The Correlation Coefficient

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

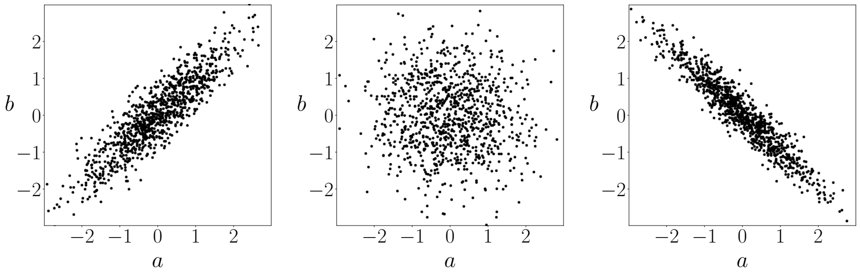
Plan

Define correlation coefficient

Show that it parametrizes dependence between Gaussian random variables

Goal

Quantify dependence between two quantities with a single number



Idea: Focus on *linear* dependence

Linear dependence

How can we quantify linear dependence between random variables $\tilde{a}$ and $\tilde{b}$?

Approximate $\tilde{b}$ using linear function of $\tilde{a}$

Standardized random variables

We focus on random variables with zero mean and unit variance

Linear estimator

Goal: Find best linear estimate $\beta \tilde{a}$ of $\tilde{b}$ given $\tilde{a}$

Assumption: $\mathrm{E}[\tilde{a}] = \mathrm{E}[\tilde{b}] = 0$, $\mathrm{Var}[\tilde{a}] = \mathrm{Var}[\tilde{b}] = 1$

We minimize the mean squared error

$$\begin{aligned}
\mathrm{MSE}(\beta) := \mathrm{E}\left[(\tilde{b} - \beta\tilde{a})^2\right] &= \mathrm{E}\left[\tilde{b}^2 - 2\beta\tilde{a}\tilde{b} + \beta^2\tilde{a}^2\right] \\
&= \mathrm{E}[\tilde{b}^2] + \beta^2\mathrm{E}[\tilde{a}^2] - 2\beta\mathrm{E}[\tilde{a}\tilde{b}] \\
&= 1 + \beta^2 - 2\beta\mathrm{E}[\tilde{a}\tilde{b}]
\end{aligned}$$

Linear minimum MSE estimator

$$\mathsf{MSE}(\beta) = 1 + \beta^2 - 2\beta \mathrm{E}[\tilde{a}\tilde{b}]$$

$$\mathsf{MSE}'(\beta) = 2\beta - 2\mathrm{E}[\tilde{a}\tilde{b}]$$

$$\mathsf{MSE}''(\beta) = 2$$

$$\beta_{\mathsf{MMSE}} = \mathrm{E}[\tilde{a}\tilde{b}] := \rho_{\tilde{a},\tilde{b}}$$

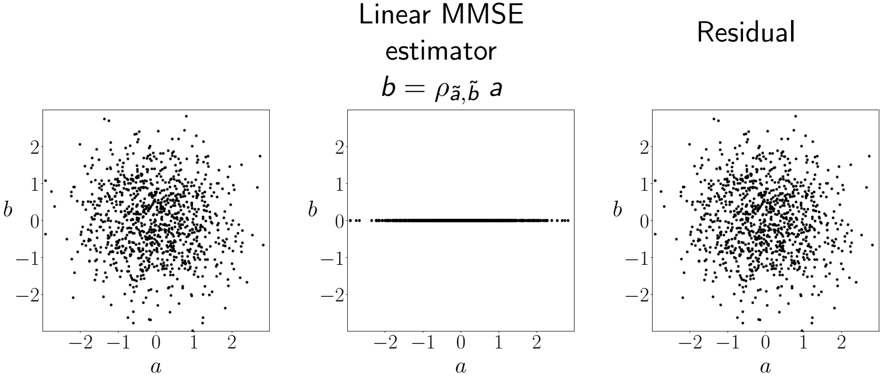
Decomposition

$$\tilde{b} = \underbrace{\rho_{\tilde{a},\tilde{b}}\ \tilde{a}}_{\text{Best linear estimate given } \tilde{a}} + \underbrace{\tilde{b} - \rho_{\tilde{a},\tilde{b}}\ \tilde{a}}_{\text{Residual}}$$

$$-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$$

$\rho_{\tilde{a},\tilde{b}} = 0$

If $\rho_{\tilde{a},\tilde{b}} = 0$, $\tilde{a}$ and $\tilde{b}$ are uncorrelated

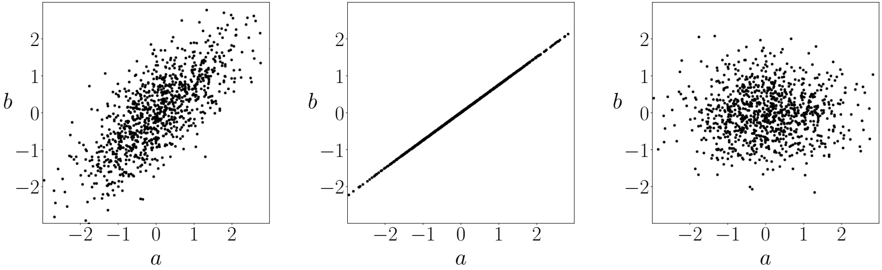


$\rho_{\tilde{a},\tilde{b}} = 0.75$

If $\rho_{\tilde{a},\tilde{b}} > 0$, $\tilde{a}$ and $\tilde{b}$ are positively correlated

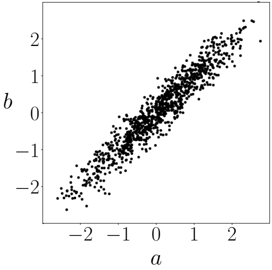
Linear MMSE estimator $b = \rho_{\tilde{a},\tilde{b}}\, a$

Residual



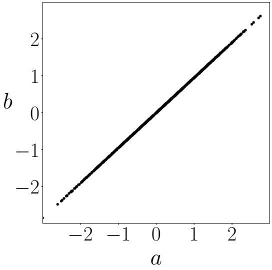
$\rho_{\tilde{a},\tilde{b}} = 0.95$

If $\rho_{\tilde{a},\tilde{b}} > 0$, $\tilde{a}$ and $\tilde{b}$ are positively correlated

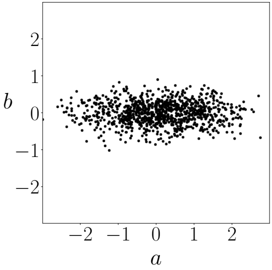


Linear MMSE estimator

$b = \rho_{\tilde{a},\tilde{b}}\ a$



Residual

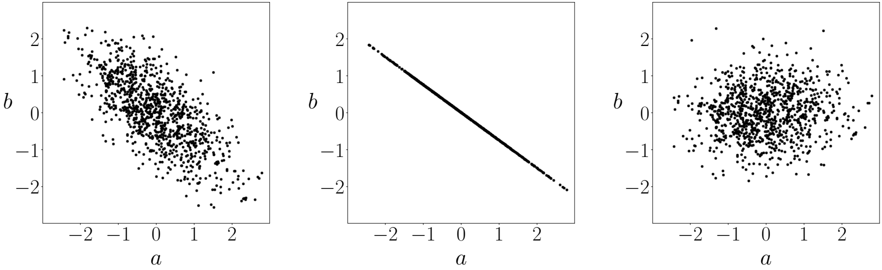


$\rho_{\tilde{a},\tilde{b}} = -0.75$

If $\rho_{\tilde{a},\tilde{b}} < 0$, $\tilde{a}$ and $\tilde{b}$ are negatively correlated

Linear MMSE estimator $b = \rho_{\tilde{a},\tilde{b}}\ a$

Residual

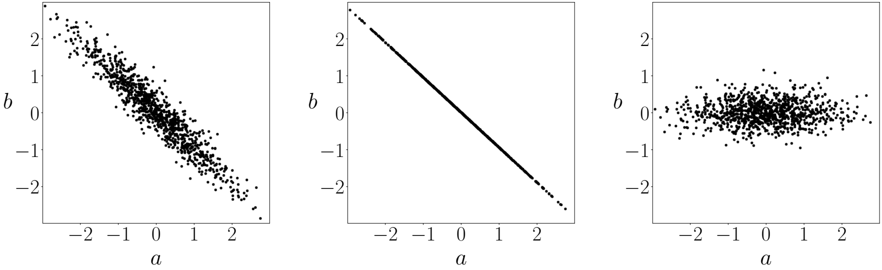


$\rho_{\tilde{a},\tilde{b}} = -0.95$

If $\rho_{\tilde{a},\tilde{b}} > 0$, $\tilde{a}$ and $\tilde{b}$ are negatively correlated

Linear MMSE estimator $b = \rho_{\tilde{a},\tilde{b}}\ a$

Residual



Gaussian random vector

A Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ is the mean and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix

$\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive definite (positive eigenvalues)

2D Gaussian

Gaussian random vector $(\tilde{a}, \tilde{b})$ with zero mean and covariance matrix

$$\Sigma := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \qquad -1 < \rho < 1$$

$$\begin{aligned} f_{\tilde{a},\tilde{b}}(a,b) &:= \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left(-\frac{1}{2} \begin{bmatrix} a \\ b \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} a \\ b \end{bmatrix}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right) \\ &= f_{\tilde{a}}(a)\, f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) \end{aligned}$$

Marginal and conditional distributions

$$
\begin{aligned}
f_{\tilde{a},\tilde{b}}(a,b) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right) \\
&= f_{\tilde{a}}(a)\, f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a)
\end{aligned}
$$

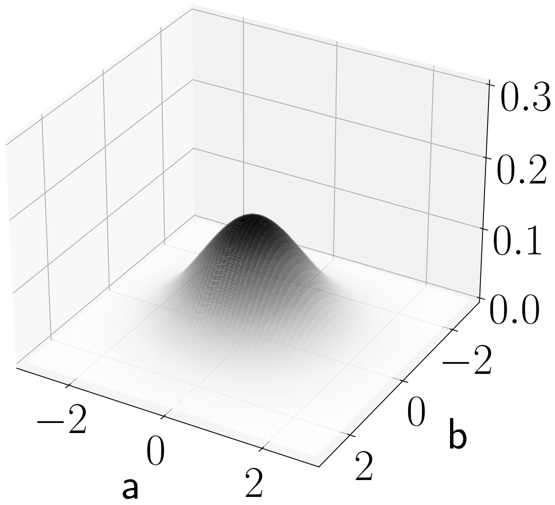
Marginal distribution of $\tilde{a}$?

Gaussian: mean = 1, standard deviation = 1

Conditional distribution of $\tilde{b}$ given $\tilde{a} = a$?

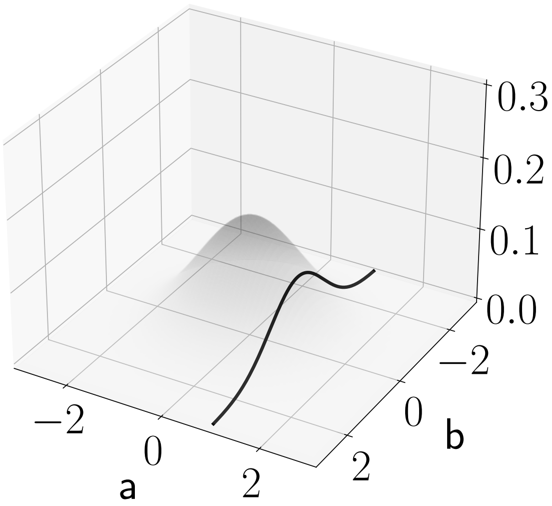
Gaussian: mean = ρa, standard deviation = $\sqrt{1-\rho^2}$

$\rho = 0$



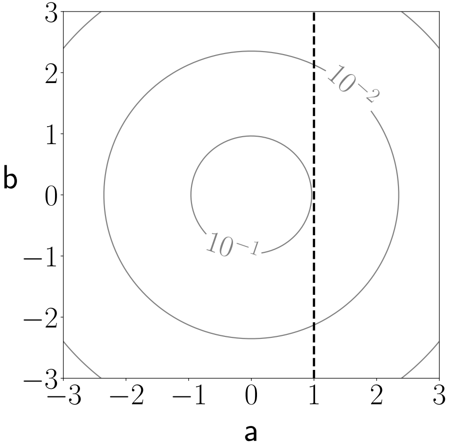
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0$

$\mu = \rho a = 0$, $\sigma = \sqrt{1 - \rho^2} = 1$



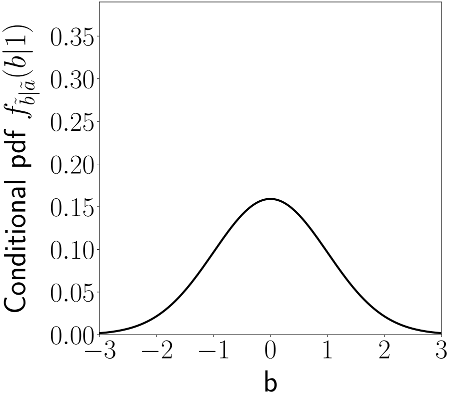
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0$

$\mu = \rho a = 0$, $\sigma = \sqrt{1 - \rho^2} = 1$

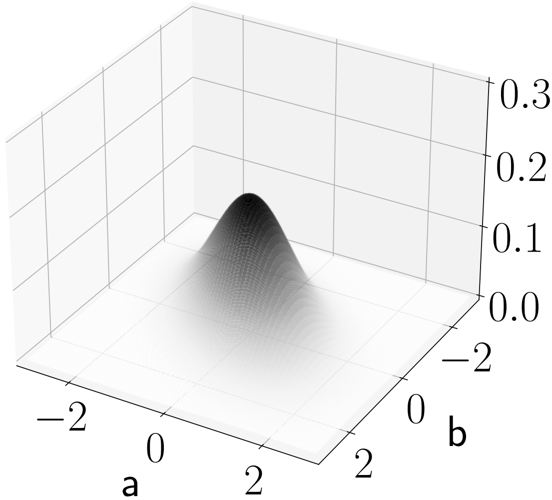


Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0$

$\mu = \rho a = 0$, $\sigma = \sqrt{1 - \rho^2} = 1$

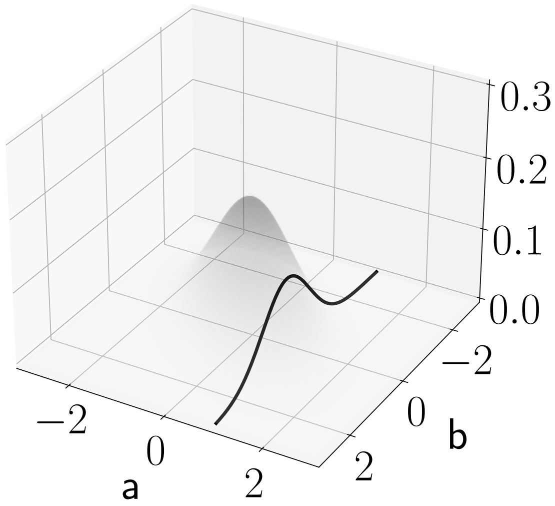


$\rho = 0.5$



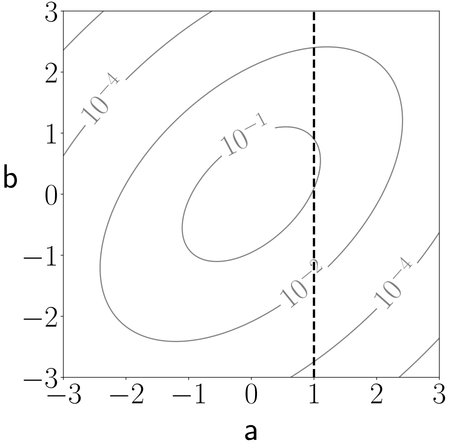
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.5$

$\mu = 0.5a$, $\sigma = \sqrt{1 - \rho^2} = 0.87$



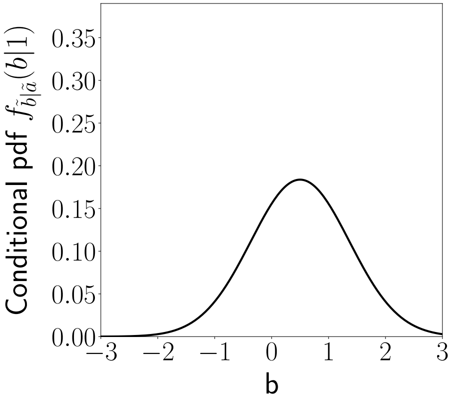
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.5$

$\mu = 0.5a$, $\sigma = \sqrt{1 - \rho^2} = 0.87$

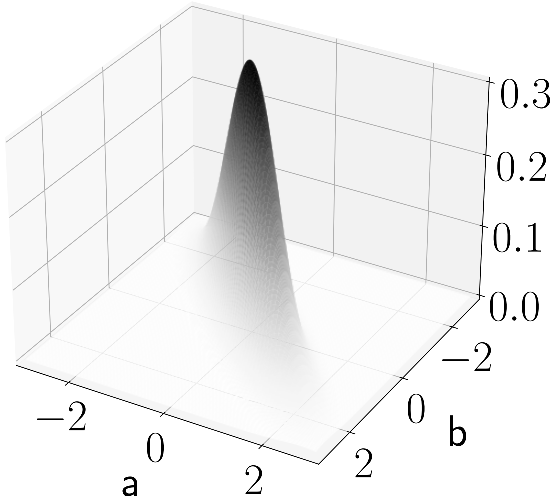


Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.5$

$\mu = 0.5a$, $\sigma = \sqrt{1 - \rho^2} = 0.87$

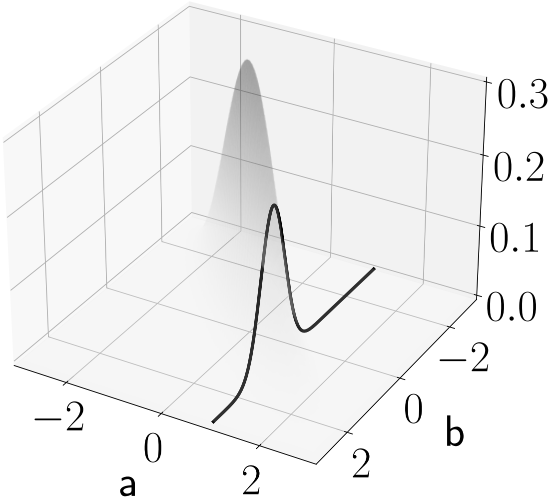


$\rho = 0.9$



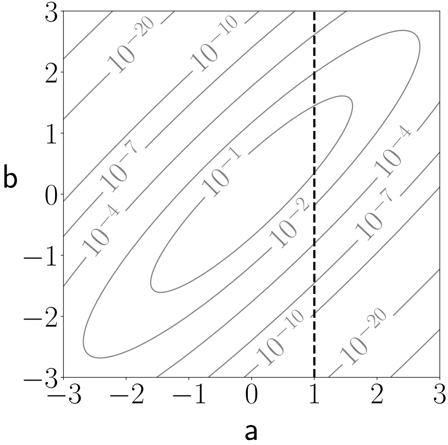
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.9$

$\mu = 0.9a$, $\sigma = \sqrt{1 - \rho^2} = 0.44$



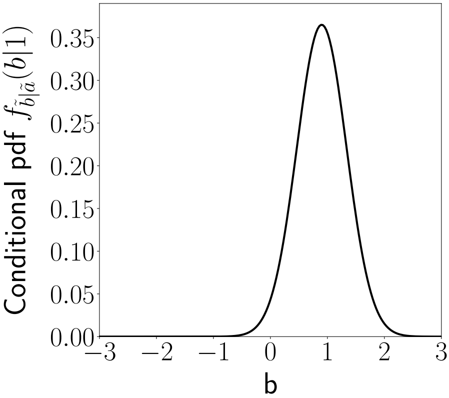
Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.9$

$\mu = 0.9a$, $\sigma = \sqrt{1 - \rho^2} = 0.44$

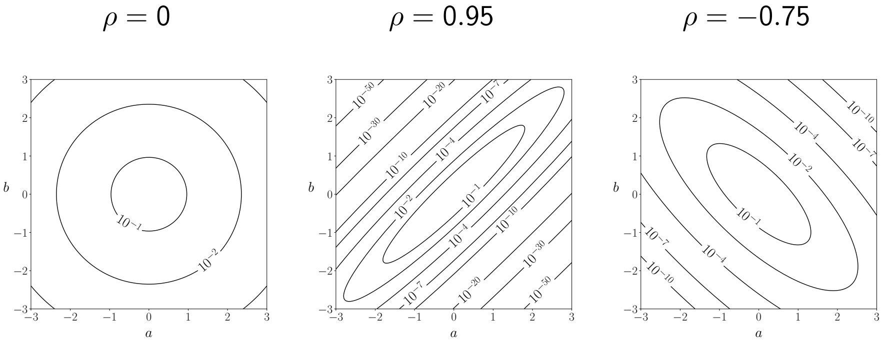


Conditional distribution of $\tilde{b}$ given $\tilde{a} = 1$ if $\rho = 0.9$

$\mu = 0.9a$, $\sigma = \sqrt{1 - \rho^2} = 0.44$



ρ dictates dependence between $\tilde{a}$ and $\tilde{b}$



Correlation coefficient of $\tilde{a}$ and $\tilde{b}$?

Marginal distribution of $\tilde{a}$ and $\tilde{b}$?

Gaussian: mean = 1, variance = 1

$$\rho_{\tilde{a},\tilde{b}} = \mathrm{E}\left[\tilde{a}\tilde{b}\right]$$

Correlation coefficient of $\tilde{a}$ and $\tilde{b}$

We apply iterated expectation

Conditional distribution of $\tilde{b}$ given $\tilde{a} = a$

Gaussian: mean $= \rho a$, standard deviation $= \sqrt{1 - \rho^2}$

$$\begin{aligned}
\mu_{\tilde{a}\tilde{b} \,|\, \tilde{a}}(a) &= \int_{a=-\infty}^{\infty} ab\, f_{\tilde{b} \,|\, \tilde{a}}(b \,|\, a)\, \mathrm{d}b \\
&= a\, \mu_{\tilde{b} \,|\, \tilde{a}}(a) \\
&= \rho a^2
\end{aligned}$$

$$\begin{aligned}
\rho_{\tilde{a},\tilde{b}} = \mathrm{E}\left[\tilde{a}\tilde{b}\right] &= \mathrm{E}\left[\mu_{\tilde{a}\tilde{b} \,|\, \tilde{a}}(\tilde{a})\right] \\
&= \mathrm{E}\left[\rho \tilde{a}^2\right] \\
&= \rho \mathrm{E}\left[\tilde{a}^2\right] \\
&= \rho
\end{aligned}$$

What have we learned?

Correlation coefficient quantifies linear dependence between random variables (with zero mean and unit variance)

Correlation coefficient parametrizes dependence between Gaussians

What have we not learned yet?

Correlation coefficient between random variables with nonzero mean or non-unit variance

Why $-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$

How to compute the correlation coefficient from data