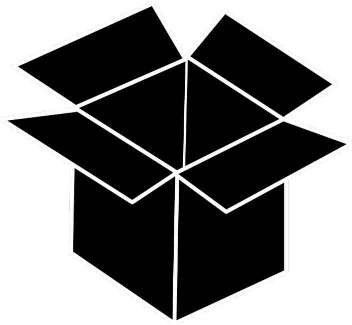
Interpretable Machine Learning

Correlation and dependencies



Learning goals

- Feature dependencies

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

Conditional distribution

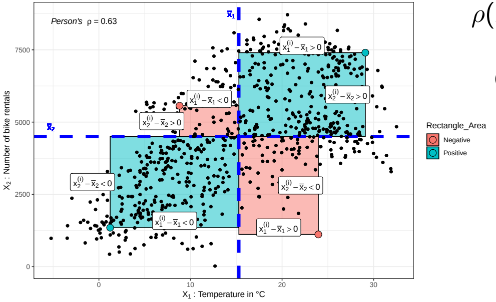
$$p_{X_1|X_2}(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

	$x_2 = 0$	$x_2 = 1$
$\mathbb{P}(X_1 = 0\|X_2 = x_2)$	0.67	0.43
$\mathbb{P}(X_1 = 1\|X_2 = x_2)$	0.33	0.57
$\sum$	1	1

$\rightsquigarrow$ Analogue in the continuous case with integrals.

PEARSON'S CORRELATION COEFFICIENT ρ

By **correlation** often Pearson's correlation is meant (measures only **linear relationship**)



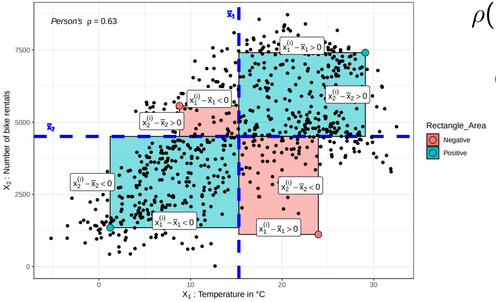
$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$

Geometric interpretation of ρ:

- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on point position
- Denominator scales the sum to $[-1, 1]$

PEARSON'S CORRELATION COEFFICIENT ρ

By **correlation** often Pearson's correlation is meant (measures only **linear relationship**)



$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$

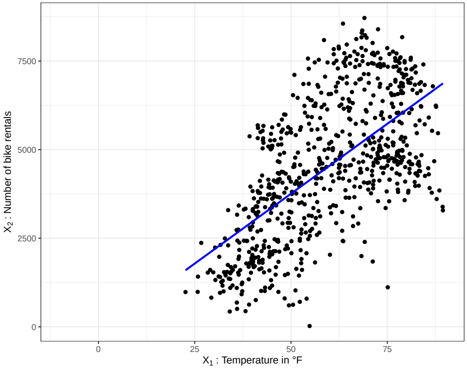
Geometric interpretation of ρ:

- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on point position
- Denominator scales the sum to $[-1, 1]$

- $\rho = 0$ if area of rectangles of all points cancels out $\rightsquigarrow X_1, X_2$ linearly uncorrelated
- $\rho > 0$ if positive areas dominate negative areas $\rightsquigarrow X_1, X_2$ positive correlated
- $\rho < 0$ if negative areas dominate positive areas $\rightsquigarrow X_1, X_2$ negative correlated

COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between variables is by calculating the R^2

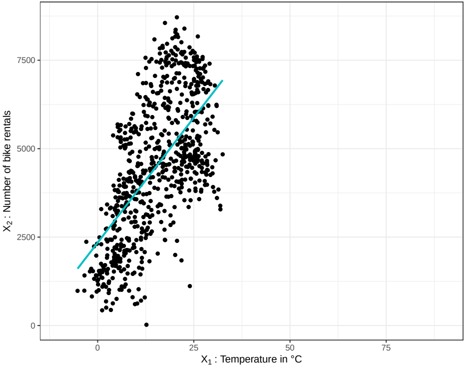


Idea for two-dimensional case:

- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
  $\rightsquigarrow$ Slope = 0 $\Rightarrow$ no dependence
  $\rightsquigarrow$ Very large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic

COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between variables is by calculating the R^2



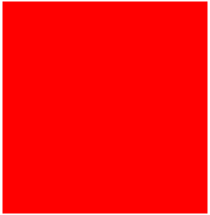
Idea for two-dimensional case:

- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

  $\rightsquigarrow$ Slope = 0 $\Rightarrow$ no dependence

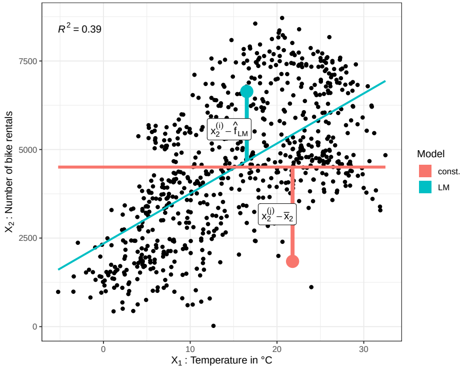
  $\rightsquigarrow$ Very large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic

  $\rightsquigarrow$ Rescaling of x_1 or x_2 changes θ_1
- e.g. ℉ $\rightarrow$ ℃ $\Rightarrow$
  $\theta_1 = 78.5 \rightarrow \theta_1^* = 141.3$



COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between variables is by calculating the R^2



Idea for two-dimensional case:

- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
  $\leadsto$ Slope = 0 $\Rightarrow$ no dependence
  $\leadsto$ Very large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
  $\leadsto$ Rescaling of x_1 or x_2 changes θ_1
- Set SSE_{LM} in relation to SSE of a constant model $\hat{f}_c = \bar{x}_2$
  $SSE_{LM} = \sum_{i=1}^{n}(x_2^{(i)} - \hat{f}_{LM}(x_1^{(i)}))^2$
  $SSE_c = \sum_{i=1}^{n}(x_2^{(i)} - \bar{x}_2)^2$

$\Rightarrow$ Measure of fitting quality of LM: $R^2 = 1 - \frac{SSE_{LM}}{SSE_c} \in [-1, 1]$

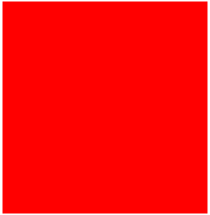
$\Rightarrow \rho(X_1, X_2) = R$

MUTUAL INFORMATION

- MI describes amount of information about one random variable obtained through another one or how different the joint distribution is from pure independence
- $MI(X_1; X_2)$ is the Kullback-Leibler distance between joint distribution and product distribution $p_{X_1} p_{X_2}$:

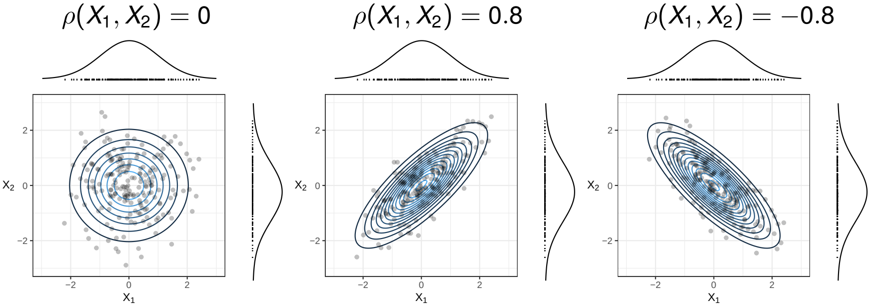
$$
\begin{aligned}
MI(X_1; X_2) &= \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) log \left(\frac{p(x_1, x_2)}{p(x_1) p(x_2)} \right) \\
&= D_{KL} \left(p(x_1, x_2) \,||\, p(x_1) p(x_2) \right) \\
&= \mathbb{E}_{p(x_1, x_2)} \left[log \left(\frac{p(x_1, x_2)}{p(x_1) p(x_2)} \right) \right]
\end{aligned}
$$

- MI measures amount of "dependence" between variables. It is zero if and only if the variables are independent.
- Unlike (Pearson) correlation, MI is not limited to real-valued random variables.



CORRELATION AND DEPENDENCE

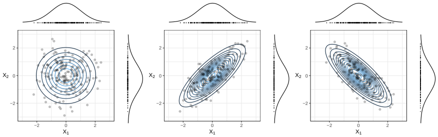
Scatterplot with multivariate distribution (contour lines) and marginal density X_1, $X_2 \sim N(0, 1)$



CORRELATION AND DEPENDENCE

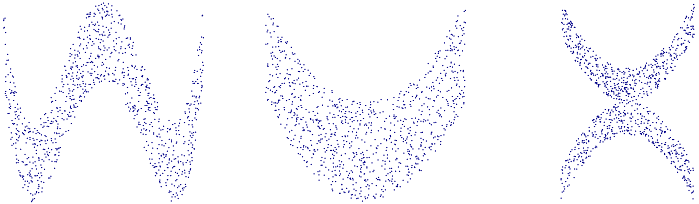
Scatterplot with multivariate distribution (contour lines) and marginal density X_1, $X_2 \sim N(0, 1)$

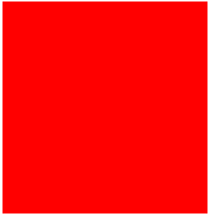
$\rho(X_1, X_2) = 0$ $\quad \rho(X_1, X_2) = 0.8$ $\quad \rho(X_1, X_2) = -0.8$



Examples with Pearson's correlation $\rho = 0$ but non-linear dependencies (MI $\neq 0$):

$\rho(X_1, X_2) = 0$, $MI(X_1, X_2) = 0.52$ $\quad \rho(X_1, X_2) = 0.01$, $MI(X_1, X_2) = 0.37$ $\quad \rho(X_1, X_2) = -0.06$, $MI(X_1, X_2) = 0.61$



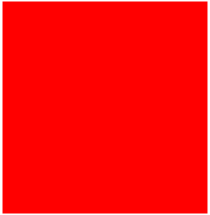


CORRELATION AND DEPENDENCE

Dependence: Describes general dependence structure of features (e.g., non-linear relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$



CORRELATION AND DEPENDENCE

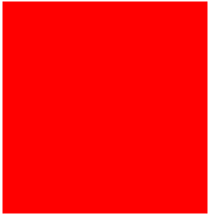
Dependence: Describes general dependence structure of features (e.g., non-linear relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k does not tell us anything about X_j and vice versa):

$$\mathbb{P}(X_j | X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k | X_j) = \mathbb{P}(X_k) \quad \text{(follows from conditional probability)}$$



CORRELATION AND DEPENDENCE

Dependence: Describes general dependence structure of features (e.g., non-linear relationships)

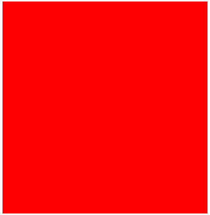
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k does not tell us anything about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \quad \text{(follows from conditional probability)}$$

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  $\leadsto$ Spearman correlation (measures monotonic dependencies via ranks)
  $\leadsto$ Information-theoretical measures like mutual information
  $\leadsto$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)



CORRELATION AND DEPENDENCE

Dependence: Describes general dependence structure of features (e.g., non-linear relationships)

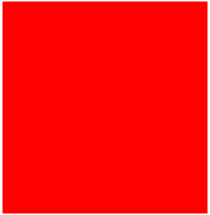
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k does not tell us anything about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \quad \text{(follows from conditional probability)}$$

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  $\leadsto$ Spearman correlation (measures monotonic dependencies via ranks)
  $\leadsto$ Information-theoretical measures like mutual information
  $\leadsto$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:** X_j, X_k independent $\Rightarrow \rho(X_j, X_k) = 0$ **but** $\rho(X_j, X_k) = 0 \not\Rightarrow X_j$, X_k independent
  But equivalency holds if distribution is jointly normal



CORRELATION AND DEPENDENCE

Dependence: Describes general dependence structure of features (e.g., non-linear relationships)

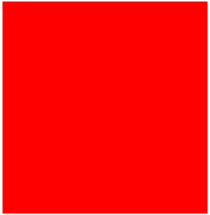
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k does not tell us anything about X_j and vice versa):

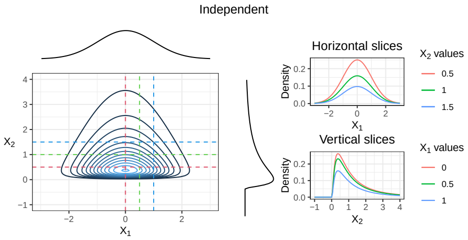
$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \quad \text{(follows from conditional probability)}$$

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:** X_j, X_k independent $\Rightarrow \rho(X_j, X_k) = 0$ **but** $\rho(X_j, X_k) = 0 \not\Rightarrow X_j$, X_k independent
  But equivalency holds if distribution is jointly normal
- $MI(X_j, X_k) = 0$ if and only if X_j, X_k independent



CORRELATION AND DEPENDENCE

Example:



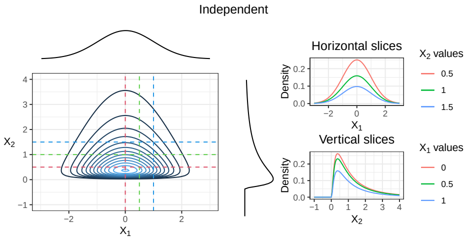
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

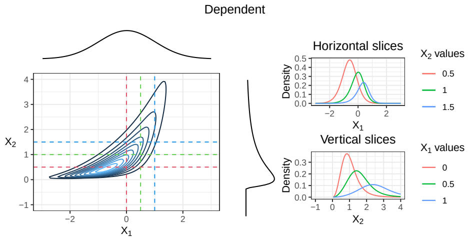
$\Rightarrow$

$\mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$ and $\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$

CORRELATION AND DEPENDENCE

Example:





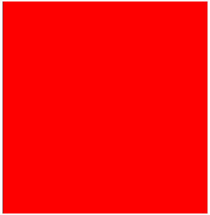
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

Conditional distributions do not match their marginal distributions

$\Rightarrow$
$\mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$ and $\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$

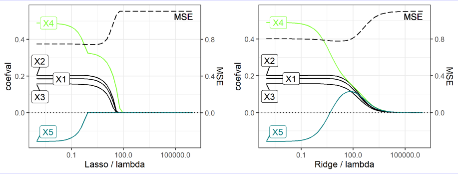
INTERPRETATIONS WITH DEPENDENT FEATURES

- Highly correlated features contain similar information
  $\rightsquigarrow$ Model might pick only one feature (regularization) (even if it is causally irrelevant)
  $\rightsquigarrow$ Produced explanations can be misleading (true to the model, but not to the data)
  $\rightsquigarrow$ Different IML models often produce different results in these situation, and not always trivial to understand which / why



INTERPRETATIONS WITH DEPENDENT FEATURES

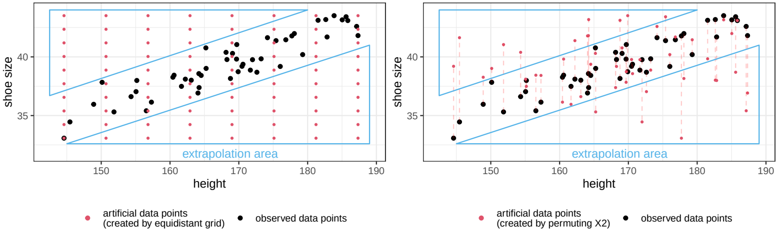
- Highly correlated features contain similar information
  $\rightsquigarrow$ Model might pick only one feature (regularization) (even if it is causally irrelevant)
  $\rightsquigarrow$ Produced explanations can be misleading (true to the model, but not to the data)
  $\rightsquigarrow$ Different IML models often produce different results in these situation, and not always trivial to understand which / why



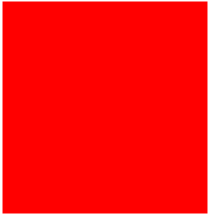
Fictional example for the model $y = 0.2X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4 + 0.2X_5 + \epsilon$ of 100 observations, $\epsilon \sim \mathcal{N}(0, 1)$. X_1-X_4 are independently drawn from different normal distributions: $X_1, X_2, X_3, X_4 \sim \mathcal{N}(0, 2)$. While X_1-X_4 have pairwise correlation coefficients of 0, X_4 and X_5 are nearly perfectly correlated: $X_5 = X_4 + \delta, \delta \sim \mathcal{N}(0, 0.3), \rho(X_4, X_5) = 0.98$.

We see that Lasso shrinks the coefficient for X_5 to zero early on, while Ridge assigns similar coefficients to X_4, X_5 for larger λ.

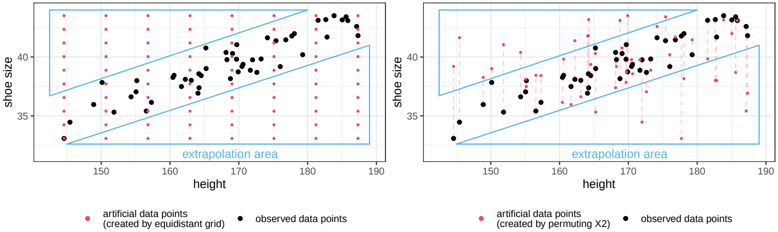
EXTRAPOLATION DUE TO DEPENDENCIES



- Many interpretation methods are based on artificially created data points
  ⇝ Many of these points can lie in low-density regions if features are dependent
  ⇝ Model predictions in such regions are subject to a high uncertainty
  ⇝ Explanations may be biased as they often rely on predictions where model extrapolated



EXTRAPOLATION DUE TO DEPENDENCIES



- Many interpretation methods are based on artificially created data points
  ⇝ Many of these points can lie in low-density regions if features are dependent
  ⇝ Model predictions in such regions are subject to a high uncertainty
  ⇝ Explanations may be biased as they often rely on predictions where model extrapolated
- There is no definition of when a model extrapolates and to what degree
  ⇝ Severity of extrapolation depends on model, some extrapolate more than others
  ⇝ Training density might serve as proxy to identify regions where extrapolation is likely
  But: Density estimation in many dimensions is often infeasible