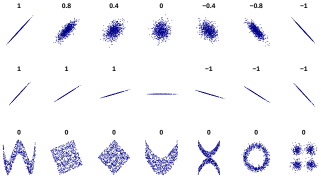
Interpretable Machine Learning

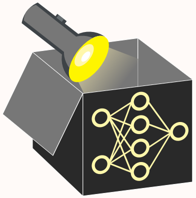
Correlation and Dependencies



Learning goals

- Difference of dependence vs. correlation
- Role of feature dependence in IML

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



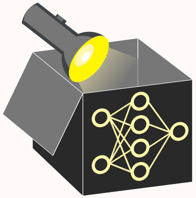
For two discrete random variables X_1, X_2:

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

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

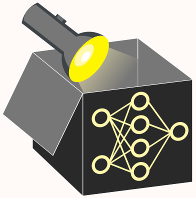
Marginal distribution

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

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

$\rightsquigarrow$ In continuous case with integrals

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

$\rightsquigarrow$ In continuous case with integrals

Conditional distribution

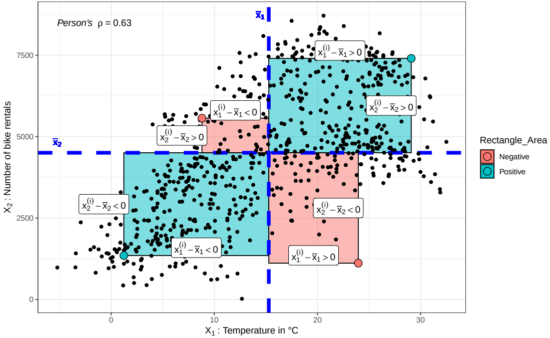
$$p_{X_1|X_2}(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2)$$
$$= \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

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

PEARSON'S CORRELATION COEFFICIENT ρ

Correlation often refers to Pearson's correlation (measures only **linear relationship**)

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



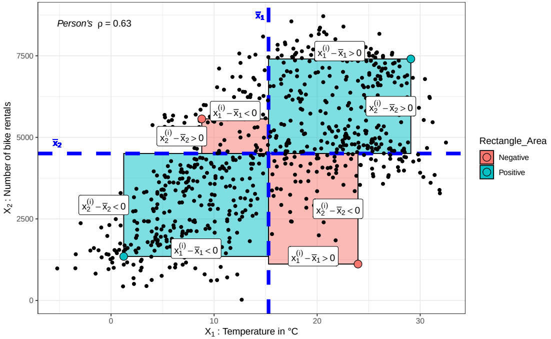
Geometric interpretation of ρ:

- 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 position
- Denominator scales the sum

PEARSON'S CORRELATION COEFFICIENT ρ

Correlation often refers to Pearson's correlation (measures only **linear relationship**)

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

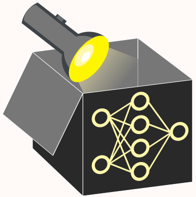


Geometric interpretation of ρ:

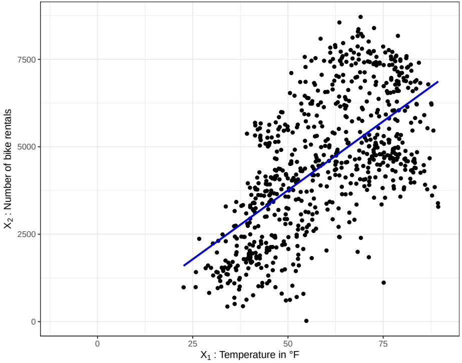
- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on position
- Denominator scales the sum

- $\rho > 0$ 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
- $\rho = 0$ if area of rectangles cancels out $\rightsquigarrow$ X_1, X_2 linearly uncorrelated

COEFFICIENT OF DETERMINATION R^2



Another method to evaluate **linear dependency** between features is R^2



Idea for two-dimensional case:

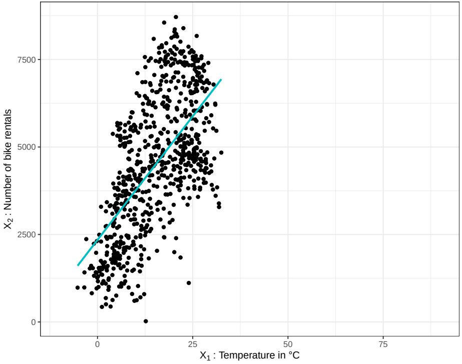
- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

$\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence

$\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence

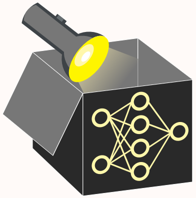
COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between features is R^2

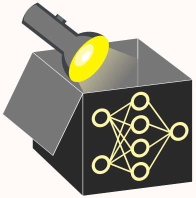


Idea for two-dimensional case:

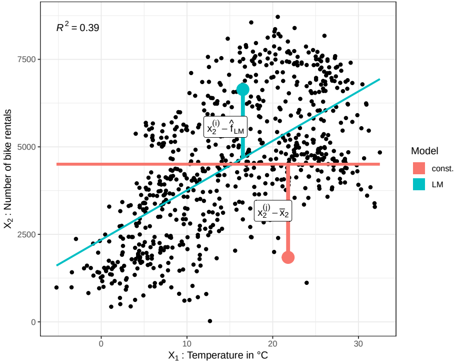
- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence
- $\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
- $\rightsquigarrow$ Re-scaling of x_1 or x_2 changes θ_1
- $\rightsquigarrow$ ℉ $\rightarrow$ ℃ $\Rightarrow \theta_1 = 78.5 \rightarrow \theta_1^* = 141.3$



COEFFICIENT OF DETERMINATION R^2



Another method to evaluate **linear dependency** between features is R^2



Idea for two-dimensional case:

- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
  - ⇝ Slope $\theta_1 = 0 \Rightarrow$ no dependence
  - ⇝ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
  - ⇝ Re-scaling 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 $p(x_1, x_2)$ and the product of their marginal distribution $p(x_1)p(x_2)$:

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

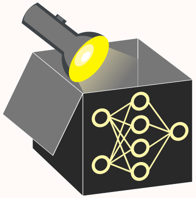
- MI measures amount of "dependence" between features
  $\rightsquigarrow$ It is zero if and only if the features are independent
- Unlike (Pearson) correlation, MI is not limited to continuous random variables

MUTUAL INFORMATION: EXAMPLE 1

$\mathbf{X_1}$	...	**Y**
yes	...	yes
yes	...	yes
yes	...	yes
yes	...	no
yes	...	no
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.5	0	0.5
$\mathbb{P}(Y = \text{no})$	0.333	0.167	0.5
p_{X_1}	0.833	0.167	1

MUTUAL INFORMATION: EXAMPLE 1



$\mathbf{X_1}$	...	**Y**
yes	...	yes
yes	...	yes
yes	...	yes
yes	...	no
yes	...	no
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.5	0	0.5
$\mathbb{P}(Y = \text{no})$	0.333	0.167	0.5
p_{X_1}	0.833	0.167	1

$$MI(X_1; Y) = \sum_{x_1 \in \mathcal{X}_1} \sum_{y \in \mathcal{Y}} p(x_1, y) log \left(\frac{p(x_1, y)}{p(x_1)p(y)} \right)$$

$$= \mathbb{P}(X_1 = \text{yes}, Y = \text{yes}) \; log \left(\frac{\mathbb{P}(X_1 = \text{yes}, Y = \text{yes})}{\mathbb{P}(X_1 = \text{yes})\mathbb{P}(Y = \text{yes})} \right)$$

$$+ \mathbb{P}(X_1 = \text{yes}, Y = \text{no}) \; log \left(\frac{\mathbb{P}(X_1 = \text{yes}, Y = \text{no})}{\mathbb{P}(X_1 = \text{yes})\mathbb{P}(Y = \text{no})} \right)$$

$$+ \mathbb{P}(X_1 = \text{no}, Y = \text{yes}) \; log \left(\frac{\mathbb{P}(X_1 = \text{no}, Y = \text{yes})}{\mathbb{P}(X_1 = \text{no})\mathbb{P}(Y = \text{yes})} \right)$$

$$+ \mathbb{P}(X_1 = \text{no}, Y = \text{no}) \; log \left(\frac{\mathbb{P}(X_1 = \text{no}, Y = \text{no})}{\mathbb{P}(X_1 = \text{no})\mathbb{P}(Y = \text{no})} \right)$$

MUTUAL INFORMATION: EXAMPLE 1

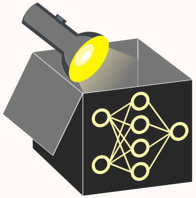
$\mathbf{X_1}$	...	**Y**
yes	...	yes
yes	...	yes
yes	...	yes
yes	...	no
yes	...	no
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.5	0	0.5
$\mathbb{P}(Y = \text{no})$	0.333	0.167	0.5
p_{X_1}	0.833	0.167	1

$$
\begin{aligned}
MI(X_1; Y) = & \; 0.5 \, log\left(\frac{0.5}{0.833 \cdot 0.5}\right) + 0.333 \, log\left(\frac{0.833}{\cdot 0.5}\right) \\
& + 0 \, log\left(\frac{0}{0.167 \cdot 0.5}\right) + 0.167 \, log\left(\frac{0.167}{0.167 \cdot 0.5}\right) \\
= & \; 0.133
\end{aligned}
$$

Note: $lim_{x \to 0} \, x \, log\left(\frac{x}{c}\right) = 0, \; c \in \mathbb{R}$

MUTUAL INFORMATION: EXAMPLE 2



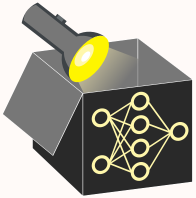
$\mathbf{X}_1$	...	**Y**
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

MUTUAL INFORMATION: EXAMPLE 2

$\mathbf{X_1}$	...	**Y**
yes	...	yes
yes	...	no
no	...	yes
no	...	no

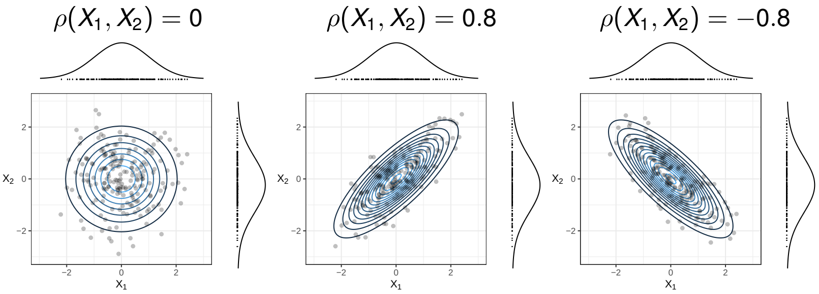
	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1



$$\begin{aligned} MI(X_1; Y) &= 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &\quad + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &= 0.25\, log\left(\frac{0.25}{0.25}\right) \cdot 4 \\ &= 0.25\, log\,(1) \cdot 4 \\ &= 0 \end{aligned}$$

CORRELATION AND DEPENDENCE

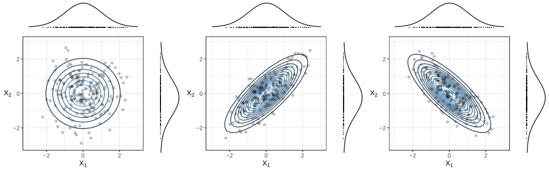
Scatterplot with multivariate distribution (contour lines) and marginal density
$X_1, X_2 \sim N(0,1)$



CORRELATION AND DEPENDENCE

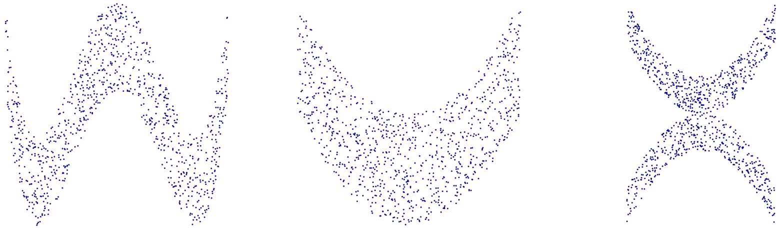
Scatterplot with multivariate distribution (contour lines) and marginal density
$X_1, X_2 \sim N(0, 1)$

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



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

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



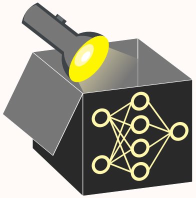
CORRELATION AND DEPENDENCE

Dependence: Describes general dependence structure (e.g., non-lin. 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 (e.g., non-lin. relationships)

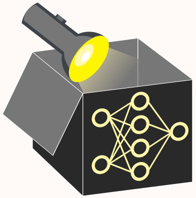
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

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

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

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

CORRELATION AND DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

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

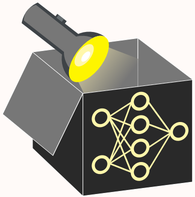
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

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

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

- 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 (e.g., non-lin. relationships)

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

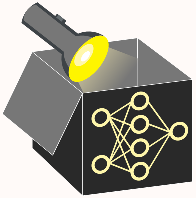
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

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

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

- 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
  Equivalency holds if distribution is jointly normal

CORRELATION AND DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

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

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

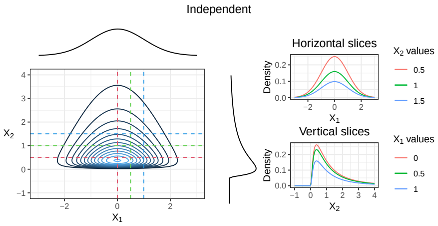
- Equivalent definition (knowing X_k does not say anything about X_j and vice versa):

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

- 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
  Equivalency holds if distribution is jointly normal
- $MI(X_j, X_k) = 0$ if and only if X_j, X_k independent

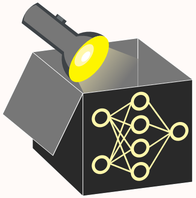
CORRELATION AND DEPENDENCE

Example:



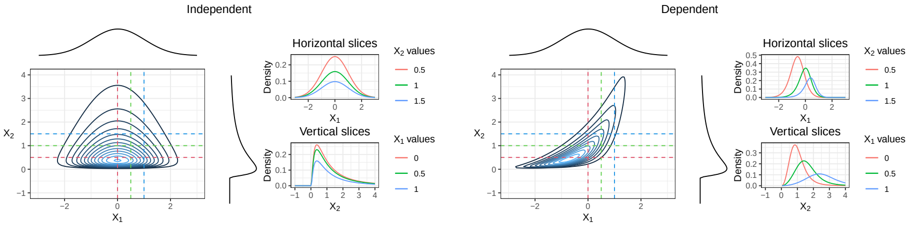
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

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



CORRELATION AND DEPENDENCE

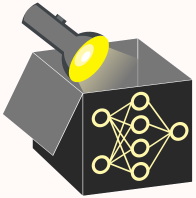
Example:



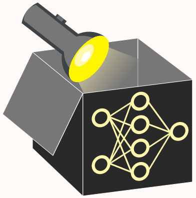
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

Conditional distributions do not match their marginal distributions

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

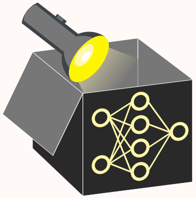


INTERPRETATIONS WITH DEPENDENT FEATURES

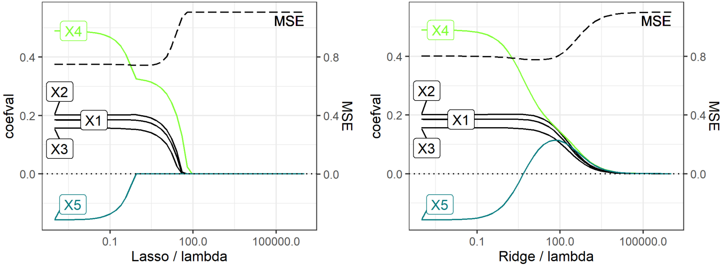


- Highly correlated features contain similar information
  $\rightsquigarrow$ Model might pick only 1 feat. (regularization), even if it is causally irrelevant
  $\rightsquigarrow$ Produced explanations can be misleading (true to model, but not to data)
  $\rightsquigarrow$ E.g., different interpretable models produce different results

INTERPRETATIONS WITH DEPENDENT FEATURES

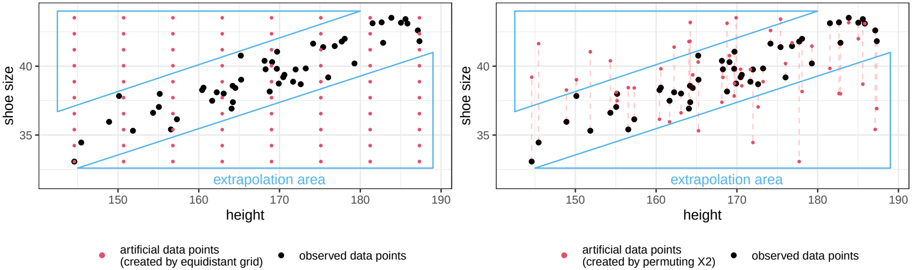


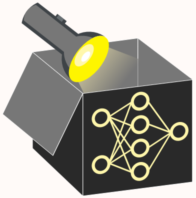
- Highly correlated features contain similar information
  - $\rightsquigarrow$ Model might pick only 1 feat. (regularization), even if it is causally irrelevant
  - $\rightsquigarrow$ Produced explanations can be misleading (true to model, but not to data)
  - $\rightsquigarrow$ E.g., different interpretable models produce different results
- **Example:** Simulate 100 obs. from DGP $Y = 0.2(X_1 + \cdots + X_5) + \epsilon, \epsilon \sim N(0, 1)$



- $X_1, \ldots, X_4 \sim N(0, 2)$ (uncorrelated)
- $X_5 = X_4 + \delta, \delta \sim N(0, 0.3) \Rightarrow \rho(X_4, X_5) = 0.98$ (highly correlated)
- LASSO: Shrinks coef. of X_5 to zero, coef. of X_4 about $1.5\times$ higher
- Ridge: Similar coef. for X_4 and X_5 for higher lambda

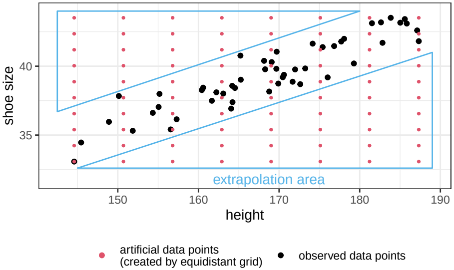
EXTRAPOLATION DUE TO DEPENDENCIES

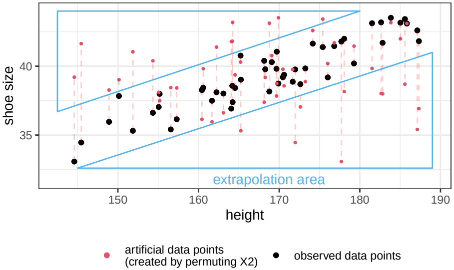


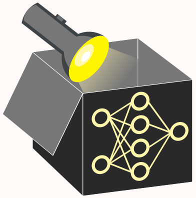


- Many interpretation methods are based on artificially created data points
  - ⇝ Many points lie in low-density regions if features are dependent
  - ⇝ Predictions in such regions have high uncertainty
  - ⇝ Explanations can be biased if they rely on pred. where model extrapolated

EXTRAPOLATION DUE TO DEPENDENCIES







- Many interpretation methods are based on artificially created data points
  - $\rightsquigarrow$ Many points lie in low-density regions if features are dependent
  - $\rightsquigarrow$ Predictions in such regions have high uncertainty
  - $\rightsquigarrow$ Explanations can be biased if they rely on pred. where model extrapolated
- There is no definition of when a model extrapolates and to what degree
  - $\rightsquigarrow$ Severity of extrapolation depends on model
  - $\rightsquigarrow$ Density of train data may helps identify regions where extrapolation is likely
    But: Density estimation in many dimensions is often infeasible