps are above 10% in all cases after millions of branch-and-bound nodes. In contrast, formulation `MICQO` is stronger, with initial gaps of 15% or less in instances **dev-3**, **dev-15** and **uni**. The branch-and-bound solver is able to leverage the stronger relaxations to prove optimality within a few minutes in all instances except two, and with only thousands of branch-and-bound nodes. Interestingly, even in instances **clu** and **rti**, where the gaps of the convex relaxation of `MICQO` is relatively high (above 50%), the resulting branch-and-bound algorithm is still substantially faster.

Table 1: Computational results for instances with $n = 200$ and different classes of outliers.

outlier	formulation	convex relaxation		branch & bound			
		time(s)	gap	time(s)	gap	nodes	#
dev-3	`MIQO`	0.03	98.0%	1,800	74.1%	6,790,725	0
	`MICQO`	0.08	13.9%	112	0.0%	41,021	5
dev-15	`MIQO`	0.04	99.5%	734	13.3%	5,745,813	3
	`MICQO`	0.10	15.5%	366	3.3%	37,389	4
uni	`MIQO`	0.04	96.9%	1,462	28.3%	9,538,118	1
	`MICQO`	0.10	7.9%	13	0.0%	4,177	5
clu	`MIQO`	0.02	96.4%	1,800	19.9%	8,979,567	0
	`MICQO`	0.07	53.0%	145	0.0%	29,980	5
tri	`MIQO`	0.05	99.5%	1,329	28.0%	1,741,649	2
	`MICQO`	0.15	62.9%	369	3.2%	5,529	4

We now test the formulations for different dimensions n and outlier class **dev-3** (similar results are obtained for other classes of outliers). Table 2 shows the result. We see that for $n = 100$, formulation `MIQO` is able to deliver optimal solutions in most instances (after 10 minutes on average) after substantial branching; nonetheless,

formulation `MICQO` is clearly superior as it delivers optimal solutions in three seconds, and the number of branch-and-bound nodes required to prove optimality is reduced by three orders-of-magnitude. In instances with $n \geq 200$ formulation `MIQO` struggles badly, resulting in end gaps of 70% or more. In contrast, formulation `MICQO` is able to deliver optimal solutions for $n = 200$ (in under two minutes), and proves stronger optimality gaps of 13% in instances with $n = 500$.

Table 2: Computational results for instances with outliers **dev-3** and different sizes.

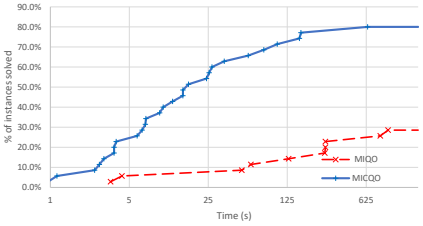
n	formulation	convex relaxation		branch & bound			
		time(s)	gap	time(s)	gap	nodes	#
100	`MIQO`	0.04	99.4%	662	2.4%	6,302,389	4
	`MICQO`	0.05	15.2%	3	0.0%	1,899	5
200	`MIQO`	0.03	98.0%	1,800	74.1%	6,790,725	0
	`MICQO`	0.08	13.9%	112	0.0%	41,021	5
500	`MIQO`	0.08	99.3%	1,800	97.3%	2,347,363	0
	`MICQO`	0.21	16.5%	1,800	12.8%	265,837	0

Summary. Figure 7(a) shows the performance profile of both `MIQO` and `MICQO` across all instances tested. The stronger formulation (4.5) is considerably faster: while `MIQO` is able to solve only 29% of the instances within the time limit of 30 minutes, `MICQO` is able to solve the same quantity of instances in only six seconds. Thus, in our computations, formulation `MICQO` is two orders-of-magnitude faster than `MIQO` in the "easy" instances that both formulations can solve to optimality. The improvement of `MICQO` is larger once the more challenging instances are accounted for. To illustrate, Figure 7(b) depicts the end gaps reported by the solvers on the instances that `MIQO` does not solve to optimality. The gaps reported by `MIQO` are large, with average and median gaps close to 50%, and in some cases gaps as large as 98% – suggesting that it would require a long time indeed to solve these problems using formulation (3.4). In contrast, the median optimality gap of `MICQO` is 0% –since most of these instances are actually solved to optimality when using formulation (4.5)–, the average gap is 3.9%, and the worst gap is under 20%.

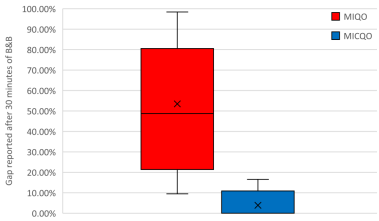
5.3. Statistical performance. In this section we discuss the statistical performance of solving the MIO problem to optimality, and test the statistical performance of using the convex relaxations of (4.5) as a proxy. We benchmark against not removing outliers at all, i.e., setting $\mathcal{Z} = \emptyset$ in (3.3), and using a greedy heuristic similar in spirit to [23, 69], described next.

Greedy heuristic. Let $\bar{S}$ be a tentative set of candidate outliers; initially, we set $\bar{S} = \emptyset$. At each iteration, we solve problem (3.3) with $\mathcal{Z} = \left\{\bar{S} \cup \{i\}\right\}_{i \in N}$. Note that, for this choice of $\mathcal{Z}$, problem (3.3) can be easily solved via enumeration. This process is repeated until $|\bar{S}| = \tau n$.

Table 3 shows the statistical **error** and **power** of the estimators obtained by either not discarding any observations (`no discard`), the greedy heuristic (`heuristic`), solving problem (3.3) to optimality (`mixed-integer optimization`), or solving the natural convex relaxations of (4.5) (`conic quadratic relaxation`) to optimality. Error is the standardized error between the estimated signal $\hat{\boldsymbol{x}}$ (corresponding to the optimal solution of an optimization problem) and the actual values of the Wiener



(a) Percentage of instances solved within a given time limit (log scale).



(b) End gaps reported after 30 minutes of branch & bound on instances that `MIQO` cannot solve to optimality.

Fig. 7: Aggregated computational results.

process $\boldsymbol{W}$: error $= \frac{\|\boldsymbol{W}-\hat{\boldsymbol{x}}\|_2^2}{\|\boldsymbol{W}\|_2^2}$; power is the probability that an anomalous point is labeled as an outlier. For the convex relaxation, we mark a point $i \in N$ as an outlier if the (fractional) value $\hat{z}_i$ of the solution is among the τn largest entries of $\hat{\boldsymbol{z}}$.

In instances **dev-3**, outliers are not significant enough to alter the inference of the underlying process, and not discarding outliers actually results in the best performance. However, we note that identifying them (by any method) results in only a slightly larger error, whereas ignoring outliers results in significantly worse performance in every other instance class.

In instances **dev-15** and **uni**, `mixed-integer optimization` consistently results in the best performance, with `heuristic` also resulting in good performance overall (and in fact matching `mixed-integer optimization` in **uni** instances). Using the `conic quadratic relaxation` results in slightly subpar performance in these two instances classes (by a fraction of a percentage point), although this method is still far superior than simply ignoring outliers.

Finally, in instances **clu** and **rti**, the errors are in general worse (for all methods). Interestingly, the `conic quadratic relaxation` results in substantially less error than other methods (by several percentage points) – although the simple rounding method we use to identify outliers is not successful, as shown by the inferior power of the method. In addition, `mixed-integer optimization` also outperforms the `heuristic` in terms of error.

In summary, we see that while the relative merits of each method depends on the instance class, we can draw the following high-level conclusions:

1. Ignoring outliers can result in large errors, whereas discarding outliers only results in small errors even when outliers are not significant.
2. In general, `mixed-integer optimization` results in better performance than the `heuristic` method, especially in terms of the error.
3. A direct inference based on the `conic quadratic relaxation` is more robust: it may result in much improved statistical properties in the more challenging instances, at the expense of slightly larger errors in simpler instances.

5.4. On outlier sparsity. Theory on robust estimators [65, 67] indicates that such estimators perform well even if the actual number of outliers is substantially less than the number of discarded points. Our experiments corroborate this theoretical result. Figure 8 illustrates this phenomenon in a **dev-15** instance with $n = 200$

Table 3: Statistical performance of different methods in instances with $n = 200$. **Error**= $\frac{\|\boldsymbol{W}-\hat{\boldsymbol{x}}\|_2^2}{\|\boldsymbol{W}\|_2^2}$ and $\hat{\boldsymbol{x}}$ is an optimal solution of the corresponding optimization problem. **Power** is the proportion of outliers that are detected. Each row represents the average over five instances generated with the same parameters.

outlier	metric	error	power
dev-3	`no discard`	**1.7%**	0.0%
	`heuristic`	1.9%	55.6%
	`conic quadratic relaxation`	2.3%	**55.9%**
	`mixed-integer optimization`	1.9%	55.6%
dev-15	`no discard`	12.6%	0.0%
	`heuristic`	1.8%	96.2%
	`conic quadratic relaxation`	2.1%	90.4%
	`mixed-integer optimization`	**1.4%**	**98.2%**
uni	`no discard`	3.7%	0.0%
	`heuristic`	**1.4%**	**66.5%**
	`conic quadratic relaxation`	1.7%	62.0%
	`mixed-integer optimization`	**1.4%**	**66.5%**
clu	`no discard`	40.8%	0.0%
	`heuristic`	21.5%	66.2%
	`conic quadratic relaxation`	**13.2%**	52.2%
	`mixed-integer optimization`	19.5%	**75.4%**
rti	`no discard`	18.8%	0.0%
	`heuristic`	4.4%	**88.8%**
	`conic quadratic relaxation`	**1.0%**	76.8%
	`mixed-integer optimization`	4.0%	84.8%

and 11 outliers (vs 20 discarded points) – corresponding to the instance of this size with the least number of outliers in our computations. Note that we observed a similar phenomenon in §5.3 with **dev-3** instances: while not discarding any points is preferable, optimally choosing 20 points to discard results in a marginally larger error, see Table 3.

Moreover, theory indicates that robust estimators may fail if there are more outliers than discarded points. In our computations we found that this may indeed be the case, but it depends on the structure of the outliers. On the one hand, in **uni** instances, where outlier observations are independent from the Wiener process but could be close to it in some cases, the `mixed-integer optimization` method is robust to miss-specification: Figure 9 presents the **uni** instance with $n = 200$ and most number of outliers.

On the other hand, if outliers are far apart from the true values of the process, then underestimating the number of outliers does result in poor performance of `mixed-integer optimization`. These settings, however, also correspond to the situations in which the `conic quadratic relaxation` results in much better statistical performance than its discrete counterpart. Figures 10 and 11 depict this phenomenon

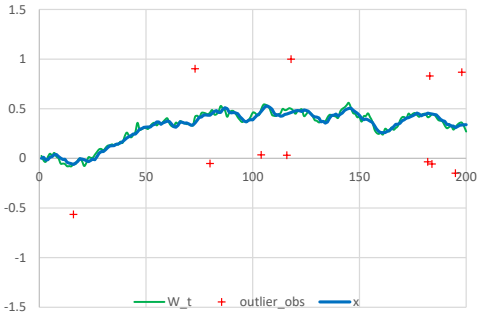


Fig. 8: Solution from `mixed-integer optimization` in a **dev-15** instance with 11 outliers and 20 discarded points. The error with respect to the true signal is 0.4%.

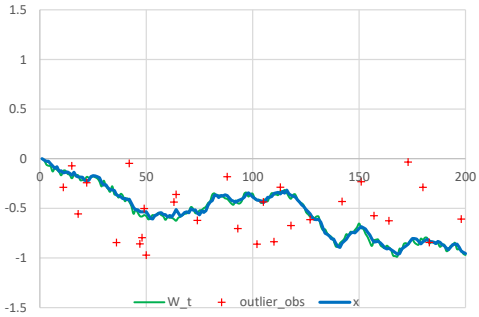
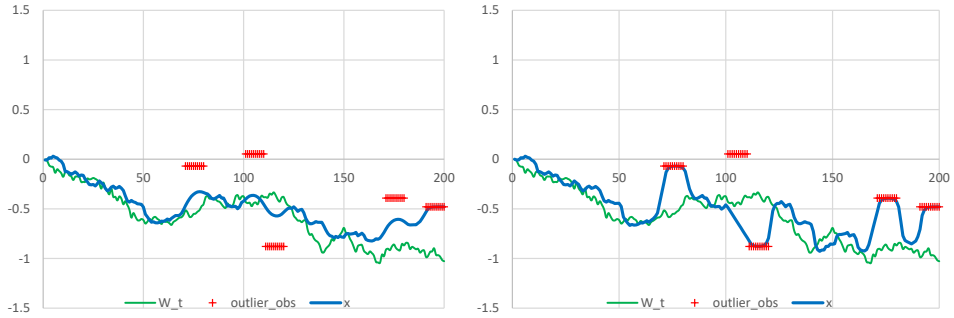


Fig. 9: Solution from `mixed-integer optimization` in a **uni** instance with 29 outliers and 20 discarded points. The error with respect to the true signal is 0.2%.

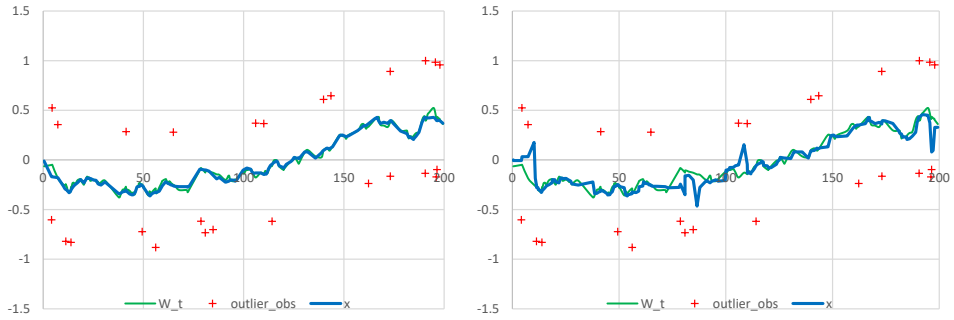
in **clu** and **rti** instances, respectively.



(a) `Conic quadratic relaxation`, error=7.1% (b) `Mixed-integer optimization`, error=14.3%

Fig. 10: Solutions in a **clu** instance with 50 outliers and 20 discarded points.

6. Proof of Theorem 4.2. This section is devoted to prove Theorem 4.2. We derive the result via a similar lifting technique as the one used in [9, 39, 59, 62].



(a) `Conic quadratic relaxation`, error=1.2% (b) `Mixed-integer optimization`, error=7.1%

Fig. 11: Solutions in a **rti** instance with 25 outliers and 20 discarded points.

6.1. Main idea and proof outline. Define

$$f(\boldsymbol{x},\boldsymbol{v}) = \frac{a_1}{2}(x_1 - v_1)^2 + \frac{(x_1 - x_2)^2}{2} + \frac{a_2}{2}(x_2 - v_2)^2, \text{ and}$$
$$C = \left\{(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) \in \mathbb{R}^2 \times \{0,1\}^2 \times \mathbb{R}^2 : v_1(1-z_1) = 0, v_2(1-z_2) = 0\right\}$$

so that $X = \{(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) \in C, s \in \mathbb{R} : f(\boldsymbol{x},\boldsymbol{v}) \leq s\}$. Moreover, for any $\boldsymbol{z} \in \{0,1\}^2$, $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{R}^2$, define the function

$$g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) \stackrel{\text{def}}{=} \min_{\boldsymbol{x},\boldsymbol{v}} f(\boldsymbol{x},\boldsymbol{v}) - \alpha_1 x_1 - \alpha_2 x_2 - \beta_1 v_1 - \beta_2 v_2 \quad \text{s.t. } (\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) \in C. \tag{6.1}$$

Function $g^{\boldsymbol{a}}$ depends on the vector $\boldsymbol{a}$ and is parametrized by $(\boldsymbol{\alpha},\boldsymbol{\beta})$. Observe that any $(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v},s) \in X$ and any $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{R}^2$ we have that

$$s - \boldsymbol{\alpha}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{v} \geq f(\boldsymbol{x},\boldsymbol{v}) - \boldsymbol{\alpha}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{v} \geq g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{6.2}$$

where the second inequality is obtained by minimizing over $(\boldsymbol{x},\boldsymbol{v})$. From (6.2), we can derive lower bounds on s by studying the function $g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) + \boldsymbol{\alpha}'\boldsymbol{x} + \boldsymbol{\beta}'\boldsymbol{v}$.

The proof of Theorem 4.2 consists of three steps:

1. We give, for any $(\boldsymbol{\alpha},\boldsymbol{\beta})$, a closed form expression of the set function $g^{\boldsymbol{a}}$ (Propositions 6.1 and 6.2).
2. We derive, for any $(\boldsymbol{\alpha},\boldsymbol{\beta})$, the closure of the convex envelope of function $g^{\boldsymbol{a}}$, i.e., the maximal convex function $\bar{g}^{\boldsymbol{a}} : [0,1]^2 :\to \mathbb{R}$ such that $\bar{g}^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) \leq g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta})$ for all $z \in \{0,1\}^n$ (Proposition 6.3). It follows from (6.2) that
$$s - \boldsymbol{\alpha}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{v} \geq \bar{g}^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{6.3}$$
3. Noting that (6.3) holds for any $(\boldsymbol{\alpha},\boldsymbol{\beta})$, we find the strong nonlinear valid inequality
$$s - \boldsymbol{\alpha}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{v} \geq \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \bar{g}^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{6.4}$$
It follows from Theorem 1 in [62] that inequality (6.4) and bound constraints describe cl conv(X) (see Proposition 6.4). Thus, to prove Theorem 4.2, it suffices to solve the lifting problem (6.4) in closed form (Proposition 6.5).

6.2. Formal proof. We now give an explicit description of function $g^{\boldsymbol{a}}$ described in (6.1).

PROPOSITION 6.1. *If $\alpha_1+\beta_1 \neq -\alpha_2-\beta_2$, then $g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) = -\infty$. Otherwise, if*

$$\alpha_1 + \beta_1 = \gamma = -\alpha_2 - \beta_2 \tag{6.5}$$

for some $\gamma \in \mathbb{R}$, then

$$g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{cases} -\frac{1}{2}\frac{(\alpha_1+\alpha_2)^2+a_2\alpha_1^2+a_1\alpha_2^2}{\bar{a}} & \text{if } \boldsymbol{z}=\boldsymbol{0} \\ -\frac{1}{2}\left(\frac{(\gamma-\alpha_1)^2}{a_1}+\gamma^2+\frac{(\gamma+\alpha_2)^2}{a_2}\right) & \text{otherwise,} \end{cases}$$

where $\bar{a}$ is given in (4.3).

Proof. We first compute $g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta})$. By taking derivatives of the objective function of (6.1) with respect to v_1 and v_2 and setting to 0, we find

$$v_1 = x_1 + \frac{1}{a_1}\beta_1$$
$$v_2 = x_2 + \frac{1}{a_2}\beta_2.$$

Thus, we obtain

$$g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{x\in\mathbb{R}^2} -\frac{\beta_1^2}{2a_1} - \frac{\beta_2^2}{2a_2} + \frac{1}{2}(x_1-x_2)^2 - (\alpha_1+\beta_1)x_1 - (\alpha_2+\beta_2)x_2. \tag{6.6}$$

If $\alpha_1+\beta_1 < -\alpha_2-\beta_2$, then setting $x_1 = x_2 = \zeta$ and letting $\zeta \to -\infty$, we find that $g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) = -\frac{\beta_1^2}{2a_1} - \frac{\beta_2^2}{2a_2} - (\alpha_1+\alpha_2+\beta_1+\beta_2)\zeta \to -\infty$. Similarly, if $\alpha_1+\beta_1 > -\alpha_2-\beta_2$, then setting $x_1 = x_2 = \zeta$ and letting $\zeta \to \infty$, $g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) \to \infty$. Otherwise, if $\alpha_1+\beta_1 = \gamma = -\alpha_2-\beta_2$, then we find by taking derivatives of (6.6) with respect to x_1, x_2 and setting to 0 that $(x_1-x_2) = \gamma$ and (6.6) reduces to

$$\begin{aligned} g^{\boldsymbol{a}}(\{1,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) &= \min_{x\in\mathbb{R}^2} -\frac{\beta_1^2}{2a_1} - \frac{\beta_2^2}{2a_2} + \frac{1}{2}(x_1-x_2)^2 - \gamma(x_1-x_2) \\ &= -\frac{\beta_1^2}{2a_1} - \frac{\beta_2^2}{2a_2} - \frac{1}{2}\gamma^2 = -\frac{(\gamma-\alpha_1)^2}{2a_1} - \frac{1}{2}\gamma^2 - \frac{(\gamma+\alpha_2)^2}{2a_2}. \end{aligned}$$

We assume that $\alpha_1+\beta_1 = \gamma = -\alpha_2-\beta_2$ in the sequel.

We now compute $g^{\boldsymbol{a}}(\{1,0\};\boldsymbol{\alpha},\boldsymbol{\beta})$. Note that $v_2 = 0$ in (6.1); by taking the derivative of (6.1) with respect to v_1 and setting to 0, we find that $v_1 = x_1 + {}^1\!/_{a_1}\beta_1$ and

$$g^{\boldsymbol{a}}(\{1,0\};\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{x\in\mathbb{R}^2} -\frac{\beta_1^2}{2a_1} + \frac{1}{2}(x_1-x_2)^2 + \frac{a_2}{2}x_2^2 - (\alpha_1+\beta_1)x_1 - \alpha_2 x_2.$$

Taking derivatives with respect to x_1 we find $x_1 = x_2 + \alpha_1 + \beta_1$ and

$$g^{\boldsymbol{a}}(\{1,0\};\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{x_2\in\mathbb{R}} -\frac{\beta_1^2}{2a_1} - \frac{1}{2}(\alpha_1+\beta_1)^2 + \frac{a_2}{2}x_2^2 - (\alpha_1+\alpha_2+\beta_1)x_2.$$

Taking derivatives with respect to x_2 we find $x_2 = \frac{\alpha_1+\alpha_2+\beta_1}{a_2}$ and

$$g^{\boldsymbol{a}}(\{1,0\};\boldsymbol{\alpha},\boldsymbol{\beta}) - \frac{\beta_1^2}{2a_1} - \frac{1}{2}(\alpha_1+\beta_1)^2 - \frac{(\alpha_1+\alpha_2+\beta_1)^2}{2a_2} = -\frac{(\gamma-\alpha_1)^2}{2a_1} - \frac{1}{2}\gamma^2 - \frac{(\gamma+\alpha_2)^2}{2a_2}.$$

Similarly, we find

$$g^{\boldsymbol{a}}(\{0,1\};\boldsymbol{\alpha},\boldsymbol{\beta}) = -\frac{(\alpha_1+\alpha_2+\beta_2)^2}{2a_1} - \frac{1}{2}(\alpha_2+\beta_2)^2 - \frac{\beta_2^2}{2a_2} = -\frac{(\gamma-\alpha_1)^2}{2a_1} - \frac{1}{2}\gamma^2 - \frac{(\gamma+\alpha_2)^2}{2a_2}.$$

Finally, we compute $g^{\boldsymbol{a}}(\{0,0\};\boldsymbol{\alpha},\boldsymbol{\beta})$. Note that $v_1 = v_2 = 0$ in (6.1); thus it follows from standard quadratic optimization arguments that $g^a(\{0,0\};\boldsymbol{\alpha},\boldsymbol{\beta}) = -1/2\boldsymbol{\alpha}'\boldsymbol{A}^{-1}\boldsymbol{\alpha}$ where

$$\boldsymbol{A} = \begin{pmatrix} 1+a_1 & -1 \\ -1 & 1+a_2 \end{pmatrix},$$

i.e.,

$$g^{\boldsymbol{a}}(\{0,0\};\boldsymbol{\alpha},\boldsymbol{\beta}) = -\frac{(1+a_2)\alpha_1^2 + 2\alpha_1\alpha_2 + (a_1+1)\alpha_2^2}{2\bar{a}},$$

and the proof is complete. ∎

Proposition 6.2 gives an alternative characterization of g as a function in terms of variables $\boldsymbol{z}$.

PROPOSITION 6.2. *For any $(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)$ satisfying (6.5), we can rewrite function g as*

$$g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) = g^{\boldsymbol{a}}(\mathbf{0};\boldsymbol{\alpha},\boldsymbol{\beta}) - \rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma}\max\{z_1,z_2\} \tag{6.7}$$

where

$$\rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} = -\frac{\Big(\bar{a}\gamma - (a_2\alpha_1 - a_1\alpha_2)\Big)^2}{2a_1a_2\bar{a}}.$$

Proof. Equation (6.7) is trivially satisfied for $\boldsymbol{z} = \mathbf{0}$. To check that it is also satisfied for $\boldsymbol{z} \neq 0$, we verify that

$$\begin{aligned}
\rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} =& g^{\boldsymbol{a}}(\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta}) - g^{\boldsymbol{a}}(\mathbf{0};\boldsymbol{\alpha},\boldsymbol{\beta}) \\
=& -\frac{1}{2}\left(\frac{(\gamma-\alpha_1)^2}{a_1} + \gamma^2 + \frac{(\gamma+\alpha_2)^2}{a_2} - \frac{(\alpha_1+\alpha_2)^2 + a_2\alpha_1^2 + a_1\alpha_2^2}{\bar{a}}\right) \\
=& -\frac{1}{2}\left(\frac{a_2\gamma^2 - 2a_2\gamma\alpha_1 + a_2\alpha_1^2 + a_1a_2\gamma^2 + a_1\gamma^2 + 2a_1\gamma\alpha_2 + a_1\alpha_2^2}{a_1a_2} - \frac{(\alpha_1+\alpha_2)^2 + a_2\alpha_1^2 + a_1\alpha_2^2}{\bar{a}}\right) \\
=& -\frac{1}{2}\left(\frac{\bar{a}\gamma^2 - 2(a_2\alpha_1 - a_1\alpha_2)\gamma + a_2\alpha_1^2 + a_1\alpha_2^2}{a_1a_2} - \frac{(\alpha_1+\alpha_2)^2 + a_2\alpha_1^2 + a_1\alpha_2^2}{\bar{a}}\right) \\
=& -\frac{1}{2}\left(\frac{(\bar{a}\gamma)^2 - 2\bar{a}(a_2\alpha_1 - a_1\alpha_2)\gamma + (a_1+a_2)(a_2\alpha_1^2 + a_1\alpha_2^2) - a_1a_2(\alpha_1+\alpha_2)^2}{a_1a_2\bar{a}}\right) \\
=& -\frac{1}{2}\left(\frac{(\bar{a}\gamma)^2 - 2\bar{a}(a_2\alpha_1 - a_1\alpha_2)\gamma + (a_2\alpha_1 - a_1\alpha_2)^2}{a_1a_2\bar{a}}\right) \\
=& -\frac{1}{2}\left(\frac{\Big(\bar{a}\gamma - (a_2\alpha_1 - a_1\alpha_2)\Big)^2}{a_1a_2\bar{a}}\right).
\end{aligned}$$

Therefore, the proposition is proven. ∎

Now consider the mixed-integer epigraph of the set function $g^{\boldsymbol{a}}$, i.e.,

$$U = \left\{ \boldsymbol{z} \in \{0,1\}^2, w \in \mathbb{R} : g^{\boldsymbol{a}}(\boldsymbol{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq w \right\}.$$

PROPOSITION 6.3. *For any* $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ *satisfying* (6.5),

$$conv(U) = \left\{ \boldsymbol{z} \in [0,1]^2, w \in \mathbb{R} : g^{\boldsymbol{a}}(\mathbf{0}; \boldsymbol{\alpha}, \boldsymbol{\beta}) - \rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} \min\{1, z_1 + z_2\} \leq w \right\}.$$

Proof. Proposition 6.3 follows directly from Proposition 6.2 and the fact that the hypograph of the maximum of two binary variables, $\bar{z} \leq \max\{z_1, z_2\}$, can be described via the inequalities $\bar{z} \leq 1$ and $\bar{z} \leq z_1 + z_2$.

From inequality (6.1) and Proposition 6.3 we find that for all $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, s) \in X$, $\bar{z} \stackrel{\text{def}}{=} \min\{1, z_1 + z_2\}$, and all $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ satisfying (6.5), the inequality

$$s - \boldsymbol{\alpha}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{v} \geq g(\boldsymbol{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq g^{\boldsymbol{a}}(\mathbf{0}; \boldsymbol{\alpha}, \boldsymbol{\beta}) - \rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} \bar{z} \tag{6.8}$$

is valid. We adopt the convention that, if (6.5) does not hold, then $g(\boldsymbol{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\infty$ and $\rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} = \infty$, in which case (6.8) holds for any $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$.

We can strengthen inequality (6.8) by finding the best $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$, i.e.,

$$s \geq \max_{(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)} \left\{ g^{\boldsymbol{a}}(\mathbf{0}; \boldsymbol{\alpha}, \boldsymbol{\beta}) - \rho^{\boldsymbol{a}}_{\boldsymbol{\alpha},\gamma} \bar{z} + \boldsymbol{\alpha}'\boldsymbol{x} + \boldsymbol{\beta}'\boldsymbol{v} \right\} \stackrel{\text{def}}{=} \bar{f}^{\boldsymbol{a}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}). \tag{6.9}$$

Obviously, optimal solutions $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ of (6.9) satisfy (6.5). As stated in Proposition 6.4, the valid inequality (6.9) is in fact *ideal*.

PROPOSITION 6.4. *Function* $\bar{f}^{\boldsymbol{a}}$ *and bound constraints describe the closure of the convex hull of* X,

$$cl\ conv(X) = \left\{ (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}, s) \in \mathbb{R}^2 \times [0,1]^2 \times \mathbb{R}^2 \times \mathbb{R} : \bar{f}^{\boldsymbol{a}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) \leq s \right\}.$$

Proof. The results follows from Theorem 1 in [62]. See also [39] for a similar result specialized to submodular functions. ∎

Therefore, in light of Proposition 6.4, it suffices to find the explicit form of $\bar{f}^{\boldsymbol{a}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v})$ to complete the proof of Theorem 4.2.

PROPOSITION 6.5. *Let functions* $\nu_1, \nu_2 : \mathbb{R}^3 \to \mathbb{R}$ *and* $\bar{\zeta} : \mathbb{R}^2_+ \to \mathbb{R}$ *be as described in Theorem 4.2. Then*

$$\bar{f}^{\boldsymbol{a}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) = \frac{a_1 \nu_1(x_1, v_1, v_2)^2 + a_2 \nu_2(x_2, v_1, v_2)^2 + (v_1 - v_2)^2}{2} + a_1 a_2 \frac{(v_1 - v_2)^2}{2\bar{a}\bar{\zeta}(z_1, z_2)}.$$

Proof. We now solve for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ in (6.9). We can write (6.9) explicitly (after

substituting $\boldsymbol{\beta}$) as

$$
\begin{aligned}
(6.10)\qquad & s \geq \max_{\boldsymbol{\alpha},\gamma} -\frac{1}{2}\frac{(\alpha_1+\alpha_2)^2 + a_2\alpha_1^2 + a_1\alpha_2^2}{\bar{a}} - \frac{\Big(\bar{a}\gamma - (a_2\alpha_1 - a_1\alpha_2)\Big)^2}{2a_1a_2\bar{a}}\bar{z} \\
& + \alpha_1 x_1 + \alpha_2 x_2 + (\gamma-\alpha_1)v_1 - (\gamma+\alpha_2)v_2 \\
(6.11)\quad & \Leftrightarrow s \geq \max_{\boldsymbol{\alpha},\gamma} -\frac{(a_1a_2 + a_1a_2^2)\alpha_1^2 + (a_1a_2 + a_1^2a_2)\alpha_2^2 + 2a_1a_2\alpha_1\alpha_2}{2a_1a_2\bar{a}} \\
& + \frac{a_2^2\bar{z}\alpha_1^2 + a_1^2\bar{z}\alpha_2^2 - 2a_1a_2\bar{z}\alpha_1\alpha_2}{2a_1a_2\bar{a}} - \frac{\bar{a}^2\bar{z}\gamma^2 - 2\bar{a}\bar{z}\gamma(a_2\alpha_1 - a_1\alpha_2)}{2a_1a_2\bar{a}} \\
& + (x_1-v_1)\alpha_1 + (x_2-v_2)\alpha_2 + (v_1-v_2)\gamma \\
(6.12)\quad & \Leftrightarrow s \geq \max_{\boldsymbol{\alpha},\gamma} -\frac{(a_1a_2 + a_1a_2^2 + a_2^2\bar{z})\alpha_1^2 + (a_1a_2 + a_1^2a_2 + a_1^2\bar{z})\alpha_2^2 + 2a_1a_2(1-\bar{z})\alpha_1\alpha_2}{2a_1a_2\bar{a}} \\
& - \frac{\bar{a}\bar{z}}{2a_1a_2}\gamma^2 + (x_1 - v_1 + \frac{\bar{z}\gamma}{a_1})\alpha_1 + (x_2 - v_2 - \frac{\bar{z}\gamma}{a_2})\alpha_2 + (v_1-v_2)\gamma \\
(6.13)\quad & \Leftrightarrow s \geq \max_{\boldsymbol{\alpha},\gamma} -\frac{a_2}{2\bar{a}}\alpha_1^2 - \frac{a_1}{2\bar{a}}\alpha_2^2 - \frac{1}{2\bar{a}}(\alpha_1+\alpha_2)^2 - \frac{\bar{z}}{2a_1a_2\bar{a}}(a_2\alpha_1 - a_1\alpha_2)^2 \\
& - \frac{\bar{a}\bar{z}}{2a_1a_2}\gamma^2 + (x_1 - v_1 + \frac{\bar{z}\gamma}{a_1})\alpha_1 + (x_2 - v_2 - \frac{\bar{z}\gamma}{a_2})\alpha_2 + (v_1-v_2)\gamma \\
(6.14)\quad & \Leftrightarrow s \geq \max_{\boldsymbol{\alpha}} -\frac{a_2}{2\bar{a}}\alpha_1^2 - \frac{a_1}{2\bar{a}}\alpha_2^2 - \frac{1}{2\bar{a}}(\alpha_1+\alpha_2)^2 - \frac{\bar{z}}{2a_1a_2\bar{a}}(a_2\alpha_1 - a_1\alpha_2)^2 \\
& + (x_1-v_1)\alpha_1 + (x_2-v_2)\alpha_2 + \frac{a_1a_2}{2\bar{a}z}\left(\frac{\bar{z}}{a_1}\alpha_1 - \frac{\bar{z}}{a_2}\alpha_2 + (v_1-v_2)\right)^2 \\
(6.15)\quad & \Leftrightarrow s \geq \max_{\boldsymbol{\alpha}} -\frac{1}{2\bar{a}}\Big((a_2+1)\alpha_1^2 + \alpha_1\alpha_2 + (a_1+1)\alpha_2^2\Big) + \Big(x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1-v_2)\Big)\alpha_1 \\
& + \Big(x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1-v_2)\Big)\alpha_2 + a_1a_2\frac{(v_1-v_2)^2}{2\bar{a}\bar{z}}.
\end{aligned}
$$

Inequality (6.11) is obtained by setting the denominator $a_1a_2\bar{a}$ on the first ratio and expanding quadratic terms; inequality (6.12) is obtained by rearranging terms, and noting that $\frac{\bar{a}^2\bar{z}\gamma^2 - 2\bar{a}\bar{z}\gamma(a_2\alpha_1 - a_1\alpha_2)}{2a_1a_2\bar{a}} = \frac{\bar{a}\bar{z}}{2a_1a_2}\gamma^2 - \frac{\bar{z}\gamma}{a_1}\alpha_1 + \frac{\bar{z}\gamma}{a_2}\alpha_2$; inequality (6.13) is obtained by grouping together terms that depend on $\bar{z}$, and simplifying the resulting expression; inequality (6.14) is obtained by noting that $\gamma^* = \big(\frac{\bar{z}}{a_1}\alpha_1 - \frac{\bar{z}}{a_2}\alpha_2 + (v_1-v_2)\big)/\big(\frac{\bar{a}\bar{z}}{a_1a_2}\big)$ in any optimal solution (which can be checked by taking derivatives with respect to γ and setting to 0); finally, inequality (6.15) is obtained by expanding the last quadratic term and regrouping terms.

Observe that optimization problem (6.15) is of the form

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha'Q\alpha} + \boldsymbol{w'\alpha}$$

where $w_1 = x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1-v_2)$, $w_2 = x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1-v_2)$ and $Q = \frac{1}{\bar{a}}\begin{pmatrix} a_2+1 & 1 \\ 1 & a_1+1 \end{pmatrix}$. Standard quadratic optimization techniques yield the optimal objective value $\frac{1}{2}\boldsymbol{w'Q^{-1}w}$, i.e., in the original space of variables:

$$
\begin{aligned}
s \geq & \frac{1}{2}\begin{pmatrix} x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1-v_2) & x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1-v_2)\end{pmatrix}\begin{pmatrix} a_1+1 & -1 \\ -1 & a_2+1\end{pmatrix}\begin{pmatrix} x_1 - v_1 + \frac{a_2}{\bar{a}}(v_1-v_2) \\ x_2 - v_2 - \frac{a_1}{\bar{a}}(v_1-v_2)\end{pmatrix} \\
& + a_1a_2\frac{(v_1-v_2)^2}{2\bar{a}\min\{1, z_1+z_2\}}.
\end{aligned}
$$

This is precisely the form given in Theorem 4.2 and Proposition 6.5 (in an extended formulation), completing the proof. ∎

7. Conclusions. In this paper we study the computation of robust estimates of a Markov process in the presence of outliers. The estimation problem is closely related to the Trimmed Least Squares procedure or the leave-k-out diagnostics from the statistical literature. We study the convexification of the natural mixed-integer quadratic optimization formulation of this problem, and derive a novel mixed-integer conic quadratic formulation via lifting. Using the new formulations with off-the-shelf solvers results in orders-of-magnitude improvements in terms of solution times or end gaps. Our computations also indicate that estimators obtained from solving just the convex relaxation of the proposed formulation, which can be done very quickly with interior point methods, results in good statistical performance, and in some cases improving upon the performance of solving the mixed-integer optimization problem to optimality.

REFERENCES

[1] *NEOS server for optimization*. https://neos-server.org/neos/. Accessed: 2019-11-19.

[2] B. Abraham and G. E. Box, *Bayesian analysis of some outlier problems in time series*, Biometrika, 66 (1979), pp. 229–236.

[3] G. Agamennoni, J. I. Nieto, and E. M. Nebot, *An outlier-robust kalman filter*, in 2011 IEEE International Conference on Robotics and Automation, IEEE, 2011, pp. 1551–1558.

[4] J. Agulló, *New algorithms for computing the least trimmed squares regression estimator*, Computational Statistics & Data Analysis, 36 (2001), pp. 425–439.

[5] H. N. Akouemo and R. J. Povinelli, *Probabilistic anomaly detection in natural gas time series data*, International Journal of Forecasting, 32 (2016), pp. 948–956.

[6] M. S. Aktürk, A. Atamtürk, and S. Gürel, *A strong conic quadratic reformulation for machine-job assignment with controllable processing times*, Operations Research Letters, 37 (2009), pp. 187–191.

[7] A. Atamtürk and A. Gómez, *Strong formulations for quadratic optimization with m-matrices and indicator variables*, Mathematical Programming, 170 (2018), pp. 141–176.

[8] A. Atamtürk and A. Gómez, *Rank-one convexification for sparse regression*, 2nd round revision at the Journal of Machine Learning Research, (2019).

[9] A. Atamtürk and A. Gómez, *Supermodularity and valid inequalities for quadratic optimization with indicators*, Submitted to Mathematical Programming, (2021).

[10] T. Bernholt, *Robust estimators are hard to compute*, tech. report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in ..., 2006.

[11] D. Bertsimas, A. King, R. Mazumder, et al., *Best subset selection via a modern optimization lens*, The Annals of Statistics, 44 (2016), pp. 813–852.

[12] D. Bertsimas, R. Mazumder, et al., *Least quantile regression via modern optimization*, The Annals of Statistics, 42 (2014), pp. 2494–2525.

[13] D. Bienstock, *Computational study of a family of mixed-integer quadratic programming problems*, Mathematical Programming, 74 (1996), pp. 121–140.

[14] D. Bienstock and A. Michalka, *Cutting-planes for optimization of convex functions over nonconvex sets*, SIAM Journal on Optimization, 24 (2014), pp. 643–677.

[15] R. E. Bixby, *A brief history of linear and mixed-integer programming computation*, Documenta Mathematica, (2012), pp. 107–121.

[16] P. Bonami, A. Lodi, A. Tramontani, and S. Wiese, *On mathematical programming with indicator constraints*, Mathematical Programming, 151 (2015), pp. 191–223.

[17] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: Forecasting and control*, John Wiley & Sons, 2015.

[18] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*, springer, 2016.

[19] R. G. Brown, P. Y. Hwang, et al., *Introduction to random signals and applied Kalman filtering*, vol. 3, Wiley New York, 1992.

[20] R. G. Brown and R. F. Meyer, *The fundamental theorem of exponential smoothing*, Operations Research, 9 (1961), pp. 673–685.

[21] A. G. Bruce and R. D. Martin, *Leave-k-out diagnostics for time series*, Journal of the Royal Statistical Society: Series B (Methodological), 51 (1989), pp. 363–401.

[22] S. Ceria and J. Soares, *Convex programming for disjunctive convex optimization*, Mathematical Programming, 86 (1999), pp. 595–614.

[23] I. Chang, G. C. Tiao, and C. Chen, *Estimation of time series parameters in the presence of outliers*, Technometrics, 30 (1988), pp. 193–204.
[24] C. Chatfield, *The holt-winters forecasting procedure*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 27 (1978), pp. 264–279.
[25] C. Chen and L.-M. Liu, *Joint estimation of model parameters and outlier effects in time series*, Journal of the American Statistical Association, 88 (1993), pp. 284–297.
[26] C. Dellacherie, S. Martinez, and J. S. Martin, *The class of inverse M-matrices associated to random walks*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 831–854.
[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), 39 (1977), pp. 1–22.
[28] H. Dong, K. Chen, and J. Linderoth, *Regularization vs. relaxation: A conic optimization perspective of statistical variable selection*, arXiv preprint arXiv:1510.06083, (2015).
[29] H. Dong and J. Linderoth, *On valid inequalities for quadratic programming with continuous variables and binary indicators*, in International Conference on Integer Programming and Combinatorial Optimization, Springer, 2013, pp. 169–180.
[30] P. H. Eilers and R. X. De Menezes, *Quantile smoothing of array CGH data*, Bioinformatics, 21 (2004), pp. 1146–1153.
[31] A. J. Fox, *Outliers in time series*, Journal of the Royal Statistical Society: Series B (Methodological), 34 (1972), pp. 350–363.
[32] A. Frangioni, F. Furini, and C. Gentile, *Approximated perspective relaxations: a project and lift approach*, Computational Optimization and Applications, 63 (2016), pp. 705–735.
[33] A. Frangioni and C. Gentile, *Perspective cuts for a class of convex 0–1 mixed integer programs*, Mathematical Programming, 106 (2006), pp. 225–236.
[34] A. Frangioni and C. Gentile, *SDP diagonalizations and perspective cuts for a class of nonseparable MIQP*, Operations Research Letters, 35 (2007), pp. 181–185.
[35] A. Frangioni, C. Gentile, E. Grande, and A. Pacifici, *Projected perspective reformulations with applications in design problems*, Operations research, 59 (2011), pp. 1225–1232.
[36] A. Frangioni, C. Gentile, and J. Hungerford, *Decompositions of semidefinite matrices and the perspective reformulation of nonseparable quadratic programs*, Mathematics of Operations Research, (2019).
[37] E. S. Gardner Jr, *Exponential smoothing: The state of the art*, Journal of Forecasting, 4 (1985), pp. 1–28.
[38] A. Giloni and M. Padberg, *Least trimmed squares regression, least median squares regression, and mathematical programming*, Mathematical and Computer Modeling, 35 (2002), pp. 1043–1060.
[39] A. Gómez, *Submodularity and valid inequalities in nonlinear optimization with indicator variables*, http://www.optimization-online.org/DB_HTML/2018/11/6925.html, (2018).
[40] O. Günlük and J. Linderoth, *Perspective reformulations of mixed integer nonlinear programs with indicator variables*, Mathematical Programming, 124 (2010), pp. 183–205.
[41] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, *Outlier detection for temporal data: A survey*, IEEE Transactions on Knowledge and Data Engineering, 26 (2013), pp. 2250–2267.
[42] F. R. Hampel, *A general qualitative definition of robustness*, The Annals of Mathematical Statistics, (1971), pp. 1887–1896.
[43] H. Hijazi, P. Bonami, G. Cornuéjols, and A. Ouorou, *Mixed-integer nonlinear programs featuring "on/off" constraints*, Computational Optimization and Applications, 52 (2012), pp. 537–558.
[44] S. C. Hillmer, W. R. Bell, and G. C. Tiao, *Modeling considerations in the seasonal adjustment of economic time series*, Applied Time Series Analysis of Economic Data, (1983), pp. 74–100.
[45] D. S. Hochbaum, *An efficient algorithm for image segmentation, Markov random fields and related problems*, Journal of the ACM (JACM), 48 (2001), pp. 686–701.
[46] D. S. Hochbaum, *Multi-label Markov random fields as an efficient and effective tool for image segmentation, total variations and regularization*, Numerical Mathematics: Theory, Methods and Applications, 6 (2013), pp. 169–198.
[47] D. S. Hochbaum and C. Lu, *A faster algorithm solving a generalization of isotonic median regression and a class of fused lasso problems*, SIAM Journal on Optimization, 27 (2017), pp. 2563–2596.
[48] H. Jeon, J. Linderoth, and A. Miller, *Quadratic cone cutting surfaces for quadratic programs with on–off constraints*, Discrete Optimization, 24 (2017), pp. 32–50.
[49] S. Jewell and D. Witten, *Exact spike train inference via ℓ_0 optimization*, The Annals of

Applied Statistics, 12 (2018), p. 2457.
[50] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering, 82 (1960), pp. 35–45.
[51] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Journal of Basic Engineering, 83 (1961), pp. 95–108.
[52] M. Kumar, D. P. Garg, and R. A. Zachery, *A method for judicious fusion of inconsistent multiple sensor data*, IEEE Sensors Journal, 7 (2007), pp. 723–733.
[53] G. M. Ljung, *On outlier detection in time series*, Journal of the Royal Statistical Society: Series B (Methodological), 55 (1993), pp. 559–567.
[54] H. Manzour, S. Küçükyavuz, and A. Shojaie, *Integer programming for learning directed acyclic graphs from continuous data*, arXiv preprint arXiv:1904.10574, (2019).
[55] K. G. Mehrotra, C. K. Mohan, and H. Huang, *Anomaly detection principles and algorithms*, Springer, 2017.
[56] H. Moonesignhe and P.-N. Tan, *Outlier detection using random walks*, in 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), IEEE, 2006, pp. 532–539.
[57] D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, *On the least trimmed squares estimator*, Algorithmica, 69 (2014), pp. 148–183.
[58] J. F. Muth, *Optimal properties of exponentially weighted forecasts*, Journal of the American Statistical Association, 55 (1960), pp. 299–306.
[59] T. T. Nguyen, J.-P. P. Richard, and M. Tawarmalani, *Deriving convex hulls through lifting and projection*, Mathematical Programming, 169 (2018), pp. 377–415.
[60] D. Peña, *Measuring the importance of outliers in ARIMA models*, New Perspectives in Theoretical and Applied Statistics, (1987), pp. 109–118.
[61] D. Peña, *Outliers, influential observations, and missing data*, A course in Time Series Analysis, (2000), pp. 136–170.
[62] J.-P. P. Richard and M. Tawarmalani, *Lifting inequalities: a framework for generating strong cuts for nonlinear programs*, Mathematical Programming, 121 (2010), pp. 61–104.
[63] A. Rinaldo et al., *Properties and refinements of the fused lasso*, The Annals of Statistics, 37 (2009), pp. 2922–2952.
[64] S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow, *Stochastic processes*, vol. 2, Wiley New York, 1996.
[65] P. J. Rousseeuw, *Least median of squares regression*, Journal of the American Statistical Association, 79 (1984), pp. 871–880.
[66] P. J. Rousseeuw and M. Hubert, *Recent developments in PROGRESS*, Lecture Notes-Monograph Series, (1997), pp. 201–214.
[67] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, vol. 589, John Wiley & Sons, 1987.
[68] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 91–108.
[69] R. S. Tsay, *Time series model specification in the presence of outliers*, Journal of the American Statistical Association, 81 (1986), pp. 132–141.
[70] R. S. Tsay, *Outliers, level shifts, and variance changes in time series*, Journal of Forecasting, 7 (1988), pp. 1–20.
[71] E. T. Whittaker, *On a new method of graduation*, Proceedings of the Edinburgh Mathematical Society, 41 (1922), pp. 63–75.
[72] P. R. Winters, *Forecasting sales by exponentially weighted moving averages*, Management Science, 6 (1960), pp. 324–342.
[73] L. A. Wolsey and G. L. Nemhauser, *Integer and combinatorial optimization*, John Wiley & Sons, 2014.
[74] B. Wu, X. Sun, D. Li, and X. Zheng, *Quadratic convex reformulations for semicontinuous quadratic programming*, SIAM Journal on Optimization, 27 (2017), pp. 1531–1553.
[75] W. Xie and X. Deng, *Scalable algorithms for the sparse ridge regression*, (2020).
[76] X. Zheng, X. Sun, and D. Li, *Improving the performance of MIQP solvers for quadratic programs with cardinality and minimum threshold constraints: A semidefinite program approach*, INFORMS Journal on Computing, 26 (2014), pp. 690–703.
[77] G. Zioutas and A. Avramidis, *Deleting outliers in robust regression with mixed integer programming*, Acta Mathematicae Applicatae Sinica, 21 (2005), pp. 323–334.
[78] G. Zioutas, L. Pitsoulis, and A. Avramidis, *Quadratic mixed integer programming and support vectors for deleting outliers in robust regression*, Annals of Operations Research,

166 (2009), pp. 339–353.

Appendix A. Modeling using mixed-integer optimization.

We now discuss how several priors on the structure of the outliers, as well a variants of (3.4) of practical interest, can be naturally modeled using mixed-integer optimization.

A.1. Outlier density. Outliers can be clustered together or be isolated, in which case (depending on the application) they may have different interpretations. For example, when monitoring the outcomes of an experiment over time, an isolated outlier may correspond to a measurement error (due to a faulty instrument) or even to a transcription error; in contrast, a cluster of outliers may corresponding to a physical phenomenon (e.g., change of temperature) or an external influence that abruptly changes the conditions of the experiment. Cluster of outliers are often more difficult to identify [56], as illustrated also in Figure 3 (b).

A decision-maker may wish to remove isolated outliers (as they are likely to be incorrect measurements) but keep cluster of outliers (as they may represent a phenomenon of interest). To accomplish this goal, given a density parameter $b \in \mathbb{Z}_+$, low-density constraints

$$\sum_{i=k}^{k+b} z_i \leq 1 \quad \text{for } k = 1, \dots, n-b$$

can be imposed, which state that outliers need to be spaced by at least b points.

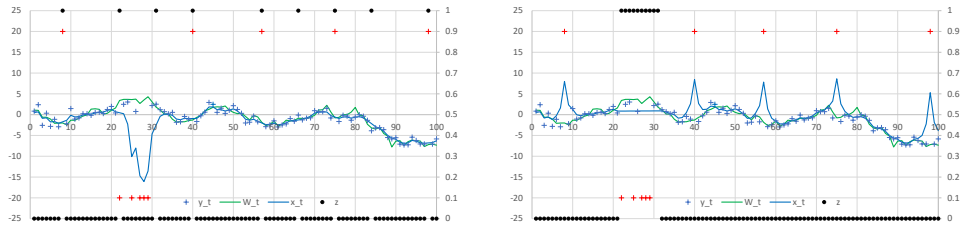
Alternatively, a decision-maker may wish to focus on detecting clusters of outliers, to identify the relevant phenomena. In this case, given a density parameter $b \in \mathbb{Z}_+$, high-density constraints

$$\sum_{i=\max\{k-b,1\}}^{\min\{k+b,n\}} z_i \geq (b+1) z_k \quad \text{for } k = 1, \dots, n$$

can be imposed, which state that outliers occur in clusters of at least $b+1$ points.

Example. Figure 12 depicts the MAP estimators with a cardinality constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$ and low/high density constraints with $b = 8$. We also use black dots (corresponding to the secondary axis) to represent the optimal values of $\boldsymbol{z}$. Note that the cardinality parameter is (purposely) miss-specified, as there are only five isolated/clustered outliers. We see that in both cases the density constraints achieve their goal of focusing on isolated/clustered outliers. An although in both cases additional outliers are incorrectly detected (due the miss-specification), the estimation corresponding to those points is not significantly affected. This examples illustrates how additional priors help the inference process and reduce the effect of choosing the "wrong" parameters.

A.2. Anomaly sparsity. As mentioned in §A.1, a cluster of outliers may be indicative of a single anomaly disrupting the process. In such cases it can be natural to assume that, in addition of the quantity of corrupted data being small, the quantity of such anomalies is also small. Such constraints can be enforced by restricting the number of times that the signal can move between anomalous states and normal



(a) Low density. (b) High density.

Fig. 12: MAP inference of the Wiener process with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$ and low/high density constraints with density parameter $b = 8$. Black dots (corresponding to the secondary axis) depict the optimal values of $\boldsymbol{z}$.

states, i.e., given a sparsity parameter $b \in \mathbb{Z}_+$,

$$\sum_{i=1}^{n-1} |z_{i+1} - z_i| \leq b. \tag{A.1}$$

Observe that constraints (A.1) have a structure similar to the fused lasso [63, 68]. However, unlike fused lasso constraints, (A.1) restrict changes in the indicator variables $\boldsymbol{z}$ instead of the continuous variables $\boldsymbol{x}$. Constraints (A.1) were previously used in signal estimation problems in [**?**].

Example. If, in addition to the cardinality constraint $\sum_{i \in N} z_i \leq 10$, an anomaly constraint (A.1) is added with $b = 2$, then the results are identical to those in Figure 12 (b). Figure 13 depicts the MAP estimators for $b = 4$ and $b = 6$ (no density constraints are used). The constraints indeed restrict the number of anomalies detected, prioritizing those corresponding to a larger number of points.

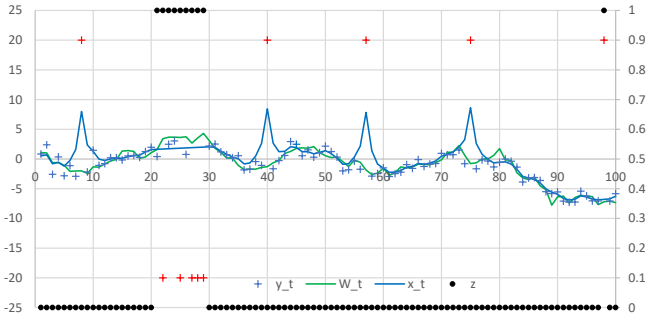
A.3. Forecasting. Thus far, we were concerned on obtained high-quality of true values of W_t using all the observations. In many applications, however, the decision-maker is interested in predicting the value of y_i using only past information $y_1, \ldots, y_{i-1}$. Note that outliers cannot be accurately forecasted (as they are not assumed to follow any particular distribution); nonetheless, one would be interested in how to deal with identify/deal with outlier data to improve the overall forecasts of non-outlier data.

Observe that if there are no outliers, $t_i = i$ for all i, $\boldsymbol{\mu} = 0$, and the noise is i.i.d, then *exponential smoothing* forecasts are ideal to estimate y_t [58]. In an exponential smoothing forecast, each prediction x_i is a weighted average of all past values $y_1, \ldots, y_{i-1}$, where recent observations carry exponentially more weight. In particular, given a smoothing factor $0 \leq b \leq 1$, the forecasts are a convex combination of the most recent forecast and observation, $x_i = by_{i-1} + (1-b)x_{i-1}$ with $x_0 = 0$.

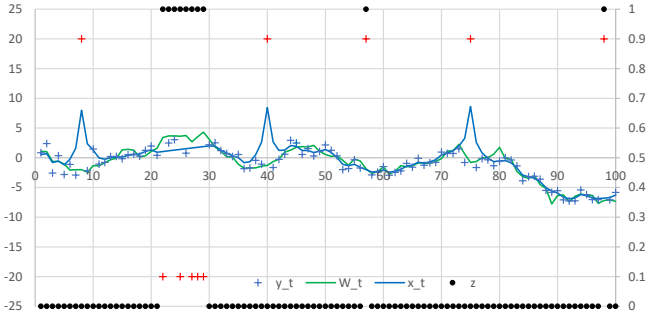
Thus, in order to remove outliers in a time series to improve forecasts, we propose to solve (3.4) with the additional constraint that the estimates $\boldsymbol{x}$ are obtained from an exponential smoothing forecast by adding the smoothing constraints

$$x_i = b(y_{i-1} + v_{i-1} - \mu_{i-1}) + (1-b)x_{t-1}, \quad i = 2, \ldots, n. \tag{A.2}$$

If constraints (A.2) are imposed, then the equality $v_i = x_i + \mu_i - y_i$ may not hold, thus in this model outliers are not discarded. Instead, the values of outlier data $i \in S$



(a) $\sum_{i=1}^{n-1} |z_{i+1} - z_i| \leq 4$.



(b) $\sum_{i=1}^{n-1} |z_{i+1} - z_i| \leq 6$.

Fig. 13: MAP inference of the Wiener process with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$ and anomaly sparsity constraints. Black dots (corresponding to the secondary axis) depict the optimal values of $\boldsymbol{z}$.

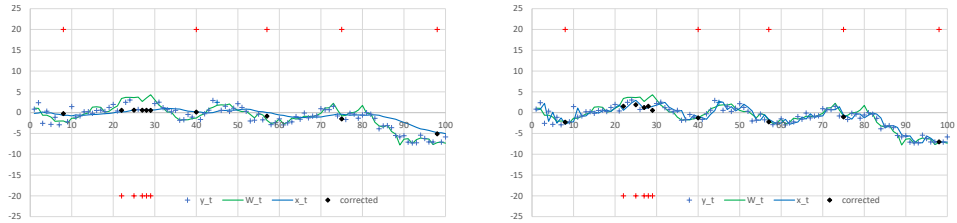
are "corrected" to a value $\bar{y}_i = y_i + v_i$ that results in the best overall MAP estimates of W_t achievable by a exponential smoothing method with parameter b and initial forecast x_1. We point out that other usual techniques such as *moving averages* or *additive seasonality* can be easily modeled as well.

Example. Figure 14 depicts the estimators with a cardinality constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$ and smoothing constraints (A.2). We also plot using black squares the corrected values $\bar{\boldsymbol{y}}$. We observe that the corrected values are not necessarily on the curve of the MAP estimator (as they account for the effect on future predictions as well) and that the corrected values depend on the smoothing parameter b.

A.4. Additional variants. Several other considerations can be modeled via minor modifications of (3.4). For example, to estimate a Wiener process $\bar{W}_t = \bar{\mu}t + \bar{\sigma}W_t$ with drift $\bar{\mu}$ and infinitesimal variance $\bar{\sigma}^2$, it suffices to update the objective value of (3.4) to

$$\frac{x_1^2}{2t_1} + \sum_{i=1}^{n-1} \frac{\Big(x_{i+1} - x_i - \bar{\mu}(t_{i+1} - t_i)\Big)^2}{2\bar{\sigma}^2(t_{i+1} - t_i)} + \sum_{i=1}^{n} \frac{(y_i + v_i - \mu_i - x_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n} \frac{\ln(2\pi\sigma_i^2)}{2}(1 - z_i).$$

The formulations can also be adapted to pairwise MRF (not necessarily one-



(a) Smoothing parameter $b = 0.1$. (b) Smoothing parameter $b = 0.9$.

Fig. 14: MAP inference of the Wiener process with constraint $\boldsymbol{e}'\boldsymbol{z} \leq 10$ and smoothing constraints. Black squares depict the corrected values $\bar{\boldsymbol{y}}$.

dimensional) estimation problems with outliers as

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}} \ & \frac{1}{2} \sum_{(i,j)\in E} c_{ij}(x_j - x_i)^2 + \frac{1}{2} \sum_{i \in N} d_i(y_i + v_i - x_i)^2 \\ \text{s.t.} \ & -M\boldsymbol{z} \leq \boldsymbol{v} \leq M\boldsymbol{z} \\ & \boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{z} \in Z, \boldsymbol{v} \in \mathbb{R}^N \end{aligned}$$

for some separation coefficients $\boldsymbol{c}$ and deviations coefficients $\boldsymbol{d}$, where E denotes the set of adjacent nodes, see [45, 46] for a detailed description of pairwise MRFs.

Several other additional constraints can be added to (3.4). For example, constraints $\boldsymbol{v} \geq 0$ (or $\boldsymbol{v} \leq 0$) can be used to incorporate the prior that outliers consistently underestimate (or overestimate) the true process W_t. A "fused lasso" constraint $\sum_{i=1}^{n-1} |v_{i+1} - v_i| \leq b$ indicates that outliers are introduced smoothly into the process. Constraints $|v_i| \leq bx_i$ indicate that the maximum perturbation induce by an outlier is bounded by the true value of the signal at that point. Constraints $x_i \leq x_{i+1}$ for all $i = 1, \ldots, n-1$ can be used for *isotonic* estimation [47] with outliers .