# OPTIMALLY MONITORING A NETWORK OF SEWAGE MANHOLES IN INFECTIOUS DISEASE SURVEILLANCE

Leyao Zhang, Yahui Zhang, Chuanwu Xi, Peter X.-K. Song

*Department of Biostatistics, University of Michigan*

*Department of Environmental Health Sciences, University of Michigan*

**Supplementary Material**

This document of Supplementary Materials include (A) the technical proof of the selection consistency of the MIO estimators; (B) software implementation details; (C) additional simulation results and (D) additional data analysis results.

## S1 Proof of Selection Consistency of the MIO estimators

For the brevity of the proof, without the loss of generality, we only consider the case when $P = 1$ and $Q = 1$, thus we have $max(P, Q) = 1$. The same statistical properties can be proved to hold for the case with $P > 1$ or $Q > 1$. We begin with notations and conditions necessitating a rigorous

presentation of the selection consistency for the MIO estimator introduced in Section 3. We write $\mathbf{y} = (y(2), \ldots, y(T))^\top$, $\mathbf{Z} = (\mathbf{1}, (y(1), \ldots, y(T))^\top)$ and the $(T-1) \times m$ design matrix $\mathbf{G} = (\mathbf{G}_w(1), \ldots, \mathbf{G}_w(T-1))^\top$, where $\mathbf{G}_w(t) = (w_1 G_1(t), \ldots, w_m G_m(t))$. In the case when $P = 1$ and $Q = 1$, we write $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{1,m})^\top$, $\beta = \frac{\beta_1}{\sum_{j=1}^m \alpha_j w_j}$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\top$. Let $a_0$ denote the true value of $a$ (e.g. $\boldsymbol{\theta}_0$ denotes the true parameter of $\boldsymbol{\theta}$). Denote the error $\boldsymbol{\varepsilon} = (\varepsilon(2), \ldots, \varepsilon(T))^\top$.

The proof of Theorem 1 under the condition of sub-Gaussian white noise is given as follows. Let $\boldsymbol{\alpha}$ be any subset solution produced by the MIO Algorithm 1. Define $\mathbf{P}_{\boldsymbol{\alpha}}$ as the projection matrix of $(\mathbf{G}_w\boldsymbol{\alpha}, \mathbf{Z})$. In the case of $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}_0$, we have

$$\begin{aligned}
&\mathbb{P}\left( \min_{\substack{\boldsymbol{\theta}=\beta\boldsymbol{\alpha} \\ \boldsymbol{\gamma}}} \|\mathbf{y} - (\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top)^\top\|_2^2 < \|\mathbf{y} - (\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}^{(ol)\top}, \boldsymbol{\gamma}^{(ol)\top})^\top\|_2^2 \right) \\
=&\mathbb{P}\Big( 2\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top + \big\|(\mathbf{I} - \mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top\big\|_2^2 - \\
&\boldsymbol{\varepsilon}^\top (\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}_0})\boldsymbol{\varepsilon} < 0 \Big). \qquad \text{(S1.1)}
\end{aligned}$$

For any $0 < \delta < 1$, Equation (S1.1) may be bounded above by

$$\begin{aligned}
\leq &\mathbb{P}\left( 2\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top + \delta\big\|(\mathbf{I} - \mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top\big\|_2^2 < 0 \right) + \\
&\mathbb{P}\left( (1-\delta)\big\|(\mathbf{I} - \mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top\big\|_2^2 - \boldsymbol{\varepsilon}^\top (\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}_0})\boldsymbol{\varepsilon} < 0 \right).
\end{aligned}$$

For any $t_1 > 0$, $t_2 > 0$, using the Markov's inequality, we have an upper

bound for equation (S1.1):

$$\leq \mathbb{E}\left[\exp\left\{-\frac{2t_1\boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top}{\sigma^2}\right\}\right]\exp\left\{-\frac{t_1\delta\left\|(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\}+ \tag{S1.2}$$

$$\mathbb{E}\left[\exp\left\{\frac{t_2\boldsymbol{\varepsilon}^\top(\mathbf{P}_{\boldsymbol{\alpha}}-\mathbf{P}_{\boldsymbol{\alpha}_0})\boldsymbol{\varepsilon}}{\sigma^2}\right\}\right]\exp\left\{-\frac{t_2(1-\delta)\left\|(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\}. \tag{S1.3}$$

By the moment generating function of the sub-Gaussian white noise, the term in equation (S1.2) becomes

$$=\exp\left\{\frac{2t_1^2\left\|(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\}\exp\left\{-\frac{t_1\delta\left\|(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\},$$

$$\leq \exp\left\{\frac{2t_1^2-t_1\delta}{\sigma^2}(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)c_{\min}\right\}.$$

Applying the geometry interpretation of projection matrix, the term in equation (S1.3) is bounded above by

$$\leq \mathbb{E}\left[\exp\left\{\frac{t_2\boldsymbol{\varepsilon}^\top P_{\mathbf{G}_w\boldsymbol{\theta}}\,\boldsymbol{\varepsilon}}{\sigma^2}\right\}\right]\exp\left\{-\frac{t_2(1-\delta)\left\|(\mathbf{I}-\mathbf{P}_{\boldsymbol{\alpha}})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\},$$

$$=(1-2t_2)^{-1/2}\exp\left\{-\frac{t_2(1-\delta)\left\|(\mathbf{I}-\mathbf{P}_{\mathbb{G}(\boldsymbol{\beta})})(\mathbf{G}_w,\mathbf{Z})(\boldsymbol{\theta}_0^\top,\boldsymbol{\gamma}_0^\top)^\top\right\|_2^2}{\sigma^2}\right\},$$

where $\mathbf{P}_{\mathbf{G}_w\boldsymbol{\theta}}$ denotes the projection matrix of vector $\mathbf{G}_w\boldsymbol{\theta}$. Given the fact that $2t_2 \geq -\log(1-2t_2)/2$ for any $0 < t_2 < 0.398$, under a restriction of $t_2 \leq 0.398$, the term in equation (S1.3) may be further bounded from

above:

$$\leq \exp(2t_2) \exp\left\{ -\frac{t_2(1-\delta)\left\|(\mathbf{I}-\mathbf{P}_{\mathbb{G}(\boldsymbol{\beta})})(\mathbf{X},\mathbf{Z})\boldsymbol{\theta}_0\right\|_2^2}{\sigma^2} \right\}$$
$$\leq \exp\left\{ 2t_2 - \frac{t_2(1-\delta)}{\sigma^2}(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)c_{\min} \right\}.$$

Combining the two terms above, and setting $t_1 = \frac{1}{5}, t_2 = \frac{3}{8}, \delta = \frac{4}{5}$, we obtain an upper bound for equation (S1.1):

$$\leq \exp\left\{ \frac{2t_1^2 - t_1\delta}{\sigma^2}(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)c_{\min} \right\} + \exp\left\{ 2t_2 - \frac{t_2(1-\delta)}{\sigma^2}(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)c_{\min} \right\},$$
$$= 2\exp\left\{ -\frac{3}{40}\frac{(T-1)d(\boldsymbol{\alpha},\boldsymbol{\alpha}_0)c_{\min}}{\sigma^2} + \frac{3}{4} \right\}.$$

Finally, we bound the probability that the MIO estimator fails to identify the true subset as follows:

$$\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \tag{S1.4}$$
$$\leq \sum_{\boldsymbol{\alpha}\in\{0,1\}^p, \boldsymbol{\alpha}\neq\boldsymbol{\alpha}_0} \mathbb{P}\left( \min_{\substack{\boldsymbol{\theta}=\beta_1\boldsymbol{\alpha} \\ \boldsymbol{\gamma}}} \|\mathbf{Y} - (\mathbf{G}_w, \mathbf{Z})(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top)^\top\|_2^2 < \|\mathbf{Y} - (\mathbf{G}_w, \mathbf{Z})(\hat{\boldsymbol{\theta}}^{(ol)\top}, \hat{\boldsymbol{\gamma}}^{(ol)\top})^\top\|_2^2 \right),$$
$$\leq \sum_{i=1}^{m} \binom{m}{i} 2\exp\left( -\frac{3(T-1)ic_{\min}}{40\sigma^2} + \frac{3}{4} \right),$$
$$\leq \sum_{i=1}^{m} m^i\, 2\exp\left( -\frac{3(T-1)ic_{\min}}{40\sigma^2} + \frac{3}{4} \right),$$
$$\leq \sum_{i=1}^{m} 2\exp\left( -\frac{3(T-1)ic_{\min}}{40\sigma^2} + i\log(m) + \frac{3}{4} \right).$$

When $c_{\min} > \frac{40\sigma^2}{3(T-1)}\log(m)$, equation (S1.4) may be further bounded by:

$$\leq 2e^{3/4}\frac{\exp\left\{-\frac{3(T-1)}{40\sigma^2}\left(c_{\min}-\frac{40\sigma^2}{3(T-1)}\log(m)\right)\right\}}{1-\exp\left\{-\frac{3n}{40\sigma^2}\left(c_{\min}-\frac{40\sigma^2}{3(T-1)}\log(m)\right)\right\}}.$$

Since $\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \leq 1$ and $2e^{3/4}\frac{x}{1-x} \leq 6.5x$ when $0 \leq 2e^{3/4}\frac{x}{1-x} \leq 1$ and $0 < x < 1$, we obtain the upper bound for equation (S1.4) as follows:

$$\leq 6.5\exp\left\{-\frac{3(T-1)}{40\sigma^2}\left(c_{\min}-\frac{40\sigma^2}{3(T-1)}\log(m)\right)\right\}.$$

# S2 Software Implementation Details

We apply the GUROBI Optimizer (Gurobi Optimization, LLC, 2021) to solve the MIO defined by equations (2.2) - (2.5) in the main paper. Algorithm 1 lists the key steps required in the pseudo code that outputs the MIO solution to the constrained optimization problem. In both simulation studies and data analyses, we use GUROBI version 9.5.2 in all numerical calculations.

# S3 Additional Simulation Results

## S3.1 Normalized Mutual Information

NMI is used to evaluate the performance of binary label variable estimation. Let $L_{true}$ and $L_{pred}$ denote the true and predicted subgroup of the sampling

**Algorithm 1:** Pseudocode of the MIO via GUROBI Solver

**Input:** $\mathbf{y}, \mathbf{G}, \mathbf{w}, m$

**begin**

$\mathcal{M}$ = gurobipy.Model("Questionnaire") // Create a new model

$\mathcal{M}$.addVars($\boldsymbol{\alpha}$, vtype = GRB.BINARY) // Create Gurobi variables of $\boldsymbol{\alpha}$

$\mathcal{M}$.addVars($\boldsymbol{\theta}, \boldsymbol{\gamma}, \gamma_0, \boldsymbol{\beta}$, vtype = GRB.CONTINUOUS) // Create Gurobi variables of $\boldsymbol{\theta}, \boldsymbol{\gamma}, \gamma_0, \boldsymbol{\beta}$

$\mathcal{M}$.setObjective($\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \gamma_0, \boldsymbol{\beta})$) // Set objective function

**for** $j \leftarrow 1$ *to* $m$ **do**

  **for** $p \leftarrow 1$ *to* $P$ **do**

    $\mathcal{M}$.addSOS(GRB.SOS_TYPE1, $[1 - \alpha_j, \theta_{p,j}]$) // Add Type 1 SOS model constraints (2.4)

    $\mathcal{M}$.addSOS(GRB.SOS_TYPE1, $[\alpha_j, \theta_{p,j} - \beta_p]$) // Add Type 1 SOS model constraints (2.5)

  $\mathcal{M}$.optimize() // Solve the model

**Output:** $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\gamma}_0, \hat{\boldsymbol{\beta}}$

nodes into two disjoint subsets. The normalized mutual information is defined as

$$NMI(L_{true}, L_{pred}) = \frac{2 \times I(L_{true}, L_{pred})}{H(L_{true}) + H(L_{pred})},$$

where $H(.)$ is the entropy of estimated labels and $I(L_{true}, L_{pred})$ is the mutual information between $L_{true}$ and $L_{pred}$. Clearly NMI varies between 0 and 1, and the larger NMI the higher accuracy of cluster label estimation.

### S3.2 Sensitivity and specificity of selection.

Table 1 and 2 present a performance comparison of the MIO algorithm against established methods, namely LASSO and ABESS, with respect to sensitivity and specificity within the simulation model (5.8). Evidently, the MIO method demonstrates superior performance in both metrics across all scenarios, establishing itself as the optimal choice among the considered methods.

### S3.3 Estimation of the auto-regression coefficient

Table 3 compares the performance of the MIO algorithm with existing methods (i.e. LASSO and ABESS) on the estimation of the auto-regression coefficient $\gamma_1$ in the simulation model (5.8). Clearly, the MIO method shows the smallest estimation bias among these methods in all scenarios. The

Table 1: Average sensitivity of selecting important manholes over 1000 replicates ranged between 0 and 1, where $m$ is the network size, $s$ is the number of important manholes, and $\beta$ is the signal strength.

| | $m = 20$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 4$ | | $s = 10$ | | $s = 16$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 1.000 | 0.999 | 1.000 | 0.999 | 1.000 | 0.999 |
| LASSO | 0.785 | 0.873 | 0.533 | 0.778 | 0.064 | 0.546 |
| ABESS | 0.981 | 0.976 | 0.948 | 0.931 | 0.925 | 0.912 |
| | $m = 50$ | | | | | |
| | $s = 10$ | | $s = 25$ | | $s = 40$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 1.000 | 0.996 | 0.999 | 0.989 | 1.000 | 0.9695 |
| LASSO | 0.730 | 0.779 | 0.519 | 0.697 | 0.035 | 0.524 |
| ABESS | 0.994 | 0.921 | 0.951 | 0.637 | 0.946 | 0.617 |
| | $m = 100$ | | | | | |
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 1.000 | 0.996 | 1.000 | 0.978 | 1.000 | 0.996 |
| LASSO | 0.461 | 0.592 | 0.023 | 0.459 | 0.000 | 0.024 |
| ABESS | 0.988 | 0.722 | 0.583 | 0.316 | 0.442 | 0.317 |

ABESS estimation is the second best, but inferior over the MIO estimation in terms of estimation bias, especially when the network size and/or the number of important monitoring sites is large.

Table 2: Average specificity of selecting important manholes over 1000 replicates ranged between 0 and 1, where $m$ is the network size, $s$ is the number of important manholes, and $\beta$ is the signal strength.

| | $m = 20$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 4$ | | $s = 10$ | | $s = 16$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.995 | 0.994 | 0.995 | 0.994 | 0.995 | 0.995 |
| LASSO | 0.593 | 0.484 | 0.630 | 0.449 | 0.951 | 0.589 |
| ABESS | 0.992 | 0.991 | 0.993 | 0.991 | 0.992 | 0.990 |
| | $m = 50$ | | | | | |
| | $s = 10$ | | $s = 25$ | | $s = 40$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 1.000 | 0.995 | 0.999 | 0.986 | 1.000 | 0.995 |
| LASSO | 0.600 | 0.506 | 0.623 | 0.487 | 0.986 | 0.932 |
| ABESS | 0.999 | 0.994 | 0.992 | 0.955 | 0.971 | 0.573 |
| | $m = 100$ | | | | | |
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 0.999 | 0.995 | 0.999 | 0.975 | 0.998 | 0.994 |
| LASSO | 0.723 | 0.625 | 0.978 | 0.638 | 1.000 | 0.977 |
| ABESS | 0.993 | 0.967 | 0.898 | 0.926 | 0.857 | 0.893 |

## S3.4 Simulations with a Fixed Network

To evaluate the reproducibility of the model performance, we consider a new simulation experiment with a fixed network, where the number of significant

Table 3: Average absolute bias AAB($\hat{\gamma}_1$) and empirical standard error ESE($\hat{\gamma}_1$) over 1000 replicates, where LASSO and ABESS estimate group label by nonzero estimate. Here $m$ is the network size, $s$ is the number of important nodes, and $\beta$ is the effect of the network-level summary biomarker on the daily number of COVID-19 confirmed cases.

| | $m = 20$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 4$ | | $s = 10$ | | $s = 16$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.007 (0.009) | 0.015 (0.018) | 0.006 (0.007) | 0.011 (0.014) | 0.003 (0.004) | 0.007 (0.009) |
| LASSO | 0.403 (0.031) | 0.382 (0.033) | 0.340 (0.025) | 0.259 (0.029) | 0.642 (0.025) | 0.250 (0.029) |
| ABESS | 0.026 (0.010) | 0.033 (0.018) | 0.038 (0.008) | 0.046 (0.014) | 0.055 (0.008) | 0.058 (0.014) |
| | $m = 50$ | | | | | |
| | $s = 10$ | | $s = 25$ | | $s = 40$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.011 (0.014) | 0.020 (0.025) | 0.009 (0.011) | 0.017 (0.020) | 0.005 (0.006) | 0.010 (0.013) |
| LASSO | 0.341 (0.036) | 0.311 (0.038) | 0.301 (0.030) | 0.208 (0.035) | 0.668 (0.030) | 0.222 (0.035) |
| ABESS | 0.014 (0.014) | 0.037 (0.025) | 0.021 (0.011) | 0.158 (0.021) | 0.018 (0.011) | 0.175 (0.021) |
| | $m = 100$ | | | | | |
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 0.008 (0.010) | 0.017 (0.019) | 0.006 (0.007) | 0.014 (0.014) | 0.003 (0.004) | 0.007 (0.008) |
| LASSO | 0.357 (0.038) | 0.295 (0.041) | 0.676 (0.008) | 0.266 (0.035) | 0.698 (0.008) | 0.677 (0.035) |
| ABESS | 0.029 (0.012) | 0.143 (0.021) | 0.225 (0.015) | 0.343 (0.015) | 0.284 (0.015) | 0.347 (0.015) |

nodes and their positions in the network are fixed. In other words, we repeat our experiments on 1000 simulated gene datasets on one simulated network where the first $s$ nodes among the $m$ nodes are fixed as important sites. The

Table 4: Average Normalized Mutual Information (NMI) value over 1000 replicates on a fixed network ranged between 0 and 1 for the accuracy of selecting important manholes, where $m$ is the network size, $s$ is the number of important manholes, and $\beta$ is the signal strength.

| | $m = 20$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 4$ | | $s = 10$ | | $s = 16$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 1.000 | 0.999 | 1.000 | 0.997 | 1.000 | 0.998 |
| LASSO | 0.077 | 0.104 | 0.031 | 0.082 | 0.055 | 0.084 |
| ABESS | 0.997 | 0.986 | 0.999 | 0.987 | 0.999 | 0.991 |
| | $m = 50$ | | | | | |
| | $s = 10$ | | $s = 25$ | | $s = 40$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.999 | 0.971 | 0.994 | 0.940 | 0.996 | 0.959 |
| LASSO | 0.032 | 0.030 | 0.010 | 0.015 | 0.026 | 0.012 |
| ABESS | 0.963 | 0.739 | 0.770 | 0.362 | 0.694 | 0.208 |
| | $m = 100$ | | | | | |
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 0.999 | 0.987 | 0.999 | 0.956 | 0.998 | 0.987 |
| LASSO | 0.009 | 0.020 | 0.017 | 0.008 | 0.000 | 0.017 |
| ABESS | 0.964 | 0.539 | 0.249 | 0.098 | 0.070 | 0.050 |

results in Table 4 and 5 once again confirm that the MIO has the highest selection accuracy and parameter estimation accuracy.

Table 5: Average absolute bias AAB($\hat{\beta}$) and empirical standard error ESE($\hat{\beta}$) by the MIO estimation over 1000 replicates on a fixed network, where LASSO and ABESS estimate group label by nonzero estimate.

| | $m = 20$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 4$ | | $s = 10$ | | $s = 16$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.001 (0.001) | 0.001 (0.001) | 0.002 (0.002) | 0.002 (0.002) | 0.002 (0.002) | 0.002 (0.002) |
| LASSO | 0.130 (0.007) | 0.058 (0.004) | 0.103 (0.004) | 0.038 (0.007) | 0.377 (0.003) | 0.077 (0.006) |
| ABESS | 0.007 (0.002) | 0.004 (0.001) | 0.011 (0.002) | 0.007 (0.002) | 0.032 (0.003) | 0.017 (0.002) |
| | $m = 50$ | | | | | |
| | $s = 10$ | | $s = 25$ | | $s = 40$ | |
| | $\beta = 0.1$ | $\beta = 0.05$ | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ |
| MIO | 0.002 (0.002) | 0.002 (0.002) | 0.002 (0.003) | 0.003 (0.003) | 0.003 (0.004) | 0.003 (0.004) |
| LASSO | 0.050 (0.005) | 0.023 (0.003) | 0.092 (0.008) | 0.031 (0.005) | 0.388 (0.003) | 0.080 (0.007) |
| ABESS | 0.002 (0.002) | 0.003 (0.002) | 0.006 (0.003) | 0.024 (0.003) | 0.011 (0.004) | 0.053 (0.004) |
| | $m = 100$ | | | | | |
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 0.002 (0.003) | 0.002 (0.003) | 0.003 (0.004) | 0.004 (0.004) | 0.004 (0.005) | 0.004 (0.005) |
| LASSO | 0.109 (0.010) | 0.044 (0.006) | 0.392 (0.003) | 0.082 (0.010) | 0.800 (0.000) | 0.392 (0.002) |
| ABESS | 0.008 (0.003) | 0.021 (0.003) | 0.131 (0.009) | 0.100 (0.004) | 0.330 (0.015) | 0.201 (0.007) |

### S3.5 Simulations Confirming the Stability of GUROBI Software

GUROBI or most of available optimization solvers might not have a guaranteed solution of the global optimality. This practical challenge deserves

some further attention and effort in the field of operations research, which is beyond the scope of the current paper. However, to inspect and confirm if a solution provided by GUROBI is globally or locally optimal, numerically we may assign many random initial values in the search and observe stability of outputs from GUROBI. In the simulation experiment conducted in the paper, we tested the capacity of GUROBI in the most challenging scenario of a sewage network containing 100 manholes. Over 1000 replicates, each being fed with 100 distinct initial values of $\alpha$, we collect the GUROBI solutions. We summarize the results into the following table (Table 1) in terms of the average Normalized Mutual Information (NMI) across the 100 GUROBI solutions throughout the 1000 replicates. The results clearly confirmed the stability of GUROBI solutions, and thus such an optimization solver is highly reliable.

## S4 Additional Data Analysis Results

We use Shapiro Wilk normality test and quantile-quantile plot (Figure 1) to check the normality of model (6.9). The p-value of the Shapiro-Wilk test is $0.19 > 0.05$, which indicates that the residuals are normally distributed.

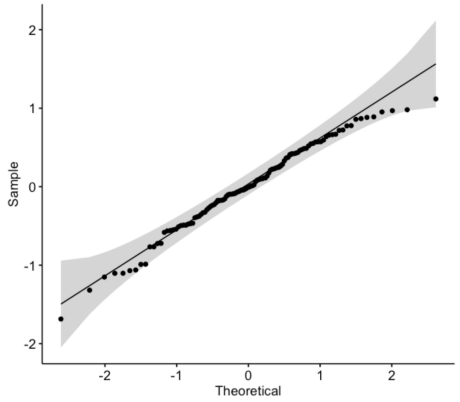


Figure 1: The Q-Q plot of the residuals in model (6.9).

Table 6: Average NMI calculated over the 100 GUROBI solutions with 100 different initial values throughout the 1000 replicates.

| | $m = 100$ | | | | | |
|---|---|---|---|---|---|---|
| | $s = 20$ | | $s = 50$ | | $s = 80$ | |
| | $\beta = 0.2$ | $\beta = 0.1$ | $\beta = 0.4$ | $\beta = 0.2$ | $\beta = 0.8$ | $\beta = 0.4$ |
| MIO | 0.996 | 0.959 | 0.998 | 0.869 | 0.998 | 0.962 |