The Annals of Statistics
2017, Vol. 45, No. 4, 1488–1515
DOI: 10.1214/16-AOS1497

CO-CLUSTERING OF NONSMOOTH GRAPHONS

BY DAVID CHOI

Carnegie Mellon University

Performance bounds are given for exploratory co-clustering/blockmodeling of bipartite graph data, where we assume the rows and columns of the data matrix are samples from an arbitrary population. This is equivalent to assuming that the data is generated from a nonsmooth graphon. It is shown that co-clusters found by any method can be extended to the row and column populations, or equivalently that the estimated blockmodel approximates a blocked version of the generative graphon, with estimation error bounded by $O_P(n^{-1/2})$. Analogous performance bounds are also given for degree-corrected blockmodels and random dot product graphs, with error rates depending on the dimensionality of the latent variable space.

1. Introduction. In the statistical analysis of network data, blockmodeling (or community detection) and its variants are a popular class of methods that have been tried in many applications, such as modeling of communication patterns [Blondel et al. (2008)], academic citations [Ji and Jin (2014)], protein networks [Airoldi et al. (2009)] and online behavior [Latouche, Birmelé and Ambroise (2011), Traud et al. (2011)].

In order to develop a theoretical understanding, many recent papers have established consistency properties for the blockmodel. In these papers, the observed network is assumed to be generated using a set of latent variables that assign the vertices into groups (the "communities"), and the inferential goal is to recover the correct group membership from the observed data. Various conditions have been established under which recovery is possible and computationally tractable [Cai and Li (2015), Chen et al. (2012), Gao et al. (2015a), Krzakala et al. (2013), Newman (2013), Sussman et al. (2012)]. Additionally, conditions are also known under which no algorithm can correctly recover the group memberships [Decelle et al. (2011), Mossel, Neeman and Sly (2013)].

The existence of a true group membership is central to these results. In particular, they assume a generative model in which all members of the same group are statistically identical. This implies that the group memberships explain the entirety of the network structure. In practice, we might not expect this assumption to even approximately hold, and the objective of finding "true communities" could be difficult to define precisely, so that a more reasonable goal might be to discover group

Received July 2015; revised March 2016.

MSC2010 subject classifications. Primary 62G05; secondary 05C80, 60B20.

Key words and phrases. Graphon, bipartite graph, co-clustering, statistical network analysis, stochastic blockmodel, degree-corrected blockmodel, random dot product graph.

labels which partially explain structure that is evident in the data. Comparatively, little work has been done to understand blockmodeling from this viewpoint.

To address this gap, we consider the problem of blockmodeling under model misspecification. We assume that the data is generated not by a blockmodel, but by a much larger nonparametric class known as a graphon. This is equivalent to assuming that the vertices are sampled from an underlying population, in which no two members are identical and the notion of a true community partition need not exist. In this setting, blockmodeling might be better understood not as a generative model, but rather as an exploratory method for finding high-level structure: by dividing the vertices into groups, we divide the network into subgraphs that can exhibit varying levels of connectivity. This is analogous to the usage of histograms to find high and low density regions in a nonparametric distribution. Just as a histogram replicates the binned version of its underlying distribution without restrictive assumptions, we will show that the blockmodel replicates the blocked version of a generative graphon.

Our results are restricted to the case of bipartite graph or binary array data. Such data arises in many settings, such as customer-product networks where connections may represent purchases, reviews or some other interaction between people and products. Examples of bipartite networks include Goh et al. (2007), Jeong et al. (2000), Newman (2001). Given nonbinary data, it may still be of practical interest to look for biclustering patterns in the thresholded or binarized data matrix [Chen et al. (2013), Harpaz et al. (2011), van Uitert, Meuleman and Wessels (2008)].

The organization of the paper is as follows. Related work is discussed in Section 2. In Section 3, we define the blockmodeling problem for bipartite data generated from a graphon, and present Theorem 1 showing that the blockmodel can detect structure in the underlying population. In Section 4, we discuss extensions of the blockmodel, such as degree-corrected blockmodels and random dot product graphs, and present Theorem 2 regarding the behavior of the excess risk in such models. Section 5 contains a sketch and proof for Theorem 1. Section 6 contains a simulation study, and Section 7 discusses future work. Auxiliary results for Theorem 1 are proven in the Appendix, and the proof of Theorem 2 is given in the supplemental material [Choi (2017)].

2. Related works. The papers Airoldi, Costa and Chan (2013), Borgs et al. (2015), Choi and Wolfe (2014), Gao, Lu and Zhou (2014), Gao et al. (2015b), Klopp, Tsybakov and Verzelen (2015), Olhede and Wolfe (2014) are most similar to the present work, in that they consider the problem of approximating a graphon by a blockmodel. The papers Airoldi, Costa and Chan (2013), Gao, Lu and Zhou (2014), Klopp, Tsybakov and Verzelen (2015), Olhede and Wolfe (2014) consider both bipartite and nonbipartite graph data, and require the generative graphon to satisfy a smoothness condition, with Gao, Lu and Zhou (2014) establishing a minimax error rate, Klopp, Tsybakov and Verzelen (2015) extending the results to a class of sparse graphon models, and Gao et al. (2015b) extending to nonbinary bipartite data with partial observation. In a similar vein, Sussman, Tang and Priebe (2012) shows consistent and computationally efficient estimation assuming a type of low rank generative model. While smoothness and rank assumptions are natural for many nonparametric regression problems, it seems difficult to judge whether they are appropriate for network data and if they are indeed necessary for good performance.

In Choi and Wolfe (2014) and in this present paper, which consider only bipartite graphs, the emphasis is on exploratory analysis. Hence, no assumptions are placed on the generative graphon. Unlike works which assume smoothness or low rank structure, the object of inference is not the generative model itself, but rather a blocked version of it (this is defined precisely in Section 3). This is reminiscent of some results for confidence intervals in nonparametric regression, in which the interval is centered not on the generative function or density itself, but rather on a smoothed or histogram-ed version [Wasserman (2006), Section 5.7 and Theorem 6.20]. The present paper can be viewed as a substantial improvement over Choi and Wolfe (2014). Specfically, Lemma 2 gives a rate of convergence of $O_P(n^{-1/2})$, versus the rate of $O_P(n^{-1/4})$ given by Theorem 4.1 of Choi and Wolfe (2014); an exponential-time fitting algorithm is no longer assumed; and the proof techniques extend to relatives of the blockmodel (degree-corrected and random dot product graphs), with different rates of convergence. However, sparse graphon models are not considered in either work.

The recent paper [Borgs et al. (2015)] also considers the problem of approximating an arbitrary graphon by a blockmodel. Exponential-time fitting methods are primarily considered, with results for a new class of sparse graphon models that—unlike previous works—allows for heavy-tailed degree distributions.

3. Co-clustering of nonsmooth graphons.

In this section, we give a formulation for co-clustering (or co-blockmodeling) in which the rows and columns of the data matrix are samples from row and column populations, and correspond to the vertices of a bipartite graph. We then present an approximation result which implies that any co-clustering of the rows and columns of the data matrix can be extended to the populations. Roughly speaking, this means that if a co-clustering "reveals structure" in the data matrix, then similar structure will also exist at the population level.

3.1. *Problem formulation.*

Data generating process. Let $A \in \{0, 1\}^{m\times n}$ denote a binary $m \times n$ matrix representing the observed data. For example, A_{ij} could denote whether person i rated movie j favorably, or whether gene i was expressed under condition j.

We assume that A is generated by the following model, in which each row and column of A is associated with a latent variable that is sampled from a population.

DEFINITION 1 (Bipartite Graphon [Diaconis and Janson (2007), Lovász (2012)]). Given m and n, let $x_1, \dots, x_m$ and $y_1, \dots, y_n$ denote i.i.d. uniform [0, 1] latent variables

$$x_1, \dots, x_m \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1] \quad \text{and} \quad y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1].$$

Let $\omega : [0, 1]^2 \mapsto [0, 1]$ specify the distribution of $A \in \{0, 1\}^{m \times n}$, conditioned the latent variables $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$,

$$A_{ij} \sim \text{Bernoulli}\big(\omega(x_i, y_j)\big), \qquad i \in [m], j \in [n],$$

where the Bernoulli random variables are independent.

We will require that ω be measurable and square-integrable, but may otherwise be arbitrarily nonsmooth. We will use $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = [0, 1]$ to denote the populations from which $\{x_i\}$ and $\{y_j\}$ are sampled.

Co-clustering. In co-clustering, the rows and columns of a data matrix A are simultaneously clustered to reveal submatrices of A that have similar values. When A is binary valued, this is also called blockmodeling (or co-blockmodeling).

Our notation for co-clustering is the following. Let K denote the number of clusters. Let $S \in [K]^m$ denote a vector identifying the cluster labels corresponding to the m rows of A, for example, $S_i = k$ means that the ith row is assigned to cluster k. Similarly, let $T \in [K]^n$ identify the cluster labels corresponding to the n rows of A. Given (S, T), let $\Phi_A(S, T) \in [0, 1]^{K \times K}$ denote the normalized sums for the submatrices of A induced by S and T:

$$\big[\Phi_A(S, T)\big]_{st} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} 1(S_i = s, T_j = t), \qquad s, t \in [K].$$

Let $\pi_S \in [0, 1]^K$ and $\pi_T \in [0, 1]^K$ denote the fraction of rows or columns in each cluster:

$$\pi_S(s) = \frac{1}{m} \sum_{i=1}^{m} 1(S_i = s) \quad \text{and} \quad \pi_T(t) = \frac{1}{n} \sum_{j=1}^{n} 1(T_j = t).$$

Let the average value of the (s, t)th submatrix be denoted by $\hat{\theta}_{st}$, given by

$$\hat{\theta}_{st} = \frac{[\Phi_A(S, T)]_{st}}{\pi_S(s)\pi_T(t)}.$$

Generally, S and T are chosen heuristically to make the entries of $\hat{\theta}$ far from the overall average of A. A common approach is to perform k-means clustering of the spectral coordinates for each row and column of A [Rohe, Qin and Yu (2012)]. Heterogeneous values of $\hat{\theta}$ can be interpreted as revealing subgroups of the rows and columns in A.

Population co-blockmodel. Given a co-clustering (S, T) of the rows and columns of A, we will consider whether similar subgroups also exist in the unobserved populations $\mathcal{X}$ and $\mathcal{Y}$. Let $\sigma : \mathcal{X} \mapsto [K]$ and $\tau : \mathcal{Y} \mapsto [K]$ denote mappings that co-cluster the row and column populations $\mathcal{X}$ and $\mathcal{Y}$. Let $\Phi_\omega(\sigma, \tau) \in [0,1]^{K\times K}$ denote the integral of ω within the induced co-clusters, or the blocked version of ω:

$$\big[\Phi_\omega(\sigma,\tau)\big]_{st} = \int_{\mathcal{X}\times\mathcal{Y}} \omega(x,y)1\big(\sigma(x)=s, \tau(y)=t\big)\,dx\,dy, \qquad s,t \in [K].$$

Let $\Phi_\omega(S, \tau) \in [0,1]^{K\times K}$ denote the integral of ω within the induced co-clusters, over $\{x_1, \dots, x_n\} \times \mathcal{Y}$:

$$\big[\Phi_\omega(S,\tau)\big]_{st} = \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{Y}} \omega(x_i,y)1\big(S_i = s, \tau(y)=t\big)\,dy.$$

Let $\pi(\sigma)$ and $\pi(\tau)$ denote the fraction of the population in each cluster:

$$\pi_\sigma(s) = \int_{\mathcal{X}} 1\big(\sigma(x)=s\big)\,dx \quad \text{and} \quad \pi_\tau(t) = \int_{\mathcal{Y}} 1\big(\tau(y)=t\big)\,dy.$$

Theorem 1 will show that for each clustering S, T, there exists $\sigma : \mathcal{X} \mapsto [K]$ and $\tau : \mathcal{Y} \mapsto [K]$ which cluster the populations $\mathcal{X}$ and $\mathcal{Y}$ such that $\Phi_A(S,T) \approx \Phi_\omega(S,\tau)$ and $\Phi_A(S,T) \approx \Phi_\omega(\sigma,\tau)$, as well as $\pi_S \approx \pi_\sigma$ and $\pi_T \approx \pi_\tau$, implying that subgroups found by co-clustering A are indicative of similar structure in the populations $\mathcal{X}$ and $\mathcal{Y}$.

3.2. *Approximation result for co-clustering.* Theorem 1 states that for each $(S,T) \in [K]^m \times [K]^n$, there exists population co-clusters $\sigma_S : \mathcal{X} \mapsto [K]$ and $\tau_T : \mathcal{Y} \mapsto [K]$ such that $\Phi_A(S,T) \approx \Phi_\omega(S,\tau_T) \approx \Phi_\omega(\sigma_S,\tau_T)$, and also $\pi_S \approx \pi_{\sigma_S}$ and $\pi_T \approx \pi_{\tau_T}$.

THEOREM 1. *Let $A \in \{0,1\}^{m\times n}$ be generated by some ω according to Definition 1, with fixed ratio m/n. Let (S,T) denote vectors in $[K]^m$ and $[K]^n$, respectively, with $K \le n^{1/2}$:*

1. *For each $T \in [K]^n$, there exists $\tau_T : \mathcal{Y} \mapsto [K]$, such that*

$$\text{(1)} \quad \max_{S,T\in[K]^m\times[K]^n} \big\|\Phi_A(S,T) - \Phi_\omega(S,\tau_T)\big\| + \|\pi_T - \pi_{\tau_T}\| = O_P\left(\sqrt{\frac{K^2\log n}{n}}\right).$$

2. *For each $S \in [K]^m$, there exists $\sigma_S : \mathcal{X} \mapsto [K]$, such that*

$$\text{(2)} \quad \sup_{S,\tau\in[K]^m\times[K]^{\mathcal{Y}}} \big\|\Phi_\omega(S,\tau) - \Phi_\omega(\sigma_S,\tau)\big\| + \|\pi_S - \pi_{\sigma_S}\| = O_P\left(\sqrt{\frac{K^2\log m}{m}}\right).$$

3. *Combining* (1) *and* (2) *yields*

$$
\begin{aligned}
&\max_{S,T\in[K]^m\times[K]^n} \|\Phi_\omega(\sigma_S,\tau_T)-\Phi_A(S,T)\| + \|\pi_T-\pi_{\tau_T}\| + \|\pi_S-\pi_{\sigma_S}\| \\
&\quad = O_P\Bigl(\sqrt{\frac{K^2\log n}{n}}\Bigr).
\end{aligned} \tag{3}
$$

Remarks for Theorem 1. To give context to Theorem 1, suppose that $A \in \{0,1\}^{m\times n}$ represents product–customer interactions, where $A_{ij}=1$ indicates that product i was purchased (or viewed, reviewed, etc.) by customer j. We assume A is generated by Definition 1, meaning that the products and customers are samples from populations. This could be literally true if A is sampled from a larger data set, or the populations might only be conceptual, perhaps representing future products and potential customers.

Suppose that we have discovered cluster labels $S\in[K]^m$ and $T\in[K]^n$ producing a density matrix $\hat\theta$ with heterogeneous values. These clusters can be interpreted as product categories and customer subgroups, with heterogeneity in $\hat\theta$ indicating that each customer subgroup may prefer certain product categories over others. We are interested in the following question: will this pattern generalize to the populations $\mathcal{X}$ and $\mathcal{Y}$? Or is it descriptive, holding only for the particular customers and products that are in the data matrix A?

An answer is given by Theorem 1. Specifically, (1) and (3) show different senses in which the co-clustering (S,T) may generalize to the underlying populations. Equation (1) implies that the customer population $\mathcal{Y}$ will be similar to the n observed customers in the data, regarding their purchases of the m observed products when aggregated by product category. Equation (3) implies a similar result, but for their purchases of the entire population $\mathcal{X}$ of products aggregated by product category, as opposed only to the m observed products in the data.

Since Theorem 1 holds for all (S,T), it applies regardless of the algorithm that is used to choose the co-blockmodel. It also applies to nested or hierarchical clusters. If (1) or (3) holds at the lowest level of hierarchy with K classes, then it also holds for the aggregated values at higher levels as well, since they correspond to K-class clusterings with one or more classes of size zero. While we have assumed that the number of row and column classes are equal to simplify the proofs, we remark that the theorem will also hold when this is not the case, with K equaling the larger of the two class counts.

Theorem 1 controls the behavior of Φ_A, π_S, and π_T, instead of the density matrix $\hat\theta$ which may be of interest. However, since $\hat\theta$ is derived from the previous quantities, it follows that Theorem 1 also implies control of $\hat\theta$ for all co-clusters involving $\gg m^{1/2}$ rows or $\gg n^{1/2}$ columns.

All constants hidden by the $O_P(\cdot)$ notation in Theorem 1 are universal, in that they do not depend on ω (but do depend on the ratio m/n).

4. Application of Theorem 1 to bipartite graph models. In many existing models for bipartite graphs, the rows and columns of the adjacency matrix $A \in \{0,1\}^{m\times n}$ are associated with latent variables that are not in $\mathcal{X}$ and $\mathcal{Y}$, but in other spaces $\mathcal{S}$ and $\mathcal{T}$ instead. In this section, we give examples of such models and discuss their estimation by minimizing empirical squared error. We define the population risk as the difference between the estimated and actual models, under a transformation mapping $\mathcal{X}$ to $\mathcal{S}$ and $\mathcal{Y}$ to $\mathcal{T}$. Theorem 2 shows that the empirical error surface converges uniformly to the population risk. The theorem does not assume a correctly specified model, but rather that the data is generated by an arbitrary ω following Definition 1.

4.1. *Examples of bipartite graph models.* We consider models in which the rows and columns of A are associated with latent variables that take values in spaces other than $\mathcal{X}$ and $\mathcal{Y}$. To describe these models, we will use $S = (S_1,\dots,S_m)$ and $T = (T_1,\dots,T_n)$ to denote the row and column latent variables, and $\mathcal{S}$ and $\mathcal{T}$ to denote their allowable values. Let Θ denote a parameter space. Given $\theta \in \Theta$, let $\omega_\theta : \mathcal{S} \times \mathcal{T} \mapsto [0,1]$ determine the distribution of A conditioned on (S,T), so that the entries $\{A_{ij}\}$ are conditionally independent Bernoulli variables, with $\mathbb{P}(A_{ij} = 1 | S, T) = \omega_\theta(S_i, T_j)$.

1. *Stochastic co-blockmodel with K classes*: Let $\mathcal{S} = \mathcal{T} = [K]$ and $\Theta = [0,1]^{K\times K}$. For $\theta \in \Theta$, let ω_θ be given by

$$\omega_\theta(s,t) = \theta_{st}, \qquad s,t \in \mathcal{S} \times \mathcal{T},$$

where $s \in \mathcal{S}$ and $t \in \mathcal{T}$ are row and column co-cluster labels.

2. *Degree-corrected co-blockmodel* [*Karrer and Newman* (2011), *Zhao, Levina and Zhu* (2012)]: Let $\mathcal{S} = \mathcal{T} = [K] \times [0,1)$ and $\Theta = [0,1]^{K\times K}$. Given $u, v \in [K]$ and $b, d \in [0,1)$, let $s = (u,b)$ and $t = (v,d)$. Let ω_θ be given by

$$\omega_\theta(s,t) = bd\theta_{uv}, \qquad s,t \in \mathcal{S} \times \mathcal{T}.$$

In this model, $u, v \in [K]$ are co-cluster labels, and $b, d \in [0,1)$ are degree parameters, allowing for degree heterogeneity within co-clusters.

3. *Random dot product* [*Hoff, Raftery and Handcock* (2002), *Sussman, Tang and Priebe* (2012)]: Let $\mathcal{S} = \mathcal{T} = \{c \in [0,1)^d : \|c\| \leq 1\}$. Let ω be given by

$$\omega(s,t) = s^T t, \qquad s,t \in \mathcal{S} \times \mathcal{T}.$$

4. *Dot product + Blockmodel*: Models 1–3 are instances of a somewhat more general model. Let $\mathcal{D} = \{c \in [0,1)^d : \|c\| \leq 1\}$. Let $\mathcal{S} = \mathcal{T} = [K] \times \mathcal{D}$ and $\Theta = [0,1]^{K\times K}$. Given $u, v \in [K]$ and $b, d \in \mathcal{D}$, let $s = (u,b)$ and $t = (v,d)$. Let ω_θ be given by

$$\omega_\theta(s,t) = b^T d\theta_{uv}. \tag{4}$$

4.2. *Empirical and population risk.* Given a data matrix $A \in \{0,1\}^{m \times n}$, and a model specification $(\mathcal{S}, \mathcal{T}, \Theta)$, one method for estimating $(S, T, \theta) \in \mathcal{S}^m \times \mathcal{T}^n \times \Theta$ is to minimize the empirical squared error R_A, given by

$$R_A(S, T; \theta) = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \bigl(A_{ij} - \omega_\theta(S_i, T_j)\bigr)^2.$$

Generally, the global minimum of R_A will be intractable to compute, so a local minimum is used for the estimate instead.

If a model (S, T, θ) is found by minimizing or exploring the empirical risk surface R_A, does it approximate the generative ω? We will define the population risk in two different ways:

1. *Approximation of ω by ω_θ*: Let σ and τ denote mappings $\mathcal{X} \mapsto \mathcal{S}$ and $\mathcal{Y} \mapsto \mathcal{T}$, and let R_ω be given by

$$R_\omega(\sigma, \tau; \theta) = \int_{\mathcal{X} \times \mathcal{Y}} \bigl[\omega(x, y) - \omega_\theta\bigl(\sigma(x), \tau(y)\bigr)\bigr]^2 \, dx \, dy,$$

denoting the error between the mapping $(x, y) \mapsto \omega_\theta(\sigma(x), \tau(y))$ and the generative ω. If there exists θ such that $R_\omega(\sigma, \tau; \theta)$ is low for some $\sigma : \mathcal{X} \mapsto \mathcal{S}$ and $\tau : \mathcal{Y} \mapsto \mathcal{T}$, then ω_θ [or more precisely, its transformation $(x, y) \mapsto \omega_\theta(\sigma(x), \tau(y))$] can be considered a good approximation to ω.

2. *Approximation of $\sigma^* = \arg\min_\sigma R_\omega(\sigma, \tau, \theta)$ by S*: Overloading notation, let $R_\omega(S, \tau, ; \theta)$ denote

$$R_\omega(S, \tau; \theta) = \frac{1}{m} \sum_{i=1}^{m} \int_{\mathcal{Y}} \bigl[\omega(x_i, y) - \omega_\theta\bigl(S_i, \tau(y)\bigr)\bigr]^2 \, dy.$$

To motivate this quantity, consider that given (τ, θ), the optimal partition $\sigma^* : [0,1] \mapsto [K]$ is the greedy assignment for each $x \in [0,1]$:

$$\sigma^*(x) = \arg\min_{s \in [K]} \int_{0,1} \bigl[\omega(x, y) - \omega_\theta\bigl(s, \tau(y)\bigr)\bigr]^2 \, dy.$$

If there exists (S, θ) such that $R_\omega(S, \tau; \theta)$ is low for some choice of τ, then S can be considered a good approximation to the corresponding $\{\sigma^*(x_i)\}_{i=1}^m$.

Theorem 2 will imply that for models of the form (4), minimizing R_A is asymptotically a reasonable proxy for minimizing R_ω (by both metrics described above), with rates of convergence depending K and d.

4.3. *Convergence of the empirical risk function.* Theorem 2 gives uniform bounds between R_A and R_ω for models of form (4). Specifically, for each choice of $(S, T) \in \mathcal{S}^m \times \mathcal{T}^n$, there exists transformations $\sigma_S : \mathcal{X} \mapsto \mathcal{S}$ and $\tau_T : \mathcal{Y} \mapsto \mathcal{T}$ such that $R_A(S, T; \theta) \approx R_\omega(\sigma_S, \tau_T; \theta) \approx R_\omega(S, \tau_T; \theta)$, up to an additive constant and with uniform convergence rates depending on d and K. As a result, minimization of $R_A(S, T; \theta)$ is a reasonable proxy for minimizing R_ω, by either measure defined in Section 4.2.

In addition, the mappings σ_S and τ_T will resemble S and T, in that they will induce similar distributions over the latent variables. To quantify this, we define the following quantities. Given $S \in [K]^m \times \mathcal{D}^m$, we will let $S = (U, B)$, where $U \in [K]^m$ and $B \in \mathcal{D}^m$, and similarly let $T = (V, D)$ where $V \in [K]^n$ and $D \in \mathcal{D}^n$. Likewise, given $\sigma : \mathcal{X} \mapsto [K] \times \mathcal{D}$, we will let $\sigma = (\mu, \beta)$, where $\mu : \mathcal{X} \mapsto [K]$ and $\beta : \mathcal{X} \mapsto \mathcal{D}$, and similarly let $\tau = (\nu, \delta)$ where $\nu : \mathcal{Y} \mapsto [K]$ and $\delta : \mathcal{Y} \mapsto \mathcal{D}$. Let Ψ_S, Ψ_T, Ψ_σ, and Ψ_τ denote the CDFs of the values given by S, T, σ and τ, which are functions $[K] \times [0, 1)^d \mapsto [0, 1]$ equaling:

$$\Psi_S(k, c) = \frac{1}{m} \sum_{i=1}^{m} 1\{U_i \le k, B_i \le c\},$$

$$\Psi_\sigma(k, c) = \int_{\mathcal{X}} 1\{\mu(x) \le k, \beta(x) \le c\}\, dx,$$

$$\Psi_T(k, c) = \frac{1}{n} \sum_{j=1}^{n} 1\{V_j \le k, D_j \le c\},$$

$$\Psi_\tau(k, c) = \int_{\mathcal{Y}} 1\{\nu(y) \le k, \delta(y) \le c\}\, dy,$$

where inequalities of the form $c \le c'$ for $c, c' \in [0, 1)^d$ are satisfied if they hold entrywise.

THEOREM 2. *Let $A \in \{0, 1\}^{m \times n}$, with fixed ratio m/n, be generated by some ω according to Definition 1. Let $(\mathcal{S}, \mathcal{T}, \Theta)$ denote a model of the form (4):*

1. *For each $T \in \mathcal{T}^n$, there exists $\tau_T : \mathcal{Y} \mapsto \mathcal{T}$ such that*

$$\begin{aligned} &\max_{S,T,\theta \in \mathcal{S}^m \times \mathcal{T}^n \times \Theta} \left| R_A(S, T; \theta) - R_\omega(S, \tau_T; \theta) - C_1 \right| + \frac{\|\Psi_T - \Psi_{\tau_T}\|^2}{Kd} \\ &\quad \le O_P\left(d^{1/2} \left(\frac{K^2 \log n}{\sqrt{n}} \right)^{\frac{1}{1+d}} \right), \end{aligned} \tag{5}$$

where $C_1 \in \mathbb{R}$ is constant in (S, T, θ).

2. *For each $S \in \mathcal{S}^m$, there exists $\sigma_S : \mathcal{X} \mapsto \mathcal{S}$ such that*

$$\begin{aligned} &\sup_{S,\tau,\theta \in \mathcal{S}^m \times \mathcal{T}^{\mathcal{Y}} \times \Theta} \left| R_\omega(S, \tau; \theta) - R_\omega(\sigma_S, \tau; \theta) - C_2 \right| + \frac{\|\Psi_S - \Psi_{\sigma_S}\|^2}{Kd} \\ &\quad \le O_P\left(d^{1/2} \left(\frac{K^2 \log n}{\sqrt{n}} \right)^{\frac{1}{1+d}} \right), \end{aligned} \tag{6}$$

where $C_2 \in \mathbb{R}$ is constant in (S, τ, θ).

3. *Combining* (5) *and* (6) *yields*

$$\max_{S,T,\theta\in\mathcal{S}^m\times\mathcal{T}^n\times\Theta}\big|R_\omega(\sigma_S,\tau_T;\theta)-R_A(S,T;\theta)-C_1-C_2\big|+\frac{\|\Psi_S-\Psi_{\sigma_S}\|^2}{Kd}$$
$$+\frac{\|\Psi_T-\Psi_{\tau_T}\|^2}{Kd}=O_P\bigg(d^{1/2}\bigg(\frac{K^2\log n}{\sqrt{n}}\bigg)^{\frac{1}{1+d}}\bigg).$$

Remarks for Theorem 2. Theorem 2 states that any assignment S and T of latent variables to the rows and columns can be extended to the populations, such that the population exhibits a similar distribution of values in $\mathcal{S}$ and $\mathcal{T}$, and the population risk as a function of θ is close to the empirical risk.

The theorem may also be viewed as an oracle inequality, in that for any fixed S and T, minimizing $\theta\mapsto R_A(S,T,\theta)$ is approximately equivalent to minimizing $\theta\mapsto R_\omega(\sigma_S,\tau_T,\theta)$, as if the model ω were known. This implies that the best parametric approximation to ω can be learned out of all possible choices for σ_S and τ_T. However, it is not known whether the mappings $S\mapsto\sigma_S$ and $T\mapsto\tau_T$ are approximately onto.

The convergence of Ψ_S to Ψ_{σ_S} is established in Euclidean norm. This implies pointwise convergence at every continuity point of Ψ_{σ_S}, thus implying weak convergence and also convergence in Wasserstein distance.

The proof of Theorem 2 is contained in the supplemental material. It is similar to that of Theorem 1, but requires substantially more notation due to the additional parameters.

5. Proof of Theorem 1.

We present a sketch of the proof for Theorem 1, which defines the most important quantities. We then present helper lemmas and give the proof of the theorem.

5.1. *Proof sketch.*

Let $W\in[0,1]^{m\times n}$ denote the expectation of A, conditioned on the latent variables $x_1,\dots,x_m$ and $y_1,\dots,y_n$:

$$W_{ij}=\omega(x_i,y_j),\qquad i\in[m],j\in[m],$$

and let $\Phi_W(S,T)$ denote the conditional expectation of $\Phi_A(S,T)$:

$$\big[\Phi_W(S,T)\big]_{st}=\frac{1}{nm}\sum_{i=1}^m\sum_{j=1}^n W_{ij}1\{S_i=s,T_j=t\}.$$

Given co-cluster labels $S\in[K]^m$ and $T\in[K]^n$, let $1_{S=s}\in\{0,1\}^m$ and $1_{T=t}\in\{0,1\}^n$ denote the indicator variables

$$1_{S=s}(i)=\begin{cases}1, & \text{if } S_i=s,\\ 0, & \text{otherwise,}\end{cases}\quad\text{and}\quad 1_{T=t}(j)=\begin{cases}1, & \text{if } T_j=t,\\ 0, & \text{otherwise.}\end{cases}$$

Let $g_{T=t} \in [0,1]^m$ denote the vector $n^{-1}W1_{T=t}$, or

$$g_{T=t}(i) = \frac{1}{n}\sum_{j=1}^{n} W_{ij}1\{T_j = t\}.$$

It can be seen that the entries of $\Phi_W(S,T)$ can be written as

$$[\Phi_W(S,T)]_{st} = \frac{1}{m}\langle 1_{S=s}, g_{T=t}\rangle, \tag{7}$$

where $\langle\cdot,\cdot\rangle$ denotes inner product. Similarly, the entries of $\Phi_\omega(S,\tau)$ can be written as

$$[\Phi_\omega(S,\tau)]_{st} = \frac{1}{m}\langle 1_{S=\text{ s}}, g_{\tau=t}\rangle, \tag{8}$$

where $g_{\tau=t} \in [0,1]^m$ is the vector

$$g_{\tau=t}(i) = \int_{\mathcal{Y}} \omega(x_i, y)1\{\tau(y) = t\}\,dy, \qquad i \in [m].$$

The proof of Theorem 1 will require three main steps:

S1: In Lemma 1, a concentration inequality will be used to show that $\Phi_A(S,T) \approx \Phi_W(S,T)$ uniformly over all possible values of (S,T).

S2: For each $T \in [K]^n$, we will show there exists $\tau : \mathcal{Y} \mapsto [K]$ such that $g_{T=t} \approx g_{\tau=t}$ for $t \in [K]$. By (7) and (8), this will imply that $\Phi_W(S,T) \approx \Phi_\omega(S,\tau)$ uniformly for all $S \in [K]^m$. The mapping τ will also satisfy $\pi_T \approx \pi_\tau$ as well, so that T and τ have similar class frequencies.

S3: Analogous to S2, we will show that for each $S \in [K]^m$, there exists $\sigma : \mathcal{X} \mapsto [K]$ such that $\Phi_\omega(S,\tau) \approx \Phi_\omega(\sigma_S,\tau)$ uniformly over τ, and also that $\pi_S \approx \pi_{\sigma_S}$.

Steps S1 and S2 correspond to (1) in Theorem 1, while step S3 corresponds to (2).

Let G_T and G_τ denote the stacked vectors in $\mathbb{R}^{mK+K}$ given by

$$G_T = \left(\frac{g_{T=1}}{\sqrt{m}}, \ldots, \frac{g_{T=K}}{\sqrt{m}}, \pi_T\right) \quad \text{and} \quad G_\tau = \left(\frac{g_{\tau=1}}{\sqrt{m}}, \ldots, \frac{g_{\tau=K}}{\sqrt{m}}, \pi_\tau\right),$$

and let $\mathcal{G}_n$ and $\mathcal{G}$ denote the set of all possible values for G_T and G_τ:

$$\mathcal{G}_n = \{G_T : T \in [K]^n\} \quad \text{and} \quad \mathcal{G} = \{G_\tau : \tau \in \mathcal{Y} \mapsto [K]\}.$$

Step S2 is established by showing that the sets $\mathcal{G}_n$ and $\mathcal{G}$ converge in Hausdorff distance. This will require the following facts. The Hausdorff distance (in Euclidean norm) between two sets $\mathcal{B}_1$ and $\mathcal{B}_2$ is defined as

$$d_{\text{Haus}}(\mathcal{B}_1, \mathcal{B}_2) = \max\Big\{\sup_{B_1\in\mathcal{B}_1}\inf_{B_2\in\mathcal{B}_2}\|B_1 - B_2\|, \sup_{B_2\in\mathcal{B}_2}\inf_{B_1\in\mathcal{B}_1}\|B_1 - B_2\|\Big\}.$$

Given a Hilbert space $\mathbb{H}$ and a set $\mathcal{B} \subset \mathbb{H}$, let $\Gamma_{\mathcal{B}} : \mathbb{H} \mapsto \mathbb{R}$ denote the support function of $\mathcal{B}$, defined as

$$\Gamma_{\mathcal{B}}(H) = \sup_{B \in \mathcal{B}} \langle H, B \rangle.$$

It is known that the convex hull $\operatorname{conv}(\mathcal{B})$ equals the intersection of its supporting hyperplanes:

$$\operatorname{conv}(\mathcal{B}) = \{H' \in \mathbb{H} : \langle H', H \rangle \leq \Gamma_{\mathcal{B}}(H) \text{ for all } H \in \mathbb{H}\},$$

and that the Hausdorff distance between $\operatorname{conv}(\mathcal{B}_1)$ and $\operatorname{conv}(\mathcal{B}_2)$ is given by Schneider [(2013), Theorem 1.8.11], Aliprantis and Border [(2006), Corollary 7.59]:

$$d_{\mathrm{Haus}}(\operatorname{conv}(\mathcal{B}_1), \operatorname{conv}(\mathcal{B}_2)) = \sup_{H : \|H\| = 1} |\Gamma_{\mathcal{B}_1}(H) - \Gamma_{\mathcal{B}_2}(H)|. \tag{9}$$

To establish S2, Lemma 2 will show that

$$\sup_{H : \|H\| = 1} |\Gamma_{\mathcal{G}_n}(H) - \Gamma_{\mathcal{G}}(H)| = O_P(K(\log n) n^{-1/2}), \tag{10}$$

and Lemma 3 will show that

$$d_{\mathrm{Haus}}(\operatorname{conv}(\mathcal{G}), \mathcal{G}) = 0. \tag{11}$$

By (9) and (10), $\operatorname{conv}(\mathcal{G}_n)$ and $\operatorname{conv}(\mathcal{G})$ converge in Hausdorff distance, which by (11) implies that $\operatorname{conv}(\mathcal{G}_n)$ and $\mathcal{G}$ converge in Hausdorff distance. This implies that for each $G_T \in \mathcal{G}_n$, there exists $G_\tau \in \mathcal{G}$ such that $\max_T \|G_T - G_\tau\| \to 0$. This will establish S2, since $G_T \approx G_\tau$ implies by (7) and (8) that $\Phi_W(S, T) \approx \Phi_\omega(S, \tau)$ uniformly over $S \in [K]^m$, and it also implies that $\pi_T \approx \pi_\tau$ as well.

The proof of S3 will be similar to S2. It can be seen that $\Phi_\omega(S, \tau)$ and $\Phi_\omega(\sigma, \tau)$ can be written as

$$[\Phi_\omega(S, \tau)]_{st} = \langle f_{S=s}, 1_{\tau=t} \rangle \quad \text{and} \quad [\Phi_\omega(\sigma, \tau)]_{st} = \langle f_{\sigma=\mathrm{s}}, 1_{\tau=t} \rangle, \tag{12}$$

where the functions $f_{S=s}$, $1_{\tau=t}$, and $f_{\sigma=s}$ are given by

$$1_{\tau=t}(y) = \begin{cases} 1, & \text{if } \tau(y) = t, \\ 0, & \text{otherwise,} \end{cases}$$

$$f_{S=s}(y) = \frac{1}{m} \sum_{i=1}^{m} \omega(x_i, y) 1\{S_i = s\},$$

$$f_{\sigma=s}(y) = \int_{\mathcal{X}} \omega(x, y) 1\{\sigma(x) = s\} \, dx.$$

Analogous to S2, we will define sets F_S and F_σ given by

$$F_S = (f_{S=1}, \ldots, f_{S=K}, \pi_S) \quad \text{and} \quad F_\sigma = (f_{\sigma=1}, \ldots, f_{\sigma=K}, \pi_\sigma),$$

whose possible values are given by

$$\mathcal{F}_n = \{F_S : S \in [K]^m\}$$

and

$$\mathcal{F} = \{F_\sigma : \sigma \in \mathcal{X} \mapsto [K]\}.$$

Lemma 2 will show that the support functions $\Gamma_{\mathcal{F}_n}$ and $\Gamma_{\mathcal{F}}$ converge, and Lemma 3 will show that $d_{\text{Haus}}(\text{conv}(\mathcal{F}), \mathcal{F}) = 0$. Using (12), this will establish S3 by arguments that are analogous to those used to prove S2.

5.2. *Intermediate results for proof of Theorem 1.* Lemmas 1–3 will be used to prove Theorem 1, and are proven in Section 5.4.

Lemma 1 states that $\Phi_A \approx \Phi_W$ for all (S, T).

LEMMA 1. *Under the conditions of Theorem 1,*

$$\max_{S,T} \|\Phi_A(S,T) - \Phi_W(S,T)\|^2 = O_P\big((\log K)n^{-1}\big). \tag{13}$$

Lemma 2 states that the support functions of $\mathcal{G}$ and $\mathcal{G}_n$ and of $\mathcal{F}$ and $\mathcal{F}_n$ converge.

LEMMA 2. *Under the conditions of Theorem 1,*

$$\sup_{\|H\|=1} \big|\Gamma_{\mathcal{G}_n}(H) - \Gamma_{\mathcal{G}}(H)\big| \le O_P\big(K(\log n)n^{-1/2}\big), \tag{14}$$

$$\sup_{\|H\|=1} \big|\Gamma_{\mathcal{F}_m}(H) - \Gamma_{\mathcal{F}}(H)\big| \le O_P\big(K(\log m)m^{-1/2}\big), \tag{15}$$

which implies

$$d_{\text{Haus}}\big(\text{conv}(\mathcal{G}_n), \text{conv}(\mathcal{G})\big) \le O_P\big(K(\log n)n^{-1/2}\big),$$
$$d_{\text{Haus}}\big(\text{conv}(\mathcal{F}_m), \text{conv}(\mathcal{F})\big) \le O_P\big(K(\log m)m^{-1/2}\big).$$

Lemma 3 states that the sets $\mathcal{F}$ and $\mathcal{G}$ are essentially convex.

LEMMA 3. *It holds that*

$$d_{\text{Haus}}\big(\text{conv}(\mathcal{G}), \mathcal{G}\big) = 0, \tag{16}$$

$$d_{\text{Haus}}\big(\text{conv}(\mathcal{F}), \mathcal{F}\big) = 0. \tag{17}$$

5.3. *Proof of Theorem* 1. We bound $\|\Phi_W(S,T) - \Phi_\omega(S,\tau)\|^2$ uniformly over S, as follows:

$$\begin{aligned}
\|\Phi_W(S,T) - \Phi_\omega(S,\tau)\|^2 &= \sum_{s=1}^K \sum_{t=1}^K \bigl([\Phi_W(S,T)]_{st} - [\Phi_\omega(S,\tau)]_{st}\bigr)^2 \\
&= \sum_{s=1}^K \sum_{t=1}^K \frac{1}{m^2} \langle 1_{S=s}, g_{T=t} - g_{\tau=t}\rangle^2 \\
&\le \sum_{s=1}^K \sum_{t=1}^K \frac{1}{m^2} \|1_{S=s}\|^2 \|g_{T=t} - g_{\tau=t}\|^2 \\
&= \left(\sum_{s=1}^K \frac{1}{m}\|1_{S=s}\|^2\right)\left(\sum_{t=1}^K \frac{1}{m}\|g_{T=t} - g_{\tau=t}\|^2\right) \qquad (18)\\
&\le \left(\sum_{t=1}^K \frac{1}{m}\|g_{T=t} - g_{\tau=t}\|^2\right),
\end{aligned}$$

where (18) holds because $m^{-1}\sum_{s=1}^K \|1_{S=s}\|^2 = 1$.

By Lemma 2 and Lemma 3, it holds that

$$d_{\text{Haus}}\bigl(\text{conv}(\mathcal{G}_n), \mathcal{G}\bigr) = O_P\bigl(K(\log n)n^{-1/2}\bigr).$$

Given T, let $\tau \equiv \tau_T$ denote the minimizer of $\|G_T - G_\tau\| = \langle G_T - G_\tau, G_T - G_\tau\rangle$. It follows that

$$\begin{aligned}
\max_T \|G_T - G_\tau\|^2 &= \max_T \sum_{t=1}^K \frac{1}{m}\|g_{T=t} - g_{\tau=t}\|^2 + \|\pi_T - \pi_\tau\|^2 \qquad (19)\\
&= O_P\left(\frac{K^2 \log n}{n}\right).
\end{aligned}$$

Combining (13), (19) and (18) yields

$$\max_{S,T} \|\Phi_A(S,T) - \Phi_\omega(S,\tau_T)\|^2 + \|\pi_T - \pi_{\tau_T}\|^2 = O_P\left(\frac{K^2 \log n}{n}\right),$$

establishing (1).

The proof of (2) proceeds in similar fashion. The quantity $\|\Phi_\omega(S,\tau) - \Phi_\omega(\sigma,\tau)\|^2$ may be bounded uniformly over τ:

$$\begin{aligned}
\|\Phi_\omega(S,\tau) - \Phi_\omega(\sigma,\tau)\|^2 &= \sum_{s=1}^K \sum_{t=1}^K \bigl([\Phi_\omega(S,\tau)]_{st} - [\Phi_\omega(\sigma,\tau)]_{st}\bigr)^2 \\
&= \sum_{s=1}^K \sum_{t=1}^K \langle f_{S=s} - f_{\sigma=s}, 1_{\tau=t}\rangle^2 \qquad (20)
\end{aligned}$$

$$\leq \left(\sum_{s=1}^{K} \| f_{S=s} - f_{\sigma=s} \|^2 \right),$$

where all steps parallel the derivation of (18). It follows from Lemma 2 and 3 that $d_{\text{Haus}}(\text{conv}(\mathcal{F}_m), \mathcal{F}) = O_P(K(\log m)m^{-1/2})$. Given S, let $\sigma \equiv \sigma_S$ denote the minimizer of $\|F_S - F_\sigma\|$, so that

$$\max_S \quad \sum_{t=1}^{K} \| f_{S=s} - f_{\sigma=s} \|^2 + \|\pi_S - \pi_\sigma\|^2 = O_P\left(\frac{K^2 \log m}{m} \right). \tag{21}$$

Combining (21) and (20) yields

$$\max_{S,\tau} \left\| \Phi_\omega(S, \tau) - \Phi_\omega(\sigma_S, \tau) \right\|^2 + \|\pi_S - \pi_{\sigma_S}\|^2 = O_P\left(\frac{K^2 \log m}{m} \right),$$

establishing (2) and completing the proof.

5.4. *Proof of Lemmas 1–3*. The proof of Lemma 2 will rely on Lemma 4, which is a very slight modification of Lemma 4.3 in Biau, Devroye and Lugosi (2008). Lemma 4 is proven in Appendix B.

LEMMA 4. *Let $\mathbb{H}$ denote a Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $g : \mathcal{Y} \mapsto \mathbb{H}$, and let $y_1, \dots, y_n \in \mathcal{Y}$ be i.i.d. Let $L_n : \mathbb{H}^K \mapsto \mathbb{R}$ be defined as*

$$L_n(H) = \frac{1}{n} \sum_{j=1}^{n} \max_{k \in [K]} \langle h_k, g(y_j) \rangle, \qquad H = (h_1, \dots, h_K) \in \mathbb{H}^K. \tag{22}$$

Let $\mathcal{H} = \{H \in \mathbb{H}^K : \|h_k\| \leq 1, t \in [K]\}$. It holds that

$$\mathbb{E} \sup_{H \in \mathcal{H}} \left| L_n(H) - \mathbb{E} L_n(H) \right| \leq 2K \left(\frac{\mathbb{E}\|g(y)\|^2}{n} \right)^{1/2}.$$

To prove Lemma 3, we will require a theorem (by Carathéodory) for finite dimensional convex hulls.

THEOREM 3 (Schneider (2013), Theorem 1.1.4). *If $\mathcal{B} \subset \mathbb{R}^d$ and $x \in$ conv$(\mathcal{B})$, there exists $B_1, \dots, B_{d+1} \in \mathcal{B}$ such that $x \in$ conv$\{B_1, \dots, B_{d+1}\}$.*

PROOF OF LEMMA 1. Given (S, T), let $\Delta \in [-1, 1]^{K \times K}$ denote the quantity

$$\Delta_{st} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} - W_{ij}) 1(S_i = s, T_j = t).$$

It holds that $\mathbb{E}[\Delta | W] = 0$, and by Hoeffding's inequality,

$$\mathbb{P}\big(|\Delta_{st}| \geq \varepsilon | W\big) \leq 2e^{-2nm\varepsilon^2}, \qquad s, t \in [K].$$

Conditioned on W, each entry of Δ is independent of the others. Given $\delta \in [-1,1]^{K \times K}$, it follows that

$$\mathbb{P}(\Delta = \delta | W) = \prod_{s=1}^{K} \prod_{t=1}^{K} \mathbb{P}(\Delta_{st} = \delta_{st} | W) \leq 2 \exp\Biggl(-2nm \sum_{s=1}^{K} \sum_{t=1}^{K} \delta_{st}^2\Biggr).$$

Let B denote the set

$$B = \biggl\{\delta \in [-1,1]^{K \times K} : \sum_{s,t} \delta_{st}^2 \geq \varepsilon, \delta \in \operatorname{supp}(\Delta)\biggr\}.$$

The cardinality of B is smaller than the support of Δ, which is less than $(nm)^{K^2}$ when conditioned on W. It follows by a union bound over B that

$$\begin{aligned}\mathbb{P}(\Delta \in B | W) &\leq 2|B| e^{-2nm\varepsilon} \\ &\leq 2(nm)^{K^2} e^{-2nm\varepsilon}.\end{aligned}$$

It can be seen that $\|\Phi_A(S,T) - \Phi_W(S,T)\|^2 = \sum_{s,t} \Delta_{st}^2$, implying that $\Delta \in B$ is equivalent to the event that $\|\Phi_A(S,T) - \Phi_W(S,T)\|^2 \geq \varepsilon$. A union bound over all S, T implies that

$$\mathbb{P}\Bigl(\max_{S,T} \bigl\|\Phi_A(S,T) - \Phi_W(S,T)\bigr\|^2 \geq \varepsilon\Bigr) \leq 2K^{n+m} (nm)^{K^2} e^{-2nm\varepsilon}.$$

Letting $\varepsilon = C(1 + n/m)(\log K)n^{-1}$ for some C proves the lemma. □

PROOF OF LEMMA 2. Let $g_y \in [0,1]^m$ denote the column of W induced by $y \in \mathcal{Y}$, and let $f_x \in [0,1]^{\mathcal{Y}}$ denote the row of ω corresponding to $x \in \mathcal{X}$:

$$g_y(i) = \omega(x_i, y), \qquad i \in [m] \quad \text{and} \quad f_x(y) = \omega(x,y), \qquad y \in \mathcal{Y}.$$

Algebraic manipulation shows that $g_{T=t}$, $g_{\tau=t}$, $f_{S=s}$, and $f_{\sigma=s}$ can be written as

$$g_{T=t} = \frac{1}{n} \sum_{j=1}^{n} g_{y_j} 1(T_j = t), \qquad g_{\tau=t} = \int_{\mathcal{Y}} g_y 1\bigl(\tau(y) = t\bigr)\, dy,$$

$$f_{S=s} = \frac{1}{m} \sum_{i=1}^{m} f_{x_i} 1(S_i = s), \qquad f_{\sigma=s} = \int_{\mathcal{X}} f_x 1\bigl(\sigma(x) = s\bigr)\, dx.$$

Given $H = (h_1, \dots, h_K, \pi_H)$, it follows that the inner products $\langle H, G_T \rangle$, $\langle H, G_\tau \rangle$, $\langle H, F_S \rangle$, and $\langle H, F_\sigma \rangle$ equal

$$\langle H, G_T \rangle = \frac{1}{n} \sum_{j=1}^{n} \biggl[\biggl\langle h_{T_j}, \frac{g_{y_j}}{\sqrt{m}} \biggr\rangle + \pi_H(T_j)\biggr],$$

$$\langle H, G_\tau \rangle = \int_{\mathcal{Y}} \biggl\langle h_{\tau(y)}, \frac{g_y}{\sqrt{m}} \biggr\rangle + \pi_H\bigl(\tau(y)\bigr)\, dy,$$

$$\langle H, F_S\rangle = \frac{1}{m}\sum_{i=1}^{m}\big[\langle h_{S_i}, f_{x_i}\rangle + \pi_H(S_i)\big],$$

$$\langle H, F_\sigma\rangle = \int_{\mathcal{X}} \langle h_{\sigma(x)}, f_x\rangle + \pi_H\big(\sigma(x)\big)\,dx,$$

and hence that the support functions equal

$$\Gamma_{\mathcal{G}_n}(H) = \frac{1}{n}\sum_{j=1}^{n}\max_{k\in[K]}\left\langle h_k, \frac{g_{y_j}}{\sqrt{m}}\right\rangle + \pi_H(k), \tag{23}$$

$$\Gamma_{\mathcal{G}}(H) = \int_{\mathcal{Y}}\max_{k\in[K]}\left\langle h_k, \frac{g_y}{\sqrt{m}}\right\rangle + \pi_H(k)\,dy, \tag{24}$$

$$\Gamma_{\mathcal{F}_m}(H) = \frac{1}{m}\sum_{i=1}^{m}\max_{k\in[K]}\langle h_k, f_{x_i}\rangle + \pi_H(k),$$

$$\Gamma_{\mathcal{F}}(H) = \int_{\mathcal{X}}\max_{k\in[K]}\langle h_k, f_x\rangle + \pi_H(k)\,dx, \tag{25}$$

which implies that $\mathbb{E}\Gamma_{\mathcal{G}_n}(H) = \Gamma_{\mathcal{G}}(H)$ and $\mathbb{E}\Gamma_{\mathcal{F}_m}(H) = \Gamma_{\mathcal{F}}(H)$.

To show (14), we observe that $\Gamma_{\mathcal{G}_n}$ can be rewritten as

$$\Gamma_{\mathcal{G}_n}(H) = \frac{1}{n}\sum_{j=1}^{n}\max_{k\in[K]}\left\langle \begin{bmatrix} h_k \\ \pi_H(k)\end{bmatrix}, \begin{bmatrix} m^{-1/2}g_{y_j} \\ 1\end{bmatrix}\right\rangle,$$

which matches (22) so that Lemma 4 can be applied. Applying Lemma 4 results in

$$\mathbb{E}\sup_{\|H\|=1}\big|\Gamma_{\mathcal{G}_n}(H) - \Gamma_{\mathcal{G}}(H)\big| \le \frac{4K}{\sqrt{n}}, \tag{26}$$

where we have used $\{H : \|H\| = 1\} \subset \mathcal{H}$ and $\|[\begin{smallmatrix} m^{-1/2}g_{y_j} \\ 1\end{smallmatrix}]\|^2 \le 2$.

Let $Z(y_1, \dots, y_n) = \sup_{\|H\|=1}|\Gamma_{\mathcal{G}_n}(H) - \Gamma_{\mathcal{G}}(H)|$. For $\ell \in [n]$, changing y_ℓ to y'_ℓ changes Z by at most $4/n$. Applying McDiarmid's inequality yields

$$\mathbb{P}\big(|Z - \mathbb{E}Z| \ge \varepsilon\big) \le 2e^{-2\varepsilon^2 n/8}.$$

Letting $\varepsilon = n^{-1/2}\log n$ implies that $Z - \mathbb{E}Z = O_P(n^{-1/2}\log n)$, which combined with (26) implies (14).

To show (15), we observe that

$$\Gamma_{\mathcal{F}_m}(H) = \frac{1}{m}\sum_{i=1}^{m}\max_{k\in[K]}\left\langle \begin{bmatrix} h_k \\ \pi_H(k)\end{bmatrix}, \begin{bmatrix} f_{x_i} \\ 1\end{bmatrix}\right\rangle,$$

so that Lemma 4 and McDiarmid's inequality can be used analogously to the proof of (14). □

We divide the proof of Lemma 3 into two sub-lemmas, one showing (16) and the other showing (17). This is because the proof of (17) will require additional work, due to the fact that the elements of $\mathcal{F}$ are infinite dimensional. We prove Lemma 5 here, and defer proof of Lemma 6 to Appendix A.

LEMMA 5. *For each* $G^* \in \operatorname{conv}(\mathcal{G})$, *there exists* $G_1, G_2, \ldots \in \mathcal{G}$ *such that* $\lim_{\ell\to\infty} \|G^* - G_\ell\| = 0$.

LEMMA 6. *For each* $F^* \in \operatorname{conv}(\mathcal{F})$, *there exists* $F_1, F_2, \ldots \in \mathcal{F}$ *such that* $\lim_{\ell\to\infty} \|F^* - F_\ell\| = 0$.

PROOF OF LEMMA 5. Recall the definition of $g_y \in [0,1]^m$ as defined in the proof of Lemma 4:

$$g_y(i) = \omega(x_i, y), \qquad i \in [m],$$

and that $g_{\tau=t}$ can be written as

$$g_{\tau=t} = \int_{\mathcal{Y}} g_y 1\{\tau(y) = t\}\, dy.$$

We note the following properties of $\{g_y : y \in \mathcal{Y}\}$:

P1: Each $G^* \in \operatorname{conv}(\mathcal{G})$ is a finite convex combination of elements in $\mathcal{G}$. This holds by Theorem 3, since $\mathcal{G}$ is a subset of $[0,1]^{mK+K}$, a finite dimensional space.

P2: For all ε, there exists a finite set $\mathcal{B}$ that is an ε-cover of $\{g_y : y \in \mathcal{Y}\}$ in Euclidean norm. This holds because $\{g_y : y \in \mathcal{Y}\}$ is a subset of the unit cube $[0,1]^m$.

By P1, each $G^* \in \operatorname{conv}(\mathcal{G})$ can be written as a finite convex combination of elements in $\mathcal{G}$, so that for some integer $N > 0$ there exists $G_{\tau_1}, \ldots, G_{\tau_N} \in \mathcal{G}$ such that

$$G^* = \sum_{i=1}^{N} \eta_i G_{\tau_i},$$

where η is in the N-dimensional unit simplex. It follows that for some $\mu : \mathcal{Y} \mapsto [0,1]^K$ satisfying $\sum_k \mu_k(y) = 1$ for all y, $G^* \equiv (g_1^*, \ldots, g_K^*, \pi_G^*)$ satisfies

$$g_k^* = \int_{\mathcal{Y}} g_y \mu_k(y)\, dy \quad \text{and} \quad \pi_G^*(k) = \int_{\mathcal{Y}} \mu_k(y)\, dy, \qquad k \in [K].$$

We now construct $\tau : \mathcal{X} \mapsto [K]$ inducing $G_\tau \in \mathcal{G}$ which approximates $G^* \in \operatorname{conv}(\mathcal{G})$. By P2, let $\mathcal{B}$ denote an ε-cover of $\{g_y : y \in \mathcal{Y}\}$, and enumerate its elements as $b_1, \ldots, b_{|\mathcal{B}|}$. For each $y \in \mathcal{Y}$, let $\ell : \mathcal{Y} \mapsto [|\mathcal{B}|]$ assign y to its closest member in $\mathcal{B}$, so that $\|g_y - b_{\ell(y)}\| \le \varepsilon$. For $i = 1, \ldots, |\mathcal{B}|$, let $\mathcal{Y}_i$ denote the set $\{y : \ell(y) = i\}$. Arbitrarily divide each region $\mathcal{Y}_i$ into K disjoint subregions

$\mathcal{Y}_{i1}, \dots, \mathcal{Y}_{iK}$ such that $\bigcup_k \mathcal{Y}_{ik} = \mathcal{Y}_i$, where the measure of each subregion is given by

$$\int_{\mathcal{Y}_{ik}} 1\, dy = \int_{\mathcal{Y}_i} \mu_k(y)\, dy, \qquad k \in [K]. \tag{27}$$

Let $\tau : \mathcal{Y} \mapsto [K]$ assign each region $\mathcal{Y}_{ik}$ to k, so that

$$\tau(y) = k \qquad \text{for all } y \in \mathcal{Y}_{ik}, i = 1, \dots, |\mathcal{B}|.$$

By (27), it holds that $\pi_\tau = \pi_G^*$, and also that

$$\begin{aligned} g_{\tau=k} - g_k^* &= \int_{\mathcal{Y}} g_y [1\{\tau(y) = k\} - \mu_k(y)]\, dy \\ &= \int_{\mathcal{Y}} [b_{\ell(y)} + g_y - b_{\ell(y)}][1\{\tau(y) = k\} - \mu_k(y)]\, dy \\ &= \sum_{i=1}^{|\mathcal{B}|} b_i \underbrace{\left[\int_{\mathcal{Y}_{ik}} 1\, dy - \int_{\mathcal{Y}_i} \mu_k(y)\, dy\right]}_{=0 \text{ by } (27)} \\ &\qquad + \int_{\mathcal{Y}} (g_y - b_\ell(y))[1\{\tau(y) = k\} - \mu_k(y)]\, dy \\ &= 0 + \int_{\mathcal{Y}} (g_y - b_{\ell(y)})[1\{\tau(y) = k\} - \mu_k(y)]\, dy, \end{aligned}$$

which implies that

$$\begin{aligned} \|g_{\tau=k} - g_k^*\| &\le \left\| \int_{\mathcal{Y}} (g_y - b_{\ell(y)}) 1\{\tau(y) = k\}\, dy \right\| + \left\| \int_{\mathcal{Y}} (g_y - b_{\ell(y)}) \mu_k(y)\, dy \right\| \\ &\le 2 \int_{\mathcal{Y}} \|g_y - b_{\ell(y)}\|\, dy \\ &\le 2\varepsilon. \end{aligned}$$

It follows that $\|G_\tau - G^*\|^2 = \sum_{k=1}^K m^{-1} \|g_{\tau=k} - g_k^*\|^2 + \|\pi_\tau - \pi_G^*\|^2 \le 4K\varepsilon^2 m^{-1}$, and hence that $\lim_{\varepsilon \to 0} \|G_\tau - G^*\| = 0$, proving the lemma. □

PROOF OF LEMMA 3. Lemma 3 follows immediately from Lemmas 5 and 6, which establish (16) and (17), respectively. □

6. Simulations.

Simulations were run for a parameterized class of generative graphons $\{\omega_\beta : \beta \ge 1\}$, in which $\omega_\beta : [0, 1]^2 \mapsto [0, 1]$ is defined as

$$\begin{aligned} \omega_\beta(x, y) &= f_\beta(x) f_\beta(y) + 1/2, \\ f_\beta(x) &= Z_\beta^{-1} \left(\frac{x^\beta}{x^\beta + (1 - x)^\beta} - \frac{1}{2} \right), \qquad 0 \le x \le 1, \\ Z_\beta &= 4 \int_0^{1/2} \left| \frac{x^\beta}{x^\beta + (1 - x)^\beta} - \frac{1}{2} \right| dx. \end{aligned} \tag{28}$$

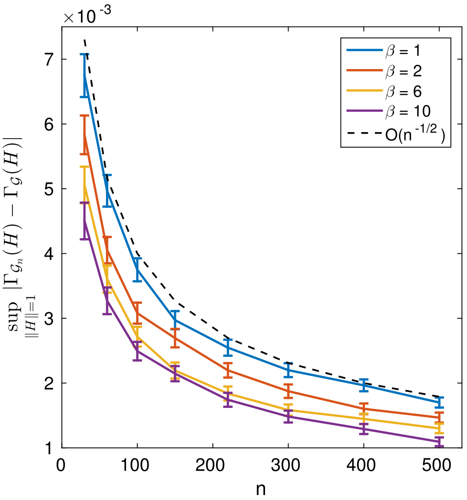
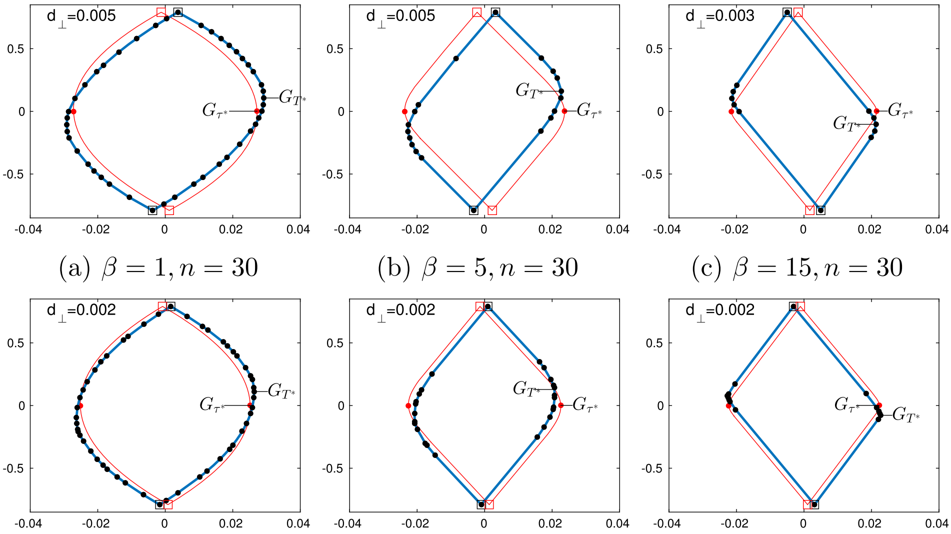


FIG. 1. *Simulated averages* (*and* 2 *standard errors*) *for* $\sup|\Gamma_{\mathcal{G}_n} - \Gamma_{\mathcal{G}}|$, *equaling the Hausdorff distance between the convex hulls of* $\mathcal{G}_n$ *and* $\mathcal{G}$. *The simulations used* ω_β *given by* (28), *with* $m = n \in [30, 500]$, $\beta \in \{1, 2, 6, 10\}$, $K = 2$ *and* 600 *simulations per data point. A dotted reference line equaling* $0.04n^{-1/2}$ *is shown, suggesting that the rate in n predicted by Lemma* 2 *may be correct up to logarithmic factors.*

The function $f_\beta(x)$ is a sigmoid that is proportional to $x - 1/2$ for $\beta = 1$, and to $1\{x > 1/2\} - 1/2$ when $\beta \to \infty$. This means that the graphon ω_β approaches a two-class blockmodel as $\beta \to \infty$. The constant Z_β enforces that the quantities $\int_a^{a+1/2} \int_b^{b+1/2} \omega_\beta(x, y)\,dx\,dy$ for $a, b \in \{0, 1/2\}$ are all constant over β. These quantities correspond to the within- and between-class densities when $\beta = \infty$.

We will use simulation to verify Lemma 2, by investigating the behavior of $\operatorname{conv}(\mathcal{G})$ and $\operatorname{conv}(\mathcal{G}_n)$. For ω_β given by (28) and $K = 2$, the sets $\mathcal{G}$ and $\mathcal{G}_n$ lie in parallel 2-dimensional affine subspaces of $\mathbb{R}^{mK+K}$, so that $d_{\text{Haus}}(\operatorname{conv}(\mathcal{G}), \operatorname{conv}(\mathcal{G}_n))$ can be efficiently computed by numerical evaluation of $\Gamma_{\mathcal{G}_n}$ and $\Gamma_{\mathcal{G}}$ as given by (23). Figure 1 shows simulation results for this quantity. We see that the observed rate seems to follow $O(n^{-1/2})$, for all values of β that were simulated. Up to logarithmic factors, this matches the dependence on n that was predicted by Lemma 2.

Since $\mathcal{G}_n$ and $\mathcal{G}$ lie in parallel 2-dimensional affine subspaces, they can be visualized in $\mathbb{R}^2$, up to an "out of the page" translation between the sets. Figure 2 shows simulated instances of $\operatorname{conv}(\mathcal{G}_n)$ and $\operatorname{conv}(\mathcal{G})$ for various choices of β and n. We observed that the extremal points of the convex hulls corresponded to mappings $T \in [K]^n$ or $\tau : \mathcal{Y} \mapsto [K]$ which had a common form, assigning $y \in [0, c]$ to cluster 1 and $y \in (c, 1]$ to cluster 2 (or vice versa) for some $c \in [0, 1]$. For large β where ω_β approached a blockmodel, the extremal points (except for two points corresponding to $c = 0$ and $c = 1$) concentrated around the assignment



(a) $\beta = 1, n = 30$ (b) $\beta = 5, n = 30$ (c) $\beta = 15, n = 30$

(d) $\beta = 1, n = 100$ (e) $\beta = 5, n = 100$ (f) $\beta = 15, n = 100$

FIG. 2. *2-D projection of simulated instances of* $\operatorname{conv}(\mathcal{G}_n)$ (*in blue, with extremal points marked*) *and* $\operatorname{conv}(\mathcal{G})$ (*in red*). *The projection was lossless up to an "out of the page" offset of distance* $d_\perp$ *between the sets. Points* G_{T^*} *and* G_{τ^*} *were induced by assignment of* $y \in [0, 1/2]$ *to class* 1 *and* $y \in (1/2, 1]$ *to class* 2. *Square markers correspond to assignment of all* $y \in [0, 1]$ *to a single cluster. Each plot was drawn by evaluating the support functions* $\Gamma_{\mathcal{G}_n}$ *and* $\Gamma_{\mathcal{G}}$ *at* 1720 *uniformly spaced directions in a* 2*-dimensional plane.*

$c = 1/2$. We speculate that this may have implications for approximation of ω_β by co-blockmodels when fitting by methods that converge to a local optima.

7. Future work. In this work, we have considered approximating a nonsmooth bipartite graphon by a blockmodel. Various possibilities for future work include the following.

Sparse graphs. Theorems 1 and 2 remain true if $\omega \equiv \omega_n$ is allowed to depend on n; however, the results are vacuous if the density of ω_n goes to zero as n increases. As an alternative to Theorem 1 for such settings, one might study convergence of $\rho_n^{-1}\|\Phi_A(S, T) - \Phi_\omega(\sigma_S, \sigma_T)\|$, where ρ_n denotes the expected density $\int \omega_n(x, y)\,dx\,dy$. We conjecture that when $\rho_n \gg \log n$, the approach of this paper can be adapted to show convergence under weaker restrictions (and perhaps slower rates) than those made in previous works such as Klopp, Tsybakov and Verzelen (2015), where $\sup_{x,y} \omega_n(x, y) = O(\rho_n)$ was required.

Construction of σ_S *and* τ_T. Theorem 1 states that for any $S \in [K]^m$ and $T \in [K]^n$, there exists $\sigma_S : \mathcal{X} \mapsto [K]$ and $\tau_T : \mathcal{Y} \mapsto [K]$ such that $\Phi_A(S, T) \approx$

$\Phi_\omega(\sigma, \tau)$. In the proof, τ_T is chosen to minimize $\|G_T - G_\tau\|$ over all $G_\tau \in \mathcal{G}$, and σ_S is chosen to minimize $\|F_S - F_\sigma\|$ over all $F_\sigma \in \mathcal{F}$. The proof is nonconstructive, as it does not explain how to actually compute σ_S and τ_T. Theorem 2 is similarly nonconstructive as well.

We propose a construction method for a limited case. Given a low rank ω that is approximated by a blockmodel with small K, it may be possible to efficiently find σ_S and τ_T that achieves Theorem 1 without resorting to heuristics. In this setting, we can construct conv($\mathcal{G}$) by numerically evaluating the support function $\Gamma_{\mathcal{G}}$,

$$\Gamma_{\mathcal{G}}(H) = \int_{\mathcal{Y}} \max_k \left\langle h_k, \frac{g_y}{\sqrt{m}} \right\rangle + \pi_H(k)\, dy, \tag{29}$$

in enough directions H to sufficiently cover the low-dimensional space spanning $\mathcal{G}$. Examples of conv($\mathcal{G}$) $= \mathcal{G}$ computed this way can be seen in Figure 2.

For each supporting hyperplane found in this manner, an extremal point of $\mathcal{G}$ and its underlying assignment τ is automatically computed as the "arg max" of (29). By a standard convex optimization formulation, the point $G_{\tau_T} \in$ conv($\mathcal{G}$) that minimizes $\|G_T - G_{\tau_T}\|$ can be computed as a convex combination of the extremal points, and the assignment τ_T can then be constructed by switching (for increasingly fine discretizations of $\mathcal{Y}$) between the extremal assignments according to their mixture weights. The assignment σ_S can be found analogously.

APPENDIX A: PROOF OF LEMMA 6

To prove Lemma 6, we require some results on Hilbert–Schmidt integral operators. A kernel function $\omega : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is Hilbert–Schmidt if it satisfies

$$\int_{\mathcal{X}\times\mathcal{Y}} |\omega(x, y)|^2\, dx\, dy < \infty.$$

It can be seen that ω defined by Definition 1 is Hilbert–Schmidt. Let Ω denote the integral operator induced by ω, given by

$$(\Omega f)(x) = \int_{\mathcal{Y}} \omega(x, y) f(y)\, dy.$$

It is known that a Hilbert–Schmidt operator Ω is a limit (in operator norm) of a sequence of finite rank operators, so that its kernel ω has singular value decomposition given by

$$\omega(x, y) = \sum_{q=1}^{\infty} \lambda_q u_q(x) v_q(y),$$

where $\{u_q\}_{q=1}^{\infty}$ and $\{v_q\}_{q=1}^{\infty}$ are sets of orthonormal functions mapping $\mathcal{X} \mapsto \mathbb{R}$ and $\mathcal{Y} \mapsto \mathbb{R}$, and $\lambda_1, \lambda_2, \ldots$ are scalars decreasing in magnitude and satisfying $\sum_{q=1}^{\infty} \lambda_q^2 < \infty$.

PROOF OF LEMMA 6. Recall the definition of $f_x : \mathcal{Y} \mapsto [0, 1]$ as defined in the proof of Lemma 4:

$$f_x(y) = \omega(x, y),$$

and that $f_{\sigma=s}$ can be written as

$$f_{\sigma=s} = \int_{\mathcal{X}} f_x 1\{\sigma(x) = s\}\, dx.$$

Because $\{f_x : x \in \mathcal{X}\}$ is not finite dimensional, the arguments of Lemma 5 do not directly apply. To circumvent this, we will approximate the space $\mathcal{F}$ by a finite dimensional $\hat{\mathcal{F}}$, such that the convex hulls $\operatorname{conv}(\mathcal{F})$ and $\operatorname{conv}(\hat{\mathcal{F}})$ converge.

For $Q = 1, 2, \ldots$, let ω_Q be the best rank-Q approximation to ω,

$$\omega_Q(x, y) = \sum_{q=1}^{Q} \lambda_q u_q(x) v_q(y).$$

Given $D > 0$, let $\hat{u}_q$ denote a truncation of u_q, defined as

$$\hat{u}_q^D(x) = \begin{cases} D, & \text{if } u_q(x) \ge D, \\ u_q(x), & \text{if } -D \le u_q(x) \le D, \\ -D, & \text{if } u_q(x) \le -D, \end{cases}$$

and let $\hat{\omega} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be defined as

$$\hat{\omega}(x, y) = \sum_{q=1}^{Q} \lambda_q \hat{u}_q(x) v_q(y).$$

Let $\hat{f}_x : \mathcal{Y} \mapsto \mathbb{R}$ and $\hat{f}_{\sigma=s}$ be defined as

$$\hat{f}_x(y) = \hat{\omega}(x, y) \quad \text{and} \quad \hat{f}_{\sigma=s} = \int_{\mathcal{X}} \hat{f}_x 1\{\sigma(x) = s\}\, dx.$$

Let $\hat{F}_\sigma$ and $\hat{\mathcal{F}}$ be defined as

$$\hat{F}_\sigma = (\hat{f}_{\sigma=1}, \ldots, \hat{f}_{\sigma=K}, \pi_\sigma) \quad \text{and} \quad \hat{\mathcal{F}} = \{\hat{F}_\sigma : \sigma \in [K]^{\mathcal{X}}\}.$$

We bound the difference $\|\hat{f}_x - f_x\|^2$:

$$\|\hat{f}_x - f_x\|^2 = \sum_{q=1}^{Q} \lambda_q^2 \bigl(\hat{u}_q(x) - u_q(x)\bigr)^2 + \sum_{q=Q+1}^{\infty} \lambda_q^2 u_q(x)^2,$$

where we used the fact $f_x = \sum_{q=1}^{\infty} \lambda_q u_q(x) v_q$, and that the functions $\{v_q\}$ are orthonormal. It follows that

$$\int_{\mathcal{X}} \|\hat{f}_x - f_x\|^2\, dx = \sum_{q=1}^{Q} \lambda_q^2 \int_{\mathcal{X}} \bigl(\hat{u}_q(x) - u_q(x)\bigr)^2\, dx + \sum_{q=Q+1}^{\infty} \lambda_q^2 \int_{\mathcal{X}} u_q(x)^2\, dx$$

$$= \sum_{q=1}^{Q} \lambda_q^2 \int_{\mathcal{X}} \big(\hat{u}_q(x) - u_q(x)\big)^2 \, dx + \sum_{q=Q+1}^{\infty} \lambda_q^2$$

$$\le \sum_{q=1}^{Q} \lambda_q^2 \int_{x:|u_q(x)| \ge D} u_q(x)^2 \, dx + \sum_{q=Q+1}^{\infty} \lambda_q^2,$$

whence it can be seen that

$$\lim_{\min(Q,D)\to\infty} \int_{\mathcal{X}} \|\hat{f}_x - f_x\|^2 \, dx = 0.$$

We use this result to bound $\|\hat{f}_{\sigma=s} - f_{\sigma=s}\|$:

$$\begin{aligned} \max_{s,\sigma} \|\hat{f}_{\sigma=s} - f_{\sigma=s}\|^2 &= \max_{s,\sigma} \left\| \int_{\mathcal{X}} (\hat{f}_x - f_x) 1_{\sigma=s}(x) \, dx \right\|^2 \\ &\le \int_{\mathcal{X}} \|\hat{f}_x - f_x\|^2 \, dx \\ &\to 0 \qquad \text{as } \min(Q, D) \to \infty. \end{aligned}$$

Since $\|\hat{F}_\sigma - F_\sigma\|^2 = \sum_{k=1}^{K} \|\hat{f}_{\sigma=k} - f_{\sigma=k}\|^2 + \|\pi_\sigma - \pi_\sigma\|^2$, it follows that for any $\varepsilon > 0$, there exists (Q, D) inducing $\hat{\mathcal{F}} = \{\hat{F}_\sigma : \sigma \in [K]^{\mathcal{X}}\}$ such that

$$\sup_\sigma \|\hat{F}_\sigma - F_\sigma\| \le \varepsilon, \tag{30}$$

so that the support functions of $\mathcal{F}$ and $\hat{\mathcal{F}}$ can be bounded by

$$\begin{aligned} \sup_{H:\|H\|=1} \big|\Gamma_{\mathcal{F}}(H) - \Gamma_{\hat{\mathcal{F}}}(H)\big| &\le \max_{\|H\|=1,\sigma} \big|\langle H, F_\sigma - \hat{F}_\sigma \rangle\big| \\ &\le \max_\sigma \|F_\sigma - \hat{F}_\sigma\| \\ &\le \varepsilon, \end{aligned}$$

implying that

$$d_{\text{Haus}}\big(\text{conv}(\mathcal{F}), \text{conv}(\hat{\mathcal{F}})\big) \le \varepsilon, \tag{31}$$

which in turn implies that for any $F^* \in \text{conv}(\mathcal{F})$, there exists $\hat{F}^* \in \text{conv}(\hat{\mathcal{F}})$ such that $\|F^* - \hat{F}^*\| \le \varepsilon$.

For any choice of (Q, D), we observe that properties P1 and P2 as described in Lemma 5 for $\mathcal{G}$ also hold for $\hat{\mathcal{F}}$:

P1: Each $\hat{F} \in \text{conv}(\hat{\mathcal{F}})$ is a finite convex combination of elements in $\hat{\mathcal{F}}$. This holds because each $\hat{f}_x$ can be written as

$$\hat{f}_x = \sum_{q=1}^{Q} \lambda_q \hat{\mu}_q(x) v_q,$$

showing that $\{\hat{f}_x : x \in \mathcal{X}\}$ is a finite dimensional subspace of $\mathcal{Y} \mapsto \mathbb{R}$, and hence $\hat{\mathcal{F}}$ is as well, allowing Theorem 3 to be applied.

P2: For all ε, there exists a finite ε-cover of $\{\hat{f}_x : x \in \mathcal{X}\}$ in Euclidean norm. This holds because the set $\{\hat{u}(x) : x \in \mathcal{X}\}$ is a subset of the hypercube $[-D, D]^Q$.

As a result, the same arguments used to prove Lemma 5 also apply to $\hat{\mathcal{F}}$, implying that for each $\hat{F} \in \operatorname{conv}(\hat{\mathcal{F}})$, there exists for any $\varepsilon > 0$ a mapping $\sigma : \mathcal{X} \mapsto [K]$ such that

$$\|\hat{F}_\sigma - \hat{F}\|^2 \leq 4K\varepsilon^2. \tag{32}$$

It thus follows that for any $\varepsilon > 0$ and $F^* \in \operatorname{conv}(\mathcal{F})$, there exists $\hat{F}^* \in \operatorname{conv}(\hat{\mathcal{F}})$ and $\sigma : \mathcal{X} \mapsto [K]$ such that

$$\begin{aligned}\|F^* - F_\sigma\| &\leq \underbrace{\|F^* - \hat{F}^*\|}_{\leq \varepsilon \text{ by (31)}} + \underbrace{\|\hat{F}^* - \hat{F}_\sigma\|}_{\leq 4K\varepsilon^2 \text{ by (32)}} + \underbrace{\|\hat{F}_\sigma - F_\sigma\|}_{\leq \varepsilon \text{ by (30)}} \\ &\leq 2\varepsilon + 4\varepsilon^2 K.\end{aligned}$$

As a result, it follows that there exists $F_1, F_2, \ldots \in \mathcal{F}$ such that $\lim_{i\to\infty} \|F^* - F_i\| = 0$. □

APPENDIX B: PROOF OF LEMMA 4

To prove Lemma 4, we will use a result from Biau, Devroye and Lugosi (2008). A proof is given in the supplemental material for self-completeness.

LEMMA 7 (Biau, Devroye and Lugosi (2008), Lemma 4.3). *Let $\mathbb{H}$ denote a Hilbert space, and let $g : \mathcal{Y} \mapsto \mathbb{H}$. Let $y_1, \ldots, y_n \in \mathcal{Y}$ be i.i.d., and let $L_n : \mathbb{H}^K \mapsto \mathbb{R}$ be defined as follows*:

$$L_n(H) = \frac{1}{n}\sum_{j=1}^n \max_{t\in[K]} \langle h_t, g(y_j)\rangle, \qquad H = (h_1, \ldots, h_K) \in \mathbb{H}^K.$$

Let $\mathcal{B} = \{H \in \mathbb{H}^K : \|h_k\| \leq 1, k \in [K]\}$. Then the following three statements hold:

$$\mathbb{E} \sup_{H\in\mathcal{B}} L_n(H) - \mathbb{E}L_n(H) \leq 2\mathbb{E} \sup_{H\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^n \varepsilon_j \max_{t\in[K]} \langle h_t, g(y_j)\rangle, \tag{33}$$

where $\varepsilon_1, \ldots, \varepsilon_j \overset{\text{i.i.d.}}{\sim} \pm 1$ *w.p.* 1/2,

$$\mathbb{E} \sup_{H\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^n \varepsilon_j \max_{t\in[K]} \langle h_t, g(y_j)\rangle \leq 2K\mathbb{E} \sup_{\|h\|=1} \frac{1}{n}\sum_{j=1}^n \varepsilon_j \langle h, g(y_j)\rangle, \tag{34}$$

and

$$\mathbb{E} \sup_{\|h\|=1} \frac{1}{n} \sum_{j=1}^{n} \varepsilon_i \langle h, g(y_j)\rangle \le \left(\frac{\mathbb{E}\|g(y)\|^2}{n}\right)^{1/2}. \tag{35}$$

PROOF OF LEMMA 4. Equation (33)–(35) imply that

$$\mathbb{E} \sup_{H\in\mathcal{B}} L_n(H) - \mathbb{E}L_n(H) \le K\left(\frac{\mathbb{E}\|g(y)\|^2}{n}\right)^{1/2}. \tag{36}$$

It also holds that

$$\begin{aligned}\mathbb{E} \inf_{H\in\mathcal{B}} L_n(H) - \mathbb{E}L_n(H) &\ge 2\mathbb{E} \inf_{H\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \max_{t\in[K]}\langle h_t, g(y_j)\rangle \\ &= -2\mathbb{E} \sup_{H\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^{n} (-\varepsilon_j) \max_{t\in[K]}\langle h_t, g(y_j)\rangle \\ &= -2\mathbb{E} \sup_{H\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \max_{t\in[K]}\langle h_t, g(y_j)\rangle \\ &\ge -2K\left(\frac{\mathbb{E}\|g(y)\|^2}{n}\right)^{1/2},\end{aligned} \tag{37}$$

where the first inequality holds by a symmetrization analogous to (33); the second by algebraic manipulation; the third because $\varepsilon_1, \dots, \varepsilon_n$ are ±1 with probability $1/2$; the fourth by (34) and (35).

Combining (36) and (37) proves the lemma. □

SUPPLEMENTARY MATERIAL

Supplement to "Co-clustering of nonsmooth graphons" (DOI: 10.1214/16-AOS1497SUPP; .pdf). The supplementary material contains a proof of Lemma 7 and Theorem 2.

REFERENCES

AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2009). Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems* 33–40.

ALIPRANTIS, C. D. and BORDER, K. C. (2006). *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, 3rd ed. Springer, Berlin. MR2378491

BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans*. *Inform*. *Theory* **54** 781–790. MR2444554

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008** P10008.

BORGS, C., CHAYES, J. T., COHN, H. and GANGULY, S. (2015). Consistent nonparametric estimation for heavy-tailed sparse graphs. Preprint. Available at arXiv:1508.06675.

CAI, T. T. and LI, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696

CHEN, A., AMINI, A. A., LEVINA, E. and BICKEL, P. J. (2012). Fitting community models to large sparse networks. *Ann. Statist.* **41** 2097–2122.

CHEN, H.-C., ZOU, W., TIEN, Y.-J. and CHEN, J. J. (2013). Identification of bicluster regions in a binary matrix and its applications. *PLoS ONE* **8** e71680.

CHOI, D. (2017). Supplement to "Co-clustering of nonsmooth graphons." DOI:10.1214/16-AOS1497SUPP.

CHOI, D. and WOLFE, P. J. (2014). Co-clustering separately exchangeable network data. *Ann. Statist.* **42** 29–63. MR3161460

DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84** 066106.

DIACONIS, P. and JANSON, S. (2007). Graph limits and exchangeable random graphs. Preprint. Available at arXiv:0712.2749.

GAO, C., LU, Y. and ZHOU, H. H. (2014). Rate-optimal graphon estimation. Preprint. Available at arXiv:1410.5837.

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2015a). Achieving optimal misclassification proportion in stochastic block model. Preprint. Available at arXiv:1505.03772.

GAO, C., LU, Y., MA, Z. and ZHOU, H. H. (2015b). Optimal estimation and completion of matrices with biclustering structures. Preprint. Available at arXiv:1512.00150.

GOH, K.-I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. and BARABÁSI, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690.

HARPAZ, R., PEREZ, H., CHASE, H. S., RABADAN, R., HRIPCSAK, G. and FRIEDMAN, C. (2011). Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clin. Pharmacol. Ther.* **89** 243–250.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. and BARABÁSI, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* **407** 651–654.

JI, P. and JIN, J. (2014). Coauthorship and citation networks for statisticians. Preprint. Available at arXiv:1410.2840.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107, 10. MR2788206

KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2015). Oracle inequalities for network models and sparse graphon estimation. Preprint. Available at arXiv:1507.04118.

KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVÁ, L. and ZHANG, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110** 20935–20940. MR3174850

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.* **5** 309–336. MR2810399

LOVÁSZ, L. (2012). *Large Networks and Graph Limits* **60**.

MOSSEL, E., NEEMAN, J. and SLY, A. (2013). A proof of the block model threshold conjecture. Preprint. Available at arXiv:1311.4115.

NEWMAN, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64** 016131.

NEWMAN, M. E. (2013). Spectral community detection in sparse networks. Preprint. Available at arXiv:1308.6494.

OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

ROHE, K., QIN, T. and YU, B. (2012). Co-clustering for directed graphs: The stochastic co-Blockmodel and spectral algorithm Di-Sim. Preprint. Available at arXiv:1204.2296.

SCHNEIDER, R. (2013). *Convex Bodies*: *The Brunn–Minkowski Theory*. Cambridge Univ. Press, Cambridge.

SUSSMAN, D. L., TANG, M. and PRIEBE, C. E. (2012). Universally consistent latent position estimation and vertex classification for random dot product graphs. Preprint. Available at arXiv:1207.6745.

SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899

TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53** 526–543. MR2834086

VAN UITERT, M., MEULEMAN, W. and WESSELS, L. (2008). Biclustering sparse binary genomic data. *J. Comput. Biol.* **15** 1329–1345. MR2461979

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

HEINZ COLLEGE OF PUBLIC POLICY AND MANAGEMENT
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVE.
HAMBURG HALL 2101C
PITTSBURGH, PENNSYLVANIA 15026
USA
E-MAIL: davidch@andrew.cmu.edu