Learning Topological Sorts of Directed Acyclic Graphs

Qiaoling Ye (yeqiaoling@g.ucla.edu) Arash A. Amini Qing Zhou

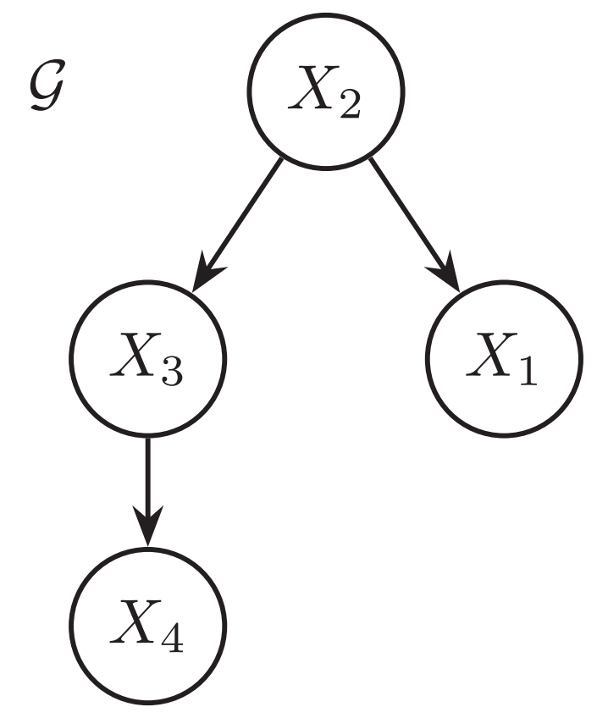
UCLA Department of Statistics

Overview

- Develop a new approach to globally search over the permutation space for a high-scoring Bayesian network by combining global simulated annealing with a fast proximal algorithm operator.

Introduction

- Bayesian network (BN) for $\{X_1, \ldots, X_p\}$, represented by a *directed acyclic graph* (DAG) $\mathcal{G}$, parameterized by (B_0, Ω_0):
  - Coefficients $B_0 := (\beta^0_{ij})$, a weighted adjacency matrix.
  - Noise variance Ω_0, a positive diagonal matrix.
- Topological sort is not unique. Consider permutation $\pi : [p] \mapsto [p]$,
  - Permutation matrix P_π: i^{th} row of P_π is $e^\top_{\pi(i)}$.
  - $B_\pi := P_\pi B_0 P_\pi^\top$.
  - B_π is strictly lower triangular. $\Leftrightarrow$ π is the reversal of a topological sort of $\mathcal{G}$.
- Gaussian BNs can be equivalently represented by a set of linear structural equation models, resulting $X \sim \mathcal{N}(0, \Sigma_0)$, where Σ_0 is positive definite and defined by (B_0, Ω_0).



$$\pi = [4, 1, 3, 2]$$

$$B_0 = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline X_1 & 0 & 0 & 0 & 0 \\ X_2 & \beta^0_{21} & 0 & \beta^0_{23} & 0 \\ X_3 & 0 & 0 & 0 & \beta^0_{34} \\ X_4 & 0 & 0 & 0 & 0 \end{array} \qquad P_\pi = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$B_\pi = P_\pi B_0 P_\pi^\top = \begin{array}{c|cccc} & X_4 & X_1 & X_3 & X_2 \\ \hline X_4 & 0 & 0 & 0 & 0 \\ X_1 & 0 & 0 & 0 & 0 \\ X_3 & \beta^0_{34} & 0 & 0 & 0 \\ X_2 & 0 & \beta^0_{21} & \beta^0_{23} & 0 \end{array}$$

An example of DAG $\mathcal{G}$, its coefficient matrix B_0, and a permutation π. B_π permutes columns and rows of B_0 and is strictly lower triangular.

Cholesky Loss

Define $\Omega_\pi := P_\pi \Omega_0 P_\pi^\top$. Drop the subscript π from P_π, B_π and Ω_π.
Define $L := (I - B)\Omega^{-\frac{1}{2}}$. Define the *Choleskly loss*:

$$\mathcal{L}_{\text{chol}}(L; A) := \frac{1}{2}\text{tr}(ALL^\top) - \log|L|, \qquad (1)$$

where $|L|$ denotes the determinant of L.

Lemma 1: Let $\mathbf{X}$ be sample of size n from a Gaussian BN. Define $\widehat{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. The negative log-likelihood of $\mathbf{X}$ is

$$\ell(L, P) := n \cdot \mathcal{L}_{\text{chol}}(L; P\widehat{\Sigma}P^\top).$$

Proposition: The unique minimizer of $\mathcal{L}_{\text{chol}}(L; A)$ is the Cholesky factor of A. $\mathcal{L}_p$: the set of $p \times p$ lower triangular matrices, and $\ell^*(P) := \min_{L \in \mathcal{L}_p} \ell(L; P)$, then

$$\ell^*(P) = n \cdot \mathcal{L}^*_{\text{chol}}(P\widehat{\Sigma}P^T) = n \cdot \mathcal{L}^*_{\text{chol}}(\widehat{\Sigma}),$$

showing that ℓ^* is invariant to permutations, hence maximum likelihood estimation does not favor any particular ordering.

Regularized Cholesky Score

- Let $\rho_\theta : \mathbb{R} \mapsto [0, \infty)$ be a nonnegative and nondecreasing regularizer with some tuning parameter(s) θ.
- To break the permutation equivalence of the maximum likelihood, we add a regularizer to the Cholesky loss to favor sparse DAGs. We consider the following penalized loss function:

$$f_\theta(L; P) := \ n \cdot \mathcal{L}_{\text{chol}}(L; P\widehat{\Sigma}P^\top) + \sum_{i>j} \rho_\theta(L_{ij}), \qquad (2)$$

  where the penalty is only applied to the off-diagonal entries of a lower triangular matrix L.
- We develop annealing on regularized Cholesky score (ARCS) to minimize (2), in which we use simulated annealing to search for a topological sort that minimizes the regularized Cholesky (RC) score, $f_\theta(P) := \min_{L \in \mathcal{L}_p} f_\theta(L; P)$.
- To obtain the RC score for a given permutation, we develop a proximal gradient algorithm to solve a continuous optimization problem.

Algorithm: Annealing on regularized Cholesky score (ARCS).

```
- Select tuning parameters $\theta$ for $f_\theta(L; P)$ by the Bayesian information criterion.
- Given an initial permutation $P_0$, repeat the following steps N times:
    a) Propose a new permutation.
    b) Estimate the corresponding lower triangular matrix.
    c) Decide to accept the proposal or not.
- Refine the estimated network structure by conditional independence tests.
```

Causal Network Learning

Assuming X_i, $i \in \mathcal{M} \subset \{1, \ldots, p\}$, is under experimental intervention, the joint density becomes

$$p(x_1, \ldots, x_p) = \prod_{i \notin \mathcal{M}} p(x_i | \Pi_i^{\mathcal{G}}) \prod_{i \in \mathcal{M}} p(x_i | \bullet),$$

where $p(x_i|\bullet)$ specifies the distribution of X_i under intervention, $\Pi_i^{\mathcal{G}}$ is the parent set of X_i in DAG $\mathcal{G}$.

Define $\mathcal{O}_{\pi(j)} \subseteq \{1, 2, \ldots, n\}$: the set of observations for which $X_{\pi(j)}$ is not under experimental intervention. Let $\mathbf{X}_{\mathcal{O}_{\pi(j)}}$ be the submatrix of $\mathbf{X}$ with rows in $\mathcal{O}_{\pi(j)}$ and

$$\widehat{\Sigma}^j := \frac{1}{|\mathcal{O}_{\pi(j)}|} \mathbf{X}^\top_{\mathcal{O}_{\pi(j)}} \mathbf{X}_{\mathcal{O}_{\pi(j)}}.$$

Lemma 2: The negative log-likelihood for experimental data can be written as

$$\ell_{\mathcal{O}}(L, P) := \sum_{j=1}^{p} |\mathcal{O}_{\pi(j)}| \mathcal{L}_{\text{chol}}\left(L_j; P\widehat{\Sigma}^j P^\top\right),$$

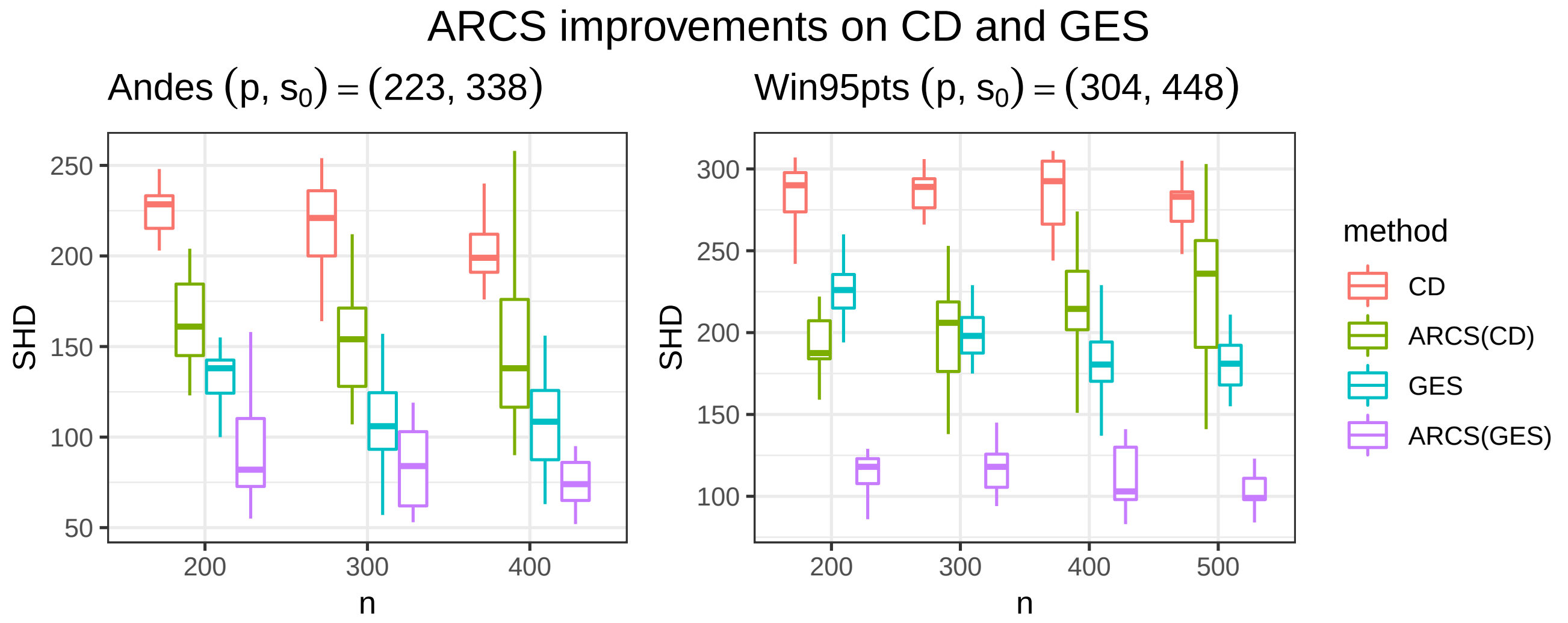
where $L_j \in \mathbb{R}^p$, $L = (L_j) \in \mathcal{L}_p$, and $|L_j| := L_{jj}$ in $\mathcal{L}_{\text{chol}}$ (1).

Networks and Accuracy Metrics

- We used real and synthetic networks to simulate data, where s_0 is the number of edges in the true DAG.
- Structural Hamming distance (SHD) measures the distance between the true and estimated DAGs.

Observational Data

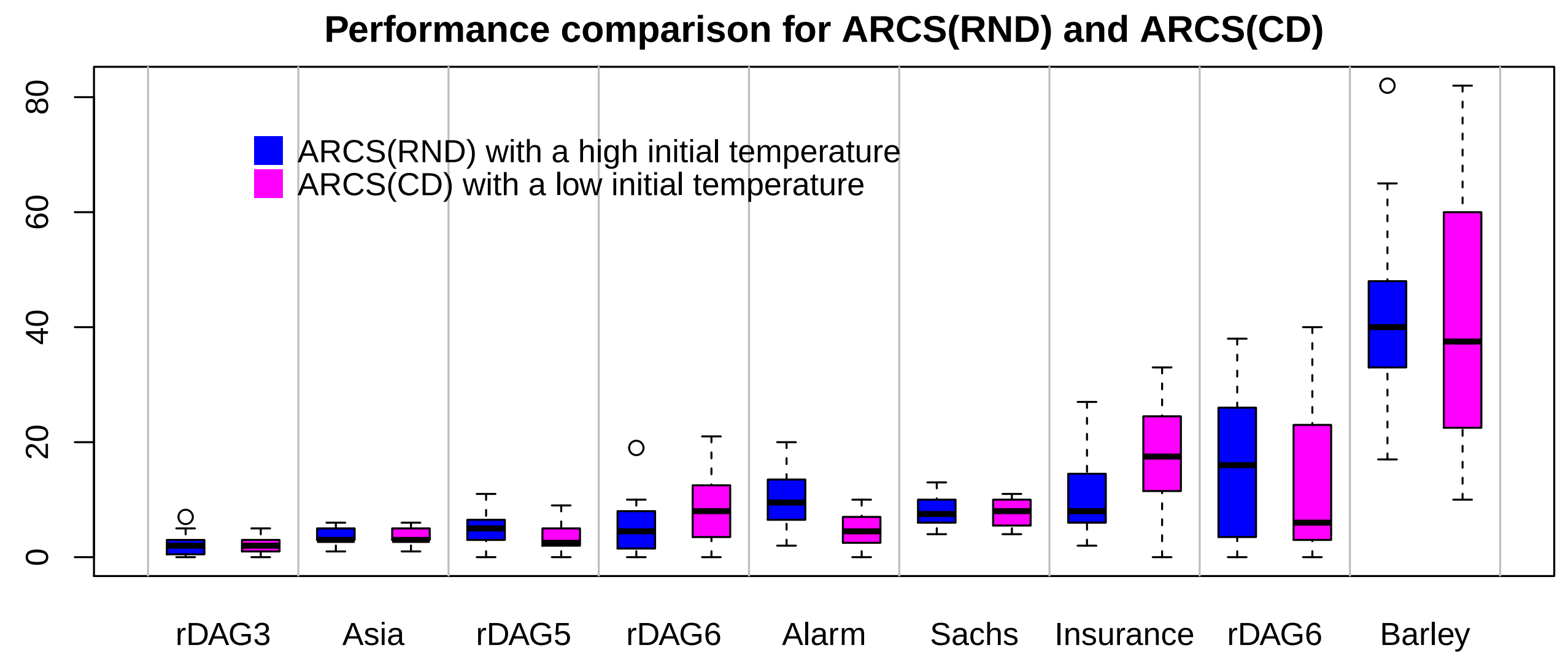
1. ARCS with local estimate initializations.
   - ARCS takes an initial permutation P_0.
   - We initialized ARCS with estimates from CD and GES algorithms, called them ARCS(CD) and ARCS(GES) respectively. Then compared their performances over 20 datasets.
   - ARCS substantially improved accuracies of local algorithms on observational data in both low-dimensional and high-dimensional cases.



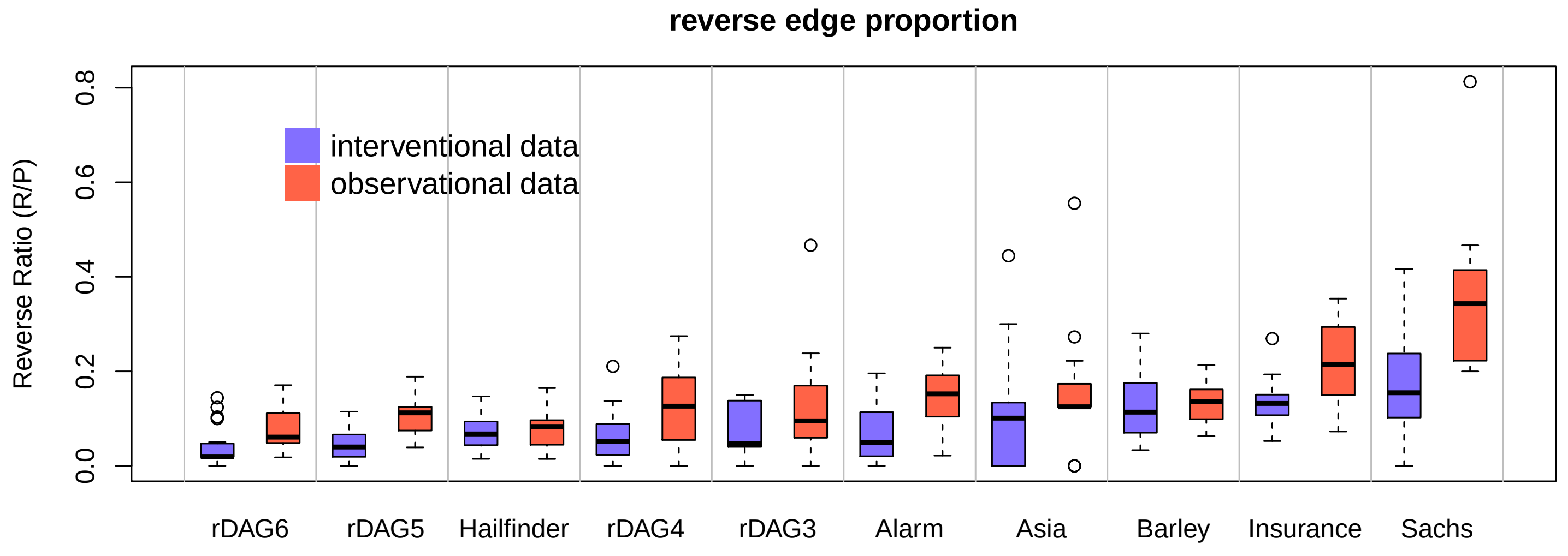
Experimental Data

1. ARCS with local estimate initializations. We initialized ARCS with CD and GIES estimates. ARCS improved accuracies of local estimates on the experimental data as well.

2. ARCS with random initializations. We initialized ARCS with random permutations, ARCS(RND), to test the global search ability of ARCS. Then compared ARCS(RND) with ARCS(CD) for median-size networks.



3. Experimental versus observational. We also compared the performance of ARCS(CD) on experimental and observational data of the same sample size to illustrate the effectiveness of experimental interventions.



Comparison of reversed edge proportion on experimental and observational data using the ARCS(CD) algorithm.