

Imbalanced Mixed Linear Regression

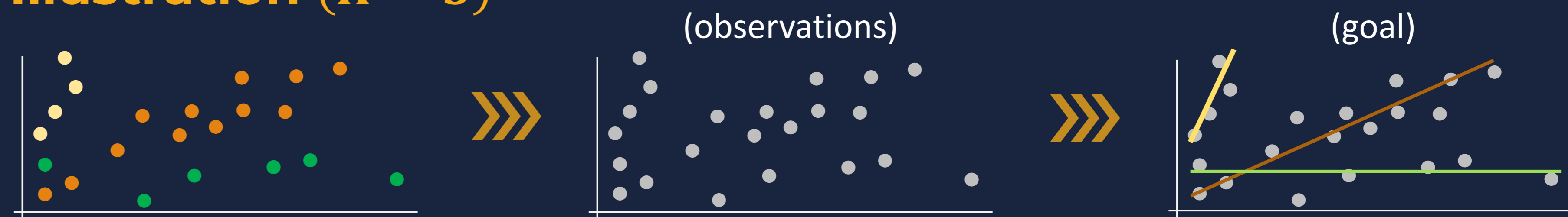
Pini Zilber, Boaz Nadler
Weizmann Institute of Science

Problem

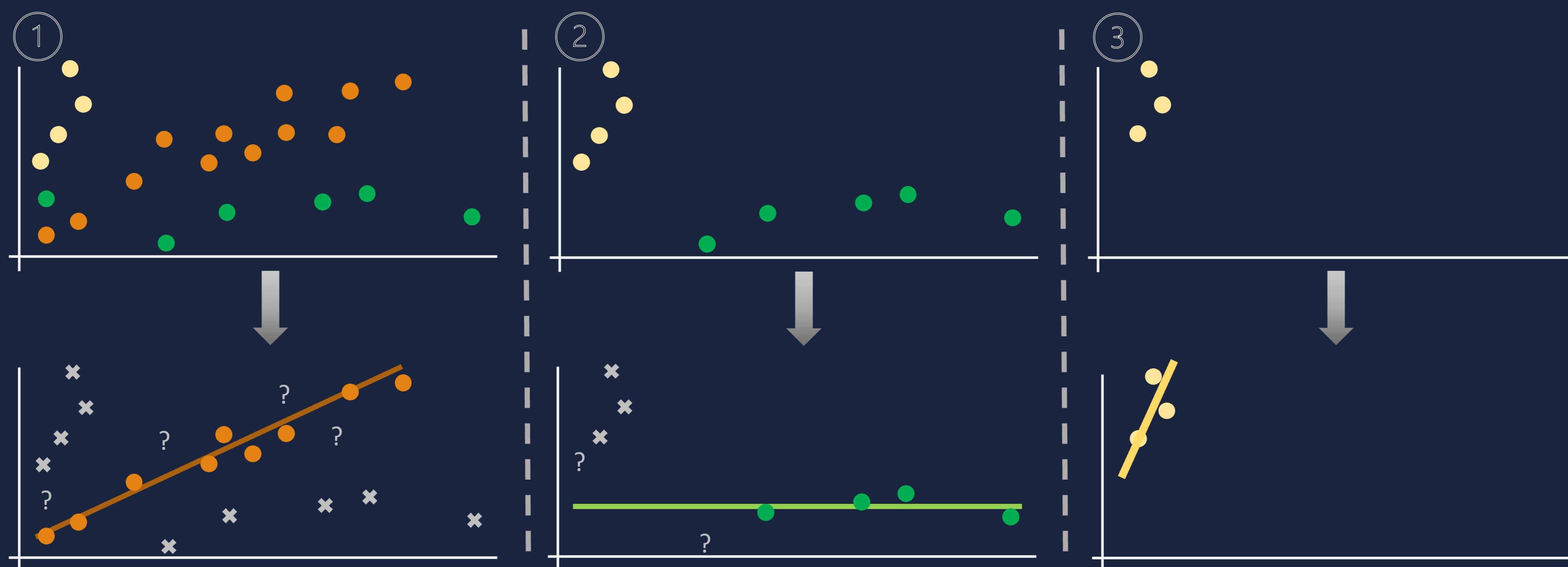
Estimate β_1, \dots, β_K from the observations
$$y_i = x_i^\top \beta_{c_i} + \epsilon_i$$

where labels $c_i \in \{1, \dots, K\}$ are unobserved

Illustration ($K = 3$)



Many methods struggle with **imbalanced** mixtures
Our method **exploits** the imbalance of the mixture to recover
the models one by one using tools from **robust regression**



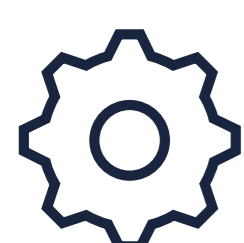
In practice,
colors are
unobserved



1. Existing methods

Expectation maximization, alternating minimization, gradient descent

- Estimate the models simultaneously
- Often fail on imbalanced mixtures
- Need to know the number of models K a-priori



2. Our Mix-IRLS method

- Sequential recovery using Iteratively Reweighted Least Squares (IRLS) to identify 'outliers'
- "I don't know" assignment to hard samples
- Number of models K can be unknown



3. Empirical performance

- Requires much fewer observations than other methods on imbalanced mixtures (Fig. 1)
- Copes better with outliers (balanced/imbalanced)
- Smaller error on real-world datasets (Fig. 2)



4. Theoretical guarantee

- Assumptions: population setting (infinite sample size), $K = 2$, imbalanced mixture, bounded noise
- Result: exact recovery of the models
- Novelty: no upper bound on the imbalance
- Other advantages: arbitrary initialization, input number of models can be unknown

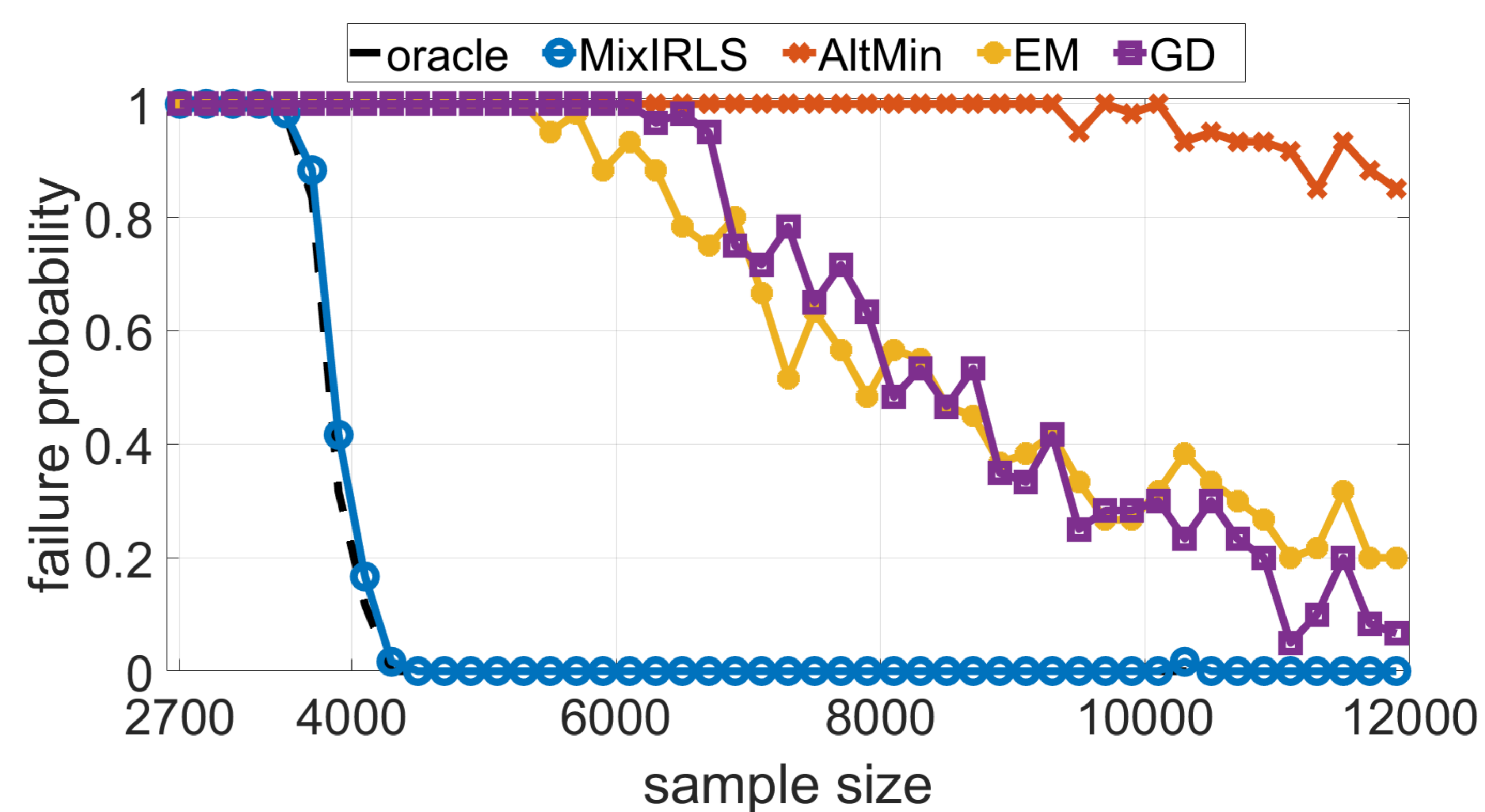


Figure 1. Results on an imbalanced mixture of $K = 3$ linear models as a function of the number of samples. Oracle knows the hidden labels c_i and performs separate linear regression for each model. Y-axis is the fraction of random initializations from which the algorithm failed.

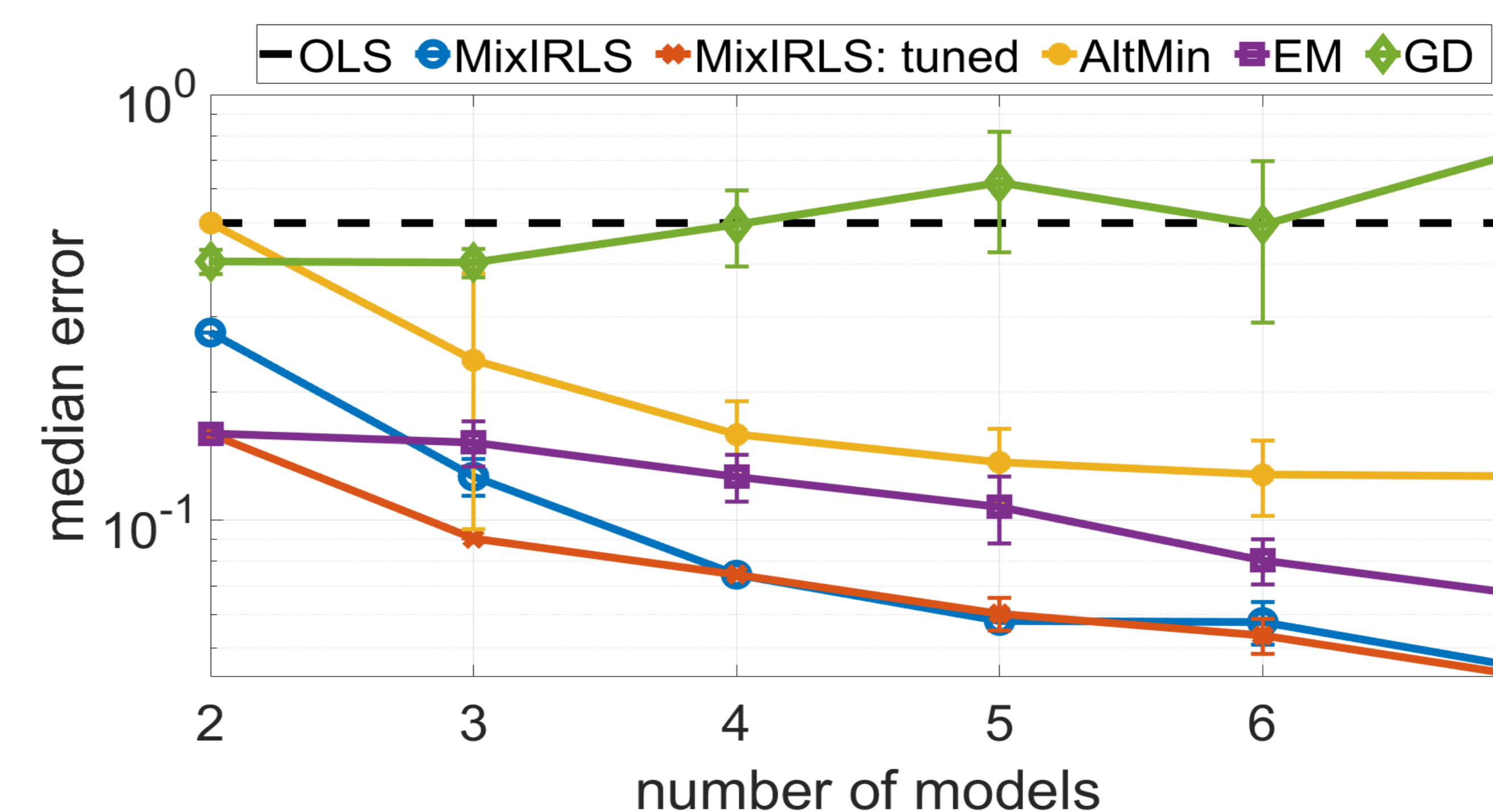


Figure 2. Results on the medical insurance cost dataset (Kaggle). Since the real number of models is unknown, depicted are the results as a function of the number of models given as input to the algorithms.

