Supplementary Material of: Covered Information Disentanglement: Correcting Permutation Feature Importance in the Presence of Covariates

João Belo Pereira [1 2] **Diogo Bastos** [1 2] **Erik Stroes** [1] **Evgeni Levin** [1 2]

We re-state theorem 0.1 here for clarity

Theorem 0.1. *For a Markov Random Field, the covered information of a r.v.* X_i *by the set of random variables* $X_I,\ I = \{1, ..., N\}\backslash\{i\}$ *w.r.t.* Y *is given by:*

$$H^{\mathcal{C}(X_I)}_{X_i \cap Y} = \\ 1 + \frac{1}{H(X_i \cap Y)} \mathbf{E}_{\sim p(x_{\sim i, \sim y})} \left[log \left(f \frac{\mathbf{d}^T \mathbf{F}\, \mathbf{e}}{\mathbf{d}^T \mathbf{F}_y \mathbf{F}_{x_i} \mathbf{e}} \right) \right] \tag{1}$$

where $p(x_{\sim i, \sim y})$ *is the joint probability of r.v.s which are neighbors to either* X_i *or* Y*,* $\mathbf{F}$ *is a matrix with the product of joint potential values* $\psi_{\mathcal{C}_F}$ *for set of cliques* $F : \{X_i, Y \in F\}$*;* f*,* $\mathbf{F}_y$ *and* $\mathbf{F}_{x_i}$ *are an entry, column and row of* $\mathbf{F}$*, respectively, while* $\mathbf{d}$ *and* $\mathbf{e}$ *are arrays with the product of potential values* $\psi_{\mathcal{C}_D}$*,* $\psi_{\mathcal{C}_E}$ *for set of cliques* $D : \{X_i \in D, Y \notin D\}$ *and* $E : \{X_i \notin E, Y \in E\}$ *with fixed* X_I*.*

Proof. Using definition 1, 2 and 3:

$$\frac{H\left(X_i \cap Y \cap \{\cup_{j \in I} X_j\}\right)}{H(X_I \cap Y)} = 1 + \frac{\overbrace{H(X_i \cup Y \cup X_I)}^{①} - \overbrace{H(X_I \cup Y)}^{②} + \overbrace{H(X_I)}^{③} - \overbrace{H(X_i \cup X_I)}^{④}}{H(X_i \cap Y)} \tag{2}$$

Representing these terms with marginal distributions:

$$① = -\sum_x p(x) log\ p(x),\ ② = -\sum_x p(x) log \sum_{x_i} p(x),\ ③ = -\sum_x p(x) log \sum_{x_i} \sum_y p(x),\ ④ = -\sum_x p(x) log \sum_y p(x) \tag{3}$$

The probability density for Markov Random fields is equal to $p(x) = \prod_{c=1}^{C} \psi_c(x_c)/\mathbf{Z}$, where $\mathbf{Z}$ is the partition function and c are cliques in the Markov network, C being the total number of cliques. Define two sets of cliques: $A : \{X_i \in A\}$ and $B : \{X_i \notin A\}$. In that case:

$$① = -\sum_x p(x) log \left[log \prod_{\mathcal{C}_B} \psi_{\mathcal{C}_B}(x_{\mathcal{C}_B}) + log \prod_{\mathcal{C}_A} \psi_{\mathcal{C}_A}(x_{\mathcal{C}_A}) \right] + log(\mathbf{Z}), \tag{4}$$

$$② = -\sum_x p(x) log \left[log \prod_{\mathcal{C}_B} \psi_{\mathcal{C}_B}(x_{\mathcal{C}_B}) + log \sum_{x_i} \prod_{\mathcal{C}_A} \psi_{\mathcal{C}_A}(x_{\mathcal{C}_A}) \right] + log(\mathbf{Z}) \tag{5}$$

$$① - ② = -\sum_x p(x) log \left(\frac{\prod_{\mathcal{C}_A} \psi_{\mathcal{C}_A}(x_{\mathcal{C}_A})}{\sum_{x_i} \prod_{\mathcal{C}_A} \psi_{\mathcal{C}_A}(x_{\mathcal{C}_A})} \right) \tag{6}$$

[1] Amsterdam University Medical Center, Meibergdreef 9 1105 AZ, Amsterdam, The Netherlands [2] Horaizon, Marshallaan 2 2625 GZ, Delft, The Netherlands. Correspondence to: João Pereira <j.p.belopereira@amsterdamumc.nl>, Evgeni Levin <e.levin@amsterdamumc.nl>.

To compute ③ − ④, define four sets of cliques: $C : \{X_i \notin C, Y \notin C\}$, $D : \{X_i \in D, Y \notin D\}$, $E : \{X_i \notin E, Y \in E\}$ and $F : \{X_i \in F, Y \in F\}$. In order to reduce the clutter, we will introduce the following functions: $d(x_i, x_I) = \prod_{j \in I, j \sim i} \psi(x_i, x_j)$, $e(y, x_I) = \prod_{j \in I, j \sim y} \psi(y, x_j)$, $f(x_i, y) = \psi(x_i, y)$, where we will abbreviate $d(x_i, x_I)$ into $d(x_i)$ and $e(y, x_I)$ into $e(y)$ when the value for random variable X_I is fixed. Then:

$$③ = -\sum_x p(x) log \left[log \prod_{\mathcal{C}_C} \psi_{\mathcal{C}_C}(x_{\mathcal{C}_C}) + log \sum_{x_i} \sum_y d(x_i) e(y) f(x_i, y) \right] + log(\mathbf{Z}), \tag{7}$$

$$④ = -\sum_x p(x) log \left[log \prod_{\mathcal{C}_C} \psi_{\mathcal{C}_C}(x_{\mathcal{C}_C}) + log \sum_y d(x_i) e(y) f(x_i, y) \right] + log(\mathbf{Z}) \tag{8}$$

$$③ - ④ = -\sum_x p(x) log \left(\frac{\sum_{x_i} \sum_y d(x_i) e(y) f(x_i, y)}{\sum_y d(x_i) e(y) f(x_i = X_i, y)} \right), \tag{9}$$

where $f(x_i = X_i, y)$ is the function f for a fixed value of the r.v. X_i. Since the set of cliques $A = \{D \cup F\}$, and denoting by $d(x_i = X_i)$, $f(x_i = X_i, Y = y)$ the functions d and f for fixed values of X_i and Y, then:

$$\begin{aligned}(① - ②) + (③ - ④) = -\sum_x p(x) log \left(\frac{\sum_{x_i} \sum_y d(x_i = X_I) f(x_i = X_i, Y = y) e(y) f(x_i, y)}{\sum_{x_i} \sum_y d(x_i = X_I) d(x_i) f(x_i, Y = y) e(y) f(X_i, y)} \right) = \\ -\mathbf{E}_{\sim p(x_{\sim i, \sim y})} \left[log\, f(x_i = X_i, Y = y) + log \left(\frac{\mathbf{d}^T \mathbf{F}\, \mathbf{e}}{\mathbf{d}^T \mathbf{F}_y \mathbf{F}_{x_i} \mathbf{e}} \right) \right],\end{aligned} \tag{10}$$

where $x_{\sim i, \sim y}$ is an instance of the set of r.v.s that are neighbors to X_i or Y, $\mathbf{d}$ and $\mathbf{e}$ are column arrays with the different values of $d(x_i)$ and $e(y)$ for fixed X_I, $\mathbf{F}$ is a matrix with all the values $f(x_i, y)$ with varying values of X_i in the rows and Y in the columns, while $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are row and column vectors of $\mathbf{F}$ corresponding to fixed Y and fixed X_i, respectively. This yields the result of the theorem.

□