The Annals of Statistics
2022, Vol. 50, No. 1, 129–146
https://doi.org/10.1214/21-AOS2094

PATTERN GRAPHS: A GRAPHICAL APPROACH TO NONMONOTONE MISSING DATA

BY YEN-CHI CHEN

Department of Statistics, University of Washington, yenchic@uw.edu

We introduce the concept of pattern graphs–directed acyclic graphs representing how response patterns are associated. A pattern graph represents an identifying restriction that is nonparametrically identified/saturated and is often a missing not at random restriction. We introduce a selection model and a pattern mixture model formulations using the pattern graphs and show that they are equivalent. A pattern graph leads to an inverse probability weighting estimator as well as an imputation-based estimator. We also study the semiparametric efficiency theory and derive a multiply-robust estimator using pattern graphs.

1. Introduction. Missing data problems are prevalent in modern scientific research (Little and Rubin (2002), Molenberghs et al. (2014)). Based on the intrinsic constraints of missing/response patterns, these problems can be categorized into monotone and nonmonotone missing data problems. In the case of monotone missing data, the missingness of variables is ordered in such a way that if a variable is missing, all following variables are missing. This occurs in a scenario in which individuals drop out of a study, which is common in longitudinal studies (Diggle et al. (2002)).

In the case of nonmonotone missing data, the missingness is not necessarily monotone, and the missingness of one variable does not necessarily place constraints on the missingness of any other variables. There have been several attempts to use the missing at random (MAR) restriction/assumption in this case (Robins (1997), Robins and Gill (1997), Sun and Tchetgen Tchetgen (2018)). However, the resulting inverse probability weighting (IPW) estimator may not be stable (Sun and Tchetgen Tchetgen (2018)), and the MAR restriction is not easy to interpret in nonmonotone cases (Robins and Gill (1997), Linero (2017)). Therefore, several attempts have been made to use missing not at random (MNAR) restrictions which are interpretable. For instance, Malinsky, Shpitser and Tchetgen (2019), Sadinle and Reiter (2017), Shpitser (2016) proposed a non-self-censoring/itemwise conditionally independent nonresponse restriction, Little (1993a) and Tchetgen Tchetgen, Wang and Sun (2018) considered a complete-case missing value (CCMV) restriction, and Linero (2017) introduced the transformed-observed-data restriction. However, each study proposed only one MNAR restriction to handle data, and it remains unclear how to construct a general class of identifying restrictions for nonmonotone missing data.

In this paper, we introduce a graphical approach to constructing identifying restrictions for nonmonotone missing data problems. This graphical approach defines an identifying restriction using a graph of response patterns; thus, the resulting graph is called a pattern graph. Formally, a pattern graph is a directed graph where nodes are possible response patterns and whose edges/arrows represent the relationship between the selection probability of patterns (also known as the missing data mechanism in Little and Rubin (2002)). A pattern graph represents an identifying restriction placing conditions on the unobserved part of data, and is

Received December 2020.
MSC2020 subject classifications. Primary 62F30; secondary 62H05, 65D18.
Key words and phrases. Missing data, nonignorable missingness, nomonotone missing, inverse probability weighting, pattern graphs, selection models.

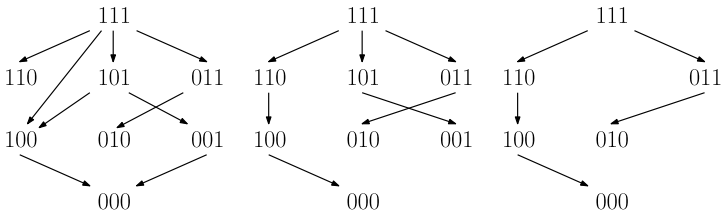

FIG. 1. *Regular pattern graphs in the case of three potentially missing variables. The binary vector indicates the response patterns, for example,* 101 *signifies that the first and the third variables are observed while the second variable is missing. The left and middle panels display examples of regular pattern graphs when all response patterns are possible. The right panel shows a regular pattern graph where there are only six possible response patterns (this occurs when* $P(R = 101) = P(R = 001) = 0$*).*

always nonparametrically identified/saturated (Theorem 3; Robins, Rotnitzky and Scharfstein (2000)); that is, it does not contradict the observed data. In general, the identifying restriction of a pattern graph is an MNAR restriction. Figure 1 provides examples of pattern graphs when three variables may be missing, and a response pattern is described by a binary vector (e.g., 110 signifies that for a variable $L = (L_1, L_2, L_3)$, L_1 and L_2 are observed and L_3 is missing). Different pattern graphs correspond to different identifying restrictions, so pattern graphs define a large class of identifying restrictions. It should be emphasized that *a pattern graph is not a conventional graphical model*.

Main results. The main results of this paper can be summarized as follows:

1. We introduce the concept of pattern graphs (Section 2) and derive a graphical criterion leading to an identifiable full-data distribution using selection odds model and pattern mixture model formulations (Theorems 1 and 3).
2. We demonstrate that the selection odds model and the pattern mixture model are equivalent (Theorem 4).
3. We introduce an IPW estimator and study its statistical properties (Theorem 5).
4. We propose a regression adjustment estimator and derive its asymptotic normality (Theorem 6).
5. We study the semiparametric theory of the pattern graph (Theorem 7) and propose a multiply robust estimator by augmenting the IPW estimator (Theorem 9).

Related work. The CCMV restriction (Little (1993a), Tchetgen Tchetgen, Wang and Sun (2018)) can be represented by a pattern graph. In monotone missing data problems, the available-case missing value restriction (Molenberghs et al. (1998)) and the neighboring-case missing value restriction (Thijs et al. (2002)) and some donor-based identifying restrictions (Chen and Sadinle (2019)) can also be represented by pattern graphs. There have been studies that utilize graphs to analyze missing data. Bhattacharya, Malinsky and Shpitser (2020), Mohan and Pearl (2014), Mohan and Pearl (2021), Mohan, Pearl and Tian (2013), Nabi, Bhattacharya and Shpitser (2020), Tian (2015) proposed methods to test missing data assumptions under graphical model frameworks. Malinsky, Shpitser and Tchetgen (2019), Sadinle and Reiter (2017), Shpitser (2016), Shpitser, Mohan and Pearl (2015) proposed a non-self-censoring graph that leads to an identifying restriction under the MNAR scenario. However, it should again be emphasized that pattern graphs are different from graphical models; thus, our graphical approach is very different from the above-mentioned studies.

Outline. In Section 2, we formally introduce the concept of (regular) pattern graphs and describe how they represent an identifying restriction. We discuss strategies for constructing

TABLE 1

Example of a hypothetical dataset with missing entries. Variable $L = (L_1, \ldots, L_3)$ represents the study variable and variable $R \in \{0, 1\}^3$ represents the response pattern. The star symbol $(*)$ indicates a missing entry

ID	L_1	L_2	L_3	R
001	5	1.3	*	110
002	6	*	1.1	101
003	*	*	1.0	001
004	5	*	*	100
005	2	2.1	0.8	111
⋮	⋮	⋮	⋮	⋮

an estimator under a pattern graph in Section 3. We discuss potential future work in Section 4. In the Supplementary Material (Chen (2022)), we present a sensitivity procedure in Appendix A, a study on the equivalence class in Appendix B, and an application to a real data in Appendix C. Technical assumptions and proofs are provided in Appendix J and K. An R script for finding the IPW estimator based on the pattern graph is in https://github.com/yenchic/PG.

2. Pattern graph and identification. Let $L \in \mathbb{R}^d$ be a vector of the study variables of interest and $R \in \{0, 1\}^d$ be a binary vector representing the response pattern. Variable $R_j = 1$ signifies that variable L_j is observed. Let $1_d = (1, 1, \ldots, 1)$ be the pattern corresponding to the completely observed case and $\bar{r} = 1_d - r$ be the reverse (flipping 0 and 1) of pattern r. We use the notation $L_r = (L_j : r_j = 1)$. For example, suppose that $L = (L_1, \ldots, L_4)$, then $L_{1010} = (L_1, L_3)$, $L_{1100} = (L_1, L_2)$ and $L_{\overline{1100}} = L_{0011} = (L_3, L_4)$. Table 1 presents an example of data with missing entries and the corresponding pattern indicator R. Both L and R are random vectors from a joint distribution $F(\ell, r)$ with a probability density function (PDF) $p(\ell, r)$, and we denote $\mathbb{S}_r$ as the support of random variable L_r. For a binary vector r, we use $|r| = \sum_j r_j$ to denote the number of nonzero elements.

Let $\mathcal{R} \subset \{0, 1\}^d$ be the collection of all possible response patterns, that is, $P(R \in \mathcal{R}) = 1$. A pattern graph is a directed graph $G = (V, E)$, where each vertex represents a response pattern (vertex/node set $V = \mathcal{R}$), and the directed edge represents associations of the distribution of (L, R) across different patterns. Figure 1 provides examples of pattern graphs. Later we will give a precise definition of how a pattern graph factorizes the underlying distribution. The joint distribution of (L, R) is called the full-data distribution and identifying the full-data distribution is a key topic in missing data problems.

When we equip the pattern set $\mathcal{R}$ with a graph G, we can define the notion of parents and children in the graph. For two patterns $r_1, r_2 \in \mathcal{R}$, if there is an arrow $r_1 \to r_2$, we say that r_1 is a parent of r_2 and r_2 is a child of r_1. Let $\mathsf{PA}_r = \{s : s \to r\}$ denote the parents of pattern/node r. A pattern/node is called a source if it has no parent.

For two patterns $s, r \in \mathcal{R}$, we say that $s > r$ if $s_j \geq r_j$ for all j and there is at least one element k such that $s_k > r_k$. For instance, $110 > 100$ and $110 > 010$; however, 110 cannot be compared with 011 or 001. An immediate result from the above ordering is that when $s > r$, the observed variables in pattern r are also observed in pattern s.

A pattern graph G is called a *regular pattern graph* if it satisfies the following conditions:

(G1) Pattern $1_d = (1, 1, \ldots, 1)$ is the only source in G.
(G2) If there is an arrow from pattern s to r (i.e., $s \to r$), then $s > r$.

Figure 1 presents three examples of regular pattern graphs when there are three variables subject to missingness. The first two panels are regular pattern graphs when all eight response

patterns are possible, and the last panel displays a regular pattern graph when only six patterns are possible.

A regular pattern graph has several interesting properties. (G1) implies that the fully observed pattern $R = 1_d$ is the only common ancestor of all patterns except for $R = 1_d$. Moreover, if s is a parent of r, then observed variables in r must be observed in s (due to (G2)). In a sense, this means that a parent pattern is more informative than its child. Condition (G2) implies the following condition:

(DAG) G is a directed acyclic graph (DAG).

Namely, a regular pattern graph is a DAG. In Appendix B, we demonstrate that replacing (G2) with (DAG) still leads to an identifiable full-data distribution.

2.1. *Pattern graph and selection odds models.* A common approach for the missing data problems is the selection model (Little and Rubin (2002)), in which we factorize the full-data density function as

$$p(\ell, r) = P(R = r|\ell)p(\ell),$$

and attempt to identify both quantities. Here, we focus on modeling the selection probability $P(R = r|\ell)$ due to its role in constructing an IPW estimator. To illustrate this, suppose that we are interested in estimating a parameter of interest θ_0 that is defined by a mean function, that is, $\theta_0 = \mathbb{E}(\theta(L))$. Using simple algebra, it can be shown that

$$\theta_0 = \mathbb{E}\big(\theta(L)\big) = \mathbb{E}\bigg(\frac{\theta(L)I(R = 1_d)}{P(R = 1_d|L)}\bigg),$$

which suggests that we can construct an IPW estimator if we know the propensity score $\pi(\ell) = P(R = 1_d|\ell)$.

To associate a pattern graph with the missing data mechanism, we consider the selection odds (Robins, Rotnitzky and Scharfstein (2000)) between a pattern r against its parents PA_r: $\frac{P(R=r|\ell)}{P(R\in\mathsf{PA}_r|\ell)}$. Formally, *the selection odds model of* (L, R) *factorizes* with respect to pattern graph G if

$$\frac{P(R = r|\ell)}{P(R \in \mathsf{PA}_r|\ell)} = \frac{P(R = r|\ell_r)}{P(R \in \mathsf{PA}_r|\ell_r)}. \tag{1}$$

Namely, we assume that the (conditional) odds of a pattern r against its parents depend only on the observed entries. Note that assumption (G2) in the regular pattern graph assumption implies that for any parent nodes of r, variable L_r is observed. Thus, factorization in terms of the selection odds implies that the selection odds are identifiable. From equation (1), it can be seen that the corresponding restriction is an MNAR restriction in general. Equation (1) is related to the MAR restriction in a more involved way (see Section 4 for a detailed discussion).

Let $O_r(\ell_r) = \frac{P(R=r|\ell_r)}{P(R\in\mathsf{PA}_r|\ell_r)}$ be the odds based on the variable ℓ_r. Equation (1) can be written as

$$P(R = r|\ell) = P(R \in \mathsf{PA}_r|\ell) \cdot O_r(\ell_r) = \sum_{s\in\mathsf{PA}_r} P(R = s|\ell) \cdot O_r(\ell_r). \tag{2}$$

Namely, the probability of observing pattern $R = r$ is the summation of the probability of observing any of its parents multiplied by the observable odds. Later in Proposition 2, we provide another interpretation of equation (1) using the path selection. A useful property of graph factorization is that the propensity score is identifiable, as described in the following theorem.

THEOREM 1. *Assume that the selection odds model of* (L, R) *factorizes with respect to a regular pattern graph* G. *Define*

$$Q_r(\ell) = \frac{P(R = r|L = \ell)}{P(R = 1_d|L = \ell)},$$

for each r *and* $Q_{1_d}(\ell) = 1$. *Then* $\pi(\ell) \equiv P(R = 1_d|\ell)$ *is identifiable and has the following recursive-form*:

$$\pi(\ell) = \frac{1}{\sum_r Q_r(\ell)}, \qquad Q_r(\ell) = O_r(\ell_r) \sum_{s \in \mathsf{PA}_r} Q_s(\ell).$$

The identifiability follows from the induction. $Q_{1_d} = 1$ is clearly identifiable, and we recursively deduce the identifiability of Q_r from $|r| = d - 1, d - 2, d - 3, \ldots, 0$. Assumption (G2) guarantees that this recursive procedure is possible. Note that with an identifiable $\pi(\ell)$, we can identify $P(R = r|\ell) = Q_r(\ell)\pi(\ell)$ and $p(\ell) = \frac{p(\ell, R=1_d)}{P(R=1_d|\ell)} = \frac{p(\ell, R=1_d)}{\pi(\ell)}$. Thus, the full-data density $p(\ell, r) = P(R = r|\ell)p(\ell)$ is identifiable.

EXAMPLE 1 (Conditional MAR). Consider the scenario in which we have a longitudinal variable Y with three time points, that is, $Y = (Y_1, Y_2, Y_3)$. In addition, we have another study variable Z that is observed once at the baseline. The total study variable $L = (Z, Y) = (Z, Y_1, Y_2, Y_3)$. Variable Y is subject to monotone missingness (dropout), and variable Z may also be missing. There are a total of six possible patterns in this case, as illustrated in the left panel of Figure 2. We use the variable $T = R_2 + R_3 + R_4$ to denote the dropout time and $R_z = R_1$ to denote the response indicator of variable Z. Suppose that we use the regular pattern graph as in the left panel of Figure 2. This graph implies the following assumptions on T and R_z (see Appendix D in Chen (2022) for the derivation):

$$P(T = t|R_z = 1, L) = P(T = t|R_z = 1, Z, Y_1, \ldots, Y_t), \quad t = 1, 2, 3,$$

$$P(T = t|R_z = 0, L) = P(T = t|R_z = 0, Y_1, \ldots, Y_t), \quad t = 1, 2, 3,$$

$$P(R_z = 0|T = 3, L) = P(R_z = 1|T = 3, L) \cdot \frac{P(R_z = 0|T = 3, Y_1, Y_2, Y_3)}{P(R_z = 1|T = 3, Y_1, Y_2, Y_3)}$$

The first two equations present the conditional MAR restriction, that is, we have MAR of Y given R_z and the observed Z. The third equation describes how the missing data mechanism of Z occurs. The graph provides a simple way to jointly model the dropout time and the missingness of variable Z.

Selection odds factorization provides an alternative interpretation of the missing data mechanism using the concept of path selection. A (directed) path $\Xi = \{r_0, \ldots, r_m\}$, is the

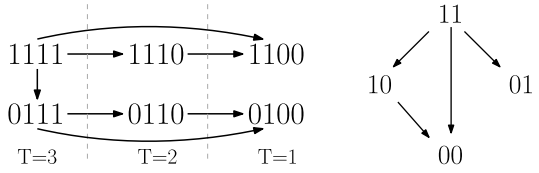

FIG. 2. *Example of regular pattern graphs*. Left: *The regular pattern graph used in Example 1, where we have a longitudinal variable with three time points* $Y = (Y_1, Y_2, Y_3)$ *and a regular variable* Z *where both are subject to missingness. The missingness of* Y *is monotone. Note that this pattern graph leads to conditional missing at random of* Y *given* Z *being observed or not. See Example 1 for further discussion*. Right: *The regular pattern graph used in Example 2.*

collection of ordered patterns

$$r_0 > r_1 > r_2 \cdots > r_m$$

such that there is an arrow from r_i to r_{i+1} in the graph. A path from s to r refers to a path where initial node $r_0 = s$ and the end node $r_m = r$. Let

$$\Pi_r = \{\text{all paths from } 1_d \text{ to } r\}, \Pi = \bigcup_r \Pi_r,$$

and operationally define $\Pi_{1_d} = \{11 \to 11\}$. If there exists a path from s to r, we call s an ancestor (pattern) of r. With the above notation, we have the following decomposition.

PROPOSITION 2. *Assume that the selection odds model of* (L, R) *factorizes with respect to a regular pattern graph* G. *Then*

$$\begin{aligned} 1 &= \sum_{\Xi \in \Pi} \pi(L) \prod_{s \in \Xi} O_s(L_s), \\ P(R = r|L) &= \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s). \end{aligned} \tag{3}$$

Proposition 2 implies

$$\pi(L) = \frac{1}{\sum_{\Xi \in \Pi} \prod_{s \in \Xi} O_s(L_s)}, \tag{4}$$

which is a closed form of the propensity score $\pi(L)$.

Proposition 2 presents an interesting interpretation of the selection odds model. Define $\kappa(\Xi|L) = \pi(L) \prod_{s \in \Xi} O_s(L_s)$ to be a path-specific score. It can be seen that $\kappa(\Xi|L) \geq 0$ and $\sum_{\Xi \in \Pi} \kappa(\Xi|L) = 1$ by the first equality in Proposition 2. Thus, $\kappa(\Xi|L)$ can be interpreted as *the probability of selecting path* Ξ *from* Π. The second equality can be written as

$$P(R = r|L) = \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s) = \sum_{\Xi \in \Pi_r} \kappa(\Xi|L),$$

which implies that the probability of observing pattern r is the summation of all path-specific probabilities corresponding to paths ending at r.

Because every path starts from 1_d, a path can be interpreted as a scenario in which the missingness occurs (from a fully observed case). A path Ξ is randomly selected with a probability of $\kappa(\Xi|L)$, and missingness occurs sequentially as the elements in Ξ. So the last element in Ξ is the observed pattern. Therefore, the probability of observing a particular pattern r is the summation of the probabilities of all possible paths that end at r. The choice of a graph is a means of incorporating our scientific knowledge of the underlying missing data mechanism; in Section C, we provide a data example to illustrate this concept.

EXAMPLE 2. Consider the pattern graph in the right panel of Figure 2, where it is generated by two variables and four patterns 11, 10, 01, 00 and has four arrows $11 \to 10 \to 00$, $11 \to 00$ and $11 \to 10$. There are five paths (including $11 \to 11$):

$$11 \to 11, \qquad 11 \to 10, \qquad 11 \to 01, \qquad 11 \to 00, \qquad 11 \to 10 \to 00$$

and each corresponds to probability

$$\begin{aligned} \kappa(11 \to 11|L) &= \pi(L), \\ \kappa(11 \to 10|L) &= \pi(L) O_{10}(L_{10}), \\ \kappa(11 \to 01|L) &= \pi(L) O_{01}(L_{01}), \\ \kappa(11 \to 00|L) &= \pi(L) O_{00}(L_{00}), \\ \kappa(11 \to 10 \to 00|L) &= \pi(L) O_{10}(L_{10}) O_{00}(L_{00}). \end{aligned}$$

Each path represents a possible scenario that generates the response pattern. Since the probability must sum to 1, we obtain

$$\pi(L) = \frac{1}{1 + O_{10}(L_{10}) + O_{01}(L_{01}) + O_{00}(L_{00}) + O_{10}(L_{10})O_{00}(L_{00})},$$

which agrees with Theorem 1. The probability of observing patterns 10 and 01 are $P(R = 10|L) = \kappa(11 \to 10|L) = \pi(L)O_{10}(L_{10})$ and $P(R = 01|L) = \kappa(11 \to 01|L) = \pi(L)O_{01}(L_{01})$, respectively. Pattern 00 occurs with a probability of

$$\begin{aligned} P(R = 00|L) &= \kappa(11 \to 00|L) + \kappa(11 \to 10 \to 00|L) \\ &= \pi(L)O_{00}(L_{00}) + \pi(L)O_{10}(L_{10})O_{00}(L_{00}). \end{aligned}$$

The first component $\pi(L)O_{00}(L_{00})$ represents scenario $11 \to 00$, that is, the individual directly drops both variables. The other component $\pi(L)O_{10}(L_{10})O_{00}(L_{00})$ corresponds to scenario $11 \to 10 \to 00$, that is, variable L_2 is missing first, and then variable L_1 is missing. Therefore, the paths in the pattern graph represent possible hidden scenarios that generate a response pattern.

REMARK 3. Robins and Gill (1997) proposed a randomized monotone missing (RMM) process to construct a class of MAR assumptions for the nonmonotone missing data problems that also admits a graph representation on how the missingness of one variable is associated with others. This method may look similar to ours; however, the two ideas (RMM and pattern graphs) are very different. First, RMM constructs a MAR assumption, whereas pattern graphs are generally MNAR (generalizations of RMM to MNAR can be found in Robins (1997) and Robins, Rotnitzky and Scharfstein (2000)). Second, each node in the RMM graph is a variable, whereas each node in a pattern graph is a response pattern. Third, in the next section, we demonstrate that the selection odds model in a pattern graph has an equivalent pattern mixture model representation; however, it is unclear whether the RMM process has a desirable pattern mixture model representation or not.

2.2. *Pattern graph and pattern mixture models.* Another common strategy for handling missing data is pattern mixture models (Little (1993b)), which factorize

$$p(\ell, r) = p(\ell|R = r)P(R = r) = p(\ell_{\bar{r}}|\ell_r, R = r)p(\ell_r|R = r)P(R = r).$$

The above factorization provides a clear separation between observed and unobserved quantities. The first part, $p(\ell_{\bar{r}}|\ell_r, R = r)$, is called the extrapolation density (Little (1993b)), which corresponds to the distribution of unobserved entries given the observed entries. This part cannot be inferred from the data without making additional assumptions. The latter part, $p(\ell_r|R = r)P(R = r)$, is called the observed-data distribution, which characterizes the distribution of the observed entries and can be estimated from the data without any identifying assumptions.

An interesting insight is that different response patterns provide information on different variables. Thus, we can associate an extrapolation density to the observed parts of another pattern. This motivates us to consider a graphical approach to factorize the distribution using pattern mixture models.

Formally, *the pattern mixture model of* (L, R) *factorizes* with respect to a pattern graph G if

$$p(x_{\bar{r}}|x_r, R = r) = p(x_{\bar{r}}|x_r, R \in \mathsf{PA}_r). \tag{5}$$

Equation (5) states that the extrapolation density of pattern r can be identified by its parent(s). Namely, we model the unobserved part of pattern r using the information from its parents.

This is a reasonable choice because condition (G2) implies that a parent pattern is more informative than its child pattern. Pattern mixture model factorization leads to the following identifiability property.

THEOREM 3. *Assume that the pattern mixture model of* (L, R) *factorizes with respect to a regular pattern graph* G*, then* $p(\ell, r)$ *is nonparametrically identifiable/saturated.*

Theorem 3 states that graph factorization using pattern mixture models implies a nonparametrically identifiable full-data distribution. Namely, the implied observed distribution of $F(\ell, r)$ coincides with the observed-data distribution that generates our data for patterns r such that $P(R = r) > 0$. Thus, the identifying restriction derived from the graph never contradicts the observed data (Robins, Rotnitzky and Scharfstein (2000)). Nonparametric identification is also known as nonparametric saturation or just-identification in Daniels and Hogan (2008), Hoonhout and Ridder (2019), Robins (1997), Vansteelandt et al. (2006).

Thus far, we have discussed two different methods of associating a pattern graph to a full-data distributions. The following theorem states that they are equivalent under the positivity condition ($p(\ell_r, r) > 0$ for all $\ell_r \in \mathbb{S}_r$ and $r \in \mathcal{R}$).

THEOREM 4. *If* G *is a regular pattern graph and* $p(\ell_r, r) > 0$ *for all* $\ell_r \in \mathbb{S}_r$ *and* $r \in \mathcal{R}$*, then the following two statements are equivalent*:

- *The selection odds model of* (L, R) *factorizes with respect to* G.
- *The pattern mixture model of* (L, R) *factorizes with respect to* G.

With Theorem 4, we can interpret the graph factorization using either the selection odds model or the pattern mixture model, both of which lead to the same full-data distribution. Because of Theorem 4, when we say (L, R) factorizes with respect to G, this factorization may be interpreted using the selection odds model or pattern mixture model. Note that this equivalence is not surprising, as Robins, Rotnitzky and Scharfstein (2000) demonstrated that certain classes of selection odds models and pattern mixture models are equivalent. Theorem 4 shows that the identifying restrictions from pattern graphs form another class of restrictions with this elegant property.

EXAMPLE 4 (Complete-case missing value restriction). The CCMV restriction (Little (1993a)) is an assumption in pattern mixture models. It requires that

$$p(\ell_{\bar{r}} | \ell_r, R = r) = p(\ell_{\bar{r}} | \ell_r, R = 1_d) \tag{6}$$

for all pattern $r \in \mathcal{R}$. The corresponding pattern graph is a graph where every node (except the node of 1_d) has only one parent: the completely-observed case; namely, $\mathsf{PA}_r = 1_d$ for all $r \neq 1_d$. The left panel in Figure 3 presents an example of the pattern graph of CCMV. Using Theorem 4 and the selection odds model, equation (6) is equivalent to

$$\frac{P(R = r | L = \ell)}{P(R = 1_d | L = \ell)} = \frac{P(R = r | L = \ell_r)}{P(R = 1_d | L = \ell_r)}, \tag{7}$$

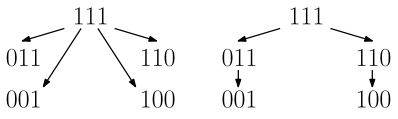

FIG. 3. *Examples of regular pattern graphs of three variables with only* 5 *possible patterns* $\mathcal{R} = \{111, 110, 100, 011, 001\}$. Left: *The left panel shows the pattern graph that CCMV restriction corresponds.* Right: *The right panel shows a pattern graph that is related to the transform-observed-data restriction in Linero (2017).*

which is the key formulation in Tchetgen Tchetgen, Wang and Sun (2018) that establishes a multiply-robust estimator.

REMARK 5 (Transform-observed-data restriction). Linero (2017) proposed a transform-observed-data restriction that is related to a particular pattern graph under a special case. Consider a three-variable scenario in which only five patterns are available 111, 110, 100, 011, 001, and there are two paths of arrows: $111 \to 110 \to 100$ and $111 \to 011 \to 001$. The right panel of Figure 3 displays this graph. The first path implies $p(x_3|x_1, x_2, 110) = p(x_3|x_1, x_2, 111)$ and $p(x_2, x_3|x_1, 100) = p(x_2, x_3|x_1, 110)$, which further implies $p(x_2|x_1, 100) = p(x_2|x_1, 110)$, which is a requirement of the transform-observed-data restriction in this case. Similarly, the other path implies $p(x_2|x_3, 001) = p(x_2|x_3, 011)$, which is another requirement of the transform-observed-data restriction.

REMARK 6 (Monotone missing data problem). Suppose that the missingness is monotone; then, the pattern graph reduces to special cases of the interior family (Thijs et al. (2002)) and donor-based identifying restriction (Chen and Sadinle (2019)). In particular, the parent set PA_r is the donor set of the dropout time $t = |r|$. The available-case missing value restriction (Molenberghs et al. (1998)) corresponds to the pattern graph with $\mathsf{PA}_r = \{s : |s| > |r|\}$, that is, the graph with all possible arrows/edges. The neighboring-case missing value restriction (Thijs et al. (2002)) is the pattern graph with $\mathsf{PA}_r = \{s : |s| = |r| + 1\}$.

3. Estimation with pattern graphs. In this section, we present several strategies for estimating the parameter of interest using the pattern graph. Here, we consider the parameter of interest that can be written in the form $\theta_0 = \mathbb{E}(\theta(L))$, where $\theta(L)$ is a known function. Note that all analyses can be applied to the case of estimating equations.

With a slight abuse of notation, the observed data are written as i.i.d. random elements

$$(L_{1,R_1}, R_1), \ldots, (L_{n,R_n}, R_n),$$

where $R_1, \ldots, R_n \in \mathcal{R}$ denote the response pattern of each observation and L_{i,R_i} denotes the observed variables of the ith individual and $L_i \in \mathbb{R}^d$ denotes the vector of study variables of the ith individual. Note that not every entry of L_i is observed; we only observe L_{i,R_i}, while $L_{i,\bar{R}_i}$ is missing.

3.1. *Inverse probability weighting.* The parameter of interest can be written as

$$\theta_0 = \mathbb{E}\big(\theta(L)\big) = \mathbb{E}\bigg(\frac{\theta(L)I(R = 1_d)}{P(R = 1_d|L)}\bigg) = \mathbb{E}\bigg(\frac{\theta(L)I(R = 1_d)}{\pi(L)}\bigg).$$

This formulation implies that as long as we can estimate $\pi(\ell)$, we can construct a consistent estimator of θ via the concept of IPW.

From Theorem 1, the propensity score can be expressed as

$$\pi(\ell) = \frac{1}{\sum_r Q_r(\ell)}, \qquad Q_{1_d}(\ell) = 1, \qquad Q_r(\ell) = O_r(\ell_r) \sum_{s \in \mathsf{PA}_r} Q_s(\ell).$$

By the above recursive property, an estimator of $O_r(\ell_r)$ leads to an estimator of $Q_r(\ell)$ and $\pi(\ell)$. The odds

$$O_r(\ell_r) = \frac{P(R = r|\ell_r)}{P(R \in \mathsf{PA}_r|\ell_r)}$$

can be estimated by comparing the distribution of patterns $R = r$ with patterns $R \in \mathsf{PA}_r$. This can be achieved by constructing a generative binary classifier (Friedman, Hastie and Tibshirani (2001)) such that label 1 refers to $R = r$ and label 0 refers to $R \in \mathsf{PA}_r$ or by a regression

function with the same binary outcome and the feature/covariate is ℓ_r. In Example 10 of Appendix J, we describe a logistic regression approach to estimate $O_r(\ell_r)$.

Suppose that we have an estimator $\hat{\pi}(\ell)$ of the propensity score. Then, we can estimate θ using the IPW approach as follows:

$$\hat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\theta(L_i) I(R_i = 1_d)}{\hat{\pi}(L_i)}.$$

As an example, suppose that we estimate $\pi(\ell)$ by placing parametric models over the odds, that is,

$$\hat{O}_r(\ell_r) = O_r(\ell_r; \hat{\eta}_r),$$

where $\hat{\eta}_r \in \Theta_r$ is the estimated parameter of the selection odds $\frac{P(R=r|\ell_r)}{P(R\in \mathsf{PA}_r|\ell_r)}$. We can estimate the selection odds using a maximum likelihood approach or moment-based approach. With the estimated selection odds, we estimate the propensity score $\hat{\pi}(\ell) = \pi(\ell; \hat{\eta})$ using the recursive relation. Let $\hat{\eta} = (\hat{\eta}_r : r \in \mathcal{R})$ be the set of the estimated parameters.

THEOREM 5. *Assume* (L1-4) *in Appendix J and that the selection odds model of* (L, R) *factorizes with respect to a regular pattern graph* G. *Then* $\hat{\theta}_{\mathsf{IPW}}$ *is a consistent estimator and satisfies*

$$\sqrt{n}(\hat{\theta}_{\mathsf{IPW}} - \theta_0) \xrightarrow{D} N\big(0, \sigma^2_{\mathrm{IPW}}\big),$$

for some $\sigma^2_{\mathrm{IPW}} > 0$.

Theorem 5 shows the asymptotic normality of the IPW estimator and can be used to construct a confidence interval. A traditional approach is to obtain a sandwich estimator of σ^2_{IPW} and use it with the normal score to construct a confidence interval. However, the actual form of σ^2_{IPW} is complex because patterns are correlated based on the graph structure and there is no simple way to disentangle them. Thus, we recommend using the bootstrap approach (Efron (1979), Efron and Tibshirani (1994)) to construct a confidence interval. This can be achieved without knowing the form of σ^2_{IPW}. Note that the bootstrap method often requires a third moment condition of the score (Hall (2013)); for smooth parametric models such as logistic regression with a bounded covariates, this condition holds.

We can rewrite the IPW estimator as

$$\hat{\theta}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \theta(L_i) I(R_i = 1_d) \sum_r Q_r(L_i; \hat{\eta}).$$

So the quantity $Q_r(L_i; \hat{\eta})$ behaves like a score from pattern r on observation L_i.

3.1.1. *Recursive computation.* Although the IPW estimator has desirable properties, the propensity score does not have a simple closed form; therefore, the computation of equation (3) is not easy. To resolve this problem, we provide a computationally friendly approach to evaluate $\pi(\ell)$ (or its estimator $\hat{\pi}(\ell)$) using the recursive relation in Theorem 1.

From Theorem 1, $\pi(\ell) = \frac{1}{\sum_r Q_r(\ell)}$; thus, it is only necessary to compute $Q_r(\ell)$. The recursive form in Theorem 1,

$$Q_{1_d}(\ell) = 1, \qquad Q_r(\ell) = O_r(\ell_r) \sum_{s\in \mathsf{PA}_r} Q_s(\ell),$$

demonstrates that we can compute $Q_r(L)$ recursively.

Algorithm 1 summarizes the procedure for computing $\hat{\pi}(L)$. We first compute cases where $|r| = d - 1$. Having computed $\{Q_r(L) : |r| = d - 1\}$, we can easily compute $\{Q_r(L) : |r| =$

Algorithm 1 Recursive computation of the propensity score

1. Input: $\hat{Q}_{1_d}(\ell) = 1$ and a given fully-observed vector L and estimators $\hat{O}_r(\ell_r)$ for each $r \in \mathcal{R}$.
2. Starting from $j = 1, \ldots, d-1$, do the following:
2-1. For each $r \in \{s \in \mathcal{R} : |s| = d - j\}$, do the following:
2-1-1. Compute $\hat{O}_r(L_r)$. In the case of logistic regression, $\hat{O}_r(L_r) = \exp(\hat{\beta}_r^T \tilde{L}_r)$.
2-1-2. Compute $\hat{Q}_r(L) = \hat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \hat{Q}_s(L)$.
3. Return: $\hat{\pi}(L) = \frac{1}{\sum_r \hat{Q}_r(L)}$.

$d-2\}$ because $\{Q_r(L) : |r| = d-2\}$ only depend on $\{Q_r(L) : |r| = d, d-1\}$ and each $O_r(L)$. Thus, by sequentially computing (noting that $Q_{1_d}(L) = 1$)

$$\big\{Q_r(L) : |r| = d-1\big\}, \qquad \big\{Q_r(L) : |r| = d-2\big\}, \qquad \ldots, \qquad \big\{Q_r(L) : |r| = 1\big\},$$

we obtain every $Q_r(L)$, which then leads to $\pi(L) = \frac{1}{\sum_r Q_r(L)}$.

Suppose that evaluating $O_r(L_r)$ takes $\Omega(1)$ units of operations; then, total cost of evaluating $\pi(L)$ using Algorithm 1 is $\Omega(\sum_r |\mathsf{PA}_r|)$ units, where $|\mathsf{PA}_r|$ is the number of parents of node r. However, if we use equation (3), the total cost is $\Omega(\sum_r \sum_{\Xi \in \Pi_r} |\Xi|)$, where $|\Xi|$ is the number of vertices in the path. It can be seen that $|\mathsf{PA}_r| \le \sum_{\Xi \in \Pi_r} |\Xi|$ and the number of parents can be much smaller than the total number of paths. Therefore, Algorithm 1 is much more efficient than directly using equation (3).

3.2. *Regression adjustments*. We can rewrite the parameter of interest as

$$\theta_0 = \mathbb{E}\big(\theta(L)\big) = \int m(\ell_r, r) P(d\ell_r, dr), \qquad m(\ell_r, r) = \mathbb{E}\big(\theta(L) | L_r = \ell_r, R = r\big).$$

Thus, if we have an estimator $\hat{m}(\ell_r, r)$ for every r, we can estimate $\mathbb{E}(\theta(L))$ using the regression adjustment approach

$$\hat{\theta}_{\mathsf{RA}} = \frac{1}{n} \sum_{i=1}^{n} \hat{m}(L_{i,R_i}, R_i).$$

In Appendix F.2, we demonstrate that a Monte Carlo approximation of this estimator is the imputation-based estimator (Little and Rubin (2002), Rubin (2004), Tsiatis (2007)).

Regression adjustment is feasible because the regression function $m(\ell_r, r) = \mathbb{E}(\theta(L)|L_r = \ell_r, R = r)$ is identifiable. To see this, using the PMM factorization in equation (5),

$$\begin{aligned} m(\ell_r, r) &= \mathbb{E}\big(\theta(L) | L_r = \ell_r, R = r\big) \\ &= \int \theta(\ell_{\bar{r}}, \ell_r) p(\ell_{\bar{r}} | \ell_r, R = r) \, d\ell_{\bar{r}} \\ &= \int \theta(\ell_{\bar{r}}, \ell_r) p(\ell_{\bar{r}} | \ell_r, R \in \mathsf{PA}_r) \, d\ell_{\bar{r}} \\ &= \mathbb{E}\big(\theta(L) | L_r = \ell_r, R \in \mathsf{PA}_r\big), \end{aligned}$$

and $p(\ell_{\bar{r}} | \ell_r, R \in \mathsf{PA}_r)$ is identifiable due to Theorem 3.

In practice, we first estimate $\hat{p}(\ell_r | R = r)$ using a parametric model for every r. With this, we then estimate $p(\ell_{\bar{r}} | \ell_r, R \in \mathsf{PA}_r)$. Note that we can use a nonparametric density estimator as well, but it often suffers from the curse of dimensionality.

For pattern r, let $\lambda_r \in \Lambda_r$ be the parameter of the model $L_r | R = r$. Namely,

$$p(\ell_r | R = r) = p(\ell_r | R = r; \lambda_r).$$

We can estimate λ_r via the maximum likelihood estimator (MLE). Let $\hat{\lambda}_r$ be the MLE. We model it in this way to avoid model conflicts; see Appendix F.1 in the Supplementary Material (Chen (2022)) for more details. Let $\lambda = (\lambda_r : r \in \mathcal{R})$ be the collection of all parameters in the model, let Λ be the corresponding parameter space, and let $\hat{\lambda}$ be the MLE. The regression function is then estimated by

$$\begin{aligned}\hat{m}(\ell_r, r) &= m(\ell_r, r; \hat{\lambda}) \\ &= \int \theta(\ell_{\bar{r}}, \ell_r) p(\ell_{\bar{r}} | \ell_r, R \in \mathsf{PA}_r; \hat{\lambda}) \, d\ell_{\bar{r}}.\end{aligned}$$

Note that in the above expression, the expression of the estimator depends on the entire set of parameters $\hat{\lambda} = (\hat{\lambda}_r : r \in \mathcal{R})$, but $\hat{m}(\ell_r, r)$ actually only depends on the parameter belonging to its ancestor. We express it using $\hat{\lambda}$ to simplify the notation.

THEOREM 6. *Assume* (*R*1-3) *in Appendix J and that the pattern mixture model of* (L, R) *factorizes with respect to a regular pattern graph* G. *Then* $\hat{\theta}_{\mathsf{RA}}$ *is a consistent estimator and satisfies*

$$\sqrt{n}(\hat{\theta}_{\mathsf{RA}} - \theta_0) \xrightarrow{D} N(0, \sigma^2_{\mathrm{RA}})$$

for some $\sigma^2_{\mathrm{RA}} > 0$.

Theorem 6 shows that if the density estimators are consistent, the resulting regression adjustment estimator is asymptotically normal. Similar to the IPW estimator, this provides a way to construct a confidence interval using the bootstrap. In Appendix F.2, we describe a Monte Carlo approach to compute $\hat{\theta}_{\mathsf{RA}}$. In addition, we show that when the pattern graph is a tree graph, there may be a closed form of the regression adjustment estimator; thus,o we do not need a numerical procedure (Appendix I).

3.3. *Semiparametric estimators*. We now study the semiparametric theory of the pattern graph and propose an efficient estimator. We start with a derivation of the efficient influence function (EIF) of $\mathbb{E}(\theta(L))$. For any pattern $r \in G$, recall that Π_r denotes all paths from 1_d to r and $\Pi = \bigcup_r \Pi_r$ is the collection of all paths.

By Theorem 1 and equation (4), the inverse of the propensity score can be written as

$$\frac{1}{\pi(L)} = \sum_r Q_r(L) = 1 + \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \prod_{s \in \Xi} O_s(L_s).$$

Thus, the IPW formulation can be decomposed as

$$\begin{aligned}\theta &= \mathbb{E}\big(\theta(L)\big) \\ &= \mathbb{E}\bigg(\frac{\theta(L) I(R = 1_d)}{\pi(L)}\bigg) \\ &= \mathbb{E}\big(\theta(L) I(R = 1_d)\big) + \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \mathbb{E}\bigg(\theta(L) I(R = 1_d) \prod_{s \in \Xi} O_s(L_s)\bigg) \\ &= \theta_{1_d} + \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \theta_\Xi.\end{aligned} \tag{8}$$

For a path $\Xi \in \Pi$ and an element $s \in \Xi$, we define

$$\begin{aligned}&\mathsf{EIF}_{\Xi,s}(L_s, R) \\ &\quad = \mu_{\Xi,s}(L_s)\big(I(R = s) - O_s(L_s) I(R \in \mathsf{PA}_s)\big) \prod_{w \in \Xi, w < s} O_w(L_w),\end{aligned} \tag{9}$$

where

$$\mu_{\Xi,s}(L_s) = \frac{m_{\Xi,s}(L_s)}{P(R \in \mathsf{PA}_s | L_s)}, \tag{10}$$

$$m_{\Xi,s}(L_s) = \mathbb{E}\Big(\theta(L) I(R = 1_d) \prod_{\tau \in \Xi, \tau > s} O_\tau(L_\tau) | L_s\Big). \tag{11}$$

The following proposition demonstrates that $\sum_{s \in \Xi} \mathsf{EIF}_{\Xi,s}(L_s, R)$ is the EIF of θ_Ξ; therefore, we obtain a closed form of the EIF of θ.

THEOREM 7 (Efficient influence function). *Suppose that the selection odds model of (L, R) factorizes with respect to a regular pattern graph G and $p(\ell_r, r) > 0$. The EIF of θ_Ξ is $\mathsf{EIF}_\Xi(L, R) - \theta_\Xi$, where*

$$\mathsf{EIF}_\Xi(L, R) = \sum_{s \in \Xi} \mathsf{EIF}_{\Xi,s}(L_s, R).$$

Thus, the EIF of θ is $\mathsf{EIF}(L, R) - \theta$ with

$$\mathsf{EIF}(L, R) = \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \mathsf{EIF}_\Xi(L, R).$$

Theorem 7 provides an analytical form of the EIF of both θ and a pathwise version of it. Theorem 7 also illustrates how a pattern graph informs the construction of the EIF. In Appendix H, we derive the expression of the EIF of Example 2. A key element in the EIF is the function $\mu_{\Xi,s}(L_s)$ defined in equation (10). In what follows, we describe how $\mu_{\Xi,s}(L_s)$ is associated with the regression adjustment estimator in Section 3.2.

PROPOSITION 8 (Relation to regression adjustment). *Let Ans_r denote the ancestors of r including r itself. For $s \in \mathsf{Ans}_r$, let $\Upsilon_{s,r}$ be the collection of all paths from s to r. Then*:

1. *Function $\mu_{\Xi,s}(L_s)$ is identifiable from $\{p(\ell_r | R = r) : r \in \mathsf{Ans}_s\}$;*
2. *$\sum_{\Xi \in \Pi_s} \mu_{\Xi,s}(\ell_s) = m(\ell_s, s)$, where $m(\ell_s, s)$ is the regression function defined in Section 3.2;*
3. *The EIF of pattern r, $\mathsf{EIF}_r = \sum_{\Xi \in \Pi_r} \mathsf{EIF}$, can be written as*

$$\mathsf{EIF}_r(L, R) = \sum_{s \in \mathsf{Ans}_r} \underbrace{m(L_s, s)\big(I(R = s) - O_s(L_s) I(R \in \mathsf{PA}_s)\big) \sum_{\zeta \in \Upsilon_{s,r}} \prod_{w \in \zeta, w < s} O_w(L_w)}_{= \mathsf{EIF}_{s,r}(L,R)}.$$

Suppose that we have a collection of models $\{p(\ell_\tau | R = \tau; \lambda_\tau) : \tau \in \mathcal{R}\}$, where λ_τ is the underlying parameters. By Proposition 8, we can identify $\mu_{\Xi,r}(L_r)$ using these models, leading to $\mu_{\Xi,r}(L_r; \lambda)$ without any knowledge of the selection odds. This insight leads to the construction of a semiparametric estimator in the next section.

In addition, Theorem 7 and Proposition 8 provide two equivalent expressions of the EIF. The first one is a *path expression*:

$$\mathsf{EIF}(L, R) = \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \sum_{s \in \Xi} \mathsf{EIF}_{\Xi,s}(L, R),$$

while the second is an *ancestor expression*:

$$\mathsf{EIF}(L, R) = \sum_{r \neq 1_d} \sum_{s \in \mathsf{Ans}_r} \mathsf{EIF}_{s,r}(L, R),$$

Algorithm 2 Monte Carlo approximation of the semiparametric estimator

Input models: $\{p(\ell_r|R=r;\hat{\lambda}_r), O_r(L_r;\hat{\eta}_r): r\in\mathcal{R}\}$.
1. Apply the multiple imputation method (Algorithm 4 in the appendix) to obtain an approximation $\tilde{m}(\ell_r, r;\hat{\lambda})$ for each r.
2. For each r and an ancestor $s\in\mathsf{Ans}_r$, compute

$$\begin{aligned}&\widetilde{\mathsf{EIF}}_{s,r}(L,R)\\&\quad=\tilde{m}(L_s,s;\hat{\lambda})(I(R=s)-O_s(L_s;\hat{\eta}_s)I(R\in\mathsf{PA}_s))\sum_{\zeta\in\Upsilon_{s,r}}\prod_{w\in\zeta,w<s}O_w(L_w;\hat{\eta}_w)\end{aligned}$$

3. Compute the EIF as $\widetilde{\mathsf{EIF}}(L,R)=\sum_{r\neq 1_d}\sum_{s\in\mathsf{Ans}_r}\widetilde{\mathsf{EIF}}_{s,r}(L,R)$.
4. Compute the propensity score $\pi(L;\hat{\eta})$ by Algorithm 1.
5. Return: $\tilde{\theta}_{\mathsf{semi}}$ as

$$\tilde{\theta}_{\mathsf{semi}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\theta(L_i)I(R_i=1_d)}{\pi(L_i;\hat{\eta})}+\widetilde{\mathsf{EIF}}(L_i,R_i).$$

where $\mathsf{EIF}_{s,r}(L,R)$ is defined in Proposition 8. The path expression provides insight into how each path's information contributes to the efficiency of a node, whereas the ancestor expression demonstrates how an ancestor improves the efficiency of its descendent. Moreover, the path expression provides a clear picture of the multiple robustness property (Section 3.3.2) while the ancestor expression leads to a simpler numerical procedure (Algorithm 2), which is a mild modification of the regression adjustment.

3.3.1. *Construction of semiparametric estimators.* With the EIF, we can derive a semiparametric estimator. Since our derivation of EIF is based on the IPW approach, the linear form of the semiparametric estimator is the IPW added to the augmentation from the EIF, that is,

$$\begin{aligned}\mathcal{L}_{\mathsf{semi}}(L,R)&=\frac{\theta(L)I(R=1_d)}{\pi(L)}+\mathsf{EIF}(L,R)\\&=\frac{\theta(L)I(R=1_d)}{\pi(L)}+\sum_{r\neq 1_d}\sum_{\Xi\in\Pi_r}\sum_{s\in\Xi}\mathsf{EIF}_{\Xi,s}(L,R)\\&=\frac{\theta(L)I(R=1_d)}{\pi(L)}+\sum_{r\neq 1_d}\sum_{s\in\mathsf{Ans}_r}\mathsf{EIF}_{s,r}(L,R).\end{aligned}$$

It can be seen that $\mathbb{E}[\mathcal{L}_{\mathsf{semi}}(L,R)]=\theta$. We use the path expression in the following derivation, as it leads to an elegant multiple robustness property (see next section).

Let $O_r(L_r;\hat{\eta}_r)$ be the estimated selection odds and let $p(\ell_r|R=r;\hat{\lambda}_r)$ be the estimated density used in the regression adjustment method. By Proposition 8, the collection $\{p(\ell_r|R=r;\hat{\lambda}_r): r\in\mathcal{R}\}$ implies the collection $\{\mu_{\Xi,s}(\ell_s,r;\hat{\lambda}): s\in\Xi, \Xi\in\Pi_r, r\neq 1_d\}$, where $\hat{\lambda}=(\hat{\lambda}_r: r\in\mathcal{R})$. In addition, let $O_r(L_r;\hat{\eta}_s)$ be the estimated selection odds of pattern r.

With these estimators, we estimate the EIF by

$$\begin{aligned}&\mathsf{EIF}_{\Xi,s}(L_s,R;\hat{\lambda},\hat{\eta})\\&\quad=\mu_{\Xi,s}(L_s;\hat{\lambda})\big[I(R=s)-O_s(L_s;\hat{\eta}_s)I(R\in\mathsf{PA}_s)\big]\prod_{w\in\Xi,w<s}O_w(L_w;\hat{\eta}_w)\end{aligned}$$

and construct the semiparametric estimator

$$\hat{\theta}_{\mathsf{semi}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\mathsf{semi}}(L_i, R_i; \hat{\lambda}, \hat{\eta})$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\theta(L_i) I(R_i = 1_d)}{\pi(L_i; \hat{\eta})}}_{\text{IPW}} + \underbrace{\sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \overbrace{\sum_{s \in \Xi} \mathsf{EIF}_{\Xi,s}(L_i, R_i; \hat{\lambda}, \hat{\eta})}^{=\mathsf{EIF}_{\Xi}(L_i, R_i; \hat{\lambda}, \hat{\eta})}}_{\text{augmentation}}. \tag{12}$$

The semiparametric estimator contains an IPW component and an augmentation component, so it is an augmented IPW estimator (see Appendix G for more details). Semiparametric theory ensures that this estimator is the most efficient estimator when both the selection odds $\{O_r(L_r; \eta_r) : r \in \mathcal{R}\}$ and the regression functions $\{\mu_{\Xi,s}(L_r; \lambda_r) : r \in \mathcal{R}\}$ are correctly specified. Algorithm 2 provides a Monte Carlo procedure to compute the semiparametric estimator, which is a combination of the recursive algorithm in Algorithm 1 and the multiple imputation in Algorithm 4 in the Supplementary Material. The key is to use the ancestor expression, which leads to a simpler form of the semiparametric estimator. Note that similar to the regression adjustment estimator, if the pattern graph is a tree graph, we can avoid using Algorithm 2 to compute the estimator; see Appendix I.

REMARK 7. In the pattern graph of the CCMV restriction, arrows are in the form $1_d \to r$ for each $r \neq 1_d$. In this case, $\Pi_r = \{r\}$ and $\Xi = r$, so

$$\mu_{\Xi,r}(\ell_r) = \frac{\mathbb{E}(\theta(L) I(R = 1_d) | L_r = \ell_r)}{P(R = 1_d | \ell_r)} = \mathbb{E}\big(\theta(L) | R = 1_d, L_r = \ell_r\big).$$

Thus, the semiparametric estimator in equation (12) is the same as the semiparametric estimator in Tchetgen Tchetgen, Wang and Sun (2018).

3.3.2. *Multiple robustness*. In many scenarios, a semiparametric estimator often exhibits a double robustness or multiple robustness property (Robins, Rotnitzky and Scharfstein (2000), Tsiatis (2007), Seaman and Vansteelandt (2018)). We demonstrate that our semiparametric estimator in equation (12) also enjoys a multiple robustness property. Here, we assume that the parameters $\hat{\lambda} \xrightarrow{P} \lambda^*$ and $\hat{\eta} \xrightarrow{P} \eta^*$. Note that equation (12) can be factorized as

$$\mathcal{L}_{\mathsf{semi}}\big(L, R; \lambda^*, \eta^*\big) = \theta(L) I(R = 1_d) + \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \mathcal{L}_{\mathsf{semi},\Xi}\big(L, R; \lambda^*, \eta^*\big),$$

$$\mathcal{L}_{\mathsf{semi},\Xi}\big(L, R; \lambda^*, \eta^*\big) = \theta(L) I(R = 1_d) \prod_{s \in \Xi} O_s\big(L_s; \eta^*\big) + \mathsf{EIF}_{\Xi}\big(L, R; \lambda^*, \eta^*\big).$$

We demonstrate the multiple robustness properties of each component $\mathcal{L}_{\mathsf{semi},\Xi}(L, R; \lambda^*, \eta^*)$. Note that we let $O_s(L_s)$ and $\mu_{\Xi,s}(L_s)$ denote the correct selection odds and regression function for each $s \in \Xi$ and each path Ξ, respectively.

THEOREM 9 (Multiple robustness). *Suppose that the selection odds model of* (L, R) *factorizes with respect to a regular pattern graph* G *and* $p(\ell_r, r) > 0$. *Let* $r \in \mathcal{R}$ *be a response pattern. For a path* $\Xi \in \Pi_r$, *if either* $O_s(L_s; \eta^*) = O_s(L_s)$ *or* $\mu_{\Xi,s}(L_s; \lambda^*) = \mu_{\Xi,s}(L_s)$ *for each* $s \in \Xi$, *then*

$$\mathbb{E}\big(\mathcal{L}_{\mathsf{semi},\Xi}\big(L, R; \lambda^*, \eta^*\big)\big) = \theta_{\Xi}.$$

Using the fact that $\theta = \theta_{1_d} + \sum_{r \neq 1_d} \sum_{\Xi \in \Pi_r} \theta_\Xi$, it is evident that if we can consistently estimate θ_Ξ for each Ξ, we can estimate θ consistently.

Let $\mathcal{M}_s^O = \{O_s(\cdot; \eta^*) = O_s(\cdot)\}$ be the case where the selection odds of pattern s is correctly specified. For $\Xi \in \Pi_r$, $r \neq 1_d$ and $s \in \Xi$, let $\mathcal{M}_{\Xi,s}^\mu = \{\mu_{\Xi,s}(\cdot; \lambda^*) = \mu_{\Xi,s}(\cdot)\}$ be the case where $\mu_{\Xi,s}$ is correctly specified. Theorem 9 shows that under the intersection of models

$$\mathcal{M}_\Xi = \bigcap_{s \in \Xi} (\mathcal{M}_s^O \cup \mathcal{M}_{\Xi,s}^\mu),$$

the quantity $\mathcal{L}_{\mathsf{semi},\Xi}(L, R; \lambda^*, \eta^*)$ leads to a consistent estimator of θ_Ξ, that is,

$$\hat{\theta}_\Xi = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\mathsf{semi},\Xi}(L_i, R_i; \hat{\lambda}, \hat{\eta}) \xrightarrow{P} \theta_\Xi.$$

Thus, to estimate $\theta = \sum_r \theta_r$, we must select a model in

$$\mathcal{M} = \bigcap_{r \neq 1_d} \bigcap_{\Xi \in \Pi_r} \bigcap_{s \in \Xi} (\mathcal{M}_s^O \cup \mathcal{M}_{\Xi,s}^\mu). \tag{13}$$

If our model falls within $\mathcal{M}$, we have $\hat{\theta}_{\mathsf{semi}} \xrightarrow{P} \theta$. This describes the multiple robustness property of the semiparametric estimator in equation (12).

Similar to a conventional multiply robust estimator (Tchetgen Tchetgen, Wang and Sun (2018)), $\hat{\theta}_{\mathsf{semi}}$ is a $\sqrt{n}$-rate efficient normal estimator of θ if for any path Ξ,

$$\sum_{s \in \Xi} \|\mu_{\Xi,s}(\cdot; \hat{\lambda}) - \mu_{\Xi,s}(\cdot)\|_{L_2(P)} \|O_s(\cdot; \hat{\eta}_s) - O_s(\cdot)\|_{L_2(P)} = o_P\left(\frac{1}{\sqrt{n}}\right),$$

where $\|f\|_{L_2(P)} = (\int |f(\ell)|^2 \, dP(\ell))^{1/2}$ is the $L_2(P)$ norm of a function f. This occurs when all (L1-L4) and (R1-R3) conditions in Appendix J hold.

4. Discussion. In this paper, we, introduce the concept of pattern graphs and use it to represent an identifying restriction for missing data problems. Pattern graphs provide a new way to construct identifying restrictions. We demonstrate that pattern graphs can be interpreted using a selection odds model or pattern mixture model. In addition, we propose various estimators using different modeling strategies and study statistical and computational properties with a pattern graph. The theories developed in Section 3.3 demonstrate the elegant association between the semiparametric theory and pattern graphs. We believe that the pattern graph approach can provide a new direction in missing data research. Below, we discuss possible future directions that are worth pursuing.

- *Choice of pattern graph.* In this paper, we mainly focus on the theoretical analysis of pattern graphs and assume that a pattern graph is given. In practice, determining how to select a pattern graph is an open problem. Since a pattern graph leads to an identifying restriction, it should be chosen based on background knowledge of how missingness occurs. In Appendix C, we provide a data analysis example and attempt to choose a pattern graph based on prior knowledge of the data generating process. In this particular example, we use the path selection interpretation of pattern graphs (Proposition 2 and related discussion) to select a plausible pattern graph. Although this approach is reasonable for this particular data, it may not apply to other problems. We plan to develop a general principle for selecting a pattern graph in future work.
- *Inference with multiple restrictions.* Although a pattern graph may be derived from scientific knowledge, sometimes there may be uncertainties regarding the graph to be used. As a result, there may be a set of possible graphs $\{G_1, \ldots, G_k\}$ that are reasonable. In this scenario, determining how to perform statistical inference is an open question. One possible

solution is to derive a nonparametric bound (Manski (1990), Horowitz and Manski (2000)) or an uncertainty interval (Vansteelandt et al. (2006)) in which we compute an estimator of each graph and use the range of these estimators as an interval estimate. Alternatively, one can consider a Bayesian approach that assigns a prior distribution over possible graphs and derives the posterior distribution of the parameter of interest. The posterior mean behaves like a Bayesian model averaging estimator (Hoeting et al. (1999)), and the posterior distribution includes uncertainties from both estimation and graphs.

- *MAR and conditional independence.* The MAR restriction can be written as a pattern graph with $\mathsf{PA}_r = \mathcal{R}\backslash\{r\}$. It is not a regular pattern graph; however, it still leads to a uniquely identified full-data distribution (Gill, van der Laan and Robins (1997)). This implies that pattern graphs that are not DAGs may still lead to an identifying restriction. Pattern graph factorization implies the following conditional independence:

$$I(R=r) \perp L_{\bar{r}} | L_r, \qquad R \in E_r, \qquad E_r = \{r\} \cup \mathsf{PA}_r \tag{14}$$

for each r. When $E_r = \mathcal{R}$, this is equivalent to the MAR restriction. The choice of E_r is equivalent to the choice of the parents, which may provide a way to study identifying restrictions beyond acyclic pattern graphs. Thus, studying the conditions on E_r that lead to an identifiable full-data distribution is a future direction that is worth pursuing.

Acknowledgement. We thank Adrian Dobra, Mathias Drton, Mauricio Sadinle, Daniel Suen, Thomas Richardson for very helpful comments on the paper.

Funding. This work is partially supported by NSF Grants DMS-1810960, DMS-1952781 and DMS-2112907, and NIH Grant U01 AG016976.

SUPPLEMENTARY MATERIAL

Pattern graphs: A graphical approach to nonmonotone missing data (DOI: 10.1214/21-AOS2094SUPP; .pdf). This document contains all proofs to the theorems and lemmas in this paper and additional topics related to this paper including sensitivity analysis, equivalence graph, and data analysis.

REFERENCES

BHATTACHARYA, R., MALINSKY, D. and SHPITSER, I. (2020). Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence* 1028–1038. PMLR.

CHEN, Y.-C. (2022). Supplement to "Pattern graphs: A graphical approach to nonmonotone missing data." https://doi.org/10.1214/21-AOS2094SUPP

CHEN, Y.-C. and SADINLE, M. (2019). Nonparametric pattern-mixture models for inference with missing data. arXiv preprint. Available at arXiv:1904.11085.

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies*: *Strategies for Bayesian Modeling and Sensitivity Analysis*. *Monographs on Statistics and Applied Probability* **109**. CRC Press/CRC, Boca Raton, FL. MR2459796 https://doi.org/10.1201/9781420011180

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. CRC press, Boca Raton.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2001). *The Elements of Statistical Learning* **1**. Springer series in statistics, New York.

GILL, R. D., VAN DER LAAN, M. J. and ROBINS, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis* 255–294.

HALL, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer, Berlin.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176 https://doi.org/10.1214/ss/1009212519

HOONHOUT, P. and RIDDER, G. (2019). Nonignorable attrition in multi-period panels with refreshment samples. *J. Bus. Econom. Statist.* **37** 377–390. MR3968379 https://doi.org/10.1080/07350015.2017.1345744

HOROWITZ, J. L. and MANSKI, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Amer. Statist. Assoc.* **95** 77–88. MR1803142 https://doi.org/10.2307/2669526

LINERO, A. R. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika* **104** 327–341. MR3698257 https://doi.org/10.1093/biomet/asx015

LITTLE, R. J. (1993a). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. (1993b). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

MALINSKY, D., SHPITSER, I. and TCHETGEN, E. J. T. (2019). Semiparametric inference for non-monotone missing-not-at-random data: The no self-censoring model. arXiv preprint. Available at arXiv:1909.01848.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

MOHAN, K. and PEARL, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems* 1520–1528.

MOHAN, K. and PEARL, J. (2021). Graphical models for processing missing data. *J. Amer. Statist. Assoc.* **116** 1023–1037. MR4270041 https://doi.org/10.1080/01621459.2021.1874961

MOHAN, K., PEARL, J. and TIAN, J. (2013). Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems* 1277–1285.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G. and DIGGLE, P. J. (1998). Monotone missing data and pattern-mixture models. *Stat. Neerl.* **52** 153–161. MR1649081 https://doi.org/10.1111/1467-9574.00075

MOLENBERGHS, G., FITZMAURICE, G., KENWARD, M. G., TSIATIS, A. and VERBEKE, G. (2014). *Handbook of Missing Data Methodology*. CRC Press/CRC, Boca Raton.

NABI, R., BHATTACHARYA, R. and SHPITSER, I. (2020). Full law identification in graphical models of missing data: Completeness results. arXiv preprint. Available at arXiv:2004.04872.

ROBINS, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat. Med.* **16** 21–37.

ROBINS, J. M. and GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* **16** 39–56.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681 https://doi.org/10.1007/978-1-4612-1284-3_1

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. MR2117498

SADINLE, M. and REITER, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika* **104** 207–220. MR3626477 https://doi.org/10.1093/biomet/asw063

SEAMAN, S. R. and VANSTEELANDT, S. (2018). Introduction to double robust methods for incomplete data. *Statist. Sci.* **33** 184–197. MR3797709 https://doi.org/10.1214/18-STS647

SHPITSER, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems* 3144–3152.

SHPITSER, I., MOHAN, K. and PEARL, J. (2015). Missing data as a causal and probabilistic problem Technical report, California Univ. Los Angeles Dept. of Computer Science.

SUN, B. and TCHETGEN TCHETGEN, E. J. (2018). On inverse probability weighting for nonmonotone missing at random data. *J. Amer. Statist. Assoc.* **113** 369–379. MR3803471 https://doi.org/10.1080/01621459.2016.1256814

TCHETGEN TCHETGEN, E. J., WANG, L. and SUN, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statist. Sinica* **28** 2069–2088. MR3839004

THIJS, H., MOLENBERGHS, G., MICHIELS, B., VERBEKE, G. and CURRAN (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3** 245–265.

TIAN, J. (2015). Missing at random in graphical models. In *Artificial Intelligence and Statistics* 977–985.

TSIATIS, A. (2007). *Semiparametric Theory and Missing Data*. Springer, Berlin.

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311