Asymptotic theory of the quadratic assignment procedure for analyzing network data

Lei Shi and Peng Ding
University of California, Berkeley
ICSDS 2024 @ Nice, France

Dyadic data in application

- Many research problems involve dyadic data, which encode pairwise relationships between units
    - Trading conditions between countries
    - Closeness in peer relationship between students
    - Similarity measures in environmental conditions between sites
- Dyadic data are usually represented by square matrices $A = (a_{ij})_{n\times n}$
    - We focus on symmetric dyadic measures with zero diagonals
- A classical question: **whether two dyadic measure matrices A and B are related**.

The quadratic assignment procedure

- QAP or quadratic assignment procedure
- A permutation test that is widely adopted in practice

1. Let $W(A, B)$ be the Pearson correlation coefficient

$$\boxed{\widehat{\rho} = \frac{\sum_{i \neq j}(a_{ij} - \overline{a})(b_{ij} - \overline{b})}{\{\sum_{i \neq j}(a_{ij} - \overline{a})^2\}^{1/2}\{\sum_{i \neq j}(b_{ij} - \overline{b})^2\}^{1/2}};}$$

2. Simultaneously permute the column and rows of A, and compute the permutational distribution

$$\boxed{\mathcal{L}(t; W^\pi) = \frac{1}{n!}\sum_{\pi \in \mathbb{S}_n} \mathbf{1}\{W^\pi \leq t\}, \ \textit{where } W^\pi = W(A_\pi, B).}$$

3. Compute the two-sided p-value p_{QAP} based on the observed value of W and its permutational distribution:

$$\boxed{p_{\text{QAP}} = 1 - \mathcal{L}(|W|; W^\pi) + \mathcal{L}(-|W|; W^\pi).}$$

Example: time-space interaction

- Testing time-space interaction is important in fields such as epidemiology
- With unit level time-space data (R_i, S_i), take closeness measure for time and space, respectively: $a_{ij} = 1\{|R_i - R_j| < \tau\}$ and $b_{ij} = 1\{|S_i - S_j| < \delta\}$
- The statistic $\sum a_{ij}b_{ij}$ measures the time-space interaction
- QAP test: permuting the time (or space) matrix and calculate the permutational p-value
- Also known as the Knox test or Mantel test

Model for dyadic data

- Without a model it is hard to study the problem…
  - How to define two networks are **related**?
  - How to justify the QAP is actually testing such relationship?
- Basic idea: dyadic measures are functions of pairs of individual features
- Let (R_i, S_i) be i.i.d. copies of some random unit-level features (R, S)
- Let $\alpha(r, r')$ and $\beta(s, s')$ be two symmetric functions encoding pairwise relationship between units
  - The features and functions can be either known (e.g. time-space interaction) or unknown (e.g. closeness in friendship)
- The dyadic measure is captured by $a_{ij} = \alpha(R_i, R_j)$ and $b_{ij} = \beta(S_i, S_j)$

Two statistical hypotheses

- Two notions for depicting unrelated dyadic measures
  - **The strong null hypothesis:** The random features that decide the two measures are **independent**

$$\boxed{\mathrm{H}_{0\mathrm{s}} : R \perp\!\!\!\perp S.}$$

  - **The weak null hypothesis:** The dyadic measures have zero correlation

$$\boxed{\mathrm{H}_{0\mathrm{w}} : \mathrm{Corr}\left\{\alpha(R, R'), \beta(S, S')\right\} = 0}$$

- The strong null hypothesis implies the weak null hypothesis

Theory of QAP

Some preliminary notions

- We need the following notations to justify the asymptotic behavior of QAP:

$$\alpha_0 = \mathbb{E}\left\{\alpha(R, R')\right\}, \quad \beta_0 = \mathbb{E}\left\{\beta(S, S')\right\},$$

$$\widetilde{\alpha}(r, r') = \alpha(r, r') - \alpha_0, \quad \widetilde{\beta}(s, s') = \beta(s, s') - \beta_0,$$

$$\eta_{2,\alpha} = \operatorname{Var}\left\{\alpha(R, R')\right\} = \mathbb{E}\left\{\widetilde{\alpha}^2(R, R')\right\}, \quad \eta_{2,\beta} = \operatorname{Var}\left\{\beta(S, S')\right\} = \mathbb{E}\left\{\widetilde{\beta}^2(S, S')\right\}.$$

Define another kernel of degree 2:

$$\phi(r, s; r', s') = \widetilde{\alpha}(r, r')\widetilde{\beta}(s, s'), \tag{3.3}$$

- Related to projection of U-statistics up to different orders

Theory of QAP

- A plug-in estimator for the Pearson correlation:

$$\boxed{\widehat{\rho} = \frac{\widehat{\phi}_0}{\sqrt{\widehat{\eta}_{1,\alpha}} \cdot \sqrt{\widehat{\eta}_{1,\beta}}}}$$

- Meaning of the notation:

Sample covariance

Sample variance

$$\widehat{\phi}_0 = \frac{1}{n(n-1)-1} \sum_{i \neq j} (a_{ij} - \overline{a})(b_{ij} - \overline{b}),$$

$$\widehat{\eta}_{2,\alpha} = \frac{1}{n(n-1)-1} \sum_{i \neq j} (a_{ij} - \overline{a})^2,$$

$$\widehat{\eta}_{2,\beta} = \frac{1}{n(n-1)-1} \sum_{i \neq j} (b_{ij} - \overline{b})^2$$

Theory of QAP

Asymptotic distribution for the Pearson correlation

- Asymptotic sampling distribution for the Pearson correlation

Theorem 3.1 (Asymptotic sampling distribution of $\widehat{\rho}$). *Assume that the covariance kernel in (3.3) is not degenerate in the sense that $\eta_{1,\phi} > 0$, and the following moments are finite:*

$$\mathbb{E}\left\{\alpha^2(R, R')\right\} < \infty, \quad \mathbb{E}\left\{\beta^2(S, S')\right\} < \infty, \quad \mathbb{E}\left\{\phi^2(R, S; R', S')\right\} < \infty.$$

1. *Under* $\mathrm{H}_{0\mathrm{W}}$ *in* (2.1), *we have:*

$$\sqrt{n}\widehat{\rho} \rightsquigarrow \mathcal{N}\left(0, v_{\mathrm{W}}\right), \quad \text{where } \boxed{v_{\mathrm{W}} = \frac{4\eta_{1,\phi}}{\eta_{2,\alpha}\eta_{2,\beta}}}.$$

2. *Under* $\mathrm{H}_{0\mathrm{S}}$ *in* (2.1), *we further have*

$$\eta_{1,\phi} = \eta_{1,\alpha}\eta_{1,\beta}, \quad \sqrt{n}\widehat{\rho} \rightsquigarrow \mathcal{N}\left(0, v_{\mathrm{S}}\right), \quad \text{where } \boxed{v_{\mathrm{S}} = \frac{4\eta_{1,\alpha}\eta_{1,\beta}}{\eta_{2,\alpha}\eta_{2,\beta}}}.$$

Theory of QAP

An example for illustration

- Random features (R, S) with mean zero, unit variance and correlation $\rho_{R,S}$
- Suppose that we have dyadic data as pairwise averages

$$\boxed{\alpha(r, r') = \frac{r + r'}{\sqrt{2}}, \quad \beta(s, s') = \frac{s + s'}{\sqrt{2}}}$$

- Applying Theorem 3.1, we obtain

$$\boxed{\sqrt{n}\widehat{\rho} \rightsquigarrow \begin{cases} \mathcal{N}(0, \mathbb{E}\{(RS)^2\}), & \textit{under } \mathrm{H_{0w}} : \rho_{R,S} = 0; \\ \mathcal{N}(0, 1), & \textit{under } \mathrm{H_{0s}} : R \perp\!\!\!\perp S. \end{cases}}$$

Theory of QAP

A variance estimator

- To estimate the asymptotic variance, we use the following variance estimator:

$$\widehat{v} = \frac{4\widehat{\eta}_{1,\phi}}{\widehat{\eta}_{2,\alpha}\widehat{\eta}_{2,\beta}}, \quad \widehat{\eta}_{1,\phi} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i}^{n}(a_{ij}-\overline{a})(b_{ij}-\overline{b})\right)^2,$$

  - Insight: approximate the first-order projection kernel by a marginal average

Theorem 3.2 (Asymptotic sampling distribution of the studentized $\widehat{\rho}$). *Under the same conditions as Theorem* $\boxed{3.1}$*, we have*

$$\widehat{\eta}_{1,\phi} \to \eta_{1,\phi}, \quad \widehat{\eta}_{1,\alpha} \to \eta_{1,\alpha}, \quad \widehat{\eta}_{1,\beta} \to \eta_{1,\beta}, \quad \widehat{v} \to \frac{4\eta_{1,\phi}}{\eta_{2,\alpha}\eta_{2,\beta}} \quad \mathbb{P}\text{-}a.s.$$

Therefore, under $\mathrm{H}_{0\mathrm{w}}$ *(and thus under* $\mathrm{H}_{0\mathrm{s}}$*) in* ($\boxed{2.1}$)*, we have*

$$\frac{\sqrt{n}\widehat{\rho}}{\widehat{v}^{1/2}} \rightsquigarrow \mathcal{N}(0,1).$$

Theory of QAP

QAP based on $\sqrt{n}\hat{\rho}$

- Under the dyadic data model, the permutation distribution of $\sqrt{n}\hat{\rho}$ can be depicted as follows:

Theorem 3.3 (Permutation distribution of $\widehat{\rho}$). *Assume the non-degeneracy of the kernels α and β: $\eta_{1,\alpha} > 0$ and $\eta_{1,\beta} > 0$. Assume there are univariate functions $\kappa_\alpha(\cdot)$ and $\kappa_\beta(\cdot)$ such that*

$$|\alpha(r, r')| \leq \kappa_\alpha(r) + \kappa_\alpha(r'), \quad |\beta(s, s')| \leq \kappa_\beta(s) + \kappa_\beta(s'), \tag{3.8}$$

with $\mathbb{E}\left\{\kappa_\alpha^4(R)\right\} < \infty$ and $\mathbb{E}\left\{\kappa_\beta^4(S)\right\} < \infty$. Then we have

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\mathcal{L}(t; \sqrt{n}\widehat{\rho}^\pi) - \mathcal{L}(t; \mathcal{N}(0, v_s))\right| = 0, \quad \mathbb{P}\text{-}a.s.$$

where $\mathcal{N}(0, v_s)$ is the asymptotic sampling distribution of $\sqrt{n}\widehat{\rho}$ under H_{0s} from (3.7).

- **The permutation test is valid for the strong null but not the weak null!**

Theory of QAP

QAP based on $\sqrt{n}\hat{\rho}/\hat{v}^{1/2}$

- Under the dyadic data model, the permutation distribution of the studentized Pearson correlation is

Theorem 3.4 (Permutation distribution of the studentized $\widehat{\rho}$). *Under the same condition as Theorem 3.3, we have*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| \mathcal{L}(t; \sqrt{n}(\widehat{\rho}/\widehat{v}^{1/2})^{\pi}) - \mathcal{L}(t; \mathcal{N}(0,1)) \right| = 0, \quad \mathbb{P}\text{-}a.s.$$

- **With studentized Pearson correlation coefficient, the permutation test is asymptotically exact for both strong and weak null hypotheses!**

Theory of QAP

Comparison: asymptotic and permutation tests

Table 2: Comparison of asymptotic and permutation tests with or without studentization

Hypothesis	Criteria	Asymptotic Test	QAP with $\sqrt{n}\widehat{\rho}$	QAP with $\sqrt{n}\widehat{\rho}/\widehat{v}^{1/2}$
H_{0s}	Finite-Sample Exact?	NO	YES	**YES**
	Asymptotically Valid?	YES	YES	**YES**
H_{0w}	Finite-Sample Exact?	NO	NO	NO
	Asymptotically Valid?	YES	NO	**YES**

Extension to MRQAP

- The discussion can be extended to linear models with dyadic measures

$$\boxed{a_{ij} = \theta_0 + \sum_j \theta_k b_{kij} + e_{ij}}$$

- Null hypotheses
  - Strong: independence between a and b
  - Weak: zero linear correlation, or $\vartheta = 0$
- **Studentized regression coefficients give robust Type-I error control!**

Discussion

- Our contribution:
  - We provide a rigorous formulation for a hypothesis testing problem in dyadic network applications
  - We establish the asymptotic theory for QAP and MRQAP. Properly studentized statistics can:
    - guarantee finite-sample exactness under the strong null hypothesis
    - control the asymptotic type one error rate under the weak null hypothesis
- Future problems
  - QAP with degenerate dyadic networks
  - QAP with multiple stratified dyadic networks