Assessment of Gene Regulatory Network Inference Algorithms Using Monte Carlo Simulations

CCBCOL V

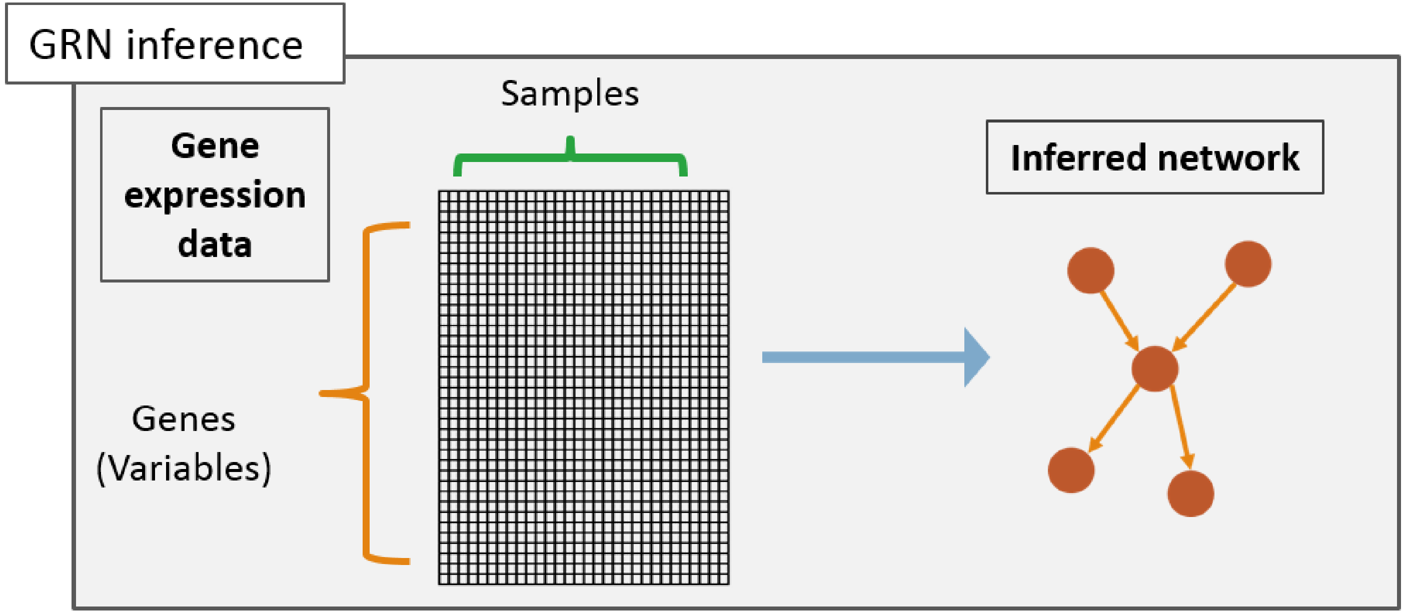
Adrián Zuur and Liliana López-Kleine

Department of Statistics, UNAL-Bogotá

November 2019

The Theoretical Problem

In bioinformatics there are many methods to build GRNs from gene expression data.



The Theoretical Problem

In bioinformatics there are many methods to build GRNs from gene expression data.

Papers introducing new methods typically test them on given gene expression datasets. They show that the methods work *in practice*.

The Theoretical Problem

In bioinformatics there are many methods to build GRNs from gene expression data.

Papers introducing new methods typically test them on given gene expression datasets. They show that the methods work *in practice*.

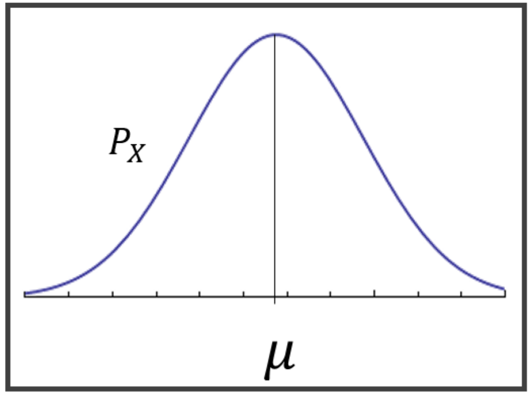
Our question is *do they work well in theory?*

- How dependent is a method on shape of regulatory relations, sample size, noise, etc.?
- How reliable is the method? Are reported results flukes?

Statistics 101

A basic example of statistical inference. We have:

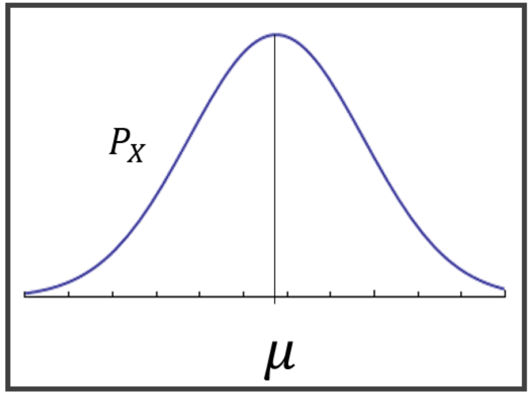
$P_X(x)$	A theoretical model with a probability distribution
$E(X) = \mu$	An unknown parameter of the model to be estimated
$X_1, X_2, \ldots, X_n$ i.i.d. $P_X(x)$	A sample from the distribution
$\overline{X_n}$	A statistic to estimate the parameter of interest



Statistics 101

A basic example of statistical inference. We have:

$P_X(x)$	A theoretical model with a probability distribution
$E(X) = \mu$	An unknown parameter of the model to be estimated
$X_1, X_2, \ldots, X_n$ i.i.d. $P_X(x)$	A sample from the distribution
$\overline{X_n}$	A statistic to estimate the parameter of interest



We ask: **is our statistic a good (reliable, accurate) estimator of our parameter?**

- Basic probability says $\bar{X}_n$ is accurate on average.
- By the Law of Large Numbers, $\bar{X}_n$ is increasingly reliable as $n \to \infty$.

The Probabilistic Model

Consider GRN inference algorithms as estimators and look at their statistical properties (unfair).

The Probabilistic Model

Consider GRN inference algorithms as estimators and look at their statistical properties (unfair).

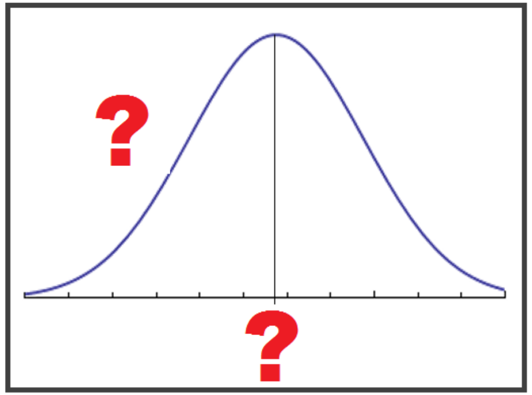
Estimators of what?

? A theoretical model with a probability distribution

? An unknown parameter of the model to be estimated

$X_1, X_2, \ldots, X_n$ i.i.d. ? A sample from the distribution

✓ A statistic to estimate the parameter of interest



The Probabilistic Model

Consider GRN inference algorithms as estimators and look at their statistical properties (unfair).

Estimators of what? **A Bayesian Network associated to a causal SEM**.

The Probabilistic Model

Consider GRN inference algorithms as estimators and look at their statistical properties (unfair).

Estimators of what? **A Bayesian Network associated to a causal SEM**.

Causal Structural Equations Model (SEM)

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= f_3(X_1, X_2, \epsilon_3) \\
X_4 &= f_4(X_3, \epsilon_4) \\
X_5 &= f_5(X_3, \epsilon_5)
\end{aligned}
$$

Each equation is a causal mechanism.

The joint distribution of noise variables ϵ_i determines a joint distribution of gene expressions. This is $P_X(x)$.

The Probabilistic Model

Consider GRN inference algorithms as estimators and look at their statistical properties (unfair).

Estimators of what? **A Bayesian Network associated to a causal SEM**.

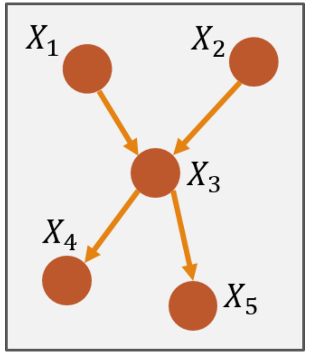
Causal Structural Equations Model (SEM)

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= f_3(X_1, X_2, \epsilon_3) \\
X_4 &= f_4(X_3, \epsilon_4) \\
X_5 &= f_5(X_3, \epsilon_5)
\end{aligned}
$$

Each equation is a causal mechanism.

The joint distribution of noise variables ϵ_i determines a joint distribution of gene expressions. This is $P_X(x)$.

Bayesian Network



Draw edges from direct causes to effects. This is a Bayesian Network, **our parameter of interest**.

Methods We Study

Methods We Study

Mutual information-based

Measure edge stregth by mutual information,

$$I(X_i, X_j) = E\left(\log \frac{f_{X_i}(X_i) f_{X_j}(X_j)}{f_{X_i X_j}(X_i, X_j)}\right).$$

Estimate mutual information with Miller-Madow estimator. Refine/threshold.

Methods We Study

Mutual information-based

Measure edge stregth by mutual information,

$$I(X_i, X_j) = E\left(\log \frac{f_{X_i}(X_i) f_{X_j}(X_j)}{f_{X_i X_j}(X_i, X_j)}\right).$$

Estimate mutual information with Miller-Madow estimator. Refine/threshold.

Regression-based

Measure edge strength with scores derived from fitting regressions.

Mutual information-based methods

Mutual information network

Estimate mutual information matrix, threshold.

ARACNe

For each triplet of variables, eliminate edge with lowest estimated MI.

MRNET

Derive 'minimum redundancy, maximum relevance' score from estimated MI.

CLR

Standardize estimated MI matrix row-wise and column-wise. Average both scores.

Regression-based methods

NARROMI

Estimate LAD-Lasso regressions. Use β as scores for edges.

TIGRESS

Estimate Least Angle Regressions (LARS) in a bootstrap (of sorts). Use estimates to compute scores of relevance in prediction.

GENIE3

Estimate an ensembles of regression trees (e.g. random forest). Use estimates to compute scores of relevance in prediction.

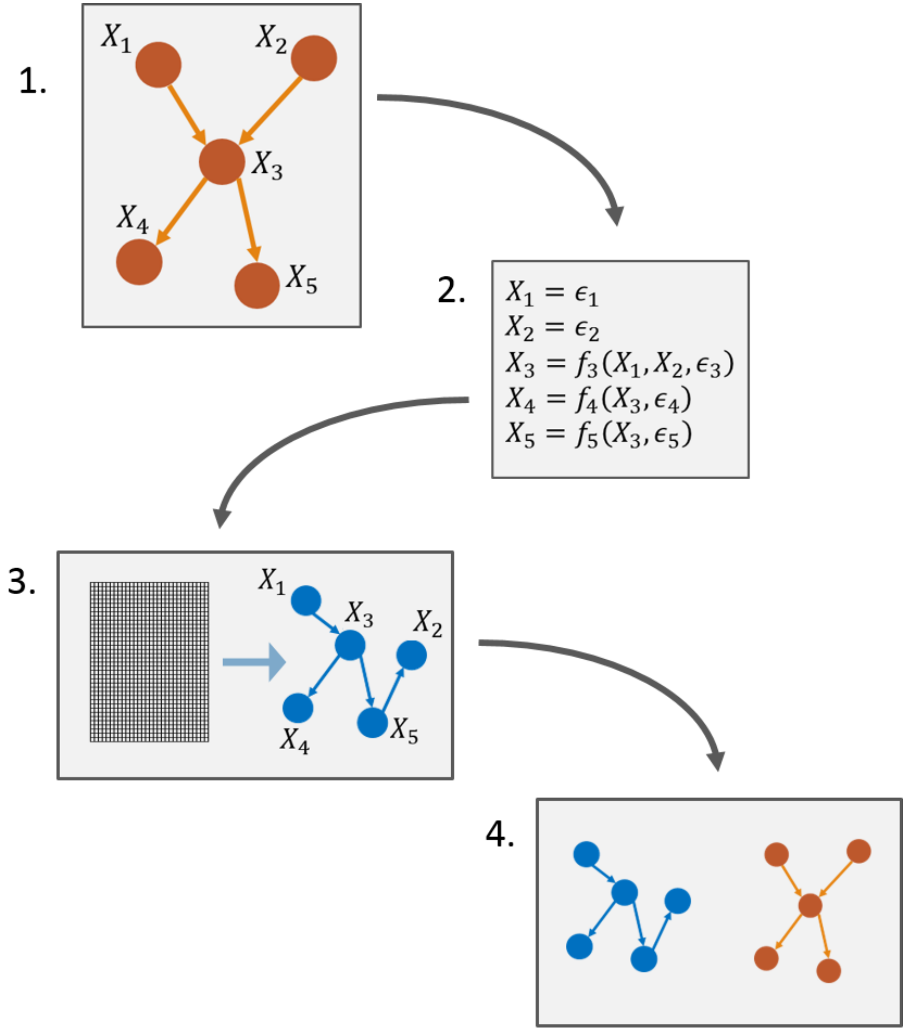
Workflow

1. Fix theoretical network.

2. Generate causal SEMs over this network.

3. Simulate data and apply algorithms.

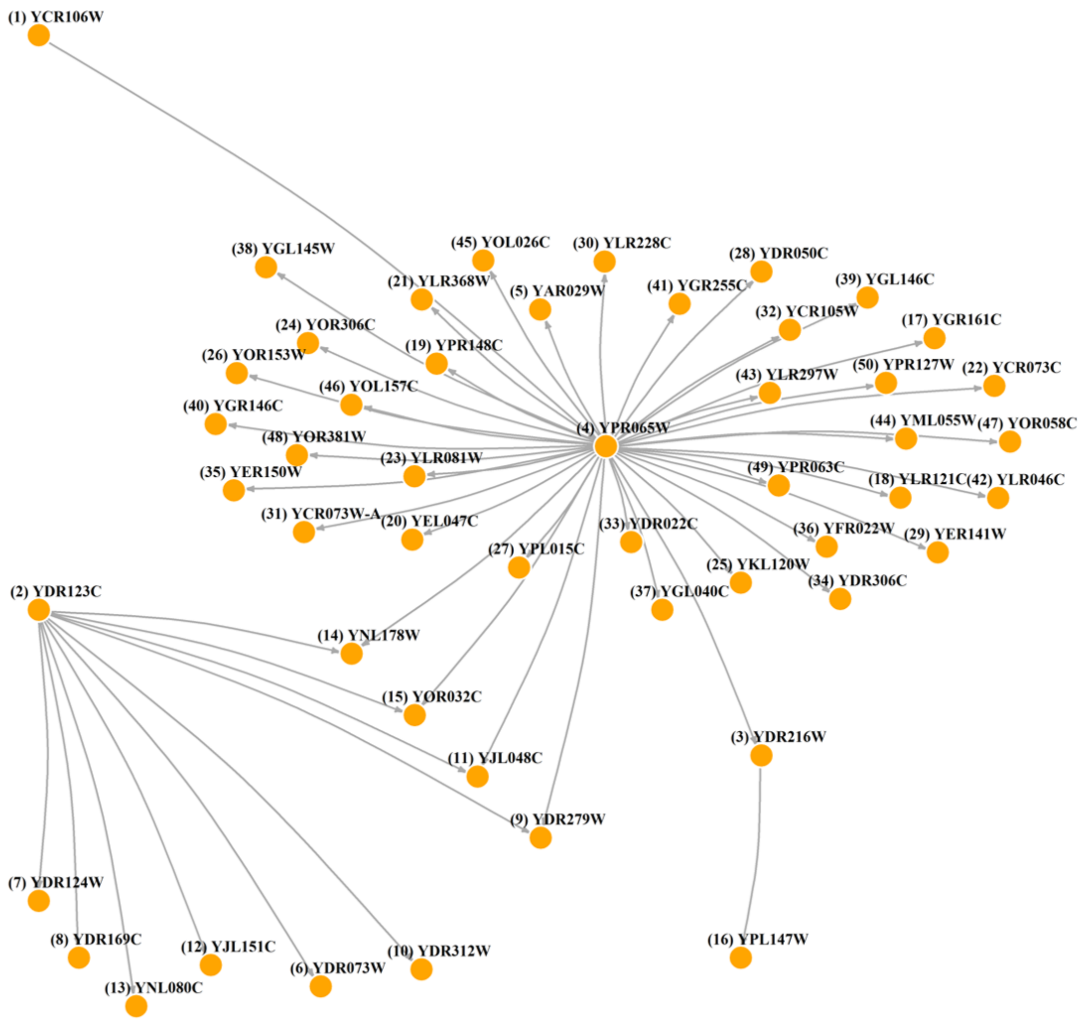
4. Evaluate algorithm outputs.



Sub-network

Source Network: Sisi Ma *et al.* (2014)

Extraction Algorithm: Marbach *et al.* (2009)



Causal SEM Definition

- Linear functional form

$$f_i\left(pa\left(X_i\right), \varepsilon_i\right) = \sum_{X_j \in pa(X_i)} \alpha_{ij} X_j + \varepsilon_i$$

Causal SEM Definition

- Linear functional form

$$f_i\left(pa\left(X_i\right), \varepsilon_i\right) = \sum_{X_j \in pa(X_i)} \alpha_{ij} X_j + \varepsilon_i$$

- Gaussian errors

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

Causal SEM Definition

- Linear functional form

$$f_i\left(pa\left(X_i\right), \varepsilon_i\right) = \sum_{X_j \in pa(X_i)} \alpha_{ij} X_j + \varepsilon_i$$

- Gaussian errors

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

- Low and high levels of noise

$$FVU_i = \frac{Var(\varepsilon_i)}{Var(X_i)} = 0.2$$

$$FVU_i = \frac{Var(\varepsilon_i)}{Var(X_i)} = 0.8$$

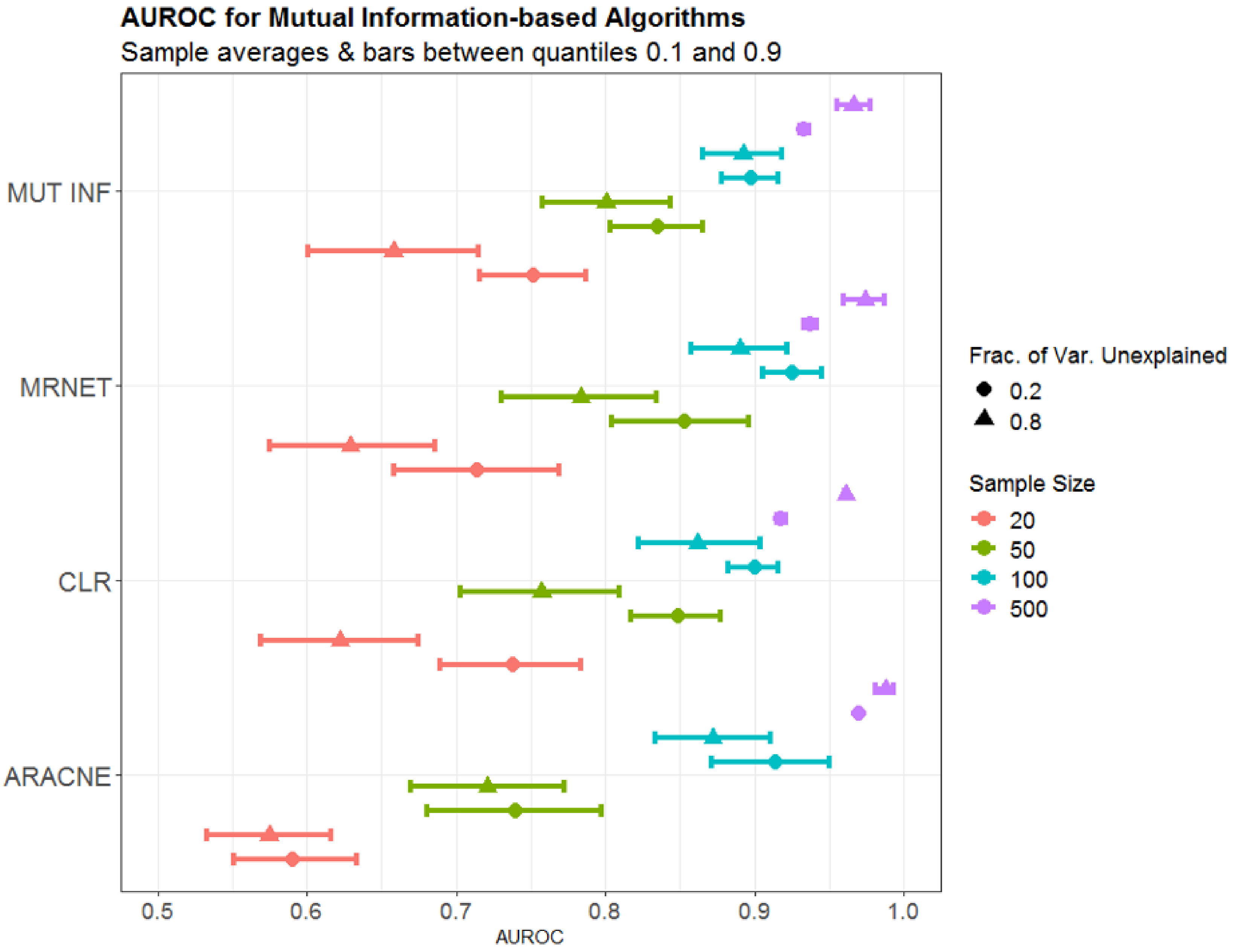
Simulations

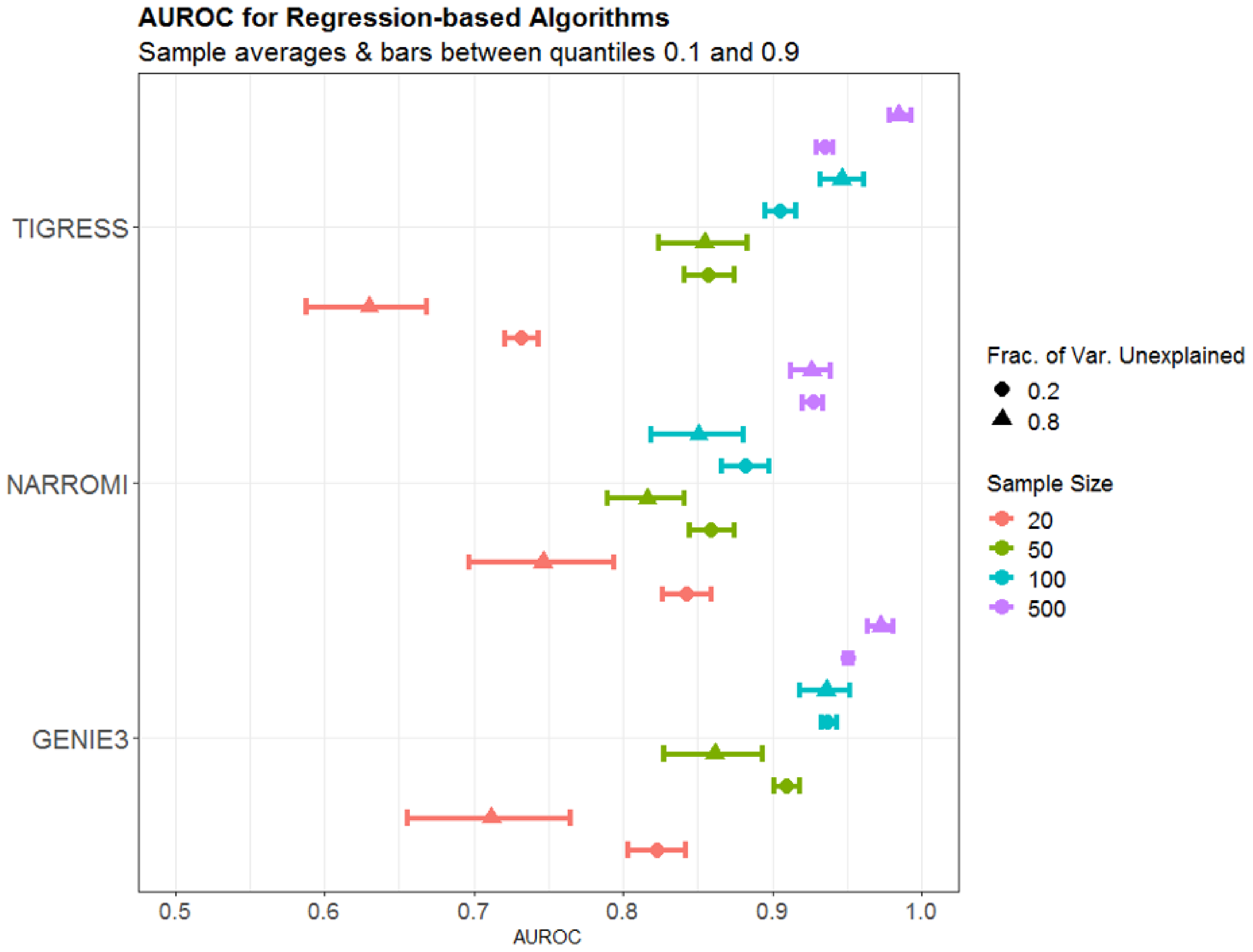
- For each causal SEM we simulate 1000 datasets of size 20, 50, 100, and 500.

Simulations

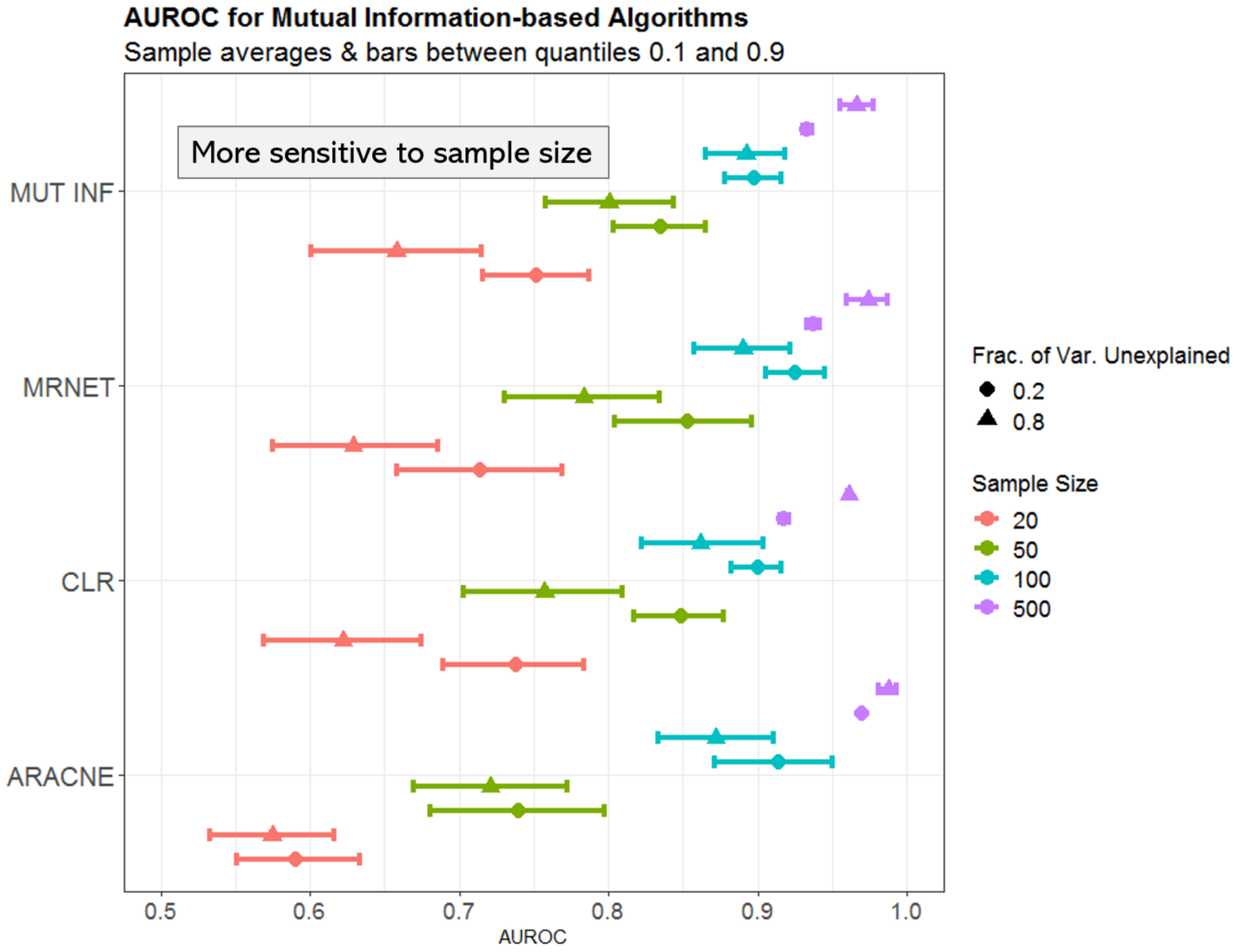
- For each causal SEM we simulate 1000 datasets of size 20, 50, 100, and 500.
- Algorithms are used "out-of-the-box", that is, using tuning parameters suggested by authors.

Results

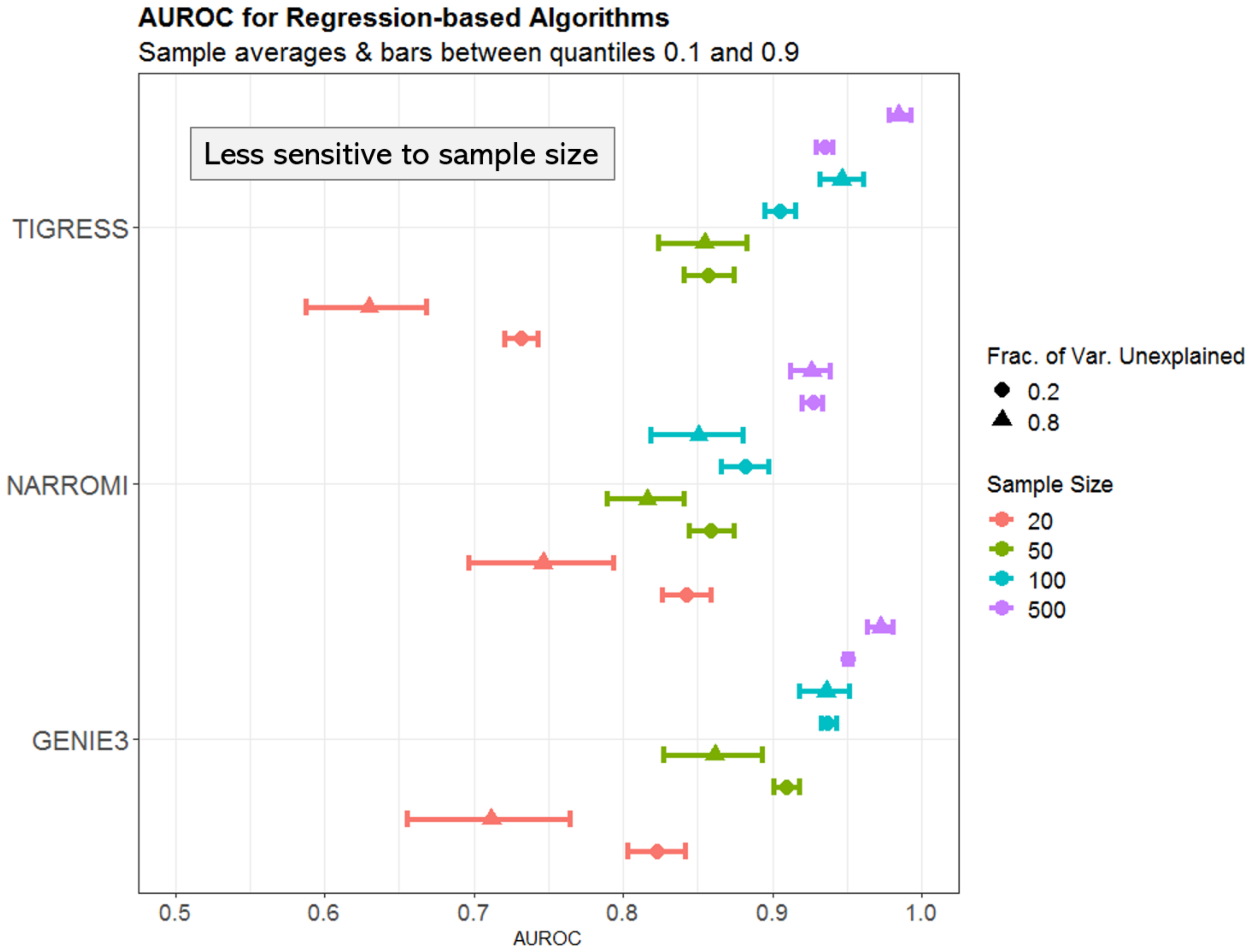

AUROC for Mutual Information-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
MUT INF
MRNET
CLR
ARACNE
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500

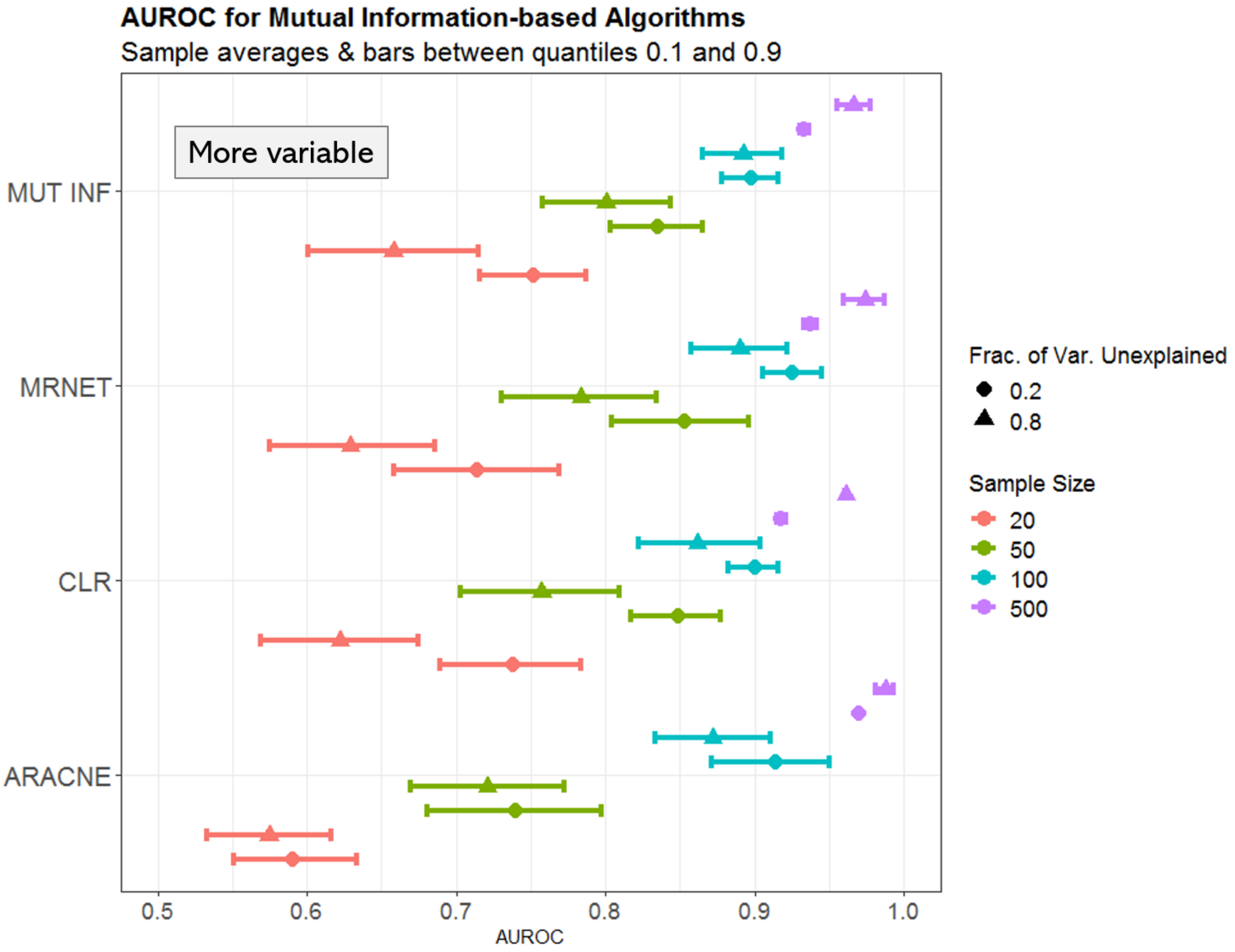


AUROC for Regression-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
TIGRESS
NARROMI
GENIE3
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500

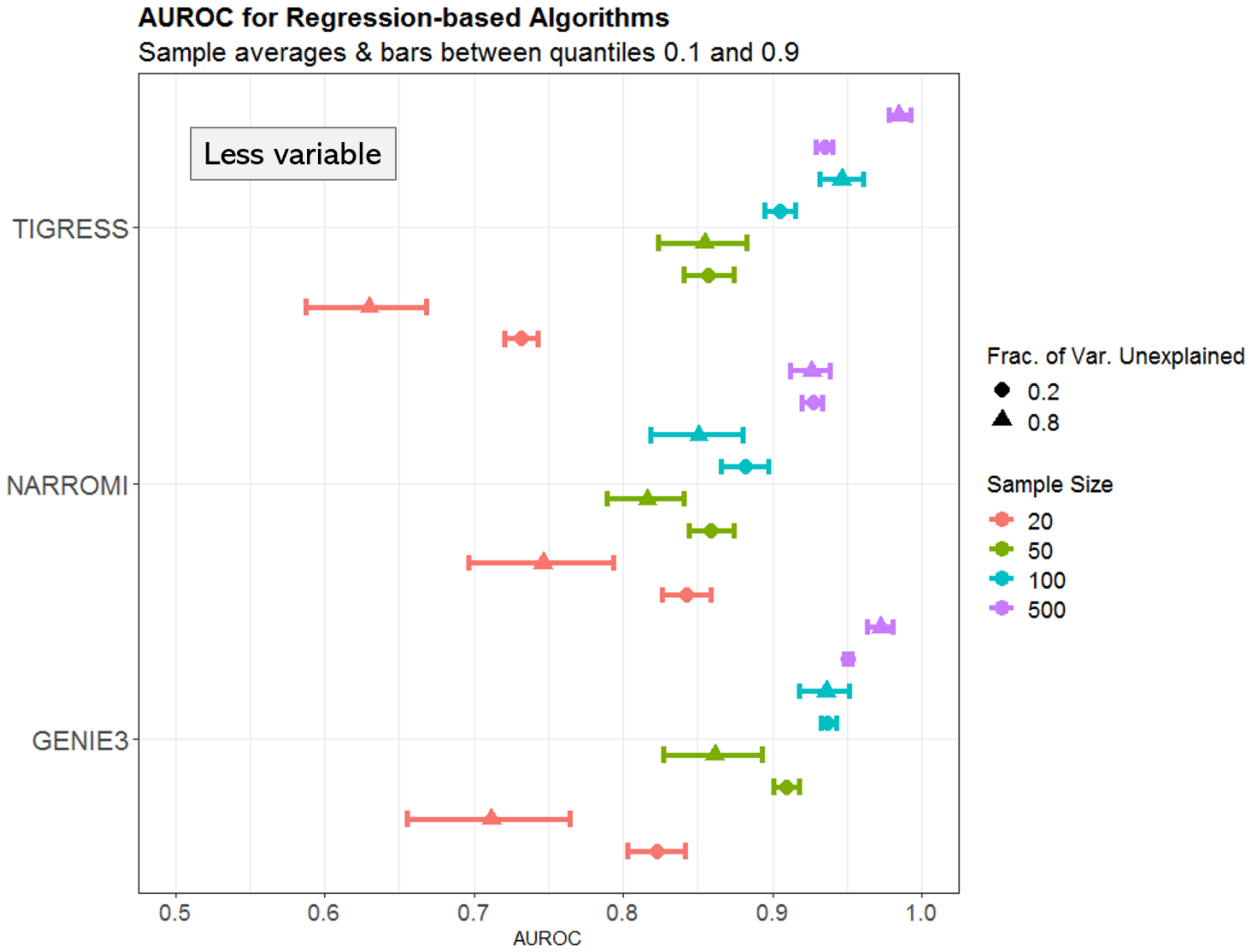


AUROC for Mutual Information-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
More sensitive to sample size
MUT INF
MRNET
CLR
ARACNE
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500

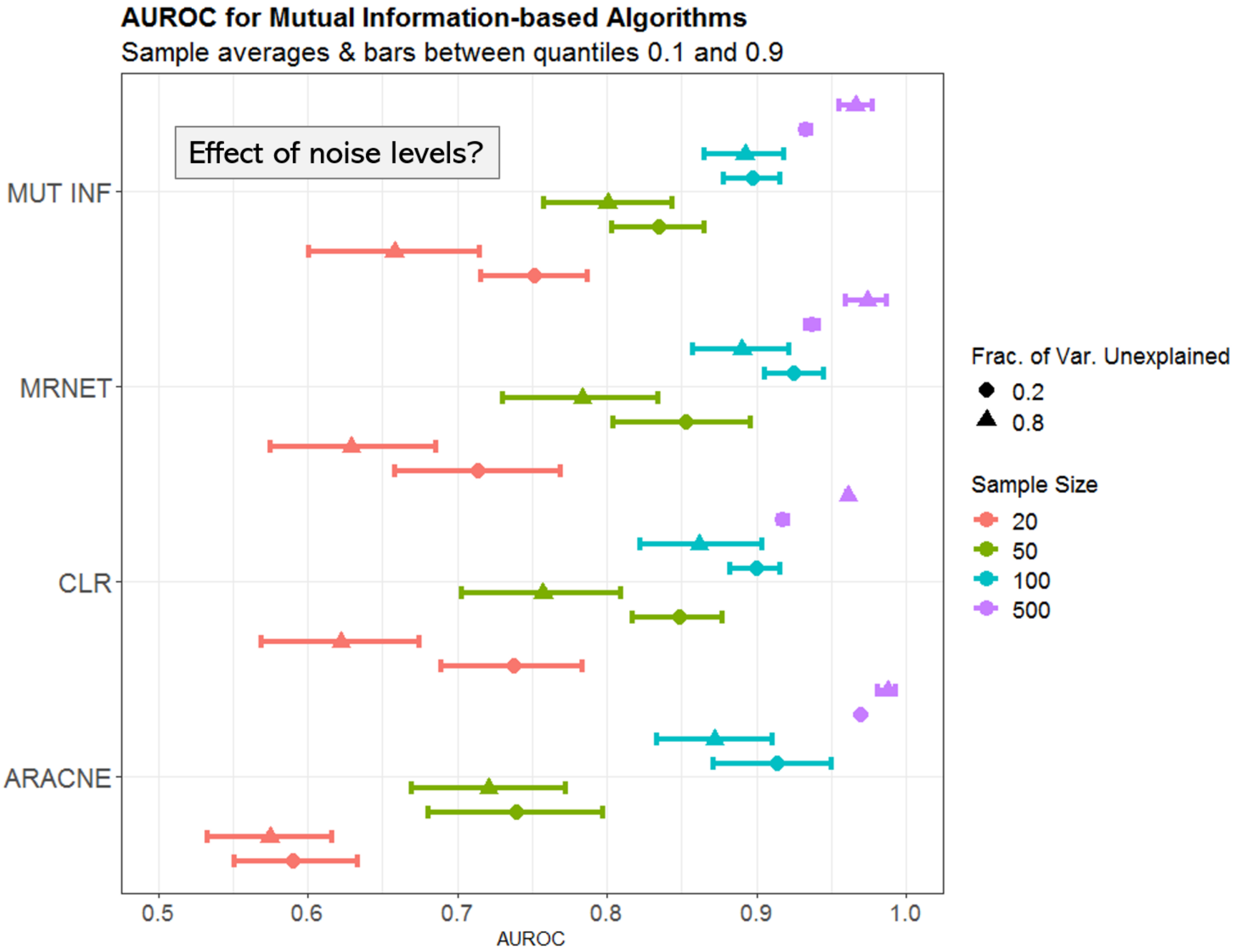

AUROC for Regression-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
Less sensitive to sample size
TIGRESS
NARROMI
GENIE3
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500

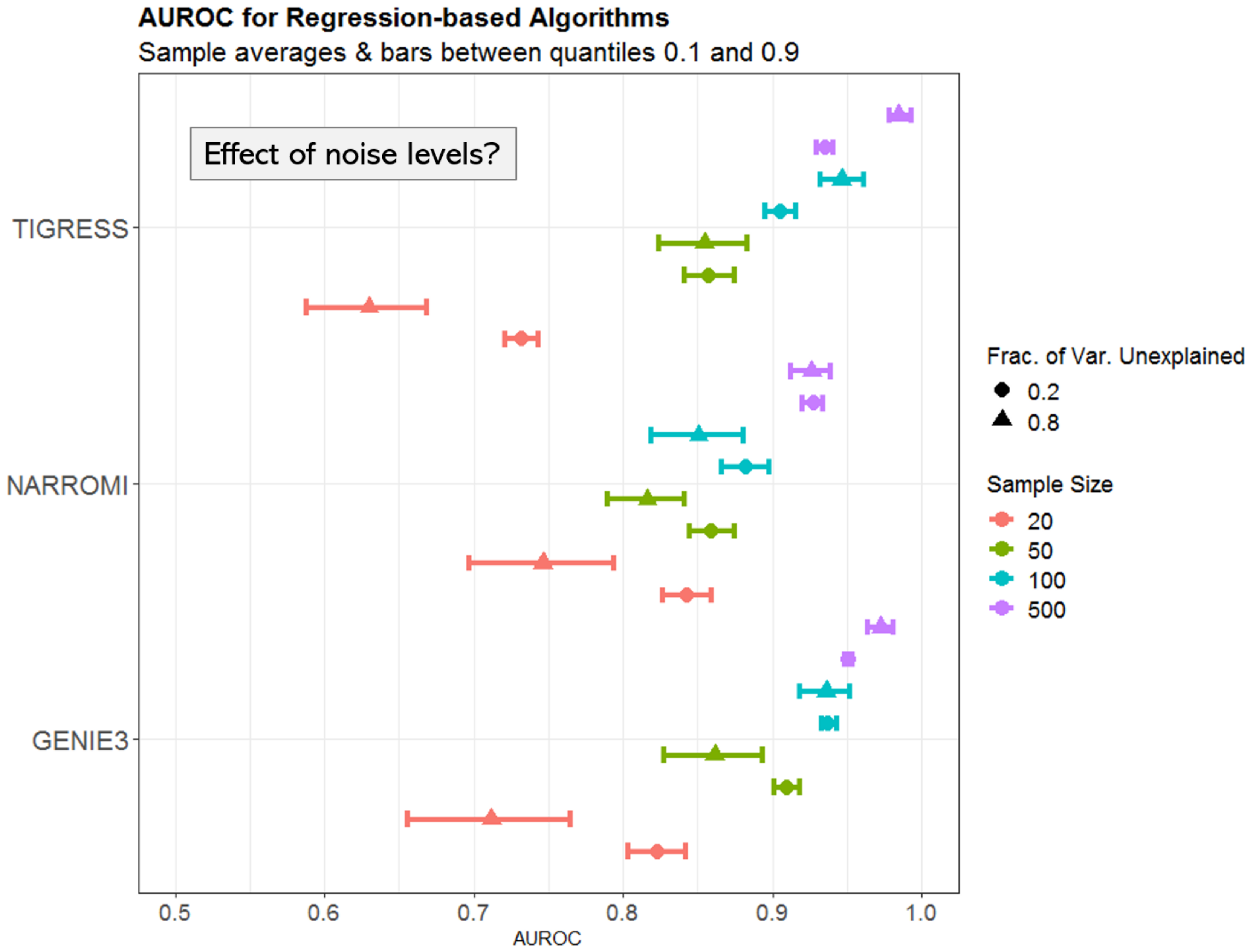


AUROC for Mutual Information-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
More variable
MUT INF
MRNET
CLR
ARACNE
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500



AUROC for Regression-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
Less variable
TIGRESS
NARROMI
GENIE3
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500


AUROC for Mutual Information-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
Effect of noise levels?
MUT INF
MRNET
CLR
ARACNE
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500



AUROC for Regression-based Algorithms
Sample averages & bars between quantiles 0.1 and 0.9
Effect of noise levels?
TIGRESS
NARROMI
GENIE3
0.5
0.6
0.7
0.8
0.9
1.0
AUROC
Frac. of Var. Unexplained
0.2
0.8
Sample Size
20
50
100
500

Thanks

Thanks.

I'm interested in comments or suggestions.

My email is agzuurp@unal.edu.co, and we can talk outside.

References

Sisi Ma et al. «De-Novo Learning of Genome-Scale Regulatory Networks in S. cerevisiae». PLOS ONE 9.9 (2014), pages. 1-20. doi: 10.1371/journal.pone.0106479. url: https://doi.org/10.1371/journal.pone.0106479.

Xiujun Zhang et al. «NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference». Bioinformatics 29.1 (2013), pages. 106-113. issn: 1367-4803. doi: 10.1093/bioinformatics/bts619.

Xin Fang et al. «Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities». Proceedings of the National Academy of Sciences (2017). issn: 0027-8424. doi: 10.1073/pnas.1702581114.

Anne-Claire Haury et al. «TIGRESS: Trustful Inference of Gene REgulation using Stability Selection». BMC Systems Biology 6.1 (2012), pág. 145. issn: 1752-0509. doi: 10.1186/1752-0509-6-145.

Emmert-Streib F, Dehmer M, Haibe-Kains B. «Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks». Front Cell Dev Biol. (2014) 2:38. doi:10.3389/fcell.2014.00038

Atul J Butte e Isaac S. Kohane. «Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.» Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2000), pages. 418-29.

Vân Anh Huynh-Thu et al. «Inferring Regulatory Networks from Expression Data Using Tree-Based Methods». PLOS ONE 5.9 (2010), pages 1-10. doi: 10.1371/journal.pone.0012776.

Adam A. Margolin et al. «ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context». BMC Bioinformatics 7 (2006), S7 -S7.

Jeremiah J Faith et al. «Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles». PLOS Biology 5.1 (2007), pages. 1-13. doi: 10.1371/journal.pbio.0050008

J. Peters, D. Janzing y B. Schölkopf. Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge, MA, USA: MIT Press, (2017).

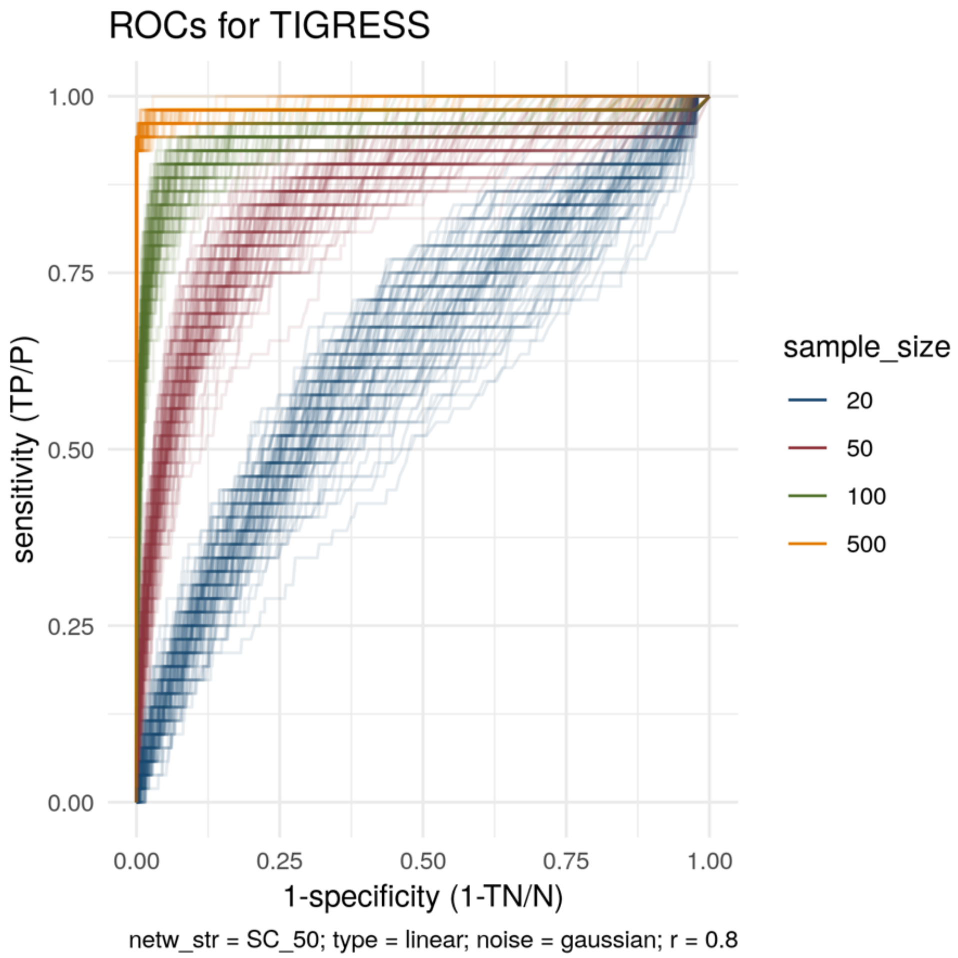
Assessment of Estimates

Each algorithm can be seen as a classifier that outputs scores s_{ij} for edges. For each threshold on scores we get

$$\text{Sensitivity} = \frac{\text{No. of correctly detected edges}}{\text{No. of true edges}}$$

$$\text{Specificity} = \frac{\text{No. of correctly detected non-edges}}{\text{No. of true non-edges}}$$

Over all thresholds, we get a parametric curve - the ROC curve. The area under it, AUROC, is the probability that a randomly sampled true edge has a score higher than that of a randomly sampled non-edge.



- Mutual information was estimated with Miller-Madow estimator.

$$\hat{H}(X) = -\sum_{b_X \in \mathrm{bins}_X} \hat{p}_{b_X} \log\left(\hat{p}_{b_X}\right)$$

$$\hat{H}(Y) = -\sum_{b_Y \in \mathrm{bins}_Y} \hat{p}_{b_Y} \log\left(\hat{p}_{b_Y}\right)$$

$$H(\hat{X}, Y) = -\sum_{b_{X\times Y} \in \mathrm{bins}_{X\times Y}} \hat{p}_{b_{X\times Y}} \log\left(\hat{p}_{b_{X\times Y}}\right)$$

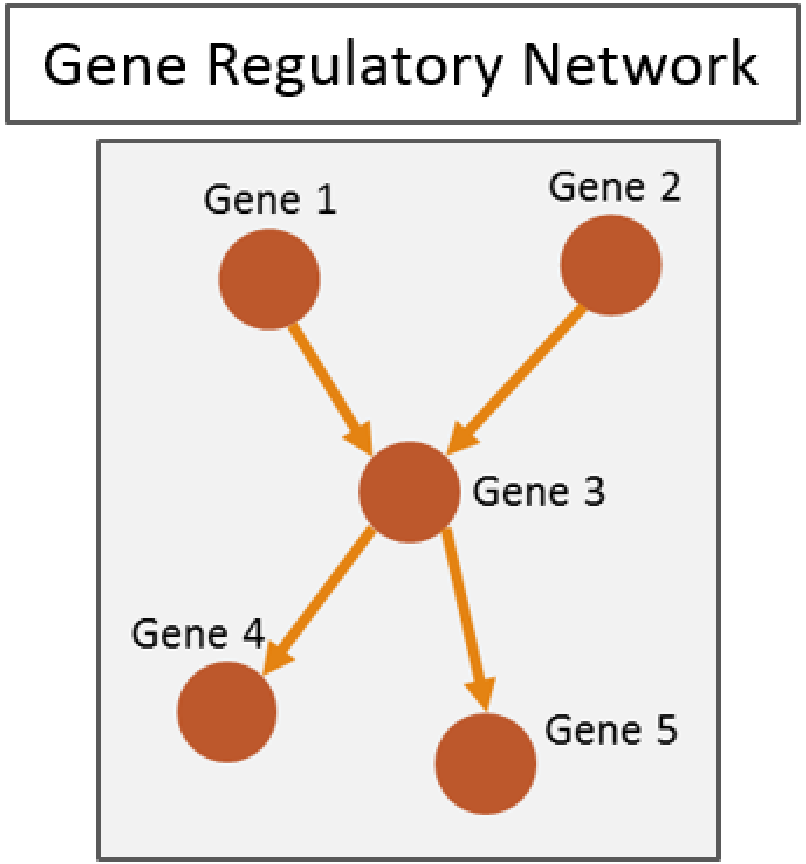
$$\hat{I}(X, Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y) + \frac{\hat{m} - 1}{n}$$

The Scientific Model

Gene regulatory networks (GRNs) are models that aim to encode the regulatory relations among genes in a genome.

- Genes are nodes, regulatory relations are edges.
- Regulatory relations are *causal* (edges are directed, indicate more than co-expression).
- Edges represent *direct* causal effects (indirect effects are directed paths).

GRNs are directed graphs (V, E), which are equivalent to an adjacency matrix.



$$\hat{\beta} = \textbf{argmin}_{\beta \in \mathbb{R}^p} |Y - \beta^\top X| + \lambda ||\beta||_1$$

Results

- MI-based algorithms are more variable than regression-based algorithms.
- MI-based algorithms are more sensitive to sample size than regression-based algorithms. ARACNe and NARROMI are the extremes.
- TIGRESS is most sensitive to FVU. ARACNe is least sensitive.
- Surprisingly good results for large n. Not so much for small n.
- Better results with less noise, except at large sample size. Bias-variance trade-off.