Model Selection over Partially Ordered Sets

Armeen Taeb[a], Peter Bühlmann[b], and Venkat Chandrasekaran[c]

[a]Department of Statistics, University of Washington
[b]Seminar for Statistics, ETH Zürich
[c]Departments of Computing and Mathematical Sciences and of Electrical Engineering, California Institute of Technology

August 24, 2023

Abstract

In problems such as variable selection and graph estimation, models are characterized by Boolean logical structure such as presence or absence of a variable or an edge. Consequently, false positive and false negative errors can be specified as the number of variables or edges that are incorrectly included/excluded in an estimated model. However, there are several other problems such as ranking, clustering, and causal inference in which the associated model classes do not admit transparent notions of false positive and false negative errors due to the lack of an underlying Boolean logical structure. In this paper, we present a generic approach to endow a collection of models with partial order structure, which leads to a hierarchical organization of model classes as well as natural analogs of false positive and false negative errors. We describe model selection procedures that provide false positive error control in our general setting and we illustrate their utility with numerical experiments.

Keywords: combinatorics | greedy algorithms | multiple testing | stability

1 Introduction

In data-driven approaches to scientific discovery, one is commonly faced with the problem of model selection. Popular examples include variable selection (which covariates influence a response?) and graph estimation (which pairs of variables have nonzero correlation or partial correlation?). As exemplified by these two problems, a common feature of most model selection problems in the literature is that the collection of models is organized according to some type of Boolean logical structure, such as presence versus absence of a variable or an edge. A consequence of such structure is that model complexity can be conveniently specified as the number of attributes (variables or edges) in a model, while false positives and false negatives correspond to the number of attributes that are incorrectly included or excluded in the model, respectively.

In many contemporary applications, models represent a far richer range of phenomena that are not conveniently characterized via Boolean logical structure. As a first example, suppose we are given observations of covariate-response pairs and we wish to order the covariates based on how much they influence a response; the collection of models is given by the set of rankings of the covariates. Second, consider a clustering problem in which we are given observations of a collection of variables and the goal is to group them according to some measure of affinity, with the number of groups and the number of variables assigned to each group not known a priori; here the model class is given by the collection of all possible partitions of the set of variables. Third, suppose we wish to identify causal relations underlying a collection of variables; the model class is the set of completed partially directed acyclic graphs. Finally, consider the blind source separation problem in which we are given a signal expressed as an additive combination of source signals and our objective is to identify the constituent sources, without prior information about the number of sources or their content; here the model class is the collection of all possible linearly independent subsets of vectors.

In these preceding examples, we lack a systematic definition of model complexity, false positives, and false negatives due to the absence of Boolean logical structure in each collection of models. In particular, in

the first three examples, valid models are characterized by structural properties such as transitivity, set partitioning, and graph acyclicity, respectively; these properties are global in nature and are not concisely modeled via separable and local characteristics such as an attribute (a variable or edge) being included in a model independently of other attributes. In the fourth example concerning blind source separation, false positive and false negative errors should not be defined merely via the inclusion or exclusion of true source vectors in an estimated set but should instead consider the degree of alignment between the estimated and true sources, which again speaks to the lack of a natural Boolean logical structure underlying the associated class of models.

As a concrete illustration of the inappropriateness of Boolean logical structure for the task of clustering, consider three items a, b, c, with the true model given by the two clusters $\{a, b\}, \{c\}$ and the estimated model given by the single cluster $\{a, b, c\}$. An incorrect perspective grounded in Boolean logical structure would suggest a false positive error of two, with the mistakes being that c is in the same cluster as a and as b. On the other hand, accounting for set partition structure would yield the more accurate false positive error value of one as a and b are in the same cluster in the true and estimated models, and therefore including c in the same cluster as $\{a, b\}$ should only incur one false discovery.

While the preceding four problems have been studied extensively, the associated methods do not systematically control for false positive error as this quantity is not formally defined. Selection procedures that yield models with small false positive error play an important role in data-driven methods for gathering evidence, rooted in the empirical philosophy and statistical testing foundations of falsification of theories and hypotheses [6, 16, 17].

1.1 Our Contributions

We begin in Section 2 by describing how collections of models may be endowed with the structure of a partially ordered set (poset). Posets are relations that satisfy reflexivity, transitivity, and antisymmetry, and they facilitate a hierarchical organization of a set of models that leads to a natural definition of model complexity. Building on this framework, we develop an axiomatic approach to defining functions over poset element pairs for evaluating similarity. This yields generalizations of well-known measures such as family-wise error and false discovery rate to an array of model selection problems in the context of ranking, causal inference, multiple change-point estimation, clustering, multi-sample testing, and blind source separation. In Section 3, we describe two generic model selection procedures that search over poset elements in a greedy fashion and that provide false discovery control in discrete model posets. The first method is based on subsampling and model averaging and it builds on the idea of stability selection [13, 22] for the variable selection problem, while the second method considers a sequence of hypothesis tests between models of growing complexity. With both these methods, the combinatorial properties of a model poset play a prominent role in determining computational and statistical efficiency. Proofs of the theorems of Section 3 are provided in Section 5. In Section 4 we provide numerical illustration via experiments on synthetic and real data. The code for implementing our methods is available at `https://github.com/armeentaeb/model-selection-over-posets`.

1.2 Related Work

Classic approaches to model selection such as the AIC and BIC assess and penalize model complexity by counting the number of attributes included in a model [1, 21]. More generally, such complexity measures facilitate a hierarchical organization of model classes, and this perspective is prevalent throughout much of the model selection literature [7, 8, 12, 18, 27]. However, these complexity measures rely on a Boolean logical structure underlying a collection of models, and are therefore not well-suited to model classes that are not characterized in this manner. The poset formalism presented in this paper is sufficiently flexible to facilitate model selection over model classes that are more complex than those characterized by Boolean logical structure (such as the illustration presented previously with clustering, see also Example 2), while being sufficiently structured to permit precise definitions of model complexity as well as false positive and false negative errors.

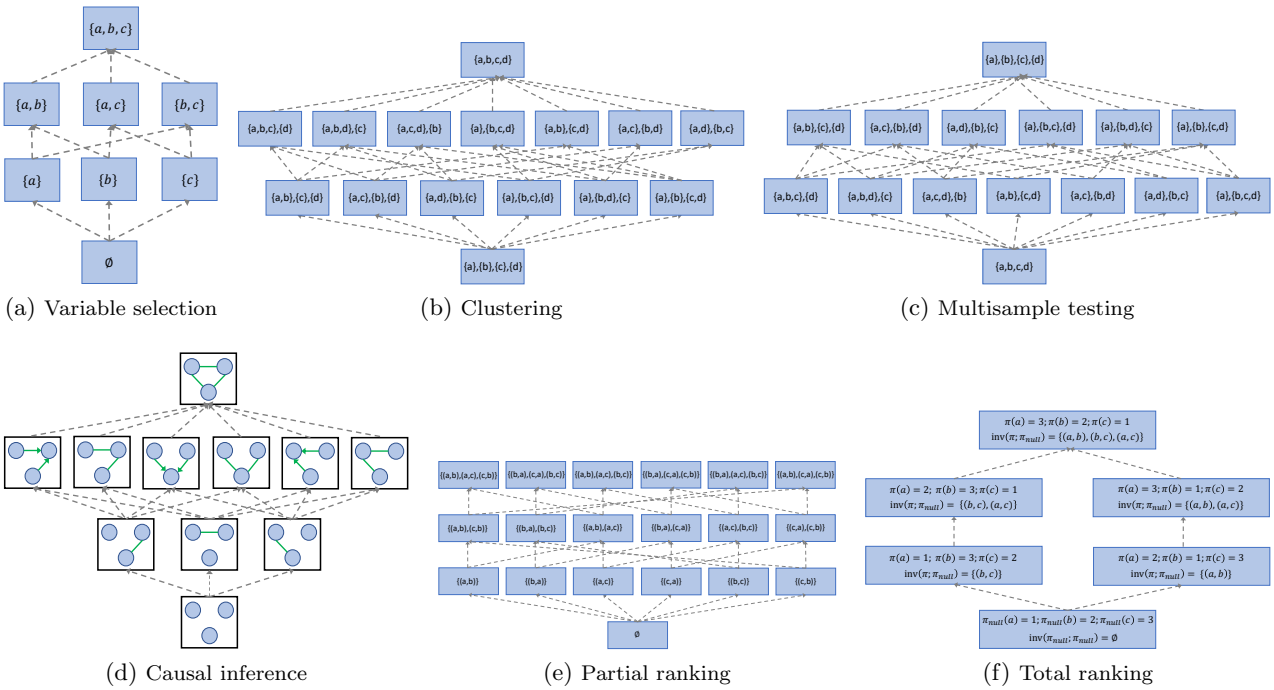
2 Poset Framework for Model Selection

We begin by describing how collections of models arising in various applications may be organized as posets. Next, we present approaches to endow poset-structured models with suitable notions of true and false discoveries.

2.1 Model Classes as Posets

We begin with some basics of posets. A *poset* $(\mathcal{L}, \preceq)$ is a collection $\mathcal{L}$ of elements and a relation $\preceq$ that is reflexive ($x \preceq x,\ \forall x \in \mathcal{L}$), transitive ($x \preceq y, y \preceq z \Rightarrow x \preceq z,\ \forall x, y, z \in \mathcal{L}$), and anti-symmetric ($x \preceq y, y \preceq x \Rightarrow x = y,\ \forall x, y \in \mathcal{L}$). An element $y \in \mathcal{L}$ *covers* $x \in \mathcal{L}$ if $x \preceq y$, $x \neq y$, and there is no $z \in \mathcal{L} \backslash \{x, y\}$ with $x \preceq z \preceq y$; we call such (x, y) a *covering pair*. A *path* from $x_1 \in \mathcal{L}$ to $x_k \in \mathcal{L}$ is a sequence $(x_1, \ldots, x_k)$ with $x_2, \ldots, x_{k-1} \in \mathcal{L}$ such that x_i covers x_{i-1} for each $i = 2, \ldots, k$. Throughout this paper, we focus on posets in which there is a *least element*, i.e., an element $x_{\text{least}} \in \mathcal{L}$ such that $x_{\text{least}} \preceq y$ for all $y \in \mathcal{L}$; such least elements are necessarily unique. Finally, a poset is *graded* if there exists a function $\text{rank}(\cdot)$ mapping poset elements to the nonnegative integers such that the rank of the least element is 0 and $\text{rank}(y) = \text{rank}(x) + 1$ for $y \in \mathcal{L}$ that covers $x \in \mathcal{L}$. In graded posets with least elements, each path from the least element to any $x \in \mathcal{L}$ has length equal to $\text{rank}(x)$. Posets are depicted visually using Hasse diagrams in which a directed arrow is drawn from $x \in \mathcal{L}$ to any $y \in \mathcal{L}$ that covers x.

Posets offer an excellent framework to formulate model selection problems as model classes in many applications possess rich partial order structure. In particular, the poset-theoretic quantities in the preceding paragraph have natural counterparts in the context of model selection – the least element corresponds to the 'null' model that represents no discoveries, the relation $\preceq$ specifies a notion of containment between simpler and more complex models, and the rank function serves as a measure of model complexity that respects the underlying containment relation. We present several concrete illustrations next; Figure 1 presents Hasse diagrams associated to several of these examples.



(a) Variable selection (b) Clustering (c) Multisample testing

(d) Causal inference (e) Partial ranking (f) Total ranking

Figure 1: Hasse diagrams for a) variable selection with 3 variables (Example 1); b) clustering 4 variables (Example 2); c) multisample testing with 4 samples (Example 3); d) causal inference with 3 variables (Example 4); e) partial ranking of 3 items (Example 6); and f) total ranking of 3 items (Example 7).

Example 1 (Variable selection)**.** *As a warm-up, consider the variable selection problem of selecting which of p variables influence a response. The poset here is the collection of all subsets of* $\{1, \ldots, p\}$ *ordered by set*

inclusion, the least element is given by the empty set, and the rank of a subset is its cardinality. This poset is called the Boolean poset *[23].*

Example 2 (Clustering)**.** *Suppose we wish to group a collection of p variables based on a given notion of similarity. The poset here is the collection of all partitions of $\{1, \ldots, p\}$ ordered by refinement, the least element is given by p groups each consisting of one variable, and the rank of a partition is equal to p minus the number of groups. Thus, higher-rank elements correspond to models specified by a small number of clusters. This poset is called the* partition poset *[23].*

Example 3 (Multisample testing)**.** *As a generalization of the classic two-sample testing problem, consider the task of grouping p samples with the objective that samples in a group come from the same distribution. Although this problem is closely related to the preceding clustering problem, it is more natural for the underlying poset here to be the reverse of the partition poset that is formed by reversing the order relation of the partition poset, i.e., the poset is the collection of all partitions of $\{1, \ldots, p\}$ ordered by coarsening. With this reverse ordering, the least element corresponds to all p samples belonging to the same group (i.e., coming from the same distribution), which generalizes the usual null hypothesis in two-sample testing. The rank of a partition is equal to the number of groups minus one. Thus, higher-rank elements correspond to the p samples arising from many distinct distributions.*

Example 4 (Causal structure learning)**.** *Causal associations among a collection of variables are often characterized by a directed acyclic graph (DAG), namely a graph with directed edges and no (directed) cycles, in which the nodes index the variables. Causal structure learning entails inferring this DAG from observations of the variables. The structure of a DAG specifies a causal model via conditional independence relations among the variables, with denser DAGs encoding fewer conditional independencies in comparison with sparser DAGs. (See [4] for details on how the structure of a DAG encodes conditional independence relations; here we describe only those aspects that pertain to a poset formulation to organize the collection of all causal models based on graph structure.) Distinct DAGs can specify the same set of conditional independence relations, and these are called Markov equivalent DAGs. We introduce some terminology to characterize Markov equivalent DAGs. The* skeleton *of a DAG is the undirected graph obtained by making all the edges undirected. A* v-structure *is a set of three nodes x, y, z such that there are directed edges from x to z and from y to z, and there is no edge between x and y. Two DAGs are Markov equivalent if and only if they have the same skeleton and the same collection of v-structures. A Markov equivalence class of DAGs can be described by a completed partially DAG (CPDAG), which is a graph consisting of both directed and undirected edges. A CPDAG has a directed edge from a node x to a node y if and only if this directed edge is present in every DAG in the associated Markov equivalence class. A CPDAG has an undirected edge between nodes x and y if the corresponding Markov equivalence class contains a DAG with a directed edge from x to y and a DAG with a directed edge from y to x. One can check that the total number of edges in a CPDAG (directed plus undirected) is equal to the number of edges in any DAG in the associated Markov equivalence class. The collection of CPDAGs on p variables may be viewed as a poset ordered by inclusion – CPDAGs $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ satisfy $\mathcal{C}^{(1)} \preceq \mathcal{C}^{(2)}$ if and only if there exist DAGs $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$ in the respective Markov equivalence classes such that $\mathcal{G}^{(1)}$ is a directed subgraph of $\mathcal{G}^{(2)}$. In other words, $\mathcal{C}^{(1)} \preceq \mathcal{C}^{(2)}$ if and only if all the conditional independencies encoded by $\mathcal{C}^{(2)}$ are also encoded by $\mathcal{C}^{(1)}$, or equivalently that all the conditional dependencies encoded by $\mathcal{C}^{(1)}$ are also encoded by $\mathcal{C}^{(2)}$. The least element is given by the CPDAG with no edges, and the rank function is equal to the number of edges. Higher-rank elements in this poset correspond to causal models exhibiting more conditional dependence relations.*

Example 5 (Multiple changepoint estimation)**.** *Consider the problem of detecting changepoints in a multivariate time series. Specifically, we observe p signals each for time instances $t = 0, \ldots, T-1$, each signal consists of at most one change (e.g., a change in the distribution or dynamics underlying the signal observations), and the objective is to identify these changes. We denote changepoints via vectors $x = (x_1, \ldots, x_p) \in \{0, \ldots, T\}^p$, with x_i denoting the time index when a change occurs in the i'th signal and $x_i = T$ corresponding to no change occurring. The poset here is the set $\{0, 1, \ldots, T\}^p$ ordered such that $x \preceq y$ if and only if $x_j \geq y_j$ for all $j = 1, \ldots, p$, the least element is $(T, \ldots, T)$, and the rank of an element is $p \cdot T$ minus the sum of the coordinates. Higher-rank elements correspond to changepoint estimates in which the changes occur early. This poset is the reverse of the (bounded)* integer poset *[23] with the product order.*

Example 6 (Partial ranking)**.** *We seek a ranking of a finite set of items given noisy observations (e.g., pairwise*

comparisons), and we allow some pairs of items to be declared as incomparable. Such a partial ranking *of the elements of a finite set S corresponds to a* strict partial order *on S, i.e., a relation $\mathcal{R}$ that is irreflexive ($(a,a) \notin \mathcal{R},\ \forall a \in S$), asymmetric ($(a,b) \in \mathcal{R} \Rightarrow (b,a) \notin \mathcal{R},\ \forall a,b \in S$), and transitive; if an element of S does not appear in $\mathcal{R}$, then that element is incomparable to any of the other elements of S in the associated partial ranking. The poset here is the collection of strict partial orders on S ordered by inclusion, the least element is the empty set, and the rank of a partial ranking is the cardinality of the associated relation. Thus, higher-rank elements correspond to partial rankings that compare many of the covariates.*

Example 7 (Total ranking)**.** *We again wish to rank a finite collection of items but now we seek a total ranking that provides an ordered list of all the items. The setting is that we are given a total ranking that represents our current state of knowledge (i.e., a 'null model') as well as a new set of noisy observations, and the goal is to identify a total ranking that represents an update of the null model to reflect the new information. Each* total ranking *of the elements of a finite set S corresponds to a one-to-one function from S to the integers $\{1,\ldots,|S|\}$. Let π_{null} be the function that describes the null ranking. A convenient way to compare total rankings and to define a poset structure over them is via the notion of an* inversion set*. For any total ranking specified by a function π, the associated inversion set (with respect to the null ranking π_{null}) is defined as $\text{inv}(\pi;\pi_{\text{null}}) \triangleq \{(x,y) \in S \times S \mid \pi_{\text{null}}(x) < \pi_{\text{null}}(y),\ \pi(x) > \pi(y)\}$. The poset here (with respect to a given null ranking π_{null}) is the collection of total rankings on S ordered by inclusion of the associated inversion sets, the least element is the null ranking π_{null}, and the rank of a total ranking is the cardinality of the associated inversion set; this rank function is also equal to the Kendall tau distance between a total ranking and π_{null}. Thus, higher-rank elements are given by total rankings that depart significantly from the null ranking π_{null}. This poset is called the* permutation poset *[23].*

Example 8 (Subspace estimation)**.** *The task is to estimate a subspace in $\mathbb{R}^p$ given noisy observations of points in the subspace. The poset is the collection of subspaces in $\mathbb{R}^p$ ordered by inclusion, the least element is the subspace $\{0\}$, and the rank of a subspace is its dimension. This poset is called the* subspace poset*.*

Example 9 (Blind source separation)**.** *We are given a signal in $\mathbb{R}^p$ that is expressed as a linear combination of some unknown source signals and the goal is to estimate these sources. The poset here is the collection of linearly independent subsets of unit-norm vectors in $\mathbb{R}^p$ ordered by inclusion, the least element is the empty set, and the rank of a linearly independent subset is equal to the cardinality of the subset.*

With respect to formalizing the notion of false positive and false negative errors, Example 1 is prominently considered in the literature, while Examples 3 and 5 are multivariate generalizations of previously studied cases [9, 11]. Finally, Example 8 was studied in [24], although that treatment proceeded from a geometric perspective rather than the order-theoretic approach presented in this paper. With the exception of Example 1, none of the other examples permit a natural formulation within the traditional multiple testing paradigm due to the lack of a Boolean logical structure underlying the associated model classes. Moreover, Examples 8-9 are model classes consisting of infinitely many elements. Nonetheless, we describe in the sequel how the poset formalism enables a systematic and unified framework for formulating model selection in all of the examples above.

2.2 Evaluating True and False Discoveries

To assess the extent to which an estimated model signifies discoveries about the true model, we describe next a general approach to quantify similarity between poset elements in a manner that respects partial order structure.

Definition 1 (similarity valuation)**.** *Let $(\mathcal{L},\preceq,\text{rank}(\cdot))$ be a graded poset. A function $\rho : \mathcal{L}\times\mathcal{L} \to \mathbb{R}$ that is symmetric, i.e., $\rho(x,y) = \rho(y,x)$ for all $x,y \in \mathcal{L}$, is called a* similarity valuation *over $\mathcal{L}$ if:*

- $0 \le \rho(x,y) \le \min\{\text{rank}(x),\text{rank}(y)\}$ *for all $x,y \in \mathcal{L}$,*
- $\rho(x,y) \le \rho(z,y)$ *for all $x \preceq z$,*
- $\rho(x,y) = \text{rank}(x)$ *if and only if $x \preceq y$.*

Remark 1. *The term 'valuation' is often used in the order-theory literature [23] to denote functions on posets that respect the underlying partial order structure, and we use it in our context for the same reason.*

In the sequel we describe similarity valuations for the various model posets discussed previously. The conditions above make similarity valuations well-suited for quantifying the amount of discovery in an estimated model with respect to a true model. The first condition states that the amount of discovery must be bounded above by the complexities of the true and estimated models (which are specified by the rank function). The second condition requires similarity valuations to respect partial order structure so that more complex models do not yield less discovery than less complex ones. The final condition expresses the desirable property that the amount of discovery contained in an estimated model is equal to the complexity of that model if and only if it is 'contained in' the true model. With these properties, we obtain the following analogs of true and false discoveries and of related quantities such as false discovery proportion.

Definition 2 (true and false discoveries)**.** *Let* $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$ *be a graded poset and let* ρ *be a similarity valuation on* $\mathcal{L}$*. Letting* $x^\star \in \mathcal{L}$ *be a true model and* $\hat{x} \in \mathcal{L}$ *be an estimate, the* true discovery*, the* false discovery*, and the* false discovery proportion *are, respectively, defined as follows:*

$$\begin{aligned}
&\mathrm{TD}(\hat{x}, x^\star) \triangleq \rho(\hat{x}, x^\star), \\
&\mathrm{FD}(\hat{x}, x^\star) \triangleq \mathrm{rank}(\hat{x}) - \rho(\hat{x}, x^\star) = \mathrm{rank}(\hat{x}) - \mathrm{TD}(\hat{x}, x^\star), \\
&\mathrm{FDP}(\hat{x}, x^\star) \triangleq \frac{\mathrm{rank}(\hat{x}) - \rho(\hat{x}, x^\star)}{\mathrm{rank}(\hat{x})} = \frac{\mathrm{FD}(\hat{x}, x^\star)}{\mathrm{rank}(\hat{x})}.
\end{aligned}$$

With these definitions, we articulate our model selection objective more precisely:

Goal: *identify the largest rank model subject to control in expectation or in probability on false discovery (proportion).*

This objective is akin to seeking the largest amount of discovery subject to control on false discovery (rate). The data available to carry out model selection vary across our examples; in Section 3 we describe methods to obtain false discovery control guarantees in various settings.

To carry out this program, a central question is the choice of a suitable similarity valuation for a graded model poset. Indeed, it is unclear whether there always exists a similarity valuation for any graded model poset $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$. To address this question, consider the following function for $x, y \in \mathcal{L}$:

$$\rho_{\mathrm{meet}}(x, y) \triangleq \max_{z \preceq x, z \preceq y} \mathrm{rank}(z). \tag{1}$$

Remark 2. *In order theory, a poset* $(\mathcal{L}, \preceq)$ *is said to possess a* meet *if for each* $x, y \in \mathcal{L}$ *there exists a* $z \in \mathcal{L}$ *satisfying* (i) $z \preceq x, z \preceq y$ *and* (ii) *for any* $w \in \mathcal{L}$ *with* $w \preceq x, w \preceq y$*, we have* $w \preceq z$*; such a* z *is called the meet of* x, y *and posets that possess a meet are called* meet semi-lattices*. Except for the poset in Example 4 on causal structure learning, the posets in the other examples are meet semi-lattices (see Appendix A). The subscript 'meet' in* (1) *signifies that* ρ_{meet} *is the rank of the meet for meet semi-lattices, although* ρ_{meet} *is well-defined even if* $(\mathcal{L}, \preceq)$ *is not a meet semi-lattice.*

One can check that ρ_{meet} is a similarity valuation on any graded poset $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$; see Appendix B for a proof. For Example 1 on variable selection, ρ_{meet} has the desirable property that it reduces to the number of common variables in two models; thus, the general model selection goal formulated above reduces to the usual problem of maximizing the number of selected variables subject to control on the number of selected variables that are null. Next we describe the model selection problems we obtain in Examples 2-6 with ρ_{meet} as the choice of similarity valuation.

In Example 2 on clustering, the value of ρ_{meet} for two partitions of p variables is equal to p minus the number of groups in the coarsest common refinement of the partitions. The model selection problem is that of partitioning the variables into the smallest number of groups subject to control on the additional number of groups in the coarsest common refinement of the estimated and true partitions compared to the number of groups in the estimated partition.

Recall that the poset in Example 3 on multisample testing is the reverse of the poset in Example 2; thus, many of the notions from the preceding paragraph are appropriately 'reversed' in Example 3. In particular, the value of ρ_{meet} in Example 3 for two partitions of p samples is equal to the number of groups in the finest

common coarsening of the partitions. The model selection problem entails partitioning the samples into the largest number of groups subject to control on the additional number of groups in the estimated partition compared to the number of groups in the finest common coarsening of the estimated and true partitions.

In Example 4 on causal structure learning, the value of ρ_{meet} for two CPDAGs $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ is equal to the maximum number of edges in a CPDAG that encodes all the conditional independencies of $\mathcal{C}^{(1)}$ and of $\mathcal{C}^{(2)}$. The model selection task is then to identify the CPDAG with the largest number of edges subject to control on the additional number of edges in the estimated CPDAG compared to the densest CPDAG that encodes all the conditional independence relationships in both the true and estimated CPDAGs.

In Example 5 on multiple changepoint estimation, suppose $x, y \in \{0, \ldots, T\}^p$ are vectors of time indices specifying changepoints in p signals. We have that $\rho_{\text{meet}}(x, y) = p \cdot T - \sum_{i=1}^{p} \max\{x_i, y_i\}$. The model selection problem entails identifying changes as quickly as possible subject to control on early detection of changes (i.e., declaring changes before they occur); this is a multivariate generalization of the classic quickest change detection problem [11].

In Example 6 on partial ranking, the value of ρ_{meet} for two partial rankings is equal to the cardinality of the intersection of the associated relations, i.e., the number of common comparisons in the two partial rankings. The associated model selection problem is that of identifying a partial ranking with the largest number of comparisons (i.e., the associated relation must have large cardinality) subject to control on the number of comparisons in the estimated partial ranking that are not in the true partial ranking.

In Examples 1-6, the function ρ_{meet} of (1) provides a convenient way to assess the amount of discovery in an estimated model with respect to a true model, thereby yielding natural formulations for model selection. However, in Examples 7-9, ρ_{meet} has some undesirable features.

Consider first the setup in Example 7 on total ranking for the set $S = \{a, b, c\}$ with the null model given by the ranking $\pi_{\text{null}}(a) = 1, \pi_{\text{null}}(b) = 2, \pi_{\text{null}}(b) = 3$, the true model given by the ranking $\pi^\star(a) = 3, \pi^\star(b) = 1, \pi^\star(c) = 2$ (Hasse diagram shown in Figure 1), and the estimated ranking given by $\hat{\pi}(a) = 2, \hat{\pi}(b) = 3, \hat{\pi}(c) = 1$. In this case, one can see from Figure 1 that $\rho_{\text{meet}}(\hat{\pi}, \pi^\star) = 0$, which suggests that no discovery is made. On the other hand, the inversion sets of these rankings are given by $\text{inv}(\pi^\star; \pi_{\text{null}}) = \{(a, b), (a, c)\}$ and $\text{inv}(\hat{\pi}; \pi_{\text{null}}) = \{(a, c), (b, c)\}$, and the element (a, c) is common to both inversion sets as the fact that item c is ranked higher than item a in the true model has been discovered in the estimated model; this reasoning suggests that a positive quantity would be a more appropriate value for the similarity valuation between $\hat{\pi}$ and $\pi^\star$. The key issue is that $\text{inv}(\pi^\star; \pi_{\text{null}}) \cap \text{inv}(\hat{\pi}; \pi_{\text{null}})$ is not an inversion set of any total ranking, but this intersection still carries valuable information about true discoveries made in $\hat{\pi}$ about $\pi^\star$. However, the similarity valuation ρ_{meet} only considers subsets of $\text{inv}(\pi^\star; \pi_{\text{null}}) \cap \text{inv}(\hat{\pi}; \pi_{\text{null}})$ that correspond to inversion sets of total rankings as the maximization in (1) is constrained to be over poset elements. Motivated by this discussion, we employ the following similarity valuation in Example 7 for total rankings $\pi, \tilde{\pi}$ (with respect to a null model π_{null}):

$$\rho_{\text{total-ranking}}(\pi, \tilde{\pi}) = |\text{inv}(\pi; \pi_{\text{null}}) \cap \text{inv}(\tilde{\pi}; \pi_{\text{null}})|. \tag{2}$$

With this similarity valuation, the model selection problem reduces to identifying a total ranking with the largest inversion set (with respect to π_{null}) subject to control on the number of comparisons in the inversion set of the estimated total ranking that are not in the inversion set of the true total ranking.

Next, in Example 8, $\rho_{\text{meet}}(\hat{x}, x^\star)$ is equal to the dimension of the intersection of the subspaces $\hat{x}, x^\star$. When these subspaces have small dimension, for example, ρ_{meet} generically equals zero regardless of the angle between the subspaces; in words, ρ_{meet} does not consider the smooth structure underlying the collection of subspaces. As discussed in [24], a more suitable measure of similarity is the sum of the squares of the cosines of the principal angles between the subspaces, which is expressed as follows using projection matrices onto subspaces $\mathcal{U}, \tilde{\mathcal{U}}$:

$$\rho_{\text{subspace}}(\mathcal{U}, \tilde{\mathcal{U}}) = \text{trace}(\mathcal{P}_{\mathcal{U}} \mathcal{P}_{\tilde{\mathcal{U}}}). \tag{3}$$

The model selection task is to identify the largest-dimensional subspace subject to control on the sum of the squares of the cosines of the principal angles between the estimated subspace and the orthogonal complement of the true subspace.

problem domain	**models**	**least element (i.e. global null)**	**partial order**	**rank (i.e. model complexity)**	**similarity valuation (i.e. true discoveries)**
variable selection	subsets of $\{1,\dots,p\}$	$\emptyset$	inclusion of subsets	cardinality of subset	subsets $x,\tilde{x}$; $\rho(x,\tilde{x}) = \lvert x \cap \tilde{x}\rvert$
clustering	partitions of $\{1,\dots,p\}$	$\{1\},\{2\},\dots,\{p\}$	refinement of partition	$p - \#\text{groups}$	partitions $x,\tilde{x}$; $\rho(x,\tilde{x}) = p - \#$ groups in coarsest common refinement
multisample testing	partitions of $\{1,\dots,p\}$	$\{1,2,\dots,p\}$	coarsening of partition	#groups	partitions $x,\tilde{x}$; $\rho(x,\tilde{x}) = \#$ groups in finest common coarsening
causal structure learning	completed partially directed acyclic graphs (CPDAG) on a set of variables	CPDAG with no edges	inclusion of conditional dependencies encoded by CPDAGs	#edges	CPDAGs $\mathcal{C},\tilde{\mathcal{C}}$; $\rho(\mathcal{C},\tilde{\mathcal{C}}) = \#$edges in densest CPDAG encoding conditional independencies of both $\mathcal{C},\tilde{\mathcal{C}}$
multiple changepoint	elements of $\{0,\dots,T\}^p$	$(T,T,\dots,T)$	entrywise reverse ordering	$p\cdot T$ minus sum of entries	changepoint vectors $x,\tilde{x}$; $\rho(x,\tilde{x}) = p\cdot T - \sum_i \max\{x_i,\tilde{x}_i\}$
partial ranking	relations specified by strict partial orders on a set of items	$\emptyset$	inclusion of sets specifying relations	cardinality of set specifying relation	sets $\mathcal{R},\tilde{\mathcal{R}}$ specifying relations; $\rho(\mathcal{R},\tilde{\mathcal{R}}) = \lvert\mathcal{R}\cap\tilde{\mathcal{R}}\rvert$
total ranking	total orders on a set of items	base ranking π_{null}	inclusion of inversion sets w.r.t. π_{null}	cardinality of inversion set w.r.t. π_{null}	total orders $\pi,\tilde{\pi}$; $\rho(\pi,\tilde{\pi}) = \lvert\text{inv}(\pi;\pi_{\text{null}})\cap\text{inv}(\tilde{\pi};\pi_{\text{null}})\rvert$
subspace estimation	subspaces in $\mathbb{R}^p$	$\{0\}$	inclusion of subspaces	dimension of subspace	subspaces $\mathcal{U},\tilde{\mathcal{U}}$; $\rho(\mathcal{U},\tilde{\mathcal{U}}) = \text{trace}(\mathcal{P}_{\mathcal{U}}\mathcal{P}_{\tilde{\mathcal{U}}})$
blind source separation	linearly independent subsets of $\mathbb{R}^p$	$\emptyset$	inclusion of subsets	cardinality of subset	subsets given by columns of $B\in\mathbb{R}^{p\times k},\tilde{B}\in\mathbb{R}^{p\times\ell},k\le\ell$; $\rho(B,\tilde{B}) = \max_{\sigma\in\text{Perm}(\ell)}\sum_{i=1}^{k}(B^T\tilde{B})^2_{i,\sigma(i)}$

Table 1: Problem classes and associated characterization of model selection via posets.

Finally, ρ_{meet} is inadequate as a similarity valuation in Example 9 for the same reasons as in Example 8 due to the underlying smooth structure, and we propose here a more appropriate alternative. Given $B\in\mathbb{R}^{p\times k}$, $\tilde{B}\in\mathbb{R}^{p\times\ell}$ (these matrices have unit-norm and linearly independent columns representing source signals), suppose without loss of generality that $k\le\ell$ (due to the symmetry of similarity valuations) and let $\text{Perm}(\ell)$ be the collection of bijections on $\{1,\dots,\ell\}$. With this notation, consider the following similarity valuation:

$$\rho_{\text{source-separation}}(B,\tilde{B}) = \max_{\sigma\in\text{Perm}(\ell)}\sum_{i=1}^{k}(B^T\tilde{B})^2_{i,\sigma(i)}. \tag{4}$$

This valuation is better suited to quantify the degree of alignment between two collections of vectors in source separation than ρ_{meet}. Model selection entails identifying the largest collection of source vectors subject to control on the difference in the number of estimated source vectors and the alignment between the true and estimated source vectors as evaluated by $\rho_{\text{source-separation}}$.

Table 1 summarizes our discussion of the various model posets and their associated similarity valuations. In conclusion, while ρ_{meet} is a similarity valuation for any model poset, it is not always the most natural choice and identifying a suitable similarity valuation that captures the essential features of an application is key to properly formulating a model selection problem. This situation is not unlike the selection of an appropriate loss function in point estimation – while there exist many candidates that are mathematically valid, the utility of an estimation procedure in the context of a problem domain depends critically on a well-chosen loss.

3 False discovery control over posets

In this section we turn our attention to the task of identifying models of large rank that provide false discovery control. We begin in Section 3.1 with a general greedy strategy for poset search that facilitates the design of model selection procedures, and we specialize this framework to specific approaches in Sections 3.2 and 3.3. Some of the discussion in Section 3.1 is relevant for all of the posets in Examples 1-9, while the methodology presented in Sections 3.2-3.3 is applicable to general discrete posets with integer-valued similarity valuations such as in Examples 1-7. Along the way, we remark on some of the challenges that arise in the two continuous cases of Examples 8-9.

3.1 Greedy Approaches to Model Selection

To make progress on the problem of identifying large rank models that provide control on false discovery, we begin by noting that the false discovery $\mathrm{FD}(\hat{x}, x^\star)$ in an estimated model $\hat{x}$ with respect to a true model $x^\star$ may be expressed as the following telescoping sum for any path $(x_0, x_1, \ldots, x_{k-1}, x_k)$ with x_0 being the least element x_{least} and $x_k = \hat{x}$:

$$\mathrm{FD}(\hat{x}, x^\star) = \sum_{i=1}^{k} 1 - [\rho(x_i, x^\star) - \rho(x_{i-1}, x^\star)]. \tag{5}$$

The term $1 - [\rho(x_i, x^\star) - \rho(x_{i-1}, x^\star)]$ may be interpreted as the "additional false discovery" incurred by the model x_i relative to the model x_{i-1}. The above decomposition of false discovery in terms of a path from the least element to an estimated model suggests a natural approach for model selection. In particular, we observe that a sufficient condition for $\mathrm{FD}(\hat{x}, x^\star)$ to be small is for each term in the above sum to be small. Thus, we will greedily grow a path starting from the least element $x_0 = x_{\text{least}}$ by adding one element x_i at a time such that each (x_{i-1}, x_i) is a covering pair and each $1 - [\rho(x_i, x^\star) - \rho(x_{i-1}, x^\star)]$ is small. We continue this process until we can no longer guarantee that $1 - [\rho(x_i, x^\star) - \rho(x_{i-1}, x^\star)]$ is small.

For such a procedure to be fruitful, we require some data-driven method to bound $1 - [\rho(x_i, x^\star) - \rho(x_{i-1}, x^\star)]$ as the true model $x^\star$ is not known. Our objective therefore is to design a data-dependent function $\Psi : \{(a, b) \mid b \text{ covers } a \text{ in } \mathcal{L}\} \to [0, 1]$ that takes as input covering pairs and outputs a number in the interval $[0, 1]$, and further satisfies the property that $\Psi(u, v)$ being small is a sufficient condition for $1 - [\rho(v, x^\star) - \rho(u, x^\star)]$ to be small (in expectation or in probability). Given such a function, we grow a path using the greedy strategy outlined above by identifying at each step a covering pair that minimizes Ψ. Algorithm 1 provides the details. In Sections 3.2 and 3.3, we present two approaches for designing suitable functions Ψ: one based on a notion of stability and the other based on testing. Proofs that both these methods control for false discoveries are presented in Section 5.

Algorithm 1 Greedy sequential algorithm for model selection

1: **Input**: poset $\mathcal{L}$, threshold $\alpha \in [0, 1]$; data-dependent function $\Psi : \{(a, b) \mid b \text{ covers } a \text{ in } \mathcal{L}\} \to [0, 1]$
2: **Greedy selection**: Set $u = x_{\text{least}}$ and perform:
   (a) find $v_{\text{opt}} \in \operatorname{argmin}_{\{(u,v) \mid v \text{ covers } u \text{ in } \mathcal{L}\}} \Psi(u, v)$.
   (b) if $\Psi(u, v_{\text{opt}}) \leq \alpha$, set $u = v_{\text{opt}}$ and repeat steps (2a-2b). Otherwise, stop.
3: **Output**: return $\hat{x} = u$

In designing a suitable function Ψ so that $1 - (\rho(v, x^\star) - \rho(u, x^\star))$ is small (in expectation or in probability) whenever $\Psi(u, v)$ is small, we note that the examples presented in Section 2 exhibit an important invariance. Specifically, in each example there are distinct covering pairs (u, v) and (u', v') such that $1 - [\rho(v, x^\star) - \rho(u, x^\star)] = 1 - [\rho(v', x^\star) - \rho(u', x^\star)]$ for every true model $x^\star$. Accordingly, it is natural that the function Ψ also satisfies the property that $\Psi(u, v) = \Psi(u', v')$; stated differently, one need only specify Ψ for a 'minimal' set of covering pairs. We present next a definition that formalizes this notion precisely.

Definition 3 (Minimal covering pairs)**.** *Consider a graded poset $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$ endowed with a similarity valuation ρ. A subset $\mathcal{S} \subset \{(a, b) \mid b \text{ covers } a \text{ in } \mathcal{L}\}$ of covering pairs in $\mathcal{L}$ is called* minimal *if the following two properties hold:*

- *For each covering pair $(u', v') \notin \mathcal{S}$, there exists $(u, v) \in \mathcal{S}$ with $\mathrm{rank}(v) \leq \mathrm{rank}(v')$ such that $\rho(v, z) - \rho(u, z) = \rho(v', z) - \rho(u', z)$ for all $z \in \mathcal{L}$.*
- *For distinct covering pairs $(u, v), (u', v') \in \mathcal{S}$, there exists some $z \in \mathcal{L}$ such that $\rho(v, z) - \rho(u, z) \neq \rho(v', z) - \rho(u', z)$.*

In words, a minimal set of covering pairs $\mathcal{S}$ for a graded poset $\mathcal{L}$ is an inclusion-minimal collection of smallest rank covering pairs for which it suffices to consider the values of Ψ. For Example 1 on variable selection with the similarity valuation ρ_{meet}, a minimal set of covering pairs is given by $\mathcal{S} = \{(\emptyset, \{i\}) \mid i = 1, \ldots, p\}$ and this minimal set is unique. In general, however, such sets are not unique; see Appendix E where we derive minimal sets of covering pairs for several examples. Minimal sets of covering pairs are significant methodologically from both computational and statistical perspectives. In particular, several of our bounds for discrete posets

depend on the cardinality $|\mathcal{S}|$ and these also involve computations that scale in number of operations with $|\mathcal{S}|$. Therefore, identifying a minimal set of covering pairs that is small in cardinality is central to the success of our proposed methods. In the remainder of this section, we assume that a minimal set of covering pairs $\mathcal{S}$ for a given model poset $\mathcal{L}$ is available.

3.2 Model selection based on stability

Our first method for designing a suitable function Ψ to employ in Algorithm 1 is based on subsampling and corresponding model averaging. We assume that we have access to a base procedure $\hat{x}_{\text{base}}$ that provides model estimates from data as well as a dataset $\mathcal{D}$ consisting of observations drawn from a probability distribution parameterized by the true model $x^\star$, and our approach is to aggregate the model estimates provided by $\hat{x}_{\text{base}}$ on subsamples of $\mathcal{D}$. The requirements on the quality of the procedure $\hat{x}_{\text{base}}$ are quite mild, and we prove bounds in the sequel on the false discovery associated to the aggregated model. In particular, the aggregation method ensures that the averaged model is 'stable' in the sense that it contains discoveries that are supported by a large fraction of the subsamples. Our method generalizes the stability selection method for variable selection [13, 22] and subspace stability selection for subspace estimation [24]. We demonstrate the broad applicability of this methodology in Section 4 by applying it to several examples from Section 2.

Formally, fix a positive even integer B and obtain $B/2$ complementary partitions of the dataset $\mathcal{D}$, each of which partitions $\mathcal{D}$ into two subsamples of equal size. Let this collection of subsamples be denoted $\{\mathcal{D}^{(\ell)}\}_{\ell=1}^{B}$, and let $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ denote the model estimate obtained by applying the base procedure to the subsample $\mathcal{D}^{(\ell)}$. For any covering pair (u, v) of a model poset $\mathcal{L}$, we define:

$$\Psi_{\text{stable}}(u, v) \triangleq 1 - \frac{1}{B}\sum_{\ell=1}^{B} \frac{\rho(v, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})) - \rho(u, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}))}{c_{\mathcal{L}}(u, v)}, \tag{6}$$

where $c_{\mathcal{L}}(u, v) \triangleq \max_{z \in \mathcal{L}} \rho(v, z) - \rho(u, z)$. Appealing to properties of similarity valuations, we have that $\rho(v, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})) - \rho(u, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})) \geq 0$ and $c_{\mathcal{L}}(u, v) \geq 1$. The term $\rho(v, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})) - \rho(u, \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}))$ measures the additional discovery about $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ in the model v relative to the model u, while the quantity $c_{\mathcal{L}}(u, v)$ serves as normalization to ensure that $\Psi_{\text{stable}}(u, v) \in [0, 1]$. In particular, $\Psi_{\text{stable}}(u, v)$ being small implies that the additional discovery represented by the model v over the model u is supported by a large fraction of the subsamples $\{\mathcal{D}^{(\ell)}\}_{\ell=1}^{B}$. Consequently, when Ψ_{stable} is employed in the context of Algorithm 1 in which we greedily grow a path, each 'step' in the path corresponds to a discovery that is supported by a large fraction of the subsamples. We provide theoretical support for this approach in Theorem 10 in the sequel and the proof proceeds by showing that $\Psi_{\text{stable}}(u, v)$ being small implies that $\mathbb{E}[1 - (\rho(u, x^\star) - \rho(v, x^\star))]$ is small; we combine this observation with the telescoping sum formula (5) to obtain a bound on the expected false discovery of the model estimated by Algorithm 1.

When Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ is specialized to Example 1 and Example 8, we obtain the stability selection procedure of [13, 22] and the subspace stability selection method of [24]. For variable selection in particular, Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ outputs the subset of variables that appear in at least a $1-\alpha$ fraction of the models estimated by the base procedure when applied to the subsamples $\{\mathcal{D}^{(\ell)}\}_{\ell=1}^{B}$. More generally, Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ also provides a procedure for model selection in Examples 2-7 corresponding to discrete model posets.

Theorem 10 (false discovery control for Algorithm 1 with $\Psi = \Psi_{\text{stable}}$)**.** *Let $(\mathcal{L}, \preceq, \text{rank}(\cdot))$ be a graded discrete model poset with integer-valued similarity valuation ρ and let $\mathcal{S}$ be an associated set of minimal covering pairs. Let $\hat{x}_{\text{base}}$ be a base estimator. Suppose the dataset $\mathcal{D}$ employed in the computation of Ψ_{stable} consists of i.i.d. observations from a distribution parametrized by the true model $x^\star \in \mathcal{L}$, and suppose $\hat{x}_{\text{sub}}$ is an estimator obtained by applying $\hat{x}_{\text{base}}$ to a subsample of $\mathcal{D}$ of size $|\mathcal{D}|/2$. Fix $\alpha \in (0, 1/2)$ and a positive, even integer B. The output $\hat{x}_{stable}$ from Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ satisfies the following false discovery bound*

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \sum_{\substack{(u,v)\in\mathcal{S} \\ \rho(v,x^\star)=\rho(u,x^\star)}} \frac{\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]^2}{(1-2\alpha)c_{\mathcal{L}}(u, v)^2}. \tag{7}$$

The summation in the false discovery bound (7) is over covering pairs $(u, v) \in \mathcal{S}$ for which there is no additional discovery in the model v over the model u with respect to the true model $x^\star$. For each such (u, v), the term $\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]^2$ characterizes the quality of the base estimator on subsamples; base estimators for which this term is small, when employed in the computation of Ψ_{stable} in the context of Algorithm 1, yield models $\hat{x}_{\text{stable}}$ with small false discovery.

Remark 3. *When specialized to Example 1 on variable selection with similarity valuation ρ_{meet}, we recover Theorem 1 of [22]. Specifically, in (7), we have that $c_{\mathcal{L}}(u, v) = 1$ for any covering pair (u, v) and $\sum_{(u,v)\in\mathcal{S},\rho(v,x^\star)=\rho(u,x^\star)} \mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]^2 = \sum_{\text{null } i} \mathbb{E}[\mathbb{I}[i \text{ selected by } \hat{x}_{\text{sub}}]]^2$.*

Theorem 10 is general in its applicability to all the discrete posets in Section 2, and it provides an intuitive bound on expected false discovery. Nonetheless, it requires a characterization of the quality of the base estimator $\hat{x}_{\text{base}}$ employed on subsamples. When such a characterization is unavailable, the false discovery bound (7) may not be easily computable in practice. To address this shortcoming and obtain easily computable bounds on false discovery, we consider natural assumptions on the estimator $\hat{x}_{\text{sub}}$ corresponding to the base estimator $\hat{x}_{\text{base}}$ applied to subsamples; these assumptions generalize those developed in [13, 24] for stability-based methods for variable selection and subspace estimation. To formulate these assumptions, we introduce some notation. Let $\mathcal{T}_{\text{null}} \triangleq \{(u, v) \text{ covering pair in } \mathcal{L} \mid \rho(v, x^\star) = \rho(u, x^\star)\}$ consist of all covering pairs (u, v) for which there is no additional discovery in the model v over the model u with respect to the true model $x^\star$, let $\text{rank}(\mathcal{L}) \triangleq \max_{u\in\mathcal{L}} \text{rank}(u)$ be the largest rank of an element in $\mathcal{L}$, and let $\mathcal{S}_k \triangleq \{(u, v) \in \mathcal{S} \mid \text{rank}(v) = k\}$ for each $k \in [\text{rank}(\mathcal{L})]$.

Assumption 1 (better than random guessing)**.** *For each $k \in [\text{rank}(\mathcal{L})]$ with $\mathcal{S}_k \neq \emptyset$, we have that*

$$\sum_{(u,v)\in\mathcal{S}_k\cap\mathcal{T}_{\text{null}}} \frac{1}{|\mathcal{S}_k \cap \mathcal{T}_{\text{null}}|} \cdot \frac{\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u, v)}$$
$$\leq \sum_{(u,v)\in\mathcal{S}_k\setminus\mathcal{T}_{\text{null}}} \frac{1}{|\mathcal{S}_k \setminus \mathcal{T}_{\text{null}}|} \frac{\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u, v)}.$$

Assumption 2 (invariance in mean)**.** *For each $k \in [\text{rank}(\mathcal{L})]$ with $\mathcal{S}_k \neq \emptyset$, we have that $\frac{\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u, v)}$ is the same for each $(u, v) \in \mathcal{S}_k \cap T_{\text{null}}$.*

In words, Assumption 1 states that the average normalized difference in similarity valuation of the estimator $\hat{x}_{\text{sub}}$ is smaller over 'null' covering pairs than over non-null covering pairs. Assumption 2 states that the expected value of the normalized difference in similarity of $\hat{x}_{\text{sub}}$ is the same for each 'null' covering pair. For the case of variable selection (Example 1), Assumption 1 reduces precisely to the 'better than random guessing' assumption employed by [13], namely that the expected number of true positives divided by the expected number of false positives selected by the estimator $\hat{x}_{\text{sub}}$ is larger than the same ratio for an estimator that selects variables at random. As a second condition, [13] required that the random variables in the collection $\{\mathbb{I}[i \in \hat{x}_{\text{sub}}] : i \text{ null}\}$ are exchangeable. Our Assumption 2 when specialized to variable selection reduces to the weaker requirement that each of the random variables in the collection $\{\mathbb{I}[i \in \hat{x}_{\text{sub}}] : i \text{ null}\}$ has the same mean. As a second illustration, consider the case of total ranking (Example 7) involving items $a_1, \ldots, a_p$, with the least element π_{null} given by $\pi_{\text{null}}(a_i) = i,\ i = 1, \ldots, p$, the true total ranking by $\pi^\star$, and the estimator on subsamples by $\hat{\pi}_{\text{sub}}$. Fix any $k \in \{1, \ldots, p-1\}$. Assumption 1 states that the expected number of pairs $(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{sub}}; \pi_{\text{null}}) \cap \text{inv}(\pi^\star; \pi_{\text{null}})$ with $j - i = k$ divided by the expected number of pairs $(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{sub}}; \pi_{\text{null}}) \setminus \text{inv}(\pi^\star; \pi_{\text{null}})$ with $j - i = k$ is larger than the same ratio for an estimator that outputs a total ranking at random. Assumption 2 states that the probability that $(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{sub}}; \pi_{\text{null}})$ is the same for all pairs (a_i, a_j) with $j - i = k$ and $(a_i, a_j) \notin \text{inv}(\pi^\star; \pi_{\text{null}})$. See Appendix F for a formal derivation.

Theorem 11 (refined false discovery control for Algorithm 1 with $\Psi = \Psi_{\text{stable}}$)**.** *Consider the setup of Theorem 10, and suppose additionally that Assumptions 1 and 2 are satisfied. The output $\hat{x}_{\text{stable}}$ from Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ satisfies the false discovery bound:*

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \sum_{k\in[\text{rank}(\mathcal{L})],\mathcal{S}_k\neq\emptyset} \frac{q_k^2}{|\mathcal{S}_k|(1-2\alpha)}, \tag{8}$$

where $q_k = \sum_{(u,v)\in\mathcal{S}_k} \mathbb{E}[\rho(v,\hat{x}_{\text{sub}}) - \rho(u,\hat{x}_{\text{sub}})]/c_{\mathcal{L}}(u,v)$.

The quantities in the bound (8) may be readily computed in practice. In particular, each $\mathcal{S}_k$ and $c_{\mathcal{L}}(\cdot,\cdot)$ depends only on the model poset $\mathcal{L}$ and each q_k can be approximated as $q_k \approx \frac{1}{B}\sum_{\ell=1}^{B}\sum_{(u,v)\in\mathcal{S}_k} \frac{\rho(v,\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}))-\rho(u,\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}))}{c_{\mathcal{L}}(u,v)}$. We give characterizations of the sets $\mathcal{S}_k$ and $c_{\mathcal{L}}(\cdot,\cdot)$ for posets corresponding to total ranking, partial ranking, clustering, and causal structure learning in Appendix E.

Remark 4. *Specializing Theorem 11 to the case of variable selection, we arrive at the bound in Theorem 1 of [13]. Specifically, note that for the Boolean poset with the similarity valuation* ρ_{meet}, $\mathcal{S}_k = \emptyset$ *for* $k \geq 2$, $|\mathcal{S}_1| = \#$ *variables, and* $q_1 = \sum_i E[\mathbb{I}[$*variable* i *selected by* $\hat{x}_{\text{sub}}]]$ *is the average number of variables selected by the estimator* $\hat{x}_{\text{sub}}$.

Turning our attention to Examples 8-9, the situation is considerably more complicated with continuous model posets. A result for these two cases under the same setup as in Theorem 10 yields the following bound for $\alpha \in (0, 1/2)$ (see Appendix D):

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \;\leq\; \frac{2\alpha + 2\sqrt{1-\alpha}}{1-\alpha}\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})] + \mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2. \tag{9}$$

The first term in the bound is a function of the average number of discoveries made by the estimator $\hat{x}_{\text{sub}}$, and this term is smaller for $\alpha \approx 0$. The second term in the bound concerns the quality of the estimator $\hat{x}_{\text{sub}}$. Specifically, note that Jensen's inequality implies $\mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2 \leq \mathbb{E}[\text{FD}(\hat{x}_{\text{sub}}, x^\star)]$, so that the improvement provided by the estimator $\hat{x}_{\text{stable}}$ based on subsampling and model averaging over the estimator $\hat{x}_{\text{sub}}$ that simply employs the base estimator on subsamples is characterized by $\text{var}(\text{FD}(\hat{x}_{\text{sub}}, x^\star))$. Thus, the key remaining task as before is to characterize the properties of the estimator $\hat{x}_{\text{sub}}$. However, the difficulty with the continuous examples is that conditions akin to Assumptions 1-2 are substantially more challenging to formulate and analyze at an appropriate level of generality. (One such effort under a limited setting for the case of subspace estimation is described in [24].) It is of interest to develop such a general framework for continuous model posets, and we leave this as a topic for future research.

3.3 Model selection based on testing

Our second approach to designing a suitable function Ψ to employ in Algorithm 1 is based on testing the following null hypothesis for each (minimal) covering pair (u, v) of a discrete model poset $\mathcal{L}$:

$$\begin{aligned} H_0^{u,v} &: \rho(v, x^\star) = \rho(u, x^\star), \\ \Psi_{\text{test}}(u, v) &\triangleq \text{p-value corresponding to } H_0^{u,v}. \end{aligned} \tag{10}$$

The null hypothesis $H_0^{u,v}$ in (10) states that there is no additional discovery about $x^\star$ in the model v relative to the model u, and small values of $\Psi_{\text{test}}(u, v)$ provide evidence for rejecting this null hypothesis and accepting the alternative that $\rho(v, x^\star) > \rho(u, x^\star)$. When Ψ_{test} is employed in the context of Algorithm 1 in which we greedily grow a path, each 'step' in the path corresponds to a discovery for which we have the 'strongest evidence' using the above test. Our next result provides theoretical support for this method.

Theorem 12 (false discovery control for Algorithm 1 with $\Psi = \Psi_{\text{test}}$)**.** *Let* $(\mathcal{L}, \preceq, \text{rank}(\cdot))$ *be a graded discrete model poset with integer-valued similarity valuation* ρ *and let* $\mathcal{S}$ *be an associated set of minimal covering pairs. The output* $\hat{x}_{\text{test}}$ *of Algorithm 1 with* $\Psi = \Psi_{test}$ *satisfies the false discovery bound* $\mathbb{P}\left(\text{FD}(\hat{x}_{\text{test}}, x^\star) > 0\right) \leq \alpha|\mathcal{S}|$.

The multiplicity factor involving the cardinality of the set of minimal covering pairs $\mathcal{S}$ is akin to a Bonferroni-type correction, and it highlights the significance of identifying a set of minimal covering pairs of small cardinality. We emphasize that although Algorithm 1 with $\Psi = \Psi_{\text{test}}$ proceeds via sequential hypothesis testing, the procedure is applicable to general model classes with no underlying Boolean logical structure; in particular, it is the graded poset structure underlying our framework that facilitates such methodology.

As an illustration of the multiplicity factor $|\mathcal{S}|$ for different model posets, we have that $|\mathcal{S}| = p(p-1)$ for partial ranking; $|\mathcal{S}| = \sum_{k=1}^{p-1}\binom{p}{k+1}\sum_{\ell=1}^{k}\binom{k+1}{\ell}$ for clustering; and $|\mathcal{S}| = \frac{p(p-1)}{2}$ for total ranking. See Appendix E for further details.

The graded poset structure of a model class can also yield more powerful model selection procedures than those obtained by the greedy procedure of Algorithm 1. We give one such illustration next in which a collection of model estimates that each exhibit zero false discovery (with high probability) can be 'combined' to derive a more complex model that also exhibits zero false discovery. Formally, a poset $(\mathcal{L}, \preceq)$ is said to possess a *join* if for each $x, y \in \mathcal{L}$ there exists a $z \in \mathcal{L}$ satisfying (i) $z \succeq x, z \succeq y$ and (ii) for any $w \in \mathcal{L}$ with $w \succeq x, w \succeq y$, we have $w \succeq z$; such a z is called the join of x, y and posets that possess a join are called *join semi-lattices* (these are dual to the notion of a meet defined in Section 2). Except for the posets in Examples 4, 6, and 9, the posets in the other examples are join semi-lattices (see Appendix A). For a model class that is a join semi-lattice, suppose we are provided estimates $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$ of a true model $x^\star$ such that $\mathrm{FD}(\hat{x}^{(j)}, x^\star) = 0,\ j = 1, \ldots, m$ (for example, by appealing to greedy methods such as Algorithm 1 or its variants). Appealing to the properties of a similarity valuation, we can conclude that the join $\hat{x}_{\mathrm{join}}$ of $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$ satisfies $\mathrm{FD}(\hat{x}_{\mathrm{join}}, x^\star) = 0$; in general, $\mathrm{rank}(\hat{x}_{\mathrm{join}})$ is larger than $\mathrm{rank}(\hat{x}^{(1)}), \ldots, \mathrm{rank}(\hat{x}^{(m)})$, and therefore, this procedure is one way to obtain a more powerful model by combining less powerful ones while still retaining control on the amount of false discovery. The following result formalizes matters.

Proposition 13 (using joins to obtain more powerful models)**.** *Let $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$ be a graded discrete model poset that is a join semi-lattice with integer-valued similarity valuation ρ and let $\mathcal{S}$ be an associated set of minimal covering pairs. Consider a collection of estimates $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$ of a true model $x^\star$ and let $\hat{x}_{\mathrm{join}}$ denote the join of $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$. Suppose for each $\hat{x}^{(j)},\ j = 1, \ldots, m$ there is a path from the least element of $\mathcal{L}$ to $\hat{x}^{(j)}$ such that every covering pair (u, v) along the path satisfies $\Psi_{\mathrm{test}}(u, v) \leq \alpha$. Then we have the false discovery bound $\mathbb{P}(\mathrm{FD}(\hat{x}_{\mathrm{join}}, x^\star) > 0) \leq \alpha |\mathcal{S}|$.*

4 Experiments

We describe the results of numerical experiments on synthetic and real data in this section. We employ Algorithm 1 with with both $\Psi = \Psi_{\mathrm{stable}}$ and $\Psi = \Psi_{\mathrm{test}}$. For the testing-based approach, the manner in which p-values are obtained is described in the context of each application and we set α equal to $0.05/|\mathcal{S}|$ for a given set $\mathcal{S}$ of minimal covering pairs. For the stability-based approach, we consider $B = 100$ subsamples obtained by partitioning a given dataset 50 times into subsamples of equal size and we set $\alpha = 0.3$.

To obtain a desired level of expected false discovery with the stability-based approach, we appeal to Theorem 11 as follows. In the bound (8), each q_k can be derived by averaging over subsamples (as explained in the discussion after the statement of Theorem 11) and all the other quantities are known. The values of these q_k's in turn depend on the model estimates returned by the base procedure $\hat{x}_{\mathrm{base}}$ employed on the subsamples; in particular, if the estimate is the least element then each q_k equals zero, and as $\hat{x}_{\mathrm{base}}$ returns models of increasing complexity, the value of each q_k generally increases. Building on this observation, we tune parameters in $\hat{x}_{\mathrm{base}}$ to return increasingly more complex models until the bound (8) is at the desired level. For causal structure learning we employ Greedy Equivalence Search as our base procedure with tuning via the regularization parameter that controls model complexity [3]. For clustering, we employ k-means [10] as the base procedure with tuning via the number of clusters. For our illustrations with ranking problems (both partial and total) in which we are provided with pairwise comparison data, our base procedure first employs the maximum-likelihood estimator associated to the Bradley-Terry model [2], which returns a vector of positive weights $\hat{w}$ of dimension equal to the number of items. Using this $\hat{w}$ we associate numerical values to covering pairs; each covering pair corresponds to increasing the complexity of a model by including a pair of items (i, j) to the inversion set (in total ranking) or to the relation specifying a strict partial order (in partial ranking), and the value we assign is the difference $\hat{w}_j - \hat{w}_i$. Our base procedure then constructs a path starting from the least element by greedily adding covering pairs of largest value at each step, provided these values are larger than a regularization parameter $\lambda > 0$; smaller values of λ yield model estimates of larger complexity, while larger values yield estimates of smaller complexity.

Finally, for causal structure learning, we restrict our search during the model aggregation phase of Algorithm 1 to paths that yield CPDAG models in which each connected component in the skeleton has diameter at most two; such a restriction facilitates a simple characterization of covering pairs. This restriction is not imposed on the output of the base procedure. Moreover, the true model can be an arbitrary CPDAG.

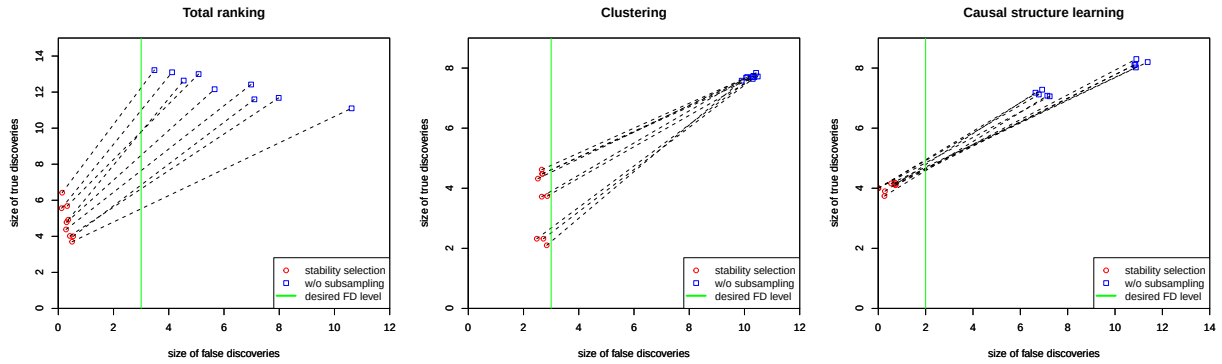


Figure 2: Comparing the performance of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ versus a non-subsampling approach for total ranking, clustering, and causal structure learning. Each problem setting corresponds to a pair of dots and a connecting line. The comparison is in terms of the amount of false and true discoveries.

4.1 Synthetic data

We describe experiments with synthetic data using Algorithm 1 with $\Psi = \Psi_{\text{stable}}$.

Total ranking: We consider a total ranking problem with $p = 30$ items. We observe n i.i.d. games between players i, j with the outcome modeled as $y_{ij\ell} \sim \texttt{Bernoulli}(w_i^\star/(w_i^\star + w_j^\star))$ for $\ell = 1, \ldots, n$, where $w^\star \in \mathbb{R}^p_{++}$ is a feature vector and $n \in \{200, 250, 300\}$. We fix $w^\star$ by first defining $\tilde{w} \in \mathbb{R}^p_{++}$ as $\tilde{w}_i = \tau^{i-1}$, $i = 1, \ldots, p$ for $\tau \in \{0.97, 0.98, 0.99\}$, and then setting $w^\star$ equal to a permutation of $\tilde{w}$ in which we swap the entries $1, 3$, the entries $8, 10$, the entries $15, 17$, the entries $20, 22$, and the entries $25, 27$. Smaller values of τ correspond to better-distinguished items, and hence to easier problem instances. The base procedure is tuned such that the expected false discovery in (8) is at most three.

Clustering: We consider a clustering problem with $p = 20$ variables. The true partition consists of 12 clusters with five variables in one cluster, another five variables in a second cluster, and the remaining variables in singleton clusters. The p variables are independent two-dimensional Gaussians. Each variable in cluster i has mean $(\mu_i, 0)$ and covariance $\frac{1}{4}I$; each $\mu_i = i/d$ for $d \in \{3, 3.5, 4\}$. Smaller values of d correspond to better-separated clusters, and hence to easier problem instances. We are provided n i.i.d. observations of these variables for $n \in \{40, 65, 90\}$. The base procedure is tuned such that the expected false discovery in (8) is at most three.

Causal structure learning: We consider a causal structure learning problem over $p = 10$ variables. The true DAG is generated by considering a random total ordering of the variables, drawing directed edges from higher nodes in the ordering to lower nodes independently with probability $v \in \{0.13, 0.18\}$, and defining a linear structural causal model in which each variable is a linear combination of its parents plus independent Gaussian noise with mean zero and variance $\frac{1}{4}$. The coefficients in the linear combination are drawn uniformly at random from the interval $[0.5, 0.7]$. Larger values of v lead to denser DAGs, and hence to harder problem instances. We obtain n i.i.d. observations from these models for $n \in \{1000, 1200, 1400, 1600, 1800\}$. The base procedure is tuned such that the expected false discovery in (8) is at most two.

For the preceding three problem classes, we compare the performance of our stability-based methodology versus that of a non-subsampled approach in which the base procedure (with suitable regularization) is applied to the entire dataset. For total ranking, the non-subsampled procedure simply extracts the ranking implied by the maximum-likelihood estimator associated to the Bradley-Terry model. For clustering, the non-subsampled approach employs k-means where the number of clusters is chosen to maximize the average silhoutte score [19]. For causal structure learning, the non-subsampled approach applies Greedy Equivalence Search with a regularization parameter chosen based on holdout validation (70% of the data is used for training and the remaining 30% for validation). Figure 2 presents the results of our experiments averaged over 50 trials, and as the plots demonstrate, our stability-based methods yield models with smaller false discovery than the corresponding non-subsampled approaches. This reduction in false discovery comes at the expense of a loss in power, which is especially significant for some of the harder problem settings. However, in all cases our stability-based method provides the desired level of control on expected false discovery.

4.2 Real data

We describe next experiments with real data.

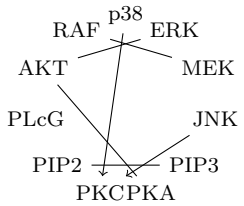
Partial ranking of tennis players: We consider the task of partially ranking six professional tennis players – Berdych, Djokovic, Federer, Murray, Nadal, and Wawrinka – based on historical head-to-head matches of these players up to the end of 2022. We apply Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ and with the base procedure tuned such that the expected false discovery in (8) is at most three. The output of our procedure is a rank-nine model given by the partial ranking {Djokovic, Nadal} > {Berdych, Murray, Wawrinka} and {Federer} > {Berdych, Wawrinka}.

Total ranking of educational systems: We consider the task of totally ordering $p = 15$ OECD countries in reading comprehension based on test results from the Programme for International Student Assessment (PISA). We take the null ranking as the ordering of the countries based on performance in 2015 (see the first row in Table 2), and we wish to update this model based on 2018 test scores (data obtained from [25]), with the number of test scores ranging from 696 to 3414. We apply Algorithm 1 with $\Psi = \Psi_{\text{test}}$ and we obtain p-values by modeling the average test score of each country as a Gaussian. We set $\alpha = 0.05/\frac{p(p-1)}{2}$ (here $\frac{p(p-1)}{2}$ is the cardinality of a set of minimal covering pairs), which yields the guarantee from Theorem 12 that the estimated model has zero false discovery with probability at least 0.95. The output of our procedure is the rank-nine model given by the total ranking in the second row in Table 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2015 base ranking	CAN	FIN	IRL	EST	KOR	JPN	NOR	NZL	DEU	POL	SvN	NLD	AUS	SWE	DNK
testing approach	FIN	IRL	EST	CAN	KOR	JPN	NOR	NZL	POL	DEU	AUS	SWE	SvN	DNK	NLD

Table 2: Ranking of nations according to PISA reading comprehension scores; the first column is the 2015 ranking of 15 OECD countries which serves as the base ranking for our analysis: based on test results in 2018, we update this ranking using Algorithm 1 based on $\Psi = \Psi_{\text{test}}$ with the result shown in the second column.

Learning causal structure among proteins: We aim to learn causal relations underlying $p = 11$ phosphoproteins and phospholipids from a mass spectroscopy dataset containing 854 measurements of abundance levels in an observational setting [20]. We apply Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ and with the base procedure tuned such that the expected false discovery in (8) is at most two. Figure 3 presents the rank-six CPDAG model obtained from our algorithm and compares to the estimates obtained from the literature [14, 20, 26]. Our CPDAG estimate has fewer edges than those in [14, 20, 26], which do not explicitly provide control on false discovery.



Edge	[20]a	[20]b	[15]	[5]	[14]a	[14]b	[26]
MEK – RAF	←	←	→	→		–	–
AKT – ERK		←	→	←	–	–	–
PIP3 – PIP2	→	→	→	←	→	→	→
AKT – PKA	←	←	←	←		←	→
P38 → PKC	←	←	←	←		–	
JNK → PKC	←	←	←	←	←	–	

Figure 3: **left**: CPDAG obtained by Algorithm 1 with $\Psi = \Psi_{\text{stable}}$; **right**: comparing the edges obtained by our algorithm (shown in the leftmost column) with different causal discovery methods (with indicated reference). The consensus network according to [20] is denoted here by "[20]a" and their reconstructed network by "[20]b"; The authors in [14] apply two methods, and the results are presented by "[14]a" and "[14]b". Here, "–" means that the edge direction is not identified.

5 Proofs

For notational ease, for a covering pair (u, v) and element z in the poset $\mathcal{L}$, we define $f(u, v; z) \triangleq \rho(v, z) - \rho(u, z)$. Recall that $\mathcal{T}_{\text{null}} \triangleq \{(u, v) \text{ covering pair in } \mathcal{L} \mid \rho(v, x^\star) = \rho(u, x^\star)\}$. Our analysis relies on the following lemmas with the proofs presented in Appendix C.

Lemma 14. *Fix a discrete model poset $\mathcal{L}$ with integer-valued similarity valuation ρ. For any model $x \in \mathcal{L}$ with $(x_0, \ldots, x_k)$ being any path from the least element $x_0 = x_{least}$ to $x_k = x$, we have that* $\text{FD}(x, x^\star) \leq$

$\sum_{i=1}^{k} \mathbb{I}[(x_{i-1}, x_i) \in \mathcal{T}_{null}]$. *As a result, we have that* $\mathrm{FD}(x, x^\star) > 0$ *implies the existence of some* i *for which* $(x_{i-1}, x_i) \in \mathcal{T}_{null}$.

Lemma 15. *For any covering pairs* (u, v) *and* (x, y) *with* $v \preceq x$*, we cannot have that* $f(u, v; z) = f(x, y; z)$ *for all* $z \in \mathcal{L}$.

5.1 Proof of Theorem 10

For notational convenience, we let $\hat{x}^{(\ell)}_{\text{base}} = \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ where $\{\mathcal{D}^{(\ell)}\}_{\ell=1}^{B}$ are the subsamples of $\mathcal{D}$. Let $\hat{x}_{\text{stable}}$ be the output of Algorithm 1 with $\mathrm{rank}(\hat{x}_{\text{stable}}) = \hat{k}$, and let $(x_0, \ldots, x_{\hat{k}})$ be the associated path from the least element $x_0 = x_{\text{least}}$ to $x_{\hat{k}} = \hat{x}_{\text{stable}}$; we have that $\frac{1}{B}\sum_{\ell=1}^{B} f(x_{i-1}, x_i; \hat{x}^{(\ell)}_{\text{base}})/c_{\mathcal{L}}(x_{i-1}, x_i) \geq (1-\alpha)$ for each $i = 1, \ldots, \hat{k}$. Let $\mathcal{C} \triangleq \{(x_{i-1}, x_i) \mid i = 1, \ldots, \hat{k}\}$. From Lemma 14, we also have that $\mathrm{FD}(\hat{x}_{\text{stable}}, x^\star) \leq |\mathcal{C} \cap \mathcal{T}_{\text{null}}|$. Combining these observations, we conclude that $\mathrm{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \sum_{(u,v)\in\mathcal{C}\cap\mathcal{T}_{\text{null}}} \mathbb{I}\left[\frac{1}{B}\sum_{\ell=1}^{B} \frac{f(u,v;\hat{x}^{(\ell)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 1-\alpha\right]$. Next, we observe that for each covering pair in $\mathcal{C}$ there exists a covering pair in the minimal set $\mathcal{S}$ with the values of f and $c_{\mathcal{L}}$ remaining the same; moreover, distinct covering pairs in $\mathcal{C}$ map to distinct covering pairs in $\mathcal{S}$ from Lemma 15. Thus, we conclude that $\mathrm{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \sum_{(u,v)\in\mathcal{S}\cap\mathcal{T}_{\text{null}}} \mathbb{I}\left[\frac{1}{B}\sum_{\ell=1}^{B} \frac{f(u,v;\hat{x}^{(\ell)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 1-\alpha\right]$. We then have the following sequence of steps:

$$\begin{aligned}
\mathrm{FD}(\hat{x}_{\text{stable}}, x^\star) &\leq \sum_{\substack{(u,v)\in \\ \mathcal{S}\cap\mathcal{T}_{\text{null}}}} \mathbb{I}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\sum_{i\in\{0,1\}} \frac{f(u,v;\hat{x}^{(2\ell-i)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 2-2\alpha\right] \\
&\leq \sum_{\substack{(u,v)\in \\ \mathcal{S}\cap\mathcal{T}_{\text{null}}}} \mathbb{I}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\prod_{i\in\{0,1\}} \frac{f(u,v;\hat{x}^{(2\ell-i)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 1-2\alpha\right].
\end{aligned} \tag{11}$$

The second inequality follows from $ab \geq a+b-1$ for $a, b \in [0,1]$, where we set $a = f(u,v;\hat{x}^{(2\ell-1)}_{\text{base}})/c_{\mathcal{L}}(u,v)$ and $b = f(u,v;\hat{x}^{(2\ell)}_{\text{base}})/c_{\mathcal{L}}(u,v)$, and note that $f(u,v;z)/c_{\mathcal{L}}(u,v) \in [0,1]$ for any $z \in \mathcal{L}$. Taking expectations on both sides of the preceding inequality, we finally seek a bound on $\mathbb{P}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\prod_{i\in\{0,1\}} \frac{f(u,v;\hat{x}^{(2\ell-i)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 1-2\alpha\right]$. We have that:

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\prod_{i\in\{0,1\}} \frac{f(u,v;\hat{x}^{(2\ell-i)}_{\text{base}})}{c_{\mathcal{L}}(u,v)} \geq 1-2\alpha\right] &\leq \frac{\mathbb{E}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\prod_{i\in\{0,1\}} \frac{f(u,v;\hat{x}^{(2\ell-i)}_{\text{base}})}{c_{\mathcal{L}}(u,v)}\right]}{1-2\alpha} \\
&= \frac{\mathbb{E}\left[f(u,v;\hat{x}_{\text{sub}})\right]^2}{c_{\mathcal{L}}(u,v)^2(1-2\alpha)}.
\end{aligned} \tag{12}$$

Here $\hat{x}_{\text{sub}}$ represents the estimator corresponding to the base procedure $\hat{x}_{\text{base}}$ applied to a subsample of $\mathcal{D}$ of size $|\mathcal{D}|/2$. The inequality follows from Markov's inequality, and the equality follows by noting that complementary bags are independent and identically distributed. Combining (11) and (12), we obtain the desired result.

5.2 Proof of Theorem 11

We have from Theorem 10 that:

$$\mathbb{E}[\mathrm{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \sum_{k=1}^{\mathrm{rank}(\mathcal{L})} \sum_{(u,v)\in\mathcal{S}_k\cap\mathcal{T}_{\text{null}}} \frac{\mathbb{E}[f(u,v;\hat{x}_{\text{sub}})]^2}{(1-2\alpha)c_{\mathcal{L}}(u,v)^2}.$$

Our goal is to bound $\mathbb{E}[f(u,v;\hat{x}_{\text{sub}})]/c_{\mathcal{L}}(u,v)$ for $(u,v) \in \mathcal{S}_k \cap \mathcal{T}_{\text{null}}$. Note that each q_k may be decomposed as

$$q_k = \sum_{\substack{(u,v)\in \\ \mathcal{S}_k\cap\mathcal{T}_{\text{null}}}} \frac{\mathbb{E}[f(u,v;\hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u,v)} + \sum_{\substack{(u,v)\in \\ \mathcal{S}_k\setminus\mathcal{T}_{\text{null}}}} \frac{\mathbb{E}[f(u,v;\hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u,v)}.$$

Appealing to Assumption 1, we have that

$$q_k \geq \left(1 + \frac{|\mathcal{S}_k \setminus \mathcal{T}_{\text{null}}|}{|\mathcal{S}_k \cap \mathcal{T}_{\text{null}}|}\right) \sum_{(u,v) \in \mathcal{S}_k \cap \mathcal{T}_{\text{null}}} \frac{\mathbb{E}[f(u, v; \hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u, v)}.$$

Rearranging the terms, we obtain that

$$\sum_{(u,v) \in \mathcal{S}_k \cap \mathcal{T}_{\text{null}}} \frac{\mathbb{E}[f(u, v; \hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u, v)} \leq \frac{q_k}{|\mathcal{S}_k|} |\mathcal{S}_k \cap \mathcal{T}_{\text{null}}|.$$

Appealing to Assumption 2, we have for each $(u, v) \in \mathcal{S}_k \cap \mathcal{T}_{\text{null}}$ that $\frac{\mathbb{E}[f(u,v;\hat{x}_{\text{sub}})]}{c_{\mathcal{L}}(u,v)} \leq \frac{q_k}{|\mathcal{S}_k|}$. Plugging this bound into the conclusion of Theorem 10 yields the desired result.

5.3 Proof of Theorem 12

Let $\hat{x}_{\text{test}}$ be the output of Algorithm 1 with $\text{rank}(\hat{x}_{\text{test}}) = \hat{k}$, and let $(x_0, \ldots, x_{\hat{k}})$ be the associated path from the least element $x_0 = x_{\text{least}}$ to $x_{\hat{k}} = \hat{x}_{\text{test}}$; we have that $\Psi_{\text{test}}(x_{i-1}, x_i) \leq \alpha$ for each $i = 1, \ldots, \hat{k}$. Let $\mathcal{C} \triangleq \{(x_{i-1}, x_i) \mid i = 1, \ldots, \hat{k}\}$. From Lemma 14, we have that $\text{FD}(\hat{x}_{\text{test}}, x^\star) > 0$ implies the existence of a covering pair $(u, v) \in \mathcal{C} \cap \mathcal{T}_{\text{null}}$ for which $\Psi_{\text{test}}(u, v) \leq \alpha$. For each covering pair in $\mathcal{C}$, there exists a covering pair in $\mathcal{S}$ with the same value of Ψ_{test}; thus, there exists $(u, v) \in \mathcal{S} \cap \mathcal{T}_{\text{null}}$ such that $\Psi_{\text{test}}(u, v) \leq \alpha$. Consequently:

$$\begin{aligned} \mathbb{P}(\text{FD}(\hat{x}_{\text{test}}, x^\star) > 0) &\leq \mathbb{P}\left(\exists (u, v) \in \mathcal{S} \cap \mathcal{T}_{\text{null}} \text{ s.t. } \Psi_{\text{test}}(u, v) \leq \alpha\right) \\ &\leq \sum_{(u,v) \in \mathcal{S} \cap \mathcal{T}_{\text{null}}} \mathbb{P}(\Psi_{\text{test}}(u, v) \leq \alpha) \leq \alpha |\mathcal{S}|. \end{aligned} \tag{13}$$

Here the second inequality follows from the union bound and the final inequality follows from the fact that the random variable $\Psi_{\text{test}}(u, v)$ is a valid p-value under the null hypothesis $\rho(v, x^\star) = \rho(u, x^\star)$.

5.4 Proof of Proposition 13

For each $\hat{x}^{(j)},\ j = 1, \ldots, m$, we are given that there is a path from x_{least} to $\hat{x}^{(j)}$ such that Ψ_{test} is bounded by α for each covering pair in the path; let $\mathcal{C}^{(j)}$ be the set of these covering pairs. As described in Section 3.3 in the discussion preceding Proposition 13, $\text{FD}(\hat{x}_{\text{join}}, x^\star) > 0$ implies that $\text{FD}(\hat{x}^{(j)}, x^\star) > 0$ for some $j = 1, \ldots, m$, which in turn implies from Lemma 14 the existence of a covering pair $(u, v) \in \mathcal{C}^{(j)} \cap \mathcal{T}_{\text{null}}$ for some $j = 1, \ldots, m$. Following the same logic as in the proof of Theorem 12, we conclude that $\text{FD}(\hat{x}_{\text{join}}, x^\star) > 0$ implies the existence of $(u, v) \in \mathcal{S} \cap \mathcal{T}_{\text{null}}$ such that $\Psi_{\text{test}}(u, v) \leq \alpha$. Using the same reasoning as in (13), we have the desired conclusion.

6 Discussion

We present a general framework to endow a collection of models with poset structure. This framework yields a systematic approach for quantifying model complexity and false positive error in an array of complex model selection tasks in which models are not characterized by Boolean logical structure (such as in variable selection). Moreover, we develop methodology for controlling false positive error in general model selection problems over posets, and we describe experimental results that demonstrate the utility of our framework.

We finally discuss some future research questions that arise from our work. On the mathematical front, a basic open question is to characterize fundamental tradeoffs between false positive and false negative errors that are achievable by any procedure in model selection over a general poset; this would generalize the Neyman-Pearson lemma on optimal procedures for testing between two hypotheses. On the computational and methodological front, it is of interest to develop new methods to control false positive error as well as false discovery rates, including in settings involving continuous model posets.

Acknowledgements

We thank Marina Meila and Lior Pachter for insightful conversations. PB received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 786461). VC was supported in part by Air Force Office of Scientific Research grant FA9550-20-1-0320 and by National Science Foundation grant DMS 2113724.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952.

[3] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[4] M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393, 2017.

[5] D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. *In Proceedings of the 10th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[6] R. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, 1932.

[7] D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B*, 70, 2008.

[8] J. J. Goeman and U. R. Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24 4:537–44, 2008.

[9] H. Hotelling. The generalization of student's ratio. *Annals of Mathematical Statistics*, 2:54–65, 1931.

[10] S. P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28:129–136, 1982.

[11] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42:1897–1908, 1971.

[12] N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95:265–278, 2008.

[13] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72, 2010.

[14] N. Meinshausen, A. Hauser, J. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceeding of National Academy of Sciences*, 113:7361–7368, 2016.

[15] J. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. *In Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence*, pages 431–439, 2013.

[16] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika*, pages 175–240, 1928.

[17] K. Popper. *Logik der forschung: Zur erkenntnistheorie der modernen naturwissenschaft*. Springer, 1935.

[18] P. R. Rosenbaum. Testing hypotheses in order. *Biometrika*, 95:248–252, 2008.

[19] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[20] K. Sachs, O. Perez, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

[21] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

[22] R. Shah and R. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B*, 75, 2013.

[23] R. P. Stanley. *Enumerative Combinatorics*, volume 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2nd edition, 2011.

[24] A. Taeb, P. Shah, and V. Chandrasekaran. False discovery and its control in low-rank estimation. *Journal of the Royal Statistical Society: Series B*, 92:997–1027, 2020.

[25] K. Wang. Learning tower, 2021. https://github.com/kevinwang09/learningtower.

[26] K. D. Yang, A. Katoff, and C. Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, 2018.

[27] D. Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103:309–316, 2008.

Appendix

A Meet semi-lattice and join semi-lattice properties and posets in Examples 1-9

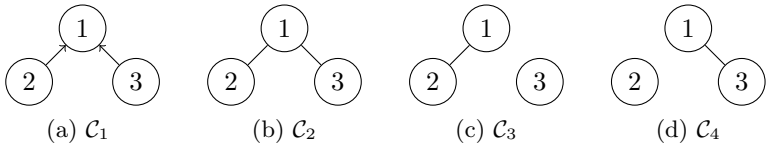
The Boolean poset (Example 1), partition poset (Examples 2-3), integer poset (Example 5), permutation poset (Example 7), and subspace poset (Example 8) are all known in the literature to be lattices (and consequently meet-semi and join semi-lattices); see [23].

We next show that for Examples 6 and 9 associated with partial ranking and blind-source separation, the corresponding posets are also meet semi-lattices. Consider the partial ranking setting in Example 6. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two relations that are irreflexive, asymmetric, and transitive. Recalling that the partial ordering is based on inclusion, it is clear that the relations $\mathcal{R} = \{(a,b) : (a,b) \in \mathcal{R}_1, (a,b) \in \mathcal{R}_2\}$ is the unique largest rank element in the partial ranking poset such that $\mathcal{R} \preceq \mathcal{R}_1$ and $\mathcal{R} \preceq \mathcal{R}_2$. Furthermore, for any $\tilde{\mathcal{R}}$ with $\tilde{\mathcal{R}} \preceq \mathcal{R}_1$ and $\tilde{\mathcal{R}} \preceq \mathcal{R}_2$, we clearly have that $\tilde{\mathcal{R}} \preceq \mathcal{R}$. Consider the blind-source separation setting in Example 9. Let x_1 and x_2 be two set of linearly independent subsets of unit norm vectors. Recalling that the partial ordering in the associated poset is based on inclusion, it is clear that the set $y = x_1 \cap x_2$ is the unique largest rank element in the partial ranking poset such that $y \preceq x_1$ and $y \preceq x_2$. Furthermore, for every z with $z \preceq x_1$ and $z \preceq x_2$, we have that $z \preceq y$.

We show that the poset corresponding to causal structure learning setting (Example 4) is not meet semi-lattice or join semi-lattice. As a counterexample, consider the CPDAGs $\mathcal{C}_i$ for $i = 1, 2, 3, 4$ shown in Figure 4. Notice that $\mathcal{C}_3 \preceq \mathcal{C}_1$, $\mathcal{C}_3 \preceq \mathcal{C}_2$, $\mathcal{C}_4 \preceq \mathcal{C}_1$, and $\mathcal{C}_4 \preceq \mathcal{C}_2$. Notice also that $\mathcal{C}_3$ and $\mathcal{C}_4$ are both CPDAGs with the largest rank that are smaller (in a partial order sense) than $\mathcal{C}_1$ and $\mathcal{C}_2$. We thus can conclude that the poset is not meet semi-lattice. Similarly, $\mathcal{C}_1$ and $\mathcal{C}_2$ are both CPDAGs with the smallest rank that are larger (in a partial order sense) than $\mathcal{C}_3$ and $\mathcal{C}_4$. We thus can conclude that the poset is not join semi-lattice.

We next show that the poset for Example 6 is not join semi-lattice with a simple counterexample. Consider as an example elements $x_1 = \{(1,2)\}$ and $x_2 = \{(2,1)\}$. Note that there does not exist an element z such that $x_1 \preceq z$ and $x_2 \preceq z$. Thus, the poset is not join semi-lattice.

Finally, we show that the poset corresponding to blind-source separation (Example 9) is not join semi-lattice. Consider a collection of $p+1$ rank-1 elements in this poset, each element consisting of a single p dimensional vector. Then, evidently, there cannot exist an element z consisting of a set of vectors that contains all of the vectors in the rank-1 elements, while satisfying the linear independence condition.



(a) $\mathcal{C}_1$ (b) $\mathcal{C}_2$ (c) $\mathcal{C}_3$ (d) $\mathcal{C}_4$

Figure 4: Four CPDAGs. Here, CPDAGs $\mathcal{C}_3$ and $\mathcal{C}_4$ are both largest complexity models that are smaller (in partial order sense) than $\mathcal{C}_1$ and $\mathcal{C}_2$. Similarly, CPDAGs $\mathcal{C}_1$ and $\mathcal{C}_2$ are the smallest complexity models that are larger (in a partial order sense) than $\mathcal{C}_3$ and $\mathcal{C}_4$.

B Proof that (1) is a similarity valuation function

Recall that

$$\rho_{\text{meet}}(x,y) = \max_{z \preceq x, z \preceq y} \text{rank}(z). \tag{14}$$

By definition, $\rho_{\text{meet}}(\cdot,\cdot)$ is a symmetric function. We will now show that it satisfies the three properties in Definition 1 for any pair of elements $x, y \in \mathcal{L}$. For the first property, we can conclude $\rho_{\text{meet}}(x,y) \geq 0$ since by definition, the rank function returns a non-negative integer for all the elements in the poset. Again, because of the property of the rank function in a graded poset, a feasible z (satisfying the constraints $z \preceq x$, $z \preceq y$) will

necessarily have $\text{rank}(z) \leq \min\{\text{rank}(x), \text{rank}(y)\}$. For the second property, consider any $w \in \mathcal{L}$ with $x \preceq w$. Note that:

$$\rho_{\text{meet}}(w, y) = \max_{z \preceq w, z \preceq y} \text{rank}(z). \tag{15}$$

Then, any feasible z in (14) is also feasible in (15) by the transitive property of posets. Therefore, $\rho_{\text{meet}}(x, y) \leq \rho_{\text{meet}}(w, y)$. For the third property, first note that if $x \preceq y$, then $z = x$ is feasible in (14) and thus $\rho_{\text{meet}}(x, y) \geq \text{rank}(x)$. Since also $\rho_{\text{meet}}(x, y) \leq \text{rank}(x)$ by the second property of similarity valuations, we have that $\rho_{\text{meet}}(x, y) = \text{rank}(x)$. Now suppose that $\rho_{\text{meet}}(x, y) = \text{rank}(x)$. By (14), we conclude that there exists a feasible z ($z \preceq x$, $z \preceq y$) such that $\text{rank}(z) = \text{rank}(x)$. By the property of the rank function, we have that if $\text{rank}(z) = \text{rank}(x)$ and $z \preceq x$, then $z = x$. Since we have additionally that $z \preceq y$, we conclude that $x \preceq y$.

C Proof of Lemmas 14-15

Proof of Lemma 14. Recall the telescoping sum decomposition (5) that $\text{FD}(x_k, x^\star) = \sum_{i=1}^{k} 1 - [f(x_{i-1}, x_i; x^\star)]$. From the first property of similarity valuation that it yields non-negative values, second property of similarity valuation that $\rho(x, y) \leq \rho(z, y)$ for $x \preceq z$, and that the ρ is an integer-valued similarity valuation, we have that $\text{FD}(x, x^\star) \leq \sum_{i=1}^{k} \mathbb{I}[(x_{i-1}, x_i) \in \mathcal{T}_{\text{null}}]$. □

Proof of Lemma 15. For any covering pairs (x, y) and (u, v) with $v \preceq x$, we cannot have that $f(x, y; z) = f(u, v; z)$ for all $z \in \mathcal{L}$. Suppose as a point of contradiction that for every $z \in \mathcal{L}$, $f(x, y; z) = f(u, v; z)$. Let $z = v$. Then, by the third property of a similarity valuation (see Definition 1), $\rho(u, z) = \text{rank}(u)$ and $\rho(v, z) = \text{rank}(v)$; thus, for this choice of z, $f(u, v; z) = 1$. On the other hand, again by the third property of a similarity valuation and for the choice of $z = v$, since $u \preceq v \preceq x \preceq y$, $\rho(x, z) = \rho(y, z) = \text{rank}(v)$ and thus $f(x, y; z) = 0$. □

D Analysis in the the continuous Examples 8 and 9

For notational ease, we let $\hat{x}_{\text{base}}^{(\ell)} = \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$. Notice that for any $l = 1, 2, \ldots, B$:

$$\begin{aligned}
\text{FD}(\hat{x}_{\text{stable}}, x^\star) &= \text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, x^\star) \\
&= \left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})\right] + \left[\text{rank}(\hat{x}_{\text{base}}^{(\ell)}) - \rho(\hat{x}_{\text{base}}^{(\ell)}, x^\star)\right] + \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}),
\end{aligned}$$

where

$$\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) := \rho(\hat{x}_{\text{base}}^{(\ell)}, x^\star) - \text{rank}(\hat{x}_{\text{base}}^{(\ell)}) + \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) - \rho(\hat{x}_{\text{stable}}, x^\star).$$

Since the choice of l was arbitrary, we note that:

$$\begin{aligned}
\text{FD}(\hat{x}_{\text{stable}}, x^\star) &= \frac{2}{B} \sum_{\ell=1}^{B/2} \min_{t \in \{0,1\}} \Bigg\{ \left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(2\ell-t)})\right] + \left[\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)\right] \\
&\quad + \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(2\ell-t)}) \Bigg\} \\
&\leq \frac{2}{B} \sum_{\ell=1}^{B/2} \min_{t \in \{0,1\}} \left\{ \left[\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)\right] \right\} + \frac{2}{B} \sum_{\ell=1}^{B} \left[\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})\right] \\
&\quad + \frac{2}{B} \sum_{\ell=1}^{B} \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) \\
&\leq \frac{2}{B} \sum_{\ell=1}^{B/2} \prod_{t \in \{0,1\}} \sqrt{\text{rank}(\hat{x}_{\text{base}}^{(2\ell-t)}) - \rho(\hat{x}_{\text{base}}^{(2\ell-t)}, x^\star)} + 2\alpha \text{rank}(\hat{x}_{\text{stable}}) + \frac{2}{B} \sum_{\ell=1}^{B} \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}).
\end{aligned}$$

Here, the second inequality follows from $\min\{a+b, c+d\} \leq \min\{a,c\} + b + d$ for $a, b, c, d \geq 0$. The third inequality follows from $\min\{a,b\} \leq \sqrt{ab}$ for $a, b \geq 0$ and

$$\frac{1}{B}\sum_{\ell=1}^{B} \text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) = \sum_{k=1}^{\text{rank}(\hat{x}_{\text{stable}})} \frac{1}{B}\sum_{\ell=1}^{B} 1 - [\rho(x_k, \hat{x}_{\text{base}}^{(\ell)}) - \rho(x_{k-1}, \hat{x}_{\text{base}}^{(\ell)})] \leq \alpha \text{rank}(\hat{x}_{\text{stable}}), \tag{16}$$

where $(x_0, x_1, \ldots, x_{\hat{k}})$ is a sequence specifying a path from the least element x_0 to $x_{\hat{k}} = \hat{x}_{\text{stable}}$ with $\text{rank}(\hat{x}_{\text{stable}}) = \hat{k}$. Thus, $\frac{1}{B}\sum_{\ell=1}^{B} \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) \geq (1-\alpha)\text{rank}(\hat{x}_{\text{stable}})$. As $\rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)}) \leq \text{rank}(\hat{x}_{\text{base}}^{(\ell)})$, we can then conclude that $\mathbb{E}[\text{rank}(\hat{x}_{\text{stable}})] \leq \frac{\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})]}{1-\alpha}$. Taking expectations and using the fact that the data across complementary bags is IID, we obtain:

$$\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2 + \frac{2\alpha}{1-\alpha}\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})] + \frac{2}{B}\sum_{\ell=1}^{B} \mathbb{E}[\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)})].$$

It remains to bound $\frac{2}{B}\sum_{\ell=1}^{B} \mathbb{E}[\kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)})]$ for subspace selection and blind-source separation.

Subspace-selection: We use the similarity valuation $\rho = \rho_{\text{subspace}}$ in (3). Note that:

$$\begin{aligned}
\text{rank}(x) - \rho(x,y) = \text{trace}\left(\mathcal{P}_x \mathcal{P}_{y^\perp}\right) &= \text{trace}\left(\mathcal{P}_x \mathcal{P}_z \mathcal{P}_{y^\perp} \mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x \mathcal{P}_{z^\perp} \mathcal{P}_{y^\perp} \mathcal{P}_{z^\perp}\right) \\
&\quad + \text{trace}\left(\mathcal{P}_x \mathcal{P}_{z^\perp} \mathcal{P}_{y^\perp} \mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x \mathcal{P}_z \mathcal{P}_{y^\perp} \mathcal{P}_{z^\perp}\right) \\
&\leq \text{trace}\left(\mathcal{P}_{y^\perp} \mathcal{P}_z\right) + \text{trace}\left(\mathcal{P}_x \mathcal{P}_{z^\perp}\right) + \text{trace}\left(\left[\mathcal{P}_x, \mathcal{P}_{z^\perp}\right]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right) \\
&= \text{rank}(z) - \rho(y,z) + \text{rank}(x) - \rho(x,z) + \text{trace}\left(\left[\mathcal{P}_x, \mathcal{P}_{z^\perp}\right]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right).
\end{aligned} \tag{17}$$

Here, for matrices $A, B \in \mathbb{R}^{p \times p}$, $[A, B] = AB - BA$ represents the commutator. Furthermore, note that:

$$\begin{aligned}
\text{trace}\left(\left[\mathcal{P}_x, \mathcal{P}_{z^\perp}\right]\left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right]\right) &\leq \| \left[\mathcal{P}_x, \mathcal{P}_{z^\perp}\right] \|_\star \| \left[\mathcal{P}_z, \mathcal{P}_{y^\perp}\right] \|_2 \\
&\leq 2\sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x,z)} \| \left[\mathcal{P}_z, \mathcal{P}_y\right] \|_2.
\end{aligned} \tag{18}$$

Combining the bounds (17) and (18), we find that:

$$\begin{aligned}
\text{rank}(x) - \rho(x,y) &\leq \text{rank}(z) - \rho(y,z) + \text{rank}(x) - \rho(x,z) + 2\sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x,z)} \| \left[\mathcal{P}_z, \mathcal{P}_y\right] \|_2 \\
&\leq \text{rank}(z) - \rho(y,z) + \text{rank}(x) - \rho(x,z) + \sqrt{\text{rank}(x)}\sqrt{\text{rank}(x) - \rho(x,z)}.
\end{aligned}$$

Here, the second inequality follows from the fact that for projection matrices A and B, $\|[A,B]\|_2 \leq 1/2$. From this inequality, we conclude that in the subspace selection setting,

$$\begin{aligned}
\frac{1}{B}\sum_{\ell=1}^{B} \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}_{\text{base}}^{(\ell)}) &\leq \sqrt{\text{rank}(\hat{x}_{\text{stable}})}\frac{1}{B}\sum_{l=1}^{B}\sqrt{\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})} \\
&\leq \sqrt{\text{rank}(\hat{x}_{\text{stable}})}\sqrt{\frac{1}{B}\sum_{l=1}^{B}\text{rank}(\hat{x}_{\text{stable}}) - \rho(\hat{x}_{\text{stable}}, \hat{x}_{\text{base}}^{(\ell)})} \\
&\leq \sqrt{\alpha}\text{rank}(\hat{x}_{\text{stable}}).
\end{aligned}$$

Here, the second equality follows from cauchy-shwartz and last inequality follows from the bound (16). Recalling that $\mathbb{E}[\text{rank}(\hat{x}_{\text{stable}})] \leq \frac{\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})]}{1-\alpha}$, we obtain the final bound:

$$\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \mathbb{E}[\sqrt{\text{FD}(\hat{x}_{\text{sub}}, x^\star)}]^2 + \frac{2\alpha + \sqrt{\alpha}}{1-\alpha}\mathbb{E}[\text{rank}(\hat{x}_{\text{sub}})].$$

Blind-source separation We use the similarity valuation $\rho = \rho_{\text{source-separation}}$ in (4). For simplicity of notation, associated with any element $z \in \mathcal{L}$, we consider a block-diagonal $p^2 \times p^2$ projection matrix where each

$p \times p$ block is a projection matrix of the subspace spanned by a vector in z. We denote this projection matrix $\mathcal{P}_z$. Then, $\rho(x,y) = \max_{\Pi \in \mathbb{S}^{p^2}_{\text{block}}} \text{trace}\left(\mathcal{P}_x \Pi \mathcal{P}_y \Pi^T\right)$ where $\mathbb{S}^{p^2}_{\text{block}}$ is the space of $p^2 \times p^2$ permutation matrices that are block-diagonal where each block is of size $p \times p$.

Note that:

$$\begin{aligned}
\text{rank}(x) - \rho(x,y) &= \min_{\Pi \in \mathbb{S}^{p^2}_{\text{block}}} \text{trace}\left(\mathcal{P}_x \Pi \mathcal{P}_{y^\perp} \Pi^T\right) \\
&\leq \min_{\tilde{\Pi} \in \mathbb{S}^{p^2}_{\text{block}}} \min_{\Pi \in \mathbb{S}^{p^2}_{\text{block}}} \text{trace}\left(\Pi \mathcal{P}_{y^\perp} \Pi^T \tilde{\Pi} \mathcal{P}_z \tilde{\Pi}^T\right) + \text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right) \\
&\qquad + 2\sqrt{\text{rank}(x)} \sqrt{\text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right)} \|[\tilde{\Pi} \mathcal{P}_z \tilde{\Pi}^T, \Pi \mathcal{P}_y \Pi^T]\|_2 \\
&\leq \min_{\tilde{\Pi} \in \mathbb{S}^{p^2}_{\text{block}}} \text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right) + 2\sqrt{\text{rank}(x)} \sqrt{\text{trace}\left(\mathcal{P}_x \tilde{\Pi} \mathcal{P}_{z^\perp} \tilde{\Pi}^T\right)} \\
&\qquad \max_{\bar{\Pi}, \bar{\tilde{\Pi}} \in \mathbb{S}^{p^2}_{\text{block}}} \|[\bar{\tilde{\Pi}} \mathcal{P}_z \bar{\tilde{\Pi}}^T, \bar{\Pi} \mathcal{P}_y \bar{\Pi}^T]\|_2 + \max_{\tilde{\Pi} \in \mathbb{S}^{p}_{\text{block}}} \min_{\Pi \in \mathbb{S}^{p}_{\text{block}}} \text{trace}\left(\Pi(\text{Id} - \mathcal{P}_y) \Pi^T \tilde{\Pi} \mathcal{P}_z \tilde{\Pi}^T\right) \\
&= [\text{rank}(x) - \rho(x,z] + [\text{rank}(z) - \rho(z,y)] \\
&\quad + 2\sqrt{\text{rank}(x)} \sqrt{\text{rank}(x) - \rho(x,z)} \max_{\bar{\Pi}, \bar{\tilde{\Pi}} \in \mathbb{S}^{p^2}_{\text{block}}} \|[\bar{\tilde{\Pi}} \mathcal{P}_z \bar{\tilde{\Pi}}^T, \bar{\Pi} \mathcal{P}_y \bar{\Pi}^T]\|_2.
\end{aligned}$$

Here, the first inequality follows from a similar analysis as arriving to (17) in subspace selection. The second inequality follows from the fact that $\min_{a,b} f(a) + g(b) \leq \min_a f(a) + \max_b f(b)$. Note that projection matrices A, B, $[A,B] \leq \frac{1}{2}$. Then, following the same exact reasoning as the subspace case, we have that in the blind-source separation setting $\frac{1}{B}\sum_{\ell=1}^{B} \kappa(\hat{x}_{\text{stable}}, x^\star, \hat{x}^{(\ell)}_{\text{base}}) \leq \sqrt{\alpha}\text{rank}(\hat{x}_{\text{sub}})$. The result follows subsequently.

E Specializing bound (8) for different problem settings

E.1 Partial ranking

Let $S = \{a_1, a_2, \ldots, a_p\}$ be the set of p elements. We use the similarity valuation $\rho := \rho_{\text{meet}}$ in (1) of the main paper.

E.1.1 Characterizing $\mathcal{S}$ for partial ranking

We construct a set $\mathcal{S}$ satisfying the properties in Definition 3 of the main paper. Specifically, we let:

$$\mathcal{S} = \{(a_i, a_j) : i \neq j\},$$

with $|\mathcal{S}_1| = p(p-1)$ and $\mathcal{S}_k = \emptyset$ for every $k \geq 2$.

We will show that set $\mathcal{S}$ as constructed above satisfies Definition 3. First, consider any covering pair $(u', v') \notin \mathcal{S}$. Here, u' and v' are relations and $v' = u' \cup (a_i, a_j)$ for some $i \neq j$. Then, for any $z \in \mathcal{L}$, it is easy to see that

$$\rho(v', z) - \rho(u', z) = \mathbb{I}[(a_i, a_j) \in z] = \rho(v, z) - \rho(u, z),$$

where $v = \{(a_i, a_j)\}$ and $u = \emptyset$. Clearly, $\text{rank}(v) \leq \text{rank}(v')$.

To show the second property, consider covering pairs $(\{(a_i, a_j)\}, \emptyset) \in \mathcal{S}$ and $(\{(a_k, a_l)\}, \emptyset) \in \mathcal{S}$. By construction of the set $\mathcal{S}$, $(a_i, a_j) \neq (a_k, a_l)$. Let $z = \{(a_i, a_j)\}$. Then, it is straightforward to see that $\rho(\{(a_i, a_j)\}, z) - \rho(\emptyset, z) = 1$ but $\rho(\{(a_k, a_l)\}, z) - \rho(\emptyset, z) = 0$.

E.1.2 Characterizing $c_{\mathcal{L}}(x,y)$ for covering pair (x,y)

Since for any z, $\rho(y,z) - \rho(x,z) = \mathbb{I}((a_i, a_j) \in z)$ for some (a_i, a_j). Thus, $c_{\mathcal{L}}(x,y) = 1$.

E.1.3 Refined false discovery bound for partial ranking

Let $\hat{x}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \frac{q_1^2}{(1-2\alpha)p(p-1)},$$

where

$$q_1 = \sum_{i \neq j} \mathbb{I}[(a_i, a_j) \in \hat{x}_{\text{sub}}].$$

Here, $\hat{x}_{\text{sub}}$ is the estimated partial ranking from supplying $n/2$ samples to base estimator. We can use the following data-driven approximation for q_1:$q_1 \approx \frac{1}{B}\sum_{\ell=1}^{B}\sum_{i\neq j}\mathbb{I}[(a_i,a_j) \in \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})]$ with $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}), l = 1, 2, \ldots, B$ representing the estimates from subsampling.

E.2 Total ranking

Let $S = \{a_1, a_2, \ldots, a_p\}$ be the set of p elements. Let $\pi_{\text{null}}(a_i) = i$ for every $i = 1, 2, \ldots, p$. We use the similarity valuation $\rho := \rho_{\text{total-ranking}}$ in (2) of the main paper. As each element in the poset corresponds to a function $\pi : S \to S$, we will use this functional notation throughout.

E.2.1 Characterizing $\mathcal{S}$ for total ranking

We construct a set $\mathcal{S}$ satisfying the properties in Definition 3 of the main paper. Initialize $\mathcal{S} = \emptyset$. Then, for every relation (a_i, a_j) with $i < j$, we augment $\mathcal{S}$ as follows:

$$\mathcal{S} = \mathcal{S} \cup (\pi_1, \pi_2),$$

where π_1, π_2 are covering pairs. Here, π_2 is any rank $j - i$ element in the poset with the relation (a_i, a_j) in its corresponding inversion set. Furthermore, we let π_1 be a rank $j - i - 1$ element that is covered by π_2 and does not contain (a_i, a_j) in its inversion set. Recalling that $\mathcal{S}_k = \{(\pi_1, \pi_2) \in \mathcal{S}, \text{rank}(\pi_2) = k\}$, we have that for every $k = 1, 2, \ldots, p-1$

$$|\mathcal{S}_k| = p - k.$$

We will show that set $\mathcal{S}$ as constructed above satisfies Definition 3. First, consider any covering pair $(\tilde{\pi}_1, \tilde{\pi}_2) \notin \mathcal{S}$. Then by definition, the corresponding inversion sets being nested, i.e. $\text{inv}(\tilde{\pi}_2; \pi_{\text{null}}) \supseteq \text{inv}(\tilde{\pi}_1; \pi_{\text{null}})$ with the difference being a single relation. We will denote this relation by (a_i, a_j) with $j > i$. Consider the covering pair $(\pi_1, \pi_2) \in \mathcal{S}$ where (a_i, a_j) is in the inversion set of π_2 but not in the inversion set of π_1. Then, for any π, we have that

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})) = \rho(\tilde{\pi}_2, \pi) - \rho(\tilde{\pi}_1, \pi).$$

Furthermore, it is straightforward to check that $\text{rank}(\tilde{\pi}_2) \geq j - i = \text{rank}(\pi_2)$. We have thus shown that $\mathcal{S}$ satisfies the first property in Definition 3.

To show the second property, consider covering pairs $(\pi_1, \pi_2) \in \mathcal{S}$ where the difference between the two inversion sets is the relation (a_i, a_j). Let $(\pi_3, \pi_4) \in \mathcal{S}$ where the difference between the two inversion sets is the relation (a_k, a_l). By construction of the set $\mathcal{S}$, $(a_i, a_j) \neq (a_k, a_l)$. Let π be a permutation with (a_i, a_j) in its inversion set. Then, as desired,

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})) \neq \rho(\pi_4, \pi) - \rho(\pi_3, \pi).$$

E.2.2 Characterizing $c_{\mathcal{L}}(\pi_1, \pi_2)$ for covering pair (π_1, π_2)

Since for any π, $\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}((a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}}))$ for some pair of elements (a_i, a_j), then $c_{\mathcal{L}}(\pi_1, \pi_2) = 1$.

E.2.3 Refined false discovery bound for total ranking

Let $\hat{\pi}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{\pi}_{\text{stable}}, \pi^\star)] \leq \sum_{k=1}^{p-1} \frac{q_k^2}{(1-2\alpha)(p-k)},$$

where

$$q_k = \sum_{(\pi_1,\pi_2)\in\mathcal{S}_k} \mathbb{E}[\rho(\pi_2, \hat{\pi}_{\text{sub}}) - \rho(\pi_1, \hat{\pi}_{\text{sub}})] = \sum_{(i,j),j-i=k} \left[\mathbb{I}[(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{sub}}; \pi_{\text{null}})]\right].$$

Here, $\hat{\pi}_{\text{sub}}$ represents ranking from supplying $n/2$ samples to base estimator. We can use the following data-driven approximation for q_k:$q_k \approx \frac{1}{B}\sum_{(i,j),j-i=k}\sum_{\ell=1}^{B}\left[\mathbb{I}[(a_i, a_j) \in \text{inv}(\hat{\pi}_{\text{base}}(\mathcal{D}^{(\ell)}); \pi_{\text{null}})]\right]$, where $\hat{\pi}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the total ranking obtained by supplying the base estimator on dataset $\mathcal{D}^{(\ell)}$.

E.3 Clustering

We have a collection of p items $\{a_1, a_2, \ldots, a_p\}$ that we wish to cluster. We let $x_0 = \{\{a_1\}, \{a_2\}, \ldots, \{a_p\}\}$ be the least element. As described in the main paper, will use the similarity valuation $\rho := \rho_{\text{meet}}$ defined in (1) of the main paper. Since the clustering poset is meet semi-lattice, ρ computes the rank of the meet of two elements; in this setting, the meet $x \wedge z$ of $x = \{G_1, \ldots, G_q\}$ and $z = \{\tilde{G}_1, \ldots, \tilde{G}_s\}$ is

$$x \wedge z = \{G_i \cap \tilde{G}_j : G_i \cap \tilde{G}_j \neq \emptyset\}.$$

Subsequently, $\rho(x, z) = \text{rank}(x \wedge y)$ is $p - \#$ groups in $x \wedge z$, which can be equivalently expressed as:

$$\rho(x, z) = \sum_{i,j:|G_i\cap\tilde{G}_j|\neq\emptyset} |G_i \cap \tilde{G}_j| - 1.$$

For sets $G_1, G_2 \subseteq \{1, 2, \ldots, p\}$ with $G_1 \cap G_2 = \emptyset$, we define:

$$\mathcal{R}_{G_1,G_2} \triangleq \{\{a_1\}, \{a_2\}, \ldots, \{a_p\}\} \setminus \{\{a_i\} : a_i \in G_1 \cup G_2\}.$$

E.3.1 Characterizing $\mathcal{S}$ for clustering

We construct a set $\mathcal{S}$ satisfying the properties in Definition 3. Initialize $\mathcal{S} = \emptyset$. Then, for every $k = 1, 2, \ldots, p-1$ and pairs of groups of variables $G_1 \subseteq \{a_1, \ldots, a_p\}$ and $G_2 \subseteq \{a_1, \ldots, a_p\}$ with $|G_1| + |G_2| = k + 1$ and $G_1 \cap G_2 = \emptyset$, we generate covering pairs (x, y) with $y = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $x = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$, and let

$$S = S \cup (x, y).$$

Recalling that $\mathcal{S}_k = \{(x, y) \in \mathcal{S}, \text{rank}(y) = k\}$, it is straightforward to check that for every $k = 1, 2, \ldots, p-1$

$$|\mathcal{S}_k| = \binom{p}{k+1} \sum_{\ell=1}^{k} \binom{k+1}{l}.$$

Here, the terms $\binom{p}{k+1}$ counts the number of possible items in $G_1 \cup G_2$ and the term $\sum_{\ell=1}^{k+1}\binom{k+1}{l}$ counts the number of possible configurations of the group G_2. We will show that the constructed set $\mathcal{S}$ satisfies Definition 3 of the main paper. Our analysis is based on the following lemma.

Lemma 11. *Consider the covering pairs (x, y) with $x = \{G_1, G_2, \ldots, G_q\}$ and $y = \{G_1 \cup G_2, G_3, \ldots, G_q\}$ where $G_i \subseteq \{1, 2, \ldots, px\}$ and $G_i \cap G_j = \emptyset$ for every $i \neq j$. Let $(\tilde{x}, \tilde{y})$ be covering pairs with $\tilde{y} = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $\tilde{x} = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$. Then, for every $z \in \mathcal{L}$, $\rho(y, z) - \rho(x, z) = \rho(\tilde{y}, z) - \rho(\tilde{x}, z)$.*

Proof of Lemma 11. Let $z = \{\tilde{G}_1, \ldots, \tilde{G}_s\}$ with $\tilde{G}_i \subseteq \{a_1, a_2, \ldots, a_p\}$ and $\tilde{G}_i \cap \tilde{G}_j = \emptyset$ for every $i \neq j$. Then:

$$\rho(y,z) = \sum_{j:(G_1 \cup G_2) \cap \tilde{G}_j \neq \emptyset} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1 + \sum_{i \geq 3, j: G_i \cap \tilde{G}_j \neq \emptyset} |G_i \cap \tilde{G}_j| - 1,$$

and

$$\rho(x,z) = \sum_{j:G_1 \cap \tilde{G}_j \neq \emptyset} |G_1 \cap \tilde{G}_j| - 1 + \sum_{j:G_2 \cap \tilde{G}_j \neq \emptyset} |G_2 \cap \tilde{G}_j| - 1 + \sum_{i \geq 3, j: G_i \cap \tilde{G}_j \neq \emptyset} |G_i \cap \tilde{G}_j| - 1.$$

Since $\mathcal{R}_{G_1,G_2}$ consists of groups of size one, we have that:

$$\rho(\tilde{y},z) = \sum_{j:(G_1 \cup G_2) \cap \tilde{G}_j \neq \emptyset} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1,$$

and

$$\rho(\tilde{x},z) = \sum_{j:G_1 \cap \tilde{G}_j \neq \emptyset} |G_1 \cap \tilde{G}_j| - 1 + \sum_{j:G_2 \cap \tilde{G}_j \neq \emptyset} |G_2 \cap \tilde{G}_j| - 1.$$

We thus can see that $\rho(y,z) - \rho(x,z) = \rho(\tilde{y},z) - \rho(\tilde{x},z)$. □

Showing $\mathcal{S}$ satisfies Definition 3 With Lemma 11 at hand, we show that out constructed $\mathcal{S}$ satisfies Definition 3 of the main paper. We start with the first property. Consider any $(u', v') \subseteq \mathcal{L}$. Without loss of generality, we take $v' = \{G_1 \cup G_2, G_3, \ldots, G_q\}$ and $u' = \{G_1, G_2, \ldots, G_q\}$. We let $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$. Then, according to Lemma 11, we have that $\rho(v',z) - \rho(u',z) = \rho(v,z) - \rho(u,z)$. Furthermore, since $\text{rank}(x) = p - \#$ groups in x, we have that $\text{rank}(v) \leq \text{rank}(v')$. Thus, the first property of $\mathcal{S}$ is satisfied. We demonstrate the second property. Consider any $(u,v) \in \mathcal{S}$ and $(u',v') \in \mathcal{S}$ that are different. Let $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$ and $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$. Additionally, let $u' = \{G'_1, G'_2, \mathcal{R}_{G'_1,G'_2}\}$ and $v' = \{G'_1 \cup G'_2, \mathcal{R}_{G'_1,G'_2}\}$. Since the covering pairs (u,v) and (u',v') are different, there must exist two items a_i, a_j such that either (a_i, a_j) are grouped together in v but are not together in u or (a_i, a_j) are grouped together in v' but are not together in u'. Let $z = \{\{a_i, a_j\}, \mathcal{R}_{\{a_i\},\{a_j\}}\}$. Since $\rho(v,z) - \rho(u,z) = \mathbb{I}[(a_i, a_j)$ grouped together in v but not in $u]$ and $\rho(v',z) - \rho(u',z) = \mathbb{I}[(a_i, a_j)$ grouped together in v' but not in $u']$, we have that $\rho(v,z) - \rho(u,z) \neq \rho(v',z) - \rho(u',z)$.

E.3.2 Characterizing $c_{\mathcal{L}}(u,v)$ for covering pair (u,v)

Lemma 12. *Let $v = \{G_1 \cup G_2, \mathcal{R}_{G_1,G_2}\}$ and $u = \{G_1, G_2, \mathcal{R}_{G_1,G_2}\}$ be a covering pair $(u,v) \in \mathcal{S}$. Then, $c_{\mathcal{L}}(u,v) = \min\{|G_1|, |G_2|\}$.*

Proof of Lemma 12. Let $z = \{\tilde{G}_1, \ldots, \tilde{G}_q\}$. Then, from proof of Lemma 11, we have that:

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j:(G_1 \cup G_2) \cap \tilde{G}_j \neq \emptyset} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1\right] - \left[\sum_{j:G_1 \cap \tilde{G}_j \neq \emptyset} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j:G_2 \cap \tilde{G}_j \neq \emptyset} |G_2 \cap \tilde{G}_j| - 1\right].$$

Let $I_1 := \{j : \tilde{G}_j \cap G_1 \neq \emptyset\}$ and $I_2 := \{j : \tilde{G}_j \cap G_2 \neq \emptyset\}$. Then,

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j \in I_1 \cup I_2} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1\right] - \left[\sum_{j \in I_1} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j \in I_2} |G_2 \cap \tilde{G}_j| - 1\right].$$

Simple manipulations yield:

$$\rho(v,z) - \rho(u,z) = \left[\sum_{j \in I_1 \cap I_2} |(G_1 \cup G_2) \cap \tilde{G}_j| - 1\right] - \left[\sum_{j \in I_1 \cap I_2} |G_1 \cap \tilde{G}_j| - 1\right] - \left[\sum_{j \in I_1 \cap I_2} |G_2 \cap \tilde{G}_j| - 1\right].$$

Clearly, if $I_1 \cap I_2 = \emptyset$, then $\rho(v,z) - \rho(u,z) = 0$. Suppose $I_1 \cap I_2 \neq \emptyset$. Then,

$$\rho(v,z) - \rho(u,z) = |I_1 \cap I_2| + \left[\sum_{j \in I_1 \cap I_2} |(G_1 \cup G_2) \cap \tilde{G}_j| - |G_1 \cap \tilde{G}_j| - |G_2 \cap \tilde{G}_j| \right] = |I_1 \cap I_2|.$$

Notice that $|I_1 \cap I_2| \leq \min\{|G_1|, |G_2|\}$. Then, the upper bound can be achieved by for example setting $z = \{N, \{\{a_1\}, \{a_2\}, \ldots, \{a_p\} \setminus N\}$ with $N = \{(a_i, a_j) : a_i \in G_1, a_j \in G_2\}$. □

E.3.3 Refined false discovery bound for clustering

Let $\hat{x}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Then:

$$\mathbb{E}[\text{FD}(\hat{x}_{\text{stable}}, x^\star)] \leq \sum_{k=1}^{p-1} \frac{q_k^2}{(1-2\alpha)\binom{p}{k+1}\sum_{\ell=1}^{k}\binom{k+1}{l}},$$

where,

$$\begin{aligned} q_k &= \sum_{(u,v) \in \mathcal{S}_k} \frac{\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]}{c(u,v)} \\ &= \sum_{\substack{G_1 \subseteq \{a_1, \ldots, a_p\}, G_2 \subseteq \{a_1, \ldots, a_p\} \\ G_1 \cap G_2 = \emptyset; |G_1| + |G_2| = k+1}} \frac{\mathbb{E}[\#\text{ groups } \hat{G}_j \text{ in } \hat{x}_{\text{sub}} \text{ satisfying } \hat{G}_j \cap G_1 \neq \emptyset \text{ and } \hat{G}_j \cap G_2 \neq \emptyset]}{\min\{|G_1|, |G_2|\}}. \end{aligned}$$

Here, $\hat{x}_{\text{sub}}$ represents clustering from supplying $n/2$ samples to base estimator. We will use the following data driven approximation to estimate q_k

$$q_k \approx \frac{1}{B} \sum_{\substack{G_1 \subseteq \{a_1, \ldots, a_p\}, G_2 \subseteq \{a_1, \ldots, a_p\} \\ G_1 \cap G_2 = \emptyset; |G_1| + |G_2| = k+1}} \sum_{\ell=1}^{B} \frac{\#\text{ groups } \hat{G}_j \text{ in } \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)}) \text{ satisfying } \hat{G}_j \cap G_1 \neq \emptyset \text{ and } \hat{G}_j \cap G_2 \neq \emptyset]}{\min\{|G_1|, |G_2|\}},$$

with $\hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the partition obtained from supplying $\mathcal{D}^{(\ell)}$ to the base estimator.

E.4 Causal structure learning

Throughout, we consider covering pairs $(\mathcal{C}_u, \mathcal{C}_v)$ where each connected component in in the skeletons of $\mathcal{C}_u, \mathcal{C}_v$ have diameter at most two. We denote this set by $\mathcal{T}$. Note that for any covering pair $(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{T}$, $\mathcal{C}_v$ is a polytree. Throughout, we will use the similarity valuation $\rho := \rho_{\text{meet}}$. Our analysis in this section will build on the the following result.

Lemma 13. *Let $\mathcal{C}_u$ and $\mathcal{C}_v$ be two CPDAGs that are polytrees with $\mathcal{C}_u \preceq \mathcal{C}_v$. Then, the following statements hold:*

(a) for any pairs of nodes $\mathcal{E}$, the set of DAGs that result from removing edges among pairs $\mathcal{E}$ in any DAG $\mathcal{G}_v$ form a Markov equivalence class.

(b) for every DAG $\mathcal{G}_v \in \mathcal{C}_v$, there exists a DAG $\mathcal{G}_u \in \mathcal{C}_u$ such that $\mathcal{G}_u$ is a directed subgraph of $\mathcal{G}_v$.

Proof of Lemma 13. We first prove part (a). By the polytree assumption, it follows that for any DAG $\mathcal{G}_v$ in the CPDAG $\mathcal{C}_v$, removing the edges among pairs in $\mathcal{E}$ does not create any v-structures, and removes the same (potentially empty) v-structures. That means that the collection of DAGs obtained by taking any DAG in $\mathcal{C}_v$ and removing the edges between the pairs of nodes $\mathcal{E}$ will have the same skeleton and same v-structures, and are thus in the same Markov equivalence class.

We next prove part (b). Let (i,j) be the pair of nodes that are connected in $\mathcal{C}_v$ but not in $\mathcal{C}_u$. Recall that $\mathcal{C}_u \preceq \mathcal{C}_v$ implies there exists a DAG $\mathcal{G}_u \in \mathcal{C}_u$ and a DAG $\mathcal{G}_v \in \mathcal{C}_v$ where $\mathcal{G}_u$ is a subgraph of $\mathcal{G}_v$, where $\mathcal{G}_u$ does not have the edge among pairs (i,j). Appealing to the result in part (a), we have that removing the edge (i,j) from any other DAG in $\mathcal{C}_v$ results in a DAG in the same equivalence class, which is $\mathcal{C}_u$. □

E.4.1 Characterizing $\mathcal{S}$ for causal structure learning

We construct the set $\mathcal{S}$ as follows. Initialize $\mathcal{S} = \emptyset$. For every reference node, and $k = 1, \ldots, p-1$, let $\mathcal{C}_y$ be a CPDAG generated with k edges, where every edge is between the reference node and another node; no other edges can be added without violating the condition that the largest undirected path has size less than or equal to two. A consequence of Lemma 13 is that there are k CPDAGs $\mathcal{C}_{x_1}, \ldots, \mathcal{C}_{x_k}$ that form a covering pair with $\mathcal{C}_y$. We then let

$$\mathcal{S} = \mathcal{S} \cup (\mathcal{C}_{x_i}, \mathcal{C}_y),$$

for every $i = 1, 2, \ldots, k$. Recall that $\mathcal{S}_k := \{(\mathcal{C}_x, \mathcal{C}_y) \in \mathcal{S}, \text{rank}(\mathcal{C}_y) = k\}$. Then,

$$|\mathcal{S}_k| = p \binom{p-1}{k} \sum_{i \in \{0,2\ldots,k\}} \binom{k}{i}.$$

The result above follows from noting that that for every reference node and k other nodes, there are $\sum_{i \in \{0,2\ldots,k\}} \binom{k}{i}$ possible CPDAGs that are polytrees can formed by connecting the k nodes to the reference node; the factor $p\binom{p-1}{k}$ comes from p total possible reference nodes and $\binom{p-1}{k}$ possible set of k nodes to connect to the reference node.

We will show that the constructed set $\mathcal{S}$ satisfies Definition 3 of the main paper. Our analysis is based on the following lemma.

Lemma 14. *Let $\mathcal{C}_{\tilde{y}}$ be a CPDAG that contains m disconnected subgraphs (both directed and undirected). Let $\mathcal{C}_{\tilde{y}_i}$ be each disconnected subgraph for $i = 1, 2, \ldots, m$. Then, for any CPDAG $\mathcal{C}_z$,*

$$\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) = \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z).$$

Proof. We will first show that $\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \leq \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Let $\mathcal{C}_{\tilde{x}} \in \text{argmax}_{\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}}, \mathcal{C}_x \preceq \mathcal{C}_z} \text{rank}(\mathcal{C}_x)$. By definition, $\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}}$ if there is a DAG $\mathcal{G}_x$ in $\mathcal{C}_x$ and a DAG $\mathcal{G}_{\tilde{y}}$ in $\mathcal{C}_{\tilde{y}}$ such that $\mathcal{G}_x$ is a subgraph of $\mathcal{G}_{\tilde{y}}$. Since $\mathcal{G}_{\tilde{y}}$ has disconnected components, so must $\mathcal{G}_x$. We let $\mathcal{C}_{\tilde{x}_i}$ be the subgraphs of $\mathcal{C}_{\tilde{x}}$ where every subgraph $\mathcal{C}_{\tilde{x}_i}$ only contains edges among nodes that are connected (to other nodes) in the graph $\mathcal{C}_{\tilde{y}_i}$. By construction, $\mathcal{C}_{\tilde{x}_i} \preceq \mathcal{C}_{\tilde{y}_i}$, $\text{rank}(\mathcal{C}_{\tilde{x}}) = \sum_{i=1}^{m} \text{rank}(\mathcal{C}_{\tilde{x}_i})$, and $\mathcal{C}_{\tilde{x}_i} \preceq \mathcal{C}_z$. Thus, $\text{rank}(\mathcal{C}_{\tilde{x}_i}) \leq \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Then, we can conclude that

$$\sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z) \geq \sum_{i=1}^{m} \text{rank}(\mathcal{C}_{\tilde{x}_i}) = \text{rank}(\mathcal{C}_{\tilde{x}}) = \rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z).$$

Now we will show that $\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \geq \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z)$. Let $\mathcal{C}_{\tilde{x}_i} \in \text{argmax}_{\mathcal{C}_x \preceq \mathcal{C}_{\tilde{y}_i}, \mathcal{C}_x \preceq \mathcal{C}_z} \text{rank}(\mathcal{C}_x)$. Now form a CPDAG $\mathcal{C}_{\bar{y}}$ by combining all the disjoint graphs $\mathcal{C}_{\tilde{x}_i}$ for every $i = 1, 2, \ldots, m$ into one graph. Since these graphs are disjoint (i.e. nodes that are connected in each graph are distinct), we have that $\mathcal{C}_{\bar{y}} \preceq \mathcal{C}_{\tilde{y}}$ and $\mathcal{C}_{\bar{y}} \preceq \mathcal{C}_z$ and that $\text{rank}(\mathcal{C}_{\bar{y}}) = \sum_{i=1}^{m} \text{rank}(\mathcal{C}_{\tilde{x}_i})$. So we conclude that

$$\rho(\mathcal{C}_{\tilde{y}}, \mathcal{C}_z) \geq \text{rank}(\mathcal{C}_{\bar{y}}) = \sum_{i=1}^{m} \text{rank}(\mathcal{C}_{\tilde{x}_i}) = \sum_{i=1}^{m} \rho(\mathcal{C}_{\tilde{y}_i}, \mathcal{C}_z).$$

□

Showing $\mathcal{S}$ satisfies Definition 3 For the first property, consider covering pairs $(\mathcal{C}_{u'}, \mathcal{C}_{v'}) \in T$. Let (i, j) be the pair of nodes that are connected in $\mathcal{C}_{v'}$ and are not connected in $\mathcal{C}_{u'}$. Since every undirected path in $\mathcal{C}_{v'}$ has size at most 2, then $\mathcal{C}_{v'}$ decouples into two disconnected CPDAGs $\mathcal{C}_v$ and $\mathcal{C}_1$, where $\mathcal{C}_v$ only involves nodes adjacent to (i, j). Similarly, $\mathcal{C}_{u'}$ decouples into two disconnected CPDAGs $\mathcal{C}_u$ and $\mathcal{C}_2$, where $\mathcal{C}_2 = \mathcal{C}_1$ and $\mathcal{C}_u$ is covered by $\mathcal{C}_v$. From Lemma 14, we have that for any CPDAG $\mathcal{C}_z$

$$\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = \rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z).$$

Notice that $(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}$. Furthermore, since the number of edges (directed and undirected) in $\mathcal{C}_{v'}$ is larger than $\mathcal{C}_v$, we have that $\text{rank}(\mathcal{C}_v) \leq \text{rank}(\mathcal{C}_{v'})$.

We next show the second property in Definition 3. Let $(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}$ and $(\mathcal{C}_{u'}, \mathcal{C}_{v'}) \in \mathcal{S}$. Our objective is to show that $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = \rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z)$ for all $\mathcal{C}_z \Leftrightarrow \mathcal{C}_u = \mathcal{C}_{u'}$ and $\mathcal{C}_v = \mathcal{C}_{v'}$. The direction $\leftarrow$ trivially holds, and hence we focus on the direction $\rightarrow$. We consider multiple scenarios; throughout the extra edge that is present in $\mathcal{C}_v$ and not in $\mathcal{C}_u$ is between the pair of nodes (i, j), and the extra edge that is present in $\mathcal{C}_{v'}$ and not in $\mathcal{C}_{u'}$ is between the pair of nodes (k, l).

(1) Suppose that the nodes (k, l) are not connected in $\mathcal{C}_v$. Letting $\mathcal{C}_z$ be a CPDAG with only an edge between nodes (k, l), we find that $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = 0$ and $\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = 1$. So this scenario cannot occur.

(2) Suppose there is an edge between pairs (s, t) in $\mathcal{C}_{u'}$ that is missing in $\mathcal{C}_v$ (and as a result in $\mathcal{C}_u$). Construct CPDAG $\mathcal{C}_z$ with two edges, one between the pair (i, j) and another between the pair (s, t) with the property that $\mathcal{C}_z \not\preceq \mathcal{C}_{v'}$; this construction is possible since $(\mathcal{C}_{u'}, \mathcal{C}_{v'}) \in \mathcal{S}$, meaning that if there is an edge between pair of nodes (i, j) in $\mathcal{C}_{v'}$, this edge is incident to the edge between the pair of nodes (s, t). Then, it is evident that $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = 1$ but $\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = 0$. So this scenario cannot occur.

(3) Suppose there is an edge between pairs (s, t) in $\mathcal{C}_{u'}$ that is missing in $\mathcal{C}_u$ but is not missing in $\mathcal{C}_v$. Let $\mathcal{C}_z$ be a CPDAG only containing an edge between (s, t). Then it follows that $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = 1$ but $\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = 0$. So this scenario cannot occur.

From the impossibilities of scenarios 1-2, and noting that a similar argument can be made by swapping $\mathcal{C}_{u'}$ with $\mathcal{C}_u$, and $\mathcal{C}_{v'}$ with $\mathcal{C}_v$, we conclude that $\mathcal{C}_v, \mathcal{C}_{v'}$ have edges between the same pairs of nodes. Combining this result with the impossibility of scenario 3, we conclude that $\mathcal{C}_u, \mathcal{C}_{u'}$ have edges between the same pairs of nodes. We then continue with the final scenario.

(4) Suppose that $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ are not identical CPDAGs. Since both $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have maximum undirected path length less than or equal to two, they both must have the same reference node i (where the other nodes are connected to). Furthermore, since $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have the same skeleton and are different, they must have strictly more than one edge, and they must have different v-structures. As a first sub-case, suppose $\mathcal{C}_{v'}$ have a v-structure $s \rightarrow i \leftarrow t$ hat is not present in $\mathcal{C}_v$, so that $s \leftarrow i$ or $s - i$ in $\mathcal{C}_v$. Then, let $\mathcal{C}_z$ be a CPDAG containing two edges between the pairs (i, j) and (i, s) with $\mathcal{C}_z \preceq \mathcal{C}_v$. By construction, $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) = 1$ but $\rho(\mathcal{C}_{v'}, \mathcal{C}_z) - \rho(\mathcal{C}_{u'}, \mathcal{C}_z) = 0$. Swapping $\mathcal{C}_{u'}$ with $\mathcal{C}_u$, and $\mathcal{C}_{v'}$ with $\mathcal{C}_v$, and following similar arguments, we arrive again at a contradiction if $\mathcal{C}_v$ has a v-structure that is not present in $\mathcal{C}_{v'}$.

From the impossibility of scenario 4, we conclude that $\mathcal{C}_v$ and $\mathcal{C}_{v'}$ have the same skeleton and v-structure and consequently $\mathcal{C}_v = \mathcal{C}_{v'}$. We thus have that $\mathcal{C}_u \preceq \mathcal{C}_v$ and $\mathcal{C}_{u'} \preceq \mathcal{C}_v$. Furthermore, since $\mathcal{C}_{u'}$ and $\mathcal{C}_u$ have the same skeleton, both are missing an edge between pair of nodes (i, j) that is connected in $\mathcal{C}_v$. Appealing to part a of Lemma 13, we conclude that $\mathcal{C}_u = \mathcal{C}_{u'}$.

E.4.2 Characterizing $c_{\mathcal{L}}(\mathcal{C}_u, \mathcal{C}_v)$ for covering pairs $(\mathcal{C}_u, \mathcal{C}_v)$

We have the following lemma.

Lemma 15. *Let $(\mathcal{C}_u, \mathcal{C}_v)$ be CPDAGs that are polytrees and form a covering pair. Then, $c_{\mathcal{L}}(\mathcal{C}_u, \mathcal{C}_v) = 1$.*

Proof. Let the pair of nodes (i, j) be connected in $\mathcal{C}_v$ and not connected in $\mathcal{C}_u$. Consider any CPDAG $\mathcal{C}_z$. Let $\mathcal{C}_{\tilde{y}} \in \text{argmax}_{\mathcal{C}_y \preceq \mathcal{C}_v, \mathcal{C}_y \preceq \mathcal{C}_z} \text{rank}(\mathcal{C}_y)$. Since the CPDAG $\mathcal{C}_v$ is a polytree, so is the CPDAG $\mathcal{C}_{\tilde{y}}$. Let $\mathcal{G}_v$ be any DAG in $\mathcal{C}_v$. Then, by Lemma 13, there exists DAGs $\mathcal{G}_{\tilde{y}}^{(1)} \in \mathcal{C}_{\tilde{y}}$ and $\mathcal{G}_u \in \mathcal{C}_u$ such that $\mathcal{G}_{\tilde{y}}^{(1)}$ and $\mathcal{G}_u$ are both subgraphs of $\mathcal{G}_v$. Suppose we remove an edge that may be present between the pair of nodes (i, j) in $\mathcal{G}_{\tilde{y}}^{(1)}$ and denote the resulting subgraph by $\mathcal{G}_x^{(1)}$. By construction, $\mathcal{G}_x^{(1)}$ is also a subgraph of $\mathcal{G}_u$. Since $\mathcal{C}_{\tilde{y}} \preceq \mathcal{C}_z$, there exists a DAG $\mathcal{G}_{\tilde{y}}^{(2)} \in \mathcal{C}_{\tilde{y}}$ and a DAG $\mathcal{G}_z \in \mathcal{C}_z$ such that $\mathcal{G}_{\tilde{y}}^{(2)}$ is a subgraph of $\mathcal{G}_z$. Suppose again we remove an edge that may be present between the pair of nodes (i, j) in $\mathcal{G}_{\tilde{y}}^{(2)}$ and denote the resulting subgraph by

$\mathcal{G}_x^{(2)}$. By Lemma 13, $\mathcal{G}_x^{(2)}$ and $\mathcal{G}_x^{(1)}$ are in the same equivalence class, which we denote by $\mathcal{C}_x$. By construction, $\mathcal{C}_x \preceq z$ and $\mathcal{C}_x \preceq \mathcal{C}_u$. Furthermore, $\text{rank}(\mathcal{C}_x) \geq \text{rank}(\mathcal{C}_{\tilde{y}}) - 1$. Thus, we have shown that for any arbitrary $\mathcal{C}_z$: $\rho(\mathcal{C}_v, \mathcal{C}_z) - \rho(\mathcal{C}_u, \mathcal{C}_z) \leq 1$. □

E.4.3 Refined false discovery bound for causal structure learning

Let $\hat{\mathcal{C}}_{\text{stable}}$ be output of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$. Let $\mathcal{C}^\star$ be the population CPDAG. Then:

$$\mathbb{E}[\text{FD}(\hat{\mathcal{C}}_{\text{stable}}, \mathcal{C}^\star)] \leq \sum_{k=1}^{p-1} \frac{q_k^2}{(1-2\alpha)p \binom{p-1}{k} \sum_{i \in \{0,2\ldots,k\}} \binom{k}{i}},$$

where,

$$q_k = \sum_{(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}_k} \mathbb{E}[\rho(\mathcal{C}_v, \hat{\mathcal{C}}_{\text{sub}}) - \rho(\mathcal{C}_u, \hat{\mathcal{C}}_{\text{sub}})].$$

Here, $\hat{\mathcal{C}}_{\text{sub}}$ represents the CPDAG from supplying $n/2$ samples to base estimator. We will use the following data driven approximation to estimate q_k

$$q_k \approx \frac{1}{B} \sum_{\ell=1}^{B} \sum_{(\mathcal{C}_u, \mathcal{C}_v) \in \mathcal{S}_k} \mathbb{E}[\rho(\mathcal{C}_v, \hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)}) - \rho(\mathcal{C}_u, \hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)}))],$$

with $\hat{\mathcal{C}}_{\text{base}}(\mathcal{D}^{(\ell)})$ represents the CPDAGs obtained from supplying dataset $\mathcal{D}^{(\ell)}$ to base estimator $\hat{\mathcal{C}}_{\text{base}}$.

F Assumptions 1 and 2 of the main paper for the total ranking problem in Example 7

Let $S = \{a_1, a_2, \ldots, a_p\}$ be the set of p elements. Let $\pi_{\text{null}}(a_i) = i$ for every $i = 1, 2, \ldots, p$. We use the similarity valuation $\rho := \rho_{\text{total-ranking}}$ in (2) of the main paper. As each element in the poset corresponds to a function $\pi : S \to S$, we will use this functional notation throughout. For a covering pair (π_1, π_2), there exists a single pair of elements $(a_i, a_j) \in \text{inv}(\pi_2; \pi_{\text{null}}) \setminus \text{inv}(\pi_1; \pi_{\text{null}})$ with $j > i$. Then, from the definition of ρ, for any permutation π, we have that

$$\rho(\pi_2, \pi) - \rho(\pi_1, \pi) = \mathbb{I}[(a_i, a_j) \in \text{inv}(\pi; \pi_{\text{null}})] = \mathbb{I}[\pi(a_j) < \pi(a_i)].$$

Let $\hat{\pi}_{\text{sub}}$ be the estimate ranking from applying a base procedure on a subsample of the data. Consider a fixed integer k with $1 \leq k \leq p-1$. Define the sets S_1 and S_2:

$$S_1 = \{(a_i, a_j) \in \text{inv}(\pi^\star; \pi_{\text{null}}) : j - i = k\},$$
$$S_2 = \{(a_i, a_j) \notin \text{inv}(\pi^\star; \pi_{\text{null}}) : j - i = k\}.$$

The set S_1 corresponds to non-null pairs (as described in the main paper) and the set S_2 corresponds to null pairs.

Then, appealing to the definition of $\mathcal{S}$ and the constant $c_{\mathcal{L}}(\cdot, \cdot)$ in the total ranking case (see Section E.2), Assumption 1 of the main paper reduces to the following inequality being satisfied

$$\frac{\sum_{(a_i, a_j) \in S_1} \mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i))}{\sum_{(a_i, a_j) \in S_2} \mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i))} \geq \frac{|S_1|}{|S_2|}. \tag{19}$$

Consider an estimator $\hat{\pi}_{\text{sub}} = \hat{\pi}_{\text{random}}$ that randomly selects a total ranking in the space of permutations. Then, for every i and j, $\mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i)) = \frac{1}{2}$. Thus, in this case, Assumption 1 in (19) is satisfied with equality.

It is also straightforward to check that Assumption 2 of the main paper reduces to

$$\mathbb{P}(\hat{\pi}_{\text{sub}}(a_j) < \hat{\pi}_{\text{sub}}(a_i)) \text{ being the same for every } (a_j, a_i) \in S_2.$$