WORKING PAPERS

N° 1574

July 2024

"Nonparametric Identification of Models for Dyadic Data"

Paul Diegert and Koen Jochmans

Toulouse School of Economics

NONPARAMETRIC IDENTIFICATION OF MODELS FOR DYADIC DATA

Paul Diegert*
Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

Koen Jochmans†
Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

This version: July 2, 2024

Abstract

Consider dyadic random variables on units from a given population. It is common to assume that these variables are jointly exchangeable and dissociated. In this case they admit a non-separable specification with two-way unobserved heterogeneity. The analysis of this type of structure is of considerable interest but little is known about their nonparametric identifiability, especially when the unobserved heterogeneity is continuous. We provide conditions under which both the distribution of the observed random variables conditional on the unit-specific heterogeneity and the distribution of the unit-specific heterogeneity itself are uniquely recoverable from knowledge of the joint marginal distribution of the observable random variables alone without imposing parametric restrictions.

Keywords: exchangeability, conditional independence, dyadic data, network, two-way heterogeneity

*Address: Toulouse School of Economics, 1 esplanade de l'Université, 31080 Toulouse, France. E-mail: paul.diegert@tse-fr.eu.

†Address: Toulouse School of Economics, 1 esplanade de l'Université, 31080 Toulouse, France. E-mail: koen.jochmans@tse-fr.eu.

Funded by the European Union (ERC, NETWORK, 101044319) and by the French Government and the French National Research Agency under the Investissements d' Avenir program (ANR-17-EURE-0010). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Dyadic data arise naturally from pairwise interaction between units belonging to a given population. Such data feature prominently in many areas of economics, psychology, and sociology, and their analysis has received an increasing amount of attention (Graham 2020 surveys recent contributions). Being a form of repeated-measurement data, dyadic data are naturally expected to be dependent. To govern this dependence the literature usually proceeds under the assumption that the array of dyadic variables is (jointly) exchangeable and dissociated. By a result of Kallenberg (1989) this can be seen to correspond to a nonparametric decomposition of the variable of interest into a pair of unit-specific random variables and a dyadic-specific random variable. As such, it yields a two-way generalization of the well-known one-way decomposition of longitudinal data.

The distribution of observable random variables conditional on the unit-specific effects and the distribution of those effects itself are often both of chief interest. In the context of one-way models this is almost invariably the case, and a variety of conditions under which these distributions are (nonparametrically) identified have been obtained (see Schennach 2020, 2021 for an overview and many references). Examples of economic applications with dyadic data are analyzed in Ahmapoor and Jones (2019), Bonhomme (2021), and Devereux (2018), for example. Contrary to the longitudinal-data case, however, the conditions under which such specifications are identified are much less clear. In this paper we provide conditions for this to be the case.

Our approach covers general non-separable representations and focusses on the case

where the unit-specific effects are continuously distributed. We exploit different types of conditional-independence restrictions implied by the assumption that the dyadic array is exchangeable and dissociated. These restrictions, when coupled with an intuitive injectivity requirement on an integral operator, are sufficient to obtain nonparametric identification of all distributions of interest.

2 Setup

Let $\mathbb{N}_1 \coloneqq \{1, 2 \ldots\}$ and $\mathbb{N}_2 \coloneqq \{(i_1, i_2) \in \mathbb{N}_1 \times \mathbb{N}_1 : i_1 < i_2\}$. We consider undirected dyadic data $(X_{i_1,i_2})_{(i_1,i_2)\in\mathbb{N}_2}$, abstracting away from the presence of covariates (to handle them it suffices to state all assumptions as holding conditional on them). The probability structure of the array is governed by two general nonparametric restrictions, stated in Assumption 1.

Assumption 1. *The array $(X_{i_1,i_2})_{(i_1,i_2)\in\mathbb{N}_2}$ is jointly exchangeable and dissociated.*

Joint exchangeability means that the distribution of the array is invariant to a relabelling of the indices. The array is dissociated if X_{i_1,i_2} and X_{i_3,i_4} are independent if their indices have no element in common. This allows for general dependence between variables that share an index.

From Kallenberg (1989), Assumption 1 implies that, almost surely,

$$X_{i_1,i_2} = H(V_{i_1}, V_{i_2}, W_{i_1,i_2}) \tag{2.1}$$

for mutually independent i.i.d. random variables $(V_i)_{i\in\mathbb{N}}$ and $(W_{i_1,i_2})_{(i_1,i_2)\in\mathbb{N}_2}$ and a function $H : \mathcal{V}\times\mathcal{V}\times\mathcal{W}\mapsto\mathcal{X}$. This representation reveals the dependency structure in the array in a transparent manner. Moreover, Equation (2.1) is a non-separable two-way decomposition of X_{i_1,i_2} into a pair of random effects (V_{i_1}, V_{i_2}) and an idiosyncratic component W_{i_1,i_2} that varies at the dyad level.

Structures as in Assumption 1 have been receiving an increasing amount of attention. Davezies, d'Haultœuille and Guyonvarch (2021), Menzel (2021), and Graham, Niu and Powell (2024), for example, are concerned with the estimation of various functionals of the marginal distribution of X_{i_1,i_2}. This distribution, which by Assumption 1 does not depend on (i_1, i_2), is nonparametrically identified. In many contexts, however, one is more interested in (functionals of) the distribution of X_{i_1,i_2} conditional on (V_{i_1}, V_{i_2}), as well as in the marginal distribution of the V_{i_1} itself. Several interesting examples are discussed in Devereux (2018), Ahmapoor and Jones (2019), and Bonhomme (2021). The extent to which these conditional distributions are nonparametrically identified is, however, not immediately clear.

The issue of identification has received attention in the context of stochastic block models for network formation, where it follows indirectly from the consistency results on spectral clustering (see von Luxburg, Belkin and Bousquet 2008, Sussman, Tang, Fishkind and Priebe 2012, and Lei and Rinaldo 2015, among others), and, to a lesser extent, in weighted versions thereof; see Allman, Matias and Rhodes (2011) and Jochmans (2024). These arguments rely heavily on the number of classes—that is, $|\mathcal{V}|$—being finite. In this

paper, our aim is to investigate the potential for identification in the case where all the variables are continuously distributed. To the best of our knowledge, no such result is currently available. Assumption 2 summarizes the continuity assumptions under which we will work.

Assumption 2. *The joint distribution of* $(X_{i_1,i_2}, V_{i_1}, V_{i_2})$ *admits a bounded density with respect to the Lebesgue measure on* $\mathcal{X} \times \mathcal{V} \times \mathcal{V}$. *The implied marginal and conditional densities are also bounded.*

We establish identification under two assumptions that involve the density of X_{i_1,i_2} conditional on V_{i_1} (or, equivalently, in light of Assumption 1, on V_{i_2}), which we denote by $f_{X|V}$ (recall that this function is independent of the indices in question). For any set $\mathcal{A}$, denote by $\mathcal{G}(\mathcal{A})$ the set of all absolutely-integrable functions with domain $\mathcal{A}$. We then let

$$[L_{X|V}\,\varphi](x) \coloneqq \int f_{X|V}(x|v)\,\varphi(v)\,dv$$

map $\varphi \in \mathcal{G}(\mathcal{V})$ to $L_{X|V}\,\varphi \in \mathcal{G}(\mathcal{X})$. We impose the following requirement on this operator.

Assumption 3. *The operator* $L_{X|V}$ *is injective.*

Injectivity assumptions of this kind are common in the analysis of latent-variable models; see, e.g., Carrasco, Florens and Renault (2007) and Hu and Schennach (2008). As explained in the latter paper, under Assumption 2, one sufficient condition for Assumption 3 to hold is the completeness of the distribution of V_{i_1} conditional on X_{i_1,i_2}. A variety of primitive conditions for this are available in the literature; see, e.g., Andrews (2017), d'Haultfœuille (2011), and Hu, Schennach and Shiu (2017).

Our fourth and final assumption arises due to the non-uniqueness of the representation in (2.1), due to the fact that the $(V_i)_{i\in\mathbb{N}_1}$ are latent and that the function H is not specified.[1]

Assumption 4. *For a known functional μ_V of $f_{X|V}$ we have that $\mu_V(v) = v$ for all $v \in \mathcal{V}$.*

This assumption can be interpreted as a location normalization.

An example that satisfies all of Assumptions 1–4 is the additively-separable specification

$$X_{i_1,i_2} = \vartheta + V_{i_1} + V_{i_2} + W_{i_1,i_2},$$

where $\vartheta := \mathbb{E}(X_{i_1,i_2})$ and both $(V_i)_{i\in\mathbb{N}_1}$ and $(W_{i_1,i_2})_{(i_1,i_2)\in\mathbb{N}_2}$ are normally distributed with variances σ_v^2 and σ_w^2, respectively. In this example, Assumption 3 is a consequence of the well-known completeness property of the exponential family; see, e.g., Newey and Powell (2003, Theorems 2.2 and 2.3) and the discussion in Hu and Schennach (2008, pp. 200) or Hu, Schennach and Shiu (2017, pp. 48). Further, here, Assumption 4 goes through for the linear functional $\mathbb{E}(X_{i_1,i_2}|V_{i_1}) - \mathbb{E}(X_{i_1,i_2})$, for example. In an error-component context, this specification can be seen as an extension of analysis of variance to dyadic data. In a clustering context one may be interested in

$$\operatorname{corr}(X_{i_1,i_2}, X_{i_1,i_3}) = \frac{\sigma_v^2}{2\sigma_v^2 + \sigma_w^2},$$

which is the two-way analog of the intraclass correlation coefficient (as in Moulton 1986).

Theorem 1 summarizes our main result.

[1] For example, if (2.1) holds for some $V_i \sim F_V$ then it also holds for the transformed random variable $F_V(V_i) \sim \text{uniform}(0,1)$ (along with a suitably modified function H) by the probability integral transform.

Theorem 1. *Let Assumptions 1–4 hold. Then the distribution of X_{i_1,i_2} conditional on (V_{i_1}, V_{i_2}) and the distribution of V_i are uniquely recoverable from knowledge of the joint distribution of $(X_{i_1,i_2})_{(i_1,i_2)\in\mathbb{S}_2}$ where $\mathbb{S}_2 \coloneqq \{(i_1,i_2) \in \mathbb{S}_1 \times \mathbb{S}_1 : i_1 < i_2\}$ and $\mathbb{S}_1$ is any subset of $\mathbb{N}_1$ with $|\mathbb{S}_1| \geq 4$.*

The proof of Theorem 1 is given in the next section.

3 Proof of Theorem 1

Throughout the proof, for a pair of (possibly multivariate) random variables $A \in \mathcal{A}$ and $B \in \mathcal{B}$, we let $f_{A,B}$ denote their joint density, f_A and f_B their marginal densities, and $f_{A|B}$ and $f_{B|A}$ their implied conditional densities. We also let

$$[L_{B|A}\,\varphi](b) \coloneqq \int f_{B|A}(b|a)\,\varphi(a)\,da$$

be the integral operator mapping $\varphi \in \mathcal{G}(\mathcal{A})$ to $L_{B|A}\,\varphi \in \mathcal{G}(\mathcal{B})$ and, finally, for any $b \in \mathcal{B}$, let $\Delta_{b,a}\,\varphi(a) \coloneqq f_{B|A}(b|a)\,\varphi(a)$ for any $\varphi \in \mathcal{G}(\mathcal{A})$.

Our proof consists of two steps. Each step is based on a different implication of conditional-independence restrictions embedded in Assumption 1. In the first step we recover the joint density of $(X_{1,2}, V_1)$. This also identifies the marginal and conditional densities. In the second step we use this knowledge to tease out the joint distribution of $X_{1,2}, V_1, V_2$, from which all conditional densities again follow. Our arguments below rely on knowledge of the joint distribution of $(X_{1,2}, X_{1,3}, X_{1,4}, X_{3,4})$, which is nonparametrically

identified. Here, the choice of indices is arbitrary and, by joint exchangeability, without loss of generality.

Consider, first, the distribution of $(X_{1,2}, X_{1,3})$ conditional on $X_{1,4}$. This distribution has associated with it the integral operator

$$[L_{X_{1,2},X_{1,3}|X_{1,4}}\,\varphi](x_{1,2};x_{1,3}) = \int f_{X_{1,2},X_{1,3}|X_{1,4}}(x_{1,2},x_{1,3}|x_{1,4})\,\varphi(x_{1,4})\,dx_{1,4},$$

where we are fixing the variable $X_{1,3}$ to a given value $x_{1,3} \in \mathcal{X}$. As is apparent from the representation of the array in Equation (2.1), Assumption 1 implies that $(X_{1,2}, X_{1,3}, X_{1,4})$ are dependent only because of their joint dependence on the variable V_1. Therefore, we have the factorization

$$f_{X_{1,2},X_{1,3},V_1|X_{1,4}}(x_{1,2},x_{1,3},v_1|x_{1,4}) = f_{X_{1,2}|V_1}(x_{1,2}|v_1)\,f_{X_{1,3}|V_1}(x_{1,3}|v_1)\,f_{V_1|X_{1,4}}(v_1|x_{1,4}).$$

Furthermore, because

$$[L_{X_{1,2},X_{1,3}|X_{1,4}}\,\varphi](x_{1,2};x_{1,3}) = \int f_{X_{1,2},X_{1,3},V_1|X_{1,4}}(x_{1,2},x_{1,3},v_1|x_{1,4})\,\varphi(x_{1,4})\,dv_1\,dx_{1,4},$$

we have the operator equivalence relation

$$L_{X_{1,2},X_{1,3}|X_{1,4}} = L_{X_{1,2}|V_1}\,\Delta_{X_{1,3},V_1}L_{V_1|X_{1,4}}. \tag{3.2}$$

Because $\int f_{X_{1,3}|V_1}(x_{1,3}|v_1)\,dx_{1,3} = 1$ for all $v_1 \in \mathcal{V}$, marginalizing with respect to $X_{1,3}$ further yields

$$L_{X_{1,2}|X_{1,4}} = L_{X_{1,2}|V_1}\,L_{V_1|X_{1,4}}. \tag{3.3}$$

By exchangeability, $L_{X_{1,2}|V_1} = L_{X|V}$ as defined in the text. By Assumption 3 this operator is injective and so, by Lemma 1 of Hu and Schennach (2008), $L^{-1}_{V_1|X_{1,2}}$ exists and is densely

defined over $\mathcal{G}(\mathcal{V})$. Therefore, Equation (3.3) yields

$$L^{-1}_{X_{1,2}|X_{1,4}} = L^{-1}_{V_1|X_{1,4}}\, L^{-1}_{X_{1,2}|V_1}. \tag{3.4}$$

Combining Equations (3.2) and (3.4) then yields the spectral decomposition

$$L_{X_{1,2},X_{1,3}|X_{1,4}} L^{-1}_{X_{1,2}|X_{1,4}} = L_{X_{1,2}|V_1}\, \Delta_{X_{1,3},V_1} L^{-1}_{X_{1,2}|V_1}.$$

By Assumptions 2-4 the arguments of Hu and Schennach (2008, Proof of Theorem 1, pp. 211-213) yield uniqueness of this decomposition, identifying the density function of $X_{1,2}$ conditional on V_1 and, with it, the operator $L_{X|V}$. Further, with this operator in hand we equally recover

$$L^{-1}_{X_{1,2}|V_1} L_{X_{1,2}|X_{1,4}} = L_{V_1|X_{1,4}},$$

and so the density of V_1 conditional on $X_{1,2}$, from re-arranging Equation (3.3). Because the marginal density of $X_{1,2}$ is identified, this equally yields the marginal density of V_1.

Consider, next, the distribution of $(X_{1,2}, X_{1,3})$ conditional on $X_{3,4}$. This distribution has associated with it the integral operator

$$[L_{X_{1,2},X_{1,3}|X_{3,4}}\, \varphi](x_{1,2}; x_{1,3}) = \int f_{X_{1,2},X_{1,3}|X_{3,4}}(x_{1,2}, x_{1,3}|x_{3,4})\, \varphi(x_{3,4})\, dx_{3,4},$$

where we are again fixing the variable $X_{1,3}$ to a given value $x_{1,3} \in \mathcal{X}$. Using Assumption 1,

$$f_{X_{1,2},X_{1,3}|X_{2,4}}(x_{1,2}, x_{1,3}|x_{3,4}) = \iint f_{X_{1,2}|V_1}(x_{1,2}|v_1)\, f_{X_{1,3},V_1|V_3}(x_{1,3}, v_1|v_3)\, f_{V_3|X_{3,4}}(v_3|x_{3,4})\, dv_1\, dv_3,$$

and so we have the operator equivalence relation

$$L_{X_{1,2},X_{1,3}|X_{3,4}} = L_{X_{1,2}|V_1} L_{X_{1,3},V_1|V_3} L_{V_3|X_{3,4}}.$$

In this equality the only operator that has not yet been shown to be identified is $L_{X_{1,3},V_1|V_3}$. The remaining operators on the right-hand side are both injective. Therefore, we can invert the above equation to obtain

$$L_{X_{1,3},V_1|V_3} = L^{-1}_{X_{1,2}|V_1} L_{X_{1,2},X_{1,3}|X_{3,4}} L^{-1}_{V_3|X_{3,4}},$$

thereby identifying the density of $(X_{1,2}, V_1)$ conditional on V_2. As the marginal density of V_2 has already been identified, we then equally recover the joint density of $(X_{1,2}, V_1, V_2)$ as well as all the implied conditional densities. The proof of the theorem is thus complete. □

References

Ahmapoor, M. and B. F. Jones (2019). Decoding team and individual impact in science and invention. *Proceedings of the National Academy of Sciences 116*, 13885–13890.

Allman, E. S., C. Matias, and J. A. Rhodes (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference 141*, 1719–1736.

Andrews, D. W. K. (2017). Examples of L^2-complete and boundedly-complete distributions. *Journal of Econometrics 199*, 213–220.

Bonhomme, S. (2021). Teams: Heterogeneity, sorting, and complementarity. Proceedings of the 15th World Congress of the Econometric Society, to appear.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems and structural econometrics: Estimation based on spectral decomposition and regularization. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 5633–5751.

Davezies, L., X. d'Haultœuille, and Y. Guyonvarch (2021). Empirical process results for exachangeable arrays. *Annals of Statistics 49*, 845–862.

Devereux, K. (2018). Identifying the value of teamwork: Application to professional tennis. Canadian Labour Economics Forum, WP 14.

d'Haultfœuille, X. (2011). On the completeness condition for nonparametric instrumental problems. *Econometric Theory 27*, 460–471.

Graham, B. S. (2020). Network data. In S. Durlauf, L. P. Hansen, J. J. Heckman, and R. Matzkin (Eds.), *Handbook of Econometrics*, Volume 7A, pp. 111–218. Elsevier.

Graham, B. S., F. Niu, and J. L. Powell (2024). Kernel density estimation for undirected dyadic data. *Journal of Econometrics 240*, 105336.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*, 195–216.

Hu, Y., S. M. Schennach, and J.-L. Shiu (2017). Injectivity of a class of integral operators with compactly supported kernels. *Journal of Econometrics 200*, 48–58.

Jochmans, K. (2024). Nonparametric identification and estimation of stochastic block models from many small networks. Forthcoming in *Journal of Econometrics*.

Kallenberg, O. (1989). On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis 30*, 137–154.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic blockmodels. *Annals of Statistics 43*, 215–237.

Menzel, K. (2021). Bootstrap with clustering in two or more dimensions. *Econometrica 89*, 2143–2188.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics 32*, 385–397.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*, 1565–1578.

Schennach, S. M. (2020). Mismeasured and unobserved variables. In S. Durlauf, L. P. Hansen, J. J. Heckman, and R. Matzkin (Eds.), *Handbook of Econometrics*, Volume 7A, pp. 487–565. Elsevier.

Schennach, S. M. (2021). Measurement systems. CeMMAP Working Paper CWP12/21. Forthcoming in *Journal of Economic Literature*.

Sussman, D. L., M. Tang, D. E. Fishkind, and C. E. Priebe (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association 107*, 1119–1128.

von Luxburg, U., M. Belkin, and O. Bousquet (2008). Consistency of spectral clustering. *Annals of Statistics 36*, 555–586.