NETWORK REGRESSION AND SUPERVISED CENTRALITY ESTIMATION

JUNHUI CAI
University of Pennsylvania

DAN YANG
The University of Hong Kong

WU ZHU
Tsinghua University

HAIPENG SHEN
The University of Hong Kong

LINDA ZHAO
University of Pennsylvania

The centrality in a network is a popular metric for agents' network positions and is often used in regression models to model the network effect on an outcome variable of interest. In empirical studies, researchers often adopt a two-stage procedure to first estimate the centrality and then infer the network effect using the estimated centrality. Despite its prevalent adoption, this two-stage procedure lacks theoretical backing and can fail in both estimation and inference. We, therefore, propose a unified framework, under which we prove the shortcomings of the two-stage in centrality estimation and the undesirable consequences in the regression. We then propose a novel supervised network centrality estimation (SuperCENT) methodology that simultaneously yields superior estimations of the centrality and the network effect and provides valid and narrower confidence intervals than those from the two-stage. We showcase the superiority of SuperCENT in predicting the currency risk premium based on the global trade network.

KEYWORDS: Hub and authority centrality, Network regression inference, Measurement error, Hyperlink Induced Topic Search (HITS) Algorithm, Global trade network, Currency risk premium.

Cai: Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA (email: junhui@wharton.upenn.edu); Yang: Innovation and Information Management, Faculty of Business and Economics, The University of Hong Kong, Hong Kong, China (email: dyanghku@hku.hk); Zhu: Department of Finance, School of Economics and Management, Tsinghua University, Beijing, 100084, China (email: zhuwu@sem.tsinghua.edu.cn); Shen: Innovation and Information Management, Faculty of Business and Economics, The University of Hong Kong, Hong Kong, China (email: haipeng@hku.hk); Zhao: Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA (email: lzhao@wharton.upenn.edu).

1. INTRODUCTION

In economics, finance, operations, sociology, and many other disciplines, agents (individual persons, firms, industries, and countries, etc.) are usually connected through different relationships. The collection of the agents and their relationships is usually represented by a network. One extremely influential concept of networks is the agents' positions, because they can induce a wide range of behaviors, including individuals' decisions on education and human capital investment (Jackson et al., 2017), information sharing and advertisement (Banerjee et al., 2019, Breza and Chandrasekhar, 2019), coalition for exchange or cooperation (Elliott and Golub, 2019), firms' investment decision-making (Allen et al., 2019), the identification of banks that are too-connected-to-fail (Gofman, 2017), and stock returns (Ahern, 2013, Richmond, 2019).

An agent's position or importance is usually captured by the so-called *centrality*, which measures how *central* the agent is in comparison to the others in a network. Since the position induces the agent's behavior and thus shapes certain outcomes, the network centrality is often used as an intermediary to measure the network effects on the outcome of interest (Ahern, 2013, Shao et al., 2018, Richmond, 2019, Allen et al., 2019, Banerjee et al., 2019, Bovet and Makse, 2019). There are many kinds of definitions of centrality (Jackson, 2010, Kolaczyk, 2010), among which we focus on the *hub* and *authority centralities* (Kleinberg, 1999), of which the *eigenvector* centrality is a special case. See a brief overview of centralities and concrete examples on the implication of centralities in Section 2.

The value of network centralities is, therefore, two-fold: first, the centralities have natural implications on the importance of agents; second, the centralities are often used as regressors in a regression to model the *network effects* on some outcome of interest. In practice, the centralities are not directly observable while the network is. Hence, the two-fold value of centralities leads to two goals of this article:

(G1) Estimate centralities from an observed network;

(G2) Estimate and conduct inference of the network effects through the centralities.

In empirical studies including many of the above-cited, these two goals are usually achieved in sequential order, to which we refer as the *two-stage* procedure throughout this article. Stage 1 targets goal **(G1)** solely, which estimates centralities by performing the singular value decomposition (SVD) on the adjacency matrix that represents the network. Stage 2 aims at goal **(G2)** next, which estimates the network effects by regressing the outcome on the *estimated* centralities from Stage 1 and conducts inference using the naive confidence intervals from the regression, an *ad-hoc* inference, ignoring the centrality estimation error. The drawbacks of such two-stage procedure are that: Stage 1 only uses the information from the network to estimate centralities, without resorting to the auxiliary information from the regression on the centralities, which leads to an inaccurate estimation of the centralities due to large observational errors in the network (see more evidence in Section 2); Stage 2 is contingent on Stage 1 – regressing the outcome on the inaccurately estimated centralities further results in an inaccurate estimation of the regression coefficients, thereby invalidating the inference. Consequently, these two stages do not achieve the two goals at their best.

Motivated by the two goals and the shortcomings of the two-stage procedure, we first propose a *unified* framework that encapsulates two models for the corresponding two goals: one network generation model based on centralities for goal **(G1)** and one network regression model for the dependency of the outcome on the centralities for goal **(G2)**. We further propose a novel *supervised network centrality estimation (SuperCENT)* methodology that accomplishes both **(G1)** and **(G2)** *simultaneously*, instead of sequentially. SuperCENT exploits information from the two models – the network regression model contains auxiliary information on the centrality in addition to the network, providing *supervision* to the centrality estimation. The supervision effect can improve the centrality estimation, which in turn benefits the network regression. In

other words, the centrality estimation and the network regression complement and empower each other.

Under the unified framework, we derive the theoretical convergence rates of the centralities and the regression coefficients estimators via the two-stage and SuperCENT, as well as their asymptotic distributions, which can be used to construct confidence intervals. Comparing the two methods, SuperCENT universally dominates the two-stage theoretically and empirically in terms of centrality estimation **(G1)** as well as the estimation and inference of the network regression coefficients **(G2)**. We summarize our contributions of this article in the following.

1. To the best of our knowledge, we are the first to provide a unified framework to study properties of centrality estimation, centrality inference, and the subsequent network regression analysis when a noisy network is observed.
2. We demonstrate that the common practice of two-stage can be problematic. For centrality estimation **(G1)**, the two-stage centrality estimates using SVD in Stage 1 are inconsistent under large noise in the network. The same inconsistency phenomenon appears when we estimate the true underlying network. This finding of inconsistency extends the phase transition phenomenon of the singular vectors (Shabalin and Nobel, 2013) and eigenvectors (Shen et al., 2016) to the network centrality problem. For the network regression **(G2)**, the centrality coefficients estimates are *biased* and *inconsistent* given the inconsistent centrality estimates from Stage 1 and the ad-hoc inference can be either *conservative* or *invalid* depending on the network noise level.
3. We show theoretically and empirically that the proposed SuperCENT dominates the two-stage universally. For **(G1)**, SuperCENT yields superior estimations of both the centralities and true network over the two-stage. For **(G2)**, SuperCENT can mitigate the coefficients estimation bias and thus boost the estimation accuracy under large network noise thanks to the superior centrality estimation. In addition, SuperCENT provides confidence intervals that are *valid* and *narrower* than the ad-hoc two-stage confidence intervals.
4. Lastly, we apply SuperCENT and the two-stage to predict the currency risk premium, based on an economic theory on the relationship between a country's currency risk premium and its importance within the global trade network. We show that a long-short trading strategy based on the SuperCENT centrality estimates yields a return three times as high as those by the two-stage procedure. Furthermore, SuperCENT can verify the economic theory via a rigorous statistical test while the two-stage fails.

As a concrete manifestation of our contributions, we perform a toy experiment as follows. The network is constructed by perturbing a true network with noise of different level σ_a. The regression model includes both the hub and authority centralities, $\boldsymbol{u}$ and $\boldsymbol{v}$, with their corresponding coefficients β_u and β_v, as well as other covariates[1]. Figure 1 shows the performance of the two-stage procedure and SuperCENT in terms of estimation of $\boldsymbol{u}$ and β_u as well as the coverage probability and the width of confidence interval (CI) for the hub centrality coefficient CI_{β_u} with varying network noise levels σ_a. Figure 1A shows the error of the hub centrality estimation. As the noise level increases, the two-stage estimate becomes increasingly inaccurate and eventually orthogonal to the truth, corroborating our inconsistency finding of the two-stage under large network noise. On the contrary, SuperCENT estimates $\boldsymbol{u}$ very accurately until the network noise becomes very large. Even then, the accuracy of the SuperCENT estimate does

[1]The detail of the network model is: $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E} \in \mathbb{R}^{256\times 256}$, where the true hub centrality is $\boldsymbol{u}$, the true authority is $\boldsymbol{v}$, both of which are scaled to have norm $\sqrt{256}$, and $\boldsymbol{E}$ contains i.i.d. normal random variables with mean zero and variance σ_a^2. For the regression model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + 16\boldsymbol{u} + \boldsymbol{v} + \boldsymbol{\epsilon} \in \mathbb{R}^{256\times 1}$ where $\boldsymbol{\beta}_x = (1, 3, 5)^\top$, the covariate matrix $\boldsymbol{X}$ consists of a column of 1's and two columns whose entries follow $N(0, 1)$ independently, and $\boldsymbol{\epsilon}$ follows $N(\boldsymbol{0}, \sigma_y^2 \boldsymbol{I}_{256})$ with $\sigma_y^2 = 2^{-4}$. We vary the network noise level $\sigma_a \in 2^{1,1.5,\ldots,5}$ and each configuration is repeated 500 times.

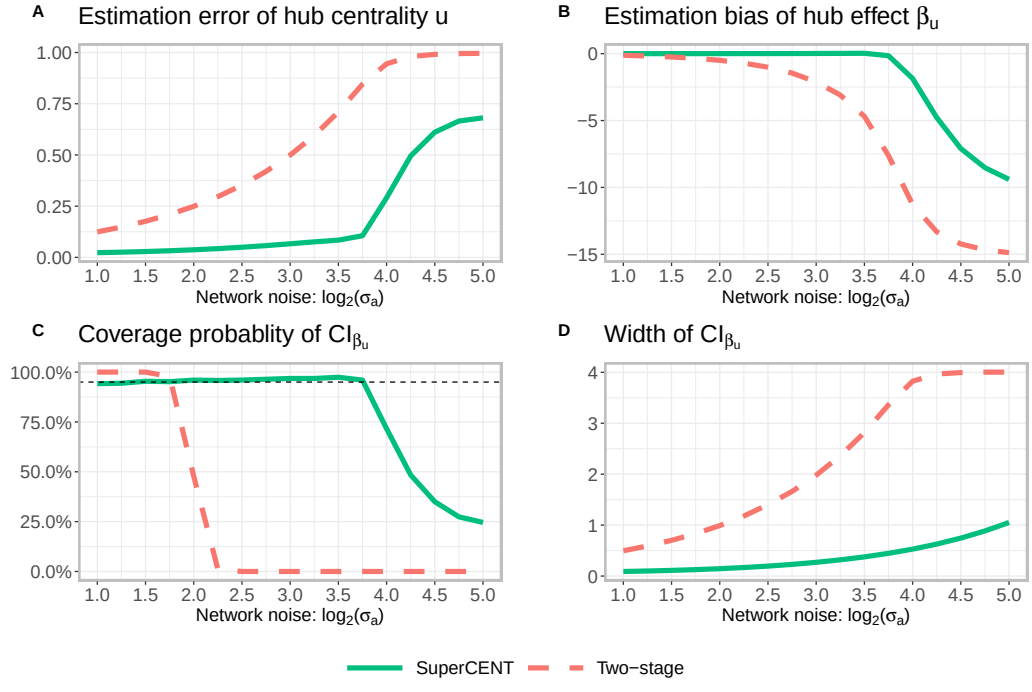


FIGURE 1.—Comparison between the two-stage and SuperCENT in terms of the estimations of $\boldsymbol{u}$ and β_u as well as the coverage probability and the width of CI_{β_u} varying the network noise level σ_a. The performance of the two-stage is shown in the red dashed line and SuperCENT in the green solid line. Subfigure A shows the estimation error of the hub centrality $\sin(\angle(\hat{\boldsymbol{u}}, \boldsymbol{u}))$, i.e., sine of the angle between the true hub centrality $\boldsymbol{u}$ and the estimate $\hat{\boldsymbol{u}}$. As the noise level increases, the two-stage hub centrality estimate using SVD becomes increasingly inaccurate and eventually orthogonal to the truth. SuperCENT estimates the centrality very accurately until σ_a becomes really large. Even then, the accuracy of the SuperCENT estimate does not deteriorate to zero thanks to the auxiliary information from the regression. Subfigure B shows the bias in estimation of β_u, i.e., $\hat{\beta}_u - \beta_u$. The inconsistency of the centrality estimation using SVD results in an attenuation bias in the estimation of β_u of Stage 2, while SuperCENT alleviates the attenuation bias. Subfigures C and D show the coverage probability and the width of the 95% confidence interval of the hub centrality coefficient, CI_{β_u}, respectively. The ad-hoc two-stage confidence interval is either conservative or invalid depending on the network noise level σ_a while SuperCENT provides a valid and narrower confidence interval until σ_a becomes unreasonably large.

not deteriorate to zero thanks to the auxiliary information from the regression. Figure 1B exhibits the estimation bias of $\hat{\beta}_u$: the two-stage estimate suffers from a severe attenuation bias while SuperCENT alleviates the bias. Figures 1C-D illustrate the coverage probability and the width of the confidence interval of β_u. Depending on the noise level in the network model, the ad-hoc two-stage inference has different undesirable consequences: when the noise is small, the ad-hoc two-stage confidence interval is still valid but conservative; when the noise is large, the ad-hoc two-stage confidence interval is invalid and wider than necessary. In contrast, SuperCENT provides a valid and narrower confidence interval until the network noise becomes unreasonably large.

Our paper contributes to several lines of inquiry in the network and econometrics literature on network modeling, network regression with centralities, covariate-assisted network modeling, network effect modeling, and measurement error. First, the proposed unified framework unites the literature on the noisy network and network regression with centralities. Most existing network literature focuses on only one of the two aspects in our unified framework. On one

hand, in the presence of noisy network generation, there are many empirical studies (Lakhina et al., 2003, Banerjee et al., 2013, Breza et al., 2020, Zhu and Yang, 2020) and many that try to estimate or recover the true network without involving centrality (Butts, 2003, Handcock and Gile, 2010, Chandrasekhar and Lewis, 2011, Le et al., 2018, Newman, 2018, Rohe, 2019). On the other hand, many researchers focus on the network regression model with centralities while ignoring the noise of the centrality estimation inherited from the noise of the network (Ahern, 2013, Hochberg et al., 2007, Shao et al., 2018, Richmond, 2019, Allen et al., 2019, Liu, 2019, Banerjee et al., 2019, Bovet and Makse, 2019, Fogli and Veldkamp, 2021).

Our unified framework is also connected to the line of research related to network with covariates supervision (Newman and Clauset, 2016, Zhang et al., 2016, Li et al., 2016, Graham, 2017, Binkiewicz et al., 2017, Yan et al., 2019, Ma et al., 2020, Chen et al., 2021). One major difference is that SuperCENT uses both the covariates and the response to supervise, instead of only the covariates. And they focus mostly on network formation or community detection.

In econometrics, there has been a significant effort to model the network effect on an outcome of interest through regression. See De Paula (2017) for a review on the econometrics of network models. One popular approach follows the pioneer work of Manski (1993), the "reflection model" (Lee, 2007, Bramoullé et al., 2009, Lee et al., 2010, Hsieh and Lee, 2016, Zhu et al., 2017). This approach handles the network effect through the observed adjacency matrix itself, not through centralities like ours. There is also a recent surge of literature in network recovery based on the reflection model (De Paula et al., 2019, Battaglini et al., 2021). Literature on this approach mainly focuses on the issue of identifiability, while our work attends to both the estimation and inference of the network effect. Nevertheless, it is possible to incorporate our model into the reflection model by assuming a low-rank structure on the true underlying network. Another popular approach assumes that the outcome depends on individual fixed effects, which are only estimable by imposing constraints or penalties, and the role of the network is cast through the Laplacian matrix, such that connected nodes share similar individual fixed effects (Jochmans and Weidner, 2019, Li et al., 2019, Le and Li, 2020). This approach emphasizes the network homophily, while ours concentrates on nodes' position or importance in the network using centralities.

Lastly, our methodology further contributes to the measurement error literature. Most literature concerns a regression setup where the covariates are observed with errors, which leads to bias in the coefficient estimation (Garber and Klepper, 1980, Griliches, 1986, Pischke, 2007, Wooldridge, 2015, Abel, 2017). We extend it to the network regression problem. Specifically, the two-stage procedure resembles the measurement error problem where the estimated centralities that are used as covariates in the regression of Stage 2 contain estimation error. Nevertheless, the derivation of the two-stage bias is not a trivial extension of the classical results because it involves the asymptotic joint distributions of the two-stage centrality estimators. Furthermore, SuperCENT corrects the bias problem in the regression coefficient estimation induced by the estimation errors and provides valid inference for the regression coefficients.

The rest of this article is organized as follows. Section 2 reviews the concept and properties of networks, network centralities, noisy networks, and provides concrete examples of network centralities and their effects on the outcome variables. Section 3 formally introduces the unified framework and makes connections with the existing work. Details of the proposed SuperCENT methodology are provided in Section 4. Theoretical properties are studied in Section 5 and the simulation study is demonstrated in Section 6. Section 7 presents the case study of the global trade network centralities and their relationships with risk premiums. Section 8 concludes with a summary and future work. Some concrete mathematical expressions, a special case of an undirected network with the eigenvector centrality, more simulation results, additional information of the case study, and the proofs are delegated to the supplementary materials. A new R package called `SuperCENT` implements the methods (https://jh-cai.com/SuperCENT).

2. NETWORK AND NETWORK CENTRALITY

In this section, we provide background knowledge and a literature review on networks and network centralities. We review how network centralities are used in various fields, particularly via the above-mentioned two-stage procedure, and point out the challenge in centralities estimation due to observational errors in networks. Sophisticated readers may skip this section.

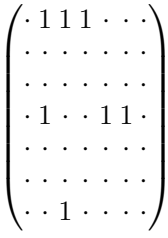
In a network, the nodes are agents involved in a network of relationships and the edges represent the relationships between the nodes. The edges can be directed or undirected depending on whether the relationships are reciprocal. For a directed network $\mathcal{G}$ composed of n nodes V and a set of weighted directed edges $E \subseteq V \times V$, it can be represented by an *asymmetric* adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ where a_{ij} is the weighted edge from node i to j.

A full description of the network depends on all the nodes and edges, whose information is too much to be thoroughly understood in empirical analysis. To feasibly analyze the network, researchers usually resort to dimension reduction or low-rank approximation to extract the characteristics and structural properties. Centralities are common low-rank summaries of the network information. Depending on goals and domain knowledge, researchers have used multiple versions of centralities. We refer readers to Chapter 2 of Jackson (2010), Chapter 4 of Kolaczyk (2010) or Chapter 1 of Graham and De Paula (2020) for a comprehensive introduction to centralities.

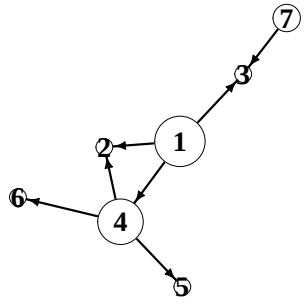
In this article, we focus on the *hub* and *authority centralities* for directed networks (Kleinberg, 1999, Benzi et al., 2013), which can be reduced to eigenvector centrality for undirected networks. For directed networks, there is a distinction between the giver and the recipient, such as the citee-citor in citation networks or webpage networks, the exporter-importer in trade networks, and the investor-investee in investment networks. The concept of "hubs and authorities" originated from web searching. Intuitively, the hub centrality of a web page depends on the total level of authority centrality of the web pages it links to, while the authority centrality of a web page depends on the total level of hub centrality of the web pages it receives links from. Similar intuition can be applied to citation networks where the hub centrality of a paper reveals the quality of a survey paper while an authoritative paper is one that is cited a lot by well-respected survey papers.

Let us use a toy example to further illuminate the intuition behind the hub and authority centralities. Consider a citation network where each paper is a node and an edge from Paper A to Paper B indicates Paper A cites Paper B. Figure 2a shows an example of the adjacency matrix of such network. Figures 2b and 2c show the same network with different node sizes: the node sizes in Figure 2b are proportional to the hub centralities while those in Figure 2c are proportional to the authority centralities.

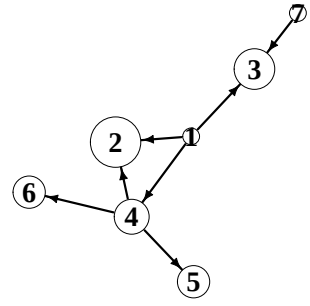
To understand the hub centrality, note that Papers 1 and 4 are the major citors: they both cite three papers with Paper 2 being the common one. Except for the common one, Paper 4 cites Papers 5 and 6, which are only cited by Paper 4, and Paper 1 cites Papers 4 and 3, among which Paper 3 is cited twice. Therefore, compared with Paper 4, Paper 1 cites the same number of papers with one being cited more than the others. This makes the hub centrality of Paper 1 larger than that of Paper 4. One can think of Paper 1 as a better survey paper than Paper 4. Paper 7 cites only one paper, which makes its hub centrality smaller than Papers 1 and 4. The rest of the papers have small hub centrality since they do not cite other papers. As for the authority centrality, attention should be given to citees. Papers 2 and 3 both have two citations, but Paper 2 is cited by Papers 1 and 4 while Paper 3 is cited by Papers 1 and 7. Observe that Paper 4 as a hub is more influential than Paper 7. So the authority centrality for Paper 2 is the highest, followed by Paper 3. For the same reason, Paper 4 has higher authority centrality than Papers 5 and 6, since Paper 4 is cited by Paper 1 while Papers 5 and 6 are cited by Paper 4.



(a) The adjacency matrix of the citation network. For readability, we denote 0 as ·.



(b) Node size by hub centralities.



(c) Node size by authority centralities.

FIGURE 2.—A toy network to illustrate the hub and authority centrality.

One can obtain the hub and authority centralities using an iterative method. Let u_i denote the hub centrality and v_i denote the authority centrality for node i. Kleinberg (1999) proposes to initiate each u_i and v_i with certain nonzero value and then iteratively update as follows:

$$\boldsymbol{v}^{(k)} \leftarrow \mathbf{A}^\top \boldsymbol{u}^{(k-1)} \quad \text{and} \quad \boldsymbol{u}^{(k)} \leftarrow \mathbf{A}\boldsymbol{v}^{(k)} \quad \text{for } k = 1, 2, 3, \ldots \tag{1}$$

This iterative algorithm is shown to converge under some regularity conditions with proper normalization; and the hub and authority centralities are the final $\boldsymbol{u}^{(\infty)}$ and $\boldsymbol{v}^{(\infty)}$ after convergence. This iterative algorithm is also well known as the power method to compute the leading left and right singular vectors of $\boldsymbol{A}$ (Van Loan and Golub, 1996). Due to the equivalence between singular value decomposition (SVD) and eigen-decomposition, the hub centrality is the leading left singular vector of $\boldsymbol{A}$ and the leading eigen-vector of $\boldsymbol{A}\boldsymbol{A}^\top$ while the authority centrality is the leading right singular vector of $\boldsymbol{A}$ and the leading eigen-vector of $\boldsymbol{A}^\top\boldsymbol{A}$.

The hub and authority centralities and their variants are widely used in many fields to study how network positions affect a particular outcome of interest. In practice, as mentioned in Section 1, one needs to estimate the centralities and then use them as regressors in the subsequent regression to estimate the network effects. To achieve these two goals, the naive two-stage procedure has been widely used in the empirical studies although it lacks statistical justifications. In the following, we showcase some concrete examples on how centralities and the two-stage procedure are used in portfolio management, finance, and social media.

In portfolio management, recent research shows that, for a trade network of firms or industries or countries, a strategy that shorts portfolios of nodes with high centralities and longs those with low centralities yields significant excess return, where they use the two-stage procedure to provide empirical evidence (Hochberg et al., 2007, Ahern, 2013, Richmond, 2019). An accurate centrality estimate, therefore, can significantly boost the investment return. As a matter of fact, our case study in Section 7 shows that such a long-short strategy based on the centrality estimated using SuperCENT yields return three times as high as the existing method. In finance, financial institutions such as banks are usually linked through debt or equity, and thus an adversarial shock to one institution can be propagated to others via the debt or equity network (See Elliott et al. (2014), Glasserman and Young (2016), Vohra et al. (2020) and references therein). When a central institution is subject to a severe adversarial shock, the shock

will propagate and the impact will be significantly amplified, resulting in a systemic risk for the whole economy. Thus, identifying the central institutions in the network is the key for policy-makers to impose additional supervision to mitigate concerns of systemic risk. To support their claims, many of the above-cited use the two-stage procedure for empirical evidence by regressing risk metrics on the estimated centrality of the financial institutions in a certain network. In social media such as Twitter or Facebook, networks often serve a role in information transmission or sharing. Individuals in the center of the social network, i.e., the "influencers", can significantly expedite information dissemination (Shao et al., 2018, Bovet and Makse, 2019). Identifying these central individuals can have a wide range of implications from marketing to information censorship.

One challenge of measuring centralities is that we often observe networks with observational error due to the cost of data collection (Lakhina et al., 2003, Banerjee et al., 2013, De Paula, 2017, Breza et al., 2020, Zhu and Yang, 2020). For example, to measure the social connections between people, researchers usually use the friendship on Facebook or Twitter to measure the tie, which is obviously not a perfect measure of the social connection strength. In particular, Banerjee et al. (2013) uses the self-reported friendship in villages of India to measure the social tie between villagers, which is subject to self-reporting and subjective bias. Zhu and Yang (2020) uses the patent citations to measure the knowledge flow between companies, which neglects the communication among workers or executives. The measurement noise of the links could significantly worsen the estimation of the centrality; see Borgatti et al. (2006), Frantz et al. (2009), Wang et al. (2012), Martin and Niemeyer (2019) and the references therein. The inaccurate centrality estimation will further affect the subsequent analysis.

3. A UNIFIED FRAMEWORK

3.1. *Set-up and notation*

We observe a sample of n observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where $y_i \in \mathbb{R}$ is the response and $\mathbf{x}_i \in \mathbb{R}^p$ is the vector of p covariates for the i-th observation as in the multivariate regression setting. The data can be represented in matrix form. Let $\boldsymbol{y} \in \mathbb{R}^n$ denote the column vector of outcome and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the design matrix with n rows and p columns. We consider the fixed design by treating $\boldsymbol{X}$ fixed. In addition, we observe a weighted and directed network $\mathcal{G} = (V, E)$ representing connections between the observations. Here, $V = (1, 2, \ldots, n)$ are the nodes corresponding to n observations and $E \subset V \times V$ is the set of directed edges. The directed network $\mathcal{G}$ can be represented by an asymmetric adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ where $a_{ij} = w_{ij}$ if $(i, j) \in E$ with edge weight $w_{ij} \in \mathbb{R} \backslash \{0\}$ and $a_{ij} = 0$ otherwise.

We adopt the following notation conventions. Given a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$, define its Frobenius norm: $\|\boldsymbol{B}\|_F = \sqrt{\sum_{ij} b_{ij}^2}$, its L_2 operator norm: $\|\boldsymbol{B}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|\boldsymbol{B}\mathbf{x}\|_2$, and the vectorization operator that converts $\boldsymbol{B}$ into a column vector of dimension mn as:

$$\mathrm{vec}(\boldsymbol{B}) = (b_{11}, \ldots, b_{m1}, b_{12}, \ldots, b_{m2}, \ldots, b_{1n}, \ldots, b_{mn})^\top.$$

$\boldsymbol{P}_{\boldsymbol{B}}$ denotes the projection matrix that projects onto the column space of $\boldsymbol{B}$.

3.2. *A unified framework*

The hub and authority centralities are widely used in many fields as shown in Section 2, thanks to their natural implication on network positions. However, literature that investigates

their statistical properties is scarce. The most direct model for them is based on the low-rank decomposition of the network as follows: the adjacency matrix for the observed directed network is generated by some underlying true hub and authority centralities corrupted with noise,

$$\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{E} = d\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E} \tag{2}$$

where the true authority and hub centralities $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are the parameters to be estimated, $\mathbf{A}_0$ is the true network, and $\boldsymbol{E}$ is the noise.

Such a low-rank mean plus noise model has been commonly adopted for matrix estimation or matrix denoise (Shabalin and Nobel, 2013, Yang et al., 2016) and matrix completion (Candes and Plan, 2010, Negahban and Wainwright, 2011). Since the hub and authority centralities are the *leading* left and right singular vectors of $\boldsymbol{A}_0$ respectively, it is natural to consider the noiseless *rank-one* structure for the network $\boldsymbol{A}_0$. This framework can be extended to general rank $r \leqslant (n-p)/2$, although the implications of the *non-leading* singular vectors as centralities are unclear.

Note that $\boldsymbol{u}$ and $\boldsymbol{v}$ are only identifiable up to a scalar. Typically, in SVD, people assume $\boldsymbol{u}$ and $\boldsymbol{v}$ have unit length. However, in our framework of network analysis, we assume $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \sqrt{n}$, in view of the fact that the network can grow and consequently the centralities should roughly be on the same scale no matter how large the network is. Furthermore, we assume the noise matrix has entries with zero mean, i.e., $\mathbb{E}[e_{ij}] = 0$ for $i, j = 1, 2, \ldots, n$.

Under Model (2), with an extra assumption that all entries of $\boldsymbol{E}$ are i.i.d. with variance σ_a^2, Shabalin and Nobel (2013) has shown that the angle between the leading left singular vector of $\boldsymbol{A}$ and that of $\boldsymbol{A}_0$, i.e., the estimated hub centrality and the true hub centrality $\boldsymbol{u}$, can converge to a nonzero quantity or even asymptotically orthogonal as n goes to infinity, if the signal-to-noise ratio d/σ_a is not large enough. This implies that the naive estimation of the centralities by implementing SVD on the observed network will fail in the presence of large noise.

Fortunately, when the naive estimation is inconsistent, it is still possible to obtain a consistent estimation of the centralities by including additional information from other sources. As we discussed in Sections 1 and 2, the positions of agents (nodes) impact the agents' behaviors and thus shape their outcomes. Since the centralities measure the positions of nodes, researchers often study the relationship between the centralities and a certain response variable of interest so as to investigate the network effect. And this relationship contains an additional source of information for the centralities.

To explore such a relationship, the common practice is to regress the response variable on the *estimated* hub and authority centralities, which are obtained through the SVD of the observed network $\boldsymbol{A}$, not the true network $\boldsymbol{A}_0$. But the generative model is actually prescribed as follows: for the i-th observation, the outcome y_i depends on the *true* hub and authority centralities u_i, v_i along with the covariates $\boldsymbol{x}_i \in \mathbb{R}^p$:

$$y_i = \beta_x^\top \boldsymbol{x}_i + \beta_u u_i + \beta_v v_i + \epsilon_i, \tag{3}$$

where $\boldsymbol{\beta}_x \in \mathbb{R}^p$ is the vector of regression coefficients and $\beta_u, \beta_v \in \mathbb{R}$ are the coefficients of the hub and authority centralities respectively. The nuance of using the *estimated* versus the *true* centralities has consequences which we will explicate later. At this stage, no extra assumptions are imposed on the distribution of the regression error, except that we assume $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathsf{Var}(\boldsymbol{\epsilon}) = \sigma_y^2 \boldsymbol{I}_n$. We further assume $n > p + 2$ and $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.

Putting (2) and the matrix version of (3) together, we propose the following unified framework that encapsulates the two models:

$$\begin{cases} \boldsymbol{A} = d\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E}, & \text{(4a)} \\ \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{u}\beta_u + \boldsymbol{v}\beta_v + \boldsymbol{\epsilon}. & \text{(4b)} \end{cases}$$

With the unified framework (4) and the observed data $\{\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}\}$, our original two goals of centrality estimation and network regression analysis can be specified as the following three: (i) estimate the true centralities $\boldsymbol{u}, \boldsymbol{v}$ and the true network $\boldsymbol{A}_0 = d\boldsymbol{u}\boldsymbol{v}^\top$; (ii) estimate the regression coefficients of the predictors $\boldsymbol{\beta}_x$ and the centralities β_u, β_v; (iii) construct *valid* confidence intervals for the centralities as well as the regression coefficients that account for the randomness in the observed network.

The unified framework unites our estimation goals and provides a theoretical framework to study the behaviors of the two-stage procedure. Furthermore, unifying the two models motivates our supervised network centrality estimation (SuperCENT) methodology, which we will describe formally in the next section. We name it the "supervised centrality estimation" because $(\boldsymbol{X}, \boldsymbol{y})$ in the regression (4b) can be thought of as the supervisors that offer additional supervision to the centrality estimation. It is expected that if the centralities indeed have strong predictive power (that is, the centrality regression coefficients β_u, β_v are large compared with the regression noise level σ_y), the estimation of the centralities will be better when combining (4a) and (4b) together instead of only considering (4a). With the improved estimation of the centralities via the supervising effect, SuperCENT can further improve the estimation and inference in the regression model. A similar idea of supervision has also been implemented in matrix decomposition (Li et al., 2016), albeit the absence of response prediction.

Remark 1. The unified framework (4) can be extended to a general case with rank r, that is,

$$\begin{cases} \boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top + \boldsymbol{E}, & \text{(5a)} \\ \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{U}\boldsymbol{\beta}_u + \boldsymbol{V}\boldsymbol{\beta}_v + \boldsymbol{\epsilon}, & \text{(5b)} \end{cases}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices of size $n \times r$ and $\boldsymbol{D}$ is a $r \times r$ diagonal matrix with the singular values as the diagonal entries. Such a natural extension may contribute to the existing literature on regression with network information, because most existing papers consider the centralities as the predictors, which are the leading singular vectors. But potentially, the few leading singular vectors can offer additional predictive power as well.

4. METHODOLOGY

Given the unified framework (4), we first elaborate on the widely used two-stage procedure which first estimates the centralities and then estimates and provides inference for the centrality effects in the regression in Section 4.1. We then describe the SuperCENT methodology that simultaneously solves the centrality estimation and network regression in Section 4.2. Section 4.3 is devoted to the prediction problem where new nodes are included together with covariates and a corresponding new network. Section 4.4 describes strategies for tuning parameter selection in SuperCENT.

4.1. *A naive two-stage procedure*

As mentioned in Sections 1 and 2, given the unified framework (4) and the observed data $\{\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}\}$, a natural and naive procedure is the two-stage estimator, which can serve as a benchmark. To make notations consistent, we now formally introduce the estimator.

In view of (4a), the first stage is to perform SVD on the observed adjacency matrix $\boldsymbol{A}$ and take its leading left and right singular vectors and rescale them to have length $\sqrt{n}$, denoted as $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$, as the estimates for the centralities $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. The superscript *ts* stands for the **t**wo-**s**tage procedure. In view of (4b), given the estimates $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$, the second stage performs the ordinary least square (OLS) regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$, treating

$\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ as given covariates. To be specific, the two-stage procedure solves the following two optimization problems sequentially,

$$\begin{cases} (\hat{d}^{ts}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}) := \underset{d, \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \sqrt{n}}{\arg\min} \|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2, & \text{(6a)} \\ \hat{\boldsymbol{\beta}}^{ts} = ((\hat{\boldsymbol{\beta}}_x^{ts})^\top, \hat{\beta}_u^{ts}, \hat{\beta}_v^{ts})^\top := \underset{\boldsymbol{\beta}_x, \beta_u, \beta_v}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \hat{\boldsymbol{u}}^{ts}\beta_u - \hat{\boldsymbol{v}}^{ts}\beta_v\|_2^2. & \text{(6b)} \end{cases}$$

Algorithm 1 outlines the two-stage procedure.

Algorithm 1: The two-stage procedure

Result: $\hat{d}^{ts}$, $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$ and $\hat{\boldsymbol{\beta}}^{ts}$.
Input: the observed network $\mathbf{A} \in \mathbb{R}^{n \times n}$, the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the response vector $\mathbf{y} \in \mathbb{R}^n$.

1. $(\hat{d}^{ts}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}) = \arg\min_{d, \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \sqrt{n}} \|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2$;
2. $\widehat{\mathbf{W}} = (\mathbf{X}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts})$;
3. $\hat{\boldsymbol{\beta}}^{ts} = (\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}})^{-1} \widehat{\mathbf{W}}^\top \mathbf{y}$.

Remark 2. Besides the estimation of the unknown parameters, valid inference is necessary to evaluate the network effect. In numerous empirical studies, researchers usually construct confidence intervals of the regression coefficients from the second stage regression by assuming that $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are fixed and noiseless. This assumption simplifies the inferential statement because it follows that $\mathrm{cov}(\hat{\boldsymbol{\beta}}^{ts}) = \sigma_y^2 (\widehat{\mathbf{W}}^\top \widehat{\mathbf{W}})^{-1}$ where $\widehat{\mathbf{W}} = (\mathbf{X}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts})$. However, the observed network $\boldsymbol{A}$ is one realization from $\boldsymbol{A}_0 + \boldsymbol{E}$ as in Model (4a), which makes its singular vectors $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ random. If one proceeds with inference ignoring the randomness, then the inference loses its justifications and the ensuing validity due to violation of the assumption. We refer to such "ad-hoc" confidence interval as the "two-stage-adhoc" method. To correct for the randomness of the estimated singular vectors $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ and make valid inferences, the asymptotic distribution of the estimator is derived rigorously in Section 5. Remarks 5 and 7 further discuss the theoretical property of the two-stage-adhoc method. The toy experiment in Figure 1 and simulation results in Section 6 show that the two-stage-adhoc method is either conservative with low network noise level or invalid with high network noise level.

4.2. *SuperCENT methodology*

From the two-stage procedure, we observe that the second step of estimation and inference in the regression model depends on the first step of centrality estimation. The more accurate the centrality estimates are, the better we are able to make inference in the regression model. On the other hand, the centralities are incorporated in the regression model as regressors, $(\boldsymbol{X}, \boldsymbol{y})$ have supervising effect on centrality estimation and can boost the estimation accuracy.

Motivated by the above intuition, we propose to optimize the following objective function to obtain the SuperCENT estimates,

$$(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{\beta}}_x, \hat{\beta}_u, \hat{\beta}_v) := \underset{\substack{\boldsymbol{\beta}_x, \beta_u, \beta_v \\ d, \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = \sqrt{n}}}{\arg\min} \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u - \boldsymbol{v}\beta_v\|_2^2 + \frac{\lambda}{n^2} \|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2. \quad (7)$$

The above objective function combines the residual sum of squares (6b) and the rank-one approximation error of the observed network $\boldsymbol{A}$ (6a). The connection between the two terms is the centralities. The tradeoff between the two terms can be tuned through a proper selection of the hyper-parameter λ. This idea is somewhat similar to the supervised SVD method in Li et al. (2016), but both the supervising mechanism and the optimization objective function of the SuperCENT are different.

To solve (7), we can use a block gradient descent algorithm by updating $(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}})$ iteratively until convergence, where $\widehat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_x^\top, \hat{\beta}_u, \hat{\beta}_v)^\top$. Such an iterative algorithm requires an initialization, which can be the first stage of the two-stage procedure, i.e., $(\hat{d}^{ts}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts})$ from the SVD of $\boldsymbol{A}$. The complete algorithm with a given tuning parameter λ is shown in Algorithm 2. We use $(\hat{d}^{(t)}, \hat{\boldsymbol{u}}^{(t)}, \hat{\boldsymbol{v}}^{(t)}, \widehat{\boldsymbol{\beta}}^{(t)})$ to denote the estimations in the t-th iteration. The derivation of Algorithm 2 and the algorithm for a symmetric network with the eigenvector centrality, a special case of Model (4), are deferred to the supplement. We will discuss the methods to choose λ in Section 4.4, including cross-validation and others.

Algorithm 2: SuperCENT$(\mathbf{A}, \mathbf{X}, \mathbf{y}, \lambda)$, an algorithm to solve (7).

Result: $\hat{d}$, $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{v}}$, and $\widehat{\boldsymbol{\beta}}$.
Input: the observed network $\mathbf{A} \in \mathbb{R}^{n\times n}$, the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, the response vector $\mathbf{y} \in \mathbb{R}^n$, the tuning penalty parameter λ, the tolerance parameter $\rho > 0$, the maximum number of iteration T ;
Initiate $(d^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) = \arg\min_{d, \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \sqrt{n}} \|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2$, $t = 1$;
while $\max\left(\|\boldsymbol{P}_{\boldsymbol{u}^{(t)}} - \boldsymbol{P}_{\boldsymbol{u}^{(t-1)}}\|_2, \|\boldsymbol{P}_{\boldsymbol{v}^{(t)}} - \boldsymbol{P}_{\boldsymbol{v}^{(t-1)}}\|_2\right) > \rho$ *and* $t < T$ **do**

1. $\mathbf{W}^{(t-1)} = (\mathbf{X}, \boldsymbol{u}^{(t-1)}, \boldsymbol{v}^{(t-1)})$;
2. $\boldsymbol{\beta}^{(t)} = (\mathbf{W}^{(t-1)^\top}\mathbf{W}^{(t-1)})^{-1}\mathbf{W}^{(t-1)^\top}\boldsymbol{y}$;
3. $d^{(t)} = \frac{1}{n^2}\boldsymbol{u}^{(t-1)^\top}\mathbf{A}\boldsymbol{v}^{(t-1)}$;
4. $\boldsymbol{u}^{(t)} = \left((\beta_u^{(t)})^2 + \lambda(d^{(t)})^2\right)^{-1}\left[\beta_u^{(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x^{(t)} - \boldsymbol{v}^{(t-1)}\beta_v^{(t)}) + \frac{1}{n}\lambda d^{(t)}\mathbf{A}\boldsymbol{v}^{(t-1)}\right]$;
5. Normalize $\boldsymbol{u}^{(t)}$ such that $\|\boldsymbol{u}^{(t)}\|_2 = \sqrt{n}$;
6. $\boldsymbol{v}^{(t)} = \left((\beta_v^{(t)})^2 + \lambda(d^{(t)})^2\right)^{-1}\left[\beta_v^{(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x^{(t)} - \boldsymbol{u}^{(t)}\beta_u^{(t)}) + \frac{1}{n}\lambda d^{(t)}\mathbf{A}^\top\boldsymbol{u}^{(t)}\right]$;
7. Normalize $\boldsymbol{v}^{(t)}$ such that $\|\boldsymbol{v}^{(t)}\|_2 = \sqrt{n}$;
8. $t \leftarrow t + 1$;

end

Note that although $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ with length $\sqrt{n}$ are only identifiable up to the sign, $\hat{\boldsymbol{u}}\hat{\beta}_u$ and $\hat{\boldsymbol{v}}\hat{\beta}_v$ are uniquely identifiable. One procedure to determine the sign of all the parameters is as follows: first, find the entry that has the largest magnitude in $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$, and make that entry positive so that the sign of either the hub centrality $\hat{\boldsymbol{u}}$ or the authority centrality $\hat{\boldsymbol{v}}$ can be fixed; then adjust the sign of the other centrality since the product $\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\top$ is identifiable, and finally determine the signs of $\hat{\beta}_u, \hat{\beta}_v$ accordingly.

4.3. *Prediction*

Once the model is fitted with training data, it can be used for prediction. Suppose there are n^* new observations, they have not only the new covariates $\boldsymbol{X}^*$ and the new network among themselves $\boldsymbol{A}^*$, but also new edges connecting them with the n training observations. The

original network $\boldsymbol{A}$ is augmented to $\boldsymbol{A}^{all}$ as follows

$$\boldsymbol{A}^{all} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}^* \end{pmatrix}, \tag{8}$$

where $\boldsymbol{A}^{all}$ is of size $(n + n^*) \times (n + n^*)$.

Given the augmented network $\boldsymbol{A}^{all}$ and new covariates $\boldsymbol{X}^*$, the task is to make prediction for $\boldsymbol{y}^*$. To make use of the regression equation $\hat{\boldsymbol{y}}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_x + \boldsymbol{u}^*\hat{\beta}_u + \boldsymbol{v}^*\hat{\beta}_v$, it is necessary to estimate $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$. Similarly as (4a), we assume $\boldsymbol{A}^{all} = d\boldsymbol{u}^{all}\boldsymbol{v}^{all\top} + \boldsymbol{E}^{all}$ where $\boldsymbol{u}^{all} = (\boldsymbol{u}^\top, \boldsymbol{u}^{*\top})^\top$ and $\boldsymbol{v}^{all} = (\boldsymbol{v}^\top, \boldsymbol{v}^{*\top})^\top$. Therefore, we obtain the model

$$\boldsymbol{A}^{all} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}^* \end{pmatrix} = d\begin{pmatrix} \boldsymbol{u}\boldsymbol{v}^\top & \boldsymbol{u}\boldsymbol{v}^{*\top} \\ \boldsymbol{u}^*\boldsymbol{v}^\top & \boldsymbol{u}^*\boldsymbol{v}^{*\top} \end{pmatrix} + \begin{pmatrix} \boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \end{pmatrix}. \tag{9}$$

In view of (9), to obtain estimates of $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$, one can either perform SVD of $\boldsymbol{A}^*$ or SVD of $\boldsymbol{A}^{all}$ and reserve only the relevant components of the singular vectors. The latter approach is more accurate and is formally described in Algorithm 3.

Algorithm 3: SVD of $\boldsymbol{A}^{all}$ to obtain $\hat{\boldsymbol{u}}^*$ and $\hat{\boldsymbol{v}}^*$

Result: $\hat{\boldsymbol{u}}^*$ and $\hat{\boldsymbol{v}}^*$.
Input: The augmented network $\boldsymbol{A}^{all}$.

1. $(\hat{\boldsymbol{u}}^{all}, \hat{\boldsymbol{v}}^{all})$ are the left and right singular vectors of $\boldsymbol{A}^{all}$;
2. $\hat{\boldsymbol{u}}^{all} = \text{sign}(\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}^{all}_{1:n})\hat{\boldsymbol{u}}^{all}$ and $\hat{\boldsymbol{v}}^{all} = \text{sign}(\hat{\boldsymbol{v}}^\top \hat{\boldsymbol{v}}^{all}_{1:n})\hat{\boldsymbol{v}}^{all}$;
3. $\hat{\boldsymbol{u}}^* = \hat{\boldsymbol{u}}^{all}_{(n+1):(n+n^*)}$ and $\hat{\boldsymbol{v}}^* = \hat{\boldsymbol{v}}^{all}_{(n+1):(n+n^*)}$;
4. Rescale $\hat{\boldsymbol{u}}^* = \frac{\sqrt{n}}{\|\hat{\boldsymbol{u}}^{all}_{1:n}\|_2}\hat{\boldsymbol{u}}^*$ and $\hat{\boldsymbol{v}}^* = \frac{\sqrt{n}}{\|\hat{\boldsymbol{v}}^{all}_{1:n}\|_2}\hat{\boldsymbol{v}}^*$.

Remark 3. (Sign and scaling issue) Since $\hat{\boldsymbol{u}}^{all}$ and $\hat{\boldsymbol{v}}^{all}$ are only identifiable up to sign, we determine their signs as Step 2 of Algorithm 3 such that the angles between the training proportions and the SuperCENT estimates are less than 90 degrees, i.e., $\text{sign}(\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}^{all}_{1:n}) > 0$ and $\text{sign}(\hat{\boldsymbol{v}}^\top \hat{\boldsymbol{v}}^{all}_{1:n}) > 0$. In addition, $\hat{\boldsymbol{u}}^*$ and $\hat{\boldsymbol{v}}^*$ need to be scaled to match with $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$. Recall that for identifiability, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are of norm $\sqrt{n}$, and β_u and β_v are of the corresponding scale. In the prediction process, we need to scale $\hat{\boldsymbol{u}}^*$ and $\hat{\boldsymbol{v}}^*$ accordingly so that $\beta_u\hat{\boldsymbol{u}}^* + \beta_v\hat{\boldsymbol{v}}^*$ is on par with $\beta_u\hat{\boldsymbol{u}} + \beta_v\hat{\boldsymbol{v}}$. Step 3 of Algorithm 3 is designed for this purpose.

4.4. *Selection of the tuning parameter λ*

The tuning parameter λ can be selected using the K-fold cross-validation. Given the prediction procedure in Section 4.3, the cross-validation procedure can be easily carried out as follows. For each fold of validation data, we first fit the model using the remaining $K-1$ folds with the corresponding induced subnetwork and obtain the estimates for the regression coefficients by implementing Algorithm 2; we then obtain the estimates of the centralities for the validation data by applying Algorithm 3; we last obtain the total prediction error for the validation data by combining the outcomes from the first two steps. The best tuning parameter λ is set to be the minimizer of the total cross-validation error that sums over all folds. Algorithm 4 outlines this procedure in more detail.

Another strategy for selecting the tuning parameter is through generalized cross-validation (GCV), which can save computational time. Denote $\hat{\boldsymbol{y}} = \boldsymbol{H}(\lambda)\boldsymbol{y}$, then the GCV criterion is

$$GCV(\lambda) = \frac{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2/n}{(1 - \mathsf{trace}(\boldsymbol{H}(\lambda))/n)^2}. \tag{10}$$

The explicit form of $\boldsymbol{H}(\lambda)$ can be derived from the proof of the theorems. The performance of GCV is left for future investigation.

As a third strategy, Remark 12 in the next section offers an alternative way to choose the tuning parameter based on the theoretical analysis of SuperCENT, which is less time-consuming than cross-validation. However, we recommend using the cross-validation strategy for the best performance based on the simulation results.

Algorithm 4: The cross-validation algorithm for SuperCENT to choose λ.

Result: λ_{min} .
Input: $\mathbf{A}, \mathbf{X}, \mathbf{y}$.
for λ *on a exponentially regular grid* **do**
  for *each fold* **do**
    0. Split the covariates and response into training $\mathbf{X}_{fold,train}$, $\mathbf{y}_{fold,train}$ and validation $\mathbf{X}_{fold,val}$, $\mathbf{y}_{fold,val}$ and denote the the induced sub-network corresponding to the training data $\mathbf{A}_{fold,train}$;
    1. $(\widehat{\boldsymbol{\beta}}_{fold,\lambda}, \hat{\boldsymbol{u}}_{fold,\lambda}, \hat{\boldsymbol{v}}_{fold,\lambda}) \leftarrow$ SuperCENT$(\mathbf{A}_{fold,train}, \mathbf{X}_{fold,train}, \mathbf{y}_{fold,train}, \lambda)$ by Algorithm 2;
    2. $\hat{\boldsymbol{u}}_{fold,val}, \hat{\boldsymbol{v}}_{fold,val} \leftarrow$ SVD$(\boldsymbol{A})$ and re-scale by Algorithm 3;
    3. $SSE_{fold,\lambda} = \|\mathbf{y}_{fold,val} - (\mathbf{X}_{fold,val}, \hat{\boldsymbol{u}}_{fold,val}, \hat{\boldsymbol{v}}_{fold,val})\, \widehat{\boldsymbol{\beta}}_{fold,\lambda}\|_2^2$;
  end
end
$\lambda_{min} = \min_\lambda \sum_{fold} SSE_{fold,\lambda}$

5. THEORETICAL PROPERTIES

We investigate the statistical properties of the two-stage procedure in Section 5.1 and SuperCENT in Section 5.2. The two main theorems provide the asymptotic distributions of the estimators under appropriate conditions, which can be used for inference. The corollaries state the convergence rates and the bias of the relevant quantities.

We first introduce some notations and assumptions. Recall that $\boldsymbol{P}_{\cdot}$ denotes the projection matrix, such as $\boldsymbol{P}_{\boldsymbol{u}}$, $\boldsymbol{P}_{\boldsymbol{v}}$ and $\boldsymbol{P}_{\boldsymbol{X}}$. Define $\tilde{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}$, $\tilde{\boldsymbol{v}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}$, which are the centralities projected onto the orthogonal space of $\boldsymbol{X}$. Denote $c = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} - (\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}})^2$ and $\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} = \begin{pmatrix} \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} & \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} \\ \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} & \tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} (\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}})$.

Assumption 1. Under the unified framework (4), the noise of the network e_{ij} independently follows $N(0, \sigma_a^2)$, and the noise of the outcome regression ϵ_i independently follows $N(0, \sigma_y^2)$.

Assumption 2. The fixed design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with $n > p + 2$ and $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. The dimension p is not diverging.

Assumption 3. The scaled network noise-to-signal ratio $\kappa = \frac{\sigma_a^2}{d^2 n} \to 0$.

In Assumption 1, the independence is assumed for simplicity. If the network noise e_{ij} or the regression noise ϵ_i are dependent with known covariance, the theorems and corollaries still hold with slight modifications; if they are dependent with unknown covariance, extra assumptions on the covariance structure need to be made and new methodologies and theories should be developed. Assumption 3 is required for the consistency of the SVD of the observed noisy network $\boldsymbol{A}$ of model (4a), which can be seen from Corollary 1 below.

5.1. *Theoretical properties for the two-stage procedure*

Under the three aforementioned assumptions, the two-stage procedure in Algorithm 1 is consistent, and the asymptotic distribution of its estimators is given in Theorem 1. The convergence rates of the estimators are given in Corollaries 1 and 2. When Assumption 3 is violated, the two-stage procedure is no longer consistent – the centralities estimation is inconsistent in Stage 1, which leads to bias in the regression coefficients estimation in Stage 2 resembling the measurement error problem as shown in Corollary 3, consequently to the detriment of the inference.

Recall that the two-stage estimates from Algorithm 1 are denoted as $\hat{d}^{ts}$, $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$ and $\widehat{\boldsymbol{\beta}}^{ts} = (\hat{\beta}_u^{ts}, \hat{\beta}_v^{ts}, (\hat{\boldsymbol{\beta}}_x^{ts})^\top)^\top$. Let $\widehat{\boldsymbol{A}}^{ts} = \hat{d}^{ts}\hat{\boldsymbol{u}}^{ts}\hat{\boldsymbol{v}}^{ts\top}$ be the estimate of $\boldsymbol{A}_0$.

THEOREM 1: *Under the unified framework (4) and Assumptions 1, 2 and 3, the two-stage estimates converge to the following normal distribution asymptotically,*

$$\begin{pmatrix} \hat{\boldsymbol{u}}^{ts} - \boldsymbol{u} \\ \hat{\boldsymbol{v}}^{ts} - \boldsymbol{v} \\ vec\left(\widehat{\boldsymbol{A}}^{ts} - \boldsymbol{A}_0\right) \\ \hat{\beta}_u^{ts} - \beta_u \\ \hat{\beta}_v^{ts} - \beta_v \\ \hat{\boldsymbol{\beta}}_x^{ts} - \boldsymbol{\beta}_x \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\boldsymbol{0}_{(2n+n^2+2+p)\times 1}, \boldsymbol{\Sigma}^{ts}\right), \tag{11}$$

where $\boldsymbol{\Sigma}^{ts} = \boldsymbol{C}^{ts}\begin{pmatrix} \sigma_y^2 \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2 \boldsymbol{I}_{n^2} \end{pmatrix}\boldsymbol{C}^{ts\top}$ *and* $\boldsymbol{C}^{ts}$ *is a function of* $(d, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{X}, \beta_u, \beta_v)$*, whose specific form is given in the supplement.*

Recall the two-stage procedure first estimates the centralities $\boldsymbol{u}$ and $\boldsymbol{v}$ and then plugs the estimated centralities into the regression model. Therefore, the asymptotic distributions for $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$ and $\widehat{\boldsymbol{A}}^{ts}$ only depend on the noise $\boldsymbol{E}$ from the network model, not the regression noise $\boldsymbol{\epsilon}$. This can also be seen in the definition of $\boldsymbol{C}^{ts}$, where the three top left blocks are zeros.

Remark 4. (Covariance of $\widehat{\boldsymbol{\beta}}^{ts}$) One important fact to emphasize is that the covariance of $\widehat{\boldsymbol{\beta}}^{ts}$ is *not* $\sigma_y^2(\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}})^{-1}$ where $\widehat{\mathbf{W}} = (\boldsymbol{X}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts})$, which is the classical results of regression when $\boldsymbol{X}, \hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ are considered fixed. This makes sense, as in our model, the observed network contains noise, which makes the estimated centralities $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ from the first stage random quantities and invalidates the traditional covariance result. As a matter of fact, the bottom right three blocks of $\boldsymbol{C}^{ts}$ are *not* zero, which highlights this phenomenon. Corollary 2 further illustrates this fact and its consequences on the inference.

In what follows, we present the convergence rates of the estimated centralities $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$, and the network $\widehat{\boldsymbol{A}}^{ts}$ in Corollary 1, and the convergence rates of the regression coefficients $\widehat{\boldsymbol{\beta}}^{ts}$ and the prediction error in Corollary 2. In Corollary 3, we show the bias of $\hat{\beta}_u$ and $\hat{\beta}_v$ when the two-stage is inconsistent.

The convergence rates of the centralities depend on the selection of the loss function. Ideally, since the scales of the centralities are not fully determined, one prefers the loss function $\|\boldsymbol{P}_{\hat{\boldsymbol{u}}} - \boldsymbol{P}_{\boldsymbol{u}}\|_2^2$, which equals the squared sine of the angle between $\hat{\boldsymbol{u}}$ and $\boldsymbol{u}$, $\sin^2\angle(\hat{\boldsymbol{u}}, \boldsymbol{u})$. However, the exact form of this loss function is not clean mathematically. Instead, we use the scaled Euclidean distance $\|\hat{\boldsymbol{u}} - \text{sign}(\hat{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}\|_2^2/n = 2 - 2\cos^2\angle(\hat{\boldsymbol{u}}, \boldsymbol{u})$, which has a cleaner expression and is connected to the squared sine through $\|\boldsymbol{P}_{\hat{\boldsymbol{u}}} - \boldsymbol{P}_{\boldsymbol{u}}\|_2^2 = (\|\hat{\boldsymbol{u}} - \text{sign}(\hat{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}\|_2^2/n)[1 - (\|\hat{\boldsymbol{u}} - \text{sign}(\hat{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}\|_2^2/n)/4]$. These two losses are approximately equivalent when the estimator is consistent and the loss goes to zero.

COROLLARY 1: *(Rate of $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$, and $\widehat{\boldsymbol{A}}^{ts}$) Under the unified framework (4) and Assumptions 1, 2 and 3, the two-stage estimators satisfy the following*

$$\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{u}}^{ts} - \boldsymbol{u}\|_2^2 = \frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{v}}^{ts} - \boldsymbol{v}\|_2^2 = \frac{\sigma_a^2(n-1)}{d^2n^2}(1 + o(1)) = O\left(\frac{\sigma_a^2}{d^2n}\right) = O(\kappa), \quad (12)$$

$$\mathbb{E}\frac{\left\|\widehat{\boldsymbol{A}}^{ts} - \boldsymbol{A}_0\right\|_F^2}{\|\boldsymbol{A}_0\|_F^2} = \frac{\sigma_a^2(2n-1)}{d^2n^2}(1 + o(1)) = O\left(\frac{\sigma_a^2}{d^2n}\right) = O(\kappa). \quad (13)$$

According to Corollary 1, the rate for $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ in (12) is $\frac{\sigma_a^2(n-1)}{d^2n^2}$ and the rate for $\widehat{\boldsymbol{A}}^{ts}$ in (13) is $\frac{\sigma_a^2(2n-1)}{d^2n^2}$. They have the same order as $\kappa = \frac{\sigma_a^2}{d^2n}$, which suggests that the noise-to-signal ratio κ is the critical quantity that determines the consistency of the two-stage procedure.

Note that, for the two-stage procedure to be consistent, one needs $\kappa \to 0$ as $n \to \infty$, which depends on the three parameters d, σ_a, n and their relationships. We first discuss two consistent scenarios: 1) the signal strength d and the noise level σ_a are of constant order while n diverges; 2) the noise level σ_a stays constant, but the signal strength d can decrease when more nodes are collected for the network (because the network edge density might decay with more nodes), in which case as long as $d^2n \to \infty$, κ still goes to zero. For real data, it is possible that the observed network gets noisier with more nodes $\sigma_a \to \infty$ and the signal strength decays $d \to 0$. In this case, it is highly likely that the two-stage procedure will be inconsistent, for example, with a fast diverging noise level; then SuperCENT can improve the performance and remain consistent as discussed below in Section 5.2.

COROLLARY 2: *(Rate of $\widehat{\boldsymbol{\beta}}^{ts}$) Under the unified framework (4) and Assumptions 1, 2 and 3, the two-stage estimators satisfy*

$$\mathbb{E}(\hat{\beta}_u^{ts} - \beta_u)^2 = \left(\frac{\sigma_y^2}{c}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}}\right. \quad (14)$$

$$\left. + \frac{\sigma_a^2}{c^2}\frac{1}{d^2n}\left[\beta_v^2\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}}\tilde{\boldsymbol{u}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} + \beta_u^2\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}\right]\right)(1 + o(1)) \quad (15)$$

$$= O\left(\frac{\sigma_y^2}{n} + \frac{\sigma_a^2(\beta_u^2 + \beta_v^2)}{d^2n^2}\right), \quad (16)$$

$$\mathbb{E}(\hat{\beta}_v^{ts} - \beta_v)^2 = \left(\frac{\sigma_y^2}{c} \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}} \right. \tag{17}$$

$$\left. + \frac{\sigma_a^2}{c^2} \frac{1}{d^2 n} \left[\beta_u^2 \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}} \tilde{\boldsymbol{v}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \tilde{\boldsymbol{v}} \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}} + \beta_v^2 \tilde{\boldsymbol{v}}^\top \tilde{\boldsymbol{u}} \tilde{\boldsymbol{u}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \tilde{\boldsymbol{u}} \tilde{\boldsymbol{v}}^\top \tilde{\boldsymbol{u}} \right] \right) (1 + o(1)) \tag{18}$$

$$= O\left(\frac{\sigma_y^2}{n} + \frac{\sigma_a^2 (\beta_u^2 + \beta_v^2)}{d^2 n^2} \right), \tag{19}$$

$$Cov\left(\hat{\boldsymbol{\beta}}_x^{ts} - \boldsymbol{\beta}_x\right) = \sigma_y^2 \left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{u}\ \boldsymbol{v}) \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \boldsymbol{u}^\top \\ \boldsymbol{v}^\top \end{pmatrix} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \right] \tag{20}$$

$$+ \sigma_a^2 \frac{1}{d^2 n} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \left[\beta_u^2 (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2 (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \right. \tag{21}$$

$$\left. + (\boldsymbol{u}\ \boldsymbol{v}) \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \beta_v^2 \tilde{\boldsymbol{u}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \tilde{\boldsymbol{u}} & 0 \\ 0 & \beta_u^2 \tilde{\boldsymbol{v}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \tilde{\boldsymbol{v}} \end{pmatrix} \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \boldsymbol{u}^\top \\ \boldsymbol{v}^\top \end{pmatrix} \right] \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (1 + o(1)) \tag{22}$$

Remark 5. (Comments on (14)-(16) for β_u) For the variance of $\hat{\beta}_u^{ts}$, the first term (14) is the typical expression for traditional regression with deterministic predictors. The additional term (15) is caused by the randomness of $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$. Note that the second term (15) is non-negative and is zero if $\sigma_a = 0$, or $\tilde{\boldsymbol{u}} \perp \tilde{\boldsymbol{v}}$. Furthermore, the first term in (16) is of order σ_y^2/n while the second term is of order $\frac{\sigma_a^2(\beta_u^2+\beta_v^2)}{d^2 n^2} = \kappa \frac{\beta_u^2+\beta_v^2}{n}$. So if β_u and β_v are of constant order and $\kappa \to 0$, the second term is of smaller order than the first term.

The above Theorem and Corollaries assume Assumptions 1-3. When Assumption 3 is violated, i.e., $\kappa \nrightarrow 0$, the two-stage cannot estimate $\boldsymbol{u}$ and $\boldsymbol{v}$ consistently, and consequently $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$ are biased.

COROLLARY 3: *(Bias of $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$ when the two-stage is inconsistent) Let $\rho = \mathrm{cor}(\boldsymbol{u}, \boldsymbol{v})$. If $\boldsymbol{\beta}_x = \boldsymbol{0}$ or $\mathrm{cov}(\boldsymbol{X}, (\boldsymbol{uv})) = \boldsymbol{0} \in \mathbb{R}^{p \times 2}$, then under the unified framework (4) and Assumptions 1-2,*

$$\mathrm{plim}\ \hat{\beta}_u^{ts} = \frac{(1 + \kappa - \rho^2)\beta_u + \kappa\rho\beta_v}{(1+\kappa)^2 - \rho^2}, \tag{23}$$

$$\mathrm{plim}\ \hat{\beta}_v^{ts} = \frac{(1 + \kappa - \rho^2)\beta_v + \kappa\rho\beta_u}{(1+\kappa)^2 - \rho^2}. \tag{24}$$

Remark 6. (Conditions for the bias of $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$) Corollary 3 assumes either $\boldsymbol{\beta}_x = \boldsymbol{0}$, namely the true regression model only involves two centrality predictors, both of which have measurement errors, or $\mathrm{cov}(\boldsymbol{X}, (\boldsymbol{uv})) = \boldsymbol{0} \in \mathbb{R}^{p \times 2}$, namely the true regression model involves p predictors without measurement error and two centrality predictors with measurement errors, but the noiseless predictors and two centrality predictors are uncorrelated. These assumptions are adopted so that the expression of the bias has intuitive explanations to be followed momentarily. In general, when the p noiseless ones and the two centralities are correlated, the bias persists as long as the signal-to-noise of the network is small when $\kappa \nrightarrow 0$, although the expressions are less comprehensive.

Remark 7. (Special cases for the bias of $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$) There are a few special cases for Corollary 3. (i) When $\kappa \to 0$, $\mathrm{plim}\ \hat{\beta}_u^{ts} = \beta_u$ and $\mathrm{plim}\ \hat{\beta}_v^{ts} = \beta_v$. That implies that when $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are consistent estimators of $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively, the existence of the estimation error of the centralities does not affect the bias-ness of the second stage regression. This

has correspondence to the errors-in-variables literature: when the relative amount of measurement error is small compared to the total variance of the observed variable, the OLS estimate is unbiased. In other words, $\frac{1}{1+\kappa}$ can be viewed as the reliability, attenuation factor, or the signal-to-total variance ratio. (ii) When $\kappa \nrightarrow 0$, but the two true centralities are uncorrelated, $\rho = 0$, we have $\text{plim}\ \hat{\beta}_u^{ts} = \frac{1}{1+\kappa}\beta_u$ and $\text{plim}\ \hat{\beta}_v = \frac{1}{1+\kappa}\beta_v$. The OLS estimate is biased towards zero, and the degree of bias depends on $\frac{1}{1+\kappa}$. (iii) When $\kappa \nrightarrow 0$ and $\rho \neq 0$, if $|\beta_u| \gg |\beta_v|$, then $\text{plim}\ \hat{\beta}_u^{ts} \approx \frac{(1+\kappa-\rho^2)\beta_u}{(1+\kappa)^2-\rho^2}$, which is equivalent to $\text{plim}\ \hat{\beta}_u^{ts} - \beta_u \approx -\frac{(1+\kappa)\kappa}{(1+\kappa)^2-\rho^2}\beta_u$. This implies that $\hat{\beta}_u^{ts}$ has attenuation bias. As for $\hat{\beta}_v^{ts}$, we obtain that $\text{plim}\ \hat{\beta}_v^{ts} - \beta_v \approx \frac{\rho\kappa}{(1+\kappa)^2-\rho^2}\beta_u$, which implies that $\hat{\beta}_v^{ts}$ is biased away from zero. (iv) When β_u and β_v have similar size, the directions of the biases depends on the β_u, β_v, ρ and κ. For $\hat{\beta}_u^{ts}$, the asymptotic bias is $\text{plim}\ \hat{\beta}_u^{ts} - \beta_u = \frac{-\kappa[(1+\kappa)\beta_u - \rho\beta_v]}{(1+\kappa)^2-\rho^2}$. Since the denominator is always larger than 0 because $\kappa > 0$ and $0 < \rho < 1$, the direction of the bias depends on the sign of $(1+\kappa)\beta_u - \rho\beta_v$: when $\beta_u > \frac{\rho}{1+\kappa}\beta_v$, $\text{plim}\ \hat{\beta}_u^{ts} - \beta_u < 0$; when $\beta_u < \frac{\rho}{1+\kappa}\beta_v$, $\text{plim}\ \hat{\beta}_u^{ts} - \beta_u > 0$. Similar conclusions can be drawn for $\hat{\beta}_v^{ts}$.

Remark 8. (Review of inference property in the framework of classical measurement error with one predictor) Consider the simplest classical population model with measurement error $\boldsymbol{y} = \boldsymbol{u}\beta_u + \boldsymbol{\epsilon}$. For simplicity and clarity, we exclude $\boldsymbol{X}$ and $\boldsymbol{v}$ from the model. We only observe $\hat{\boldsymbol{u}} = \boldsymbol{u} + \boldsymbol{\delta}_u$ with measurement error $\boldsymbol{\delta}_u$ instead of the true $\boldsymbol{u}$. It can be shown that the OLS estimate that regresses $\boldsymbol{y}$ on $\hat{\boldsymbol{u}}$ satisfies that $\hat{\beta}_u \to \frac{1}{1+\kappa}\beta_u := \gamma\beta_u$, where $\kappa = \frac{\sigma_a^2}{d^2 n}$ is our scaled noise-to-signal ratio and γ is the attenuation factor.

Furthermore, with the OLS estimate, the residual sum of squares has limit $n(\sigma_y^2 + (1-\gamma)^2\beta_u^2 + \gamma^2\beta_u^2\kappa) = n(\sigma_y^2 + \beta_u^2\kappa/(1+\kappa))$. Hence, the traditional estimate of σ_y^2, $\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2/(n-1)$, over-estimates σ_y^2, the larger the κ and β_u, the larger the over-estimation. It can also be shown that the standard error of OLS $\hat{\beta}_u$ converges to $\gamma\sigma_y^2 + \gamma(1-\gamma)\beta_u^2$. Combined, the t-ratio goes to $\sqrt{\gamma}\frac{\beta_u}{\sqrt{\sigma_y^2+(1-\gamma)\beta_u^2}}$, which is smaller than β_u/σ_y. So the traditional inference ignoring the measurement error is conservative. Please refer to Pischke (2007) and Wooldridge (2015) for more details on measurement error.

Remark 9. (CI for β_u from the ad-hoc two-stage method) Remark 8 has a few implications on the confidence interval (CI) for β_u under the unified framework (4) when the two-stage estimator is consistent ($\kappa \to 0$). On one hand, when all of the quantities in (14)-(15) (including $\sigma_a, \sigma_y, d, n, \beta_u, \beta_v, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}$) are known, one should use both terms to make valid inference. If one uses (14) alone while assuming noiseless $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}$ to construct the CI, i.e. the "two-stage-adhoc" method, the inference is invalid unless $\sigma_a = 0$, or $\tilde{\boldsymbol{u}} \perp \tilde{\boldsymbol{v}}$. But the degree of invalidity is typically small because (15) $\ll$ (14) when $\kappa \to 0$. On the other hand, when the quantities are unknown and need to be estimated and plugged into these two terms, the stories are different. Note that from the classical measurement error literature, $\hat{\sigma}_y^2$ from the two-stage over-estimates σ_y^2 and the inference based on the first estimated term (14) alone is already conservative. As a consequence, the inference based on the two estimated terms would be even more conservative with unnecessarily large width.

In the simulation study, we consider three relevant methods: two-stage-oracle uses both terms with the true parameters, two-stage uses both terms with the estimated parameters, and the two-stage-adhoc uses only the first term with the estimated parameters. The results are consistent with the above discussion.

To further add to Remark 4 on the covariance of $\hat{\boldsymbol{\beta}}_x^{ts}$, due to the randomness of $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$, the covariance of $\hat{\boldsymbol{\beta}}_x^{ts}$ in Corollary 2 involves three terms, where (20) is the term involving σ_y^2 and would be the traditional covariance of $\hat{\boldsymbol{\beta}}_x^{ts}$ if $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ were indeed fixed, the two additional terms (21)-(22) are the extra covariance caused by the randomness of $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$.

5.2. *Theoretical properties for SuperCENT*

This section states the theoretical properties of our proposed SuperCENT method. Intuitively, we expect SuperCENT to be superior to the two-stage procedure when the signal-to-noise ratio of the regression model is large (so that the supervision information from the regression model is strong enough), especially when the observed network is noisy. Theorem 2 shows the asymptotic distribution of the SuperCENT estimators under the same set of conditions as the two-stage and Corollaries 4 and 5 provide their convergence rates. Comparing SuperCENT with the two-stage, we further explicate the discrepancy between SuperCENT and the two-stage and the conditions such that SuperCENT outperforms the two-stage, particularly when the two-stage is inconsistent. Note that the SuperCENT estimates from Algorithm 2 with a given tuning parameter λ are denoted as $\hat{d}$, $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{v}}$, and $\widehat{\boldsymbol{\beta}} = ((\hat{\boldsymbol{\beta}}_x)^\top, \hat{\beta}_u, \hat{\beta}_v)^\top$. Let $\widehat{\boldsymbol{A}} = \hat{d}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\top$ be the estimate of $\boldsymbol{A}_0$.

THEOREM 2: *Under the unified framework* (4) *and Assumptions 1, 2 and 3, the SuperCENT estimators converge to the following normal distribution asymptotically,*

$$\begin{pmatrix} \hat{\boldsymbol{u}} - \boldsymbol{u} \\ \hat{\boldsymbol{v}} - \boldsymbol{v} \\ vec\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\right) \\ \hat{\beta}_u - \beta_u \\ \hat{\beta}_v - \beta_v \\ \hat{\boldsymbol{\beta}}_x - \boldsymbol{\beta}_x \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\mathbf{0}_{(2n+n^2+2+p)\times 1}, \boldsymbol{\Sigma}\right), \tag{25}$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}\begin{pmatrix} \sigma_y^2\boldsymbol{I}_n & \mathbf{0}_{n\times n^2} \\ \mathbf{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2} \end{pmatrix}\boldsymbol{C}^\top$ *and* $\boldsymbol{C}$ *is a function of* $(d, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{X}, \beta_u, \beta_v, \lambda)$*, whose specific form is given in the supplement.*

Corollaries 4 and 5 provide the convergence rates of the SuperCENT estimators. To explicate the difference between the two-stage and SuperCENT, let

$$\delta_{ts,sc} = \frac{1}{(\lambda d^2 + \beta_u^2 + \beta_v^2)^2}\left[\frac{2\lambda d^2 + \beta_u^2 + \beta_v^2}{d^2 n}\sigma_a^2 - \sigma_y^2\right]. \tag{26}$$

COROLLARY 4: *(Rate of* $\hat{\boldsymbol{u}}$*,* $\hat{\boldsymbol{v}}$ *and* $\widehat{\boldsymbol{A}}$*). Under the unified framework* (4) *and Assumptions 1, 2 and 3, the SuperCENT estimators satisfy the following,*

$$\mathbb{E}\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2/n = \left(\frac{\sigma_a^2(n-1)}{d^2n^2} - \frac{n-p-2}{n}\beta_u^2\delta_{ts,sc}\right)(1 + o(1)) \tag{27}$$

$$= O\left(\frac{\sigma_a^2}{d^2 n} - \beta_u^2\delta_{ts,sc}\right), \tag{28}$$

$$\mathbb{E}\|\hat{\boldsymbol{v}}-\boldsymbol{v}\|_2^2/n=\left(\frac{\sigma_a^2(n-1)}{d^2n^2}-\frac{n-p-2}{n}\beta_v^2\delta_{ts,sc}\right)(1+o(1)) \tag{29}$$

$$=O\left(\frac{\sigma_a^2}{d^2n}-\beta_v^2\delta_{ts,sc}\right), \tag{30}$$

$$\mathbb{E}\frac{\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_0\right\|_F^2}{\|\boldsymbol{A}_0\|_F^2}=\left(\frac{\sigma_a^2(2n-1)}{d^2n^2}-\frac{n-p-2}{n}(\beta_u^2+\beta_v^2)\delta_{ts,sc}\right)(1+o(1)) \tag{31}$$

$$=O\left(\frac{\sigma_a^2}{d^2n}-(\beta_u^2+\beta_v^2)\delta_{ts,sc}\right). \tag{32}$$

Remark 10. (The role of $\delta_{ts,sc}$) Let us consider the estimation of $\boldsymbol{u}$. Similar messages can be obtained for $\boldsymbol{v}$ and $\boldsymbol{A}_0$. Comparing the rate of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{u}}^{ts}$ in (28) and (12), $\mathbb{E}\|\hat{\boldsymbol{u}}^{ts}-\boldsymbol{u}\|_2^2/n-\mathbb{E}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|_2^2/n=\beta_u^2\delta_{ts,sc}$, where $\delta_{ts,sc}$ is defined in (26). As one can see, $\delta_{ts,sc}$ is the crucial quantity that measures the discrepancy between the two-stage and SuperCENT estimators for $\boldsymbol{u}$. When $\delta_{ts,sc}>0$, SuperCENT outperforms two-stage and vice versa. The positiveness $\delta_{ts,sc}>0$ requires $d^2n\sigma_y^2<(2\lambda d^2+\beta_u^2+\beta_v^2)\sigma_a^2$. On one hand, this inequality can be satisfied when λ is not too small and is satisfied when λ takes the optimal value $\lambda_0=n\sigma_y^2/\sigma_a^2$ given in the remark below. It implies that when the tuning parameter is properly selected, SuperCENT performs better than the two-stage. On the other hand, this inequality is more likely to hold when the signal of the regression β_u,β_v is large, or the noise of the regression σ_y is small, or the signal of the network d is small, or the noise of the network σ_a is large. This exactly corresponds to our intuition: when the signal-to-noise ratio in the regression model is high, we gain information from the regression model to assist centrality estimation; and the advantage is more pronounced when the signal-to-noise ratio in the network is low, which is exactly when the two-stage behaves poorly.

Remark 11. (Optimal λ) SuperCENT achieves the best performance with the following λ value

$$\lambda_0=\frac{n\sigma_y^2}{\sigma_a^2}. \tag{33}$$

SuperCENT with this λ_0 also obtains the most improvement over the two-stage procedure. It can be directly derived by minimizing (28), (30), or (32). Plugging the optimal value λ_0 into the SuperCENT objective function (7) leads to

$$(\hat{d},\hat{\boldsymbol{u}},\hat{\boldsymbol{v}},\hat{\boldsymbol{\beta}}_x,\hat{\beta}_u,\hat{\beta}_v)=\underset{\substack{\boldsymbol{\beta}_x,\beta_u,\beta_v\\ d,\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=\sqrt{n}}}{\arg\min}\ \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\boldsymbol{u}\beta_u-\boldsymbol{v}\beta_v\|_2^2}{\sigma_y^2}+\frac{\|\boldsymbol{A}-d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2}{\sigma_a^2}, \tag{34}$$

which is -2 times log likelihood when the errors $\boldsymbol{\epsilon}$ and $\boldsymbol{E}$ are normally distributed. Nevertheless, the objective function is just the scaled residual sum of squares of the regression and the scaled rank-one approximation error of the observed network, which does not require the normality assumption.

Remark 12. (SuperCENT-$\hat{\lambda}_0$ and SuperCENT-$\hat{\lambda}_{cv}$) The benefit of the optimal value λ_0 is twofold: 1) to benchmark the cross-validation procedure in Algorithm 4; 2) to avoid the time-consuming cross-validation by supplying a candidate for the tuning parameter λ. We can obtain a crude estimate of λ_0 by plugging in the two-stage estimates of σ_y^2 and σ_a^2. To be specific, after we obtain $\widehat{\boldsymbol{A}}^{ts}$ and $\hat{\boldsymbol{y}}^{ts}=\boldsymbol{X}\hat{\boldsymbol{\beta}}_x^{ts}+\hat{\boldsymbol{u}}^{ts}\hat{\beta}_u^{ts}+\hat{\boldsymbol{v}}^{ts}\hat{\beta}_v^{ts}$ from the two-stage procedure

by Algorithm 1, we estimate $(\hat{\sigma}_y^{ts})^2 = \frac{1}{n-p-2}(\hat{\boldsymbol{y}}^{ts} - \boldsymbol{y})^2$ and $(\hat{\sigma}_a^{ts})^2 = \frac{1}{n^2}\|\widehat{\boldsymbol{A}}^{ts} - \boldsymbol{A}_0\|_F^2$. We then plug in and obtain $\hat{\lambda}_0 = n(\hat{\sigma}_y^{ts})^2/(\hat{\sigma}_a^{ts})^2$. We refer to the SuperCENT with given $\hat{\lambda}_0$ as SuperCENT-$\hat{\lambda}_0$, whose empirical performance is given in the simulation study. Furthermore, $\hat{\lambda}_0$ can be used as a guide to lay out the cross-validation grid points in Algorithm 4, to obtain $\hat{\lambda}_{cv}$ and SuperCENT-$\hat{\lambda}_{cv}$.

Remark 13. (Comparison of the estimation of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{A}_0$ when two-stage is *inconsistent*) For the two-stage procedure, $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}, \widehat{\boldsymbol{A}}^{ts}$ is consistent if and only if $\kappa = \frac{\sigma_a^2}{d^2 n} \to 0$, which implies the network signal-to-noise ratio has to be large enough for the two-stage to be consistent. When $\kappa = O(1)$, the two-stage procedure is inconsistent. Can the SuperCENT estimates remain consistent under this regime?

The answer is positive. Plugging in the optimal λ_0, the rate of $\mathbb{E}\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|_2^2/n$ in (28) becomes

$$\kappa \frac{1 + \kappa \frac{\beta_v^2}{\sigma_y^2}}{1 + \kappa\left(\frac{\beta_u^2}{\sigma_y^2} + \frac{\beta_v^2}{\sigma_y^2}\right)}, \tag{35}$$

which is obviously smaller than κ, the rate of $\mathbb{E}\|\hat{\boldsymbol{u}}^{ts} - \boldsymbol{u}\|_2^2/n$ in (12). We want the above rate converges to 0 when $\kappa = O(1)$. Given (35), the convergence of $\hat{\boldsymbol{u}}$ boils down to the signal-to-noise ratio of $\boldsymbol{u}$ and $\boldsymbol{v}$ in the network regression model, i.e., $\frac{\beta_u^2}{\sigma_y^2}$ and $\frac{\beta_v^2}{\sigma_y^2}$. One sufficient condition for convergence is then $\frac{\beta_u^2}{\sigma_y^2} = O(n^c),\ c > 0$ and $\frac{\beta_v^2}{\sigma_y^2} = O(1)$, meaning the signal for $\boldsymbol{u}$ has to be stronger than both the signal for $\boldsymbol{v}$ and the noise σ_y to guarantee convergence of $\hat{\boldsymbol{u}}$.

If we want to guarantee the convergence of $\hat{\boldsymbol{v}}$ under this regime, one sufficient condition is $\frac{\beta_v^2}{\sigma_y^2} = O(n^c),\ c > 0$ and $\frac{\beta_u^2}{\sigma_y^2} = O(1)$. This conflicts with the requirement of the convergence of $\hat{\boldsymbol{u}}$. Fortunately, the rates of both $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are smaller than those of $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$, so SuperCENT always improves the estimation: when $\beta_u^2/\beta_v^2 = o(1)$ or $\beta_v^2/\beta_u^2 = o(1)$, one of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ will be consistent. We will demonstrate this phenomenon in the simulation.

Lastly, for the estimation of $\boldsymbol{A}_0$ with λ_0, the rate of $\mathbb{E}\|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F^2/\|\boldsymbol{A}_0\|_F^2$ in (32) becomes

$$\kappa \frac{1}{1 + \kappa\left(\frac{\beta_u^2}{\sigma_y^2} + \frac{\beta_v^2}{\sigma_y^2}\right)},$$

which is much smaller than κ, the rate of $\mathbb{E}\|\widehat{\boldsymbol{A}}^{ts} - \boldsymbol{A}_0\|_F^2/\|\boldsymbol{A}_0\|_F^2$ in (13). Better yet, to ensure $\widehat{\boldsymbol{A}}$ is consistent, we only require either $\frac{\beta_u^2}{\sigma_y^2} = O(n^c),\ c >$ or $\frac{\beta_v^2}{\sigma_y^2} = O(n^c),\ c > 0$. This means that, as long as one of SuperCENT $\hat{\boldsymbol{u}}$ or $\hat{\boldsymbol{v}}$ is consistent, SuperCENT $\widehat{\boldsymbol{A}}$ is consistent as well, while two-stage $\widehat{\boldsymbol{A}}^{ts}$ is only consistent when both $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are consistent.

When the two-stage estimator is consistent, the supervision effect of SuperCENT only takes place for the estimation of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{A}$ and the in-sample prediction, but not for $\beta_u, \beta_v, \boldsymbol{\beta}_x$, as shown in Corollary 5.

COROLLARY 5: *(Rate of $\widehat{\boldsymbol{\beta}}^{ts}$) Under the unified framework (4) and Assumptions 1, 2 and 3, the SuperCENT estimators satisfy the following,*

$$\mathbb{E}(\hat{\beta}_u - \beta_u)^2 = \mathbb{E}(\hat{\beta}_u^{ts} - \beta_u)^2 = O\left(\frac{\sigma_y^2}{n} + \frac{\sigma_a^2(\beta_u^2 + \beta_v^2)}{d^2 n^2}\right),$$

$$\mathbb{E}(\hat{\beta}_v - \beta_v)^2 = \mathbb{E}(\hat{\beta}_v^{ts} - \beta_v)^2 = O\left(\frac{\sigma_y^2}{n} + \frac{\sigma_a^2(\beta_u^2 + \beta_v^2)}{d^2 n^2}\right),$$

$$Cov\left(\widehat{\boldsymbol{\beta}}_x - \boldsymbol{\beta}_x\right) = Cov\left(\hat{\boldsymbol{\beta}}_x^{ts} - \boldsymbol{\beta}_x\right).$$

Remark 14. (Comparison of the estimation of β_u, β_v) When $\kappa \to 0$ and the two-stage is consistent, from the perspective of regression coefficient estimation, SuperCENT and the two-stage are similar. However, when $\kappa \nrightarrow 0$ and the two-stage regression coefficient estimation is biased as shown in Corollary 3. While for SuperCENT, if the conditions in Remark 13 hold, the SuperCENT coefficient estimates are still consistent and satisfy Corollary 5.

6. SIMULATION

In this section, we investigate the empirical performances, including the *estimation accuracy* and *inference property* of the two-stage and SuperCENT estimators under various simulation setups. We describe our simulation setup and overview the simulation results in Section 6.1 and show the simulation results for the inconsistent regime of the two-stage procedure in Section 6.2. The simulation results for the consistent regime of the two-stage procedure is deferred to the supplement.

6.1. *Simulation setup and results overview*

We generate the network as $\boldsymbol{A} = d\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E} \in \mathbb{R}^{n\times n}$, where $\boldsymbol{u} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{v} \in \mathbb{R}^{n\times 1}$ are vectors of the hub and authority centralities and all the entries of $\boldsymbol{E}$ follow $N(0, \sigma_a^2)$ independently. The elements of $\boldsymbol{u}$ are first generated from i.i.d. $N(0,1)$ and $\boldsymbol{v} = 0.5\boldsymbol{u} + \boldsymbol{\epsilon_v}$ where $\boldsymbol{\epsilon_v}$ are generated from i.i.d. $N(0,1)$. $\boldsymbol{u}$ and $\boldsymbol{v}$ are then re-scaled to have norm $\sqrt{n}$. For the regression model, $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}_x + \boldsymbol{u}\beta_u + \boldsymbol{v}\beta_v + \boldsymbol{\epsilon}$, we set $p = 3$, the coefficients for the covariates $\boldsymbol{\beta}_x = (1,3,5)^\top$, the covariate matrix $\mathbf{X}$ consists of a column of 1's and $p-1$ columns whose entries follow $N(0,1)$ independently, and $\boldsymbol{\epsilon} \sim N(0, \sigma_y^2 \boldsymbol{I}_n)$.

Since only the signal-to-noise of the network d/σ_a and the signal-to-noise of the regression $(\beta_u/\sigma_y, \beta_v/\sigma_y)$ matter to the properties of our estimator and inference, we fix $n = 2^8$, $d = 1$, and $\beta_v = 1$ to study the effect of σ_a, σ_y, and β_u. To study the effect of the signal-to-noise of the regression, we vary $\sigma_y \in 2^{-4,-2,0}$ and $\beta_u \in 2^{0,2,4}$ so that $\frac{\beta_u^2}{\sigma_y^2} = O(n^c),\ c \geqslant 0$ and $\frac{\beta_v^2}{\sigma_y^2} = O(1)$. Now that the signal-to-noise of the network is solely controlled by σ_a, we vary σ_a to differentiate the regimes when the two-stage estimator is consistent with small σ_a and inconsistent with large σ_a. Specifically, for the *consistent regime of the two-stage*, i.e., when the network noise-to-signal ratio $\kappa = \frac{\sigma_a^2}{d^2 n} \to 0$, we vary $\sigma_a \in 2^{-4,-2}$ to keep $\kappa < 2^{-12}$. For the *inconsistent regime of the two-stage*, i.e., $\kappa = O(1)$, we vary $\sigma_a \in 2^{0,2}$ so that $\kappa \in 2^{-8,-4} = O(1)$.

For each setting, we compare several two-stage-based and SuperCENT-based procedures in terms of estimation accuracy and inference property as follows.

Estimation accuracy For the estimation accuracy, we compare the following procedures:

1. **Two-stage**: the two-stage procedure as in Algorithm 1;
2. **SuperCENT-λ_0**: SuperCENT Algorithm 2 with the optimal $\lambda_0 = n\sigma_y^2/\sigma_a^2$ as in Remark 11, where the true σ_y, σ_a are used. In other words, SuperCENT-λ_0 is not implementable for real data and it only serves as the benchmark;
3. **SuperCENT-$\hat{\lambda}_0$**: SuperCENT with estimated tuning parameter $\hat{\lambda}_0 = n(\hat{\sigma}_y^{ts})^2/(\hat{\sigma}_a^{ts})^2$ based on Remark 12, where $(\hat{\sigma}_y^{ts})^2 = \frac{1}{n-p-2}\|\hat{\boldsymbol{y}}^{ts} - \boldsymbol{y}\|_2^2$ and $(\hat{\sigma}_a^{ts})^2 = \frac{1}{n^2}\|\widehat{\boldsymbol{A}}^{ts} - \boldsymbol{A}_0\|_F^2$ are estimated from the two-stage procedure. This way, SuperCENT-$\hat{\lambda}_0$ is implementable for real data;
4. **SuperCENT-$\hat{\lambda}_{cv}$**: SuperCENT with tuning parameter $\hat{\lambda}_{cv}$ chosen by 10-fold cross-validation as in Algorithm 4.

In the SuperCENT Algorithm 2, the tolerance parameter for the stopping criterion is set to be $\rho = 10^{-4}$. The following five performance metrics are used to compare these four procedures: the first two are from the perspective of network and centralities, and the last three are from the network regression.

1. The loss for estimating $\boldsymbol{u}$ and $\boldsymbol{v}$, $l(\hat{\boldsymbol{u}}, \boldsymbol{u}) = \|\boldsymbol{P}_{\hat{\boldsymbol{u}}} - \boldsymbol{P}_{\boldsymbol{u}}\|_2^2$ and $l(\hat{\boldsymbol{v}}, \boldsymbol{v}) = \|\boldsymbol{P}_{\hat{\boldsymbol{v}}} - \boldsymbol{P}_{\boldsymbol{v}}\|_2^2$;
2. The loss for estimating $\boldsymbol{A}_0$, $l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0) = \|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F^2/\|\boldsymbol{A}_0\|_F^2$;
3. The normalized squared error loss for estimating β_u and β_v, $l(\hat{\beta}_u, \beta_u) = (\hat{\beta}_u - \beta_u)^2/\beta_u^2$ and $l(\hat{\beta}_v, \beta_v) = (\hat{\beta}_v - \beta_v)^2/\beta_v^2$;
4. The estimation bias for regression coefficients $\hat{\beta}_u - \beta_u$ and $\hat{\beta}_v - \beta_v$.

Inference property For the inference property, let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$-quantile of the standard normal distribution and we consider the following procedures to construct the confidence intervals (CIs) for the regression coefficient, CI_{β_u} and CI_{β_v}:

1. **Two-stage-adhoc**: $\hat{\beta}^{ts} \pm z_{1-\alpha/2}\hat{\sigma}^{OLS}(\hat{\beta}^{ts})$, where $\hat{\beta}^{ts}$ is the two-stage estimate of β and $\hat{\sigma}^{OLS}(\hat{\beta}^{ts})$ is the standard error from OLS, assuming $\hat{\boldsymbol{u}}^{ts}, \hat{\boldsymbol{v}}^{ts}$ are fixed predictors;
2. **Two-stage-oracle**: $\hat{\beta}^{ts} \pm z_{1-\alpha/2}\sigma(\hat{\beta}^{ts})$, where $\sigma(\hat{\beta}^{ts})$ is the standard error of $\hat{\beta}^{ts}$, whose mathematical expressions are given in (14)-(15) or (17)-(18) and the true parameters are plugged into those expressions;
3. **Two-stage**: $\hat{\beta}^{ts} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\beta}^{ts})$, where $\hat{\sigma}(\hat{\beta}^{ts})$ is the standard error of $\hat{\beta}^{ts}$ by plugging all the two-stage estimators into (14)-(15) or (17)-(18).
4. **SuperCENT-λ_0-oracle**; $\hat{\beta}^{\lambda_0} \pm z_{1-\alpha/2}\sigma(\hat{\beta}^{\lambda_0})$, where $\hat{\beta}^{\lambda_0}$ is the estimate of β by SuperCENT-λ_0 and $\sigma(\hat{\beta}^{\lambda_0})$ follows (14)-(15) or (17)-(18), with the true parameters plugged in;
5. **SuperCENT-$\hat{\lambda}_{cv}$**: $\hat{\beta}^{\hat{\lambda}_{cv}} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\beta}^{\hat{\lambda}_{cv}})$, where $\hat{\beta}^{\hat{\lambda}_{cv}}$ is the estimate of β by SuperCENT-$\hat{\lambda}_{cv}$ and $\hat{\sigma}(\hat{\beta}^{\hat{\lambda}_{cv}})$ is obtained by plugging the SuperCENT-$\hat{\lambda}_{cv}$ estimates into (14)-(15) or (17)-(18).

For these five methods, we compare the empirical coverage probability (CP) and the average width of the confidence intervals. The experiments are repeated 500 times. In parallel, the same five procedures can be used to study the inference regarding the true network $\boldsymbol{A}_0$. The only difference is that the inference is made for each entry of $\boldsymbol{A}_0$, which we denote as $CI_{a_{ij}}$; and the CP and the average width reported are the average over all the entries a_{ij}.

To give an overview of the simulation results, Table I summarizes the comparison of the two-stage method and SuperCENT in the consistent and the inconsistent regimes of the two-stage from the perspectives of both estimation and inference. SuperCENT universally outperforms the two-stage in terms of centrality estimation, regression coefficients estimation, and infer-

TABLE I

COMPARISON OF THE TWO-STAGE AND SUPERCENT IN THE CONSISTENT AND INCONSISTENT REGIMES OF THE TWO-STAGE WHEN $\beta_u^2/\sigma_y^2 = O(n^c),\ c > 0$ AND $\beta_v^2/\sigma_y^2 = O(1)$. IF $\beta_v^2/\sigma_y^2 = O(n^c),\ c > 0$ AND $\beta_u^2/\sigma_y^2 = O(1)$, THE RESULTS FOR $\boldsymbol{u}$ AND $\boldsymbol{v}$, β_u AND β_v WILL BE SWITCHED. IN THE ESTIMATION PANEL, ✓ INDICATES THAT THE ESTIMATION IS CONSISTENT, ✗ INDICATES THAT THE ESTIMATION IS INCONSISTENT. IN THE INFERENCE PANEL, ✓ INDICATES THAT THE EMPIRICAL COVERAGE OF CONFIDENCE INTERVAL IS NO LESS THAN THE NOMINAL LEVEL, ✗ INDICATES THAT THE CONFIDENCE INTERVAL FAILS TO REACH THE NOMINAL LEVEL. IN EACH ROW, SUPERCENT IS UNDERLINED WHENEVER IT OUTPERFORMS THE TWO-STAGE.

	$\kappa \to 0$: two-stage consistent		$\kappa \not\to 0$: two-stage inconsistent	
	Two-stage	SuperCENT	Two-stage	SuperCENT
		Estimation		
$\boldsymbol{u}$	✓	Improved	✗	✓
$\boldsymbol{v}$	✓	Slightly Improved	✗	✗ (Slightly Improved)
$\boldsymbol{A}_0$	✓	Improved	✗	✓
β_u	✓	✓	✗ (Biased)	✓
β_v	✓	✓	✗ (Biased)	✗ (Biased)
		Inference		
CI_{β_u}	✓ (Conservative)	✓ (Shorter)	✗	✓
CI_{β_v}	✓	✓	✗	✗
$CI_{a_{ij}}$	✓	✓ (Shorter)	✗	✓

ence. In what follows, we focus on the inconsistent regime of the two-stage and defer the consistent regime of the two-stage to the supplement.

6.2. *Simulation results for the inconsistent regime of the two-stage procedure*

In the inconsistent regime of two-stage where $\kappa = \frac{\sigma_a^2}{d^2 n} = O(1)$, the two-stage procedure cannot consistently estimate $\boldsymbol{u}$ and $\boldsymbol{v}$, i.e., $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are inconsistent. Recall that we fix $\beta_v = 1$ and vary $\sigma_y \in 2^{-4,-2,0}$, $\beta_u \in 2^{0,2,4}$, and $\sigma_a \in 2^{0,2}$, so that the scaled noise-to-signal of the network $\kappa \in 2^{-8,-4} = O(1)$, and the signal-to-noise of the regression $\frac{\beta_u^2}{\sigma_y^2} = O(n^c),\ c > 0$ and $\frac{\beta_v^2}{\sigma_y^2} = O(1)$. In such a range, we expect the two-stage to be inaccurate for the estimation of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{A}_0$ and SuperCENT to be much more accurate for $\boldsymbol{u}, \boldsymbol{A}_0$, as mentioned in Remark 13. In addition, the two-stage estimates of β_u, β_v are biased as shown in Corollary 3 while SuperCENT estimates remain consistent as in Remark 14. As to the inference property, the two-stage-based confidence intervals are expected to be under-covered and wider than necessary while SuperCENT confidence intervals are valid and narrower.

Estimation accuracy Figure 3 shows the boxplot of the logarithm of $l(\hat{\boldsymbol{u}}, \boldsymbol{u})$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. The rows correspond to $\log_2(\sigma_a)$ and the columns correspond to $\log_2(\beta_u)$. For each panel, the x-axis is $\log_2(\sigma_y)$ and the y-axis is $\log_{10}(l(\hat{\boldsymbol{u}}, \boldsymbol{u}))$. The super-imposed red symbols show the theoretical rates of $\hat{\boldsymbol{u}}^{ts}$ in Corollary 1 and that of $\hat{\boldsymbol{u}}$ in Corollary 4. The two-stage estimator performs the same no matter how large σ_y and β_u are, and it has a smaller error with smaller σ_a. The performance of SuperCENT is better with smaller σ_a, σ_y or larger β_u. As expected, the three SuperCENT-based methods estimate $\boldsymbol{u}$ much more accurately than the two-stage procedure. In particular, the supervision effect of $(\boldsymbol{X}, \boldsymbol{y})$ is more pronounced when the noise of the outcome regression, σ_y, is small, or when the signal of

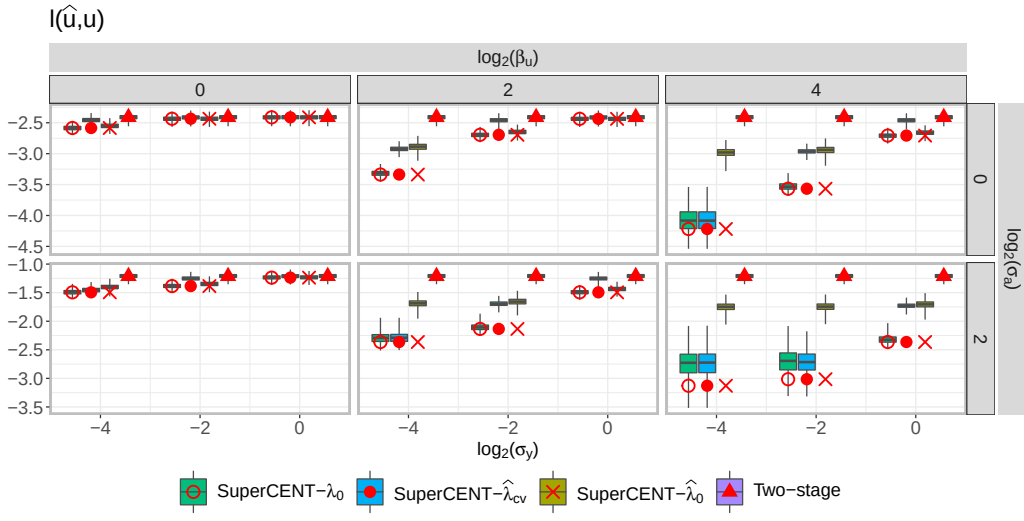


FIGURE 3.—Inconsistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\boldsymbol{u}}, \boldsymbol{u}))$ for the four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\hat{\boldsymbol{u}}, \boldsymbol{u}) = \|\boldsymbol{P}_{\hat{\boldsymbol{u}}} - \boldsymbol{P}_{\boldsymbol{u}}\|_2^2$. The super-imposed red symbols show the theoretical rates of the two-stage in Corollary 1 and SuperCENT in Corollary 4.

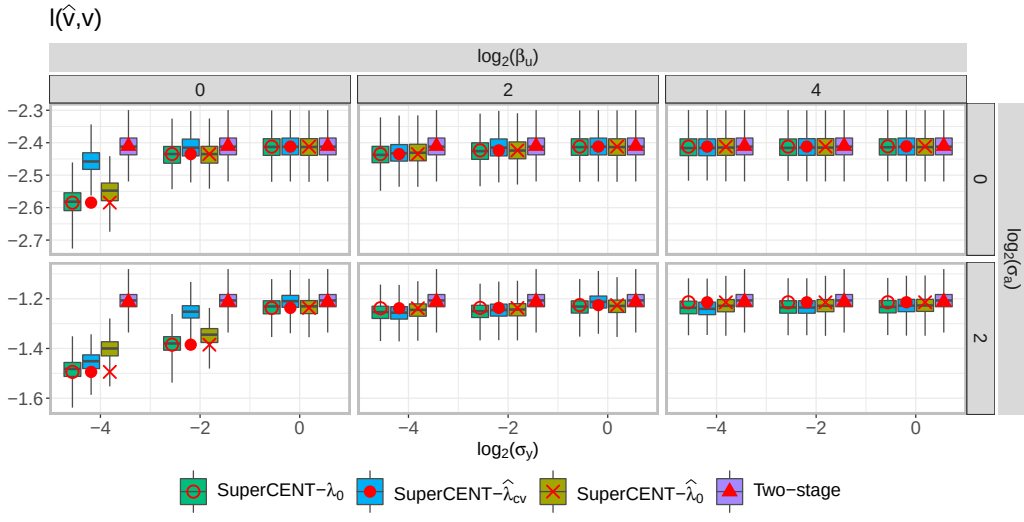


FIGURE 4.—Inconsistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\boldsymbol{v}}, \boldsymbol{v}))$ for the four estimators across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. where $l(\hat{\boldsymbol{v}}, \boldsymbol{v}) = \|\boldsymbol{P}_{\hat{\boldsymbol{v}}} - \boldsymbol{P}_{\boldsymbol{v}}\|_2^2$. The super-imposed red symbols show the theoretical rates of the two-stage in Corollary 1 and SuperCENT in Corollary 4.

the outcome regression, β_u, is large, or when the network noise-to-signal, $\sigma_a/d = \sigma_a$ is large. This validates Remarks 10 and 13 on the theoretical comparison of the estimators. Comparing the three SuperCENT-based methods, the benchmark SuperCENT-λ_0 is always the best, SuperCENT-$\hat{\lambda}_{cv}$ and SuperCENT-$\hat{\lambda}_0$ are sometimes worse than SuperCENT-λ_0, but still better than the two-stage. SuperCENT-$\hat{\lambda}_0$ is typically comparable to or worse than SuperCENT-$\hat{\lambda}_{cv}$, because SuperCENT-$\hat{\lambda}_0$ fails to locate the optimal λ_0 due to inaccurate estimate of σ_a and σ_y from the two-stage procedure.

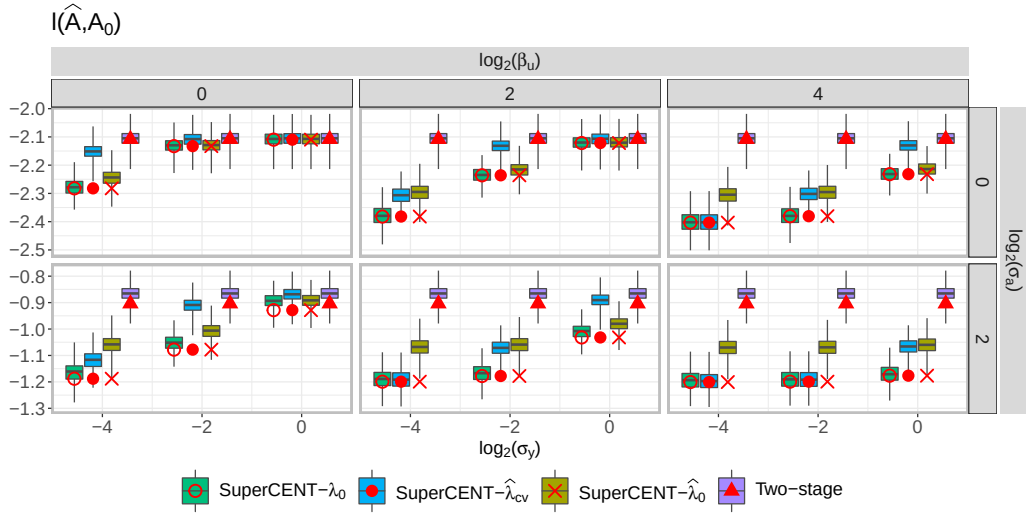


FIGURE 5.—Inconsistent regime of two-stage: The boxplot of $\log_{10}(l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0))$ for four estimators across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$ where $l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0) = \|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F^2 / \|\boldsymbol{A}_0\|_F^2$. The super-imposed red symbols show the theoretical rates of the two-stage in Corollary 1 and SuperCENT in Corollary 4.

For the estimation of $\boldsymbol{v}$ shown in Figure 4, the improvement of SuperCENT over two-stage is not as large as that of the estimation of $\boldsymbol{u}$ when $\beta_u \in 2^{2,4}$, because $\frac{\beta_u^2}{\sigma_y^2} = O(n^c),\ c > 0$ and $\frac{\beta_v^2}{\sigma_y^2} = O(1)$. But the improvement is still quite significant when $\beta_u = 2^0 \approx \beta_v = 1$. It is worth noting that the supervised effect to $\hat{\boldsymbol{v}}$ shrinks as β_u increases, leading to a different trend comparing Figures 3 and 4. This phenomena aligns with Remark 13 where we discuss the estimation of $\boldsymbol{u}$ and $\boldsymbol{v}$ when the two-stage is inconsistent. Specifically, the roles of $\boldsymbol{u}$ and $\boldsymbol{v}$ are not exchangeable, because here we have $\beta_v \leqslant \beta_u$ by fixing $\beta_v = 1$ and varying $\beta_u \in 2^{0,2,4}$. On the other hand, when $\beta_v \gg \beta_u$ we should expect the improvement in estimating $\boldsymbol{v}$ to increase.

The conclusion for the estimation of $\boldsymbol{A}_0$ is similar to that of $\boldsymbol{u}$ as shown in Figure 5. With the improvement from estimating $\boldsymbol{u}$ and $\boldsymbol{v}$, SuperCENT-$\hat{\lambda}_{cv}$ estimates $\boldsymbol{A}_0$ more accurately across all the settings. As claimed in Remark 13, the convergence of $\widehat{\boldsymbol{A}}$ in this regime only requires $\frac{{\beta_u}^2}{\sigma_y^2} = O(n^c),\ c > 0$ or $\frac{{\beta_v}^2}{\sigma_y^2} = O(n^c),\ c > 0$. Therefore, with $\beta_v = 1$, $\beta_u > 1$, $\widehat{\boldsymbol{A}}$ converges and $l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0) < l(\widehat{\boldsymbol{A}}^{ts}, \boldsymbol{A}_0)$. Comparing Figures 3, 4 and 5 altogether, when $\beta_u = 2^0$, SuperCENT improves the estimation of both $\boldsymbol{u}$ and $\boldsymbol{v}$ significantly; when $\beta_u \in 2^{2,4}$, SuperCENT improves the estimation of $\boldsymbol{u}$ a lot; therefore, SuperCENT improves the estimation of $\boldsymbol{A}_0$ a lot for all the ranges of β_u.

The attention is next turned to the regression coefficient β_u. Based on Corollary 3 on the issues of measurement error, the two-stage coefficient estimates tend to have bias under the inconsistent regime for the two-stage and the directions of the bias depend on the size of β_u, β_v, κ, and the correlation ρ between $\boldsymbol{u}$ and $\boldsymbol{v}$. Figure 6 shows the estimation bias $\hat{\beta}_u - \beta_u$. With large σ_a or large β_u, the two-stage estimates suffer from sever attenuation bias, while SuperCENT can alleviate the bias. The attenuation bias by the two-stage can be explained by Remark 7 as follows. In this regime where $\kappa = 2^{-8,-4} \nrightarrow 0$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are correlated with $\rho = \frac{2^{-1}}{\sqrt{1.25}}$, and $\beta_u \in 2^{2,4} > 1 \cdot 2^{-1.2} = \beta_v \frac{\rho}{(1+\kappa)}$, then plim $\hat{\beta}_u^{ts} - \beta_u < 0$. Hence, $\hat{\beta}_u^{ts}$ has an attenuation bias and the bias becomes larger as β_u increases. On the other hand, this also

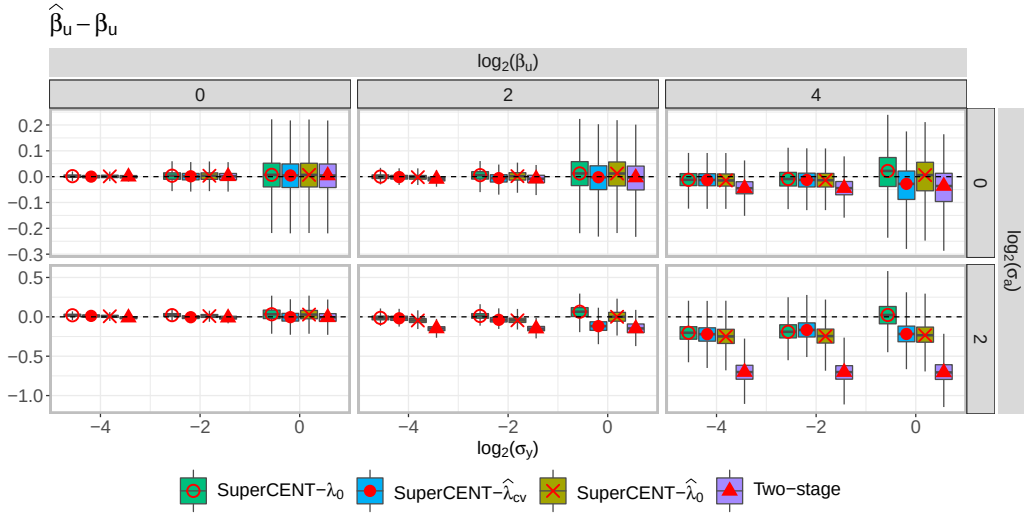


FIGURE 6.—Inconsistent regime of two-stage: The boxplot of the bias $\hat{\beta}_u - \beta_u$ across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$. The dashed lines show $\hat{\beta}_u - \beta_u = 0$. The super-imposed red points show the median of the bias.

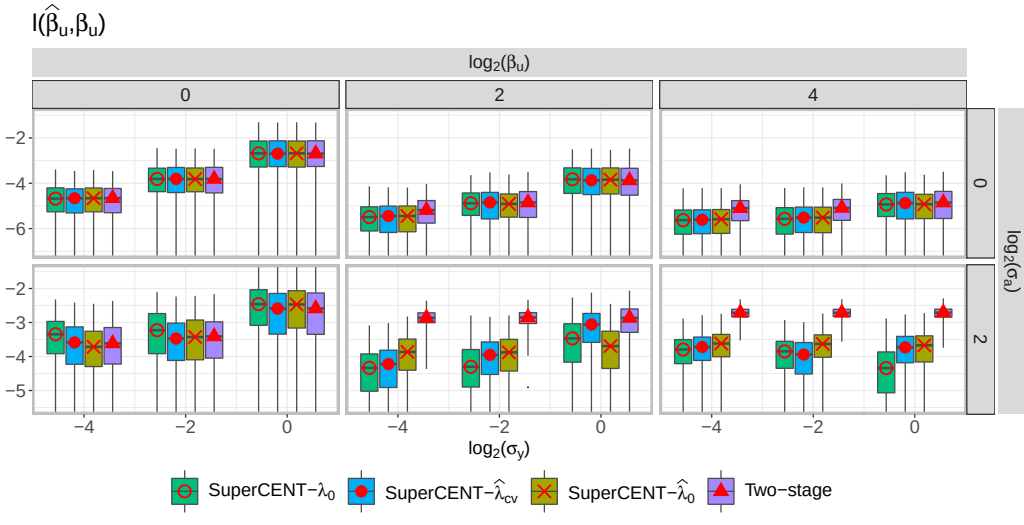


FIGURE 7.—Inconsistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\beta}_u, \beta_u))$ for four estimators across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$ where $l(\hat{\beta}_u, \beta_u) = (\hat{\beta}_u - \beta_u)^2/\beta_u^2$. The super-imposed red points show the median of $\log_{10}(l(\hat{\beta}_u, \beta_u))$.

implies that the two-stage estimation of β_v is biased away from zero and the bias is also larger as β_u increases as shown in the supplement. The improvement of SuperCENT over the two-stage is relatively small and sometimes negligible for β_v.

In terms of the squared error loss of the estimation of β_u as shown in Figure 7, since $\hat{\beta}_u^{ts}$ suffers from an attenuation bias with large σ_a and β_u and SuperCENT can alleviate the bias, SuperCENT improves over two-stage in the mean squared error $l(\hat{\beta}_u, \beta_u)$, corroborating Remark 14.

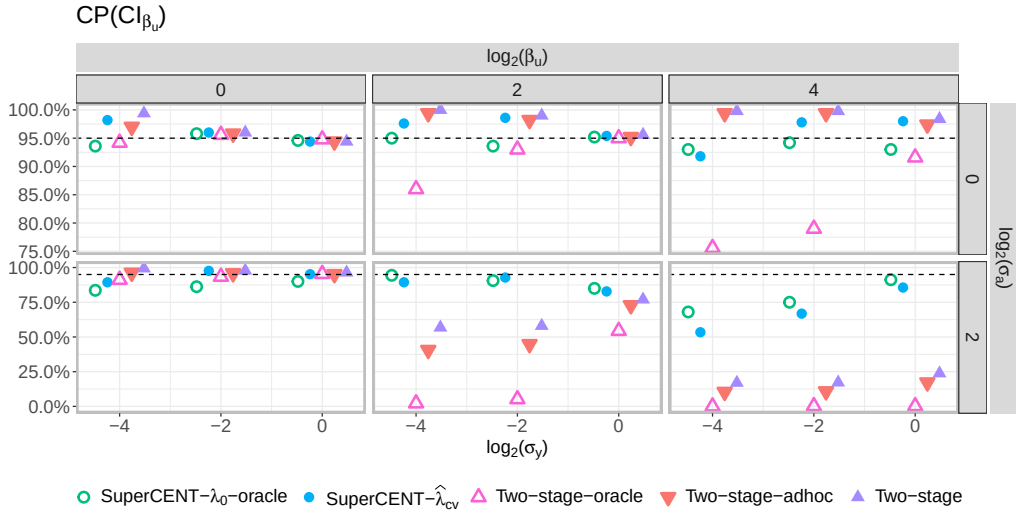


FIGURE 8.—Inconsistent regime of two-stage: Empirical coverage of CI_{β_u} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles. The dashed lines show the nominal confidence level 0.95.

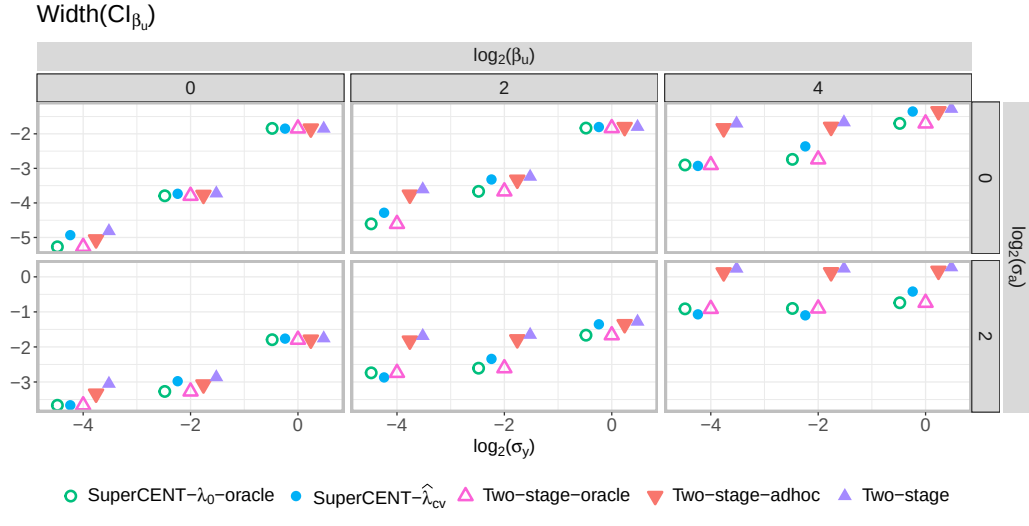


FIGURE 9.—Inconsistent regime of two-stage: The $\log_{10}$ of the width of CI_{β_u} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles.

Inference property We now switch gears to the inference property. The bias in the estimation of β_u by the two-stage further affects its confidence interval. Figures 8 and 9 show the empirical coverage and the average width of the 95% confidence interval for β_u respectively. For the empirical coverage, when β_u is small (leftmost column), all the methods are close to the nominal level. When β_u increases and σ_a remains small (top right two panels), all the methods remain valid except for two-stage-oracle, but different methods remain valid for different rea-

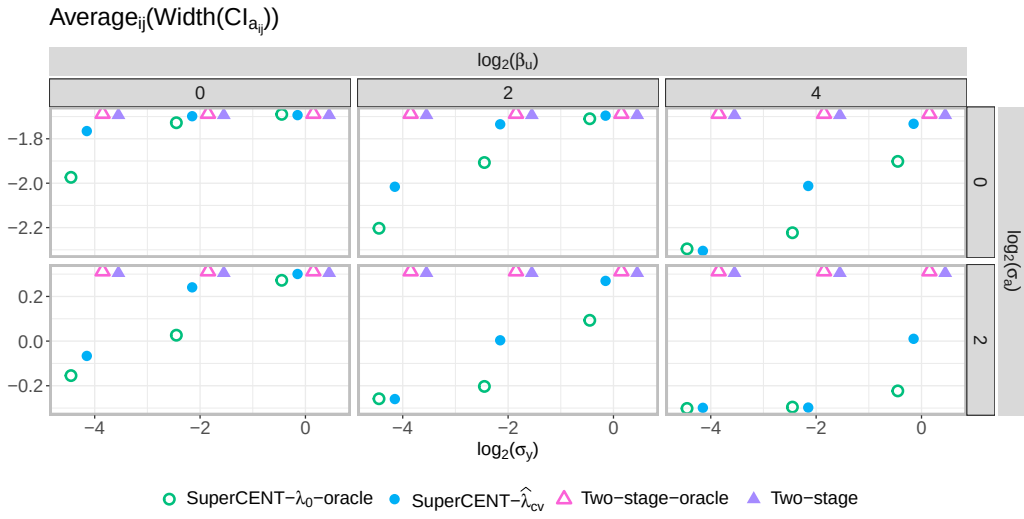


FIGURE 10.—Inconsistent regime of two-stage: log_{10} of the average width of $CI_{a_{ij}}$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles.

sons. The two SuperCENT-based methods remain valid because there is no estimation bias and the estimation of the standard error is accurate. Two-stage and two-stage-adhoc remain valid mainly because they over-estimate σ_y^2, and this conservative-ness covers up the issue of bias. Two-stage-oracle uses the true σ_y^2 and the issue of bias uncovers itself, consequently invalidating the inference. When β_u increases and σ_a gets large as well (bottom right two panels), the over-estimation of σ_y^2 can no longer conceal the issue of bias and all two-stage related methods are not valid anymore. Again, the SuperCENT can mitigate the bias and the coverage probability is closer to the nominal level.

As for the width of CI_{β_u}, Figure 9 shows that the confidence intervals by the SuperCENT-based methods have better coverage and are narrower than those by the two-stage methods. The improvement in the width is more significant with larger β_u, σ_a.

For the confidence interval of the authority centrality coefficient, CI_{β_v}, the improvement of SuperCENT over two-stage is relatively small in terms of both the coverage probability and width as shown in the supplement, a similar phenomena as of the estimation accuracy of β_v.

Finally, we investigate the average coverage and the average width of confidence intervals for all the entries a_{ij} of $\boldsymbol{A}_0$ respectively. The average coverage probability of all the methods, $\text{Average}_{ij}\big(\text{CP}(\text{CI}a_{ij})\big)$, achieves the nominal level of 95% as shown in the supplement. The coverage tends to be slightly below the nominal coverage as σ_a increases, because the estimation becomes worse and the theorem only holds up to $1 + o(1)$. SuperCENT-$\hat{\lambda}_{cv}$ is the closest to the nominal coverage in all the settings compared to the others. Figure 10 shows the $\log_{10}$ of the average width of the CIs, $\text{Average}_{ij}\big(\text{Width}(\text{CI}a_{ij})\big)$. SuperCENT-$\lambda_0$-oracle provides the shortest width among the four methods, followed by SuperCENT-$\hat{\lambda}_{cv}$. The widths of the confidence intervals of both SuperCENT-based methods are shorter than those of the two-stage methods. Again, the improvement of SuperCENT over the two-stage increases as σ_a and β_u increase or σ_y decreases.

7. CASE STUDY: GLOBAL TRADE NETWORK AND CURRENCY RISK PREMIUM

We consider a real case study with a triplet of $\{\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}\}$, where $\boldsymbol{A}$ is the country-level trade network, $\boldsymbol{y}$ is the currency risk premium, and $\boldsymbol{X}$ is GDP share, whose detailed information and construction will be given shortly. In this case study, we demonstrate that SuperCENT can provide more accurate estimation of the centrality, which is closely related to currency risk premium, and hence has a profound and lucrative implication on portfolio management. We further show that the SuperCENT method outperforms the two-stage methods in the inference of regression coefficients, providing less biased estimates and narrower CIs, and thus strengthens a related economic theory.

In the literature of international finance, economists have been studying currency risk premium extensively and are puzzled by its driving forces. The currency risk premium is formally defined as the excess return from holding foreign currency compared to holding the US dollar. Specifically, for an investor going long in a country/region i, the log risk premium "rx" at time $t+1$ is

$$\mathrm{rx}_{i,t+1} := r_{it} - r_t - \Delta q_{i,t+1}, \tag{36}$$

where r_{it} is log interest rate of country/region i, r_t is the log interest rate of the U.S. and $\Delta q_{i,t+1}$ is the appreciation of U.S. dollar.

In this case study, we investigate how the global trade network drives the currency risk premium and build a regression model that regresses the currency risk premium on the centrality from the international trade network. Such a predictive relationship is motivated by Richmond (2019), which developed a general equilibrium with international trade between countries and showed that countries' positions in the trade network can explain the difference in currency premiums across countries. Specifically, he showed an economic theory that countries that are central in the trade network exhibit lower currency risk premiums. This has two implications: (i) the regression coefficients for the centralities should be negative; (ii) international investors can obtain profit through taking a long-short strategy for foreign exchange – take a long position in currencies of countries with low centralities and a short position in currencies of countries with high centralities. Therefore, if the centralities can be estimated accurately, one can yield a significant investment return based on the strategy.

To verify the economic theory and construct a profitable portfolio, we first need to compute the currency risk premium and construct the global trade network following Richmond (2019), because these data are not directly available. We then apply the developed methodologies and theories to compare the two-stage and SuperCENT. We focus on the period between 1999 and 2013[2].

To compute the currency risk premium, we obtain the interest rates and the exchange rates from DataStream. The currency risk premium can be calculated by plugging the interest rates and exchange rates into the definition of risk premium (36). Only 25 countries/regions have exchange rates available during the period of interest. We further exclude the region of Europe as it is not comparable to the others in the trade network, resulting in 24 countries/regions in the end[3]. We use a 5-year moving average of the currency risk premium. Specifically, when considering year t, all the relevant quantities are the averages from year $t-4$ to year t. Figure 11 shows the time series plot of the rank of the 5-year moving average of risk premium from 2003 to 2012[4] for the 24 countries/regions. In each year, we rank the 24 countries/regions from

[2]Euro was first adopted in 1999. Exchange rate for Malaysia (MYS) is not available from 1999 to 2004. The bilateral trade data is only available till 2013.

[3]The list of country acronyms is provided in the supplement.

[4]We leave the last available year 2013 for the validation purpose.

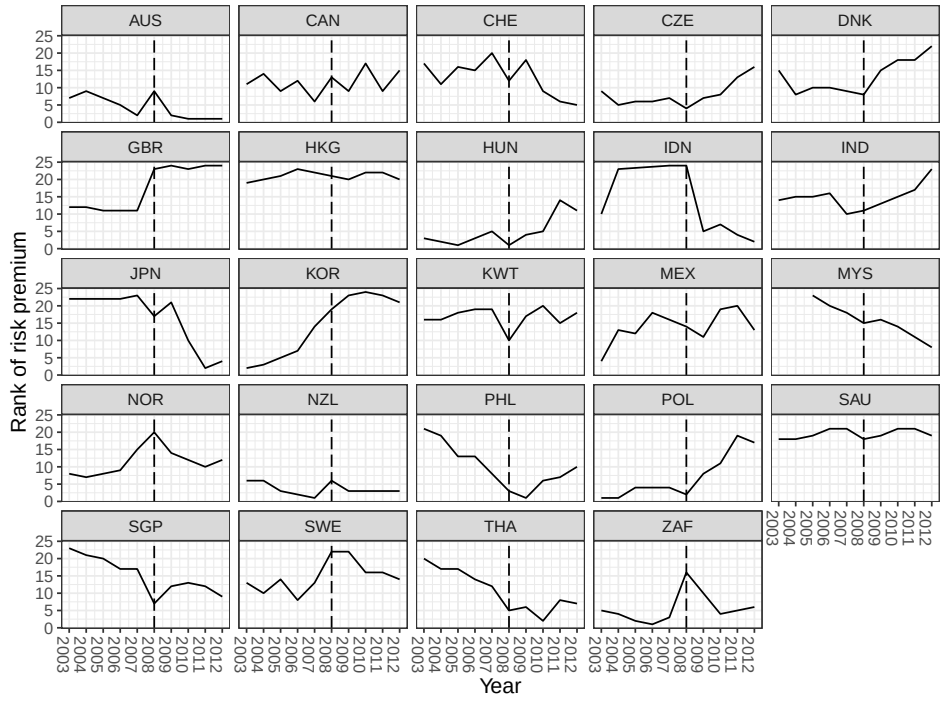


FIGURE 11.—Time series of the risk premium ranking in descending order from 2003 to 2012. The vertical dashed line indicates 2008, the year of the financial crisis.

1 to 24 from the largest risk premium to the smallest, i.e., in descending order of the risk premium.

Richmond (2019) defined the trade linkage as the trade amount normalized by the pair-wise total GDP, which represents the relative trade (export/import) intensity between two countries. Specifically, the trade linkage between two countries is computed as

$$a_{ijt} = \frac{S_{ijt}}{GDP_{it} + GDP_{jt}}, \tag{37}$$

where S_{ijt} is the dollar value of goods and commodities exported from country i to country j at time t, and GDP_{it} is the GDP of country i at time t in U.S. dollar. The bilateral trade data come from the correlates of war project (COW) (Barbieri et al., 2009) and the International Monetary Fund (IMF) Direction of Trade Statistics[5]. Current U.S. dollar GDP (using 2015 as the base year) data are from the World Bank's World Development Indicators[6]. Same as the currency risk premium, we also use the 5-year moving average in the following analysis. In the supplement, we show a circular plot to visualize the average trade volume from 2003 to 2012.

As neither the two-stage nor SuperCENT is applicable for panel data, we will repeat the analysis for each year from 2003 to 2012. Besides the network and the response variable, we also include the predictor of GDP share, which is defined as the percentage of country/region GDP among the world GDP, where the world GDP is the total GDP of all available countries

[5]https://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85
[6]https://databank.worldbank.org/source/world-development-indicators

in the sample for that year. In summary, the models are, for each t,

$$a_{ijt} = d\,\text{Hub}_{it} \times \text{Authority}_{jt} + e_{ijt}, \tag{38}$$

$$\text{rx}_{it} = \alpha + \beta_{ut} \cdot \text{Hub}_{it} + \beta_{vt} \cdot \text{Authority}_{it} + \beta_{xt} \cdot \text{GDP share}_{it} + \epsilon_{it}. \tag{39}$$

In Sections 5 and 6, we have demonstrated that the two-stage procedure is problematic under large network noise. In this case study, the observational error of the network comes from two sources: GDPs and the trade volumes, because each entry of the observed network a_{ijt} is defined as (37), i.e., the trade tie S_{ijt} between country i and country j normalized by their GDPs. GDPs and the trade volumes are often measured with errors. For the GDP, its accounting has been a challenge in macroeconomics (Landefeld et al., 2008). For the trade volume, the measurement errors are mostly due to (i) underground or illegal import and export; (ii) not including service trade; (iii) trade cost like transportation or taxes (Lipsey, 2009). Consequently, the observed trade network $\boldsymbol{A}$ can be very noisy and the two-stage method can perform badly.

On the other hand, SuperCENT can significantly improve over the two-stage when the network noise is large. In what follows, we focus on SuperCENT-$\hat{\lambda}_{cv}$ using 10-fold cross-validation. We will refer SuperCENT-$\hat{\lambda}_{cv}$ to SuperCENT for simplicity and use the superscript sc for the SuperCENT-$\hat{\lambda}_{cv}$ estimates. Figure 12 shows the time series plots of the ranking of the hub centrality estimated by two-stage and SuperCENT for the 24 countries/regions, together with the ranking of the currency risk premium. Figure 13 is for the authority centrality. We rank the centrality in ascending order and the risk premium in descending order. Based on the negative relationship between centralities and risk premium established in Richmond (2019), the closer the trends of rankings between centralities and risk premium are, the better the centralities capture the time variation in the risk premium. In general, the centrality estimated by the two-stage procedure is relatively more stable over time compared to SuperCENT, since SuperCENT incorporates information of both the GDP share and currency risk premium, which is more volatile than the trade network itself. Asian trade hubs such as Hong Kong (HKG) and Singapore (SGP) are the most central; while countries like South Africa (ZAF) and New Zealand (NZL) are peripheral. Comparing with the ranking of risk premium, the time variation is not reflected in the centrality estimated by the two-stage procedure, while it can be captured by SuperCENT. In 2008, the year of the financial crisis, the SuperCENT centralities fluctuate together with risk premium while the two-stage centralities almost remained unchanged.

To emphasize the importance of an accurate centrality estimation for portfolio management, we examine whether a long-short strategy based on our estimated centrality can significantly boost investment performance. Specifically, Richmond (2019) showed in theory that a country with low centrality exhibits a higher expected currency premium than the ones with high centralities. We have shown in theory and simulations that the two-stage centrality estimation can be off from or even orthogonal to the truth when the signal-to-noise ratio of the network is low. Thus, if we can estimate the centralities more accurately than the two-stage procedure, we expect to obtain a higher return through longing countries with low centralities and shorting countries with high centralities.

To illustrate this insight, for each method, we take a long position on the currencies with the lowest 3 centralities (bottom 10%) and a short position on the currencies with the highest 3 centralities (top 10%). That is, we obtain a long-short return based on the estimated centrality of the period between year $t-4$ and t. Figure 14 shows the year $t+1$ return based on this strategy. The return based on the centrality estimated by SuperCENT is much higher than that of the two-stage procedure. Table II shows the 10-year average return based on this strategy with the top and bottom 3, 4, and 5 currencies, respectively. The 10-year average return based

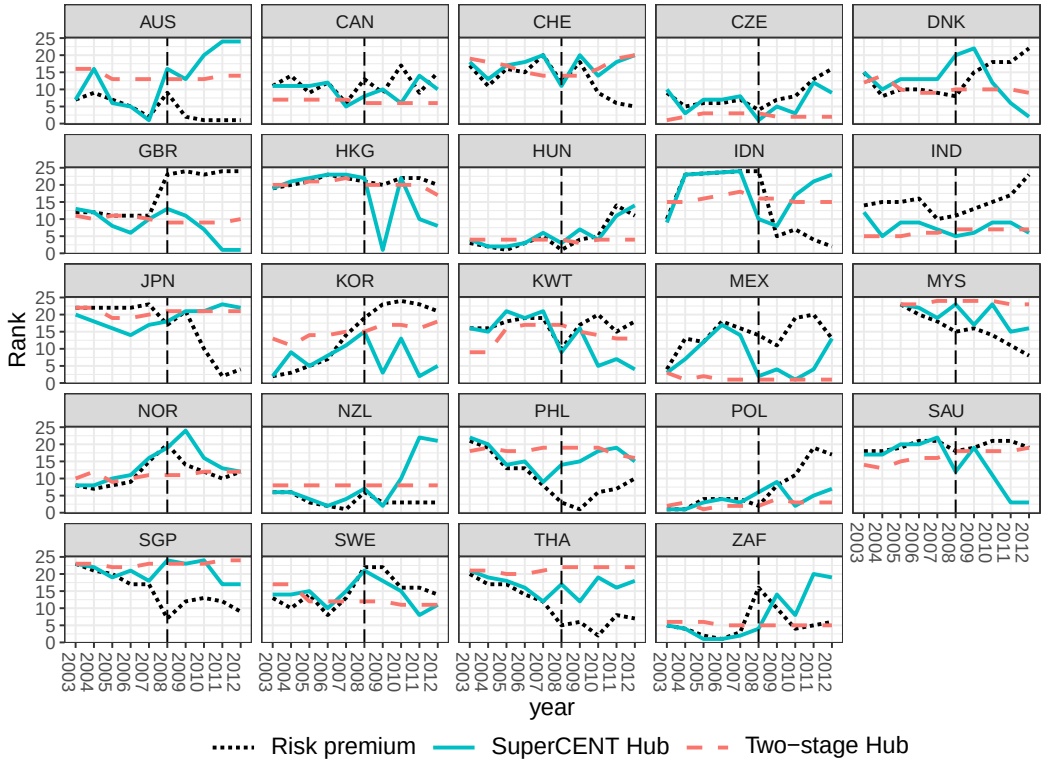


FIGURE 12.—Time series of hub centrality ranking in ascending order from 2003 to 2012. The ranking of risk premium is in descending order, same as Figure 11. If the trend of centralities is close to the trend of risk premium, then the centralities capture the time variation of risk premium, based on the negative relationship between the two as claimed in Richmond (2019). The vertical dashed line indicates 2008, the year of the financial crisis.

on the SuperCENT centralities increased more than twice from that of the two-stage procedure. Thus, an accurate estimate of the centrality can significantly boost the average portfolio return.

We further demonstrate the superiority of SuperCENT in inference. Again since our method is not directly applicable to longitudinal data, we take the 10-year average of trade volume and GDP to construct a 10-year trade network and GDP share. Similarly, we take the 10-year average of risk premium as the response.

To better understand the behavior of the two-stage and SuperCENT estimators, it is crucial to know which regime the trade network belongs to. However, the true noise-to-signal ratio κ of the trade network is unknown, so we estimate it using SuperCENT: $\hat{\kappa}^{sc} = 0.154 \approx 2^{-3}$, which falls in the inconsistent regime of the two-stage. Note that for the simulation study, when $\kappa = 2^{-8}$, two-stage already shows inconsistency.

To further comprehend the behavior of SuperCENT and gauge how much improvement SuperCENT can potentially achieve in the inconsistent regime, we estimate the signal-to-noise ratio of the regression: $(\hat{\beta}_u^{sc}/\hat{\sigma}_y^{sc})^2 = 7.6 \times 10^6 \approx 2^{23}$ and $(\hat{\beta}_v^{sc}/\hat{\sigma}_y^{sc})^2 = 1.8 \times 10^5 \approx 2^{17}$. Compared with the simulation settings in the inconsistent regime where $\kappa = 2^{-4}$, $\beta_u^2/\sigma_y^2 \leqslant 2^{16}$ and $\beta_v^2/\sigma_y^2 \leqslant 2^8$, we should expect SuperCENT to improve enormously over two-stage for both the estimation and inference of β_u, thanks to a large $(\hat{\beta}_u^{sc}/\hat{\sigma}_y^{sc})^2$ and $\hat{\beta}_u^{sc} \gg \hat{\beta}_v^{sc}$ under a relatively large $\hat{\kappa}^{sc}$.

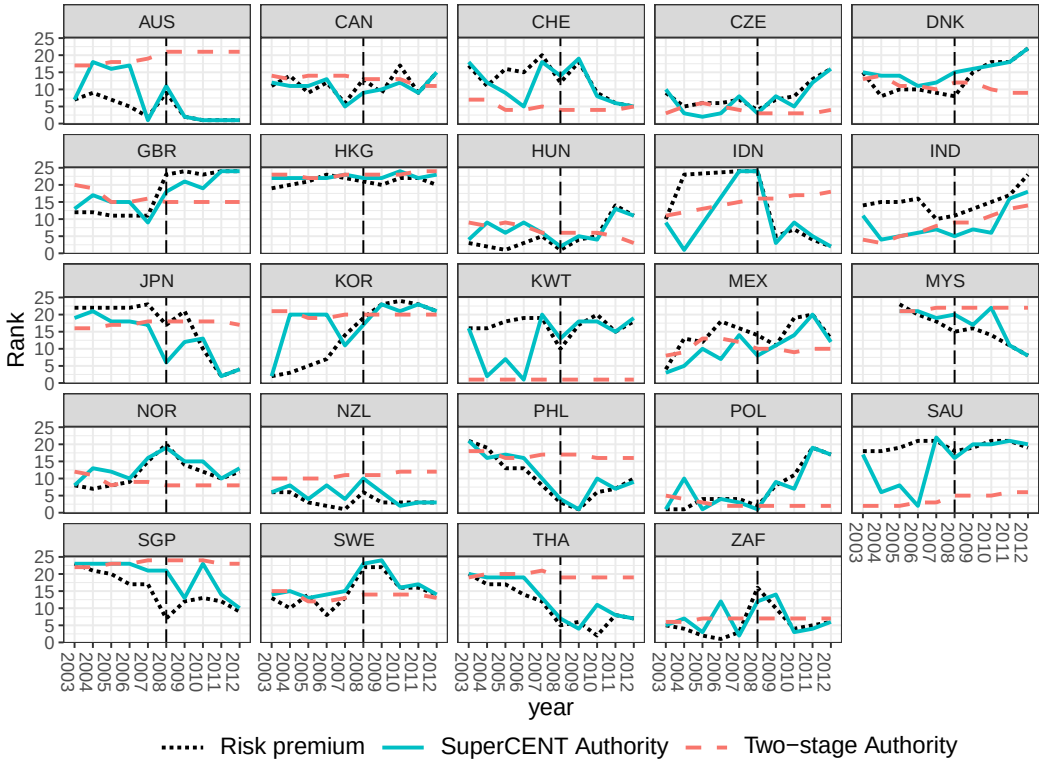


FIGURE 13.—Time series of authority centrality ranking in ascending order from 2003 to 2012. The ranking of risk premium is in descending order, same as Figure 11. Similar to the hub centrality, if the trend of centralities is close to the trend of risk premium, then the centralities capture the time variation of risk premium, based on the negative relationship between the two as claimed in Richmond (2019). The vertical dashed line indicates 2008, the year of the financial crisis.

The three columns of Table III show the coefficient estimation, standard error and significant level for the two-stage-adhoc, two-stage, and SuperCENT, respectively. For the hub centrality β_u, (i) the estimate from two-stage-adhoc and two-stage is -0.0011, while the estimate from SuperCENT is -0.0021, which demonstrates the severe bias of two-stage in the inconsistent regime and the bias is towards zero because $|\hat{\beta}_u^{sc}| = 0.0021 \gg 0.0003 = |\hat{\beta}_v^{sc}|$[7]; (ii) the standard errors from two-stage-adhoc and two-stage are close to 0.0006, much larger than 0.0002 from SuperCENT, which reinforces the problem of overestimation of σ_y^2 in two-stage-adhoc and two-stage; (iii) the above two facts combined make the confidence intervals by two-stage-adhoc and two-stage unnecessarily wide, yet still invalid: consequently the hub centrality β_u is barely significant at level 0.1 using two-stage and is insignificant using two-stage-adhoc; (iv) the two facts in (i) and (ii) also lead to a valid but narrower confidence interval for SuperCENT, making the hub centrality a significant factor at level 0.01 for the currency risk premium; (v) conclusions drawn from the two-stage-adhoc and two-stage methods contradict the theory in Richmond (2019), while SuperCENT supports the theory.

[7]Specifically, $\hat{\beta}_u^{sc} = -0.0021 < -0.0003 \times \frac{0.673}{1+0.154} \approx -0.0002 = \hat{\beta}_v^{sc} \frac{\rho^{sc}}{1+\kappa^{sc}}$, then plim $\hat{\beta}_u^{ts} - \beta_u > 0$ as in Remark 7, and therefore the two-stage estimate is biased towards zero.

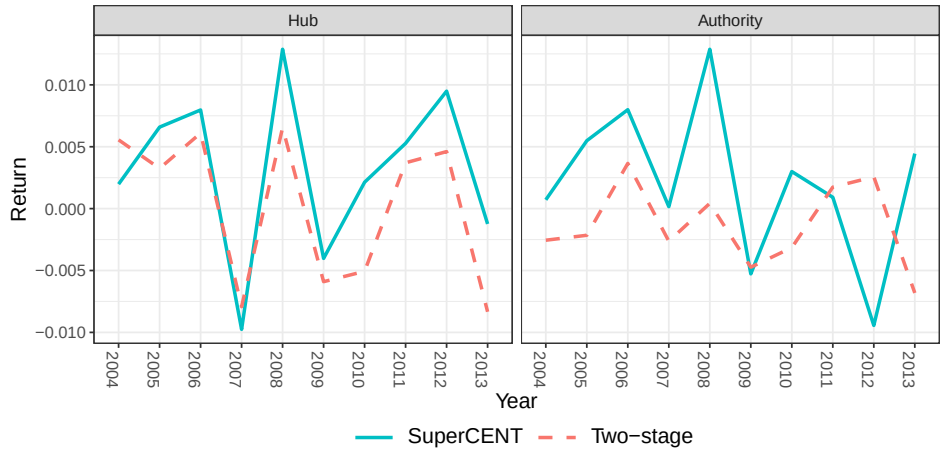


FIGURE 14.—Time series of the next-year return from 2004 to 2013 based on the long-short strategy that takes a long position on the currencies with the lowest 3 centralities and a short position on the currencies with the highest 3 centralities estimated from 2003 to 2012 respectively.

TABLE II

THE 10-YEAR AVERAGE RETURN

	Top/Bottom 3		Top/Bottom 4		Top/Bottom 5	
	Hub	Authority	Hub	Authority	Hub	Authority
SuperCENT	0.0031	0.0021	0.0036	0.0019	0.0033	0.0014
Two-stage	0.0003	-0.0014	0.0008	-0.001	0.001	-0.0006
Relative difference	1 136%	253%	338%	285%	237%	320%

Let us consider other regression coefficients. For β_v, the estimate from two-stage-adhoc and two-stage is -0.0005, while the estimate from SuperCENT is -0.0003, implying a bias away from zero of two-stage in the inconsistent regime due to $|\hat{\beta}_v^{sc}| \ll |\hat{\beta}_u^{sc}|$. SuperCENT still improves its estimation and confidence interval, even though the improvement is not as large as β_u due to $(\hat{\beta}_v^{sc})^2 \ll (\hat{\beta}_u^{sc})^2$ and the nonexchangeable roles of $\boldsymbol{u}$ and $\boldsymbol{v}$. For $\boldsymbol{\beta}_x$, the estimates from two-stage and SuperCENT are comparable, as there is no attenuation bias, but the widths of the confidence intervals from two-stage-adhoc and two-stage are much larger than that from SuperCENT, again as a result of the over-estimation of σ_y^2. Hence, $\boldsymbol{\beta}_x$ is barely significant when using two-stage-adhoc and two-stage, while very significant by SuperCENT.

8. CONCLUSION AND DISCUSSION

Motivated by the rising use of centralities in empirical literature, we highlighted and examined the centrality estimation and inference problems on a noisy network and presented the impact of the estimated centralities on the subsequent network regression. We showed that the commonly used two-stage procedure could yield inaccurate centrality estimates, biased regression coefficient estimates, and invalid inference, especially when the noise-to-signal ratio of the network is large. We proposed SuperCENT which incorporates the network-generation model with the network regression model to simultaneously estimate the centralities and the effects of centralities on the outcome. The additional information from the regression model supervises

TABLE III

THE SUMMARY TABLE OF THE OUTCOME REGRESSION COMPARING THREE METHODS IN TERMS OF COEFFICIENT ESTIMATION, STANDARD ERROR (IN PARENTHESIS) AND THE SIGNIFICANT LEVEL (BY ASTERISKS).

	Dependent variable: Risk premium		
	Two-stage-adhoc	Two-stage	SuperCENT-$\hat{\lambda}_{cv}$
GDP share $\boldsymbol{\beta}_x$	−0.0159*	−0.0159*	−0.0162***
	(0.0083)	(0.0083)	(0.0037)
Hub β_u	−0.0011	−0.0011*	−0.0021***
	(0.0006)	(0.0006)	(0.0002)
Authority β_v	−0.0005	−0.0005	−0.0003
	(0.0006)	(0.0006)	(0.0003)
Note:			*p<0.1; **p<0.05; ***p<0.01
			Year 1999-2008

centrality estimations and thus boosts the accuracy of the estimations. The better centrality estimations, in turn, benefits the network regression. We further derived the convergence rate and distribution of the SuperCENT estimator and provided valid confidence intervals for all the parameters of interest. The theoretical results are corroborated with extensive simulations and a real case study.

The SuperCENT model can be extended in multiple directions. One can consider a generalized linear model for the outcome model and extend SuperCENT to generalized SuperCENT. In the case when only a subset of covariates and outcomes are observed, semi-supervised SuperCENT can be developed. In the case when the network is partially observed, we can perform matrix completion with supervision. SuperCENT can be extended to a longitudinal model with additional assumptions by using techniques from Tensor decomposition as well as functional data analysis to obtain centralities that are smooth over time. In the case of ultra high-dimensional problems, sparsity can be imposed on centralities due to the existence of abundant peripheral nodes.

REFERENCES

Abel, A. B. (2017). Classical measurement error with several regressors. Technical report, Tech Rep. 2017. Tech Rep) Working Paper.

Ahern, K. R. (2013). Network centrality and the cross section of stock returns. *Available at SSRN 2197370*.

Allen, F., Cai, J., Gu, X., Qian, J., Zhao, L., and Zhu, W. (2019). Ownership network and firm growth: What do five million companies tell about chinese economy. *Available at SSRN 3465126*.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144).

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490.

Barbieri, K., Keshk, O. M., and Pollins, B. M. (2009). Trading data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science*, 26(5):471–491.

Battaglini, M., Patacchini, E., and Rainone, E. (2021). Endogenous social interactions with unobserved networks. *The Review of Economic Studies*.

Benzi, M., Estrada, E., and Klymko, C. (2013). Ranking hubs and authorities using matrix functions. *Linear Algebra and its Applications*, 438(5):2447–2474.

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377.

Borgatti, S. P., Carley, K. M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks*, 28(2):124–136.

Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55.

Breza, E. and Chandrasekhar, A. G. (2019). Social networks, reputation, and commitment: evidence from a savings monitors experiment. *Econometrica*, 87(1):175–216.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84.

Butts, C. T. (2003). Network inference, error, and informant (in) accuracy: a bayesian approach. *social networks*, 25(2):103–140.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Chandrasekhar, A. and Lewis, R. (2011). Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*.

Chen, M., Fernández-Val, I., and Weidner, M. (2021). Nonlinear factor models for network and panel data. *Journal of Econometrics*, 220(2):296–324.

De Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications: Eleventh World Congress*, volume 1, pages 268–323. Cambridge University Press Cambridge.

De Paula, Á., Rasul, I., and Souza, P. (2019). Identifying network ties from panel data: theory and an application to tax competition. *arXiv preprint arXiv:1910.07452*.

Elliott, M. and Golub, B. (2019). A network approach to public goods. *Journal of Political Economy*, 127(2):730–776.

Elliott, M., Golub, B., and Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review*, 104(10):3115–53.

Fogli, A. and Veldkamp, L. (2021). Germs, social networks, and growth. *The Review of Economic Studies*, 88(3):1074–1100.

Frantz, T. L., Cataldo, M., and Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328.

Garber, S. and Klepper, S. (1980). Extending the classical normal errors-in-variables model. *Econometrica: Journal of the Econometric Society*, pages 1541–1546.

Glasserman, P. and Young, H. P. (2016). Contagion in financial networks. *Journal of Economic Literature*, 54(3):779–831.

Gofman, M. (2017). Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions. *Journal of Financial Economics*, 124(1):113–146.

Graham, B. and De Paula, A. (2020). *The Econometric Analysis of Network Data*. Academic Press.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063.

Griliches, Z. (1986). Economic data issues. *Handbook of econometrics*, 3:1465–1514.

Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5.

Hochberg, Y. V., Ljungqvist, A., and Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance*, 62(1):251–301.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jackson, M. O., Rogers, B. W., and Zenou, Y. (2017). The economic consequences of social-network structure. *Journal of Economic Literature*, 55(1):49–95.

Jochmans, K. and Weidner, M. (2019). Fixed-effect regressions on network data. *Econometrica*, 87(5):1543–1560.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Kolaczyk, E. D. (2010). Statistical analysis of network data: Methods and models.

Lakhina, A., Byers, J. W., Crovella, M., and Xie, P. (2003). Sampling biases in ip topology measurements. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 1, pages 332–341. IEEE.

Landefeld, J. S., Seskin, E. P., and Fraumeni, B. M. (2008). Taking the pulse of the economy: Measuring gdp. *Journal of Economic Perspectives*, 22(2):193–216.

Le, C. M., Levin, K., and Levina, E. (2018). Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740.

Le, C. M. and Li, T. (2020). Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*.

Lee, L.-F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2):333–374.

Lee, L.-f., Liu, X., and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176.

Li, G., Yang, D., Nobel, A. B., and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 146:7–17.

Li, T., Levina, E., Zhu, J., et al. (2019). Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164.

Lipsey, R. E. (2009). *1. Measuring International Trade in Services*. University of Chicago Press.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics*, 134(4):1883–1948.

Ma, Z., Ma, Z., and Yuan, H. (2020). Universal latent space model fitting for large networks with edge covariates. *J. Mach. Learn. Res.*, 21:4–1.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Martin, C. and Niemeyer, P. (2019). Influence of measurement errors on networks: Estimating the robustness of centrality measures. *Network Science*, 7(2):180–195.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Newman, M. E. (2018). Estimating network structure from unreliable measurements. *Physical Review E*, 98(6):062321.

Newman, M. E. and Clauset, A. (2016). Structure and inference in annotated networks. *Nature communications*, 7(1):1–11.

Pischke, S. (2007). Lecture notes on measurement error. *London School of Economics, London*.

Richmond, R. J. (2019). Trade network centrality and currency risk premia. *The Journal of Finance*, 74(3):1315–1361.

Rohe, K. (2019). A critical threshold for design effects in network sampling. *The Annals of Statistics*, 47(1):556–582.

Shabalin, A. A. and Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Shen, D., Shen, H., and Marron, J. (2016). A general framework for consistency of principal component analysis. *The Journal of Machine Learning Research*, 17(1):5218–5251.

Van Loan, C. F. and Golub, G. (1996). Matrix computations (johns hopkins studies in mathematical sciences).

Vohra, R., Xing, Y., and Zhu, W. (2020). The network effects of agency conflicts. *Available at SSRN 3556298*.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 114(526):857–868.

Yang, D., Ma, Z., and Buja, A. (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. *The Journal of Machine Learning Research*, 17(1):3163–3189.

Zhang, Y., Levina, E., and Zhu, J. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178.

Zhu, W. and Yang, Y. (2020). Networks and business cycles. *Available at SSRN*.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.

SUPPLEMENT TO "NETWORK REGRESSION AND SUPERVISED CENTRALITY ESTIMATION"

JUNHUI CAI
University of Pennsylvania

DAN YANG
The University of Hong Kong

WU ZHU
Tsinghua University

HAIPENG SHEN
The University of Hong Kong

LINDA ZHAO
University of Pennsylvania

This supplementary material contains more details and proofs that are deferred from the main text and is organized as follows. Section S1 provides the derivation of Algorithm 2. Section S2 presents the algorithm of SuperCENT for an undirected network with the eigenvector centrality. Section S3 gives the explicit mathematical expressions of the covariance matrices in Theorems 1 and 2. Section S4 shows more simulation results. Section S5 provides additional information for the case study. Finally, the proofs of the theoretical results are in Section S6.

S1. DERIVATION OF ALGORITHM 2

In the paper, the proposed SuperCENT obtains estimates by optimizing the following objective function which combines the network model and the network regression model in the unified framework (4),

$$(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{\beta}}_x, \hat{\beta}_u, \hat{\beta}_v) := \underset{\substack{\boldsymbol{\beta}_x, \beta_u, \beta_v \\ d, \|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\sqrt{n}}}{\arg\min} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u - \boldsymbol{v}\beta_v\|_2^2 + \frac{\lambda}{n^2}\|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2. \quad \text{(S1)}$$

To solve (S1), we use a block gradient descent algorithm by updating $(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_x^\top, \hat{\beta}_u, \hat{\beta}_v)^\top$. The derivation of each step in each iteration of Algorithm 2 is described below. Denote $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \beta_u, \beta_v)$, and $\mathcal{L} := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u - \boldsymbol{v}\beta_v\|_2^2 + \frac{\lambda}{n^2}\|\boldsymbol{A} - d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2$. Given λ, we minimize the objection function (S1) by setting the partial derivatives of all the parameters as zero. The partial derivatives are as follows.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\frac{2}{n}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}), \quad \text{(S2)}$$

Junhui Cai: junhui@wharton.upenn.edu
Dan Yang: dyanghku@hku.hk
Wu Zhu: zhuwu@sem.tsinghua.edu.cn
Haipeng Shen: haipeng@hku.hk
Linda Zhao: lzhao@wharton.upenn.edu

$$\frac{\partial \mathcal{L}}{\partial d} = 2\lambda d - 2\lambda \boldsymbol{u}^\top \mathbf{A}\boldsymbol{v}/n^2, \tag{S3}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{u}} = -\frac{2}{n}\beta_u(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u - \boldsymbol{v}\beta_v) + \frac{2}{n}\lambda d^2\boldsymbol{u} - \frac{2}{n^2}\lambda d\mathbf{A}\boldsymbol{v}, \tag{S4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}} = -\frac{2}{n}\beta_v(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u - \boldsymbol{v}\beta_v) + \frac{2}{n}\lambda d^2\boldsymbol{v} - \frac{2}{n^2}\lambda d\mathbf{A}^\top\boldsymbol{u}. \tag{S5}$$

Setting the partial derivatives above as zeros yields the estimates

$$\hat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{W}}^\top\widehat{\boldsymbol{W}})^{-1}\widehat{\boldsymbol{W}}^\top\boldsymbol{y}, \tag{S6}$$

$$\hat{d} = \frac{1}{n^2}\hat{\boldsymbol{u}}^\top\boldsymbol{A}\hat{\boldsymbol{v}}, \tag{S7}$$

$$\hat{\boldsymbol{u}} = \left(\hat{\beta}_u^2 + \lambda\hat{d}^2\right)^{-1}\left[\hat{\beta}_u(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_x - \hat{\boldsymbol{v}}\hat{\beta}_v) + \frac{1}{n}\lambda\hat{d}\mathbf{A}\hat{\boldsymbol{v}}\right], \tag{S8}$$

$$\hat{\boldsymbol{v}} = \left(\hat{\beta}_v^2 + \lambda\hat{d}^2\right)^{-1}\left[\hat{\beta}_v(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_x - \hat{\boldsymbol{u}}\hat{\beta}_u) + \frac{1}{n}\lambda\hat{d}\mathbf{A}^\top\hat{\boldsymbol{u}}\right], \tag{S9}$$

with constraints

$$\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}} = n \quad \text{and} \quad \hat{\boldsymbol{v}}^\top\hat{\boldsymbol{v}} = n \tag{S10}$$

where $\widehat{\boldsymbol{W}} = (\boldsymbol{X}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$. Denote $(\hat{d}^{(t)}, \hat{\boldsymbol{u}}^{(t)}, \hat{\boldsymbol{v}}^{(t)}, \widehat{\boldsymbol{\beta}}^{(t)})$ as the estimations from the t-th iteration. Combining (S8)-(S10) and substituting the corresponding estimates from the previous updates, we obtain each update step in each iteration.

S2. SUPERCENT FOR AN UNDIRECTED NETWORK

When the network is undirected with the eigenvector centrality, it can be represented by a symmetric matrix $\boldsymbol{A}$. Denote $\boldsymbol{u}$ as the eigenvector centrality. The objective function of SuperCENT estimation is a special case of (S1), i.e.,

$$(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}_x, \hat{\beta}_u) := \underset{\substack{\boldsymbol{\beta}_x, \beta_u \\ d, \|\boldsymbol{u}\|_2 = \sqrt{n}}}{\arg\min} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u\|_2^2 + \frac{\lambda}{n^2}\|\mathbf{A} - d\boldsymbol{u}\boldsymbol{u}^\top\|_F^2. \tag{S11}$$

To solve (S11), we adopt a similar strategy as Algorithm 2 – a block gradient descent algorithm by updating $(\hat{d}, \hat{\boldsymbol{u}}, \widehat{\boldsymbol{\beta}})$ iteratively until convergence, where $\widehat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_x^\top, \hat{\beta}_u)^\top$. The initialization can be obtained from the eigen decomposition of $\boldsymbol{A}$. Given a tuning parameter λ,

Algorithm S1 describes the algorithm for a symmetric matrix $\boldsymbol{A}$. Similarly, the tuning parameters λ can be chosen using cross-validation and others as described in Section 4.4.

Algorithm S1: SuperCENT$(\mathbf{A}, \mathbf{X}, \mathbf{y}, \lambda)$ to solve (S11) for a symmetric $\boldsymbol{A}$.

Result: $\hat{d}$, $\hat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{\beta}}$
Input: the observed network $\mathbf{A} \in \mathbb{R}^{n\times n}$, the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, the response vector $\mathbf{y} \in \mathbb{R}^n$, the tuning penalty parameter λ, the tolerance parameter $\rho > 0$, the maximum number of iteration T ;
Initiate $(d^{(0)}, \boldsymbol{u}^{(0)}) = \arg\min_{d,\|\boldsymbol{u}\|_2=\sqrt{n}} \|A - d\boldsymbol{u}\boldsymbol{u}^\top\|_F^2$, $t = 1$;
while $\|\boldsymbol{P}_{\boldsymbol{u}^{(t)}} - \boldsymbol{P}_{\boldsymbol{u}^{(t-1)}}\|_2 > \rho$ *and* $t < T$ **do**

1. $\mathbf{W}^{(t-1)} = (\mathbf{X}, \boldsymbol{u}^{(t-1)})$
2. $\boldsymbol{\beta}^{(t)} = (\mathbf{W}^{(t-1)\top}\mathbf{W}^{(t-1)})^{-1}\mathbf{W}^{(t-1)\top}\mathbf{Y}$
3. $d^{(t)} = \boldsymbol{u}^{(t-1)\top}\mathbf{A}\boldsymbol{u}^{(t-1)}/n^2$
4. $\boldsymbol{u}^{(t)} = \left((\beta_u^{(t)})^2\boldsymbol{I} + \lambda(d^{(t)})^2\boldsymbol{I} - 2\lambda d^{(t)}\boldsymbol{A}/n\right)^{-1}\beta_u^{(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x^{(t)})$
5. Normalize $\boldsymbol{u}^{(t)}$ such that $\|\boldsymbol{u}^{(t)}\|_2 = \sqrt{n}$
6. $t \leftarrow t + 1$

end

The derivation of Algorithm 1 is similar to Section S1. Denote $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{u})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \beta_u)$, and $\mathcal{L}_{sym} := \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u\|_2^2 + \frac{\lambda}{n^2}\|\mathbf{A} - d\boldsymbol{u}\boldsymbol{u}^\top\|_F^2$ where the subscript sym denotes the objective function for a symmetric matrix $\boldsymbol{A}$. Given λ, we minimize the objection function (S1) by setting the partial derivatives with all the parameters as zero. The partial derivatives are as follows.

$$\frac{\partial \mathcal{L}_{sym}}{\partial \boldsymbol{\beta}} = -\frac{2}{n}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}), \tag{S12}$$

$$\frac{\partial \mathcal{L}_{sym}}{\partial d} = 2\lambda d - 2\lambda \boldsymbol{u}^\top\mathbf{A}\boldsymbol{u}/n^2, \tag{S13}$$

$$\frac{\partial \mathcal{L}_{sym}}{\partial \boldsymbol{u}} = -\frac{2}{n}\beta_u(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{u}\beta_u) + \frac{2}{n}\lambda d^2\boldsymbol{u} - \frac{4}{n^2}\lambda d\mathbf{A}\boldsymbol{u}. \tag{S14}$$

Setting the partial derivatives above as zero yields the estimates

$$\hat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{W}}^\top\widehat{\boldsymbol{W}})^{-1}\widehat{\boldsymbol{W}}^\top\boldsymbol{y}, \tag{S15}$$

$$\hat{d} = \frac{1}{n^2}\hat{\boldsymbol{u}}^\top\boldsymbol{A}\hat{\boldsymbol{u}}, \tag{S16}$$

$$\hat{\boldsymbol{u}} = \left(\hat{\beta}_u^2\boldsymbol{I} + \lambda\hat{d}^2\boldsymbol{I} - \frac{2}{n}\lambda\hat{d}\mathbf{A}\right)^{-1}\hat{\beta}_u(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_x), \tag{S17}$$

with a constraint

$$\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}} = n, \tag{S18}$$

where $\widehat{\boldsymbol{W}} = (\boldsymbol{X}, \hat{\boldsymbol{u}})$. Similarly, denote $(\hat{d}^{(t)}, \hat{\boldsymbol{u}}^{(t)}, \widehat{\boldsymbol{\beta}}^{(t)})$ as the estimations from the t-th iteration. Taking together (S15)-(S18) and substituting corresponding estimates from the previous update, we obtain each update step in each iteration.

S3. EXPLICIT MATHEMATICAL EXPRESSIONS OF COVARIANCES IN THEOREMS 1 AND 2

Let $\boldsymbol{K}$ be the $n^2 \times n^2$ commutation matrix such that $\text{vec}(\boldsymbol{E}^\top) = \boldsymbol{K}\,\text{vec}(\boldsymbol{E})$ and $\otimes$ denote the Kronecker product. Recall $\tilde{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}$, $\tilde{\boldsymbol{v}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}$ and $\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} = \begin{pmatrix} \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} & \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} \\ \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} & \tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} \end{pmatrix}$.

S3.1. *Specific form of $\boldsymbol{\Sigma}^{ts}$ and $\boldsymbol{C}^{ts}$ in Theorem 1*

The asymptotic variance of the two-stage estimator in (11) $\boldsymbol{\Sigma}^{ts} = \boldsymbol{C}^{ts}\begin{pmatrix} \sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2} \end{pmatrix}\boldsymbol{C}^{ts\top}$.

Denote $\boldsymbol{C}^{ts} = \begin{pmatrix} \boldsymbol{0}_{n\times n} & \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{0}_{n\times n} & \boldsymbol{C}_{22}^{ts} \\ \boldsymbol{0}_{n^2\times n} & \boldsymbol{C}_{32}^{ts} \\ \boldsymbol{C}_{41}^{ts} & \boldsymbol{C}_{42}^{ts} \\ \boldsymbol{C}_{51}^{ts} & \boldsymbol{C}_{52}^{ts} \\ \boldsymbol{C}_{61}^{ts} & \boldsymbol{C}_{62}^{ts} \end{pmatrix}$ whose specific forms are as follows. The matrices related to $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are

$$\begin{pmatrix} \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{C}_{22}^{ts} \end{pmatrix} = (dn)^{-1}\begin{pmatrix} \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \\ (\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}))\,\boldsymbol{K} \end{pmatrix}, \tag{S19}$$

the matrix related to $\widehat{\boldsymbol{A}}^{ts}$ is

$$\boldsymbol{C}_{32}^{ts} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \otimes \boldsymbol{P}_{\boldsymbol{u}} + \boldsymbol{P}_{\boldsymbol{v}} \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \boldsymbol{P}_{\boldsymbol{v}} \otimes \boldsymbol{P}_{\boldsymbol{u}}, \tag{S20}$$

the matrices related to $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$ are

$$\begin{pmatrix} \boldsymbol{C}_{41}^{ts} & \boldsymbol{C}_{42}^{ts} \\ \boldsymbol{C}_{51}^{ts} & \boldsymbol{C}_{52}^{ts} \end{pmatrix} = \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix}(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; \boldsymbol{I}_n)\begin{pmatrix} \boldsymbol{C}_{11}^{ts} & \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{C}_{21}^{ts} & \boldsymbol{C}_{22}^{ts} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \end{pmatrix}, \tag{S21}$$

and the matrices related to $\hat{\boldsymbol{\beta}}_x^{ts}$ are

$$(\,\boldsymbol{C}_{61}^{ts}\;\boldsymbol{C}_{62}^{ts}\,) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; -\boldsymbol{u} \;\; -\boldsymbol{v} \;\; \boldsymbol{I}_n)\begin{pmatrix} \boldsymbol{C}_{11}^{ts} & \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{C}_{21}^{ts} & \boldsymbol{C}_{22}^{ts} \\ \boldsymbol{C}_{41}^{ts} & \boldsymbol{C}_{42}^{ts} \\ \boldsymbol{C}_{51}^{ts} & \boldsymbol{C}_{52}^{ts} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \end{pmatrix}. \tag{S22}$$

S3.2. *Specific form of $\boldsymbol{\Sigma}$ and $\boldsymbol{C}$ in Theorem 2*

The asymptotic variance of the SuperCENT estimator in (25) $\boldsymbol{\Sigma} = \boldsymbol{C}\begin{pmatrix} \sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2} \end{pmatrix}\boldsymbol{C}^\top$.

Denote $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11}\, \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\, \boldsymbol{C}_{22} \\ \boldsymbol{C}_{31}\, \boldsymbol{C}_{32} \\ \boldsymbol{C}_{41}\, \boldsymbol{C}_{42} \\ \boldsymbol{C}_{51}\, \boldsymbol{C}_{52} \\ \boldsymbol{C}_{61}\, \boldsymbol{C}_{62} \end{pmatrix}$ whose specific forms are given as follows. The matrices related to $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are

$$\begin{pmatrix} \boldsymbol{C}_{11}\, \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\, \boldsymbol{C}_{22} \end{pmatrix} = (\lambda d^2)^{-1} \left(\boldsymbol{I}_{2n} - (\lambda d^2 + {\beta_u}^2 + {\beta_v}^2)^{-1} \begin{pmatrix} {\beta_u}^2 & \beta_u\beta_v \\ \beta_u\beta_v & {\beta_v}^2 \end{pmatrix} \otimes \Big(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \Big) \right) \begin{pmatrix} \beta_u \Big(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \Big) & \lambda d n^{-1} \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \\ \beta_v \Big(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \Big) & \lambda d n^{-1} \left(\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \right) \boldsymbol{K} \end{pmatrix} \tag{S23}$$

where $\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}$ is the projection matrix that projects onto the column space of $(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v})$, the matrices related to $\widehat{\boldsymbol{A}}$ are

$$(\, \boldsymbol{C}_{31}\, \boldsymbol{C}_{32}\,) = \Big(d(\boldsymbol{v} \otimes \boldsymbol{I}_n) \;\; d\boldsymbol{K}(\boldsymbol{u} \otimes \boldsymbol{I}_n) \;\; \boldsymbol{P}_{\boldsymbol{v}} \otimes \boldsymbol{P}_{\boldsymbol{u}} \Big) \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{0}_{n^2 \times n} & \boldsymbol{I}_{n^2} \end{pmatrix}, \tag{S24}$$

the matrices related to $\hat{\beta}_u$ and $\hat{\beta}_v$ are

$$\begin{pmatrix} \boldsymbol{C}_{41}\, \boldsymbol{C}_{42} \\ \boldsymbol{C}_{51}\, \boldsymbol{C}_{52} \end{pmatrix} = \boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} (-\beta_u \boldsymbol{I}_n \;\; -\beta_v \boldsymbol{I}_n \;\; \boldsymbol{I}_n) \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n \times n^2} \end{pmatrix}, \tag{S25}$$

and the matrices related to $\hat{\boldsymbol{\beta}}_x$ are

$$(\, \boldsymbol{C}_{61}\, \boldsymbol{C}_{62}\,) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (-\beta_u \boldsymbol{I}_n \;\; -\beta_v \boldsymbol{I}_n \;\; -\boldsymbol{u} \;\; -\boldsymbol{v} \;\; \boldsymbol{I}_n) \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{C}_{41} & \boldsymbol{C}_{42} \\ \boldsymbol{C}_{51} & \boldsymbol{C}_{52} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n \times n^2} \end{pmatrix}. \tag{S26}$$

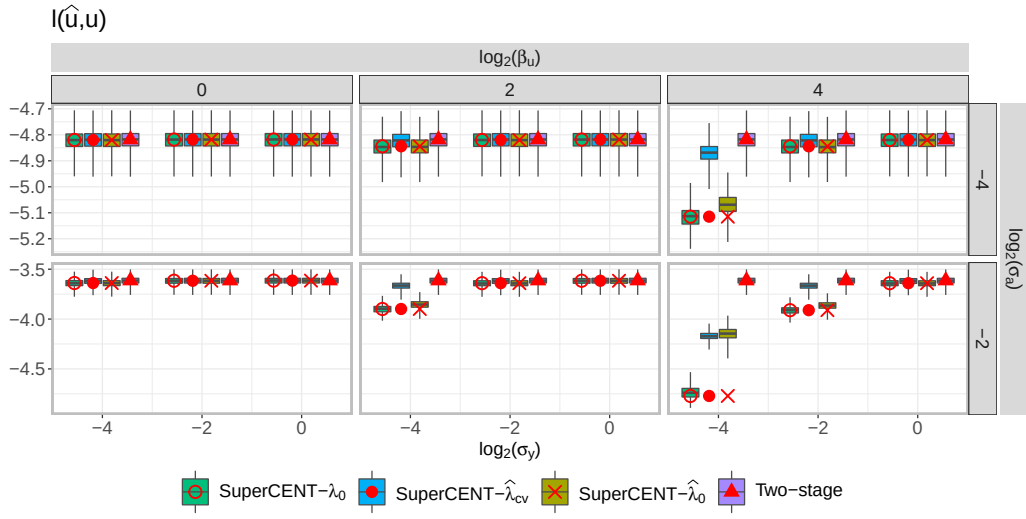


FIGURE S1.—Consistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\boldsymbol{u}}, \boldsymbol{u}))$ for the four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\hat{\boldsymbol{u}}, \boldsymbol{u}) = \|\boldsymbol{P}_{\hat{\boldsymbol{u}}} - \boldsymbol{P}_{\boldsymbol{u}}\|_2^2$. The super-imposed red symbols show the theoretical rates of the two-stage in Corollary 1 and SuperCENT in Corollary 4.

S4. MORE ON SIMULATIONS

In this section, we show more simulation results that are deferred from Section 6. Section S4.1 is for the consistent regime of two-stage and Section S4.2 is for the inconsistent regime of two-stage.

S4.1. *Consistent regime of two-stage*

We present the consistent regime of the two-stage procedure, i.e., when the network noise-to-signal ratio $\kappa = \frac{\sigma_a^2}{d^2 n} \to 0$. For the consistency of the two-stage procedure, we keep $\kappa < 2^{-12}$ (recall $d = 1$ and $n = 2^8$) by varying $\sigma_a \in 2^{-4,-2}$. The configuration is exactly the same as the consistent regime except for σ_a. We expect that SuperCENT improves over the two-stage in terms of both estimation and inference.

Figure S1 shows the boxplot of the logarithm of $l(\hat{\boldsymbol{u}}, \boldsymbol{u})$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. The panel structure remains the same: the rows correspond to $\log_2(\sigma_a)$ and the columns correspond to $\log_2(\beta_u)$; for each panel, the x-axis is $\log_2(\sigma_y)$ and the y-axis is $\log_{10}(l(\hat{\boldsymbol{u}}, \boldsymbol{u}))$. The super-imposed red symbols show the theoretical rates of $\hat{\boldsymbol{u}}^{ts}$ in Corollary 1 and that of $\hat{\boldsymbol{u}}$ in Corollary 4. Some messages are consistent between Figure S1 and Figure 3: i) SuperCENT-λ_0 performs the best, Two-stage is always the worst, SuperCENT-$\hat{\lambda}_{cv}$ and SuperCENT-$\hat{\lambda}_0$ are in-between; ii) The two-stage estimator performs the same no matter how large σ_y and β_u are, and it has smaller error with smaller σ_a; ii) The improvement of SuperCENT over two-stage is stronger with larger β_u, σ_a and smaller σ_y, which aligns with Remark 10. For the comparison of the three SuperCENT-based methods, the benchmark SuperCENT-λ_0 is always the best, SuperCENT-$\hat{\lambda}_{cv}$ and SuperCENT-$\hat{\lambda}_0$ are sometimes worse than SuperCENT-λ_0, but still better than two-stage.

Figure S2 shows the boxplot of logarithm of $l(\hat{\boldsymbol{v}}, \boldsymbol{v})$ across different σ_a, σ_y and β_u. Other than a noticeable improvement when $\beta_u \in 2^{0,2}$ and $\sigma_y = 2^{-4}$, the improvement of SuperCENT over two-stage is negligible in other settings when $\beta_u > \beta_v$ and the network signal-to-noise

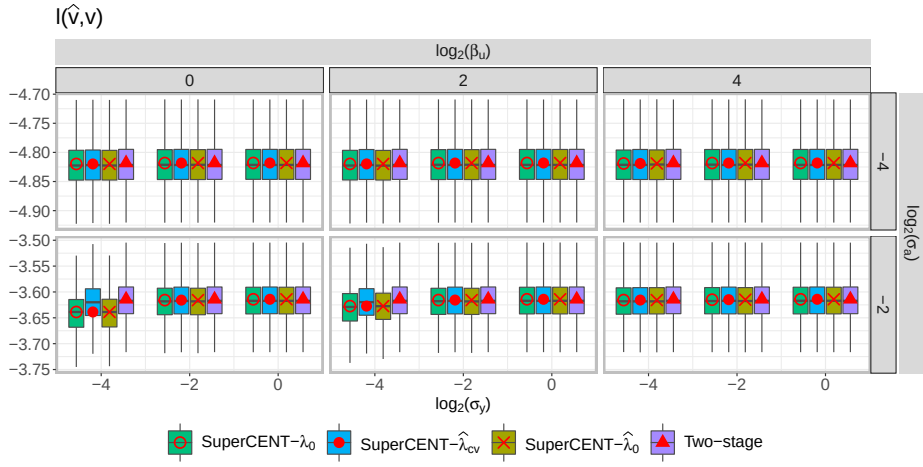


FIGURE S2.—Consistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\boldsymbol{v}}, \boldsymbol{v}))$ for four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\hat{\boldsymbol{v}}, \boldsymbol{v}) = \|\boldsymbol{P}_{\hat{\boldsymbol{v}}} - \boldsymbol{P}_{\boldsymbol{v}}\|_2^2$. The super-imposed red points show the theoretical rates of the two-stage and SuperCENT algorithms.

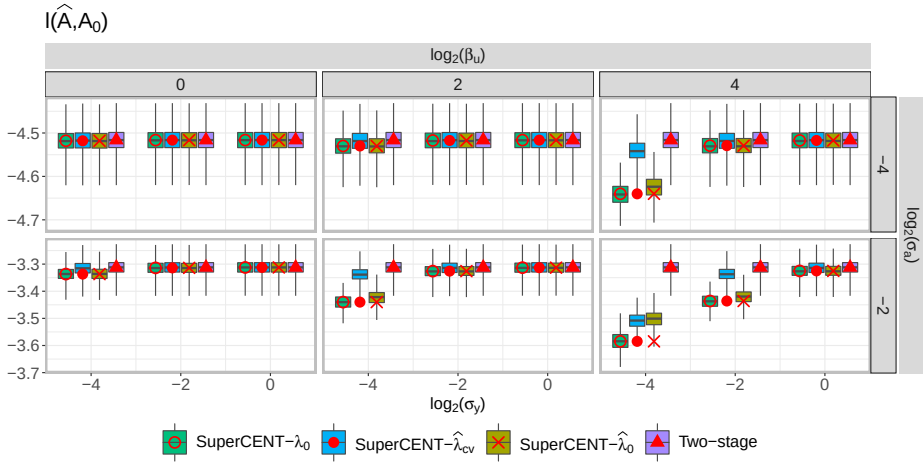


FIGURE S3.—Consistent regime of two-stage: The boxplot of $\log_{10}(l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0))$ for four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0) = \|\widehat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F^2 / \|\boldsymbol{A}_0\|_F^2$. The super-imposed red points show the theoretical rates of the two-stage and SuperCENT algorithms.

ratio is relatively large. Comparing Figures S1 and S2, the supervised effect to $\hat{\boldsymbol{u}}$ increases as β_u increases while the supervised effect to $\hat{\boldsymbol{v}}$ shrinks as β_u increases. This phenomena aligns with the nonexchangeable roles $\boldsymbol{u}$ and $\boldsymbol{v}$ as in Remark 13 – noted that here we have $\beta_v \leqslant \beta_u$ by fixing $\beta_v = 1$ and varying $\beta_u \in 2^{0,2,4}$. Similarly, when $\beta_v \gg \beta_u$ we should expect the improvement in estimating $\boldsymbol{v}$ to increase.

Figure S3 shows the boxplot of logarithm of $l(\widehat{\boldsymbol{A}}, \boldsymbol{A}_0)$ across different σ_a, σ_y and β_u. With the improvement in estimating $\boldsymbol{u}$ and $\boldsymbol{v}$ by SuperCENT as shown in Figures S1 and S2, it is not surprising to see the improvement in estimating $\boldsymbol{A}_0$. The trend of improvement is similar to that of $\boldsymbol{u}$ in Figure S1. Comparing Figures S1,S2 and S3 altogether, when $\beta_u = 2^0$ and $\sigma_y = 2^{-4}$,

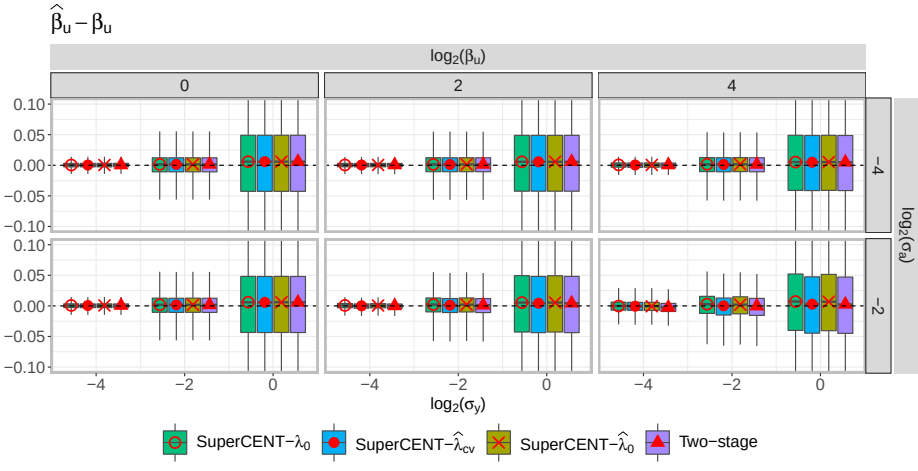


FIGURE S4.—Consistent regime of two-stage: The boxplot of the bias $\hat{\beta}_u - \beta_u$ across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$. The dashed lines show $\hat{\beta}_u - \beta_u = 0$. The super-imposed red points show the median of the bias.

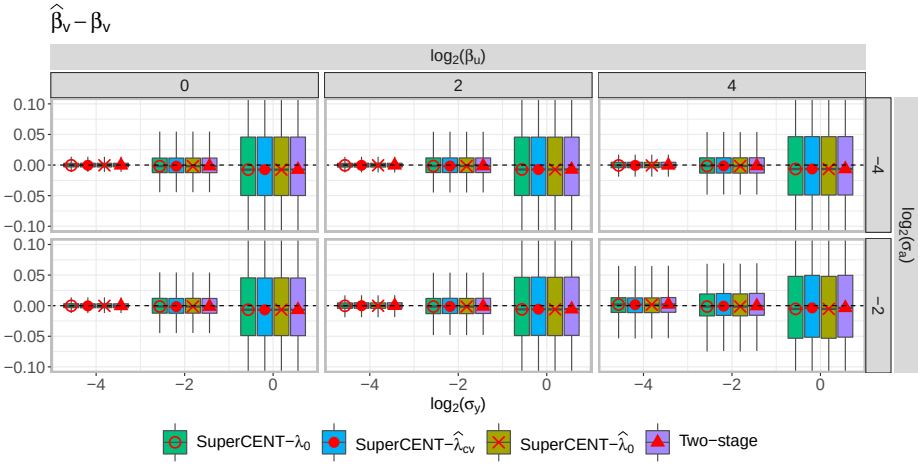


FIGURE S5.—Consistent regime of two-stage: The boxplot of the bias $\hat{\beta}_v - \beta_v$ across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$. The dashed lines show $\hat{\beta}_v - \beta_v = 0$. The super-imposed red points show the median of the bias.

SuperCENT improves both the estimation of $\boldsymbol{u}$ and $\boldsymbol{v}$; when $\beta_u \in 2^{2,4}$, SuperCENT improves the estimation of $\boldsymbol{u}$ a lot, and therefore, SuperCENT improves the estimation of $\boldsymbol{A}_0$ for most ranges of β_u, especially when σ_y is small. The comparison of the three SuperCENT-based methods is similar, with the benchmark SuperCENT-λ_0 always being the best, SuperCENT-$\hat{\lambda}_{cv}$ and SuperCENT-$\hat{\lambda}_0$ being worse than SuperCENT-λ_0 sometimes but still not worse than two-stage.

Figures S4 and S5 show the boxplots of the bias in estimating β_u and β_v, $\hat{\beta}_u - \beta_u$ and $\hat{\beta}_v - \beta_v$, respectively across different σ_a, σ_y and β_u. The performance of all methods are

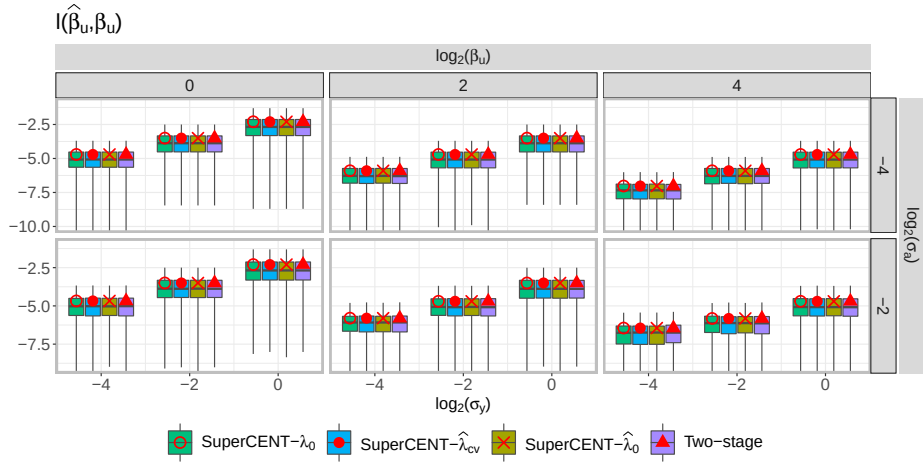


FIGURE S6.—Consistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\beta}_u, \beta_u))$ for four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\hat{\beta}_u, \beta_u) = (\hat{\beta}_u - \beta_u)^2/\beta_u^2$. The super-imposed red points show the theoretical rates of the two-stage in Corollary 2 and SuperCENT algorithms in Corollary 5.

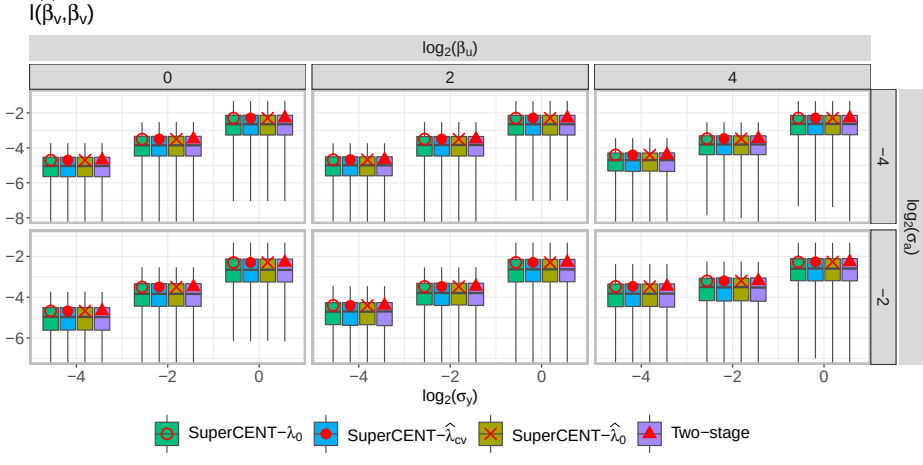


FIGURE S7.—Consistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\beta}_v, \beta_v))$ for four estimators across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$ where $l(\hat{\beta}_v, \beta_v) = (\hat{\beta}_v - \beta_v)^2/\beta_v^2$. The super-imposed red points show the theoretical rates of the two-stage in Corollary 2 and SuperCENT algorithms in Corollary 5.

similar with negligible bias. This is aligned with Remark 7 where we claim that there exists no asymptotic bias when two-stage is consistent.

Figures S6 and S7 show the boxplots of $l(\hat{\beta}_u, \beta_u)$ and $l(\hat{\beta}_v, \beta_v)$ respectively across different σ_a, σ_y and β_u. The performance of all methods are similar in estimating β_u and β_v. This echoes with Corollary 5 where we prove that despite better estimations of $\boldsymbol{u}, \boldsymbol{v}$ by SuperCENT, the rates for estimating β_u and β_v are surprisingly the same for SuperCENT and two-stage.

We now turn our attention to the inference property. Figures S8 and S9 show the empirical coverage and $\log_{10}$ of the average width, respectively, of the 95% confidence interval for β_u from the five methods. In terms of the empirical coverage, both SuperCENT-λ_0-oracle and

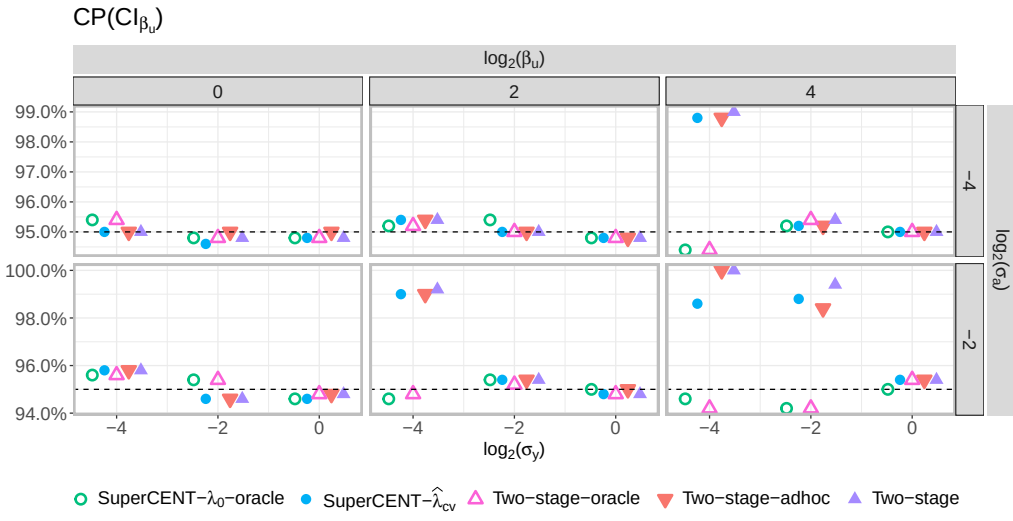


FIGURE S8.—Consistent regime of two-stage: The empirical coverage of CI_{β_u} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles. The dashed lines show the nominal confidence level 0.95.

two-stage-oracle are very close to the nominal level across different settings. Their widths are the same because the same rate of $\hat{\beta}_u^{ts}$ and $\hat{\beta}_u$ are shown in Corollary 5 and the true parameters are plugged in. For the non-oracle methods, the two two-stage related methods and SuperCENT-$\hat{\lambda}_{cv}$ are either close to nominal or tend to be conservative. When all of them are conservative, SuperCENT-$\hat{\lambda}_{cv}$ always has the smallest width among the three. Again, the relative narrower width of SuperCENT-$\hat{\lambda}_{cv}$ is more obvious with large σ_a, β_u or small σ_y. Furthermore, two-stage tends to have wider CI than two-stage-adhoc and is more conservative, which supports Remark 5.

Figures S10 and S11 show the empirical coverage and $\log_{10}$ of the average width, respectively, of 95% confidence interval for β_v from five methods. In terms of empirical coverage, both SuperCENT-λ_0-oracle and two-stage-oracle are very close to the nominal level across different settings. Their widths are the same due to the same rate of $\hat{\beta}_v^{ts}$ and $\hat{\beta}_v$ in Corollary 5 with the true parameters plugged in. For the non-oracle methods, two-stage and SuperCENT-$\hat{\lambda}_{cv}$ are mostly close to the nominal level although tend to be conservative when $\beta_u = 2^4$ and $\sigma_y = 2^{-4}$; while two-stage-adhoc is always close to the nominal. It is worth noting that as β_u increases, the first term of (18) dominates (17) because $\kappa = \frac{\sigma_a^2}{d^2 n} = O(1)$ and $\beta_u^2 \gg \sigma_y^2$. Therefore, even with σ_y^2 being overestimated, two-stage-adhoc is on par with two-stage-oracle because two-stage-adhoc leaves out (18). This is different from the behavior of CI_{β_u} where two-stage-adhoc is more conservative than the two-stage-oracle. The phenomena is more pronounced when σ_a increases (bottom right two panels). In terms of the width, when two-stage and SuperCENT-$\hat{\lambda}_{cv}$ are conservative, their widths are always wider than two-stage-adhoc with two-stage being the widest. Furthermore, the fact of two-stage having wider CI and being more conservative than two-stage-adhoc supports Remark 5.

Finally, we investigate the average coverage and the average width of confidence intervals for all the entries a_{ij} of $\boldsymbol{A}_0$ respectively. Figure S12 shows that the average coverage probability of all the methods, $\text{Average}_{ij}\big(\text{CP}(\text{CI}a_{ij})\big)$, achieves the nominal level of 95%. Figure S13 shows

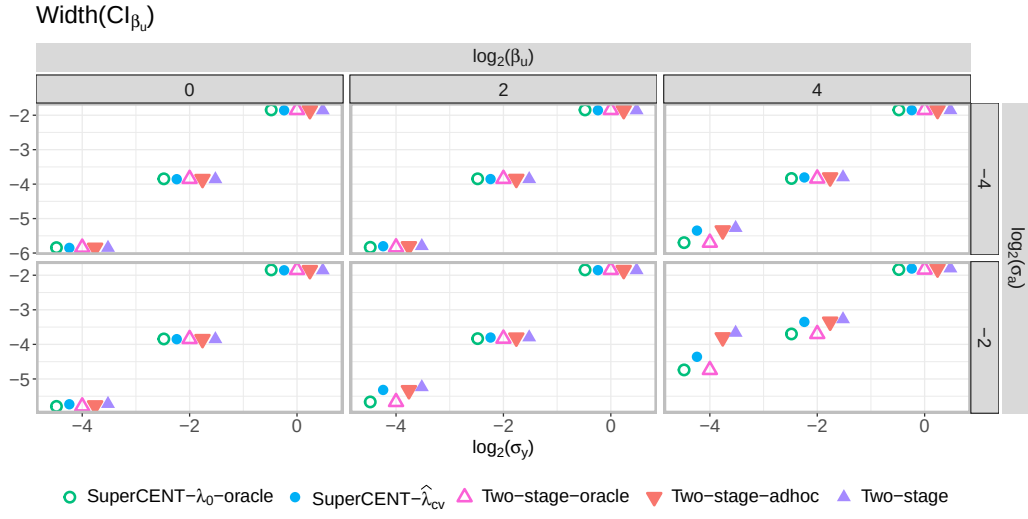


FIGURE S9.—Consistent regime of two-stage: $\log_{10}$ of the width of CI_{β_u} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles.

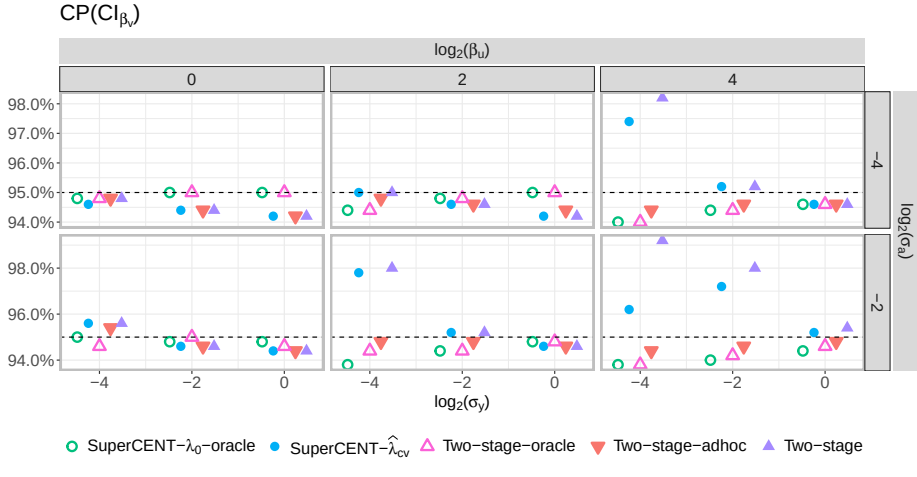


FIGURE S10.—Consistent regime of two-stage: Empirical coverage of β_v across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles. The dashed lines show the nominal confidence level 0.95.

the $\log_{10}$ of the average width of the CIs, $\text{Average}_{ij}\big(\text{Width}(\text{CI}a_{ij})\big)$. SuperCENT-$\lambda_0$-oracle provides the shortest width among the four methods, followed by SuperCENT-$\hat{\lambda}_{cv}$, which in turn dominates the two two-stage related methods. Again, the improvement of SuperCENT over two-stage increases as σ_a and β_u increase or σ_y decreases.

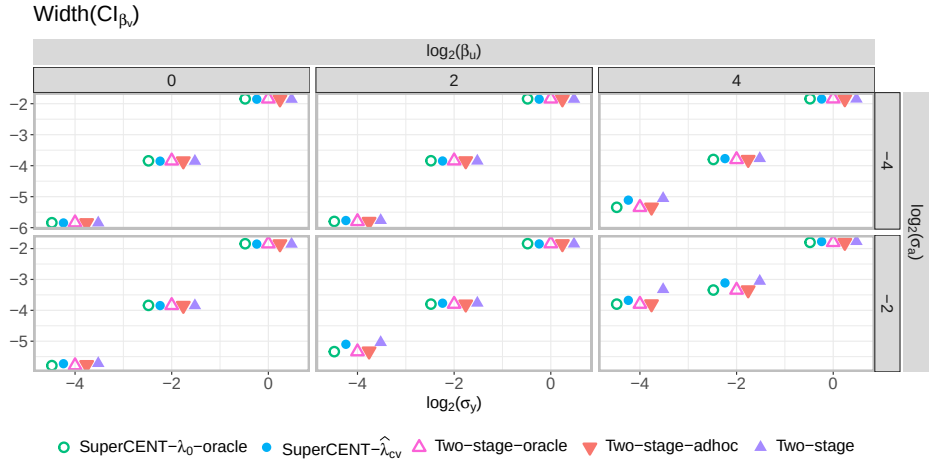


FIGURE S11.—Consistent regime of two-stage: Width of CI_{β_v} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles.

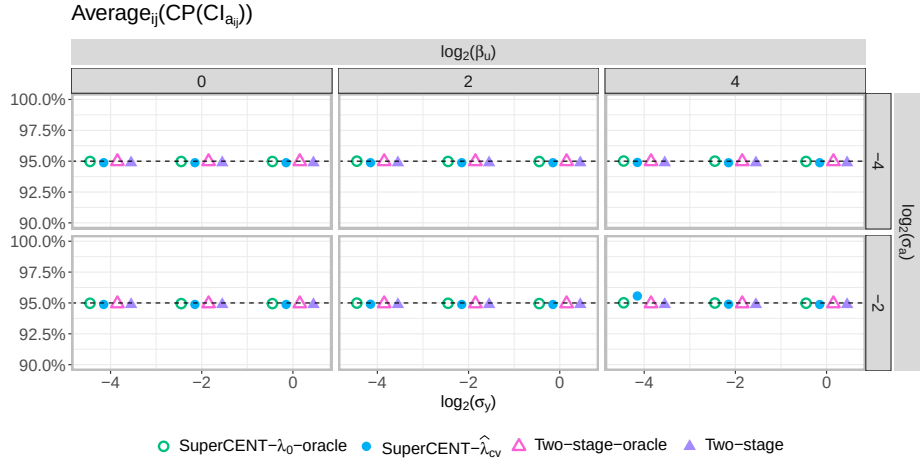


FIGURE S12.—Consistent regime of two-stage: The average empirical coverage of $CI_{a_{ij}}$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▲). The dashed lines show the nominal confidence level 0.95.

S4.2. *Inconsistent regime of two-stage*

In the inconsistent regime of two-stage procedure, $\kappa = \frac{\sigma_a^2}{d^2 n} = O(1)$, $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ are inconsistent. Section 6.2 shows that SuperCENT enormously improves over two-stage in terms of estimation of $\boldsymbol{u}$, $\boldsymbol{v}$ and β_u as well as the inference of β_u and $\boldsymbol{A}_0$. In this section, we demonstrate the behaviors of the SuperCENT-based and the two-stage-based estimators of β_v and their corresponding confidence intervals. We also show the plot of the average coverage of the confidence interval of $\boldsymbol{A}_0$, which is deferred from the main text.

Figure S14 shows the bias in estimating β_v, i.e., $\hat{\beta}_v - \beta_v$. We observe a bias away from zero in estimating β_v when β_u is large, which is different from the attenuation bias in $\hat{\beta}_u$. As in

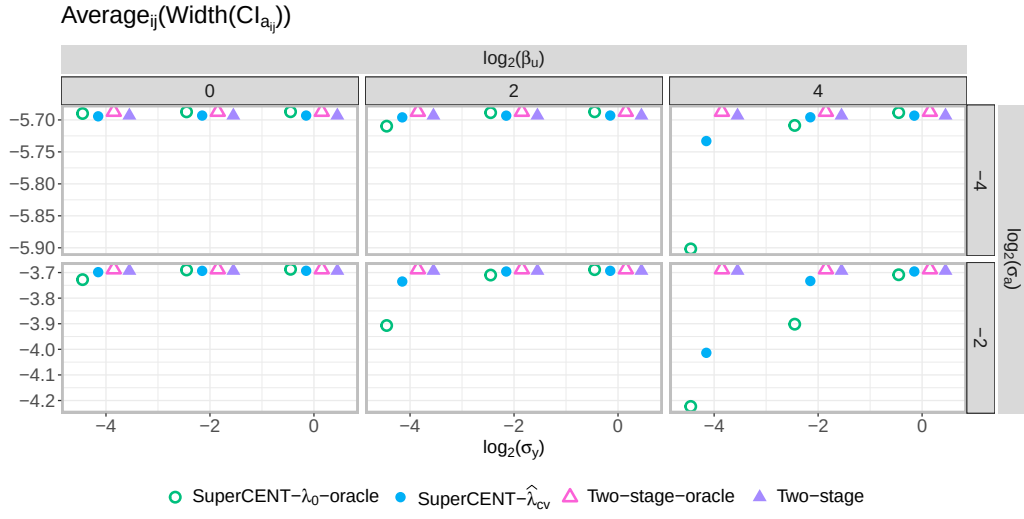


FIGURE S13.—Consistent regime of two-stage: $\log_{10}$ of the average width of $CI_{a_{ij}}$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▲).

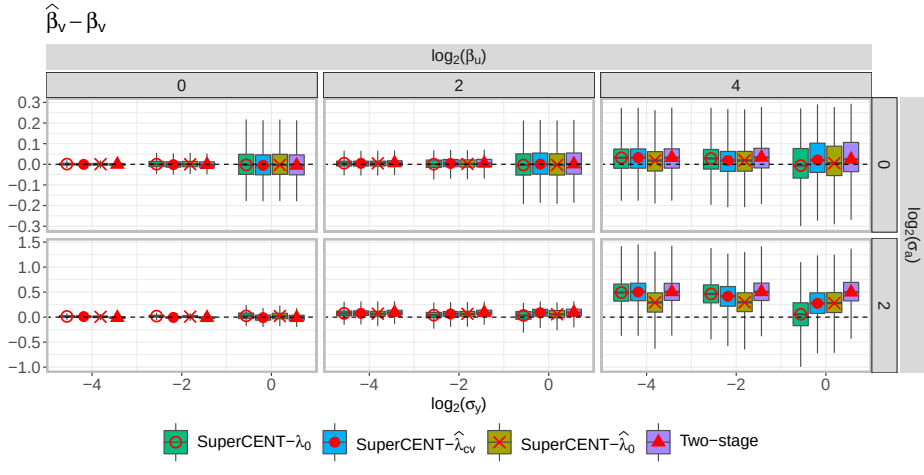


FIGURE S14.—Inconsistent regime of two-stage: The boxplot of the bias $\hat{\beta}_v - \beta_v$ across different σ_a, σ_y and β_u with fixed $d = 1$ and $\beta_v = 1$. The dashed lines show $\hat{\beta}_v - \beta_v = 0$. The super-imposed red points show the median of the bias.

Remark 7, the bias is away from zero because $\kappa = 2^{-8,-4} \nrightarrow 0$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are correlated with $\rho = \frac{2^{-1}}{\sqrt{1.25}}$, and $\beta_v = 1 = 2^0 < \{2^2 \cdot 2^{-1.2}, 2^4 \cdot 2^{-1.2}\} = \beta_u \frac{\rho}{(1+\kappa)}$, leading to $\operatorname{plim} \hat{\beta}_v^{ts} - \beta_v > 0$. In addition, the bias is larger as β_u increases. SuperCENT in this case can still alleviate the bias but the improvement is not as large as that of β_u since the improvement in estimation of $\boldsymbol{v}$ when β_u is large is limited as shown in Figure 4, thereby slightly improving over the two-stage from the perspective of the estimation bias in Figure S14 as well as the squared error loss in Figure S15.

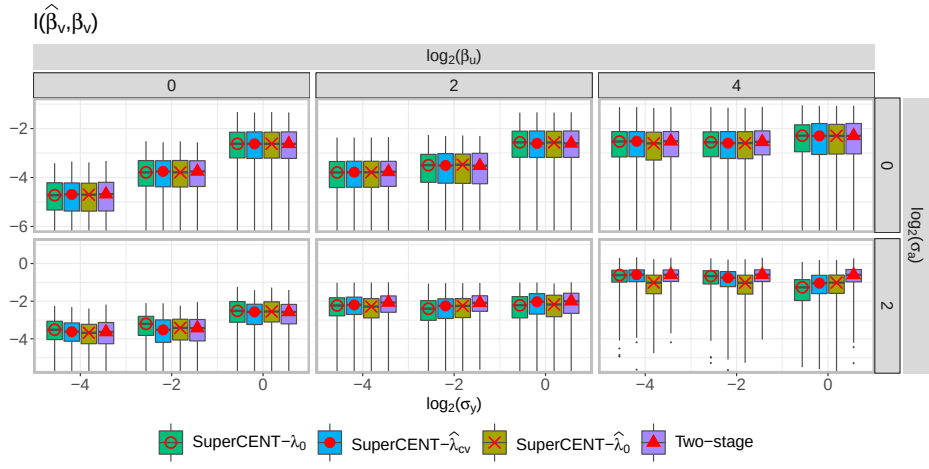


FIGURE S15.—Inconsistent regime of two-stage: The boxplot of $\log_{10}(l(\hat{\beta}_v, \beta_v))$ across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$ where $l(\hat{\beta}_v, \beta_v) = (\hat{\beta}_v - \beta_v)^2/\beta_v^2$. The super-imposed red points show the median of $\log_{10}(l(\hat{\beta}_v, \beta_v))$.

Similar to β_u, the bias in the estimation of β_v further affects its confidence interval. Figures S16 and S17 show the empirical coverage and $\log_{10}$ of the average width, respectively, of the 95% confidence interval for β_v. For the empirical coverage, when σ_a remains small (top panels), all methods remain close to the nominal level, again with different reasons for different methods as discussed for the coverage of CI_{β_u}. When σ_a remains small, the bias in $\hat{\beta}_v$ is relatively small, leaving a relatively small impact on the coverage. As β_u increases, the oracles tend to under cover since the bias increases. Two-stage is still conservative because σ_y^2 is overestimated, covering up the issue of bias. Similar to the phenomena we observed in the consistent regime, the two-stage-adhoc is on par with the two-stage-oracle even with σ_y^2 being overestimated and thus below the nominal level. When σ_a increases and the bias is not too severe with $\beta_u = 2^2$ (the mid-bottom panel), two-stage is still close to the nominal level while two-stage-oracle and two-stage-adhoc are no longer valid. SuperCENT does not improve much from two-stage as the improvement of estimating β_v is limited. The trade-off between β_u and β_v for SuperCENT is desirable – SuperCENT provides valid and shorter intervals for both β_u and β_v if β_u and β_v are similar; if β_u and β_v differ a lot, SuperCENT provides a valid and shorter interval for the larger effect which is more of one's interest.

As for the width of the CI for β_v, Figure S17 shows that when the SuperCENT methods reach the nominal level, the widths are shorter than two-stage.

Finally, Figure S18 shows the average coverage of the confidence intervals of $\boldsymbol{A}_0$. All methods remain at the nominal level but tends to be slightly below the nominal coverage as σ_a increases. SuperCENT-$\hat{\lambda}_{cv}$ is the closest to the nominal coverage in all the settings compared to the others. At the same time, SuperCENT-$\hat{\lambda}_{cv}$ retains its advantage in widths compared to the two-stage as shown in 10.

S5. ADDITIONAL INFORMATION FOR THE CASE STUDY IN SECTION 7

Table SI provides the country acronyms and full names. Figure S19 shows the average trade volume from 2003 to 2012 among the 24 countries/regions. The arrows reflect the trade directions and the widths represent the volume.

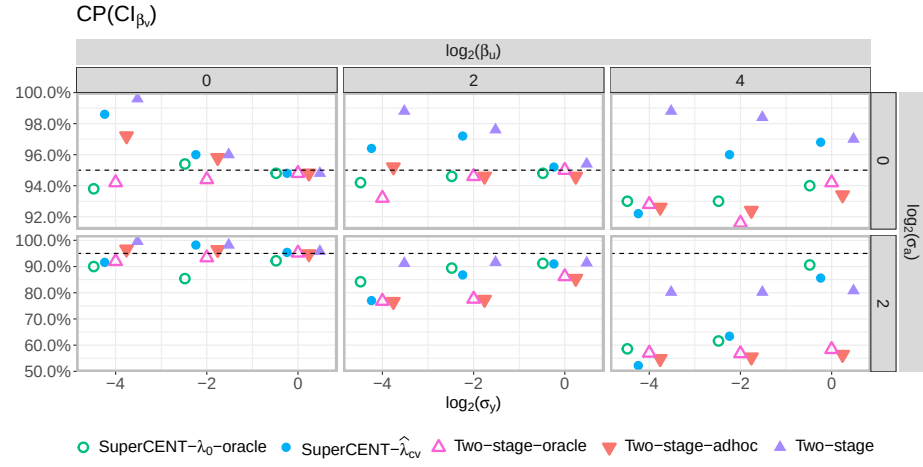


FIGURE S16.—Inconsistent regime of two-stage: Empirical coverage of CI_{β_v} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles. The dashed lines show the nominal confidence level 0.95.

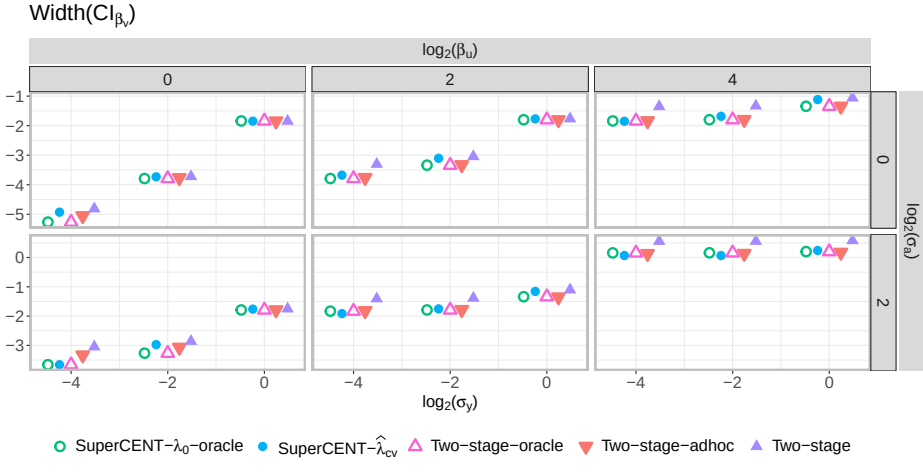


FIGURE S17.—Inconsistent regime of two-stage: Width of CI_{β_v} across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles.

S6. PROOF OF THE THEORETICAL RESULTS IN SECTION 5

In this section, we collect the proofs of Theorem 1, Corollaries 1, 2, 3 in Section S6.1 and Theorem 2 and Corollaries 4, 5 in Section S6.2.

We begin by providing some basic properties of the Kronecker product and the commutation matrix. The Kronecker product of $\boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{m\times n}$ and $\boldsymbol{N} = (n_{ij}) \in \mathbb{R}^{p\times q}$, denoted by

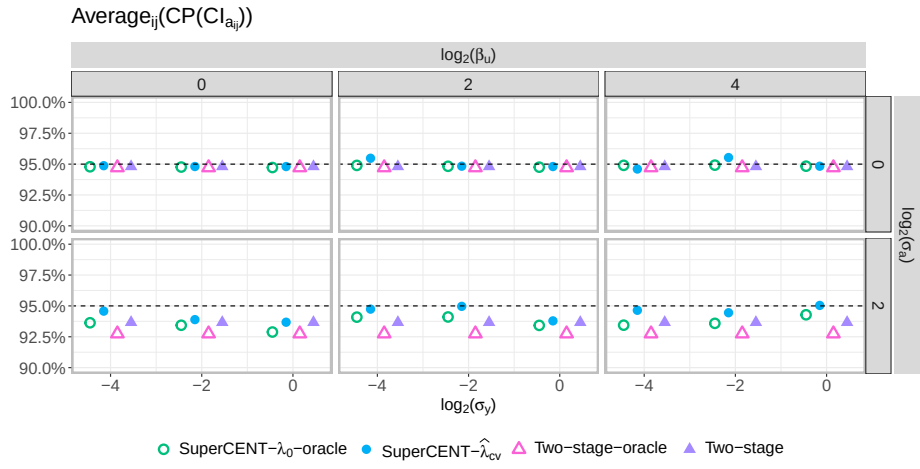


FIGURE S18.—Inconsistent regime of two-stage: The average empirical coverage Average$_{ij}$(CP(CI$_{a_{ij}}$)) across different σ_a, σ_y and β_u with $d = 1$ and $\beta_v = 1$. SuperCENT variants are labelled as circles (○ ●) and the two-stage variants are labelled as triangles (△ ▼ ▲). The hollow ones are for oracles and the solid ones are for non-oracles. The dashed lines show the nominal confidence level 0.95.

Code	Country
AUS	Australia
CAN	Canada
CHE	Switzerland
CZE	Czech Republic
DNK	Denmark
GBR	United Kingdom
HKG	Hong Kong
HUN	Hungary
IDN	Indonesia
IND	India

Code	Country
JPN	Japan
KOR	Korea
KWT	Kuwait
MEX	Mexico
MYS	Malaysia
NOR	Norway
NZL	New Zealand
PHL	Philippines
POL	Poland
SAU	Saudi Arabia

Code	Country
SGP	Singapore
SWE	Sweden
THA	Thailand
ZAF	South Africa

TABLE SI

LIST OF COUNTRY ACRONYMS.

$\boldsymbol{M} \otimes \boldsymbol{N}$, is defined as

$$\boldsymbol{M} \otimes \boldsymbol{N} = \begin{pmatrix} m_{11}\boldsymbol{N} & \cdots & m_{1n}\boldsymbol{N} \\ \vdots & \ldots & \vdots \\ m_{m1}\boldsymbol{N} & \cdots & m_{mn}\boldsymbol{N} \end{pmatrix} \in \mathbb{R}^{mp \times nq}. \tag{S27}$$

Denote $\boldsymbol{K}_{mn} \in \{0,1\}^{mn \times mn}$ as the commutation matrix such that

$$\operatorname{vec}(\boldsymbol{M}^\top) = \boldsymbol{K}_{mn} \operatorname{vec}(\boldsymbol{M}). \tag{S28}$$

We list the following facts about the Kronecker product and the commutation matrix, which are used in the section without specific references. Proofs of these facts can be found in Magnus and Neudecker (1979).

Let $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, $\boldsymbol{N} \in \mathbb{R}^{p \times q}$, $\boldsymbol{P} \in \mathbb{R}^{n \times t}$, $\boldsymbol{Q} \in \mathbb{R}^{q \times s}$, and $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$.

(i) $(\boldsymbol{M} \otimes \boldsymbol{N})^\top = \boldsymbol{M}^\top \otimes \boldsymbol{N}^\top$.

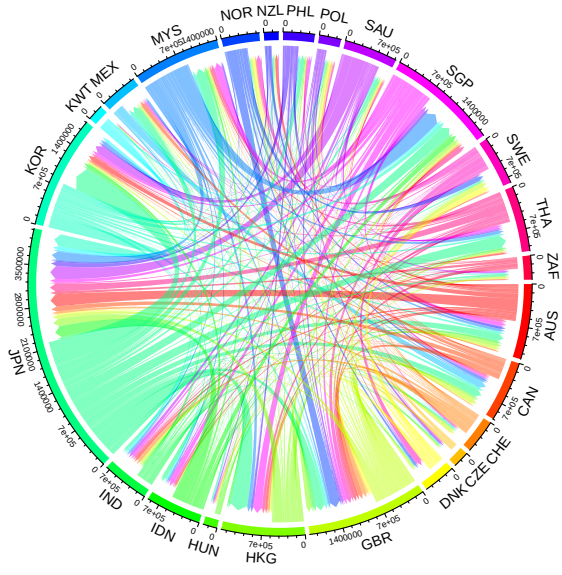


FIGURE S19.—The average trade volume from 2003 to 2012 among the 24 countries/regions. Each country is in different color.

(ii) $(\boldsymbol{M}\otimes\boldsymbol{N})(\boldsymbol{P}\otimes\boldsymbol{Q})=(\boldsymbol{MP})\otimes(\boldsymbol{NQ})$.
(iii) $\mathrm{vec}(\boldsymbol{MZN})=(\boldsymbol{N}^\top\otimes\boldsymbol{M})\,\mathrm{vec}(\boldsymbol{Z})$.
(iv) $\mathrm{vec}(\boldsymbol{MP})=(\boldsymbol{I}\otimes\boldsymbol{M})\,\mathrm{vec}(\boldsymbol{P})=(\boldsymbol{P}^\top\otimes\boldsymbol{I})\,\mathrm{vec}(\boldsymbol{M})$.
(v) $\boldsymbol{K}_{mn}^\top=\boldsymbol{K}_{nm}$.
(vi) $\boldsymbol{K}_{mn}^\top\boldsymbol{K}_{mn}=\boldsymbol{K}_{mn}^\top\boldsymbol{K}_{mn}=\boldsymbol{I}$.
(vii) $\boldsymbol{K}_{mp}(\boldsymbol{M}\otimes\boldsymbol{N})\boldsymbol{K}_{qn}=(\boldsymbol{N}\otimes\boldsymbol{M})$. Equivalently, $\boldsymbol{K}_{mp}(\boldsymbol{M}\otimes\boldsymbol{N})=(\boldsymbol{N}\otimes\boldsymbol{M})\boldsymbol{K}_{qn}$.
(viii) $(\boldsymbol{M}\otimes\boldsymbol{N})\boldsymbol{K}_{nq}(\boldsymbol{P}\otimes\boldsymbol{Q})=((\boldsymbol{MP})\otimes(\boldsymbol{NQ}))\boldsymbol{K}_{st}=\boldsymbol{K}_{mp}((\boldsymbol{NQ})\otimes(\boldsymbol{MP}))$.
(ix) $\mathrm{tr}\,(\boldsymbol{K}_{mn}(\boldsymbol{M}\otimes\boldsymbol{N}))=\mathrm{tr}\,(\boldsymbol{MN})$ where $\boldsymbol{M},\boldsymbol{N}\in\mathbb{R}^{m\times n}$.

S6.1. *Proof of the theoretical results of two-stage*

S6.1.1. *Proof of Theorem 1*

PROOF: The naive two-stage procedure first estimates the centralities $\boldsymbol{u}$ and $\boldsymbol{v}$ by the leading left and right singular vectors, rescaled to have norm $\sqrt{n}$ and denoted as $\hat{\boldsymbol{u}}^{ts},\hat{\boldsymbol{v}}^{ts}$, from the SVD on the observed adjacency matrix $\boldsymbol{A}$, and then performs ordinary least square (OLS) regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\hat{\boldsymbol{u}}^{ts},\hat{\boldsymbol{v}}^{ts}$, treating $\hat{\boldsymbol{u}}^{ts},\hat{\boldsymbol{v}}^{ts}$ as given covariates. It is, therefore, equivalent to solve the following two optimization problems sequentially,

$$\begin{cases} (\hat{d}^{ts},\hat{\boldsymbol{u}}^{ts},\hat{\boldsymbol{v}}^{ts}):=\underset{d,\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\sqrt{n}}{\arg\min}\|\boldsymbol{A}-d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2, & \text{(S29a)}\\ \hat{\boldsymbol{\beta}}^{ts}=((\hat{\boldsymbol{\beta}}_x^{ts})^\top,\hat{\beta}_u^{ts},\hat{\beta}_v^{ts})^\top:=\underset{\boldsymbol{\beta}_x,\beta_u,\beta_v}{\arg\min}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\hat{\boldsymbol{u}}^{ts}\beta_u-\hat{\boldsymbol{v}}^{ts}\beta_v\|_2^2. & \text{(S29b)}\end{cases}$$

Denote $\mathcal{L}_1:=\|\boldsymbol{A}-d\boldsymbol{u}\boldsymbol{v}^\top\|_F^2$ and $\mathcal{L}_2:=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\hat{\boldsymbol{u}}^{ts}\beta_u-\hat{\boldsymbol{v}}^{ts}\beta_v\|_2^2$. In the following, we refer to solving (S29a) as the first stage and (S29b) as the second stage.

(1) First stage We minimize the objection function (S29a) of the first stage by setting the partial derivatives with all the parameters as zero. The partial derivatives are as follows.

$$\frac{\partial \mathcal{L}_1}{\partial d} = 2dn^2 - 2\boldsymbol{u}^\top \mathbf{A}\boldsymbol{v}, \tag{S30}$$

$$\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{u}} = 2d^2 n\boldsymbol{u} - 2d\mathbf{A}\boldsymbol{v}, \tag{S31}$$

$$\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{v}} = 2d^2 n\boldsymbol{v} - 2d\mathbf{A}^\top \boldsymbol{u}. \tag{S32}$$

By setting the partial derivatives above as zero, we have the estimates

$$\hat{d}^{ts} = \hat{\boldsymbol{u}}^{ts\top} \boldsymbol{A}\hat{\boldsymbol{v}}^{ts}/n^2, \tag{S33}$$

$$\hat{d}^{ts}\hat{\boldsymbol{u}}^{ts} = \mathbf{A}\hat{\boldsymbol{v}}^{ts}/n, \tag{S34}$$

$$\hat{d}^{ts}\hat{\boldsymbol{v}}^{ts} = \mathbf{A}^\top \hat{\boldsymbol{u}}^{ts}/n \tag{S35}$$

with constraints

$$\hat{\boldsymbol{u}}^{ts\top}\hat{\boldsymbol{u}}^{ts} = n \quad \text{and} \quad \hat{\boldsymbol{v}}^{ts\top}\hat{\boldsymbol{v}}^{ts} = n. \tag{S36}$$

Together they lead to the first order expansion

$$(\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts})^\top (\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts}) \approx n, \tag{S37}$$

$$(\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts})^\top (\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts}) \approx n, \tag{S38}$$

$$d + \delta_d^{ts} \approx (\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts})^\top (d\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E})(\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts})/n^2, \tag{S39}$$

$$(d + \delta_d^{ts})(\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts}) - (d\boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{E})(\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts})/n \approx 0, \tag{S40}$$

$$(d + \delta_d^{ts})(\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts}) - (d\boldsymbol{v}\boldsymbol{u}^\top + \boldsymbol{E}^\top)(\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts})/n \approx 0. \tag{S41}$$

After simplification and dropping the second order terms, (S37) and (S38) become

$$\boldsymbol{u}^\top \delta_{\boldsymbol{u}}^{ts}/n \approx 0 \quad \text{and} \quad \boldsymbol{v}^\top \delta_{\boldsymbol{v}}^{ts}/n \approx 0. \tag{S42}$$

Further simplifying the rest and using (S42), we have

$$\delta_d^{ts} \approx d\boldsymbol{u}^\top \delta_{\boldsymbol{u}}^{ts} + d\boldsymbol{v}^\top \delta_{\boldsymbol{v}}^{ts} + \boldsymbol{u}^\top \boldsymbol{E}\boldsymbol{v}/n^2 \approx \boldsymbol{u}^\top \boldsymbol{E}\boldsymbol{v}/n^2, \tag{S43}$$

$$nd\delta_{\boldsymbol{u}}^{ts} \approx (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})\boldsymbol{E}\boldsymbol{v}, \tag{S44}$$

$$nd\delta_{\boldsymbol{v}}^{ts} \approx (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})\boldsymbol{E}^\top \boldsymbol{u}. \tag{S45}$$

Combining (S44) and (S45), we obtain

$$\begin{pmatrix} \delta_{\boldsymbol{u}}^{ts} \\ \delta_{\boldsymbol{v}}^{ts} \end{pmatrix} \approx (dn)^{-1} \begin{pmatrix} \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \\ (\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}))\,\boldsymbol{K} \end{pmatrix} \mathrm{vec}(\boldsymbol{E}) \stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{C}_{22}^{ts} \end{pmatrix} \mathrm{vec}(\boldsymbol{E}). \tag{S46}$$

Let $\widehat{\boldsymbol{A}}^{ts} = \hat{d}^{ts}\hat{\boldsymbol{u}}^{ts}\hat{\boldsymbol{v}}^{ts\top}$. Then the first order expansion leads to

$$\mathrm{vec}\left(\delta_{\boldsymbol{A}}^{ts}\right) = \mathrm{vec}(\hat{d}^{ts}\hat{\boldsymbol{u}}^{ts}\hat{\boldsymbol{v}}^{ts\top}) - \mathrm{vec}(d\boldsymbol{u}\boldsymbol{v}^\top) \tag{S47}$$

$$= \mathrm{vec}(d + \delta_d^{ts})(\boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts})(\boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts})^\top - \mathrm{vec}(d\boldsymbol{u}\boldsymbol{v}^\top) \tag{S48}$$

$$= d\,\mathrm{vec}(\boldsymbol{u}\delta_{\boldsymbol{v}}^{ts\top}) + d\,\mathrm{vec}(\delta_{\boldsymbol{u}}^{ts}\boldsymbol{v}^\top) + \delta_d^{ts}\,\mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top) \tag{S49}$$

$$= d\boldsymbol{K}\,\mathrm{vec}(\delta_{\boldsymbol{v}}^{ts}\boldsymbol{u}^\top) + d\,\mathrm{vec}(\delta_{\boldsymbol{u}}^{ts}\boldsymbol{v}^\top) + \delta_d^{ts}\,\mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top) \tag{S50}$$

$$= d\boldsymbol{K}(\boldsymbol{u} \otimes \boldsymbol{I})\delta_{\boldsymbol{v}}^{ts} + d(\boldsymbol{v} \otimes \boldsymbol{I})\delta_{\boldsymbol{u}}^{ts} + \delta_d^{ts}\,\mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top) \tag{S51}$$

$$= n^{-1}\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})\left(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right)\boldsymbol{K}\,\mathrm{vec}(\boldsymbol{E}) \tag{S52}$$

$$+n^{-1}(\boldsymbol{v}\otimes\boldsymbol{I})\left(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\right)\mathrm{vec}(\boldsymbol{E})+n^{-2}\boldsymbol{u}^\top\boldsymbol{E}\boldsymbol{v}\,\mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top) \tag{S53}$$

$$= [(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}+\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}]\,\mathrm{vec}(\boldsymbol{E}) \tag{S54}$$

$$\stackrel{def}{=} \boldsymbol{C}_{32}\,\mathrm{vec}(\boldsymbol{E}). \tag{S55}$$

Putting together (S46) and (S55)

$$\begin{pmatrix}\hat{\boldsymbol{u}}^{ts}-\boldsymbol{u}\\ \hat{\boldsymbol{v}}^{ts}-\boldsymbol{v}\\ \mathrm{vec}\left(\hat{\boldsymbol{A}}^{ts}-\boldsymbol{A}_0\right)\end{pmatrix}\approx\begin{pmatrix}\boldsymbol{C}_{12}^{ts}\\ \boldsymbol{C}_{22}^{ts}\\ \boldsymbol{C}_{32}^{ts}\end{pmatrix}\mathrm{vec}(\boldsymbol{E}). \tag{S56}$$

This finishes the proof of the first stage.

(2) Second stage. For the second stage, we plug in $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ from the first stage into the objection function (S29b) of the second stage. We then minimize the objection function by setting the partial derivatives with the regression coefficients as zero,

$$\frac{\partial\mathcal{L}_2}{\partial\beta_u}=-\frac{2}{n}\hat{\boldsymbol{u}}^{ts\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\hat{\boldsymbol{u}}^{ts}\beta_u-\hat{\boldsymbol{v}}^{ts}\beta_v)=0, \tag{S57}$$

$$\frac{\partial\mathcal{L}_2}{\partial\beta_v}=-\frac{2}{n}\hat{\boldsymbol{v}}^{ts\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\hat{\boldsymbol{u}}^{ts}\beta_u-\hat{\boldsymbol{v}}^{ts}\beta_v)=0, \tag{S58}$$

$$\frac{\partial\mathcal{L}_2}{\partial\boldsymbol{\beta}_x}=-\frac{2}{n}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x-\hat{\boldsymbol{u}}^{ts}\beta_u-\hat{\boldsymbol{v}}^{ts}\beta_v)=0. \tag{S59}$$

After some algebra, we obtain the estimates as

$$\hat{\beta}_u^{ts}=\frac{1}{n}\hat{\boldsymbol{u}}^{ts\top}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_x^{ts}-\hat{\boldsymbol{v}}^{ts}\hat{\beta}_v^{ts}), \tag{S60}$$

$$\hat{\beta}_v^{ts}=\frac{1}{n}\hat{\boldsymbol{v}}^{ts\top}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_x^{ts}-\hat{\boldsymbol{u}}^{ts}\hat{\beta}_u^{ts}), \tag{S61}$$

$$\hat{\boldsymbol{\beta}}_x^{ts}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y}-\hat{\boldsymbol{u}}^{ts}\hat{\beta}_u^{ts}-\hat{\boldsymbol{v}}^{ts}\hat{\beta}_v^{ts}), \tag{S62}$$

which leads to the first order expansion

$$\beta_u+\delta_{\beta_u}^{ts}\approx\frac{1}{n}(\boldsymbol{u}+\delta_{\boldsymbol{u}}^{ts})^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{u}\beta_u+\boldsymbol{v}\beta_v+\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\beta}_x-\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}^{ts}-\boldsymbol{v}\beta_v-\boldsymbol{v}\delta_{\beta_v}^{ts}-\delta_{\boldsymbol{v}}^{ts}\beta_v), \tag{S63}$$

$$\beta_v+\delta_{\beta_v}^{ts}\approx\frac{1}{n}(\boldsymbol{v}+\delta_{\boldsymbol{v}}^{ts})^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{u}\beta_u+\boldsymbol{v}\beta_v+\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\beta}_x-\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}^{ts}-\boldsymbol{u}\beta_u-\boldsymbol{u}\delta_{\beta_u}^{ts}-\delta_{\boldsymbol{u}}^{ts}\beta_u), \tag{S64}$$

$$\boldsymbol{\beta}_x+\delta_{\boldsymbol{\beta}_x}^{ts}\approx(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{\epsilon}-\boldsymbol{u}\delta_{\beta_u}^{ts}-\delta_{\boldsymbol{u}}^{ts}\beta_u-\boldsymbol{v}\delta_{\beta_v}^{ts}-\delta_{\boldsymbol{v}}^{ts}\beta_v). \tag{S65}$$

After simplification and dropping the second order terms, we have

$$\boldsymbol{u}^\top\delta_{\boldsymbol{v}}^{ts}\beta_v+\boldsymbol{u}^\top\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}^{ts}+n\delta_{\beta_u}^{ts}+\boldsymbol{u}^\top\boldsymbol{v}\delta_{\beta_v}^{ts}\approx\boldsymbol{u}^\top\boldsymbol{\epsilon}, \tag{S66}$$

$$\boldsymbol{v}^\top\delta_{\boldsymbol{u}}^{ts}\beta_u+\boldsymbol{v}^\top\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}^{ts}+\boldsymbol{u}^\top\boldsymbol{v}\delta_{\beta_u}^{ts}+n\delta_{\beta_v}^{ts}\approx\boldsymbol{v}^\top\boldsymbol{\epsilon}, \tag{S67}$$

$$\delta_{\boldsymbol{\beta}_x}^{ts}\approx(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{\epsilon}-\boldsymbol{u}\delta_{\beta_u}^{ts}-\delta_{\boldsymbol{u}}^{ts}\beta_u-\boldsymbol{v}\delta_{\beta_v}^{ts}-\delta_{\boldsymbol{v}}^{ts}\beta_v). \tag{S68}$$

Plugging (S68) into (S66, S67) and using (S42), we have

$$\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta^{ts}_{\boldsymbol{u}}\beta_u+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta^{ts}_{\boldsymbol{v}}\beta_v+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\delta^{ts}_{\beta_u}+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}\delta^{ts}_{\beta_v} \tag{S69}$$

$$\approx \boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon}, \tag{S70}$$

$$\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta^{ts}_{\boldsymbol{u}}\beta_u+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta^{ts}_{\boldsymbol{v}}\beta_v+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\delta^{ts}_{\beta_u}+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}\delta^{ts}_{\beta_v} \tag{S71}$$

$$\approx \boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon}. \tag{S72}$$

Denote

$$\tilde{\boldsymbol{u}}=(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u} \quad \text{and} \quad \tilde{\boldsymbol{v}}=(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v} \tag{S73}$$

and

$$\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}=\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} & \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}\\ \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} & \tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}}\end{pmatrix}. \tag{S74}$$

Then (S70)-(S72) can be written as

$$\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\delta^{ts}_{\beta_u}\\ \delta^{ts}_{\beta_v}\end{pmatrix}\approx\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; \boldsymbol{I}_n)\begin{pmatrix}\delta^{ts}_{\boldsymbol{u}}\\ \delta^{ts}_{\boldsymbol{v}}\\ \boldsymbol{\epsilon}\end{pmatrix}. \tag{S75}$$

Plugging (S46) into (S75) and solving for $\begin{pmatrix}\delta^{ts}_{\beta_u}\\ \delta^{ts}_{\beta_v}\end{pmatrix}$, we obtain

$$\begin{pmatrix}\delta^{ts}_{\beta_u}\\ \delta^{ts}_{\beta_v}\end{pmatrix}\approx\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; \boldsymbol{I}_n)\begin{pmatrix}\boldsymbol{0}_{n\times n} & \boldsymbol{C}^{ts}_{12}\\ \boldsymbol{0}_{n\times n} & \boldsymbol{C}^{ts}_{22}\\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2}\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix} \tag{S76}$$

$$\stackrel{def}{=}\begin{pmatrix}\boldsymbol{C}^{ts}_{41} & \boldsymbol{C}^{ts}_{42}\\ \boldsymbol{C}^{ts}_{51} & \boldsymbol{C}^{ts}_{52}\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix} \tag{S77}$$

where explicitly

$$\begin{pmatrix}\boldsymbol{C}^{ts}_{41}\\ \boldsymbol{C}^{ts}_{51}\end{pmatrix}=\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix} \tag{S78}$$

and

$$\begin{pmatrix}\boldsymbol{C}^{ts}_{42}\\ \boldsymbol{C}^{ts}_{52}\end{pmatrix}=(dn)^{-1}\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\beta_u\boldsymbol{v}^\top\otimes\left(\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\right)\\ \beta_v\boldsymbol{u}^\top\otimes\left(\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right)\boldsymbol{K}\end{pmatrix}. \tag{S79}$$

For $\delta^{ts}_{\boldsymbol{\beta}_x}$, plugging (S46) and (S77) into (S68),

$$\delta^{ts}_{\boldsymbol{\beta}_x}\approx(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{\epsilon}-\boldsymbol{u}\delta^{ts}_{\beta_u}-\delta^{ts}_{\boldsymbol{u}}\beta_u-\boldsymbol{v}\delta^{ts}_{\beta_v}-\delta^{ts}_{\boldsymbol{v}}\beta_v) \tag{S80}$$

$$=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; -\boldsymbol{u} \;\; -\boldsymbol{v} \;\; \boldsymbol{I}_n)\begin{pmatrix}\delta^{ts}_{\boldsymbol{u}}\\ \delta^{ts}_{\boldsymbol{v}}\\ \delta^{ts}_{\beta_u}\\ \delta^{ts}_{\beta_v}\\ \boldsymbol{\epsilon}\end{pmatrix} \tag{S81}$$

$$=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(-\beta_u\boldsymbol{I}_n \;\; -\beta_v\boldsymbol{I}_n \;\; -\boldsymbol{u} \;\; -\boldsymbol{v} \;\; \boldsymbol{I}_n)\begin{pmatrix}\boldsymbol{0}_{n\times n} & \boldsymbol{C}^{ts}_{12}\\ \boldsymbol{0}_{n\times n} & \boldsymbol{C}^{ts}_{22}\\ \boldsymbol{C}^{ts}_{41} & \boldsymbol{C}^{ts}_{52}\\ \boldsymbol{C}^{ts}_{51} & \boldsymbol{C}^{ts}_{52}\\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2}\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix} \tag{S82}$$

$$\stackrel{def}{=} \left(\boldsymbol{C}_{61}^{ts}\ \boldsymbol{C}_{62}^{ts}\right)\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix}. \tag{S83}$$

Finally, putting (S46), (S55), (S77) and (S83) together, we have

$$\begin{pmatrix}\hat{\boldsymbol{u}}^{ts}-\boldsymbol{u}\\ \hat{\boldsymbol{v}}^{ts}-\boldsymbol{v}\\ \mathrm{vec}\left(\widehat{\boldsymbol{A}}^{ts}-\boldsymbol{A}_0\right)\\ \hat{\beta}_u^{ts}-\beta_u\\ \hat{\beta}_v^{ts}-\beta_v\\ \hat{\boldsymbol{\beta}}_x^{ts}-\boldsymbol{\beta}_x\end{pmatrix}=\begin{pmatrix}\delta_{\boldsymbol{u}}^{ts}\\ \delta_{\boldsymbol{v}}^{ts}\\ \mathrm{vec}\left(\delta_{\boldsymbol{A}}^{ts}\right)\\ \delta_{\beta_u}^{ts}\\ \delta_{\beta_v}^{ts}\\ \delta_{\boldsymbol{\beta}_x}^{ts}\end{pmatrix}\approx\begin{pmatrix}\boldsymbol{0}_{n\times n} & \boldsymbol{C}_{12}^{ts}\\ \boldsymbol{0}_{n\times n} & \boldsymbol{C}_{22}^{ts}\\ \boldsymbol{0}_{n^2\times n} & \boldsymbol{C}_{32}^{ts}\\ \boldsymbol{C}_{41}^{ts} & \boldsymbol{C}_{42}^{ts}\\ \boldsymbol{C}_{51}^{ts} & \boldsymbol{C}_{52}^{ts}\\ \boldsymbol{C}_{61}^{ts} & \boldsymbol{C}_{62}^{ts}\end{pmatrix}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix}\stackrel{def}{=}\boldsymbol{C}^{ts}\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix}. \tag{S84}$$

Recall that we assume

$$\begin{pmatrix}\boldsymbol{\epsilon}\\ \mathrm{vec}(\boldsymbol{E})\end{pmatrix}\sim N\left(\boldsymbol{0}_{(n+n^2)\times 1},\begin{pmatrix}\sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2}\\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2}\end{pmatrix}\right). \tag{S85}$$

Therefore, the two-stage estimators converge to the following normal distribution asymptotically,

$$\begin{pmatrix}\hat{\boldsymbol{u}}^{ts}-\boldsymbol{u}\\ \hat{\boldsymbol{v}}^{ts}-\boldsymbol{v}\\ \mathrm{vec}\left(\widehat{\boldsymbol{A}}^{ts}-\boldsymbol{A}_0\right)\\ \hat{\beta}_u^{ts}-\beta_u\\ \hat{\beta}_v^{ts}-\beta_v\\ \hat{\boldsymbol{\beta}}_x^{ts}-\boldsymbol{\beta}_x\end{pmatrix}\xrightarrow{\mathcal{D}}N\left(\boldsymbol{0}_{(2n+n^2+2+p)\times 1},\boldsymbol{C}^{ts}\begin{pmatrix}\sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2}\\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2}\end{pmatrix}\boldsymbol{C}^{ts\top}\right). \tag{S86}$$

Q.E.D.

S6.1.2. *Proof of Corollary 1*

PROOF: From Theorem 1, we have

$$\begin{pmatrix}\delta_{\boldsymbol{u}}^{ts}\\ \delta_{\boldsymbol{v}}^{ts}\\ \delta_{\boldsymbol{A}}^{ts}\end{pmatrix}\approx\begin{pmatrix}\boldsymbol{C}_{12}^{ts}\\ \boldsymbol{C}_{22}^{ts}\\ \boldsymbol{C}_{32}^{ts}\end{pmatrix}\mathrm{vec}(\boldsymbol{E}) \tag{S87}$$

where $\boldsymbol{C}_{12}^{ts}=(dn)^{-1}\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})$, $\boldsymbol{C}_{22}^{ts}=(dn)^{-1}\left(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right)\boldsymbol{K}$ and $\boldsymbol{C}_{32}^{ts}=[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}+\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}]$.

(i) Rate of $\hat{\boldsymbol{u}}^{ts}$, $\hat{\boldsymbol{v}}^{ts}$.

$$\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{u}}^{ts}-\boldsymbol{u}\|_2^2\approx\frac{1}{n}\mathrm{tr}(\sigma_a^2\boldsymbol{C}_{12}^{ts}\boldsymbol{C}_{12}^{ts\top}) \tag{S88}$$

$$=\frac{1}{n}\frac{\sigma_a^2}{(dn)^2}\mathrm{tr}\left(\left(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\right)(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\right) \tag{S89}$$

$$=\frac{1}{n}\frac{\sigma_a^2}{(dn)^2}\mathrm{tr}\left(\boldsymbol{v}^\top\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\right) \tag{S90}$$

$$=\frac{1}{n}\frac{\sigma_a^2}{(dn)^2}\mathrm{tr}\left(n\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top\right) \tag{S91}$$

$$=\frac{\sigma_a^2}{d^2n^2}(n-1) \tag{S92}$$

$$=O(\kappa). \tag{S93}$$

Similarly for the rate of $\hat{\boldsymbol{v}}^{ts}$,

$$\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{v}}^{ts}-\boldsymbol{v}\|_2^2 \approx \frac{1}{n}\mathrm{tr}(\sigma_a^2\boldsymbol{C}_{22}^{ts}\boldsymbol{C}_{22}^{ts\top}) \tag{S94}$$

$$= \frac{1}{n}\frac{\sigma_a^2}{(dn)^2}\mathrm{tr}\left(\Big(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\Big)(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))\right) \tag{S95}$$

$$= \frac{\sigma_a^2}{d^2n^2}(n-1) \tag{S96}$$

$$= O(\kappa) \tag{S97}$$

where the second equality is due to $\boldsymbol{K}\boldsymbol{K}^\top=\boldsymbol{I}$.

(ii) Rate of $\hat{\boldsymbol{A}}^{ts}$. Note that

$$[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}][(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}]^\top = (\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}} \tag{S98}$$

$$[\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})][\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})]^\top = \boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}) \tag{S99}$$

$$[\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}][\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}]^\top = \boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}} \tag{S100}$$

$$[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}][\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})]^\top = \boldsymbol{0} \tag{S101}$$

$$[\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})][(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}]^\top = \boldsymbol{0} \tag{S102}$$

$$[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}][\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}]^\top = \boldsymbol{0} \tag{S103}$$

$$[\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})][\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}]^\top = \boldsymbol{0}. \tag{S104}$$

Putting together, we have

$$\mathbb{E}\frac{\left\|\hat{\boldsymbol{A}}^{ts}-\boldsymbol{A}_0\right\|_F^2}{\|\boldsymbol{A}_0\|_F^2} \approx \frac{1}{d^2n^2}\mathrm{tr}(\sigma_a^2\boldsymbol{C}_{32}^{ts}\boldsymbol{C}_{32}^{ts\top}) \tag{S105}$$

$$= \frac{1}{d^2n^2}\sigma_a^2\mathrm{tr}((\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}+\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}) \tag{S106}$$

$$= \frac{\sigma_a^2}{d^2n^2}\left[2(n-1)+1\right] = \frac{\sigma_a^2}{d^2n^2}(2n-1) = O(\kappa). \tag{S107}$$

Q.E.D.

S6.1.3. *Proof of Corollary 2*

PROOF: We now prove the rate of $\hat{\boldsymbol{\beta}}^{ts}$ of two-stage.

(1) Rate of $\hat{\beta}_u^{ts}$ *and* $\hat{\beta}_v^{ts}$. Recall that

$$\begin{pmatrix}\delta_{\beta_u}^{ts}\\ \delta_{\beta_v}^{ts}\end{pmatrix} \approx \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\boldsymbol{\epsilon}-(\delta_{\boldsymbol{u}}^{ts}\beta_u+\delta_{\boldsymbol{v}}^{ts}\beta_v)\right] \tag{S108}$$

$$= \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\boldsymbol{\epsilon}-\frac{1}{dn}\left(\beta_u\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\boldsymbol{K}\right)\mathrm{vec}(\boldsymbol{E})\right] \tag{S109}$$

$$\stackrel{def}{=} \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\boldsymbol{\epsilon}+\boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E})\right] \tag{S110}$$

$$\stackrel{def}{=} \boldsymbol{D}_1\boldsymbol{\epsilon}+\boldsymbol{D}_2\,\mathrm{vec}(\boldsymbol{E}). \tag{S111}$$

Note that

$$\boldsymbol{D}_1\boldsymbol{D}_1^\top = \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}(\,\tilde{\boldsymbol{u}}\;\tilde{\boldsymbol{v}}\,)\,\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} = \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S112}$$

and

$$\boldsymbol{D}_2\boldsymbol{D}_2^\top = \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\boldsymbol{B}\boldsymbol{B}^\top(\,\tilde{\boldsymbol{u}}\;\tilde{\boldsymbol{v}}\,)\,\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S113}$$

where

$$\boldsymbol{B}\boldsymbol{B}^\top = \frac{1}{(dn)^2}\Big[\beta_u^2(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})) + \beta_v^2(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})) \tag{S114}$$

$$+\beta_u\beta_v(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\boldsymbol{K}^\top(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})) + \beta_u\beta_v(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))\boldsymbol{K}(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\Big] \tag{S115}$$

$$= \frac{1}{d^2n}\left[\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right] \tag{S116}$$

since $(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\boldsymbol{K}^\top(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})) = (\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\boldsymbol{u})\boldsymbol{K} = 0$ and $(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))\boldsymbol{K}(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})) = (\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\boldsymbol{v})\boldsymbol{K} = 0$.

Plugging in (S112), (S113) and (S116),

$$Cov\begin{pmatrix}\delta_{\beta_u}^{ts}\\ \delta_{\beta_v}^{ts}\end{pmatrix} \approx \sigma_y^2\boldsymbol{D}_1\boldsymbol{D}_1^\top + \sigma_a^2\boldsymbol{D}_2\boldsymbol{D}_2^\top \tag{S117}$$

$$= \sigma_y^2\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S118}$$

$$+\sigma_a^2\frac{1}{d^2n}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right](\,\tilde{\boldsymbol{u}}\;\tilde{\boldsymbol{v}}\,)\,\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S119}$$

$$= \sigma_y^2\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} + \sigma_a^2\frac{1}{d^2n}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}. \tag{S120}$$

Therefore, we obtain the rate of $\hat{\beta}_u^{ts}$ and $\hat{\beta}_v^{ts}$ from the diagonal entries of $Cov\begin{pmatrix}\delta_{\beta_u}^{ts}\\ \delta_{\beta_v}^{ts}\end{pmatrix}$ as

$$\mathbb{E}(\hat{\beta}_u^{ts}-\beta_u)^2 = \frac{\sigma_y^2}{c}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} \tag{S121}$$

$$+\frac{\sigma_a^2}{c^2}\frac{1}{d^2n}\left[\beta_v^2\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}}\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} + \beta_u^2\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}\right](1+o(1)) \tag{S122}$$

$$\mathbb{E}(\hat{\beta}_v^{ts}-\beta_v)^2 = \frac{\sigma_y^2}{c}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} \tag{S123}$$

$$+\frac{\sigma_a^2}{c^2}\frac{1}{d^2n}\left[\beta_u^2\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} + \beta_v^2\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{u}}\right](1+o(1)) \tag{S124}$$

where $c = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} - (\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}})^2$.

(2) Rate of $\hat{\boldsymbol{\beta}}_x^{ts}$. Recall that

$$\delta_{\boldsymbol{\beta}_x}^{ts} \approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{\epsilon} - \boldsymbol{u}\delta_{\beta_u}^{ts} - \delta_{\boldsymbol{u}}^{ts}\beta_u - \boldsymbol{v}\delta_{\beta_v}^{ts} - \delta_{\boldsymbol{v}}^{ts}\beta_v). \tag{S125}$$

From (S110), we have

$$\boldsymbol{\epsilon} - (\delta_{\boldsymbol{u}}^{ts}\beta_u + \delta_{\boldsymbol{v}}^{ts}\beta_v) \overset{def}{=} \boldsymbol{\epsilon} + \boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E}) \tag{S126}$$

and

$$\begin{pmatrix}\delta_{\beta_u}^{ts}\\ \delta_{\beta_v}^{ts}\end{pmatrix} \approx \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}[\boldsymbol{\epsilon} + \boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E})]. \tag{S127}$$

Then we have

$$\boldsymbol{u}\delta^{ts}_{\beta_u} + \boldsymbol{v}\delta^{ts}_{\beta_v} = (\,\boldsymbol{u}\ \boldsymbol{v}\,)\begin{pmatrix}\delta^{ts}_{\beta_u}\\ \delta^{ts}_{\beta_v}\end{pmatrix} \approx (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}[\boldsymbol{\epsilon} + \boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E})]\,. \tag{S128}$$

Let

$$\tilde{\boldsymbol{G}} = (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix} \text{ and } \boldsymbol{G} = (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}. \tag{S129}$$

Plugging (S110) and (S128) into (S125) yields

$$\begin{aligned}\delta^{ts}_{\boldsymbol{\beta}_x} &\approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\left[\boldsymbol{I} - (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\right](\boldsymbol{\epsilon} + \boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E})) &\text{(S130)}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \tilde{\boldsymbol{G}})(\boldsymbol{\epsilon} + \boldsymbol{B}\,\mathrm{vec}(\boldsymbol{E})) &\text{(S131)}\\ &\overset{def}{=} \boldsymbol{F}_1\boldsymbol{\epsilon} + \boldsymbol{F}_2\,\mathrm{vec}(\boldsymbol{E}). &\text{(S132)}\end{aligned}$$

Hence, the variance-covariance matrix of $\delta^{ts}_{\boldsymbol{\beta}_x}$ is

$$Cov\left(\delta^{ts}_{\boldsymbol{\beta}_x}\right) \approx \sigma_y^2\boldsymbol{F}_1\boldsymbol{F}_1^\top + \sigma_a^2\boldsymbol{F}_2\boldsymbol{F}_2^\top \tag{S133}$$

where we will derive the explicit form of $\boldsymbol{F}_1\boldsymbol{F}_1^\top$ and $\boldsymbol{F}_2\boldsymbol{F}_2^\top$ in the following.

(a) $\boldsymbol{F}_1\boldsymbol{F}_1^\top$.

Since

$$(\boldsymbol{I} - \tilde{\boldsymbol{G}})(\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top = \boldsymbol{I} - \tilde{\boldsymbol{G}} - \tilde{\boldsymbol{G}}^\top + \boldsymbol{G} \tag{S134}$$

and

$$\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\boldsymbol{X} = \boldsymbol{0} \text{ and thus } \tilde{\boldsymbol{G}}\boldsymbol{X} = \boldsymbol{0}, \tag{S135}$$

consequently

$$\boldsymbol{F}_1\boldsymbol{F}_1^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{G}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{S136}$$

(b) $\boldsymbol{F}_2\boldsymbol{F}_2^\top$.

Recall (S116) where

$$\boldsymbol{B}\boldsymbol{B}^\top = \frac{1}{d^2n}\left[\beta_u^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})\right].$$

Plugging in we have,

$$\begin{aligned}&(\boldsymbol{I} - \tilde{\boldsymbol{G}})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top &\text{(S137)}\\ =\ &\frac{1}{d^2n}\Bigg[\beta_u^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) - \tilde{\boldsymbol{G}}\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{B}\boldsymbol{B}^\top\tilde{\boldsymbol{G}}^\top &\text{(S138)}\\ &+ (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\Bigg]. &\text{(S139)}\end{aligned}$$

Together with (S136) and using (S135), we obtain the variance-covariance matrix of $\delta^{ts}_{\boldsymbol{\beta}_x}$ as follows.

$$\begin{aligned}&Cov\left(\hat{\boldsymbol{\beta}}^{ts}_x - \boldsymbol{\beta}_x\right) &\text{(S140)}\\ \approx\ &\sigma_y^2\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right] &\text{(S141)}\\ &+\sigma_a^2\frac{1}{d^2n}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Bigg[\beta_u^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) &\text{(S142)}\\ &+ (\,\boldsymbol{u}\ \boldsymbol{v}\,)\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\Bigg]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. &\text{(S143)}\end{aligned}$$

Q.E.D.

S6.1.4. *Proof of Corollary 3*

PROOF: For simplicity, let us first consider $\boldsymbol{y} = \boldsymbol{u}\beta_u + \boldsymbol{v}\beta_v + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$ independently. Denote the correlation between $\boldsymbol{u}$ and $\boldsymbol{v}$ as ρ.

The two-stage procedure first estimates $\boldsymbol{u}$ and $\boldsymbol{v}$ from the network $\boldsymbol{A}$ as $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$, and then regresses $\boldsymbol{y}$ on $\hat{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts}$ to obtain $\widehat{\boldsymbol{\beta}}^{ts\top} = (\hat{\beta}_u^{ts}, \hat{\beta}_v^{ts})$ in the second stage. From Theorem 1, we have $\hat{\boldsymbol{u}}^{ts} = \boldsymbol{u} + \delta_{\boldsymbol{u}}^{ts}$ and $\hat{\boldsymbol{v}}^{ts} = \boldsymbol{v} + \delta_{\boldsymbol{v}}^{ts}$ where

$$\begin{pmatrix} \delta_{\boldsymbol{u}}^{ts} \\ \delta_{\boldsymbol{v}}^{ts} \end{pmatrix} \approx (dn)^{-1} \begin{pmatrix} \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \\ (\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}))\boldsymbol{K} \end{pmatrix} \mathrm{vec}(\boldsymbol{E}) \stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{12}^{ts} \\ \boldsymbol{C}_{22}^{ts} \end{pmatrix} \mathrm{vec}(\boldsymbol{E}), \tag{S144}$$

$\boldsymbol{u} \perp \delta_{\boldsymbol{u}}^{ts}$ and $\boldsymbol{v} \perp \delta_{\boldsymbol{v}}^{ts}$. Furthermore $\delta_{\boldsymbol{u}}^{ts} \perp \delta_{\boldsymbol{v}}^{ts}$ because

$$\delta_{\boldsymbol{u}}^{ts\top} \delta_{\boldsymbol{v}}^{ts} = \frac{1}{(dn)^2} \mathrm{vec}\,(\boldsymbol{E})^\top (\boldsymbol{v} \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})) \left(\left(\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \right) \right) \boldsymbol{K}\, \mathrm{vec}\,(\boldsymbol{E}) \tag{S145}$$

$$= \frac{\sigma_a^2}{(dn)^2} \mathrm{tr} \left(\left(\boldsymbol{v}\boldsymbol{u}^\top \right) \otimes ((\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}))\, \boldsymbol{K} \right) \tag{S146}$$

$$= \frac{\sigma_a^2}{(dn)^2} \mathrm{tr} \left((\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \boldsymbol{v}\boldsymbol{u}^\top \right) \tag{S147}$$

$$= 0. \tag{S148}$$

Denote $\widetilde{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{\delta}$ where $\boldsymbol{W} = \begin{pmatrix} \boldsymbol{u} & \boldsymbol{v} \end{pmatrix}$ and $\boldsymbol{\delta}^{ts} = \begin{pmatrix} \delta_{\boldsymbol{u}}^{ts} & \delta_{\boldsymbol{v}}^{ts} \end{pmatrix}$, then

$$\hat{\boldsymbol{\beta}}^{ts} = \left(\widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{W}} \right)^{-1} \widetilde{\boldsymbol{W}}^\top \boldsymbol{y} = \left(\boldsymbol{W}^\top \boldsymbol{W} + \boldsymbol{\delta}^\top \boldsymbol{\delta} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{W} \begin{pmatrix} \beta_u \\ \beta_v \end{pmatrix}. \tag{S149}$$

Since the correlation between $\boldsymbol{u}$ and $\boldsymbol{v}$ is ρ, the covariance matrix of $\boldsymbol{W}$ is $\Sigma_{\boldsymbol{uv}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The covariance matrix of $\boldsymbol{\delta}^{ts}$ is $\Sigma_{\boldsymbol{\delta}^{ts}} = \begin{pmatrix} 0 & \kappa \\ \kappa & 0 \end{pmatrix}$ due to Corollary 1 and $\delta_{\boldsymbol{u}}^{ts} \perp \delta_{\boldsymbol{v}}^{ts}$. Taking plim of both sides of (S149), we have

$$\mathrm{plim}\ \hat{\boldsymbol{\beta}}^{ts} = \left(\Sigma_{\boldsymbol{uv}} + \Sigma_{\boldsymbol{\delta}^{ts}} \right)^{-1} \Sigma_{\boldsymbol{uv}} \begin{pmatrix} \beta_u \\ \beta_v \end{pmatrix} \tag{S150}$$

$$= \frac{1}{(1+\kappa)^2 - \rho^2} \begin{pmatrix} 1 + \kappa - \rho^2 & \kappa\rho \\ \kappa\rho & 1 + \kappa - \rho^2 \end{pmatrix} \begin{pmatrix} \beta_u \\ \beta_v \end{pmatrix}. \tag{S151}$$

Specifically, we obtain

$$\mathrm{plim}\ \hat{\beta}_u^{ts} = \frac{(1 + \kappa - \rho^2)\beta_u + \kappa\rho\beta_v}{(1+\kappa)^2 - \rho^2}, \tag{S152}$$

$$\mathrm{plim}\ \hat{\beta}_v^{ts} = \frac{(1 + \kappa - \rho^2)\beta_v + \kappa\rho\beta_u}{(1+\kappa)^2 - \rho^2}. \tag{S153}$$

When $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{u}\beta_u + \boldsymbol{v}\beta_v + \boldsymbol{\epsilon}$ and $\mathrm{cov}(\boldsymbol{X}, (\boldsymbol{uv})) = \boldsymbol{0} \in \mathbb{R}^{p \times 2}$, the above results remain the same due to the property of ordinary least square estimator.

Q.E.D.

S6.2. *Proof of the theoretical results of SuperCENT*

S6.2.1. *Proof of Theorem 2*

PROOF: In Section S1, we derive the estimates of $(\hat{d}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}})$ as (S6)-(S10). Together they lead to the first order expansion

$$(\boldsymbol{u}+\delta_{\boldsymbol{u}})^\top(\boldsymbol{u}+\delta_{\boldsymbol{u}}) \approx n \tag{S154}$$

$$(\boldsymbol{v}+\delta_{\boldsymbol{v}})^\top(\boldsymbol{v}+\delta_{\boldsymbol{v}}) \approx n \tag{S155}$$

$$\beta_u+\delta_{\beta_u} \approx \frac{1}{n}(\boldsymbol{u}+\delta_{\boldsymbol{u}})^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{u}\beta_u+\boldsymbol{v}\beta_v+\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\beta}_x-\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}-\boldsymbol{v}\beta_v-\boldsymbol{v}\delta_{\beta_v}-\delta_{\boldsymbol{v}}\beta_v) \tag{S156}$$

$$\beta_v+\delta_{\beta_v} \approx \frac{1}{n}(\boldsymbol{v}+\delta_{\boldsymbol{v}})^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{u}\beta_u+\boldsymbol{v}\beta_v+\boldsymbol{\epsilon}-\boldsymbol{X}\boldsymbol{\beta}_x-\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}-\boldsymbol{u}\beta_u-\boldsymbol{u}\delta_{\beta_u}-\delta_{\boldsymbol{u}}\beta_u) \tag{S157}$$

$$\boldsymbol{\beta}_x+\delta_{\boldsymbol{\beta}_x} \approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}_x+\boldsymbol{\epsilon}-\boldsymbol{u}\delta_{\beta_u}-\delta_{\boldsymbol{u}}\beta_u-\boldsymbol{v}\delta_{\beta_v}-\delta_{\boldsymbol{v}}\beta_v) \tag{S158}$$

$$d+\delta_d \approx (\boldsymbol{u}+\delta_{\boldsymbol{u}})^\top(d\boldsymbol{u}\boldsymbol{v}^\top+\boldsymbol{E})(\boldsymbol{v}+\delta_{\boldsymbol{v}})/n^2 \tag{S159}$$

$$(\beta_u+\delta_{\beta_u})(-\boldsymbol{\epsilon}+\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+\boldsymbol{u}\delta_{\beta_u}+\delta_{\boldsymbol{u}}\beta_u+\boldsymbol{v}\delta_{\beta_v}+\delta_{\boldsymbol{v}}\beta_v) \tag{S160}$$

$$+\lambda(d+\delta_d)^2(\boldsymbol{u}+\delta_{\boldsymbol{u}})-\lambda(d+\delta_d)(d\boldsymbol{u}\boldsymbol{v}^\top+\boldsymbol{E})(\boldsymbol{v}+\delta_{\boldsymbol{v}})/n \approx 0 \tag{S161}$$

$$(\beta_v+\delta_{\beta_v})(-\boldsymbol{\epsilon}+\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+\boldsymbol{u}\delta_{\beta_u}+\delta_{\boldsymbol{u}}\beta_u+\boldsymbol{v}\delta_{\beta_v}+\delta_{\boldsymbol{v}}\beta_v) \tag{S162}$$

$$+\lambda(d+\delta_d)^2(\boldsymbol{v}+\delta_{\boldsymbol{v}})-\lambda(d+\delta_d)(d\boldsymbol{v}\boldsymbol{u}^\top+\boldsymbol{E}^\top)(\boldsymbol{u}+\delta_{\boldsymbol{u}})/n \approx 0 \tag{S163}$$

After simplification and dropping the second order terms,

$$\boldsymbol{u}^\top\delta_{\boldsymbol{u}}/n \approx 0, \tag{S164}$$

$$\boldsymbol{v}^\top\delta_{\boldsymbol{v}}/n \approx 0, \tag{S165}$$

$$\boldsymbol{u}^\top\delta_{\boldsymbol{v}}\beta_v+\boldsymbol{u}^\top\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+n\delta_{\beta_u}+\boldsymbol{u}^\top\boldsymbol{v}\delta_{\beta_v} \approx \boldsymbol{u}^\top\boldsymbol{\epsilon}, \tag{S166}$$

$$\boldsymbol{v}^\top\delta_{\boldsymbol{u}}\beta_u+\boldsymbol{v}^\top\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+\boldsymbol{u}^\top\boldsymbol{v}\delta_{\beta_u}+n\delta_{\beta_v} \approx \boldsymbol{v}^\top\boldsymbol{\epsilon}, \tag{S167}$$

$$\delta_{\boldsymbol{\beta}_x} \approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{\epsilon}-\boldsymbol{u}\delta_{\beta_u}-\delta_{\boldsymbol{u}}\beta_u-\boldsymbol{v}\delta_{\beta_v}-\delta_{\boldsymbol{v}}\beta_v), \tag{S168}$$

$$\delta_d \approx d\boldsymbol{u}^\top\delta_{\boldsymbol{u}}+d\boldsymbol{v}^\top\delta_{\boldsymbol{v}}+\boldsymbol{u}^\top\boldsymbol{E}\boldsymbol{v} \approx \boldsymbol{u}^\top\boldsymbol{E}\boldsymbol{v}/n^2, \tag{S169}$$

$$(\beta_u^2+\lambda d^2)\delta_{\boldsymbol{u}}+\beta_u\beta_v\delta_{\boldsymbol{v}}+\beta_u\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+\beta_u\boldsymbol{u}\delta_{\beta_u}+\beta_u\boldsymbol{v}\delta_{\beta_v} \approx \lambda d(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\boldsymbol{E}\boldsymbol{v}/n+\beta_u\boldsymbol{\epsilon}, \tag{S170}$$

$$\beta_u\beta_v\delta_{\boldsymbol{u}}+(\beta_v^2+\lambda d^2)\delta_{\boldsymbol{v}}+\beta_v\boldsymbol{X}\delta_{\boldsymbol{\beta}_x}+\beta_v\boldsymbol{u}\delta_{\beta_u}+\beta_v\boldsymbol{v}\delta_{\beta_v} \approx \lambda d(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\boldsymbol{E}^\top\boldsymbol{u}/n+\beta_v\boldsymbol{\epsilon} \tag{S171}$$

where (S166)-(S167) have used (S164)-(S165).

Plugging (S168) into (S166)-(S167) and using (S164)-(S165)

$$\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta_{\boldsymbol{u}}\beta_u+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta_{\boldsymbol{v}}\beta_v+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\delta_{\beta_u}+\boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}\delta_{\beta_v} \tag{S172}$$

$$\approx \boldsymbol{u}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon}, \tag{S173}$$

$$\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta_{\boldsymbol{u}}\beta_u+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\delta_{\boldsymbol{v}}\beta_v+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\delta_{\beta_u}+\boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}\delta_{\beta_v} \tag{S174}$$

$$\approx \boldsymbol{v}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\epsilon}. \tag{S175}$$

Recall $\tilde{\boldsymbol{u}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}$, $\tilde{\boldsymbol{v}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v}$ and $\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} = \begin{pmatrix} \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} & \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} \\ \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} & \tilde{\boldsymbol{v}}^\top\tilde{\boldsymbol{v}} \end{pmatrix}$. Then (S172)-(S175) can be written as

$$\begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} (-\beta_u \boldsymbol{I}_n \quad -\beta_v \boldsymbol{I}_n) \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} + \boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} \begin{pmatrix} \delta_{\beta_u} \\ \delta_{\beta_v} \end{pmatrix} \approx \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} \boldsymbol{\epsilon}. \tag{S176}$$

Solving for δ_{β_u}, δ_{β_v} gives

$$\begin{pmatrix} \delta_{\beta_u} \\ \delta_{\beta_v} \end{pmatrix} \approx \boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} (-\beta_u \boldsymbol{I}_n \quad -\beta_v \boldsymbol{I}_n \quad \boldsymbol{I}_n) \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \\ \boldsymbol{\epsilon} \end{pmatrix}. \tag{S177}$$

On the other hand, plugging (S168) into (S170, S171)

$$\begin{pmatrix} \beta_u^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) + \lambda d^2 \boldsymbol{I} & \beta_u\beta_v(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) \\ \beta_u\beta_v(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) & \beta_v^2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) + \lambda d^2 \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} + \begin{pmatrix} \beta_u \\ \beta_v \end{pmatrix} (\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}}) \begin{pmatrix} \delta_{\beta_u} \\ \delta_{\beta_v} \end{pmatrix} \tag{S178}$$

$$\approx \begin{pmatrix} \beta_u(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) & \lambda d \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})/n \\ \beta_v(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) & \lambda d\,(\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})/n)\,\boldsymbol{K} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix}. \tag{S179}$$

Combining (S177)-(S179) and after some algebra, we have

$$\begin{pmatrix} \beta_u^2\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) + \lambda d^2 \boldsymbol{I} & \beta_u\beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \\ \beta_u\beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \beta_v^2\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) + \lambda d^2 \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} \tag{S180}$$

$$\approx \left[\lambda d^2 \boldsymbol{I}_{2n} + \begin{pmatrix} \beta_u^2\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \beta_u\beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \\ \beta_u\beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \beta_v^2\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \end{pmatrix} \right] \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} \tag{S181}$$

$$\approx \left[\lambda d^2 \boldsymbol{I}_{2n} + \begin{pmatrix} \beta_u^2 & \beta_u\beta_v \\ \beta_u\beta_v & \beta_v^2 \end{pmatrix} \otimes \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \right] \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} \tag{S182}$$

$$\approx \begin{pmatrix} \beta_u\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \lambda d \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})/n \\ \beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \lambda d\,(\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})/n)\,\boldsymbol{K} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix} \tag{S183}$$

because $(\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}})\,\boldsymbol{C}^{-1}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} = \boldsymbol{P}_{(\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}})}$ and $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{(\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}})} = \boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}$.

Using the Woodbury identity and the property of projection matrix,

$$\left(\lambda d^2 \boldsymbol{I}_{2n} + \begin{pmatrix} \beta_u^2 & \beta_u\beta_v \\ \beta_u\beta_v & \beta_v^2 \end{pmatrix} \otimes \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)^{-1} \tag{S184}$$

$$= (\lambda d^2)^{-1} \left(\boldsymbol{I}_{2n} - (\lambda d^2 + \beta_u^2 + \beta_v^2)^{-1} \begin{pmatrix} \beta_u^2 & \beta_u\beta_v \\ \beta_u\beta_v & \beta_v^2 \end{pmatrix} \otimes \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right). \tag{S185}$$

Hence,

$$\begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \end{pmatrix} \approx (\lambda d^2)^{-1} \left(\boldsymbol{I}_{2n} - (\lambda d^2 + \beta_u^2 + \beta_v^2)^{-1} \begin{pmatrix} \beta_u^2 & \beta_u\beta_v \\ \beta_u\beta_v & \beta_v^2 \end{pmatrix} \otimes \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right) \tag{S186}$$

$$\begin{pmatrix} \beta_u\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \lambda d \boldsymbol{v}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}})/n \\ \beta_v\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) & \lambda d\,(\boldsymbol{u}^\top \otimes (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}})/n)\,\boldsymbol{K} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix} \tag{S187}$$

$$\stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{11}\, \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\, \boldsymbol{C}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix}. \tag{S188}$$

Let $\widehat{\boldsymbol{A}} = \hat{d}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\top$. Plugging in (S169) and (S188), the first order expansion leads to

$$\begin{aligned}
\text{vec}\left(\delta_{\boldsymbol{A}}\right) &= \text{vec}(\hat{d}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\top) - \text{vec}(d\boldsymbol{u}\boldsymbol{v}^\top) && \text{(S189)}\\
&= \text{vec}(d+\delta_d)(\boldsymbol{u}+\delta_{\boldsymbol{u}})(\boldsymbol{v}+\delta_{\boldsymbol{v}})^\top - \text{vec}(d\boldsymbol{u}\boldsymbol{v}^\top) && \text{(S190)}\\
&= d\,\text{vec}(\boldsymbol{u}\delta_{\boldsymbol{v}}^\top) + d\,\text{vec}(\delta_{\boldsymbol{u}}\boldsymbol{v}^\top) + \boldsymbol{u}^\top\boldsymbol{E}\boldsymbol{v}\,\text{vec}(\boldsymbol{u}\boldsymbol{v}^\top)/n^2 && \text{(S191)}\\
&= d\boldsymbol{K}\,\text{vec}(\delta_{\boldsymbol{v}}\boldsymbol{u}^\top) + d\,\text{vec}(\delta_{\boldsymbol{u}}\boldsymbol{v}^\top) + \boldsymbol{P}\boldsymbol{v}\otimes\boldsymbol{P}\boldsymbol{u} && \text{(S192)}\\
&= d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I}_n)\delta_{\boldsymbol{v}} + d(\boldsymbol{v}\otimes\boldsymbol{I}_n)\delta_{\boldsymbol{u}} + \boldsymbol{P}\boldsymbol{v}\otimes\boldsymbol{P}\boldsymbol{u} && \text{(S193)}\\
&= \begin{pmatrix} d(\boldsymbol{v}\otimes\boldsymbol{I}_n) & d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I}_n) & \boldsymbol{P}\boldsymbol{v}\otimes\boldsymbol{P}\boldsymbol{u} \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{0}_{n^2\times n} & \boldsymbol{I}_{n^2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix} && \text{(S194)}\\
&\stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{31}\, \boldsymbol{C}_{32} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix}. && \text{(S195)}
\end{aligned}$$

For δ_{β_u} and δ_{β_v}, plugging (S188) into (S177)

$$\begin{aligned}
\begin{pmatrix} \delta_{\beta_u} \\ \delta_{\beta_v} \end{pmatrix} &\approx \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} \begin{pmatrix} -\beta_u\boldsymbol{I}_n & -\beta_v\boldsymbol{I}_n & \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \\ \boldsymbol{\epsilon} \end{pmatrix} && \text{(S196)}\\
&= \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} \begin{pmatrix} -\beta_u\boldsymbol{I}_n & -\beta_v\boldsymbol{I}_n & \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix} && \text{(S197)}\\
&\stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{41}\, \boldsymbol{C}_{42} \\ \boldsymbol{C}_{51}\, \boldsymbol{C}_{52} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix}. && \text{(S198)}
\end{aligned}$$

Lastly for $\delta_{\boldsymbol{\beta}_x}$, we plug (S188) and (S198) into (S168)

$$\begin{aligned}
\delta_{\boldsymbol{\beta}_x} &\approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{\epsilon} - \boldsymbol{u}\delta_{\beta_u} - \delta_{\boldsymbol{u}}\beta_u - \boldsymbol{v}\delta_{\beta_v} - \delta_{\boldsymbol{v}}\beta_v) && \text{(S199)}\\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \begin{pmatrix} -\beta_u\boldsymbol{I}_n & -\beta_v\boldsymbol{I}_n & -\boldsymbol{u} & -\boldsymbol{v} & \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \delta_{\boldsymbol{u}} \\ \delta_{\boldsymbol{v}} \\ \delta_{\beta_u} \\ \delta_{\beta_v} \\ \boldsymbol{\epsilon} \end{pmatrix} && \text{(S200)}\\
&\approx (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \begin{pmatrix} -\beta_u\boldsymbol{I}_n & -\beta_v\boldsymbol{I}_n & -\boldsymbol{u} & -\boldsymbol{v} & \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \\ \boldsymbol{C}_{31} & \boldsymbol{C}_{32} \\ \boldsymbol{C}_{41} & \boldsymbol{C}_{42} \\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix} && \text{(S201)}\\
&\stackrel{def}{=} \begin{pmatrix} \boldsymbol{C}_{51}\, \boldsymbol{C}_{52} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \text{vec}(\boldsymbol{E}) \end{pmatrix}. && \text{(S202)}
\end{aligned}$$

Finally, putting (S188), (S195), (S198) and (S202) together

$$\begin{pmatrix} \hat{\boldsymbol{u}}-\boldsymbol{u} \\ \hat{\boldsymbol{v}}-\boldsymbol{v} \\ \mathrm{vec}\left(\widehat{\boldsymbol{A}}-\boldsymbol{A}_0\right) \\ \hat{\beta}_u-\beta_u \\ \hat{\beta}_v-\beta_v \\ \hat{\boldsymbol{\beta}}_x-\boldsymbol{\beta}_x \end{pmatrix} = \begin{pmatrix} \delta\boldsymbol{u} \\ \delta\boldsymbol{v} \\ \mathrm{vec}\left(\delta\boldsymbol{A}\right) \\ \delta_{\beta_u} \\ \delta_{\beta_v} \\ \delta\boldsymbol{\beta}_x \end{pmatrix} \approx \begin{pmatrix} \boldsymbol{C}_{11}\ \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\ \boldsymbol{C}_{22} \\ \boldsymbol{C}_{31}\ \boldsymbol{C}_{32} \\ \boldsymbol{C}_{41}\ \boldsymbol{C}_{42} \\ \boldsymbol{C}_{51}\ \boldsymbol{C}_{52} \\ \boldsymbol{C}_{61}\ \boldsymbol{C}_{62} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix} = \boldsymbol{C} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix}. \tag{S203}$$

Recall that we assume

$$\begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix} \sim N\left(\boldsymbol{0}_{(n+n^2)\times 1}, \begin{pmatrix} \sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2} \end{pmatrix}\right). \tag{S204}$$

Therefore, the SuperCENT estimators converge to the following normal distribution asymptotically,

$$\begin{pmatrix} \hat{\boldsymbol{u}}-\boldsymbol{u} \\ \hat{\boldsymbol{v}}-\boldsymbol{v} \\ \mathrm{vec}\left(\widehat{\boldsymbol{A}}-\boldsymbol{A}_0\right) \\ \hat{\beta}_u-\beta_u \\ \hat{\beta}_v-\beta_v \\ \hat{\boldsymbol{\beta}}_x-\boldsymbol{\beta}_x \end{pmatrix} \xrightarrow{\mathcal{D}} N\Big(\boldsymbol{0}_{(2n+n^2+2+p)\times 1}, \boldsymbol{C}\begin{pmatrix} \sigma_y^2\boldsymbol{I}_n & \boldsymbol{0}_{n\times n^2} \\ \boldsymbol{0}_{n^2\times n} & \sigma_a^2\boldsymbol{I}_{n^2} \end{pmatrix}\boldsymbol{C}^\top\Big). \tag{S205}$$

Q.E.D.

S6.2.2. *Proof of Corollary 4*

PROOF: We first show the rate of $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{v}}$ then $\widehat{\boldsymbol{A}}$.

(1) Rate of $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{v}}$. From Theorem 2, we have

$$\begin{pmatrix} \delta\boldsymbol{u} \\ \delta\boldsymbol{v} \end{pmatrix} \approx \begin{pmatrix} \boldsymbol{C}_{11}\ \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21}\ \boldsymbol{C}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathrm{vec}(\boldsymbol{E}) \end{pmatrix} \tag{S206}$$

where

$$\begin{pmatrix} \boldsymbol{C}_{11} \\ \boldsymbol{C}_{21} \end{pmatrix} = (\lambda d^2+\beta_u^2+\beta_v^2)^{-1} \begin{pmatrix} \beta_u \\ \beta_v \end{pmatrix} \otimes \left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \tag{S207}$$

and

$$\begin{pmatrix} \boldsymbol{C}_{12} \\ \boldsymbol{C}_{22} \end{pmatrix} = \frac{1}{dn}\left[\begin{pmatrix} \boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}) \\ (\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))\boldsymbol{K} \end{pmatrix}\right. \tag{S208}$$

$$\left. -\frac{1}{(\lambda d^2+\beta_u^2+\beta_v^2)} \begin{pmatrix} \beta_u^2\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\beta_u\beta_v\left(\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K} \\ \beta_u\beta_v\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\beta_v^2\left(\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K} \end{pmatrix}\right] \tag{S209}$$

For the rate of $\hat{\boldsymbol{u}}$,

$$\mathbb{E}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2 \approx \mathrm{tr}(\sigma_y^2\boldsymbol{C}_{11}\boldsymbol{C}_{11}^\top+\sigma_a^2\boldsymbol{C}_{12}\boldsymbol{C}_{12}^\top) \tag{S210}$$

$$= \frac{\sigma_y^2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\mathrm{tr}\left(\beta_u^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right) \tag{S211}$$

$$+\frac{\sigma_a^2}{n^2d^2}\mathrm{tr}\left(\boldsymbol{v}^\top\boldsymbol{v}\right)\mathrm{tr}\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}\right) \tag{S212}$$

$$+\frac{\sigma_a^2}{n^2d^2(\lambda d^2+\beta_u^2+\beta_v^2)^2}\mathrm{tr}\Bigg(\beta_u{}^4\boldsymbol{v}^\top\boldsymbol{v}\Big(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\Big)+ \tag{S213}$$

$$2\beta_u^3\beta_v\boldsymbol{u}\boldsymbol{v}^\top\otimes\Big(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\Big)\boldsymbol{K}^\top+\beta_u^2\beta_v^2\boldsymbol{u}\boldsymbol{u}^\top\otimes\Big(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\Big)\Bigg) \tag{S214}$$

$$-\frac{2\sigma_a^2}{nd^2(\lambda d^2+\beta_u^2+\beta_v^2)}\mathrm{tr}\Bigg(\beta_u^2/n\Big(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{u})\Big)\Big(\boldsymbol{v}\otimes\Big(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\Big)\Big) \tag{S215}$$

$$\beta_u\beta_v/n\Big(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{u})\Big)\Big(\boldsymbol{K}^\top\boldsymbol{u}\otimes\Big(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\Big)\Big)\Bigg) \tag{S216}$$

$$=\frac{\sigma_y^2\beta_u^2(n-p-2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}+\left(\frac{\sigma_a^2}{d^2}-\frac{\sigma_a^2}{d^2n}\right) \tag{S217}$$

$$+\frac{\sigma_a^2\beta_u^2(n-p-2)}{n^2d^2(\lambda d^2+\beta_u^2+\beta_v^2)^2}\Big(\beta_u^2n+2\beta_u\beta_v\mathrm{tr}(\boldsymbol{v}^\top\boldsymbol{u})+\beta_v^2n\Big) \tag{S218}$$

$$-\frac{2\sigma_a^2(n-p-2)}{nd^2(\lambda d^2+\beta_u^2+\beta_v^2)}\beta_u^2 \tag{S219}$$

$$=\left(\frac{\sigma_a^2}{d^2}-\frac{\sigma_a^2}{d^2n}\right) \tag{S220}$$

$$-\frac{\beta_u^2(n-p-2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\frac{2\lambda d^2+\beta_u^2+\beta_v^2}{d^2n}\sigma_a^2-\sigma_y^2\right]. \tag{S221}$$

Then,

$$\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2\approx\frac{\sigma_a^2(n-1)}{d^2n^2}-\frac{\beta_u^2(n-p-2)}{n(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\frac{2\lambda d^2+\beta_u^2+\beta_v^2}{d^2n}\sigma_a^2-\sigma_y^2\right] \tag{S222}$$

$$=O\left(\frac{\sigma_a^2}{d^2n}-\beta_u^2\delta_{ts,sc}\right) \tag{S223}$$

where $\delta_{ts,sc}=\frac{1}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\frac{2\lambda d^2+\beta_u^2+\beta_v^2}{d^2n}\sigma_a^2-\sigma_y^2\right]$.

To get the optimal λ in Remark 11, we take the partial derivative of $\ell_u\triangleq$ (S221) with respect to λ yields

$$\frac{\partial\ell_u}{\partial\lambda}=\frac{\beta_u^2(n-p-2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^3}\left[2d^2\sigma_y^2+\frac{\sigma_a^2}{n}\Big(2(\lambda d^2+\beta_u^2+\beta_v^2)-4\lambda d^2-2\beta_u^2-2\beta_v^2\Big)\right] \tag{S224}$$

$$=\frac{\beta_u^2(n-p-2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^3}\left[2d^2\sigma_y^2-\frac{2d^2}{n}\sigma_a^2\lambda\right]. \tag{S225}$$

Setting $\frac{\partial\ell_u}{\partial\lambda}=0$ yields

$$\lambda_0=\frac{n\sigma_y^2}{\sigma_a^2}. \tag{S226}$$

When $\lambda\in(0,\lambda_0]$, ℓ_u increases as λ increases; $\lambda\in(\lambda_0,\infty)$, ℓ_u decreases and converges to 0 as λ increases. The maximum of ℓ_u is then taken at λ_0.

Similarly, we derive the rate of $\hat{\boldsymbol{v}}$ as

$$\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{v}}-\boldsymbol{v}\|^2 \approx \frac{\sigma_a^2(n-1)}{d^2n^2} - \frac{\beta_v^2(n-p-2)}{n(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\frac{2\lambda d^2+\beta_u^2+\beta_v^2}{d^2n}\sigma_a^2-\sigma_y^2\right] \tag{S227}$$

$$= O\left(\frac{\sigma_a^2}{d^2n} - \beta_v^2\delta_{ts,sc}\right). \tag{S228}$$

(2) Rate of $\widehat{\boldsymbol{A}}$. From Theorem 2, we have

$$\operatorname{vec}\left(\delta_{\boldsymbol{A}}\right) \approx d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I}_n)\delta\boldsymbol{v} + d(\boldsymbol{v}\otimes\boldsymbol{I}_n)\delta\boldsymbol{u} + \boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}. \tag{S229}$$

Using (S188), we have

$$\operatorname{vec}\left(\delta_{\boldsymbol{A}}\right) \tag{S230}$$

$$\approx \frac{d}{\lambda d^2+\beta_u^2+\beta_v^2}\left(\beta_v\boldsymbol{K}\boldsymbol{u}\otimes\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}+\beta_u\boldsymbol{v}\right)\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{\epsilon} \tag{S231}$$

$$+\Bigg\{[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}+\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})] \tag{S232}$$

$$-\frac{1}{\lambda d^2+\beta_u^2+\beta_v^2}\left[\beta_v^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\otimes\boldsymbol{P}_{\boldsymbol{u}}+\beta_u^2\boldsymbol{P}_{\boldsymbol{v}}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right] \tag{S233}$$

$$-\frac{\beta_u\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\left[\boldsymbol{K}\boldsymbol{u}\boldsymbol{v}^\top/n\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\boldsymbol{v}\boldsymbol{u}^\top/n\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{K}\right] \tag{S234}$$

$$+\boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}}\Bigg\}\operatorname{vec}(\boldsymbol{E}) \tag{S235}$$

$$\overset{def}{=} \boldsymbol{H}_1\boldsymbol{\epsilon}+(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4+\boldsymbol{H}_5)\operatorname{vec}(\boldsymbol{E}) \tag{S236}$$

because

$$d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})\delta\boldsymbol{v} \tag{S237}$$

$$= d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})[\boldsymbol{C}_{21}\boldsymbol{\epsilon}+\boldsymbol{C}_{22}\operatorname{vec}(\boldsymbol{E})] \tag{S238}$$

$$= d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})\frac{\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{\epsilon} \tag{S239}$$

$$+d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})\frac{1}{dn}\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\boldsymbol{K}\operatorname{vec}(\boldsymbol{E}) \tag{S240}$$

$$-d\boldsymbol{K}(\boldsymbol{u}\otimes\boldsymbol{I})\frac{1}{dn}\frac{1}{(\lambda d^2+\beta_u^2+\beta_v^2)}\Bigg[\beta_u\beta_v\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+ \tag{S241}$$

$$\beta_v^2\left(\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}\Bigg]\operatorname{vec}(\boldsymbol{E}) \tag{S242}$$

$$= \frac{d\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\boldsymbol{K}\left(\boldsymbol{u}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{\epsilon} \tag{S243}$$

$$+\left[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}-\frac{\beta_v^2}{\lambda d^2+\beta_u^2+\beta_v^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\otimes\boldsymbol{P}_{\boldsymbol{u}}\right]\operatorname{vec}(\boldsymbol{E}) \tag{S244}$$

$$-\frac{\beta_u\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\boldsymbol{K}\left(\boldsymbol{u}\boldsymbol{v}^\top/n\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\operatorname{vec}(\boldsymbol{E}) \tag{S245}$$

and similarly

$$d(\boldsymbol{v}\otimes\boldsymbol{I})\delta\boldsymbol{u} \tag{S246}$$

$$= \frac{d\beta_u}{\lambda d^2+\beta_u^2+\beta_v^2}\boldsymbol{v}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{\epsilon} \tag{S247}$$

$$+\left[\boldsymbol{P}\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{u})-\frac{\beta_u^2}{\lambda d^2+\beta_u^2+\beta_v^2}\boldsymbol{P}\boldsymbol{v}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right]\mathrm{vec}(\boldsymbol{E}) \tag{S248}$$

$$-\frac{\beta_u\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\left(\boldsymbol{v}\boldsymbol{u}^\top/n\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}\,\mathrm{vec}(\boldsymbol{E}). \tag{S249}$$

In order to obtain the rate of $\widehat{\boldsymbol{A}}$, we need $\boldsymbol{H}_1\boldsymbol{H}_1^\top$ and $(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4+\boldsymbol{H}_5)(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4+\boldsymbol{H}_5)^\top$:

$$\boldsymbol{H}_1\boldsymbol{H}_1^\top = \frac{d^2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\Big\{\beta_v^2\boldsymbol{K}\left(\boldsymbol{u}\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}^\top+\beta_u^2\boldsymbol{v}\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right) \tag{S250}$$

$$+\beta_u\beta_v\boldsymbol{K}\left(\boldsymbol{u}\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)+\beta_u\beta_v\left(\boldsymbol{v}\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}^\top\Big\} \tag{S251}$$

$$\boldsymbol{H}_2\boldsymbol{H}_2^\top = (\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\otimes\boldsymbol{P}_{\boldsymbol{u}}+\boldsymbol{P}_{\boldsymbol{v}}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}) \tag{S252}$$

$$\boldsymbol{H}_3\boldsymbol{H}_3^\top = \frac{1}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[{\beta_u}^4\boldsymbol{P}_{\boldsymbol{v}}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+{\beta_v}^4\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\otimes\boldsymbol{P}_{\boldsymbol{u}}\right] \tag{S253}$$

$$\boldsymbol{H}_4\boldsymbol{H}_4^\top = \frac{\beta_u^2\beta_v^2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\boldsymbol{K}\boldsymbol{P}_{\boldsymbol{u}}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{K}^\top+\boldsymbol{P}_{\boldsymbol{v}}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right] \tag{S254}$$

$$\boldsymbol{H}_5\boldsymbol{H}_5^\top = \boldsymbol{P}_{\boldsymbol{v}}\otimes\boldsymbol{P}_{\boldsymbol{u}} \tag{S255}$$

$$\boldsymbol{H}_2\boldsymbol{H}_3^\top = \boldsymbol{H}_3\boldsymbol{H}_2^\top \tag{S256}$$

$$= -\frac{1}{\lambda d^2+\beta_u^2+\beta_v^2}\left[\beta_u^2\boldsymbol{P}_{\boldsymbol{v}}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\beta_v^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\otimes\boldsymbol{P}_{\boldsymbol{u}}\right] \tag{S257}$$

$$\boldsymbol{H}_2\boldsymbol{H}_4^\top = \boldsymbol{H}_4\boldsymbol{H}_2^\top \tag{S258}$$

$$= -\frac{\beta_u\beta_v}{\lambda d^2+\beta_u^2+\beta_v^2}\left[\boldsymbol{K}\left(\frac{\boldsymbol{u}\boldsymbol{v}^\top}{n}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)+\left(\frac{\boldsymbol{v}\boldsymbol{u}^\top}{n}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}^\top\right] \tag{S259}$$

$$\boldsymbol{H}_3\boldsymbol{H}_4^\top = \frac{\beta_u\beta_v}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\beta_u^2\left(\frac{\boldsymbol{v}\boldsymbol{u}^\top}{n}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\boldsymbol{K}^\top+\beta_v^2\boldsymbol{K}\left(\frac{\boldsymbol{u}\boldsymbol{v}^\top}{n}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)\right] \tag{S260}$$

$$\boldsymbol{H}_4\boldsymbol{H}_3^\top = \frac{\beta_u\beta_v}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[\beta_u^2\boldsymbol{K}\left(\frac{\boldsymbol{u}\boldsymbol{v}^\top}{n}\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right)+\beta_v^2\boldsymbol{K}\left(\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\otimes\frac{\boldsymbol{v}\boldsymbol{u}^\top}{n}\right)\right] \tag{S261}$$

$$\boldsymbol{H}_2\boldsymbol{H}_5^\top = \boldsymbol{H}_3\boldsymbol{H}_5^\top = \boldsymbol{H}_4\boldsymbol{H}_5^\top = \boldsymbol{H}_5\boldsymbol{H}_2^\top = \boldsymbol{H}_5\boldsymbol{H}_3^\top = \boldsymbol{H}_5\boldsymbol{H}_4^\top = 0. \tag{S262}$$

Plugging in (S251) to (S262) yields the rate of $\widehat{\boldsymbol{A}}$ as follows.

$$\mathbb{E}\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_0\right\|_F^2 \tag{S263}$$

$$\approx \sigma_y^2\mathrm{tr}\left(\boldsymbol{H}_1\boldsymbol{H}_1^\top\right)+\sigma_a^2\mathrm{tr}\left((\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4+\boldsymbol{H}_5)(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4+\boldsymbol{H}_5)^\top\right) \tag{S264}$$

$$= \sigma_y^2\frac{d^2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left(\beta_u^2+\beta_v^2\right)n(n-p-2) \tag{S265}$$

$$+\sigma_a^2\left[2(n-1)+1\right] \tag{S266}$$

$$+\sigma_a^2\frac{n-p-2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left[{\beta_u}^4+{\beta_v}^4+2\beta_u^2\beta_v^2-2(\beta_u^2+\beta_v^2)(\lambda d^2+\beta_u^2+\beta_v^2)\right] \tag{S267}$$

$$=\sigma_a^2(2n-1) \tag{S268}$$

$$-\frac{n-p-2}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}(\beta_u^2+\beta_v^2)\left[(2\lambda d^2+\beta_u^2+\beta_v^2)\sigma_a^2-nd^2\sigma_y^2\right]. \tag{S269}$$

Therefore,

$$\mathbb{E}\frac{\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_0\right\|_F^2}{\|\boldsymbol{A}_0\|_F^2}\approx\frac{\sigma_a^2(2n-1)}{d^2n}-\frac{n-p-2}{n(\lambda d^2+\beta_u^2+\beta_v^2)^2}(\beta_u^2+\beta_v^2)\left[\frac{(2\lambda d^2+\beta_u^2+\beta_v^2)}{d^2n}\sigma_a^2-\sigma_y^2\right] \tag{S270}$$

$$=O\left(\frac{\sigma_a^2}{d^2n}-(\beta_u^2+\beta_v^2)\delta_{ts,sc}\right). \tag{S271}$$

S6.2.3. *Proof of Corollary 5*

We now prove the rate of $\hat{\beta}$ of SuperCENT.

(1) Rate of $\hat{\beta}_u$ and $\hat{\beta}_v$. From Theorem 2, we have

$$\boldsymbol{\epsilon}-(\delta_{\boldsymbol{u}}\beta_u+\delta_{\boldsymbol{v}}\beta_v) \tag{S272}$$

$$=(\boldsymbol{I}-\beta_u\boldsymbol{C}_{11}+\beta_v\boldsymbol{C}_{12})\boldsymbol{\epsilon}+(\beta_u\boldsymbol{C}_{12}+\beta_v\boldsymbol{C}_{22})\,\mathrm{vec}(\boldsymbol{E}) \tag{S273}$$

$$=\left[\boldsymbol{I}-\frac{\beta_u^2+\beta_v^2}{\lambda d^2+\beta_u^2+\beta_v^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\right)\right]\boldsymbol{\epsilon} \tag{S274}$$

$$-\frac{1}{dn}\left[\beta_u\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v\left(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right)\boldsymbol{K}\right]\mathrm{vec}(\boldsymbol{E}) \tag{S275}$$

$$+\frac{\beta_u^2+\beta_v^2}{dn(\lambda d^2+\beta_u^2+\beta_v^2)}\left[\beta_u\boldsymbol{v}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\right)+\right. \tag{S276}$$

$$\left.\beta_v\left(\boldsymbol{u}^\top\otimes\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\right)\boldsymbol{K}\right)\right]\mathrm{vec}(\boldsymbol{E}) \tag{S277}$$

$$\stackrel{def}{=}\boldsymbol{A}_1\boldsymbol{\epsilon}+(\boldsymbol{C}_1+\boldsymbol{C}_2)\,\mathrm{vec}(\boldsymbol{E}). \tag{S278}$$

Plugging (S278) into (S177), we obtain

$$\begin{pmatrix}\delta_{\beta_u}\\ \delta_{\beta_v}\end{pmatrix}\approx\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\boldsymbol{\epsilon}-(\delta_{\boldsymbol{u}}\beta_u+\delta_{\boldsymbol{v}}\beta_v)\right] \tag{S279}$$

$$=\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\boldsymbol{A}_1\boldsymbol{\epsilon}+(\boldsymbol{C}_1+\boldsymbol{C}_2)\,\mathrm{vec}(\boldsymbol{E})\right] \tag{S280}$$

$$\stackrel{def}{=}\boldsymbol{B}_1\boldsymbol{\epsilon}+\boldsymbol{B}_2\,\mathrm{vec}(\boldsymbol{E}). \tag{S281}$$

To get the rate of $\hat{\beta}_u$ and $\hat{\beta}_v$, we next calculate $\boldsymbol{B}_1\boldsymbol{B}_1^\top$ and $\boldsymbol{B}_2\boldsymbol{B}_2^\top$.
(a) $\boldsymbol{B}_1\boldsymbol{B}_1^\top$.
Since

$$\boldsymbol{A}_1\boldsymbol{A}_1^\top=\boldsymbol{I}-\frac{(\beta_u^2+\beta_v^2)(2\lambda d^2+\beta_u^2+\beta_v^2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{Xuv})}\right) \tag{S282}$$

and

$$\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\tilde{\boldsymbol{u}}=\boldsymbol{0}, \tag{S283}$$

consequently

$$\boldsymbol{B}_1\boldsymbol{B}_1^\top=\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\boldsymbol{A}_1\boldsymbol{A}_1^\top\left(\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}=\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}. \tag{S284}$$

(b) $\boldsymbol{B}_2\boldsymbol{B}_2^\top$.
Since

$$\boldsymbol{C}_1\boldsymbol{C}_1^\top \tag{S285}$$

$$=\frac{1}{(dn)^2}\Big[\beta_u^2(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))+\beta_v^2(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})) \tag{S286}$$

$$+\beta_u\beta_v(\boldsymbol{v}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\boldsymbol{K}^\top(\boldsymbol{u}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))+\beta_u\beta_v(\boldsymbol{u}^\top\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))\boldsymbol{K}(\boldsymbol{v}\otimes(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}}))\Big] \tag{S287}$$

$$=\frac{1}{d^2n}\left[\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right], \tag{S288}$$

$$\boldsymbol{C}_2\boldsymbol{C}_2^\top=n\left(\frac{\beta_u^2+\beta_v^2}{dn(\lambda d^2+\beta_u^2+\beta_v^2)}\right)^2\left[\beta_u^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\beta_v^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right] \tag{S289}$$

and

$$\boldsymbol{C}_1\boldsymbol{C}_2^\top=\boldsymbol{C}_2\boldsymbol{C}_1^\top=-\frac{1}{d^2n}\frac{\beta_u^2+\beta_v^2}{\lambda d^2+\beta_u^2+\beta_v^2}\left[\beta_u^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)+\beta_v^2\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right], \tag{S290}$$

consequently

$$(\boldsymbol{C}_1+\boldsymbol{C}_2)(\boldsymbol{C}_1+\boldsymbol{C}_2)^\top \tag{S291}$$

$$=\frac{1}{d^2n}\left[(\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}}))-\frac{(\beta_u^2+\beta_v^2)^2(2\lambda d^2+\beta_u^2+\beta_v^2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\right] \tag{S292}$$

Therefore,

$$\boldsymbol{B}_2\boldsymbol{B}_2^\top=\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}(\boldsymbol{C}_1+\boldsymbol{C}_2)(\boldsymbol{C}_1+\boldsymbol{C}_2)^\top\left(\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S293}$$

$$=\frac{1}{d^2n}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left(\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right)\left(\tilde{\boldsymbol{u}}\ \tilde{\boldsymbol{v}}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S294}$$

$$=\frac{1}{d^2n}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S295}$$

where the first equality is due to (S283).
Combining (S284) and (S295),

$$Cov\begin{pmatrix}\delta_{\beta_u}\\ \delta_{\beta_v}\end{pmatrix}\approx\sigma_y^2\boldsymbol{B}_1\boldsymbol{B}_1^\top+\sigma_a^2\boldsymbol{B}_2\boldsymbol{B}_2^\top \tag{S296}$$

$$=\sigma_y^2\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S297}$$

$$+\sigma_a^2\frac{1}{d^2n}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\tilde{\boldsymbol{u}}^\top\\ \tilde{\boldsymbol{v}}^\top\end{pmatrix}\left[\beta_u^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{u}})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{v}})\right]\begin{pmatrix}\tilde{\boldsymbol{u}}\\ \tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \tag{S298}$$

$$= \sigma_y^2 \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} + \sigma_a^2 \frac{1}{d^2 n} \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \beta_v^2 \tilde{\boldsymbol{u}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) \tilde{\boldsymbol{u}} & 0 \\ 0 & \beta_u^2 \tilde{\boldsymbol{v}}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) \tilde{\boldsymbol{v}} \end{pmatrix} \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \quad \text{(S299)}$$

where is the same as $Cov \begin{pmatrix} \delta_{\beta_u}^{ts} \\ \delta_{\beta_v}^{ts} \end{pmatrix}$ in (S120).

Therefore,

$$\mathbb{E}(\hat{\beta}_u - \beta_u)^2 = \mathbb{E}(\hat{\beta}_u^{ts} - \beta_u)^2 \quad \text{and} \quad \mathbb{E}(\hat{\beta}_v - \beta_v)^2 = \mathbb{E}(\hat{\beta}_v^{ts} - \beta_v)^2. \quad \text{(S300)}$$

(2) Rate of $\hat{\boldsymbol{\beta}}_x$. Recall that

$$\tilde{\boldsymbol{G}} = (\boldsymbol{u}\ \boldsymbol{v})\, \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} \quad \text{and} \quad \boldsymbol{G} = (\boldsymbol{u}\ \boldsymbol{v})\, \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \boldsymbol{u}^\top \\ \boldsymbol{v}^\top \end{pmatrix}. \quad \text{(S301)}$$

By plugging (S188) and (S198) into (S168), we have

$$\begin{aligned} \delta_{\boldsymbol{\beta}_x} &\approx (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{\epsilon} - \boldsymbol{u}\delta_{\beta_u} - \delta_{\boldsymbol{u}}\beta_u - \boldsymbol{v}\delta_{\beta_v} - \delta_{\boldsymbol{v}}\beta_v) && \text{(S302)} \\ &\approx (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \left[\boldsymbol{I} - (\boldsymbol{u}\ \boldsymbol{v})\, \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \tilde{\boldsymbol{u}}^\top \\ \tilde{\boldsymbol{v}}^\top \end{pmatrix} \right] [\boldsymbol{A}_1 \boldsymbol{\epsilon} + (\boldsymbol{C}_1 + \boldsymbol{C}_2)\, \mathrm{vec}(\boldsymbol{E})] && \text{(S303)} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{I} - \tilde{\boldsymbol{G}})\, [\boldsymbol{A}_1 \boldsymbol{\epsilon} + (\boldsymbol{C}_1 + \boldsymbol{C}_2)\, \mathrm{vec}(\boldsymbol{E})] && \text{(S304)} \\ &\stackrel{def}{=} \boldsymbol{F}_1 \boldsymbol{\epsilon} + \boldsymbol{F}_2\, \mathrm{vec}(\boldsymbol{E}). && \text{(S305)} \end{aligned}$$

To get the rate of $\hat{\boldsymbol{\beta}}_x$, we next calculate $\boldsymbol{F}_1 \boldsymbol{F}_1^\top$ and $\boldsymbol{F}_2 \boldsymbol{F}_2^\top$.
(a) $\boldsymbol{F}_1 \boldsymbol{F}_1^\top$.
Since

$$(\boldsymbol{I} - \tilde{\boldsymbol{G}})(\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top = \boldsymbol{I} - \tilde{\boldsymbol{G}} - \tilde{\boldsymbol{G}}^\top + \boldsymbol{G} \quad \text{(S306)}$$

and

$$\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right) \tilde{\boldsymbol{u}} = \boldsymbol{0} \quad \text{and} \quad \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right) \boldsymbol{u} = \boldsymbol{0}, \quad \text{(S307)}$$

consequently

$$(\boldsymbol{I} - \tilde{\boldsymbol{G}}) \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top = (\boldsymbol{I} - \tilde{\boldsymbol{G}})(\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top - \frac{(\beta_u^2 + \beta_v^2)(2\lambda d^2 + \beta_u^2 + \beta_v^2)}{(\lambda d^2 + \beta_u^2 + \beta_v^2)^2} \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right). \quad \text{(S308)}$$

Further because $\left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right) \boldsymbol{X} = \boldsymbol{0}$ and $\tilde{\boldsymbol{u}}^\top \boldsymbol{X} = \boldsymbol{0}$,

$$\boldsymbol{F}_1 \boldsymbol{F}_1^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{u}\ \boldsymbol{v})\, \boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1} \begin{pmatrix} \boldsymbol{u}^\top \\ \boldsymbol{v}^\top \end{pmatrix} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \quad \text{(S309)}$$

(b) $\boldsymbol{F}_2 \boldsymbol{F}_2^\top$.
Note that due to (S307)

$$(\boldsymbol{I} - \tilde{\boldsymbol{G}}) \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right) (\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top = \left(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})} \right). \quad \text{(S310)}$$

Combining with (S292), we have

$$\begin{aligned} &(\boldsymbol{I} - \tilde{\boldsymbol{G}})(\boldsymbol{C}_1 + \boldsymbol{C}_2)(\boldsymbol{C}_1 + \boldsymbol{C}_2)^\top (\boldsymbol{I} - \tilde{\boldsymbol{G}})^\top && \text{(S311)} \\ = \frac{1}{d^2 n} \Big[&\beta_u^2 (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}}) + \beta_v^2 (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{v}}) && \text{(S312)} \end{aligned}$$

$$- \left(\beta_v^2(\boldsymbol{I}-\boldsymbol{P_v})\tilde{\boldsymbol{u}}\ \beta_u^2(\boldsymbol{I}-\boldsymbol{P_u})\tilde{\boldsymbol{v}}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix} - \left(\boldsymbol{u}\ \boldsymbol{v}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2(\boldsymbol{I}-\boldsymbol{P_v})\tilde{\boldsymbol{u}}\\ \beta_u^2(\boldsymbol{I}-\boldsymbol{P_u})\tilde{\boldsymbol{v}}\end{pmatrix} \tag{S313}$$

$$+ \left(\boldsymbol{u}\ \boldsymbol{v}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P_v})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P_u})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix} \tag{S314}$$

$$-\frac{(\beta_u^2+\beta_v^2)(2\lambda d^2+\beta_u^2+\beta_v^2)}{(\lambda d^2+\beta_u^2+\beta_v^2)^2}\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\Bigg]. \tag{S315}$$

Because $\left(\boldsymbol{I}-\boldsymbol{P}_{(\boldsymbol{X}\boldsymbol{u}\boldsymbol{v})}\right)\boldsymbol{X}=\boldsymbol{0}$, $\boldsymbol{X}^\top\tilde{\boldsymbol{u}}=\boldsymbol{0}$ and $\boldsymbol{X}^\top\tilde{\boldsymbol{v}}=\boldsymbol{0}$,

$$\boldsymbol{F}_2\boldsymbol{F}_2^\top \tag{S316}$$

$$= \frac{1}{d^2 n}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Bigg[\beta_u^2(\boldsymbol{I}-\boldsymbol{P_u})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P_v}) \tag{S317}$$

$$+ \left(\boldsymbol{u}\ \boldsymbol{v}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P_v})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P_u})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\Bigg]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{S318}$$

Together with (S309) and (S318), we obtain the variance-covariance matrix of $\delta_{\boldsymbol{\beta}_x}$ as follows.

$$Cov\left(\delta_{\boldsymbol{\beta}_x}\right) = \sigma_y^2\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\left(\boldsymbol{u}\ \boldsymbol{v}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]+ \tag{S319}$$

$$\sigma_a^2\frac{1}{d^2 n}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Bigg[\beta_u^2(\boldsymbol{I}-\boldsymbol{P_u})+\beta_v^2(\boldsymbol{I}-\boldsymbol{P_v}) \tag{S320}$$

$$+ \left(\boldsymbol{u}\ \boldsymbol{v}\right)\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\beta_v^2\tilde{\boldsymbol{u}}^\top(\boldsymbol{I}-\boldsymbol{P_v})\tilde{\boldsymbol{u}} & 0\\ 0 & \beta_u^2\tilde{\boldsymbol{v}}^\top(\boldsymbol{I}-\boldsymbol{P_u})\tilde{\boldsymbol{v}}\end{pmatrix}\boldsymbol{C}_{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}}^{-1}\begin{pmatrix}\boldsymbol{u}^\top\\ \boldsymbol{v}^\top\end{pmatrix}\Bigg]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{S321}$$

which is the same as $Cov\left(\delta^{ts}_{\boldsymbol{\beta}_x}\right)$ in (S143).

Q.E.D.

REFERENCES

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *The Annals of Statistics*, pages 381–394.