STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 10: Part II

1 Knockoffs: the random design case

In contrast to the setting in the previous lecture, we focus here on the cases where the covariates are random. Our goal is to construct a set of control variables $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ (a.k.a Model-X knockoffs) with the following properties:

1. Pairwise Exchangeability: for any index set $S \subset \{1, 2, \ldots, p\}$, we have that $[\mathbf{X}, \tilde{\mathbf{X}}] \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}$ or equivalently
$$\mathbf{X}_S, \mathbf{X}_{-S}, \tilde{\mathbf{X}}_S, \tilde{\mathbf{X}}_{-S} \stackrel{d}{=} \tilde{\mathbf{X}}_S, \mathbf{X}_{-S}, \mathbf{X}_S, \tilde{\mathbf{X}}_{-S}$$
2. Response Independence: $\tilde{\mathbf{X}}$ are constructed such that $\mathbf{Y} \perp \tilde{\mathbf{X}}|\mathbf{X}$. This is guaranteed if $\tilde{\mathbf{X}}$ is constructed without looking at $\mathbf{Y}$

Here $\mathbf{X}_S$ is the submatrix of $\mathbf{X}$ with the columns in the set S. When $S = \{j\}$, we require that
$$\mathbf{X}_j, \mathbf{X}_{-j}, \tilde{\mathbf{X}}_j, \tilde{\mathbf{X}}_{-j} \stackrel{d}{=} \tilde{\mathbf{X}}_j, \mathbf{X}_{-j}, \mathbf{X}_j, \tilde{\mathbf{X}}_{-j}.$$

Importantly, it is not sufficient to choose a permutation of the rows of $\mathbf{X}$. Let $\tilde{\mathbf{X}}$ be a permutation of the rows of matrix $\mathbf{X}$, then Pairwise Exchangeability does not hold because the correlation structure between $\mathbf{X}_j$ (the jth column of $\mathbf{X}$) and $\mathbf{X}_{-j}$ is not preserved when replacing $\mathbf{X}_j$ with $\tilde{\mathbf{X}}_j$.

If the distribution of X (denoted by P_X) is exactly known, it is possible to construct knockoff variables $\tilde{\mathbf{X}}$ that satisfy the above properties. We shall discuss methods for doing so.

Lemma. For any subset $S \subseteq \mathcal{N}_0$ (i.e., the null set), we have
$$[\mathbf{X}, \tilde{\mathbf{X}}]|\mathbf{Y} \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}|\mathbf{Y}.$$

Proof. It is equivalent to show that
$$([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{Y}) \stackrel{d}{=} ([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}, \mathbf{Y}).$$
Without loss of generality, let us assume $S = \{1, 2 \ldots, m\} \subseteq \mathcal{N}_0$ for $m \leq n$. By row independence, it suffices to show that
$$((X, \tilde{X}), Y) \stackrel{d}{=} ((X, \tilde{X})_{\text{swap}(S)}, Y).$$
By our assumption that $(X, \tilde{X}) \stackrel{d}{=} (X, \tilde{X})_{\text{swap}(S)}$, we only need to show that
$$Y|(X, \tilde{X}) \stackrel{d}{=} Y|(X, \tilde{X})_{\text{swap}(S)}.$$
Letting $P_{Y|X}(y|x)$ be the conditional distribution of Y given X, observe that
$$P_{Y|(X,\tilde{X})_{\text{swap}(S)}}(y|(x, \tilde{x})) = P_{Y|(X,\tilde{X})}(y|(x, \tilde{x})_{\text{swap}(S)}) = P_{Y|X}(y|x'),$$
where $x'_i = \tilde{x}_i$ for $i \in S$ and $x'_i = x_i$ otherwise. The second equality above comes from the fact that Y is conditionally independent of $\tilde{X}$ given X. Since $Y \perp X_1|X_{2:p}$, we have
$$P_{Y|X}(y|\tilde{x}_1, x'_{2:p}) = P_{Y|X_{2:p}}(y|x'_{2:p}) = P_{Y|X}(y|x_1, x'_{2:p}) = P_{Y|(X,\tilde{X})_{\text{swap}(S\setminus\{1\})}}(y|(x, \tilde{x})).$$

This shows that

$$Y|(X,\tilde{X})_{\text{swap}(S)} \overset{d}{=} Y|(X,\tilde{X})_{\text{swap}(S\setminus\{1\})}.$$

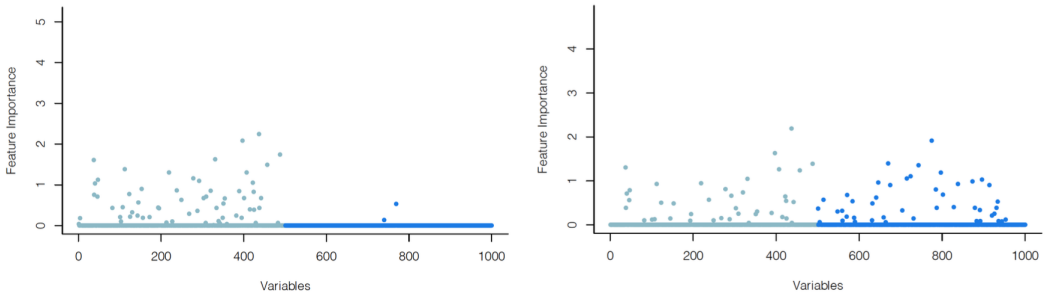
We can repeat this argument with the second variable, the third, and so on until S is empty.

Given the knockoff matrix $\tilde{\mathbf{X}}$, we can construct feature importance statistics $Z_j := T_j(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}})$ for measuring the importance of X_j and $\tilde{Z}_j := T_{p+j}(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}})$ for measuring the importance of $\tilde{X}_j$.

Why is it important to use both $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as control variables inputted into $T_j(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}})$? It is instructive to consider the example where

$$Y = X_2 + \epsilon, \quad \epsilon \sim N(0,1)$$

where ϵ is independent of (X_1, X_2) and X_1, X_2 are strongly correlated. If we just compared $T_j(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2)$ to $T_j(\mathbf{Y}, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ to determine variable importance, then since $\mathbf{X}_1$ is correlated with $\mathbf{X}_2$, it may be that $T_1(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2)$ would be large since the null variable $\mathbf{X}_1$ is correlated with the non-null variable $\mathbf{X}_2$, while $T_1(\mathbf{Y}, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ would be small since $\mathbf{Y} \perp (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ by construction; as such, we would spuriously "reject" H_1.



(a) Permuted Control Variables (b) Knockoffs

Figure 1: Plots of LASSO coefficient magnitudes estimated from data containing 500 null variables colored light blue concatenated with either 500 permuted control variables colored dark blue or knockoff variables.

Let

$$T(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}}) = (Z, \tilde{Z}) = (Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p).$$

Assume the natural property that switching a variable with its knockoff simply switches the components of T in the same way, namely, for each $S \subseteq \{1, ..., p\}$,

$$(Z, \tilde{Z})_{\text{swap}(S)} = T(\mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}).$$

We can show the following result.

Exercise 10.3. For any index set $S \subset \mathcal{N}_0$, we have

$$(Z, \tilde{Z})_{\text{swap}(S)} \overset{d}{=} (Z, \tilde{Z}).$$

1.1 The FDR controlling procedure

We construct the knockoffs-adjusted scores $W_j = w_j(Z_j, \tilde{Z}_j)$ via some anti-symmetric function $w(\cdot, \cdot)$ such that $w(x, y) = -w(y, x)$. One example is that $w(x, y) = x - y$.

Lemma. For any null index $j \in \mathcal{N}_0$, the distribution of W_j is symmetric about zero. Therefore, $P(\text{sign}(W_j) = 1) = P(\text{sign}(W_j) = -1) = 1/2$, i.e., W_j is a Rademacher random variable. Moreover, conditional on $\{|W_j| : 1 \leq j \leq p\}$, sign$(W_j)$s are i.i.d Rademacher random variables.

Proof. We only show the first fact. Consider any measurable set $A \subseteq \mathbb{R}$. Because $(Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j)$ by the exercise,

$$\Pr(w_j(Z_j, \tilde{Z}_j) \in A) = \Pr((Z_j, \tilde{Z}_j) \in w_j^{-1}(A)) = \Pr((\tilde{Z}_j, Z_j) \in w_j^{-1}(A)) = \Pr(-w_j(Z_j, \tilde{Z}_j) \in A),$$

which shows that W_j is symmetric about zero.

Since W_j is symmetrically distributed under the null, it is equally likely that $W_j \geq t$ and $W_j \leq -t$. Let

$$S^+(t) := \{j : W_j \geq t\} \quad \text{and} \quad S^-(t) := \{j : W_j \leq -t\}.$$

Under the alternative, we expect W_j to take a large positive value. Similar to the fixed design case, we can construct the following "conservative" estimate of $\mathrm{FDP}(t)$:

$$\mathrm{FDP}(t) = \frac{|\{j \in \mathcal{N}_0 : W_j \geq t\}|}{1 \vee |S^+(t)|} \approx \frac{|\{j \in \mathcal{N}_0 : W_j \leq -t\}|}{1 \vee |S^+(t)|} \leq \frac{1 + |S^-(t)|}{1 \vee |S^+(t)|} := \widehat{\mathrm{FDP}}(t).$$

Define

$$\tau_q = \min \left\{ t : \widehat{\mathrm{FDP}}(t) \leq q \right\}.$$

Then, we reject all H_j with $W_j \geq \tau_q$.

Theorem. The knockoff procedure controls the FDR.

Proof. The proof is based on the optional stopping theorem. First, since we reject all H_j such that $j \in S^+(\tau_q)$, we can write the FDP using the threshold τ_q as follows:

$$\begin{aligned}
\mathrm{FDP}(\tau_q) =& \frac{|\mathcal{N}_0 \cap S^+(\tau_q)|}{|S^+(\tau_q)| \vee 1} \\
=& \frac{|\mathcal{N}_0 \cap S^+(\tau_q)|}{1 + |\mathcal{N}_0 \cap S^-(\tau_q)|} \frac{1 + |\mathcal{N}_0 \cap S^-(\tau_q)|}{|S^+(\tau_q)| \vee 1} \\
\leq& \frac{|\mathcal{N}_0 \cap S^+(\tau_q)|}{1 + |\mathcal{N}_0 \cap S^-(\tau_q)|} \frac{1 + |S^-(\tau_q)|}{|S^+(\tau_q)| \vee 1} \\
\leq& q \frac{|\mathcal{N}_0 \cap S^+(\tau_q)|}{1 + |\mathcal{N}_0 \cap S^-(\tau_q)|}.
\end{aligned}$$

As such, let $V^+(\tau_q) := |\mathcal{N}_0 \cap S^+(\tau_q)|$ and $V^-(\tau_q) := |\mathcal{N}_0 \cap S^-(\tau_q)|$. To show that $\mathrm{FDR}(\tau_q) = \mathbb{E}[\mathrm{FDP}(\tau_q)] \leq q$, it suffices to show that

$$\mathbb{E}\left[\frac{|\mathcal{N}_0 \cap S^+(\tau_q)|}{1 + |\mathcal{N}_0 \cap S^-(\tau_q)|} \right] = \mathbb{E}\left[\frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \right] \leq 1. \tag{1}$$

Next, akin to the argument used in the empirical process perspective-based proof that the BH procedure controls FDR, we will argue that $V^+(t)/(1 + V^-(t))$ is a supermartingale with respect to the filtration $\mathcal{F}_t := \{\sigma(V^{\pm}(u))\}_{u \leq t}$ with t increasing from 0 so that we can apply Doob's Optional Stopping Theorem. Consider any $s \geq t$, and note that, conditional on $V^+(s) + V^-(s)$, $V^+(s)$ has a hypergeometric distribution. Then it can be shown that (Why? Check the supplement of Barber and Candès (2015))

$$\mathbb{E}\left[\frac{V^+(s)}{1 + V^-(s)} \middle| V^{\pm}(t), V^+(s) + V^-(s) \right] \leq \frac{V^+(t)}{1 + V^-(t)}, \tag{2}$$

which is exactly what is required for $V^+(t)/(1 + V^-(t))$ to be a supermartingale. As a prelude to applying Doob's Optional Stopping Theorem, recall that if $Y \sim \mathrm{Bin}(n_0, 1/2)$, $\mathbb{E}[Y/(1 + n_0 - Y)] \leq 1$ (why?). Since $V^+(0) \sim \mathrm{Bin}(|\mathcal{N}_0|, 1/2)$ (why?), then by combining Doob's Optional Stopping Theorem with this fact, we have that

$$\mathrm{FDR} \leq q\mathbb{E}\left[\frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \right] \leq q\mathbb{E}\left[\frac{V^+(0)}{1 + V^-(0)} \right] = q\mathbb{E}\left[\frac{V^+(0)}{1 + |\mathcal{N}_0| - V^+(0)} \right] \leq q, \tag{3}$$

as required.

2 Constructing knockoff copies

2.1 The Gaussian case

We first consider the case where X is Gaussian. For simplicity, let $X = (X_1, X_2) \sim N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{i,j=1}^2$. Let $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)$ be our knockoff copy. By the pairwise exchangeability,

$$(X_1, X_2, \tilde{X}_1, \tilde{X}_2) \stackrel{d}{=} (\tilde{X}_1, \tilde{X}_2, X_1, X_2)$$

and thus

$$(X_1, X_2) \stackrel{d}{=} (\tilde{X}_1, \tilde{X}_2).$$

The joint distribution of $(X_1, X_2, \tilde{X}_1, \tilde{X}_2)$ is $N(\mathbf{0}, G)$, where G should have the structure:

$$G = \begin{pmatrix} \Sigma & * \\ * & \Sigma \end{pmatrix}.$$

Again, by the pairwise exchangeability,

$$(X_1, X_2, \tilde{X}_1, \tilde{X}_2) \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_1, X_2)$$

and hence

$$(X_1, X_2) \stackrel{d}{=} (X_1, \tilde{X}_2) \stackrel{d}{=} (\tilde{X}_1, X_2).$$

We can show that

$$G = \begin{pmatrix} \sigma_{11} & \sigma_{12} & * & \sigma_{12} \\ \sigma_{12} & \sigma_{22} & \sigma_{12} & * \\ * & \sigma_{12} & \sigma_{11} & \sigma_{12} \\ \sigma_{12} & * & \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma & \Sigma - \mathrm{diag}(s) \\ \Sigma - \mathrm{diag}(S) & \Sigma \end{pmatrix},$$

where $s = (s_1, s_2)$. where s is a 2-vector. Ideally, we would like the $*$ elements of G to be 0. We would then have that X_1 and $\tilde{X}_1$ are independent, and X_2 and $\tilde{X}_2$ are independent. Although this is sometimes possible, it is not always the case. Structurally, we know that G must be a symmetric, positive definite matrix to be a covariance matrix. A possible relaxation is to solve the optimization problem

$$\min_{s_1, s_2} |\sigma_{11} - s_1| + |\sigma_{22} - s_2|$$

subject to that G is positive semi-definite.

Once we determine the value of s, we can derive the conditional distribution $\tilde{X}|X$. We can then sample $\tilde{X}_i$ from this distribution given $X = X_i$, where X_i is the ith observation with $1 \leq i \leq n$.

2.2 The General case

Candès et al. (2018, JRSSB) and Sesia et al. (2018, Biometrika) presented an algorithm (Sequential Conditional Independence Pairs or SCIP) for generating knockoffs in the general case. Let $X = (X_1, \ldots, X_p) \sim P_X$, where P_X is known. SCIP is motivated by the following observation.

Observation. The random variables $(\tilde{X}_1, \ldots, \tilde{X}_p)$ are model-X knockoffs for $(X_1, \ldots, X_p)$ if and only if for any $j \in \{1, 2, \ldots, p\}$, the pair $(X_j, \tilde{X}_j)$ is exchangeable conditional on all the other variables and their knockoffs, i.e., $X_{-j}, \tilde{X}_{-j}$.

If the components of the vector X are independent, then any independent copy of X would work; that is, any vector $\tilde{X}$ independently sampled from the same joint distribution as X would work. In general, we can consider the sequential procedure.

- For $j = 1, 2, \ldots, p$, sample $\tilde{X}_j$ from the law of $X_j | X_{-j}, \tilde{X}_{1:j-1}$.

Theorem. The variables generated in this way are Model-X knockoffs.

Proof. We only prove the results when the variables are discrete. The proof is based on

Induction hypothesis: At the jth step, $(X_k, \tilde{X}_k)$ are exchangeable in the joint distribution of $X_1, \ldots, X_p$ and $\tilde{X}_1, \ldots, \tilde{X}_j$ for $k = 1, 2 \ldots, j$.

Clearly, X_1 and $\tilde{X}_1$ are exchangeable conditional on X_{-1}. Suppose the induction hypothesis is true for $j-1$. We hope to show the induction hypothesis for j.

Below, we denote the probability mass function (PMF) of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $P(X_{-j}, X_j, \tilde{X}_{1:j-1})$. Note that $\tilde{X}_j$ is sampled from the law of $X_j | X_{-j}, \tilde{X}_{1:j-1}$. Thus, the law of $\tilde{X}_j$ conditional on $X_{-j}, \tilde{X}_{1:j-1}$ is given by

$$\frac{P(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u P(X_{-j}, u, \tilde{X}_{1:j-1})}.$$

Thus the joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is given by

$$\begin{aligned} P(X_{1:p}, \tilde{X}_{1:j}) =& P(\tilde{X}_j | X_{1:p}, \tilde{X}_{1:j-1}) P(X_{1:p}, \tilde{X}_{1:j-1}) \\ =& P(\tilde{X}_j | X_{-j}, \tilde{X}_{1:j-1}) P(X_{1:p}, \tilde{X}_{1:j-1}) \\ =& \frac{P(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1}) P(X_{-j}, X_j, \tilde{X}_{1:j-1})}{\sum_u P(X_{-j}, u, \tilde{X}_{1:j-1})}, \end{aligned}$$

which is symmetric in X_j and $\tilde{X}_j$. By the induction hypothesis, $P(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})$ is symmetric in $(X_k, \tilde{X}_k)$ for $k = 1, 2, \ldots, j-1$. Therefore, $P(X_{1:p}, \tilde{X}_{1:j})$ is symmetric in $(X_k, \tilde{X}_k)$ for all $k = 1, 2, \ldots, j$.

Although this method allows for sampling knockoffs from arbitrary known distributions, it may be infeasible to compute. Consider the example (Ising model) where

$$P(X = x) \propto \exp\left(-\sum_{i,j} \beta_{ij} x_i x_j - \sum_i \alpha_i x_i \right),$$

where $x = (x_1, \ldots, x_p)$ with $x_i = \pm 1$. In this case, the algorithm is infeasible to compute if p is large.

2.3 Markov models

Instead, if X is described by a Markov or hidden Markov model, then SCIP is much easier to compute. Let $X = (X_1, \ldots, X_p) \sim \mathrm{MC}(q_1, Q)$ be a Markov chain, where q_1 is the distribution of X_1 and Q is the transition distribution. The Markov property dictates that $X_i | X_{1:(i-1)} \stackrel{d}{=} X_i \mid X_{i-1}$. It follows by the definition of conditional probability that the probability density of X is

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j \mid X_{j-1}).$$

In this case, we can implement SCIP efficiently. We consider the $p = 4$ case, although this procedure will hold for all integers $p \geq 1$. Let $\mathcal{X}$ be the state space of X. We first sample $\tilde{X}_1$ from $X_1 \mid X_{-1}$. It follows by Bayes's theorem and the Markov property that

$$p(X_1 = \tilde{x}_1 \mid X_{-1} = x_{-1}) \propto q_1(\tilde{x}_1) Q_2(x_2 \mid \tilde{x}_1).$$

We can calculate the normalizing constant as $\mathcal{N}_1(x_2) = \sum_{l \in \mathcal{X}} q_1(l) Q_2(x_2 \mid l)$. We can now sample $\tilde{X}_2$ from $X_2 \mid X_{-2}, \tilde{X}_1$ as

$$p(\tilde{X}_2 = \tilde{x}_2 \mid X_{-2} = x_{-2}, \tilde{X}_1 = \tilde{x}_1) \propto Q_2(\tilde{x}_2 \mid x_1) Q_3(x_3 \mid \tilde{x}_2) \frac{Q_2(\tilde{x}_2 \mid \tilde{x}_1)}{\mathcal{N}_1(\tilde{x}_2)}.$$

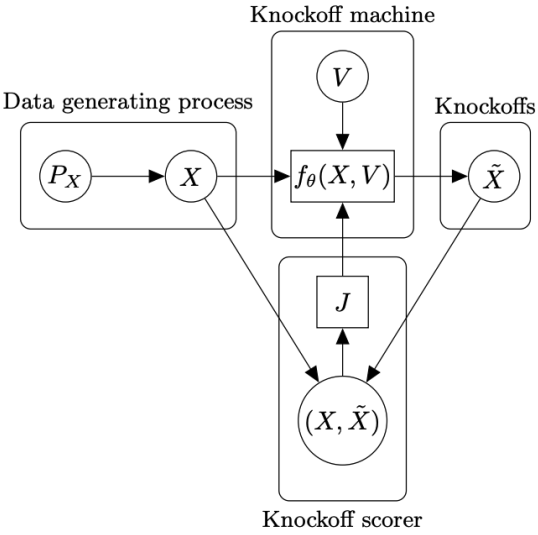


Figure 2: Schematic representation of the learning mechanism of a knockoff machine.

It follows that the normalization constant is $\mathcal{N}_2(x_2) = \sum_{l \in \mathcal{X}} Q_2(l \mid x_1) Q_3(x_3 \mid l) \frac{Q_2(l \mid \tilde{x}_1)}{\mathcal{N}_1(l)}$. This pattern continues as

$$p(\tilde{X}_3 = \tilde{x}_3 \mid X_{-3} = x_{-3}, \tilde{X}_1 = \tilde{x}_1, \tilde{X}_2 = \tilde{x}_2) \propto Q_3(\tilde{x}_3 \mid x_2) Q_4(x_4 \mid \tilde{x}_3) \frac{Q_3(\tilde{x}_3 \mid \tilde{x}_2)}{\mathcal{N}_2(\tilde{x}_3)},$$

with

$$\mathcal{N}_3(x_2, \tilde{x}_2, x_4) = \sum_{l \in \mathcal{X}} Q_3(l \mid x_2) Q_4(x_4 \mid l) \frac{Q_3(l \mid \tilde{x}_2)}{\mathcal{N}_2(l)},$$

and

$$p(\tilde{X}_4 = \tilde{x}_4 \mid X_{-4} = x_{-4}, \tilde{X}_1 = \tilde{x}_1, \tilde{X}_2 = \tilde{x}_2, \tilde{X}_3 = \tilde{x}_3) \propto Q_4(\tilde{x}_4 \mid x_3) \frac{Q_4(\tilde{x}_4 \mid \tilde{x}_3)}{\mathcal{N}_3(\tilde{x}_4)},$$

with

$$\mathcal{N}_4(x_3, \tilde{x}_3) = \sum_{l \in \mathcal{X}} Q_4(l \mid x_3) \frac{Q_4(l \mid \tilde{x}_3)}{\mathcal{N}_3(l)}.$$

2.4 Deep learning based approaches

More recently, novel research has been conducted on generating knockoffs of X when the distribution of X is unknown using neural networks, such as KnockoffGAN, and Deep knockoffs based on the so-called Maximum Mean Discrepancy. Please refer to Jordon et al. (2019, ICLR) and Romano et al. (2020, JASA). Here, we provide a high-level description of the idea.

Given n independent p-dimensional samples $\{X_i\}_{i=1}^n$ from an unknown distribution P_X, a generative model approximating the true PX is sought to synthesize new observations that could plausibly belong to the training set while being sufficiently different to be non-trivial. Modern approaches include

- variational autoencoders;
- generative adversarial networks;
- diffusion models.

Let V_i be an external random variable (e.g., $V_i \sim N(0, I_p)$). We generate $\tilde{X}_i = f_\theta(X_i, V_i)$, where θ is the parameter to be optimized. Let $\tilde{\mathbf{X}} = [\tilde{X}_1, \ldots, \tilde{X}_n]^\top$. We need a measure to quantify the discrepancy between the distributions of $[\mathbf{X}, \tilde{\mathbf{X}}]$ and $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\{j\})}$. Let J be such a measure. Examples include

- maximum mean discrepancy [Gretton et al. (2012, JMLR)] and energy distance;
- Wasserstein metric;
- Jensen-Shannon divergence.

Then, we can define

$$\sum_{j=1}^{p} J([\mathbf{X}, \tilde{\mathbf{X}}], [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\{j\})}).$$

and find θ to minimize this objective function. In practice, the optimization is often performed through stochastic gradient descent (SGD) or Adam.