Individualized Inference using Bayesian Quantile Directed Acyclic Graphical Models

Ksheera Sagar K. N**

Department of Statistics
Purdue University

August 11, 2022

**collaborative work with Veera B (UMich), Yang Ni (Texas A&M) and Anindya Bhadra (Purdue).

Outline

1. Quantile DAGs
   - Introduction and motivation
   - Intuition to the model
   - Related works and our contributions
   - qDAGx: The model
   - Prior formulation
   - Synthetic data and simulation settings
   - Simulation results
   - Application of qDAGx in lung cancer data
2. Conclusion and future scope

Introduction and motivation

Challenges with GGMs:

- What if the interacting variables are not jointly Gaussian?
- And consequent lack of robustness in model misspecification.

Our solution:

- Circumvent the Gaussian assumption on the likelihood.
- Model association between variables at any given quantile level, $\tau \in (0, 1)$.

Real life motivation:

- Directed acyclic graphs are important to understand protein–protein interaction (PPI) networks.
- Personalized PPI's can help in a better understanding of diseases like cancer, and therefore finding applications in precision medicine.

Intuition to the model

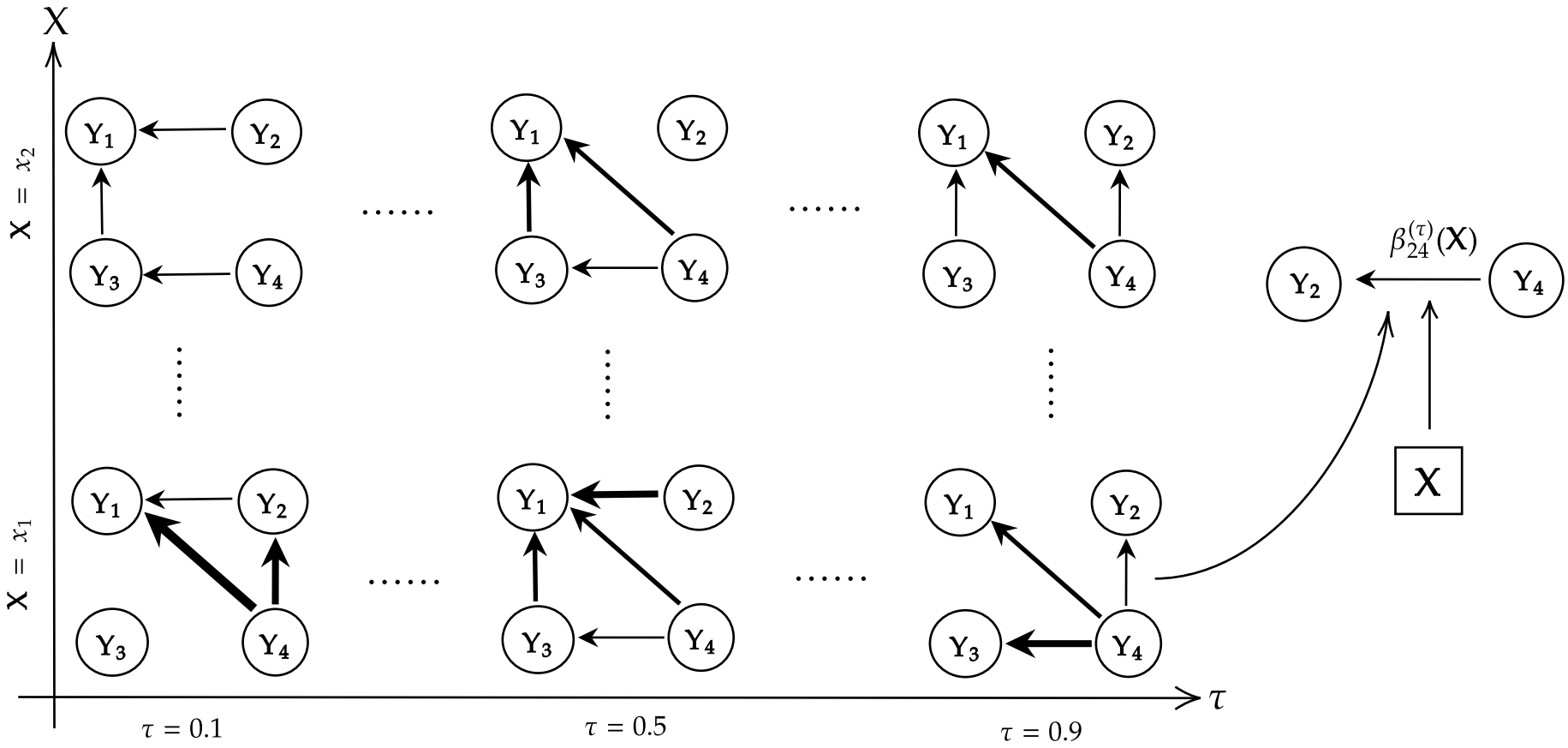

Figure: The directed acyclic graph $\mathcal{QG}^{(\tau)}$, for two observations, on four vertices $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, \boldsymbol{Y}_4)$ is presented, for univariate $\boldsymbol{X}$ and for given quantile levels $\tau = 0.1, 0.5$ and 0.9.

Related works and our contributions

Related works inspired by varying coefficient models (Hastie and Tibshirani, 1993):

- DAG inference using node conditional varying coefficient models (Ni et al., 2019). Drawbacks: Gaussian likelihood and known ordering assumption.

- Varying coefficient Bayesian quantile regression (Das et al., 2021). Drawbacks: lack of support to graphical models.

Related work in quantile-graphs:

- Undirected quantile-graphs constructed from node-wise quantile regression (Guha et al., 2020). Drawbacks: not individualized.

Related works and our contributions (cont.)

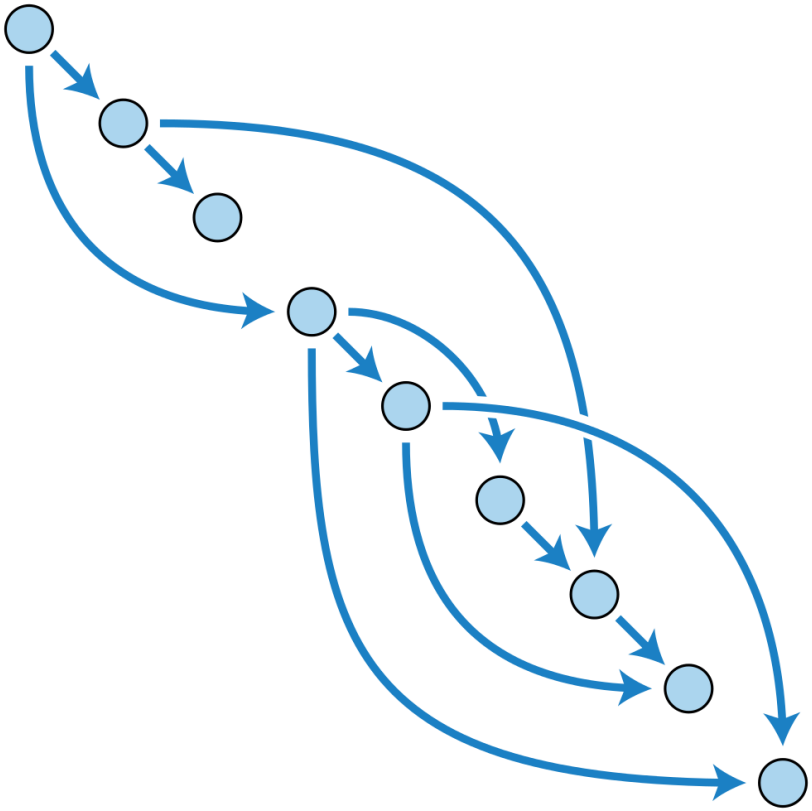

Figure: A sample DAG which is Topologically sorted.

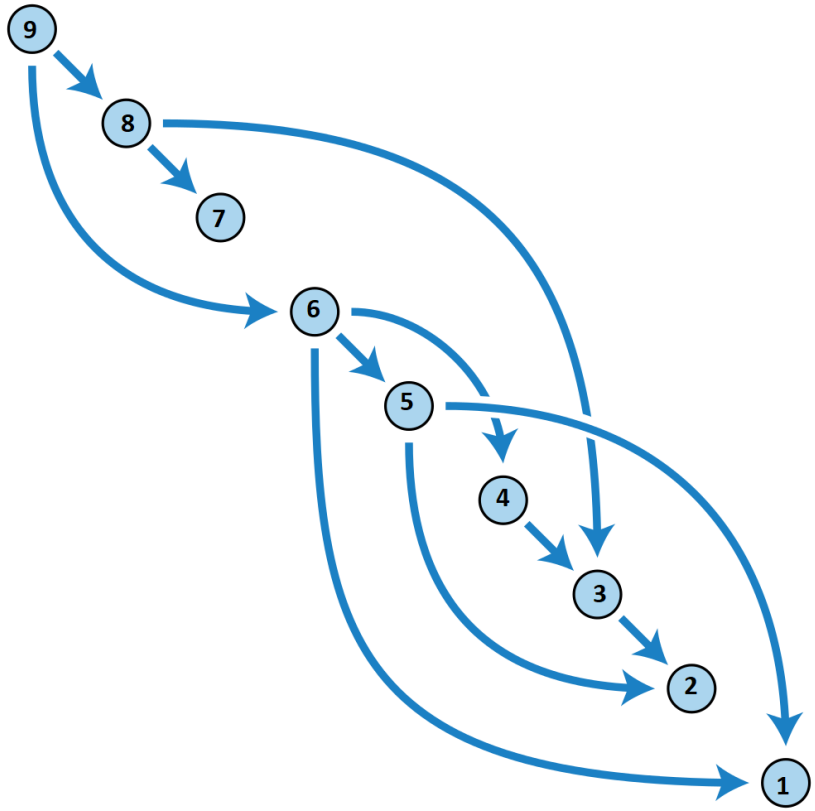

Figure: Labels given to nodes WLOG in the topologically sorted DAG.

Picture credits: **https://en.wikipedia.org/wiki/Directed_acyclic_graph**

Related works and our contributions (cont.)

Our contributions:

1. **qDAGx**: Learning ***individual-specific*** DAG's at any quantile level $\tau \in (0, 1)$, with no assumptions on the data likelihood or on the ordering of nodes.

2. Infer for the first time *individual-specific* protein–protein interaction networks in patients with lung adenocarcinoma and lung squamous cell carcinoma.
   - Model the protein-protein association in each patient at a quantile level τ, as a function of external covariates mRNA and methylation.

3. Structural identifiability of the quantile-DAGs, properties of prior which aid in sparse quantile DAG discovery and posterior consistency of node conditional fitted densities.

qDAGx: The model

Notations:

- First level covariates: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p$ and for $h \in \{1, \ldots, p\}$, $\boldsymbol{Y}_h \in \mathbb{R}^n$.
- Second level covariates: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ and for $k \in \{1, \ldots, q\}$, $\boldsymbol{X}_k \in \mathbb{R}^n$.
- Y_{ih}, X_{ik}: first, second level covariate values for i^{th} observation, $i \in \{1, \ldots, n\}$.

Model for conditional quantile:

$$Q_{Y_{ih}}(\tau \mid Y_{ij}, \boldsymbol{X}_{i\cdot}) = \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) + \sum_{j \in pa(h)} Y_{ij} \beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$$

qDAGx: The model

Notations:

- First level covariates: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p$ and for $h \in \{1, \ldots, p\}$, $\boldsymbol{Y}_h \in \mathbb{R}^n$.
- Second level covariates: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ and for $k \in \{1, \ldots, q\}$, $\boldsymbol{X}_k \in \mathbb{R}^n$.
- Y_{ih}, X_{ik}: first, second level covariate values for i^{th} observation, $i \in \{1, \ldots, n\}$.

Model for conditional quantile:

$$Q_{Y_{ih}}(\tau \mid Y_{ij}, \boldsymbol{X}_{i\cdot}) = \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) + \sum_{j \in pa(h)} Y_{ij} \beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$$

$$\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \cdot \mathbb{1}\left(|\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})| > t_{hj}\right) \text{ where,}$$

qDAGx: The model

Notations:

- First level covariates: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p$ and for $h \in \{1, \ldots, p\}$, $\boldsymbol{Y}_h \in \mathbb{R}^n$.
- Second level covariates: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ and for $k \in \{1, \ldots, q\}$, $\boldsymbol{X}_k \in \mathbb{R}^n$.
- Y_{ih}, X_{ik}: first, second level covariate values for i^{th} observation, $i \in \{1, \ldots, n\}$.

Model for conditional quantile:

$$Q_{Y_{ih}}(\tau \mid Y_{ij}, \boldsymbol{X}_{i\cdot}) = \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) + \sum_{j \in pa(h)} Y_{ij} \beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$$

$$\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \cdot \mathbb{1}\left(|\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})| > t_{hj}\right) \text{ where,}$$

$$\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \sum_{k=1}^{q} f_{hjk}^{(\tau)}(X_{ik}).$$

qDAGx: The model (cont.)

Union-DAG condition: Let $\mathcal{QG}_i^{(\tau)}$ be the adjacency matrix of quantile-DAG of i^{th} observation at quantile level τ. Then,

$$\mathcal{QG}_u^{(\tau)} = \bigcup_{i=1}^{n} \mathcal{QG}_i^{(\tau)} \text{ is a DAG, where } \mathcal{QG}_i^{(\tau)} = \Big(\Big(\beta_{hj}(\boldsymbol{X}_{i\cdot}) \neq 0\Big)\Big).$$

loss function → negative log-likelihood of ALD (Koenker and Bassett Jr, 1978) → Joint likelihood.

Joint likelihood:

$$\pi(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}^{(\tau)}, \tau) = \prod_{i=1}^{n}\prod_{h=1}^{p} \tau(1-\tau)\exp\left(-\psi_\tau\left(Y_{ih} - \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) - \sum_{j\in pa(h)} Y_{ij}\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})\right)\right) \times \mathbb{1}\left(\mathcal{QG}_u^{(\tau)} \text{ is a DAG}\right).$$

Where, $\psi_\tau(x) = \tau\mathbb{1}(x \geq 0) - (1-\tau)\mathbb{1}(x < 0)$.

Prior formulation

$$\boldsymbol{\theta}_{hj}^{(\tau)}(\boldsymbol{X}) = \sum_{k=1}^{q} f_{hjk}^{(\tau)}(\boldsymbol{X}_k) = \mu_{hj}\mathbf{1}_n + \sum_{k=1}^{q} \widetilde{\boldsymbol{X}_k}^{*}\boldsymbol{\alpha}_{hjk}^{*} + \sum_{k=1}^{q} \boldsymbol{X}_k\alpha_{hjk}^{0},$$

$$\text{peNMHS}: \begin{cases} \boldsymbol{\alpha}_{hjk}^{*} & = \eta_{hjk}\boldsymbol{\xi}_{hjk}\,,\ \eta_{hjk} \sim \mathcal{N}(0, T_{hj}^2 L_{hjk}^2), \\ \boldsymbol{\xi}_{hjk} & = \left(\xi_{hjk}^{(1)}, \ldots, \xi_{hjk}^{(B_k^*)}\right)^T, \\ \xi_{hjk}^{(l)} & \sim \mathcal{N}(m_{hjk}^{(l)}, 1)\,, \text{ for } l \in \{1, \ldots, B_k^*\}, \\ m_{hjk}^{(l)} & \sim 0.5 \cdot \delta_1(m_{hjk}^{(l)}) + 0.5 \cdot \delta_{-1}(m_{hjk}^{(l)}), \\ T_{hj} & \sim \mathcal{C}^{+}(0,1),\ L_{hjk} \sim \mathcal{C}^{+}(0,1). \end{cases}$$

$\alpha_{hjk}^{0} \sim$ peNMHS prior analogous to $\boldsymbol{\alpha}_{hjk}^{*}$.

$\mu_{hj} \sim \mathcal{N}(0, \sigma_\mu^2)$ and $t_{hj} \sim$ Gamma(shape $= a$, rate $= b$), $1 \le h, j \le p$.

Synthetic data and simulation settings

Problem dimensions: $n \in \{100, 250\}$, $p \in \{25, 50, 100\}$, $q \in \{2, 5\}$.

Data generating mechanism:

- $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ from a multivariate normal $\mathcal{N}(0, \boldsymbol{I}_q)$.
- WLOG, true order is $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p$ and for each $\boldsymbol{Y}_h$ choose $\max\left\{1, \lfloor \frac{p-h}{5} \rfloor\right\}$ number of parents from $\{\boldsymbol{Y}_{h+1}, \ldots, \boldsymbol{Y}_p\}$.
- Compute values of corresponding $\theta_{hj}(\cdot)$'s where $j \in pa(h)$, which are functions of τ, $\boldsymbol{X}$
- When $q = 2$, all thresholds $= 0.5$ and when $q = 5$, all thresholds $= 1$.
- Compute $\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \cdot \mathbb{1}\left(|\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})| > t_{hj}\right)$
- Generate n random samples for $\boldsymbol{Y}_h$:

$$Q_{Y_{ih}}(\tau \mid Y_{ij}, \boldsymbol{X}_{i\cdot}) = \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) + \sum_{j \in pa(h)} Y_{ij} \beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}).$$

Simulation results

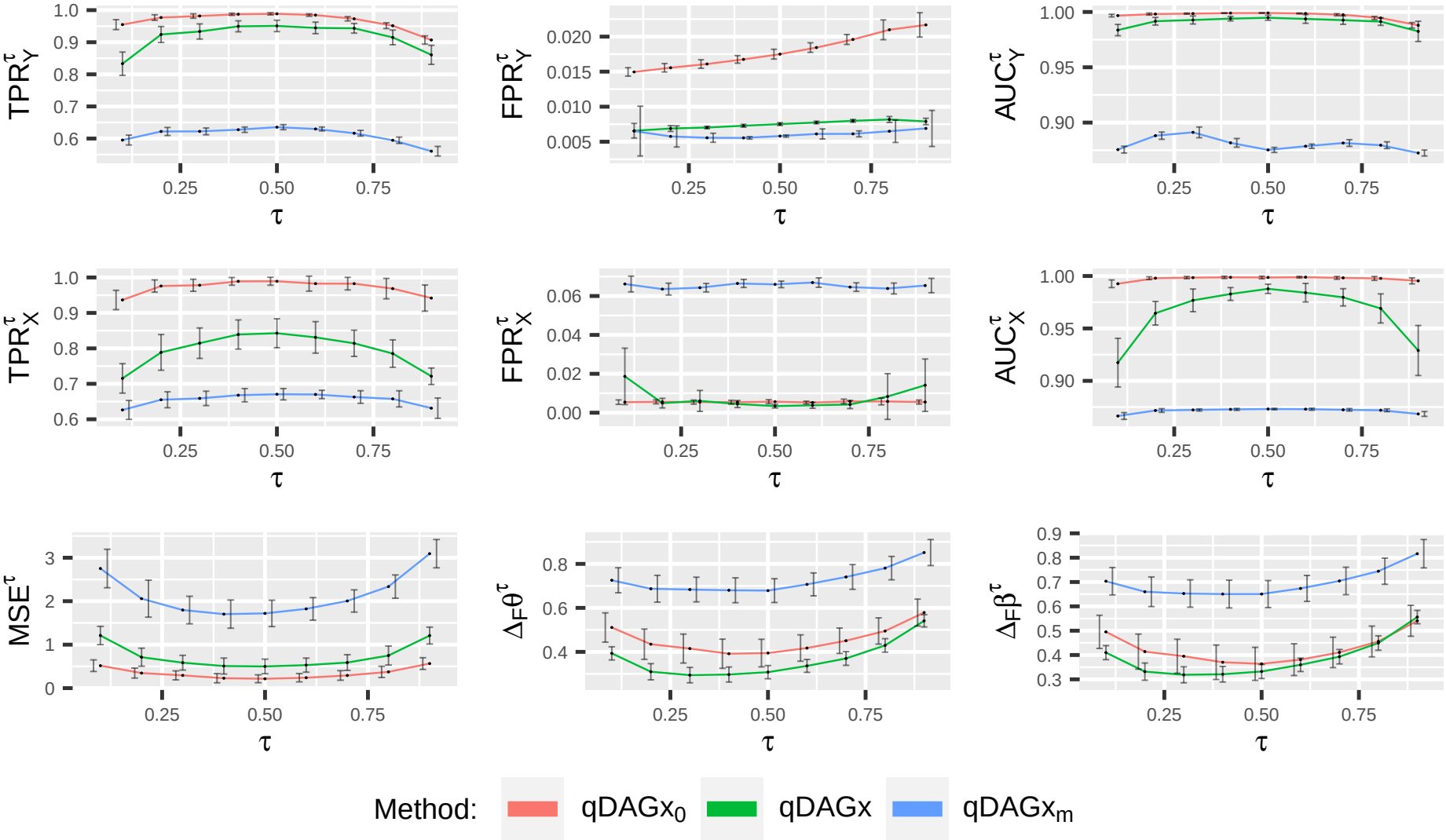

Figure: $p = 25, q = 5, n = 250$. Kendall's' T for the misspecified sequence is 0.5

qDAGx in Lung Cancer data

Overview:

- Protein expressions of 67 proteins, analogous to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{p=67}$.
- $\boldsymbol{X}_1$, $\boldsymbol{X}_2$: mRNA expression and methylation.
- $n = 306$ patients: Lung adenocarcinoma (LUAD).
- $n = 278$ patients: Lung squamous cell carcinoma (LUSC).
- Estimate quantile-DAGs at $\tau \in \{0.1, \ldots, 0.9\}$.
- Aggregate DAGs at each quantile level *for visualization proposes* and show edges present in $\geq n/2$ patients and node size $\propto$ in-degree.

qDAGx in Lung Cancer data (cont.)

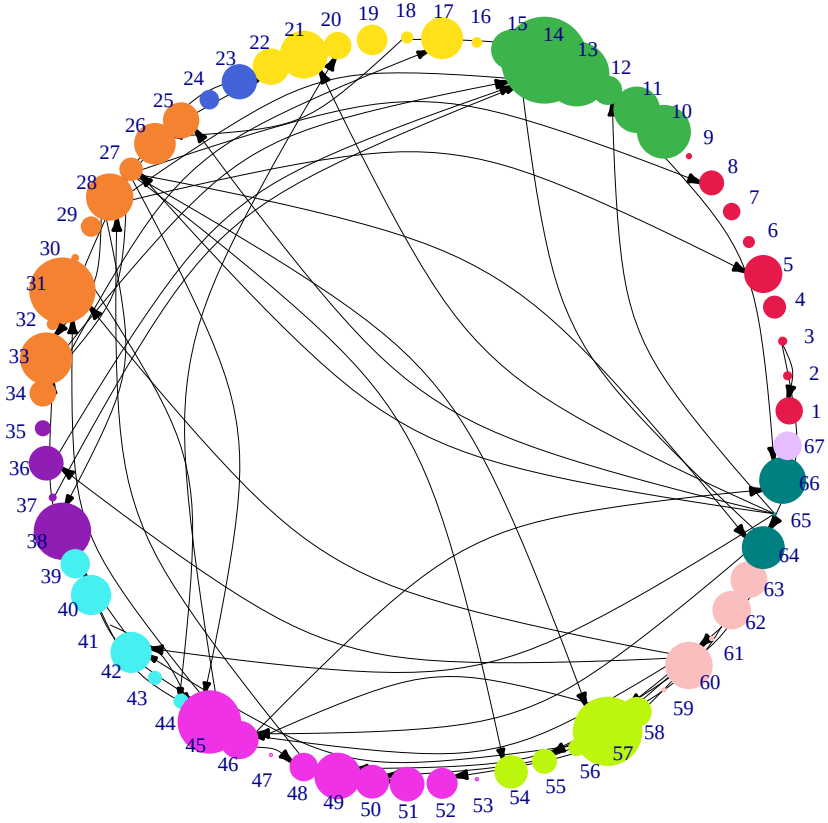

Figure: Graph of $\boldsymbol{E}^{(0.1)}_{\text{LUAD}}$.

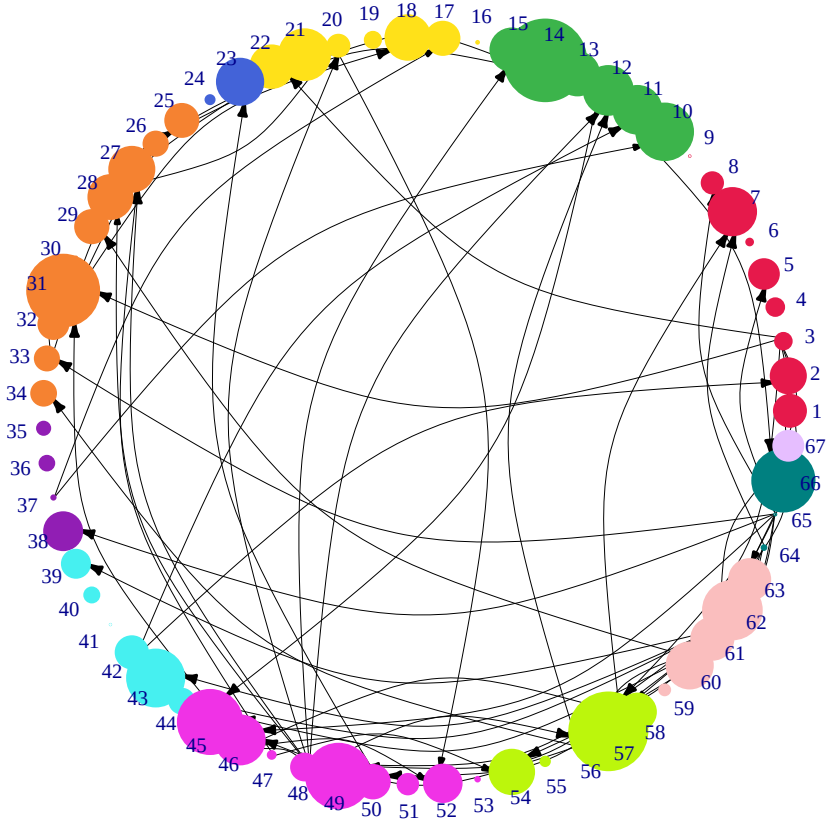

Figure: Graph of $\boldsymbol{E}^{(0.1)}_{\text{LUSC}}$.

qDAGx in Lung Cancer data (cont.)

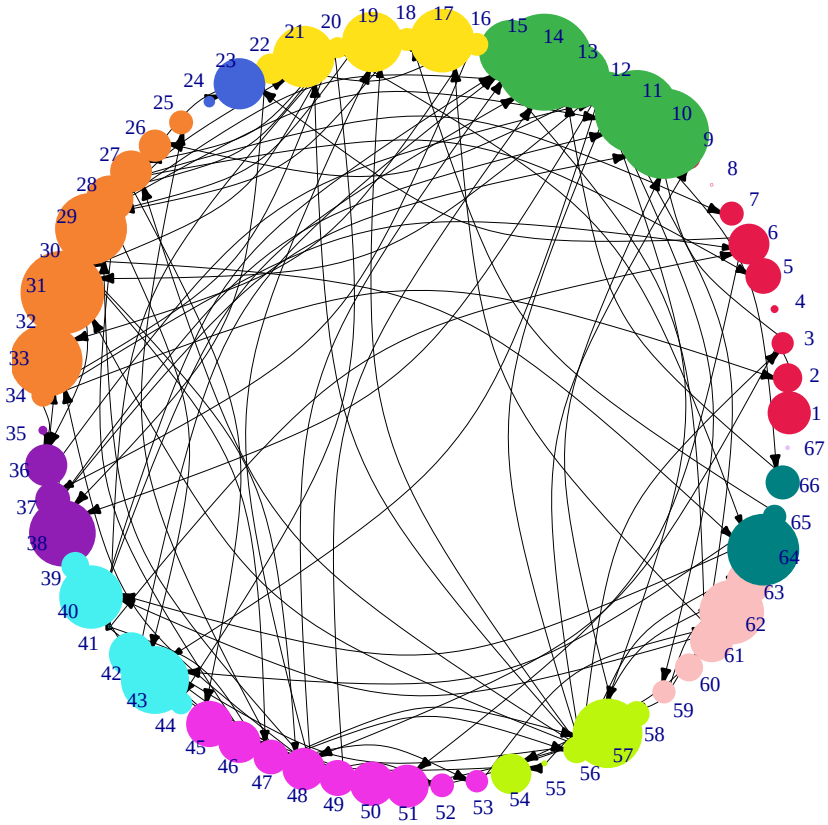

Figure: Graph of $\boldsymbol{E}^{(0.5)}_{\text{LUAD}}$.

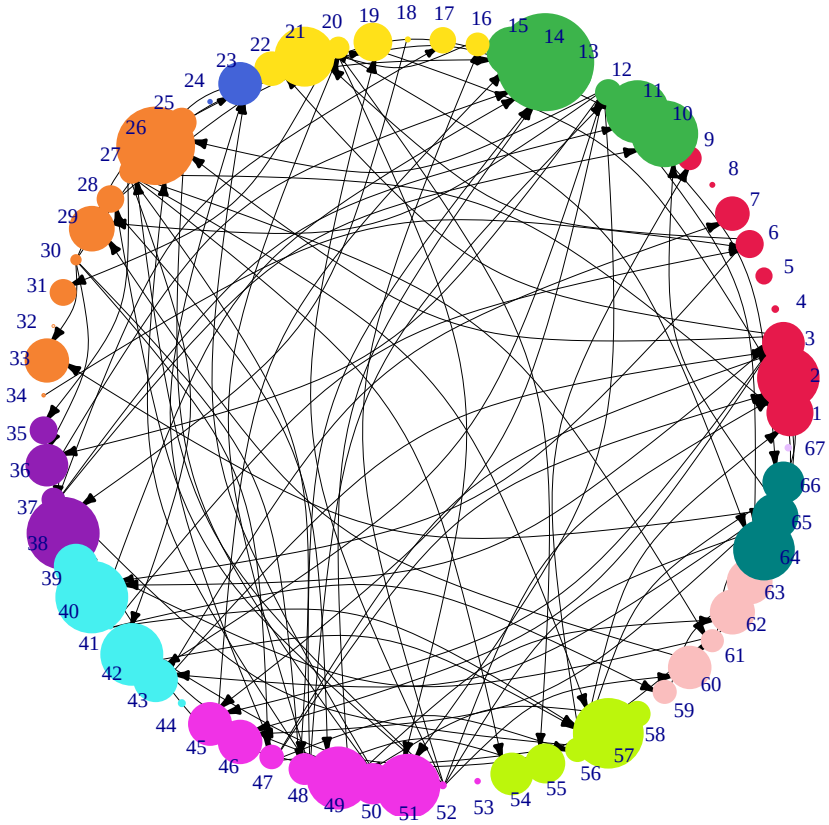

Figure: Graph of $\boldsymbol{E}^{(0.5)}_{\text{LUSC}}$.

qDAGx in Lung Cancer data (cont.)

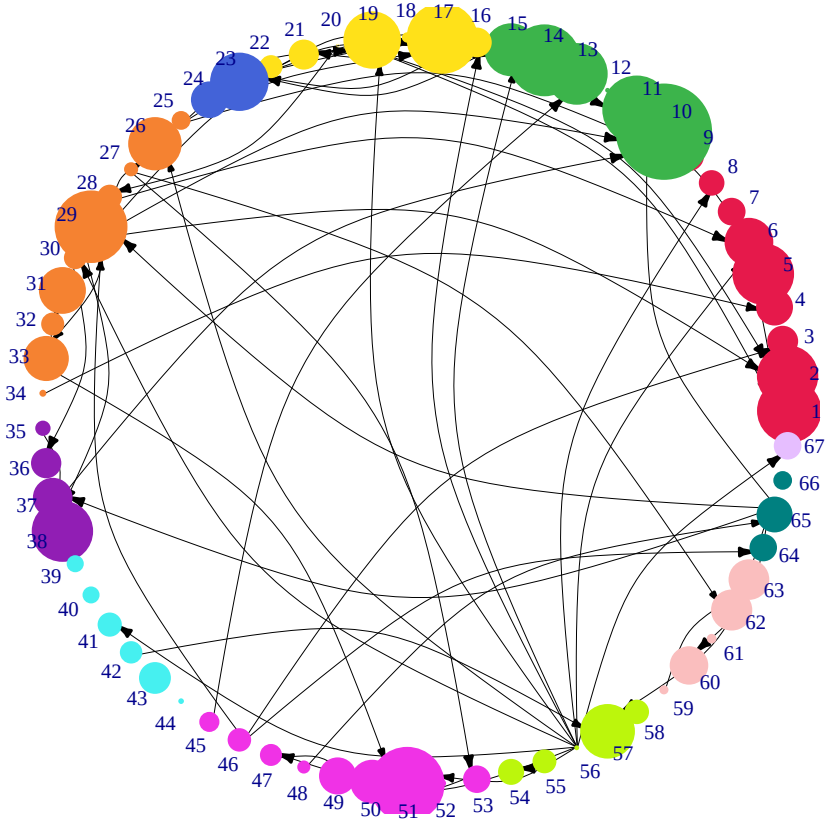

Figure: Graph of $\boldsymbol{E}^{(0.9)}_{\text{LUAD}}$.

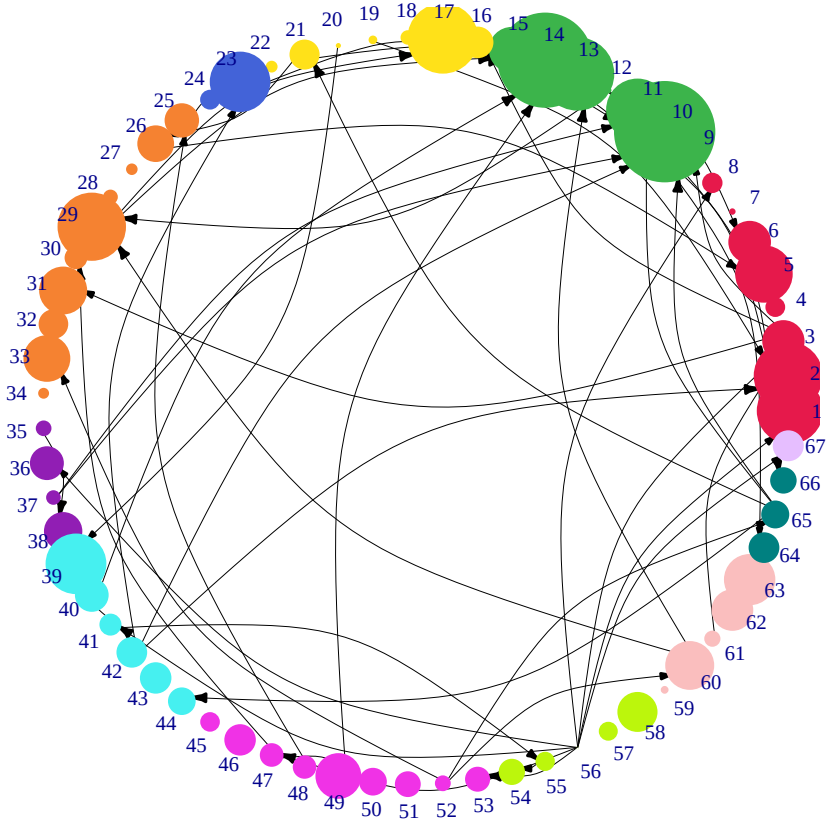

Figure: Graph of $\boldsymbol{E}^{(0.9)}_{\text{LUSC}}$.

qDAGx in Lung Cancer data (cont.)

Table: Directed edges in quantile-DAG estimates which are present in at least 50% of patients and across five out of nine quantile levels, $\tau \in \{0.1, \ldots, 0.9\}$. Common edges in LUAD and LUSC are highlighted.

Lung adenocarcinoma (LUAD)			Lung squamous cell carcinoma (LUSC)		
BAK1←BID	BAD←ATK1S1	BID←ERBB3	BAK1←BID	AKT1, AKT2, AKT3←AKT1S1	CAV1←PGR
CAV1←COL6A1	EGFR←ERBB2	GAPDH←CDH2	CAV1←COL6A1	EGFR←ERBB2	CCNB1←COL6A1
JUN←ERBB3	MAPK1, MAPK3←MAP2K1	MYH11←COL6A1	MTOR←PGR	MAPK1, MAPK3←MAP2K1	MYH11←COL6A1
PCNA←CHEK1	RPS6KB1←PGR		MYH11←FOXM1	RPS6KB1←PGR	RAD51←PGR

Table: Percentage of edges, mean (sd), influenced by second level covariates in quantile-DAG estimates of all patients, across the quantiles $\tau \in \{0.1, \ldots, 0.9\}$.

	only mRNA	only methylation	both
LUAD	13.7 (0.78)	28 (0.86)	58.3 (1.55)
LUSC	13.6 (0.66)	28.1 (0.61)	58.3 (0.85)

qDAGx in Lung Cancer data (cont.)

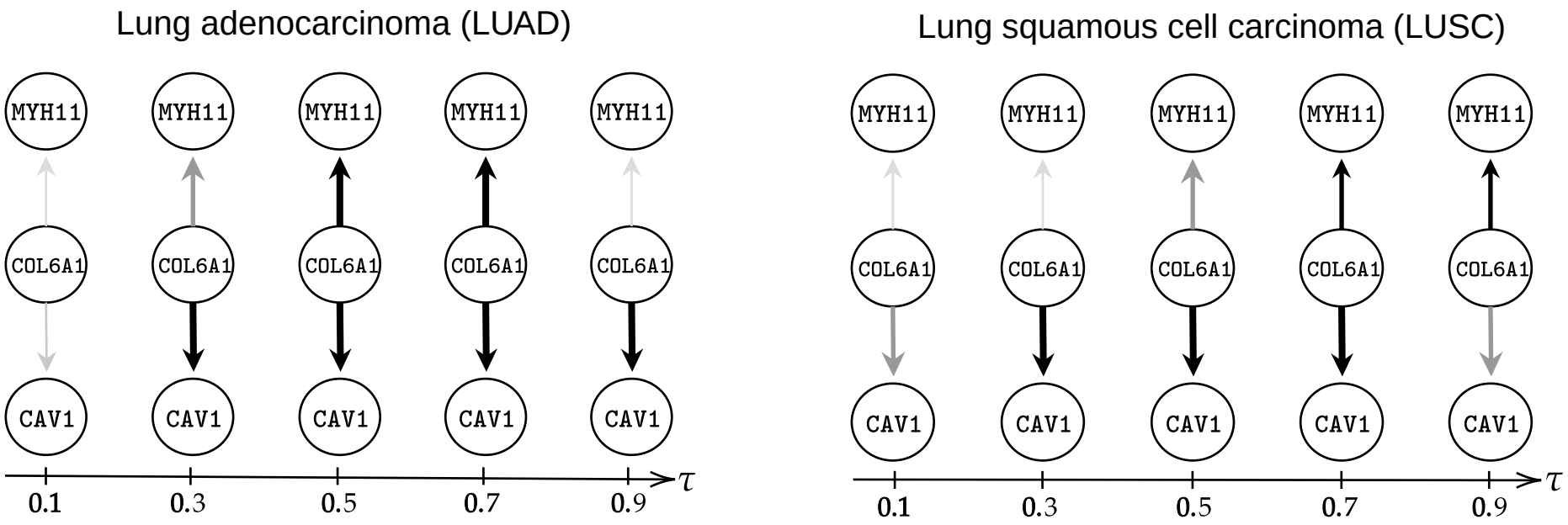

Figure: Prevalence of CAV1←COL6A1 and MYH11←COL6A1 in LUAD and LUSC. Boldness of the edge is proportional to the number of patients in whom the edge was inferred at the specific quantile level τ.

Conclusion and future scope

Contributions:

- **qDAGx**: quantile-DAG learning framework with neither assumptions on likelihood nor on ordering of the nodes.
- Individualized inference via a varying coefficient framework.
- Demonstration of qDAGx in patients with LUAD and LUSC → usefulness in precision medicine.

Future work:

- Theoretical guarantees of estimating quantile-DAG structure similar to Cao et al. (2019); DAG estimation consistency in GGMs.
- Incorporate conditions for preserving increasing nature of quantile estimates (Ali et al., 2016; Yang and Tokdar, 2017).
- Mixture quantile-DAG modeling, possibly by modeling threshold parameter t_{hj} as function of categorical variables like type of cancer, gender etc. And/or relaxing the union-DAG condition.

Ali, A., Kolter, J. Z., and Tibshirani, R. J. (2016). The multiple quantile graphical model. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Cao, X., Khare, K., and Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. *The Annals of Statistics*, 47(1):319–348.

Das, P., Christine, B. P., Ni, Y., Reuben, A., Zhang, J., Zhang, J., Do, K.-A., and Baladandayuthapani, V. (2021). Bayesian hierarchical quantile regression for precision immuno-oncology. *TBD*, 00(0):123–123.

Guha, N., Baladandyauthapani, V., and Mallick, B. (2020). Quantile graphical models: a bayesian approach. *Journal of machine learning research*.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019). Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197.

Yang, Y. and Tokdar, S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, 112(519):1107–1120.

Supplementary Material

Functional formulation of $\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$ in simulations:

1. For $q^* = 0$, $\{\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})\} = (1+\tau^2)\mathbf{1}_n$, where $\mathbf{1}_n$ is the unit vector of dimension n.
2. For $q^* = 1$, $\{\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})\} = \boldsymbol{X}_{k_1}^2 + \log((1+\tau^2)\mathbf{1}_n)$, where k_1 is randomly chosen from $\{1,\ldots,q\}$.
3. For $q^* = 2$, $\{\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})\} = \boldsymbol{X}_{k_1}^2 + \log((1+\tau^2)\mathbf{1}_n) + \exp(\boldsymbol{X}_{k_2})$, where k_1, k_2 are distinct and randomly chosen from $\{1,\ldots,q\}$.
4. For $q^* = 3$, $\{\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})\} = \boldsymbol{X}_{k_1}^2 + \log((1+\tau^2)\mathbf{1}_n) + \exp(\boldsymbol{X}_{k_2}) + \log|\boldsymbol{X}_{k_3}|$, where k_1, k_2, k_3 are distinct and randomly chosen from $\{1,\ldots,q\}$.

Supplementary Material (cont.)

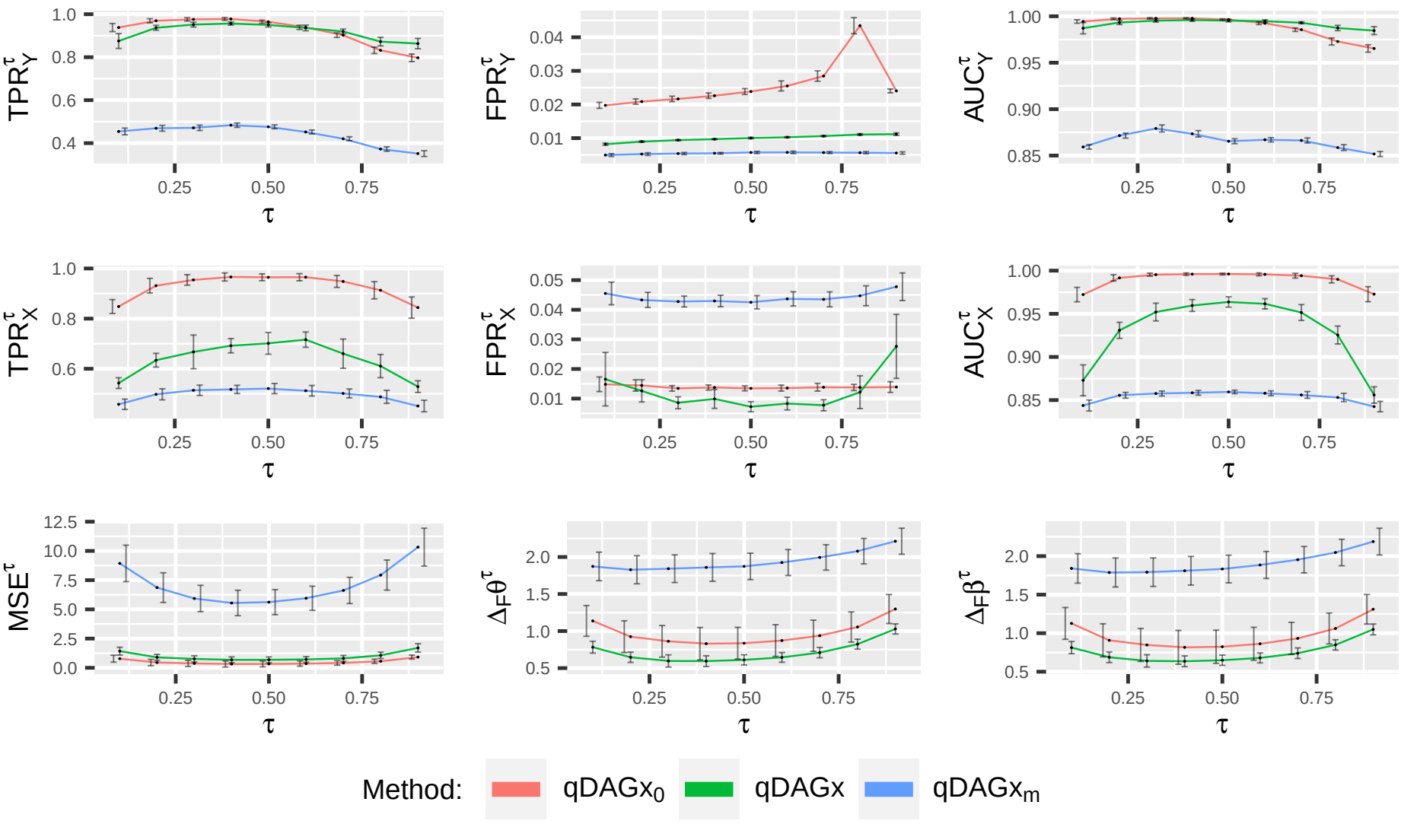

Figure: $p = 50, q = 2, n = 250$. Kendall's' T for the misspecified sequence is 0.25