A. Theoretical Analysis of Local Feature FDR

Definition A.1. *We call a mapping $\sigma : \mathbb{R}^d \to \mathcal{P}([d])$ a generic swap or a swap. In addition, we say that a swap is a local swap if for any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$,*

$$\boldsymbol{x}_{[d]\setminus\sigma(\boldsymbol{x})} = \boldsymbol{z}_{[d]\setminus\sigma(\boldsymbol{x})} \quad \Rightarrow \quad \sigma(\boldsymbol{x}) = \sigma(\boldsymbol{z})$$

Given a mapping

$$(\boldsymbol{F}, \tilde{\boldsymbol{F}}) : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d \\ (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mapsto \begin{array}{r} \big(F_1(\boldsymbol{x},\tilde{\boldsymbol{x}}), \dots, F_d(\boldsymbol{x}, \tilde{\boldsymbol{x}}), \\ \tilde{F}_1(\boldsymbol{x}, \tilde{\boldsymbol{x}}), \dots, \tilde{F}_d(\boldsymbol{x}, \tilde{\boldsymbol{x}})\big) \end{array} \end{cases}$$

and swap σ, define the operation $[\boldsymbol{F}, \tilde{\boldsymbol{F}}]_{swap(\sigma)}$ as the mapping $[\boldsymbol{F}, \tilde{\boldsymbol{F}}]_{swap(\sigma)} : (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mapsto \big[(\boldsymbol{F}, \tilde{\boldsymbol{F}})(\boldsymbol{x}, \tilde{\boldsymbol{x}})\big]_{swap(\sigma(\boldsymbol{x}))}$.

It is important to clearly identify the input space of the swap σ. The output of the swap operation on a mapping is again a mapping with the same input space, so for example we can iterate swap operations on a given mapping $(\boldsymbol{F}, \tilde{\boldsymbol{F}})$: for σ_1, σ_2 swaps the following is well defined: $\big[[\boldsymbol{F}, \tilde{\boldsymbol{F}}]_{swap(\sigma_1)}\big]_{swap(\sigma_2)}$. Both σ_1, σ_2 are evaluated on a same given point $\boldsymbol{x}$ of the space, and the final mapping corresponds to a point-wise concatenation of the swaps. As a consequence, the order of the swap operations does not matter. In particular, for $\sigma = \sigma_1 = \sigma_2$, we have

$$\big[[\boldsymbol{F}, \tilde{\boldsymbol{F}}]_{swap(\sigma)}\big]_{swap(\sigma)} = (\boldsymbol{F}, \tilde{\boldsymbol{F}})$$

Such mappings from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d \times \mathbb{R}^d$ can be concatenated: for $(\boldsymbol{F}, \tilde{\boldsymbol{F}})$, and $(\boldsymbol{G}, \tilde{\boldsymbol{G}})$ we denote $(\boldsymbol{F}, \tilde{\boldsymbol{F}}) \circ (\boldsymbol{G}, \tilde{\boldsymbol{G}})$ the concatenated mapping. Note that concatenation and swap operations do not behave well. For example, in the general case,

$$[(\boldsymbol{F}, \tilde{\boldsymbol{F}}) \circ (\boldsymbol{G}, \tilde{\boldsymbol{G}})]_{swap(\sigma)} \neq (\boldsymbol{F}, \tilde{\boldsymbol{F}}) \circ [\boldsymbol{G}, \tilde{\boldsymbol{G}}]_{swap(\sigma)}$$
$$[(\boldsymbol{F}, \tilde{\boldsymbol{F}}) \circ (\boldsymbol{G}, \tilde{\boldsymbol{G}})]_{swap(\sigma)} \neq [\boldsymbol{F}, \tilde{\boldsymbol{F}}]_{swap(\sigma)} \circ (\boldsymbol{G}, \tilde{\boldsymbol{G}})$$

One particular case is the identity mapping that we denote $(\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id})$, i.e. $\boldsymbol{F}^{Id}_j(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = x_j$ and $\tilde{\boldsymbol{F}}^{Id}_j(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \tilde{x}_j$. Whenever we consider random variables $\boldsymbol{X}, \tilde{\boldsymbol{X}}$ we denote $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)} := [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)}(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ the random variable resulting from applying the swapped identity map. In addition, if σ is constant equal to $S \subset [d]$, then we go back to the previous definition of swap $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}$ in (Candès et al., 2018). Whenever we consider an identity mapping and a local swap σ, we have the immediate following result:

$$(\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}) = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)} \circ [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)}$$

Our goal now is to show that the exchangeability condition that defines a knockoff variable implies a stronger distributional result.

Proposition A.1. *Let σ be a local swap. If $\tilde{\boldsymbol{X}}$ is a knockoff random variable for $\boldsymbol{X}$ (i.e. satisfies exchangeability), then*

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)} \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}] \tag{8}$$

which we refer to as local exchangeability. If $\sigma \subset \mathcal{H}_0^0$, then

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y \tag{9}$$

We prove this result in Section B.1. This result extends the exchangeability property and the Lemma 3.2 in (Candès et al., 2018). Instead of swapping a fixed set of features, we now allow the swapping indices to depend on the features. Notice that the knockoffs $\tilde{\boldsymbol{X}}$ are constructed as in the general case, the local exchangeability does not require a different definition for knockoff variables. We extend the definition of a swap to probability distributions: for $\mu \in Pr(\mathbb{R}^d \times \mathbb{R}^d)$, we denote $\mu_{swap(\sigma)} := \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)})$ whenever $\mu = \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}])$. Abusing notation, whenever $\mu = \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y)$ we will still denote $\mu_{swap(\sigma)} := \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y)$.

Local Feature Statistics The next step is to extend the construction of feature statistics to the local setting.

Definition A.2. *Define local importance scores as a mapping:*

$$\Phi : \begin{cases} Pr(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}) \longrightarrow \big(\mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d\big) \\ \mu \mapsto (\boldsymbol{T}_\mu, \tilde{\boldsymbol{T}}_\mu) \end{cases} \tag{10}$$

where

$$(\boldsymbol{T}_\mu, \tilde{\boldsymbol{T}}_\mu) : \begin{cases} \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d \\ \boldsymbol{z} \mapsto (\boldsymbol{T}_\mu(\boldsymbol{z}), \tilde{\boldsymbol{T}}_\mu(\boldsymbol{z})) \end{cases} \tag{11}$$

such that, for any $S \subset [d]$, we have

$$\Phi(\mu_{swap(S)}) = [\Phi(\mu)]_{swap(S)}$$

For $r > 0$, we say that such importance scores Φ are r-local if, for any $\mu \in Pr(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$, we have that $\Phi(\mu)(\boldsymbol{z}) = (\boldsymbol{T}_\mu(\boldsymbol{z}), \tilde{\boldsymbol{T}}_\mu(\boldsymbol{z}))$ only depends on μ through the restriction of μ to $B(\boldsymbol{z}, r) \times B(\boldsymbol{z}, r) \times \mathbb{R}$. That is, if μ, μ' are two probability measures on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ such that they coincide on $B(\boldsymbol{z}, r) \times B(\boldsymbol{z}, r) \times \mathbb{R}$, then $(\boldsymbol{T}_\mu(\boldsymbol{z}), \tilde{\boldsymbol{T}}_\mu(\boldsymbol{z})) = (\boldsymbol{T}_{\mu'}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu'}(\boldsymbol{z}))$.

The next goal consists in translating the swap operation in $\mu_{swap(\sigma)} = \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y)$ into a swap of $[\boldsymbol{T}_\mu, \tilde{\boldsymbol{T}}_\mu]_{swap(\sigma)}$. This step does not require $\tilde{\boldsymbol{X}}$ to be a knockoff of $\boldsymbol{X}$: in what follows we do not make any assumption on μ. Notice that the swap operation has been defined (Definition 5.1) as a transformation of a mapping $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$, but it can be immediately extended to mappings $\mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$. We are able to relate $[\boldsymbol{T}_{\mu_{swap(\sigma)}}, \tilde{\boldsymbol{T}}_{\mu_{swap(\sigma)}}]$ and $[\boldsymbol{T}_\mu, \tilde{\boldsymbol{T}}_\mu]_{swap(\sigma)}$ if we assume locality constraints on the importance scores.

Definition A.3. *For $\sigma : \mathbb{R}^d \to \mathcal{P}([d])$ local swap, and $r > 0$, define*

$$A^r = \{\boldsymbol{z} \in \mathbb{R}^d,\ \sigma(\boldsymbol{z}) = \sigma(\boldsymbol{y})\ \forall \boldsymbol{y} \in B(\boldsymbol{z}, r)\}$$
$$\sigma^r : \boldsymbol{z} \mapsto \begin{cases} \sigma(\boldsymbol{z}) & \text{if } \boldsymbol{z} \in A^r \\ \emptyset & \text{else} \end{cases}$$

Proposition A.2. *Assume there exists $r > 0$ such that the importance scores are r-local. Then, assuming A^r is non-empty, for $\boldsymbol{z} \in A^r$, we have:*

$$[(\boldsymbol{T}_\mu(\boldsymbol{z}), \tilde{\boldsymbol{T}}_\mu(\boldsymbol{z}))]_{swap(\sigma(\boldsymbol{z}))} = [\boldsymbol{T}_{\mu_{swap(\sigma)}}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu_{swap(\sigma)}}(\boldsymbol{z})]$$

We prove this result in Section B.2. Whenever $\mu = \mathcal{L}([\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y)$ where $\tilde{\boldsymbol{X}}$ is a knockoff of $\boldsymbol{X}$, consider a local swap σ such that $\sigma \subset \mathcal{H}_0^0$. By proposition 5.1, we get that $\mu = \mu_{swap(\sigma)}$, and therefore $(\boldsymbol{T}_\mu, \tilde{\boldsymbol{T}}_\mu) = (\boldsymbol{T}_{\mu_{swap(\sigma)}}, \tilde{\boldsymbol{T}}_{\mu_{swap(\sigma)}})$. In practice, we can imagine that instead of using μ as an input when constructing importance scores, we use $\hat{\mu}_n$, an empirical measure defined by the dataset of n i.i.d. samples from μ that we feed as input to the algorithm. Therefore a consequence of proposition 5.1 will be that, taking also into account the eventual randomness when generating importance scores, we have for any $\boldsymbol{z} \in \mathbb{R}^d$,

$$(\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z})) \overset{d}{=} (\boldsymbol{T}_{\hat{\mu}_{n,swap(\sigma)}}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\hat{\mu}_{n,swap(\sigma)}}(\boldsymbol{z}))$$

We now combine this equality with that of Proposition 5.2. Fix $r > 0$ such that we have r-local importance scores. Consider a set of L points $\boldsymbol{z}_1, \dots \boldsymbol{z}_N \in \mathbb{R}^d$ that are pairwise $2r$ far apart, that is, for any $1 \leq l, l' \leq N$, $\|\boldsymbol{z}_l - \boldsymbol{z}_{l'}\|_\infty \geq 2r$.

Proposition A.3. *σ^r is a local swap and if $\sigma \subset \mathcal{H}_0^0$, then $\sigma^r \subset \mathcal{H}_0^r$. Furthermore,*

$$\left[\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)\right]_{1 \leq l \leq L} \overset{d}{=} \left[[\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_{swap(\sigma^r(\boldsymbol{z}_l))}\right]_{1 \leq l \leq L}$$

We prove this proposition in Section B.3. This allows us to conclude. Fix a target FDR level $q \in (0, 1)$. Indeed, Proposition B.3 directly implies the flip-sign condition of Lemma 3.3 in (Candès et al., 2018). Independently for each $\boldsymbol{z}_l$, consider an independent random variable $\epsilon_l = (\epsilon_{l,1}, \dots, \epsilon_{l,d})$, where for each $1 \leq l \leq L$, and $1 \leq j \leq d$ we have $\epsilon_j^l = 1$ if $j \notin \sigma^r(\boldsymbol{z}_l)$, and a Rademacher random variable if $j \in \sigma^r(\boldsymbol{z}_l)$. Then denoting $\sigma_\epsilon^r(\boldsymbol{z}_l) = \sigma_r(\boldsymbol{z}_l) \cap \{j : \epsilon_{l,j} = -1\}$, we have:

$$(\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)) \overset{d}{=} [\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_{swap(\sigma_\epsilon^r(\boldsymbol{z}_l))}$$

As a consequence, denoting

$$\boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l) = \boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l) - \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)$$

we get that

$$\epsilon_l \odot \boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l) \overset{d}{=} \boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l)$$

where the symbol $\odot$ indicates component-wise multiplication. Now, given that $\|\boldsymbol{z}_l - \boldsymbol{z}_{l'}\|_\infty \geq 2r$, this equality holds uniformly for $1 \leq l \leq L$. The random choice of the swap $\sigma_\epsilon^r(\boldsymbol{z}_l)$ is done independently of a random swap $\sigma_\epsilon^r(\boldsymbol{z}_{l'})$ at another point $\boldsymbol{z}_{l'}$. We conclude that the knockoff selection procedure now applies to each of these vectors in an independent way: that is, for each $1 \leq l \leq L$, setting

$$\hat{\tau}_l = \min\left\{t > 0 : \frac{1 + \#\{j : [\boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_j \leq -t\}}{\#\{j : [\boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_j\}} \leq q\right\}$$

allows to construct selection sets $\hat{\mathcal{S}}_l = \{j : [\boldsymbol{W}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_j \geq \hat{\tau}_l\}$, that control FDR^r given that initially $\sigma^r(\boldsymbol{z}) \subset \mathcal{H}_0^r(\boldsymbol{z})$.

That is, according to Theorem 3.4 in (Candès et al., 2018), the set $\hat{\mathcal{S}}_l$ is such that

$$\mathbb{E}\left[\frac{|\hat{\mathcal{S}}_l \cap \mathcal{H}_0^r(\boldsymbol{x})|}{1 \vee |\hat{\mathcal{S}}_l|}\right] \leq q$$

hence the result.

B. Proofs

B.1. Proof of Proposition 5.1

Proof. We begin the proof with two lemmas:

Lemma B.1. *We can decompose a local swap $\sigma : \mathbb{R}^d \to \mathcal{P}([d])$ into $\sigma_i : \mathbb{R}^d \to \mathcal{P}([d])$ such that for every $i \in [d]$:*

$$\begin{cases} Im(\sigma_i) = \{\emptyset, \{i\}\} \\ \sigma(\boldsymbol{x}) = \bigsqcup_{i=1}^d \sigma_i(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathbb{R}^d \\ \sigma_i \quad \text{is a local swap} \end{cases}$$

We will denote by $\sigma = \bigsqcup_{i=1}^d \sigma_i$ the property $\sigma(\boldsymbol{x}) = \bigsqcup_{i=1}^d \sigma_i(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathbb{R}^d$.

Proof. Define, for every $i \in [d]$,

$$\sigma_i(x) = \begin{cases} \emptyset & \text{if } i \notin \sigma(x) \\ \{i\} & \text{if } i \in \sigma(x) \end{cases}$$

We need to show that σ_i is a local swap. Let $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ such that $\boldsymbol{x}_{[d] \setminus \sigma_i(\boldsymbol{x})} = \boldsymbol{z}_{[d] \setminus \sigma_i(\boldsymbol{x})}$. If $\sigma_i(\boldsymbol{x}) = \emptyset$ then $\boldsymbol{x} = \boldsymbol{z}$ and therefore $\sigma_i(\boldsymbol{x}) = \sigma_i(\boldsymbol{z})$. If $\sigma_i(\boldsymbol{x}) = \{i\}$, then $i \in \sigma(\boldsymbol{x})$, so $\boldsymbol{x}_{[d] \setminus \sigma_i(\boldsymbol{x})} = \boldsymbol{z}_{[d] \setminus \sigma_i(\boldsymbol{x})}$ implies $\boldsymbol{x}_{[d] \setminus \sigma(\boldsymbol{x})} = \boldsymbol{z}_{[d] \setminus \sigma(\boldsymbol{x})}$ and therefore $\sigma(\boldsymbol{x}) = \sigma(\boldsymbol{z})$ given that σ is a local swap. But then we have $\sigma_i(\boldsymbol{x}) = \sigma_i(\boldsymbol{z})$ by definition of σ_i. □

Lemma B.2. *Assume that we have a partition of a local swap σ into two local swaps σ^a, σ^b: $\sigma = \sigma^a \bigsqcup \sigma^b$. Then we have :*

$$[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)} = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma^b)} \circ [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma^a)}$$

Proof. It is crucial to notice that, given our previous swap definition for an identity mapping, the $swap(\sigma^b)$ operator applies to the output of the swapped vector by σ^a. In order to prove the result we need to show that $\sigma^b(\boldsymbol{x}) = \sigma^b(\boldsymbol{z})$ where $\boldsymbol{z}$ is the vector of the first d coordinates of $[\boldsymbol{x}, \tilde{\boldsymbol{x}}]_{swap(\sigma^a(\boldsymbol{x}))}$. Given that, for any $\boldsymbol{x} \in \mathbb{R}^d$ we have $\boldsymbol{x}_{[d]\setminus\sigma^a(\boldsymbol{x})} = \boldsymbol{z}_{[d]\setminus\sigma^a(\boldsymbol{x})}$ and $\sigma^a(\boldsymbol{x}) \subset \sigma(\boldsymbol{x})$, we get $\sigma(\boldsymbol{x}) = \sigma(\boldsymbol{z})$ and same equality with σ^a, as both σ, σ^a are local swaps. That implies $\sigma^b(\boldsymbol{x}) = \sigma^b(\boldsymbol{z})$ as $\sigma = \sigma^a \bigsqcup \sigma^b$. Notice that the order does not matter when composing the two swapped identity mappings. □

In order to prove equation 4, we write it in terms of mappings: we want to show that if $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ satisfy exchangeability, then

$$\begin{aligned}[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)} &= [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \\ &\overset{d}{=} [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}](\boldsymbol{X}, \tilde{\boldsymbol{X}}) = [\boldsymbol{X}, \tilde{\boldsymbol{X}}]\end{aligned}$$

With Lemma B.1 we decompose $\sigma = \bigsqcup_{i=1}^{d} \sigma_i$, and by recursively using Lemma B.2 we get that

$$\begin{aligned}[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)} = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)} \circ \dots \\ \dots \circ [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_d)}\end{aligned}$$

It then suffices to show the equality in distribution for just one swap operation, so that we can recursively apply the swapped identity mappings while keeping the equality in distribution. We then need to prove that :

$$[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \overset{d}{=} [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}](\boldsymbol{X}, \tilde{\boldsymbol{X}})$$

Equivalently, if we condition on $\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}$ we need to show that

$$\begin{aligned}[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)}(\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1} \\ \overset{d}{=} [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}](\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}\end{aligned}$$

Here crucially we use the fact that σ_1 is a local swap. Indeed, whenever we condition on $\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}$, the input values to the mapping $[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)}$ can be seen as constant with respect to $\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}$. Given that σ_1 can only be equal to $\emptyset$ or $\{1\}$, and that therefore its value is determined by $\boldsymbol{X}_{-1}$, hence constant when we condition on $\boldsymbol{X}_{-1}$, we get that either

$$\begin{aligned}[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)}(\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1} \\ = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}](\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}\end{aligned}$$

which therefore holds also in distribution or

$$\begin{aligned}[\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma_1)}(\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1} \\ = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\{1\})}(\boldsymbol{X}, \tilde{\boldsymbol{X}})|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}\end{aligned}$$

In that case, it simplifies into

$$X_1, \tilde{X}_1|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1} \overset{d}{=} \tilde{X}_1, X_1|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}$$

which is a consequence of the fact that $\boldsymbol{X}, \tilde{\boldsymbol{X}}$ satisfy exchangeability, hence the result.

To prove $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}], Y$, we assume that for every $\boldsymbol{x} \in \mathbb{R}^d$, $\sigma(\boldsymbol{x}) \subset \mathcal{H}_0^0(\boldsymbol{x})$. Jointly taking Y with $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$, the proof is the same up to proving that the following holds whenever $1 \in \mathcal{H}_0^0(\boldsymbol{X})$:

$$X_1, \tilde{X}_1, Y|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1} \overset{d}{=} \tilde{X}_1, X_1, Y|\boldsymbol{X}_{-1}, \tilde{\boldsymbol{X}}_{-1}$$

By the properties of $\mathcal{H}_0^0$, $1 \in \mathcal{H}_0^0(\boldsymbol{X})$ holds regardless of the value of X_1 when conditioning on $\boldsymbol{X}_{-1}$. Now if we write down the densities (as we assumed that the joint distribution has a positive density with respect to a product measure):

$$\begin{aligned}p(\boldsymbol{x}, \tilde{\boldsymbol{x}}, y) &= p(y|\boldsymbol{x}, \tilde{\boldsymbol{x}})p(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \\ &= p(y|\boldsymbol{x})p(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \quad \text{as} \quad \tilde{\boldsymbol{X}} \perp\!\!\!\perp Y|\boldsymbol{X} \\ &= p(y|\boldsymbol{x}_{-1})p(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \quad \text{as} \quad X_1 \perp\!\!\!\perp Y|\boldsymbol{X}_{-1} = \boldsymbol{x}_{-1} \\ &= p(y|\boldsymbol{x}_{-1})p([\boldsymbol{x}, \tilde{\boldsymbol{x}}]_{swap(\{1\})}) \quad \text{by exchangeability} \\ &= p([\boldsymbol{x}, \tilde{\boldsymbol{x}}]_{swap(\{1\})}, y)\end{aligned}$$

Hence the result. □

B.2. Proof of Proposition 5.2

Proof. Fix σ local swap, $r > 0$, and $\boldsymbol{z} \in A^r$. Let $S := \sigma(\boldsymbol{z})$, by definition of local importance scores we have that

$$[(\boldsymbol{T}_{\mu}(\cdot), \tilde{\boldsymbol{T}}_{\mu}(\cdot))]_{swap(S)} = [(\boldsymbol{T}_{\mu_{swap(S)}}(\cdot), \tilde{\boldsymbol{T}}_{\mu_{swap(S)}}(\cdot))]$$

By definition of r-local importance scores, the value of $[\boldsymbol{T}_{\mu_{swap(\sigma)}}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu_{swap(\sigma)}}(\boldsymbol{z})]$ depends on $\mu_{swap(\sigma)}$ only through $B(\boldsymbol{z}, r) \times B(\boldsymbol{z}, r) \times \mathbb{R}$. Therefore if $\mu_{swap(\sigma)}$ and $\mu_{swap(S)}$ coincide on $B(\boldsymbol{z}, r) \times B(\boldsymbol{z}, r) \times \mathbb{R}$, then we have

$$\begin{aligned}[\boldsymbol{T}_{\mu_{swap(\sigma)}}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu_{swap(\sigma)}}(\boldsymbol{z})] \\ = [\boldsymbol{T}_{\mu_{swap(S)}}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu_{swap(S)}}(\boldsymbol{z})] \\ = [(\boldsymbol{T}_{\mu}(\boldsymbol{z}), \tilde{\boldsymbol{T}}_{\mu}(\boldsymbol{z}))]_{swap(\sigma(\boldsymbol{z}))}\end{aligned}$$

Given that $\boldsymbol{z} \in A^r$, we have that $\forall \boldsymbol{u}, \boldsymbol{v} \in B(\boldsymbol{z}, r)$, $\sigma(\boldsymbol{u}) = S$. We want to show that, for any $y \in \mathbb{R}$,

$$\begin{aligned}
&\mu_{swap(\sigma)}(\boldsymbol{u}, \boldsymbol{v}, y) = \mu_{swap(S)}(\boldsymbol{u}, \boldsymbol{v}, y) \\
\Leftrightarrow\ &\mathbb{P}\big([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y = \boldsymbol{u}, \boldsymbol{v}, y\big) \\
&\qquad = \mathbb{P}\big([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}, Y = \boldsymbol{u}, \boldsymbol{v}, y\big) \\
\Leftrightarrow\ &\mathbb{P}\big([\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)}(\boldsymbol{X}, \tilde{\boldsymbol{X}}), Y = \boldsymbol{u}, \boldsymbol{v}, y\big) \\
&\qquad = \mathbb{P}\big([\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(S)}(\boldsymbol{X}, \tilde{\boldsymbol{X}}), Y = \boldsymbol{u}, \boldsymbol{v}, y\big) \\
\Leftrightarrow\ &\mathbb{P}\big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(\sigma)}(\boldsymbol{u}, \boldsymbol{v}), y\big) \\
&\qquad = \mathbb{P}\big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y = [\boldsymbol{F}^{Id}, \tilde{\boldsymbol{F}}^{Id}]_{swap(S)}(\boldsymbol{u}, \boldsymbol{v}), y\big) \\
\Leftrightarrow\ &\mathbb{P}\big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y = [\boldsymbol{u}, \boldsymbol{v}]_{swap(\sigma(\boldsymbol{u}))}, y\big) \\
&\qquad = \mathbb{P}\big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y = [\boldsymbol{u}, \boldsymbol{v}]_{swap(S)}, y\big)
\end{aligned}$$

Hence the result. □

B.3. Extended Flip-Sign Property for Local Swaps and Local Importance Scores

Proposition B.3. *σ^r is a local swap and if $\sigma \subset \mathcal{H}_0^0$, then $\sigma^r \subset \mathcal{H}_0^r$. Furthermore,*

$$\begin{aligned}
&\big[\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)\big]_{1 \le l \le L} \\
&\qquad \overset{d}{=} \big[[\boldsymbol{T}_{\hat{\mu}_n}(\boldsymbol{z}_l), \tilde{\boldsymbol{T}}_{\hat{\mu}_n}(\boldsymbol{z}_l)]_{swap(\sigma^r(\boldsymbol{z}_l))}\big]_{1 \le l \le L}
\end{aligned}$$

Proof. Let $\boldsymbol{x}, \boldsymbol{z}$ such that $\boldsymbol{x}_{[d]\setminus\sigma^r(\boldsymbol{x})} = \boldsymbol{z}_{[d]\setminus\sigma^r(\boldsymbol{x})}$. We want to show that $\sigma^r(\boldsymbol{x}) = \sigma^r(\boldsymbol{z})$.

If $\boldsymbol{x} \notin A^r$, then $\sigma^r(\boldsymbol{x}) = \emptyset$, so $\boldsymbol{x} = \boldsymbol{z}$ and $\sigma^r(\boldsymbol{x}) = \sigma^r(\boldsymbol{z})$.

If $\boldsymbol{x} \in A^r$, let us first show that it implies $\boldsymbol{z} \in A^r$. Let $\boldsymbol{y} \in B(\boldsymbol{z}, r)$, show that $\sigma(\boldsymbol{y}) = \sigma(\boldsymbol{z})$. We have that $\boldsymbol{y} - (\boldsymbol{z} - \boldsymbol{x}) \in B(\boldsymbol{x}, r)$, and given that $\boldsymbol{x} \in A^r$, we get $\sigma(\boldsymbol{x}) = \sigma(\boldsymbol{y} - (\boldsymbol{z} - \boldsymbol{x}))$. As σ is a local swap, we have $\sigma(\boldsymbol{x}) = \sigma(\boldsymbol{z})$, and we also get $\sigma(\boldsymbol{y} - (\boldsymbol{z} - \boldsymbol{x})) = \sigma(\boldsymbol{y})$ because

$$\begin{aligned}
&[\boldsymbol{y} - (\boldsymbol{z} - \boldsymbol{x})]_{[d]\setminus\sigma(\boldsymbol{y}-(\boldsymbol{z}-\boldsymbol{x}))} = \boldsymbol{y}_{[d]\setminus\sigma(\boldsymbol{y}-(\boldsymbol{z}-\boldsymbol{x}))} \\
\Leftrightarrow\quad &[\boldsymbol{y} - (\boldsymbol{z} - \boldsymbol{x})]_{[d]\setminus\sigma(\boldsymbol{x})} = \boldsymbol{y}_{[d]\setminus\sigma(\boldsymbol{x})} \\
\Leftrightarrow\quad &[(\boldsymbol{z} - \boldsymbol{x})]_{[d]\setminus\sigma(\boldsymbol{x})} = 0 \\
\Leftrightarrow\quad &\boldsymbol{x}_{[d]\setminus\sigma^r(\boldsymbol{x})} = \boldsymbol{z}_{[d]\setminus\sigma^r(\boldsymbol{x})}
\end{aligned}$$

We can now conclude: $\sigma^r(\boldsymbol{x}) = \sigma(\boldsymbol{x})$ as $\boldsymbol{x} \in A^r$, and given that σ is a local swap we get that $\sigma(\boldsymbol{x}) = \sigma(\boldsymbol{z})$. Finally, as $\boldsymbol{z} \in A^r$, we have $\sigma(\boldsymbol{z}) = \sigma^r(\boldsymbol{z})$ and therefore $\sigma^r(\boldsymbol{x}) = \sigma^r(\boldsymbol{z})$.

Assume $\sigma \subset \mathcal{H}_0^0$. Fix $\boldsymbol{z} \in \mathbb{R}^d$, assume that $\sigma^r(\boldsymbol{z}) \neq \emptyset$ and take $j \in \sigma^r(\boldsymbol{z})$. That implies $\boldsymbol{z} \in A^r$, and therefore $\forall \boldsymbol{y} \in B(\boldsymbol{z}, r)$ we have $j \in \sigma^r(\boldsymbol{z}) = \sigma(\boldsymbol{z}) = \sigma(\boldsymbol{y}) \subset \mathcal{H}_0^0(\boldsymbol{y})$. Therefore $j \in \cap_{\boldsymbol{y} \in B(\boldsymbol{z}, r)} \mathcal{H}_0^0(\boldsymbol{y}) = \mathcal{H}_0^r(\boldsymbol{z})$. We then conclude that $\sigma^r \subset \mathcal{H}_0^r$.

The last statement is a concatenation of Proposition 5.2 and the fact that $\mu = \mu_{swap(\sigma^r)}$. □

C. Semi-synthetic Data Experiments

Our simulations previously described were entirely based on synthetic data. Alternatively, using real SNPs data and then fitting a HMM model yields the same experimental results, which we did for data from the 1000 Genomes Project (Consortium et al., 2015), where we obtained around 2000 individual samples for 27 distinct segments of chromosome 19 containing an average of 50 SNPs per segment, and filtered out SNPs that are extremely correlated (above 0.95). This is because HMMs can truly capture the covariate distribution of a SNP dataset and are a good model for a downstream feature selection with the knockoff procedure. For simplicity, and in order to scale with the number of samples (which is limited with real data), we described simulations based on synthetic covariates.

D. Saliency-based Partitioning

Saliency maps (Lipton, 2016) have emerged as a popular tool for interpretability in Neural Networks. A saliency map allows to identify, under a trained model that minimizes some loss L, which input variation has the strongest impact on the loss at a given training point. Training a neural network on the concatenated vector $\boldsymbol{X}, \tilde{\boldsymbol{X}}$ to predict Y, saliency maps can be used as importance scores at any given training point. That is, denoting $g_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a classifier parametrized by θ (such as a neural network), consider $\hat{\theta}_n$ the output of training such model on the actual data. We can now compute the saliency scores:

$$\mathbb{S}, \tilde{\mathbb{S}} : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d \\ \boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i \mapsto \nabla_{\boldsymbol{X}, \tilde{\boldsymbol{X}}} L(g_{\hat{\theta}_n}(\boldsymbol{X}_i, \hat{\boldsymbol{X}}_i) - Y_i) \end{cases}$$

Notice that the saliency scores are only computed for training points $\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i, Y_i$, but our definition of $\mathbb{S}, \tilde{\mathbb{S}}$ assumes that we can expand the saliency scores to the whole feature space (for example, through smoothing). The reason why these mappings can not be immediately used as local importance scores is because of the training process: the output of the training process are the parameter estimates in $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y)$, which are constructed based on the global training set. Even though the saliency is local at a point the swap operation can not go through the training process, i.e. we can not relate $g_{\hat{\theta}_n(\boldsymbol{X}, \tilde{\boldsymbol{X}}, Y)}([\boldsymbol{x}, \tilde{\boldsymbol{x}}]_{swap(\sigma)})$ and $g_{\hat{\theta}_n([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(\sigma)}, Y)}([\boldsymbol{x}, \tilde{\boldsymbol{x}}])$. This is due to the influence that a training point lying in one region of the space may have at another point (during evaluation) at a different region (Koh & Liang, 2017). Still, we can use these saliency scores to partition the space based on a subsample of the whole initial dataset, and then run the Knockoff procedure with local importance scores at points located in each subregion. This method has the advantage of being computationally less expensive than the previous one, especially in high dimensions.