PNAS RESEARCH ARTICLE | APPLIED MATHEMATICS

Model selection over partially ordered sets

Armeen Taeb[a,1], Peter Bühlmann[b], and Venkat Chandrasekaran[c,d]

Edited by Peter Bickel, University of California, Berkeley, CA; received August 20, 2023; accepted October 19, 2023

In problems such as variable selection and graph estimation, models are characterized by Boolean logical structure such as the presence or absence of a variable or an edge. Consequently, false-positive error or false-negative error can be specified as the number of variables/edges that are incorrectly included or 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.

combinatorics | greedy algorithms | multiple testing | stability

Significance

The increasing complexity of modern datasets has been accompanied by the use of sophisticated modeling paradigms in which the task of model selection is a significant challenge. In particular, models specified by structures such as permutations (for ranking) or directed acyclic graphs (for causal inference) are not characterized by an underlying Boolean logical structure, which leads to difficulties with formalizing and controlling false-positive error. We address this challenge by organizing classes of models as partially ordered sets, which leads to systematic approaches for defining natural generalizations of false-positive error and methodology for controlling this error.

In data-driven approaches to scientific discovery, one is commonly faced with the problem of model selection. Popular examples include variable selection (which covariates predict 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 the 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-positive error or false-negative error corresponds to the number of attributes that are incorrectly included or excluded in the model.

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 well they predict 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-positive error, and false-negative error 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 of 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 model class.

As a concrete illustration of the inappropriateness of Boolean logical structure for clustering, consider three items a, b, c, with the null model given by the three clusters $\{a\}, \{b\}, \{c\}$, the true model by the two clusters $\{a, b\}, \{c\}$, and the estimated model by

Author affiliations: [a]Department of Statistics, University of Washington, Seattle, WA 98195; [b]Seminar for Statistics, Eidgenossische Technische Hochschule Zürich, Zurich, CH-8092, Switzerland; [c]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125; and [d]Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125

Author contributions: A.T., P.B., and V.C. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: ataeb@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2314228121/-/DCSupplemental.

Published February 16, 2024.

the single cluster $\{a, b, c\}$. An incorrect perspective grounded in Boolean logical structure suggests a false-positive error of two, with the "false discoveries" being that c is in the same cluster as a and as b. However, accounting for set partition structure yields the more accurate false-positive error value of one as a and b are in the same cluster in the true and estimated models; hence, 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 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 (1–3).

1. Our Contributions

We begin in Section 3 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, multisample testing, and blind source separation. In Section 4, 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 (4, 5) 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 4 are provided in Section 6. In Section 5, we provide numerical illustration via experiments on synthetic and real data.

2. Related Work

Classic approaches to model selection such as the Akaike information criterion and Bayesian information criterion assess and penalize model complexity by counting the number of attributes included in a model (6, 7). More generally, such complexity measures facilitate a hierarchical organization of model classes, and this perspective is prevalent throughout much of the model selection literature (8–13). 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.

3. 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.

A. 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 antisymmetric ($x \preceq y, y \preceq x \Rightarrow x = y,\ \forall x, y \in \mathcal{L}$). 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 structures. 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; Fig. 1 presents Hasse diagrams associated with several of these examples.

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 (14).

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 (14).

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

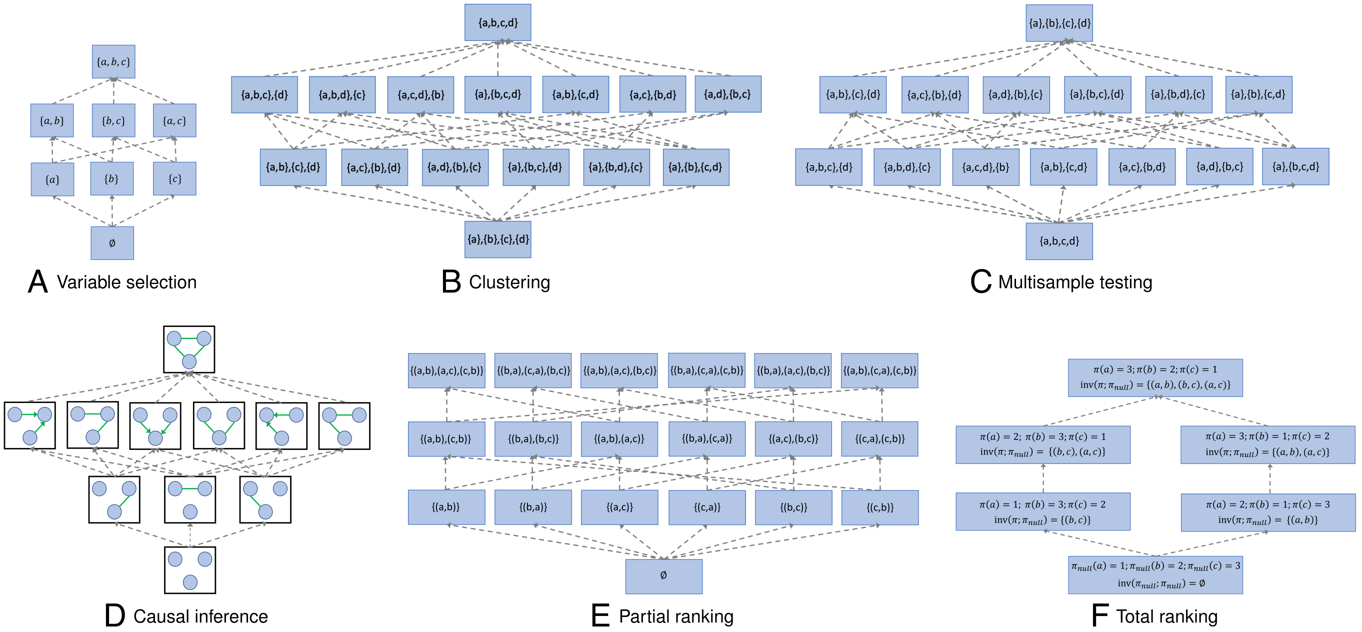


Fig. 1. Hasse diagrams for (*A*) variable selection with three variables (Example 1); (*B*) clustering four variables (Example 2); (*C*) multisample testing with four samples (Example 3); (*D*) causal inference with three variables (Example 4); (*E*) partial ranking of three items (Example 6); and (*F*) total ranking of three items (Example 7).

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 ref. 15 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 (16). A Markov equivalence class of DAGs can be described by a completed partially directed acyclic graph (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. In some causal inference contexts, it may be more natural to view the fully connected CPDAG as the null model. Our framework can accommodate this perspective by reversing the preceding model poset. Specifically, in this reversed poset, the partial order is given by inclusion of conditional independencies, the least element is the fully connected CPDAG, and the rank function is $p(p-1)/2$ minus the number of edges. Higher-rank elements in this poset correspond to causal models exhibiting more conditional independence 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

https://doi.org/10.1073/pnas.2314228121

poset is the reverse of the (bounded) integer poset (14) 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 $\text{inv}(\pi;\pi_{\text{null}}) := \{(x,y) \in S \times S \mid \pi_{\text{null}}(x) < \pi_{\text{null}}(y),\ \pi(x) > \pi(y)\}$. 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 (14).

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 (17, 18). Finally, Example 8 was studied in ref. 19, 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 and 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.

B. 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 the 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 (14) 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 $(\mathcal{L}, \preceq, \text{rank}(\cdot))$ 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:

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

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

B.1. 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 4, 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 $x, y \in \mathcal{L}$:

$$\rho_{\mathrm{meet}}(x, y) := \max_{z \preceq x, z \preceq y} \mathrm{rank}(z). \qquad [1]$$

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 semilattices. Except for the poset in Example 4 on causal structure learning, the posets in the other examples are meet semilattices (*SI Appendix*, section 1). The subscript "meet" in Eq. **1** signifies that ρ_{meet} is the rank of the meet for meet semilattices, although ρ_{meet} is well defined even if $(\mathcal{L}, \preceq)$ is not a meet semilattice.

One can check that ρ_{meet} is a similarity valuation on any graded poset $(\mathcal{L}, \preceq, \mathrm{rank}(\cdot))$; see *SI Appendix*, section 2, 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 $\rho_{\mathrm{meet}}(x, y) = p \cdot T - \sum_{i=1}^{p} \max\{x_i, y_i\}$. 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 (18).

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 Eq. **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 $\pi_{\mathrm{null}}(a) = 1, \pi_{\mathrm{null}}(b) = 2, \pi_{\mathrm{null}}(b) = 3$, the true model given by the ranking $\pi^\star(a) = 3, \pi^\star(b) = 1, \pi^\star(c) = 2$ (Hasse diagram shown in Fig. 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 Fig. 1 that $\rho_{\mathrm{meet}}(\hat{\pi}, \pi^\star) = 0$, which suggests that no discovery is made. On the other hand, the inversion sets of these rankings are given by $\mathrm{inv}(\pi^\star; \pi_{\mathrm{null}}) = \{(a, b), (a, c)\}$ and $\mathrm{inv}(\hat{\pi}; \pi_{\mathrm{null}}) = \{(a, c), (b, c)\}$, 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 $\mathrm{inv}(\pi^\star; \pi_{\mathrm{null}}) \cap \mathrm{inv}(\hat{\pi}; \pi_{\mathrm{null}})$ 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 $\mathrm{inv}(\pi^\star; \pi_{\mathrm{null}}) \cap \mathrm{inv}(\hat{\pi}; \pi_{\mathrm{null}})$ that correspond to inversion sets of total rankings as the maximization in Eq. **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}):

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

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, $\rho_{\mathrm{meet}}(\hat{x}, x^\star)$ is equal to the dimension of the intersection of the subspaces $\hat{x}, x^\star$. When these subspaces have small dimensions, 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 ref. 19, 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 $\mathcal{U}, \tilde{\mathcal{U}}$:

$$\rho_{\mathrm{subspace}}(\mathcal{U}, \tilde{\mathcal{U}}) = \mathrm{trace}(\mathcal{P}_{\mathcal{U}} \mathcal{P}_{\tilde{\mathcal{U}}}). \qquad [3]$$

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.

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 $k \leq \ell$ (due to the symmetry of similarity valuations) and let $\mathrm{Perm}(\ell)$ be the collection of bijections on $\{1, \dots, \ell\}$. With this notation, consider the following similarity valuation:

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

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.

4. 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 A 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 B and C. Some of the discussion in Section A is relevant for all of the posets in Examples 1–9, while the methodology presented in Sections B and C 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 and 9.

A. 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 $(x_0, x_1, \dots, x_{k-1}, x_k)$ with x_0 being the least element x_{least} and $x_k = \hat{x}$:

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

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 B and C, 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 6.

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 3 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 $\mathcal{S} = \{(\emptyset, \{i\}) \mid i = 1, \dots, p\}$ and this minimal set is unique. In general, however, such sets are not unique; see *SI Appendix*, section 3, 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

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

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 $\{1,\ldots,p\}$	$\emptyset$	Inclusion of subsets	Cardinality of subset	Subsets $x,\tilde{x}$; $\rho(x,\tilde{x}) = \lvert x\cap\tilde{x}\rvert$
clustering	Partitions of $\{1,\ldots,p\}$	$\{1\},\{2\},\ldots,\{p\}$	Refinement of partition	$p - \#$groups	Partitions $x,\tilde{x}$; $\rho(x,\tilde{x}) = p - \#$ groups in coarsest common refinement
Multisample testing	Partitions of $\{1,\ldots,p\}$	$\{1,2,\ldots,p\}$	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 $\{0,\ldots,T\}^p$	$(T,T,\ldots,T)$	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 set 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 $B\in\mathbb{R}^{p\times k},\tilde{B}\in\mathbb{R}^{p\times\ell},k\le\ell$; $\rho(B,\tilde{B}) = \max_{\sigma\in\text{Perm}(\ell)}\sum_{i=1}^{k}(B^T\tilde{B})^2_{i,\sigma(i)}$

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.

B. Model Selection via 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 with 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 (4, 5) and subspace stability selection for subspace estimation (19). We demonstrate the broad applicability of this methodology in Section 5 by applying it to several examples from Section 3.

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:

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

where $c_{\mathcal{L}}(u,v) := \max_{z\in\mathcal{L}}\rho(v,z) - \rho(u,z)$. 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 4 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 Eq. **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 (4, 5) and the subspace stability selection method of ref. 19. 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 4 (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 $\hat{x}_{\text{stable}}$ from Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ satisfies the following false discovery bound*

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

Here, the set $\mathcal{T}_{\text{null}} := \{(u, v) \text{ covering pair in } \mathcal{L} \mid \rho(v, x^\star) = \rho(u, x^\star)\}$ consists 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 $x^\star$.

In the bound Eq. **7**, the numerator $\mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]^2$ of each summand 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 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 ref. 5. Specifically, in Eq. **7**, we have that $c_{\mathcal{L}}(u, v) = 1$ for any covering pair (u, v) and $\sum_{(u,v)\in\mathcal{S}\cap\mathcal{T}_{\text{null}}} \mathbb{E}[\rho(v, \hat{x}_{\text{sub}}) - \rho(u, \hat{x}_{\text{sub}})]^2 = \sum_{\text{null } i} \mathbb{P}[\text{variable } i \text{ selected by } \hat{x}_{\text{sub}}]^2$.

Theorem 4 is general in its applicability to all the discrete posets in Section 3, 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 Eq. **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 refs. 4 and 19 for stability-based methods for variable selection and subspace estimation. To formulate these assumptions, we introduce some notation. Let $\text{rank}(\mathcal{L}) := \max_{u\in\mathcal{L}} \text{rank}(u)$ be the largest rank of an element in $\mathcal{L}$ and let $\mathcal{S}_k := \{(u, v) \in \mathcal{S} \mid \text{rank}(v) = k\}$ 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 $(u, v) \in \mathcal{S}_k \cap \mathcal{T}_{\text{null}}$.*

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 nonnull 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 (4), 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, (4) 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 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 *SI Appendix*, section 4, for a formal derivation.

Theorem 5 (Refined false discovery control for Algorithm 1 with $\Psi = \Psi_{\text{stable}}$). *Consider the setup of Theorem 4, 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:*

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

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 Eq. **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 *SI Appendix*, section 3.

Remark 4: Specializing Theorem 5 to the case of variable selection, we arrive at the bound in Theorem 1 of ref. 4. Specifically, note that for the Boolean poset with the similarity valuation ρ_{meet}, $\mathcal{S}_k = \emptyset$ for $k \geq 2$, $|\mathcal{S}_1| = \#\,\text{variables}$, and $q_1 = \sum_i \mathbb{P}[\text{variable } i \text{ selected by } \hat{x}_{\text{sub}}]$ is the average number of variables selected by the estimator $\hat{x}_{\text{sub}}$.

Turning our attention to Examples 8 and 9, the situation is considerably more complicated with continuous model posets. A result for these two cases under the same setup as in Theorem 4 yields the following bound for $\alpha \in (0, 1/2)$ (*SI Appendix*, section 5):

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

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 and 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 ref. 19.) It is of interest to develop such a general framework for continuous model posets, and we leave this as a topic for future research.

C. Model Selection via 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 $\mathcal{L}$:

$$\begin{aligned} H_0^{u,v} &: \rho(v, x^\star) = \rho(u, x^\star), \\ \Psi_{\text{test}}(u, v) &:= P\text{-value corresponding to } H_0^{u,v}. \end{aligned} \qquad \textbf{[10]}$$

The null hypothesis $H_0^{u,v}$ in Eq. **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 6 (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 $\Psi = \Psi_{\text{test}}$ satisfies the false discovery bound $\mathbb{P}\left(\text{FD}(\hat{x}_{\text{test}}, x^\star) > 0\right) \le \alpha|\mathcal{S}|$.*

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 settings, 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 *SI Appendix*, section 3, 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 exhibits 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 semilattices (these are dual to the notion of a meet defined in Section 3). Except for the posets in Examples 4, 6, and 9, the posets in the other examples are join semilattices (*SI Appendix*, section 1). For a model class that is a join semilattice, suppose we are provided estimates $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$ of a true model $x^\star$ such that $\text{FD}(\hat{x}^{(j)}, x^\star) = 0,\ j = 1, \ldots, m$ (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 $\hat{x}_{\text{join}}$ of $\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}$ satisfies $\text{FD}(\hat{x}_{\text{join}}, x^\star) = 0$; in general, $\text{rank}(\hat{x}_{\text{join}})$ is larger than $\text{rank}(\hat{x}^{(1)}), \ldots, \text{rank}(\hat{x}^{(m)})$, 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 7 (Using joins to obtain more powerful models). *Let $(\mathcal{L}, \preceq, \text{rank}(\cdot))$ be a graded discrete model poset that is a join semilattice 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 $\hat{x}_{\text{join}}$ 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 $\Psi_{\text{test}}(u, v) \le \alpha$. Then, we have the false discovery bound $\mathbb{P}(\text{FD}(\hat{x}_{\text{join}}, x^\star) > 0) \le \alpha|\mathcal{S}|$.*

5. Experiments

We describe the results of numerical experiments on synthetic and real data in this section. We employ Algorithm 1 with both $\Psi = \Psi_{\text{stable}}$ and $\Psi = \Psi_{\text{test}}$. 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 5 as follows. In the bound Eq. **8**, each q_k can be derived by averaging over subsamples (as explained in the discussion after the statement of Theorem 5), 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 $\hat{x}_{\text{base}}$ employed on the subsamples; in particular, if the estimate is the least element then each q_k equals zero, and as $\hat{x}_{\text{base}}$ returns models of increasing complexity, the value of each q_k generally increases. Building on this observation, we tune parameters in $\hat{x}_{\text{base}}$ to return increasingly more complex models until the bound Eq. **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 (20). For clustering, we employ k-means (21) 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 with the Bradley–Terry model (22), 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 a 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.

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

A.1. 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 $y_{ij\ell} \sim \text{Bernoulli}(w_i^\star/(w_i^\star + w_j^\star))$ 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 Eq. **8** is at most three.

A.2. 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 Eq. **8** is at most three.

A.3. Causal structure learning. We consider a causal structure learning problem over $p = 10$ variables. We generate the true DAG as follows: we obtain an Erdös-Renyi graph with edge probability $v \in \{0.08, 0.16\}$ and then orient the edges according to a random total ordering of the variables. A linear structural causal model is defined where 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 $n \in \{1{,}000, 1{,}200, 1{,}400, 1{,}600, 1{,}800\}$. The base procedure is tuned such that the expected false discovery in Eq. **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 nonsubsampled procedure simply extracts the ranking implied by the maximum-likelihood estimator associated with the Bradley–Terry model. For clustering, the nonsubsampled approach employs k-means where the number of clusters is chosen to maximize the average silhouette score (23). For causal structure learning, the nonsubsampled 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). Fig. 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 nonsubsampled 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.

B. Real Data. We describe the next experiments with real data.

B.1. 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 Eq. **8** is at most three. The output of our procedure is a rank-eight model given by the partial ranking {Djokovic, Nadal} > {Berdych, Murray, Wawrinka} and {Federer} > {Berdych, Wawrinka}.

B.2. 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 ref. 24), with the number of test scores ranging from 696 to 3,414. 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 6 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.

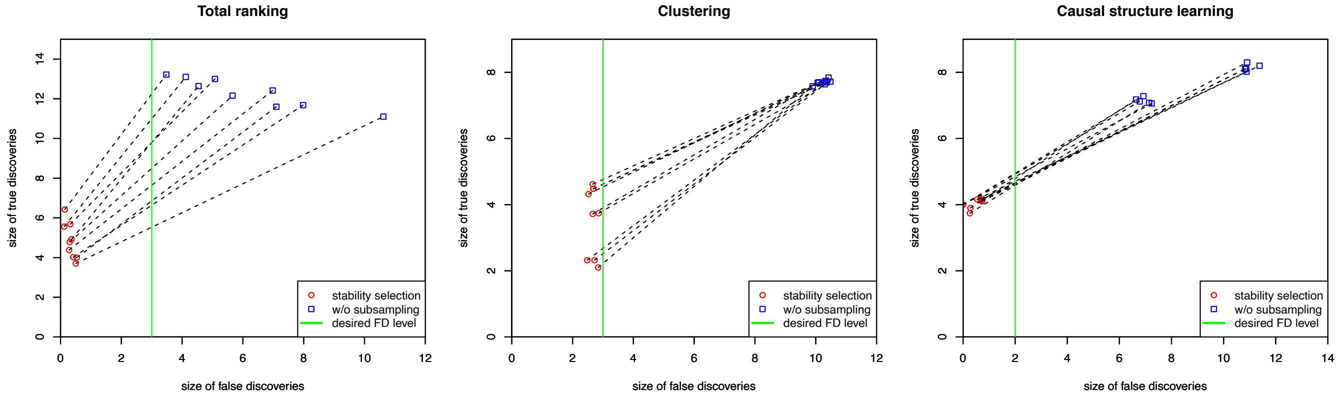


Fig. 2. Comparing the performance of Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ vs. a nonsubsampling approach for total ranking, clustering, and causal structure learning. Each problem setting corresponds to a pair of dots and a connecting line. The comparison is in terms of the amount of false and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

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

	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

B.3. 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 (25); data obtained from the supplementary material of ref. (25). We apply Algorithm 1 with $\Psi = \Psi_{\text{stable}}$ and with the base procedure tuned such that the expected false discovery in Eq. **8** is at most two. Fig. 3 presents the rank-six CPDAG model obtained from our algorithm and compares it to the estimates obtained from the literature (25–27). Our CPDAG estimate has fewer edges than those in refs. 25–27, which do not explicitly provide control on false discovery.

6. Proofs

For notational ease, for a covering pair (u, v) and element z in the poset $\mathcal{L}$, we define $f(u, v; z) := \rho(v, z) - \rho(u, z)$. Recall that $\mathcal{T}_{\text{null}} := \{(u, v) \text{ covering pair in } \mathcal{L} \mid \rho(v, x^\star) = \rho(u, x^\star)\}$. Our analysis relies on the following lemmas with the proofs presented in *SI Appendix*, section 6.

Lemma 8. *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 $x_0 = x_{\text{least}}$ to $x_k = x$, we have that $\text{FD}(x, x^\star) \leq \sum_{i=1}^{k} \mathbb{I}[(x_{i-1}, x_i) \in \mathcal{T}_{\text{null}}]$. As a result, we have that $\text{FD}(x, x^\star) > 0$ implies the existence of some i for which $(x_{i-1}, x_i) \in \mathcal{T}_{\text{null}}$.*

Lemma 9. *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 $z \in \mathcal{L}$.*

A. Proof of Theorem 4.

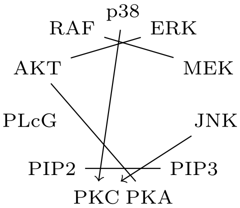
For notational convenience, we let $\hat{x}_{\text{base}}^{(\ell)} = \hat{x}_{\text{base}}(\mathcal{D}^{(\ell)})$ where $\{\mathcal{D}^{(\ell)}\}_{\ell=1}^{B}$ are the subsamples of $\mathcal{D}$. Let $\hat{x}_{\text{stable}}$ be the output of Algorithm 1 with $\text{rank}(\hat{x}_{\text{stable}}) = \hat{k}$, 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 $\frac{1}{B}\sum_{\ell=1}^{B} f(x_{i-1}, x_i; \hat{x}_{\text{base}}^{(\ell)})/c_{\mathcal{L}}(x_{i-1}, x_i) \geq (1 - \alpha)$ for each $i = 1, \ldots, \hat{k}$. Let $\mathcal{C} := \{(x_{i-1}, x_i) \mid i = 1, \ldots, \hat{k}\}$. From Lemma 8, we also have that $\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq |\mathcal{C} \cap \mathcal{T}_{\text{null}}|$. Combining these observations, we conclude that $\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \sum_{(u,v)\in\mathcal{C}\cap\mathcal{T}_{\text{null}}} \mathbb{I}\left[\frac{1}{B}\sum_{\ell=1}^{B} \frac{f(u,v;\hat{x}_{\text{base}}^{(\ell)})}{c_{\mathcal{L}}(u,v)} \geq 1 - \alpha\right]$. 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 9. Thus, we conclude that $\text{FD}(\hat{x}_{\text{stable}}, x^\star) \leq \sum_{(u,v)\in\mathcal{S}\cap\mathcal{T}_{\text{null}}} \mathbb{I}\left[\frac{1}{B}\sum_{\ell=1}^{B} \frac{f(u,v;\hat{x}_{\text{base}}^{(\ell)})}{c_{\mathcal{L}}(u,v)} \geq 1 - \alpha\right]$. Then:

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

The second inequality follows from $ab \geq a + b - 1$ for $a, b \in [0, 1]$, where we set $a = f(u, v; \hat{x}_{\text{base}}^{(2\ell-1)})/c_{\mathcal{L}}(u, v)$ and $b = f(u, v; \hat{x}_{\text{base}}^{(2\ell)})/c_{\mathcal{L}}(u, v)$, 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 $\mathbb{P}\left[\frac{1}{B/2}\sum_{\ell=1}^{B/2}\prod_{i\in\{0,1\}} \frac{f(u,v;\hat{x}_{\text{base}}^{(2\ell-i)})}{c_{\mathcal{L}}(u,v)} \geq 1 - 2\alpha\right]$. We have that

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

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 Eq. **11** and Eq. **12**, we obtain the desired result.



Edge	(25)a	(25)b	(28)	(29)	(26)a	(26)b	(27)
MEK – RAF	←	←	→	→		–	–
AKT – ERK		←	→	←	–	–	–
PIP3 – PIP2	→	→	→	←	→	→	→
AKT – PKA	←	←	←	←		←	→
P38 → PKC	←	←	←	←		–	
JNK → PKC	←	←	←	←	←	–	

Fig. 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 ref. 25 is denoted here by Sachs (25)a and their reconstructed network by Sachs (25)b; The authors in ref. 26 apply two methods, and the results are presented by Meinshausen (26) a and b. Here, "–" means that the edge direction is not identified.

B. Proof of Theorem 5.

We have from Theorem 4 that

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

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

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

Rearranging the terms, we obtain that

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

Appealing to Assumption 2, we have for each $(u,v) \in \mathcal{S}_k \cap \mathcal{T}_{\mathrm{null}}$ that $\frac{\mathbb{E}[f(u,v;\hat{x}_{\mathrm{sub}})]}{c_{\mathcal{L}}(u,v)} \leq \frac{q_k}{|\mathcal{S}_k|}$. Plugging this bound into the conclusion of Theorem 4 yields the desired result.

C. Proof of Theorem 6. Let $\hat{x}_{\mathrm{test}}$ be the output of Algorithm 1 with $\mathrm{rank}(\hat{x}_{\mathrm{test}}) = \hat{k}$, and let $(x_0, \ldots, x_{\hat{k}})$ be the associated path from the least element $x_0 = x_{\mathrm{least}}$ to $x_{\hat{k}} = \hat{x}_{\mathrm{test}}$; we have that $\Psi_{\mathrm{test}}(x_{i-1}, x_i) \leq \alpha$ for each $i = 1, \ldots, \hat{k}$. Let $\mathcal{C} := \{(x_{i-1}, x_i) \mid i = 1, \ldots, \hat{k}\}$. From Lemma 8, we have that $\mathrm{FD}(\hat{x}_{\mathrm{test}}, x^\star) > 0$ implies the existence of a covering pair $(u,v) \in \mathcal{C} \cap \mathcal{T}_{\mathrm{null}}$ for which $\Psi_{\mathrm{test}}(u,v) \leq \alpha$. 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 $(u,v) \in \mathcal{S} \cap \mathcal{T}_{\mathrm{null}}$ such that $\Psi_{\mathrm{test}}(u,v) \leq \alpha$. Consequently:

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

Here, the second inequality follows from the union bound and the final inequality from $\Psi_{\mathrm{test}}(u,v)$ being a valid P-value under the null hypothesis $\rho(v, x^\star) = \rho(u, x^\star)$.

D. Proof of Proposition 7. 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 C in the discussion preceding Proposition 7, $\mathrm{FD}(\hat{x}_{\mathrm{join}}, x^\star) > 0$ implies that $\mathrm{FD}(\hat{x}^{(j)}, x^\star) > 0$ for some $j = 1, \ldots, m$, which in turn implies from Lemma 8 the existence of a covering pair $(u,v) \in \mathcal{C}^{(j)} \cap \mathcal{T}_{\mathrm{null}}$ for some $j = 1, \ldots, m$. Following the same logic as in the proof of Theorem 6, we conclude that $\mathrm{FD}(\hat{x}_{\mathrm{join}}, x^\star) > 0$ implies the existence of $(u,v) \in \mathcal{S} \cap \mathcal{T}_{\mathrm{null}}$ such that $\Psi_{\mathrm{test}}(u,v) \leq \alpha$. Using the same reasoning as in Eq. **13**, we have the desired conclusion.

7. 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 a 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 methods to control false-positive error as well as false discovery rates, including in settings involving continuous model posets.

Data, Materials, and Software Availability. All study data are included in the article. The code for implementing our methods is available at https://github.com/armeentaeb/model-selection-over-posets (30).

ACKNOWLEDGMENTS. We thank Marina Meila and Lior Pachter for insightful conversations. A.T. received funding from the Royalty Research Fund at the University of Washington. P.B. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 786461). V.C. was supported in part by Air Force Office of Scientific Research grants FA9550-22-1-0225 and FA9550-23-1-0204 and by NSF grant DMS 2113724.

1. J. Neyman, E. S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika* **20**, 175–240 (1928).
2. R. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, 1932).
3. K. Popper, *Logik der forschung: Zur erkenntnistheorie der modernen naturwissenschaft* (Springer, 1935).
4. N. Meinshausen, P. Bühlmann, Stability selection. *J. Roy. Stat. Soc. B* **72**, 417–473 (2010).
5. R. Shah, R. Samworth, Variable selection with error control: Another look at stability selection. *J. Roy. Stat. Soc. B* **75**, 55–80 (2013).
6. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974).
7. G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
8. J. J. Goeman, U. R. Mansmann, Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24**, 537–44 (2008).
9. N. Meinshausen, Hierarchical testing of variable importance. *Biometrika* **95**, 265–278 (2008).
10. D. Yekutieli, Hierarchical false discovery rate-controlling methodology. *J. Am. Stat. Assoc.* **103**, 309–316 (2008).
11. R. P. Rosenbaum, Testing hypotheses in order. *Biometrika* **95**, 248–252 (2008).
12. D. P. Foster, R. A. Stine, α-investing: A procedure for sequential control of expected false discoveries. *J. Roy. Stat. Soc. B* **70**, 429–444 (2008).
13. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289–300 (1995).
14. R. P. Stanley, *Enumerative Combinatorics, Volume 1 of Cambridge Studies in Advanced Mathematics* (Cambridge University Press, ed. 2, 2011). 10.1017/CBO9781139058520.
15. M. Drton, M. H. Maathuis, Structure learning in graphical modeling. *Ann. Rev. Stat. Appl.* **4**, 365–393 (2017).
16. T. Verma, J. Pearl, "Equivalence and synthesis of causal models" in *Proceedings of the Sixth Annual Conference of Uncertainty in Artificial Intelligence*, P. Bonissone, M. Henrion, L. Kanal, J. Lemmer, Eds. (1990), pp. 255–270.
17. H. Hotelling, The generalization of student's ratio. *Ann. Math. Stat.* **2**, 54–65 (1931).
18. G. Lorden, Procedures for reacting to a change in distribution. *Ann. Math. Stat.* **42**, 1897–1908 (1971).
19. A. Taeb, P. Shah, V. Chandrasekaran, False discovery and its control in low-rank estimation. *J. Roy. Stat. Soc. B* **92**, 997–1027 (2020).
20. D. M. Chickering, Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3**, 507–554 (2002).
21. S. P. Lloyd, Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–136 (1982).
22. R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324 (1952).
23. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
24. K. Wang, Learning tower. GitHub Repository. https://github.com/kevinwang09/learningtower. Accessed 20 August 2023.
25. K. Sachs, O. Perez, D. Lauffenburger, G. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. https://www.science.org/doi/10.1126/science.1105809#supplementary-materials. Accessed 20 August 2023.
26. N. Meinshausen *et al.*, Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7361–7368 (2016).
27. K. D. Yang, A. Katoff, C. Uhler, "Characterizing and learning equivalence classes of causal DAGs under interventions" in *Proceedings of the Thirty-Fifth International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (2018).
28. J. Mooij, T. Heskes, "Cyclic causal discovery from continuous equilibrium data" in *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence*, A. Nicholson, P. Smyth, Eds. (2013), pp. 431–439.
29. D. Eaton, K. Murphy, "Exact Bayesian structure learning from uncertain interventions" in *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2007).
30. A. Taeb, P. Bühlmann, V. Chandrasekaran, Model selection over posets. GitHub Repository. https://github.com/armeentaeb/model-selection-over-posets. Deposited 20 August 2023.