Learning Hierarchical Riffle Independent Groupings from Rankings: Supplemental Material

Jonathan Huang JCH1@CS.CMU.EDU
CMU

Carlos Guestrin JCH1@CS.CMU.EDU
CMU

Shared independence structure

While Figures 1(a) and 1(b) (on the next page) encode distinct families of distributions, they share a common subset of independence assumptions. It turns out that any distributions consistent with either of the two hierarchies must also be consistent with what we call a 3-*way decomposition*. We define a d-way decomposition to be a distribution with a single level of hierarchy, but instead of partitioning the entire item set into just two subsets, one partitions into d subsets, then interleaves the relative rankings of each of the d subsets together to form a joint ranking of items. Any distribution consistent with either Figure 1(b) or 1(a) must also be consistent with the structure of Figure 1(c). More generally, we have:

Proposition 1. *If h is a hierarchical riffle independent model with d leaf sets, then h can also be written as a d-way decomposition.*

Proof. We proceed by induction. Suppose the result holds for $S_{n'}$ for all $n' < n$. We want to establish that the result also holds for S_n. If h factors according to a hierarchical riffle independent model, then it can be written as $h = m \cdot f_A \cdot f_B$, where m is the interleaving distribution, and f_A, f_B themselves factor as hierarchical riffle independent distributions with, say, d_1 and d_2 leaf sets, respectively (where $d_1 + d_2 = d$). By the hypothesis, since $|A|, |B| < n$, we can factor both f_A and g_B as d_1 and d_2-way decompositions respectively. We can therefore write f_A and g_B as:

$$f_A(\pi_A) = m_A(\tau_{A_1,\ldots,A_{d_1}}) \cdot \prod_{i=1}^{d_1} f_{A_i}\left(\phi_{A_i}(\pi_A)\right)$$

$$g_B(\pi_B) = m_B(\tau_{B_1,\ldots,B_{d_2}}) \cdot \prod_{i=1}^{d_2} g_{B_i}\left(\phi_{B_i}(\pi_B)\right)$$

Substituting these decompositions into the factorization of the distribution h, we have:

$$\begin{aligned} h(\sigma) &= m(\tau_{A,B}(\sigma)) f_A(\phi_A(\sigma)) g_B(\phi_B(\sigma)) \\ &= \left(m(\tau_{A,B}(\sigma)) m_A(\tau_{A_1\ldots,A_{d_1}}) m_B(\tau_{B_1,\ldots,B_{d_2}})\right) \\ &\quad \cdot \prod_{i=1}^{d_1} f_{A_i}\left(\phi_{A_i}(\phi_A(\sigma))\right) \prod_{i=1}^{d_2} g_{B_i}\left(\phi_{B_i}(\phi_B(\sigma))\right) \\ &= \tilde{m}(\tau_{A_1,\ldots,A_{d_1},B_1\ldots,B_{d_2}}) \\ &\quad \cdot \prod_{i=1}^{d_1} f_{A_i}\left(\phi_{A_i}(\sigma)\right) \prod_{i=1}^{d_2} g_{B_i}\left(\phi_{B_i}(\sigma)\right), \end{aligned}$$

where the last line follows because any legitimate interleaving of the sets A and B is also a legitimate interleaving of the sets $A_1, \ldots, A_{d_1}, B_1, \ldots, B_{d_2}$ and since $\phi_{A_i}(\phi_A(\sigma)) = \phi_{A_i}(\sigma)$. This shows that the distribution h factors as a $d_1 + d_2$-way decomposition, and concludes the proof. □

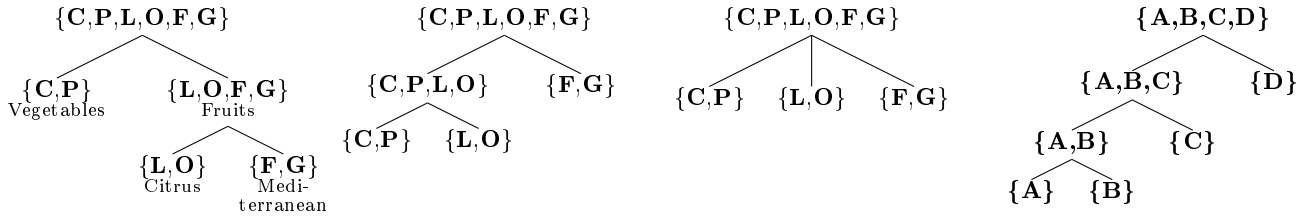
Riffled independence criterion

Our objective is defined as:

$$\mathcal{F}(A) \equiv I(\sigma(A)\ ;\ \phi_B(\sigma)) + I(\sigma(B)\ ;\ \phi_A(\sigma)), \quad (0.1)$$

Proposition 2. *$\mathcal{F}(A) = 0$ is a necessary and sufficient criterion for a subset $A \subset \{1, \ldots, n\}$ to be riffle independent of its complement, B.*

Proof. Suppose A and B are riffle independent. We first claim that $\sigma(A)$ and $\phi_B(\sigma)$ are independent. To see this, observe that the absolute ranks of A, $\sigma(A)$, are determined by the relative rankings of A, $\phi_A(\sigma)$ and the interleaving $\tau_{A,B}(\sigma)$. By the assumption that A and B are riffle independent, we know that the relative rankings of A and B ($\phi_A(\sigma)$ and $\phi_B(\sigma)$), and the interleaving $\tau_{A,B}(\sigma)$ are independent, establishing the claim. The argument that $\sigma(B)$ and $\phi_A(\sigma)$ are independent is similar, thus establishing one direction of the proposition.



(a) Example of hierarchical riffled independence structure on S_6

(b) Another example, not equivalent to (a)

(c) 3-way decomposition for S_6 (generalizes the class of distributions parameterized by (a), (b)

(d) Hierarchical decomposition into singleton subsets (each leaf set consists of a single item)

Figure 1. Examples of distinct hierarchical riffle independent structures

To establish the reverse direction, assume that Equation 0.1 evaluates to zero on sets A and B. It follows that $\sigma(A) \perp \phi_B(\sigma)$ and $\phi_A(\sigma) \perp \sigma(B)$. Now, as a converse to the observation from above, note that the absolute ranks of A *determine* the relative ranks of A, $\phi_A(\sigma)$, as well as the interleaving $\tau_{A,B}(\sigma)$. Similarly, $\sigma(B)$ determines $\phi_B(\sigma)$ and $\tau_{A,B}(\sigma)$. Thus, $(\phi_A(\sigma), \tau_{A,B}(\sigma)) \perp \phi_B(\sigma)$ and $\phi_A(\sigma) \perp (\tau_{A,B}(\sigma), \phi_B(\sigma))$. It then follows that $\phi_A(\sigma) \perp \tau_{A,B}(\sigma) \perp \phi_B(\sigma)$. □

Sample complexity analysis

Lemma 3 (adapted from (Höffgen, 1993))**.** *The entropy of a discrete random variable with arity R can be estimated to within accuracy Δ with probability $1-\beta$ using $O\left(\frac{R^2}{\Delta^2}\log^2\frac{R}{\Delta}\log\frac{R}{\beta}\right)$ i.i.d samples and the same time.*

Lemma 4. *The collection of mutual informations $I_{i;j,k}$ can be estimated to within accuracy Δ for all triplets (i,j,k) with probability at least $1-\gamma$ using $S(\Delta,\gamma) \equiv O\left(\frac{n^2}{\Delta^2}\log^2\frac{n}{\Delta}\log\frac{n^4}{\gamma}\right)$ i.i.d. samples and the same amount of time.*

Proof. Fix a $0 < \gamma \leq 1$ and Δ. For any fixed triplet (i,j,k), Hoffgen's result (Lemma 3) implies that $H(\sigma_i; \sigma_j < \sigma_k)$ can be estimated with accuracy Δ with probability at least $1-\gamma/n^3$ using $O\left(\frac{n^2}{\Delta^2}\log^2\frac{n}{\Delta}\log\frac{n^4}{\gamma}\right)$ i.i.d. samples since the variable $(\sigma_i, \sigma_j < \sigma_k)$ has arity $2n$ and setting $\beta \equiv \frac{\gamma}{n^3}$.

Estimating the mutual information for the same triplet therefore requires the same sample complexity by the expansion: $I_{i;j,k} = H(\sigma_i) + H(\sigma_j < \sigma_k) - H(\sigma_i; \sigma_j < \sigma_k)$. Now we use a simple union bound to bound the probability that the collection of mutual informations over all triplets is estimated to within Δ accuracy. Define $\Delta_{i,j,k} \equiv I_{i;j,k} - \hat{I}_{i;j,k}$.

$$\begin{aligned} P(|\Delta_{i,j,k}| < \Delta,\ \forall(i,j,k)) &\geq 1 - \sum_{i,j,k} P(|\Delta_{i,j,k}| \geq \Delta), \\ &\geq 1 - n^3 \cdot \frac{\gamma}{n^3}, \\ &\geq 1 - \gamma. \end{aligned}$$

□

Lemma 5. *Fix $k \leq n/2$. and let A be a k-subset of $\{1,\ldots,n\}$ with A riffle independent of its complement B. Let A' be a k-subset with $A' \neq A$ or B. If A and B are each ϵ-third order strongly connected, we have $\tilde{\mathcal{F}}(A') = \tilde{\mathcal{F}}(B') > \psi(n,k)\cdot\epsilon$, where $\psi(n,k) \equiv (n-k)(n-2k)$.*

Proof. Let us first establish some notation. Given a subset $X \subset \{1,\ldots,n\}$, define

$$\Omega_X^{int} \equiv \{(x;y,z) : x,y,z \in X\}.$$

Thus Ω_A^{int} and Ω_B^{int} are the sets of triplets whose indices are all internal to A or internal to B respectively. We define $\Omega_{A',B'}^{cross}$ to be the set of triplets which "cross" between the sets A and B:

$$\Omega_{A',B'}^{cross} \equiv \{(x;y,z) : x \in A, y,z \in B, \text{ or } x \in B, y,z \in A\}.$$

The goal of this proof is to use the strong connectivity assumptions to lower bound $\tilde{\mathcal{F}}(A')$. In particular, due to strong connectivity, each triplet inside $\Omega_{A',B'}^{cross}$ that also lies in either Ω_A^{int} or Ω_B^{int} must contribute at least ϵ to the objective function $\tilde{\mathcal{F}}(A')$. It therefore suffices to lower bound the number of triplets which cross between A' and B', but are internal to either A or B (i.e., $|\Omega_{A',B'}^{cross} \cap (\Omega_A^{int} \cup \Omega_B^{int})|$). Define $\ell \equiv |A \cap A'|$ and note that $0 \leq \ell < k$. It is straightforward to check that: $|A \cap B'| = k - \ell$, $|B \cap A'| = k - \ell$, and $|B \cap B'| = (n-k) - (k-\ell) = n + \ell - 2k$.

$$\begin{aligned}|\Omega^{cross}_{A',B'} \cap (\Omega^{int}_A \cup \Omega^{int}_B)| &= |\Omega^{cross}_{A',B'} \cap \Omega^{int}_A| + |\Omega^{cross}_{A',B'} \cap \Omega^{int}_B| \\ &\geq \ell(k-\ell)^2 + \ell^2(k-\ell) \\ &\quad + (k-\ell)(n+\ell-2k)^2 \\ &\quad + (n+\ell-2k)(k-\ell)^2 \\ &\geq (k-\ell)\left((n-k)(n-2k) + \ell n\right) \\ &\geq k\left((n-k)(n-2k) + kn\right).\end{aligned}$$

We do want the bound above to depend on ℓ. Intuitively, for a fixed k and n, the above expression is minimized when either $\ell = 0$ or $k-1$ (a more formal argument is shown below in the proof of Lemma 6). Plugging $\ell = 0$ and $k-1$ and bounding from below yields:

$$\begin{aligned}|\Omega^{cross}_{A',B'} \cap (\Omega^{int}_A \cup \Omega^{int}_B)| &\geq \min\left(k(n-k)(n-2k),\right. \\ &\qquad \left.(n-k)(n-2k) + n(k-1)\right]) \\ &\geq (n-k)(n-2k).\end{aligned}$$

Finally due to strong connectivity, we know that for each triplet in $\Omega^{int}_A \cup \Omega^{int}_B$, we have $I_{x;y,z} > \epsilon$, thus each edge in $\Omega^{cross}_{A',B'} \cap (\Omega^{int}_A \cup \Omega^{int}_B)$ contributes at least ϵ to $\tilde{\mathcal{F}}(A')$, establishing the desired result. □

Lemma 6. *Under the same assumptions as Lemma 5, $p(n,k,\ell) = (k-\ell)\left((n-k)(n-2k) + \ell n\right)$ is minimized at either $\ell = 0$ or $k-1$.*

Proof. Let $\alpha = (n-k)(n-2k)$. We know that $\alpha \geq 0$ since $k \leq n/2$ by assumption (and equals zero only when $k = n/2$). We want to find the $\ell \in \{0,\ldots,k-1\}$ which minimizes the concave quadratic function $p(\ell) = (k-\ell)(\alpha + \ell n)$, the roots of which are $\ell = k$ and $\ell = -\alpha/n$ (note that $-\alpha/n \leq 0$. The minimizer is thus the element of $\{0,\ldots,k-1\}$ which is closest to either of the roots. □

Theorem 7. *Let A be a k-subset of $\{1,\ldots,n\}$ with A riffle independent of its complement B. If A and B are each ϵ-third order strongly connected, then given $S(\Delta,\epsilon) \equiv O\left(\frac{n^4}{\epsilon^2}\log^2\frac{n^2}{\epsilon}\log\frac{n^4}{\gamma}\right)$ i.i.d. samples, the minimum of $\hat{\mathcal{F}}$ (evaluated over all k-subsets of $\{1,\ldots,n\}$) is achieved at exactly the subsets A and B with probability at least $1-\gamma$.*

Proof. Let A' be a k-subset with $A' \neq A$ or B. Our goal is to show that $\hat{\mathcal{F}}(A') > \hat{\mathcal{F}}(A)$.

Denote the error between estimated mutual information and true mutual information by $\Delta_{i;j,k} \equiv \hat{I}_{i;j,k} - I_{i;j,k}$. We have:

$$\begin{aligned}\hat{\mathcal{F}}(A') - \hat{\mathcal{F}}(A) &= \left(\sum_{(i,j,k)\in\Omega^{cross}_{A',B'}} \hat{I}_{i;j,k}\right) - \left(\sum_{(i,j,k)\in\Omega^{cross}_{A,B}} \hat{I}_{i;j,k}\right) \\ &= \tilde{\mathcal{F}}(A') - \tilde{\mathcal{F}}(A) + \sum_{(i,j,k)\in\Omega^{cross}_{A',B'}} \Delta_{i;j,k} \\ &\quad - \sum_{(i,j,k)\in\Omega^{cross}_{A,B}} \Delta_{i;j,k} \\ &\geq \psi(n,k)\cdot\epsilon + \sum_{(i,j,k)\in\Omega^{cross}_{A',B'}} \Delta_{i;j,k} \\ &\quad - \sum_{(i,j,k)\in\Omega^{cross}_{A,B}} \Delta_{i;j,k} \\ &\qquad \text{(by Lemma 5 and } \tilde{\mathcal{F}}(A) = 0)\end{aligned}$$

Now suppose assume that all of the estimation errors Δ are uniformly bounded as:

$$|\Delta_{i;j,k}| \leq \frac{\epsilon}{4}\left(\frac{\psi(n,k)}{n^2k - k^2n}\right). \tag{0.2}$$

And note that $|\Omega^{cross}_{A',B'}| = |\Omega^{cross}_{A,B}| = k^2(n-k) + k(n-k)^2 = n^2k - k^2n$. We have:

$$\begin{aligned}\sum_{(i,j,k)\in\Omega^{cross}_{A',B'}} |\Delta_{i;j,k}| - \sum_{(i,j,k)\in\Omega^{cross}_{A,B}} |\Delta_{i;j,k}| &\leq 2\cdot(n^2k - k^2n) \\ &\quad \cdot\frac{\epsilon}{4}\left(\frac{\psi(n,k)}{n^2k-k^2n}\right) \\ &\leq \frac{\epsilon\psi(n,k)}{2} \\ &\leq \epsilon\cdot\psi(n,k)\end{aligned}$$

Combining this bound on the estimation errors with the bound on $\hat{\mathcal{F}}(A') - \hat{\mathcal{F}}(A)$ yields:

$$\begin{aligned}\hat{\mathcal{F}}(A') - \hat{\mathcal{F}}(A) &\geq \epsilon\psi(n,k) \\ &\quad - \left(\sum_{(i,j,k)\in\Omega^{cross}_{A',B'}} |\Delta_{i;j,k}| - \sum_{(i,j,k)\in\Omega^{cross}_{A,B}} |\Delta_{i;j,k}|\right) \\ &\geq \frac{\epsilon\psi(n,k)}{2} \\ &> 0,\end{aligned}$$

which is almost what we want to show. There remains one thing to address. How many samples do we require to achieve the bound assumed in Equation 0.2 with high probability? Observe that the bound simplifies as,

$$\begin{aligned}\frac{\epsilon}{4}\left(\frac{\psi(n,k)}{n^2k-k^2n}\right) &= \frac{\epsilon}{4}\left(\frac{(n-k)(n-2k)}{nk(n-k)}\right) \\ &= \frac{\epsilon}{4}\left(\frac{n-2k}{nk}\right),\end{aligned}$$

which behaves $O(\epsilon)$ when k is $O(1)$, but like $O\left(\frac{\epsilon}{n}\right)$ when k is $O(n)$. Applying the sample complexity result of Lemma 4 with $\Delta = O(\epsilon/n)$, we see that given $O\left(\frac{n^4}{\epsilon^2}\log^2\frac{n^2}{\epsilon}\log\frac{n^4}{\gamma}\right)$ i.i.d. samples, the bound in Equation 0.2 holds with probability $1-\gamma$, concluding the proof. □

Running time complexity discussion

We need to compute the mutual information quantities $I_{i;j,k}$ for all triplets i, j, k from m samples. These can be computed in $O(mn^3)$ time.

The exhaustive method for finding the k-subset which minimizes $\hat{\mathcal{F}}$ requires evaluating the objective function at $\binom{n}{k} = O(n^k)$ subsets. What is the complexity of evaluating $\hat{\mathcal{F}}$ at a particular partition A, B? We need to sum the precomputed mutual informations over the number of triangles that cross between A and B. If $|A| = k$ and $|B| = n-k$, then we can bound the number of such triangles by $k(n-k)^2 + k^2(n-k) = O(kn^2)$. Thus, we require $O(n^k + kn^2)$ optimization time, leading to a bound of $O(kn^{k+2}+mn^3)$ total time.

The anchors method requires us to (again) precompute mutual informations. The other seeming bottleneck is the last step, in which we must evaluate the objective function $\hat{\mathcal{F}}$ at $O(n^2)$ partitions. In reality, if $|A|$ and $|B|$ are both larger than 1, then a_1 can be held fixed at any arbitrary element, and we must only optimize over $O(n)$ partitions. When $|A| = |B| = 1$, then $n = 2$, in which case the two sets are trivially riffle independent (independent of the actualy distribution). As we showed in the previous paragraph, evaluating $\hat{\mathcal{F}}$ requires $O(kn^2)$ time, and thus optimization using the anchors method $= O(n^3(k+m))$ total time. Since k is much smaller than m (in any meaningful training set), we can drop it from the big-O notation to get $O(mn^3)$ time complexity, showing that the anchors method is dominated by the time that is required to precompute and cache mutual informations.

Encouraging balanced partitions

We have found the following *normalized cut* based variation of our objective (Shi & Malik, 2000)) to be useful for detecting riffled independence when the size k is unknown (see Appendix for details).

$$\mathcal{F}^{balanced}(A) \equiv \frac{\sum_{\Omega^{cross}_{A,B}} I_{i;j,k}}{\sum_{\Omega^{cross}_{A,B}} I_{i;j,k} + \sum_{\Omega^{int}_{A}} I_{i;j,k}} + \frac{\sum_{\Omega^{cross}_{B,A}} I_{i;j,k}}{\sum_{\Omega^{cross}_{B,A}} I_{i;j,k} + \sum_{\Omega^{int}_{B}} I_{i;j,k}}. \quad (0.3)$$

Intuitively, the denominator in Equation 0.3 penalizes subsets whose interiors have small weight. Note that there exist many variations on the objective function that encourage balance, but $\mathcal{F}^{balanced}$ is the one that we have used in our experiments.

References

Höffgen, K. U. Learning and robust learning of product distributions. In *Sixth Annual Conference on Learning Theory*, 1993.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 22(8), 2000.