Bayesian Covariate-Dependent Quantile Directed Acyclic Graphical Models for Individualized Inference

Anindya Bhadra
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- This talk is about *individualized inference* in *quantile DAG models*.
- The *DAG structure* and *edge strengths* are modeled as a function of *individual-specific covariates*.
- We will also discuss an application in precision medicine by modeling individual-specific protein–protein interaction network in lung cancer.
- *Joint work with Ksheera Sagar (Purdue), Yang Ni (Texas A&M) and Veera Baladandayuthapani (Michigan). Supported by NSF Grant DMS-2014371.*

Preliminaries: the Gaussian DAG model

- DAGs provide a natural way to model multivariate interactions in gene or protein interaction networks.
- The most popular approach is a Gaussian DAG.
- Denote n to be the sample size and p to be the number of nodes. The Gaussian DAG model is:

$$Y_{ih} = \sum_{j \in pa(h)} \beta_{hj} Y_{ij} + \varepsilon_{ih}, \quad i \in \{1, \ldots, n\}; \;\; j, h \in \{1, \ldots, p\},$$

where $\varepsilon_{ih} \sim \mathcal{N}(0, \sigma_h^2)$ and $pa(h)$ denotes the parent set of $\boldsymbol{Y}_h$, i.e., the set of nodes $\boldsymbol{Y}_j$s for which there exists an edge $\boldsymbol{Y}_h \leftarrow \boldsymbol{Y}_j$.

Features of a Gaussian DAG (that can be limitations)

- An assumption of *Gaussian* likelihood: susceptible to model mis-specification.
- The parent set $pa(h)$ is assumed known for each h, which can be unrealistic. Or it is extracted by something like the PC algorithm (Spirtes et al., 2000), which does not give a unique DAG.
- Not possible to infer β_{hj} without modeling the entire distribution. For certain diseases, we may want to selectively focus on certain quantiles.
- The parameter β_{hj} is does not depend on individual-specific covariates (no i in it). Unappealing in precision medicine.

Summary of our contributions

- A quantile based approach: free from any specific likelihood assumption (Gaussian or otherwise).
- The DAG structure is inferred (not assumed known)
- Model coefficients ($\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$) are:
  - Specific to a given quantile τ.
  - Depend on individual-specific covariates ($\boldsymbol{X}_{i\cdot}$), using the varying coefficients framework of Hastie and Tibshirani (1993, JRSSB).
- An application in precision medicine of lung cancer.

The individualized model

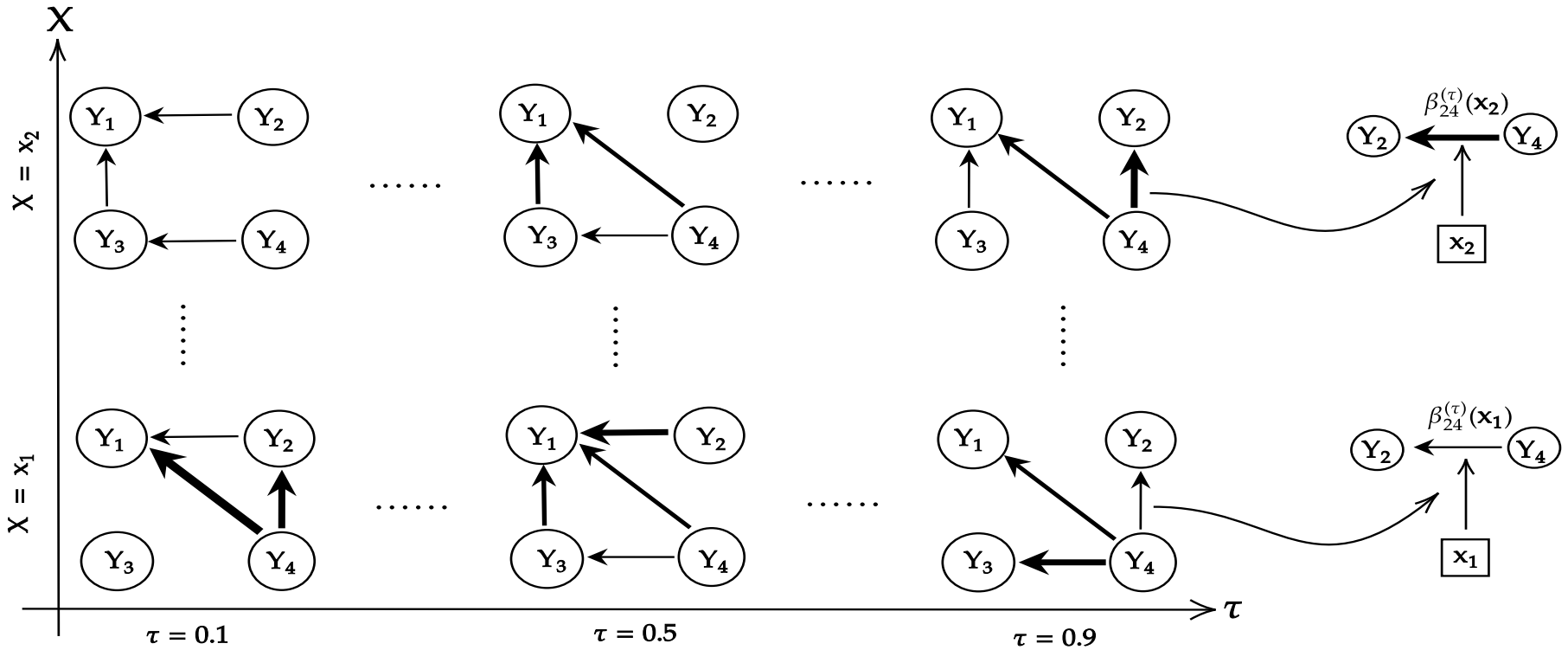
- We model the τth quantile of Y_{ih} as:

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

$$\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \cdot \mathbb{1}\big(|\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})| > t_{hj}\big), \quad \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \sum_{k=1}^{q} f_{hjk}^{(\tau)}(X_{ik}).$$

- Key features:
  - $\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})$ is specific to given τ and covariates $\boldsymbol{X}_{i\cdot}$ for individual i.
  - β is a thresholded version of θ, which is a smooth function of $\boldsymbol{X}_{i\cdot} \in \mathbb{R}^q$.
  - The DAG is learned, along with model parameters.

Overall modeling strategy



- A schematic for $n = 2$ samples, $p = 4$ variables, $q = 1$ covariates, at $\tau = 0.1, 0.5, 0.9$.
- If β is zero (recall: $\beta =$ thresholded θ) then the edge is missing.
- For a certain $\beta_{24}^{(\tau)}(\boldsymbol{X}_{i\bullet}) \neq 0$, the magnitude depends on $\boldsymbol{X}_{i\bullet}$.

Modeling and estimation details

- We impose the *union-DAG condition*. Let

$$\mathcal{QG}_u^{(\tau)} = \bigcup_{i=1}^{n} \mathcal{QG}_i^{(\tau)}.$$

  We restrict $\mathcal{QG}_u^{(\tau)}$ to be a DAG.
  - Reason 1: Same directionality acros all i is biologically justified.
  - Reason 2: Only need to check if union graph is a DAG, not for each i.

- Loss function:

$$L(\tau) = \sum_{i=1}^{n} \sum_{h=1}^{p} \psi_\tau \left(Y_{ih} - \beta_{h0}^{(\tau)}(\boldsymbol{X}_{i\cdot}) - \sum_{j \in pa_i(h)} Y_{ij} \beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \right),$$

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

- Likelihood (asymmetric Laplace):

$$\pi(\boldsymbol{Y} \mid \boldsymbol{X}, \tau, \boldsymbol{\beta}^{(\tau)}) \propto \exp(-L(\tau)) \times \mathbb{1}\left(\mathcal{QG}_u^{(\tau)} \text{ is a DAG} \right).$$

Model for the smooth function θ

- Recall:

$$\beta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) \cdot \mathbb{1}\big(|\theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot})| > t_{hj}\big), \quad \theta_{hj}^{(\tau)}(\boldsymbol{X}_{i\cdot}) = \sum_{k=1}^{q} f_{hjk}^{(\tau)}(X_{ik}).$$

- We model:

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

  - $\widetilde{\boldsymbol{X}_k}^{*}$ is a nonlinear basis expansion for $\boldsymbol{X}_k$.
  - A parameter expanded version of the horseshoe prior on $\boldsymbol{\alpha}_{hjk}^{*}$ and α_{hjk}^{0} for sparse estimation.
  - Normal prior on μ_{hj} (no need to shrink the intercept).
  - Gamma prior on the threshold t_{hj}.

Theoretical properties

- Identifiability: There do not exist $\boldsymbol{\beta}^{(\tau)'} \neq \boldsymbol{\beta}^{(\tau)}$ such that $\pi(\boldsymbol{Y} \mid \boldsymbol{X}, \tau, \boldsymbol{\beta}^{(\tau)}) \equiv \pi(\boldsymbol{Y} \mid \boldsymbol{X}, \tau, \boldsymbol{\beta}^{(\tau)'})$.

- Properties of the marginal prior: The marginal prior on β is a two component mixture of a delta function at zero, and a non-local prior (Johnson and Rossell, 2010).

- Posterior concentration: The posterior of the node-conditional fitted densities concentrate around the truth (rate of convergence in paper).

Simulation results: data generation

- We consider $n \in \{100, 250\},\ p = \{25, 50, 100\},\ q \in \{2, 5\}$.
- Data generation:
  - Covariates $X_1, \ldots, X_q$ are i.i.d. standard normal.
  - WLOG, select an order $Y_1, \ldots, Y_p$ and a true DAG that is 80% sparse.
  - Set true θ as a combination of a variety of linear and nonlinear functions.
  - Set $\beta = \mathbf{1}(\theta > 0.5)$.
  - Calculate the quantile function of $Y_h \mid pa(h)$.
  - Simulate Y_h by inverse cdf method.

Simulation results: competing methods and performance metrics

- Methods under consideration:
  - qDAGx (qDAGx with unknown ordering that is inferred)
  - qDAGx$_0$ (qDAGx with oracle true ordering supplied)
  - qDAGx$_m$ (qDAGx with a known but false ordering fed to it).
  - Additional comparisons with lasso-QR (Wu et al., 2009) in paper, but it does not support individualized coefficients.
- Performance metrics:
  - TPR and FPR in variable (Y) and covariate (X) selection.
  - Area under curve (AUC).
  - Frobenius norm in estimating β and θ.
  - MSE in estimating the true quantile function.

Results

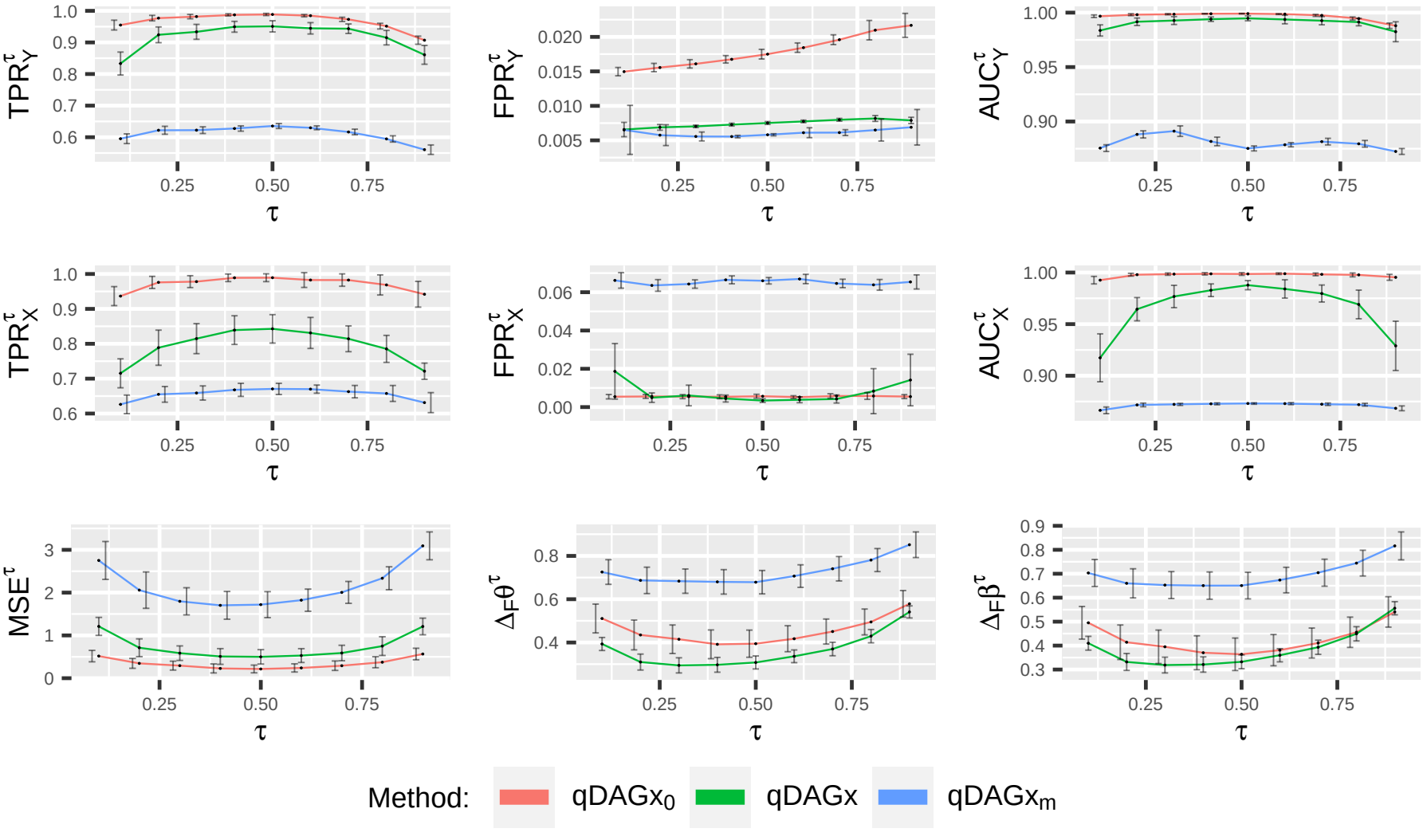


Figure: $p = 25, q = 5, n = 250$. Kendall's' T for the misspecified sequence is 0.5

An application in precision medicine of lung cancer

- Protein expressions of $p = 67$ proteins, denoted as: $Y_1, \ldots, Y_{67}$.
- $q = 2$ covariates: $(X_1, X_2) =$ (mRNA expression, DNA 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.

Individual level inference

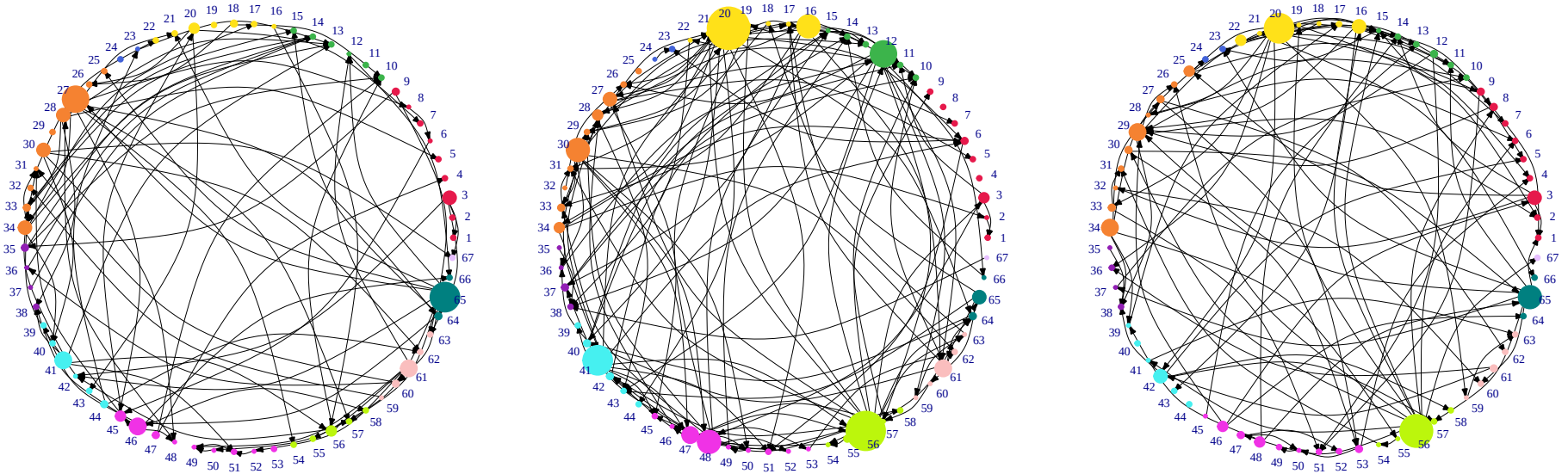


Figure: Quantile graphs at $\tau = 0.1$, 0.5 and 0.9 for a random patient with LUAD.

Table: Map between pathways and colors.

Apoptosis		Breast reactive		Cell cycle		Core reactive		DNA damage response		EMT	
PI3K/AKT		RAS/MAPK		RTK		TSC/mTOR		Hormone receptor		Hormone signaling (Breast)	

Individual level inference

Table: Directed edges present in at least 50% of patients and across 5 out of 9 quantile levels, $\tau \in \{0.1, \ldots, 0.9\}$. Common edges in LUAD and LUSC in blue.

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: Mean (sd) for the percentage of edges influenced by covariates (only mRNA, only methylation, both mRNA and methylation)

	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)

Population level inference

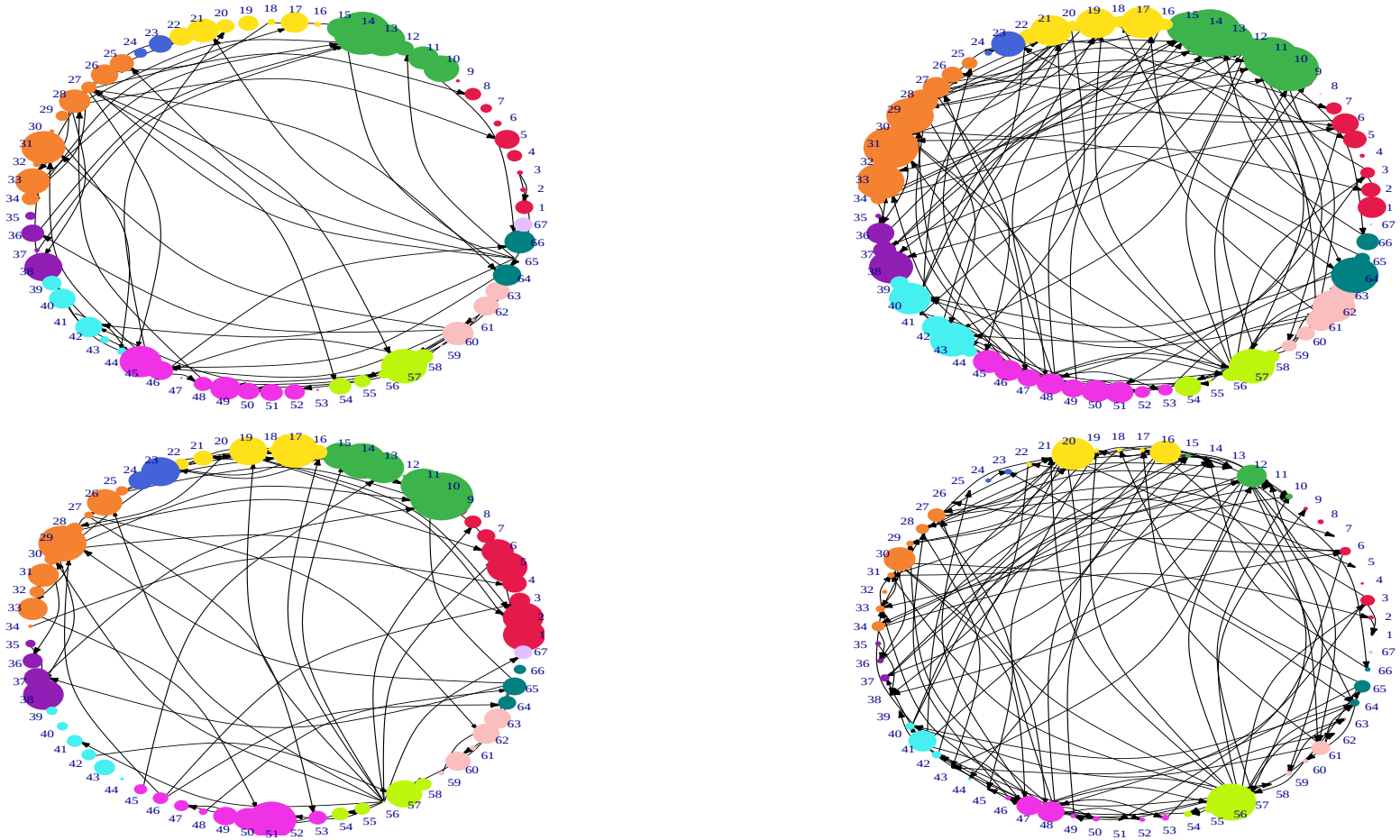


Figure: First 3 panels: aggregated qDAGs, $\boldsymbol{E}_{\text{LUAD}}^{(\tau)}$, for $\tau = 0.1, 0.5, 0.9$. Last panel: $\boldsymbol{E}_{\text{LUAD}}^{(0.5)}$, with node size proportional to out-degree.

Concluding remarks

- We proposed a quantile-based approach for individualized inference in DAG models.
- We do not impose a likelihood, a known ordering of nodes, or global coefficients.
- Presence/absence of edges and their strengths are modeled as functions of individual-specific covariates.
- Currently quite computation-intensive. Increasing scalability, especially in terms of the dimension of covariates (q) should be useful.
- Currently, quantile crossing is not addressed.

References

- Sagar, K., Ni, Y., Baladandayuthapani, V. and **Bhadra, A.** (2023+). Bayesian Covariate-Dependent Quantile Directed Acyclic Graphical Models for Individualized Inference. *(submitted).* [arXiv:2210.08096]