Solution 1:

(a) **Can we factorize the joint distribution $\mathbb{P}(x)$ as $\mathbb{P}(x_S)\mathbb{P}(x_{-S})$? How can we factorize the joint distribution so that the distribution is preserved? Formally prove your answer.**

In general, **no**, we cannot factorize the joint distribution as $\mathbb{P}(x) = \mathbb{P}(x_S)\mathbb{P}(x_{-S})$ without losing information about the distribution.

Formal proof:

By the definition of conditional probability:

$$\mathbb{P}(x) = \mathbb{P}(x_S, x_{-S}) = \mathbb{P}(x_S|x_{-S}) \cdot \mathbb{P}(x_{-S}) = \mathbb{P}(x_{-S}|x_S) \cdot \mathbb{P}(x_S)$$

The factorization $\mathbb{P}(x) = \mathbb{P}(x_S)\mathbb{P}(x_{-S})$ would only be valid if:

$$\mathbb{P}(x_S|x_{-S}) = \mathbb{P}(x_S) \quad \text{for all } x_{-S}$$

or equivalently:

$$\mathbb{P}(x_{-S}|x_S) = \mathbb{P}(x_{-S}) \quad \text{for all } x_S$$

This is the definition of independence: $x_S \perp\!\!\!\perp x_{-S}$. When features are dependent, this factorization does not preserve the joint distribution.

Correct factorization (chain rule) that preserves the distribution:

$$\mathbb{P}(x) = \mathbb{P}(x_S|x_{-S}) \cdot \mathbb{P}(x_{-S}) \quad \text{or} \quad \mathbb{P}(x) = \mathbb{P}(x_{-S}|x_S) \cdot \mathbb{P}(x_S)$$

(b) **Let $x_S \perp\!\!\!\perp x_{-S}$. Does the factorization now preserve the joint distribution? Formally prove your answer.**

Yes, when $x_S \perp\!\!\!\perp x_{-S}$, the factorization $\mathbb{P}(x) = \mathbb{P}(x_S)\mathbb{P}(x_{-S})$ exactly preserves the joint distribution.

Formal proof:

If $x_S \perp\!\!\!\perp x_{-S}$, then by definition:

$$\mathbb{P}(x_S|x_{-S}) = \mathbb{P}(x_S) \quad \text{for all } x_{-S}$$

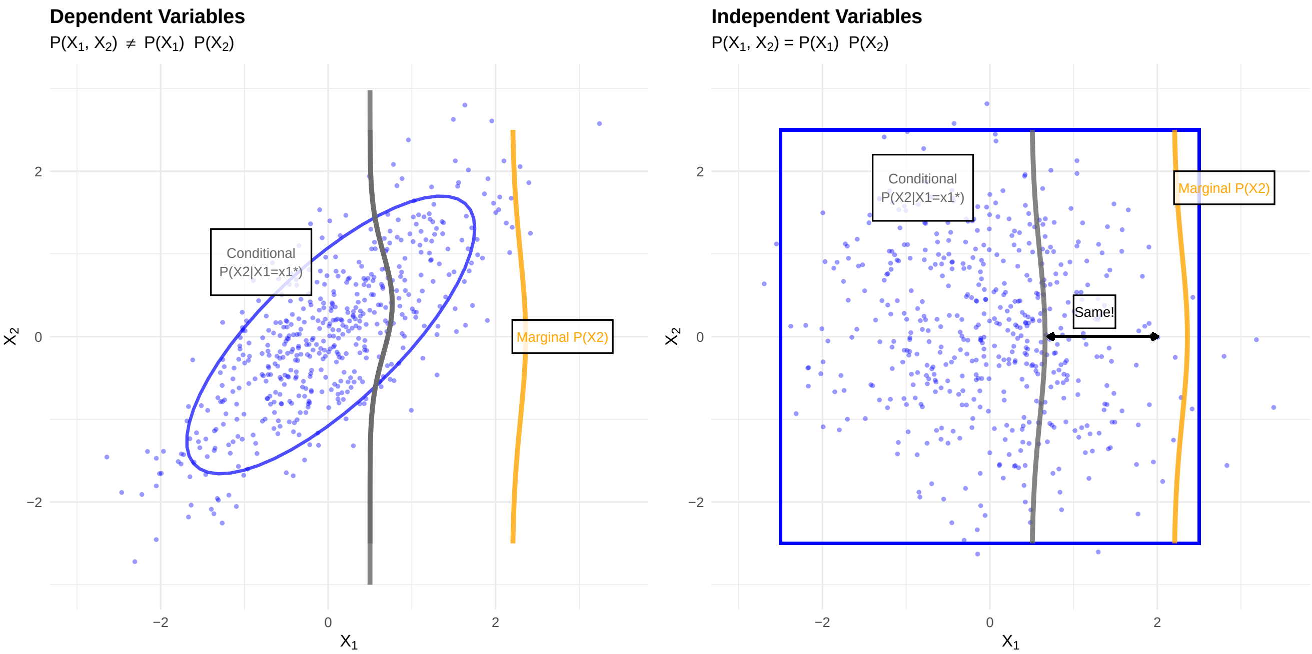
and

$$\mathbb{P}(x_{-S}|x_S) = \mathbb{P}(x_{-S}) \quad \text{for all } x_S$$

Therefore:

$$\begin{aligned}
\mathbb{P}(x) &= \mathbb{P}(x_S, x_{-S}) \\
&= \mathbb{P}(x_S|x_{-S}) \cdot \mathbb{P}(x_{-S}) \\
&= \mathbb{P}(x_S) \cdot \mathbb{P}(x_{-S}) \quad \text{(by independence)}
\end{aligned}$$

(c) **Illustrate the two factorizations in a schematic drawing. *Hint:* You can draw a 2D scatterplot with two dependent variables. Given a fixed value for the conditioned variable, draw the range of values that conditional and marginal sampling consider.**



Solution 2:

(a) **Over which distributions does PFI evaluate the model? Under which assumptions is the model evaluated outside the domain?**

PFI evaluates the model over the marginal distribution. This causes an extrapolation issue when features are dependent. The marginal sampling creates feature combinations (x_S, x_{-S}) that may be unrealistic. The model is then evaluated on these unrealistic data points that lie outside the original data domain.

(b) **Over which distributions does CFI evaluate the model? Does the method extrapolate?**

CFI evaluates the model over the original joint distribution by using conditional sampling. Thanks to this, CFI does not create unrealistic feature combinations.

(c) **What distributions does LOCO consider? Do extrapolation or data outside the domain occur here?**

LOCO compares performance of the model on full data versus data with one feature removed. Because it evaluates the model on the same data distribution (just with/without one feature), it does not create unrealistic feature combinations.

(d) **For both PFI and CFI, evaluate whether/when the perturbed variables are dependent/independent of the target variable.**

PFI: The perturbed feature values are created by shuffling, so they're independent of the target (and of the other features) by design.

CFI: The perturbed feature values are sampled from the conditional distribution $\mathbb{P}(X_S|X_{-S})$. If X_S is dependent on Y even after conditioning on X_{-S}, then the perturbed values will still be dependent on Y. Similarly, if X_S is independent of Y given X_{-S}, then the perturbed values will be independent of Y.

(e) **What does that mean for the interpretation of PFI and CFI?**

ToDo

(f) **Can a feature be relevant for CFI but not relevant for PFI?**

ToDo. We could talk about the case where the opposite happens (for example x_1 is temperature in Celsius, x_2 is temperature in Fahrenheit), but for this case, I feel like theoretically we could construct a case where CFI is relevant but PFI is not, but I can't think of a concrete example.

Solution 3:

Discuss with your neighbor. Which of the aforementioned methods is superior? PFI or the extrapolation-free alternatives?

(a) **Which method is most suitable for situations where we aim to understand the model's mechanism? If any?**

Prefer CFI (and report PFI alongside). CFI conditionally resamples X_S from $P(X_S|X_{-S})$, which preserves the joint feature distribution, keeps inputs realistic, and measures the extra predictive information in X_S beyond what the other features already provide. PFI, by contrast, breaks all links by shuffling X_S; its score therefore includes interactions with X_{-S} but can be inflated by unrealistic pairs when features are correlated. Read them together: if PFI $\approx$ CFI > 0, the feature adds unique signal; if PFI $\gg$ CFI with correlated features, suspect extrapolation or reliance on interactions/proxies; if both ≈ 0, the feature is likely redundant. Always compute importance on a held-out test set - PFI on train often reflects overfitting.

(b) **Which method is most suitable for situations where we want to understand the data generating mechanism?**

(i) **In order to find features that are informative of the prediction target?**

TODO. Perhaps should be CFI (on test set) but I'm not super sure. The reasoning is similar to the point above.

(ii) **In order to select the smallest possible set of features, which would enable the same prediction performance?**

LOCO is best for finding the smallest feature set with the same accuracy because it answers exactly that question at the learner level: drop x_j, **retrain** the learner, and see whether performance changes; if a feature is redundant (e.g., perfectly substituted by others), the retrained model will recover performance, whereas irreplaceable features cause a performance drop, and therefore answer the "can the learner do just as well without it?" question directly.

Alternative consideration: For computational efficiency in high-dimensional settings, CFI might be preferred as it doesn't require retraining models. However, LOCO provides the most direct answer to the feature selection question.

(iii) **In order to find variables that are causal for the prediction target?**

All discussed methods have a fundamental limitation: they measure *associational* rather than *causal* relationships. Correlation $\neq$ Causation, high feature importance doesn't imply causal influence, and important features might be correlated with true causal variables without being causal themselves.

Feature importance methods can provide *hints* about potential causal relationships by identifying strong associations, but they cannot establish causation and should be combined with proper causal inference techniques and domain expertise.

Example: High CFI for "ice cream sales" when predicting "drowning incidents" doesn't mean ice cream causes drowning - both are caused by hot weather (confounding variable).