Web Appendix A: Proofs of results and technical lemmas

Lemma A1 describes the atoms obtained from Algorithm 1:

LEMMA A1: *By using Algorithm 1, a variable m is assigned to an atom $A(m)$, which is defined as:*

$$A(m) = \bigcap_{S \in \mathcal{S}_m} S \setminus \Big(\bigcup_{S \notin \mathcal{S}_m} S \Big),$$

where $\mathcal{S}_m$ is the collection of sets among $S_1, \ldots, S_K$ which contain m.

THEOREM A1: *The atoms resulting from the steps of Algorithm 1 uniquely satisfy properties 1-3 for the units of a collection of sets.*

Proof. We first note that property 2 is satisfied by construction, since every time a new atom is created using Algorithm 1, the variables in it are removed from $\mathcal{M}_A$.

Proof that property 1 holds. Let S be one of the original sets. We want to show that we can write $S = \bigcup_{\ell \in \mathcal{A}} A_\ell$, for some $\mathcal{A} \subset \{1, \ldots, L\}$. Denote the variables in set S by $m_1, \ldots, m_R$. Each variable m_r will be mapped to an atom $A(m_r)$ by the algorithm (note that the atoms may not necessarily be unique). We now show that $A(m_1) \bigcup \ldots \bigcup A(m_R) \subset S$:

Let $m \in A(m_1) \bigcup \ldots \bigcup A(m_R)$. Then the variable m is part of some atom $A(m_r)$, according to the algorithm, for a variable m_r. Since m and m_r are in the same atom, they appear in exactly the same sets, according to the atom construction in the algorithm, so in particular $m \in S$, since $m_r \in S$. Thus, $A(m_1) \bigcup \ldots \bigcup A(m_R) \subset S$. Since $S \subset A(m_1) \bigcup \ldots \bigcup A(m_R)$ trivially, we have shown that $S = A(m_1) \bigcup \ldots \bigcup A(m_R)$. Taking only the unique atoms, this ensures that property 1 holds.

Proof that property 3 and uniqueness hold. Let $\{A_1, \ldots, A_L\}$ be the collection of atoms produced by the algorithm. Let $\{B_1, \ldots, B_N\}$ be a different collection of atoms which satisfies properties 1 and 2. We prove that this collection must then have a higher cardinality than

$\{A_1, \ldots, A_L\}$, i.e. $K > L$, which shows that $\{A_1, \ldots, A_L\}$ fulfills property 3, and is also the only collection of atoms to fulfill properties 1-3.

Consider an atom B_n. We show that it must be a subset of some atom A_l, i.e. there exists $1 \leqslant l \leqslant L$ such that $B_n \subset A_l$. We look at the collection of all sets that atom B_n is part of, i.e. $\mathcal{S}_{B_n} = \{S : B_n \subset S\}$. We note that we must have $B_n \subset C_{B_n}$, where we define $C_{B_n} = \bigcap_{S \in \mathcal{S}_{B_n}} S \setminus \bigcup_{S \notin \mathcal{S}_{B_n}} S$. There is no set S out of the original sets such that $S \subset C_{B_n}$: If such a set S existed, and $B_n \setminus S \neq \emptyset$, B_n would not be an atom (since S can be written as a union of atoms in $\{B_1, \ldots, B_N\}$, and all these atoms are disjoint); and if such a set existed and $B_n \setminus S = \emptyset$, there would be a contradiction with the definition of C_{B_n}. As a result, the variables which are in C_{B_n} all belong to exactly the same original sets. This means that they form an atom obtained through Algorithm 1, i.e. $C_{B_n} = A_l$, for some $1 \leqslant l \leqslant L$. As a result, $B_n \subset A_l$. Since this holds for any n, we have shown that the collection $K > L$, so we have proven that property 3 holds, and also that uniqueness holds.

Proof. [Proof of Eq. (3)] Consider a set U which is a union of atoms. Since $U \setminus \tau = \bigcup_{A_l \subset U} (A_l \setminus \tau)$ and all $A_l \setminus \tau$ are disjoint, then:

$$\sum_{m \in U \setminus \tau} d(m, \tau) = \sum_{A_l \subset U} \sum_{m \in A_l \setminus \tau} d(m, \tau)$$

Since the set of variables is completely partitioned by the set of atoms (i.e. all variables are annotated to some set, and therefore to some atom), then $\tau \setminus U = \tau \cap \left[\bigcup_{A_l \not\subset U} A_l\right] = \bigcup_{A_l \not\subset U} (\tau \cap A_l)$, so by the disjointness of $\tau \cap A_l$:

$$\sum_{m \in \tau \setminus U} d(m, U) = \sum_{A_l \not\subset U} \sum_{m \in \tau \cap A_l} d(m, U).$$

Thus, we obtain:

$$(1 - w) \sum_{A_l \subset U} \sum_{m \in A_l \setminus \tau} d(m, \tau) + w \sum_{A_l \not\subset U} \sum_{m \in \tau \cap A_l} d(m, U).$$

We introduce the following notation: We start by defining the posterior probability that a set U is a proper subset of the set of interesting variables τ.

$$P(U \in \tau|\mathbf{X},\mathbf{Y}) \equiv p^*_U = \sum_{\tau\in 2^{\mathcal{M}}, U\subset\tau} p_\tau.$$

For specific cases we can simply write out the variables in the set, e.g. $p_{12} = p_{21} = p_{\{1,2\}}$. The marginal posterior probabilities for each variable are a specific case, where the set represents a single variable: $p^*_m = \sum_{\tau\in 2^{\mathcal{M}}, m\in\tau} p_\tau$.

In Lemma A2 we derive a simplified form of the two components of the posterior expected loss function. Note, in particular, that $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in\tau\setminus U} d(m,U)\}$ can be written as a linear function of marginal variable-level posterior probabilities.

LEMMA A2: *Under the loss function described in equation (2), in the case of a general discrepancy measure d and single linkage, the following simplified forms of* $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ *and* $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in\tau\setminus U} d(m,U)\}$ *are obtained:*

$$E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\} = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in U\setminus\tau} d(m,\tau)p_\tau = \sum_{m\in U}\sum_{\tau\in 2^{\mathcal{M}}, m\notin\tau} d(m,\tau)p_\tau$$
$$E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in\tau\setminus U} d(m,U)\} = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in\tau\setminus U} d(m,U)p_\tau = \sum_{m\notin U} d(m,U)p^*_m$$

Proof.

$$\sum_{m\in U}\sum_{\tau\in 2^{\mathcal{M}}, m\notin\tau} d(m,\tau)p_\tau = \sum_{\tau\in 2^{\mathcal{M}}, m\in U\setminus\tau} d(m,\tau)p_\tau = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in U\setminus\tau} d(m,\tau)p_\tau$$
$$\sum_{m\notin U} d(m,U)p^*_m = \sum_{m\notin U} d(m,U)\sum_{\tau\in 2^{\mathcal{M}}, m\in\tau} p_\tau = \sum_{m\in\tau\setminus U,\tau\in 2^{\mathcal{M}}} d(m,U)p_\tau$$
$$= \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in\tau\setminus U} d(m,U)p_\tau$$

Proof. [Proof of Theorem 1] Using Lemma A2:

$$\mathrm{EFD}(U) = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in U\setminus\tau} p_\tau = \sum_{m\in U}\sum_{\tau\in 2^{\mathcal{M}}, m\notin\tau} p_\tau = \sum_{m\in U}(1-p^*_m) = |U| - \sum_{m\in U} p^*_m$$
$$\mathrm{EMD}(U) = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in\tau\setminus U} p_\tau = \sum_{m\notin U} p^*_m$$

Proof. [Proof of Corollary 1] By the mixture model in Equation (1), when considered as functions of the data, the marginal variable-level posterior probabilities can be rewritten as:

$$\begin{aligned} p_m^* &= P(\text{variable } m \text{ is from the alternative distribution}|\mathbf{X}, \mathbf{Y}) \\ &= P(\text{variable } m \text{ is from the alternative distribution}|\mathbf{X}_m, \mathbf{Y}) \\ &= 1 - P(\text{variable } m \text{ is from the null distribution}|T_m, \mathbf{Y}) \\ &= 1 - \pi_0 \frac{f_0(T_m|\mathbf{Y})}{f(T_m|\mathbf{Y})} \\ &= 1 - \text{fdr}(T_m|\mathbf{Y}) \end{aligned}$$

Thus, the results of Theorem 1 may be rewritten as:

$$\begin{aligned} \text{EFD}(U) &= \sum_{m \in U} \text{fdr}(T_m|\mathbf{Y}_m) \\ \text{EMD}(U) &= \sum_{m \notin U} [1 - \text{fdr}(T_m|\mathbf{Y}_m)] \end{aligned}$$

In Lemma A3, we show how the posterior expected loss can be written in terms of the expected numbers of false discoveries for each atom, for the 0-1 discrepancy measure. When written in this form, it is possible to show that optimizing the posterior expected loss is equivalent to the well-known "knapsack" problem in computer science (Garey and Johnson, 1979). Using this result, we show in Theorem 2 that the Bayes estimator for a fixed value of the weight w is given by thresholding the atomic false discovery rate.

LEMMA A3: *The posterior expected loss $\mathcal{L}(U)$ which results from the $0-1$ dissimilarity measure may be rewritten as:*

$$\sum_{l=1}^{L} \delta_l[(1-w)EFD(A_l) - w\{n_l - EFD(A_l)\}] + w\sum_{l=1}^{L}\{n_l - EFD(A_l)\},$$

where δ_l is the indicator of whether atom A_l is in U and n_l is the number of variables in A_l,

i.e. $n_l = |A_l|$.

Using vector notation with $\boldsymbol{\delta} = [\delta_l]_{l=1}^L$, $\mathbf{1} = [1]_{l=1}^L$, $\mathbf{n} = [n_l]_{l=1}^L$, *and* $\mathbf{EFD}_A = [EFD(A_l)]_{l=1}^L$, *this is the same as:*

$$\boldsymbol{\delta}' \{(1-w)\mathbf{EFD}_A - w(\mathbf{n} - \mathbf{EFD}_A)\} + w(\mathbf{n} - \mathbf{EFD}_A).$$

Proof.

$$\begin{aligned}
\mathcal{L}(U) &= (1-w) \sum_{T \in 2^{\mathcal{M}}} \sum_{A_l \subset U} \sum_{m \in A_l \setminus T} p_T + w \sum_{T \in 2^{\mathcal{M}}} \sum_{A_l \not\subset U} \sum_{m \in T \cap A_l} d(m,U) p_T \\
&= (1-w) \sum_{A_l \subset U} \sum_{T \in 2^{\mathcal{M}}} \sum_{m \in A_l \setminus T} p_T + w \sum_{A_l \not\subset U} \sum_{T \in 2^{\mathcal{M}}} \sum_{m \in T \cap A_l} p_T \\
&= (1-w) \sum_{A_l \subset U} \sum_{m \in A_l} (1 - p_m^*) + w \sum_{A_l \not\subset U} \sum_{m \in A_l} p_m^* \\
&= (1-w) \sum_{A_l \subset U} EFD(A_l) + w \sum_{A_l \not\subset U} \{n_l - EFD(A_l)\} \\
&= (1-w) \sum_{l=1}^{L} \delta_l EFD(A_l) + w \sum_{l=1}^{L} (1-\delta_l)\{n_l - EFD(A_l)\} \\
&= \sum_{l=1}^{L} \delta_l [(1-w) EFD(A_l) - w\{n_l - EFD(A_l)\}] + w \sum_{l=1}^{L} \{n_l - EFD(A_l)\},
\end{aligned}$$

using Lemma A2 and Theorem 1.

Thus, the parametrization is linear in $\boldsymbol{\delta}$. In Lemma A4, we show that similar parametrizations exist for the posterior expected loss functions $\mathcal{L}_f^\lambda$ and $\mathcal{L}_a^\xi$, which correspond to the loss functions $L_f^\lambda(\tau, U)$ and $L_a^\xi(\tau, U)$ (from Web Appendix C).

LEMMA A4: *Up to an additive constant,* $\mathcal{L}$ *can be written as:*

$$\sum_l \delta_l [(1-w) EFD(A_l) - w c_2(A_l)\{n_l - EFD(A_l)\}] + w \sum_l \{n_l - EFD(A_l)\}.$$

$\mathcal{L}$ *corresponds to* $c_1(A_l) = c_2(A_l) = 1$; $\mathcal{L}_f^\lambda$ *corresponds to* $c_1(A_l) = \frac{1-w+\lambda}{1-w}$ *and* $c_2(A_l) = \frac{w-\lambda}{w}$; $\mathcal{L}_a^\xi$ *corresponds to* $c_1(A_l) = \frac{1-w+\frac{\xi}{n_l}}{1-w}$ *and* $c_2(A_l) = \frac{w-\frac{\xi}{n_l}}{w}$.

Proof. [Proof of Theorem 2] The posterior expected loss of L can be parametrized as an affine function of $\boldsymbol{\delta}$, for a fixed w between 0 and 1. Any affine function of δ_l, $h(\boldsymbol{\delta}) = \boldsymbol{\delta}'\mathbf{a} + b, \boldsymbol{\delta} \in$

$\{0,1\}^L, \mathbf{a} \in \mathbb{R}^L, b \in \mathbb{R}$ is minimized when $\delta_l = 1\{a_l \leqslant 0\}$, since $h(t)$ is minimized when $\boldsymbol{\delta}'\mathbf{a}$ is minimized. This is a linear function in each component δ_j of $\boldsymbol{\delta}$, and if we minimize it in each component we also minimize it overall. As a result, it is minimized by choosing to sum only over those components of $\mathbf{a}$ which are negative or zero.

The proof of Lemma C1 is similar to that of Theorem 2, since the posterior expected losses of L_f^λ, and L_a^ξ can also be parametrized as affine functions of $\boldsymbol{\delta}$, for a fixed w between 0 and 1.

Proof. [Proof of Theorem C1]

$$\begin{aligned}
\mathcal{L}_r(\boldsymbol{\delta}) &= (1-w)\frac{\sum_{m\in U}(1-p_m^*)}{|U|} + w\frac{\sum_{n\notin U} p_m^*}{M-|U|} \\
&= (1-w)\frac{\sum_{A_l\in U}\mathrm{EFD}(A_l)}{\sum_{A_j\in U} n_l} + w\frac{\sum_{A_l\notin U}(n_l - \mathrm{EFD}(A_l))}{\sum_{A_l\notin U} n_l} \\
&= (1-w)\frac{\boldsymbol{\delta}'\mathbf{EFD}_A}{\boldsymbol{\delta}'\mathbf{n}} + w\frac{(\mathbf{1}-\boldsymbol{\delta})'(\mathbf{n}-\mathbf{EFD}_A)}{M-\boldsymbol{\delta}'\mathbf{n}} \\
&= \boldsymbol{\delta}'\left\{\frac{(1-w)}{\boldsymbol{\delta}'\mathbf{n}}\mathbf{EFD}_A - \frac{w}{(M-\boldsymbol{\delta}'\mathbf{n})}(\mathbf{n}-\mathbf{EFD}_A)\right\} + \frac{w}{M-\boldsymbol{\delta}'\mathbf{n}}(\mathbf{n}-\mathbf{EFD}_A)
\end{aligned}$$

We now show that $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ cannot be written as an affine function of marginal variable-level posterior probabilities for a general discrepancy measure d which takes into account how far or close variables are to each other, in the case where the single linkage property holds. Thus, in general, to calculate the posterior expected loss, we need to model the joint distribution of all the variables. This also leads to much more complex computations.

COROLLARY A1: *Under the loss function described in equation (2), in the case of a general discrepancy measure d and single linkage,* $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ *cannot be written as an affine function of marginal variable-level posterior probabilities. Therefore* $\mathcal{L}(U)$ *also cannot be written as an affine function of marginal variable-level posterior probabilities.*

Proof. We show that we cannot write $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ as an affine function of the marginal variable-level posterior probabilities.

Step 1. We first show that, for any proper subset $\nu \subsetneq 2^{\mathcal{M}}$, setting any affine of the posterior probabilities of the sets in ν equal to 0 forces all the coefficients to be 0, i.e.:

Denote the elements in ν by $\tau_{l_1},\ldots,\tau_{l_{|\nu|}}$. We will show that setting any affine function of these elements to 0 implies that all the coefficients are 0. Thus, we have:

$$\sum_{\tau\in\nu} a_\tau p_\tau + b = 0 \tag{1}$$

We note that if $\nu = \{\tau_{l_1},\ldots,\tau_{l_{|\nu|}}\} \subsetneq 2^{\mathcal{M}}$, then $p_{\tau_{l_1}}+\ldots+p_{\tau_{l_{|\nu|}}} \leqslant 1$ and $p_{\tau_{l_1}} \geqslant 0,\ldots,p_{\tau_{l_{|\nu|}}} \geqslant 0$. Plugging in $p_{\tau_{l_1}} = 1, p_{\tau_{l_2}} = \ldots = p_{\tau_{l_{|\nu|}}} = 0$, followed by $p_{\tau_{l_1}} = \frac{1}{2}, p_{\tau_{l_2}} = \ldots = p_{\tau_{l_{|\nu|}}} = 0$, and solving the resulting system of equations in $a_{\tau_{l_1}}$ and b results in $a_{\tau_{l_1}} = b = 0$. From here on, plugging in only one non-zero probability for each $\tau \in \nu$ in turn will result in $a_{\tau_{l_2}} = \ldots = a_{\tau_{l_{|\nu|}}} = 0$.

Step 2. We now apply the result in *Step 1* to show that $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ can in general not be written as an affine function of the marginal variable-level posterior probabilities. We note that we have:

$$E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\} = \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\in U\setminus\tau} d(m,\tau)p_\tau = \sum_{\tau\in 2^{\mathcal{M}},\tau\neq U}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}p_\tau \tag{2}$$

since $d(m,\tau) = 0$ if $m \in \tau$. Using a simple transformation, we note that showing that $E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ we need to show that we can find a_m and b such that:

$$\begin{aligned} E_{\tau|\mathbf{X},\mathbf{Y}}\{\sum_{m\in U\setminus\tau} d(m,\tau)\} &= \sum_{m\in\mathcal{M}} a_m(1-p^*_m) + b \\ &= \sum_{m\in\mathcal{M}} a_m \sum_{\tau\in 2^{\mathcal{M}},m\notin\tau} p_\tau + b \\ &= \sum_{\tau\in 2^{\mathcal{M}}}\sum_{m\notin\mathcal{M}\setminus\tau} a_m p_\tau + b \\ &= \sum_{\tau\in 2^{\mathcal{M}},\tau\neq\mathcal{M}}\{\sum_{m\in\mathcal{M}\setminus\tau} a_m\}p_\tau + b \end{aligned} \tag{3}$$

The coefficients a_m and b are more accurately written as $a_m(U)$ and $b(U)$, but we use the

simpler notation here. Setting the expressions in 2 and 3 equal to each other, we get:

$$\sum_{\tau \in 2^{\mathcal{M}}, \tau \neq U} \{ \sum_{m \in U \setminus \tau} d(m, \tau) \} p_\tau = \sum_{\tau \in 2^{\mathcal{M}}, \tau \neq \mathcal{M}} \{ \sum_{m \in \mathcal{M} \setminus \tau} a_m \} p_\tau + b$$

We may now take $\nu = 2^{\mathcal{M}} \setminus \mathcal{M}$, which is a proper subset of $2^{\mathcal{M}}$. Using the result in *Step 1*, we get:

$$\sum_{m \in U \setminus \tau} a_m \quad = \quad \sum_{m \in U \setminus \tau} d(m, \tau),$$

all the other coefficients being 0. Now consider cycling through all the sets τ such that $U \setminus \tau$ consists of a single element. We thus obtain:

$$d(m, \tau) = a_m \text{ for all } m \in U \setminus \tau$$

regardless of how many elements there are in $\tau \setminus U$ and how far away they are from the elements in $U \setminus \tau$.

To illustrate this last portion of the proof, consider $\mathcal{M} = \{1, 2, 3\}$ and $U = \{1, 2\}$. Then:

$$\tau = \{2\} \quad => \quad d(1, \{2\}) = a_1$$

$$\tau = \{2, 3\} \quad => \quad d(1, \{2, 3\}) = a_1$$

Given our use of the single-linkage property, $d(1, \{2, 3\}) = min\{d(1, \{2\}), d(1, \{3\})\}$. So if $d(1, \{3\}) \geqslant d(1, \{2\})$, then $d(1, \{2\}) = d(1, \{3\})$, which means that the discrepancy measure d does not take into account how far or close variables are to each other.

References

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability: a guide to NP-completeness*. WH Freeman and Company, San Francisco.