EPFL

la manno lab

Joint and Bivariate Distributions

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

March 2025

EPFL - BMI - UPLAMANNO

Contents

Joint Probability Distributions

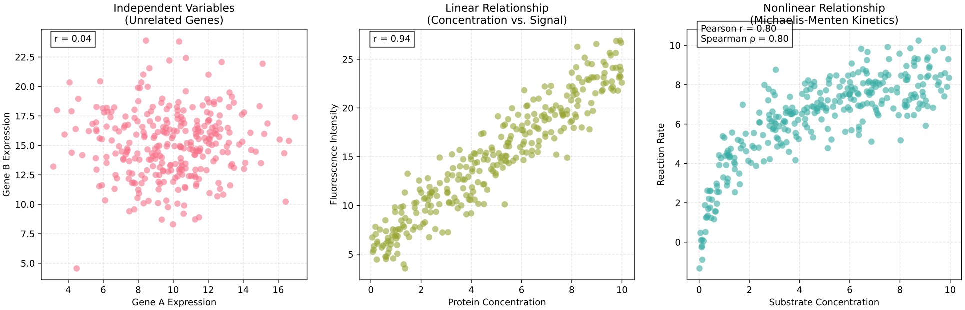
From Individual Variables to Relationships

In biological systems, variables rarely exist in isolation:

- Gene expression levels influence each other
- Protein concentrations depend on multiple factors
- Cellular behaviors arise from interacting components

To understand these complex relationships, we need to move beyond studying single random variables in isolation.

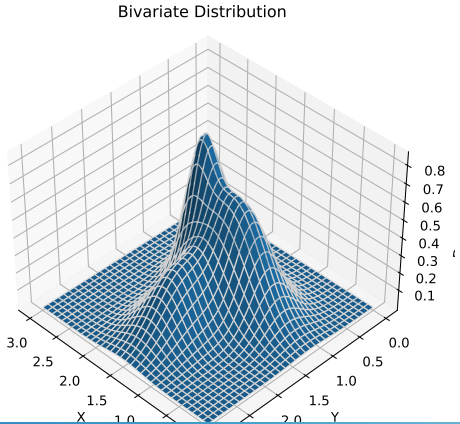
Patterns of Relationship in Biological Data

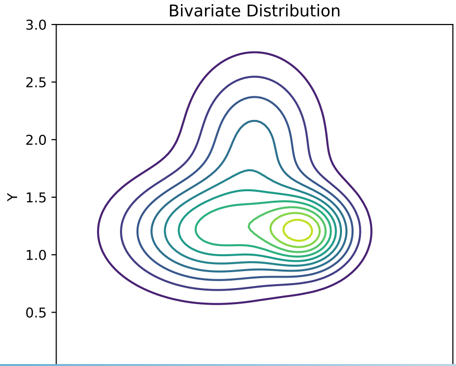


Visualizing Joint Distributions

For univariate data: 2D plots (value vs density) For bivariate data: 3D plots (two values vs density)

Alternative 2D representations: contour plots and heat maps



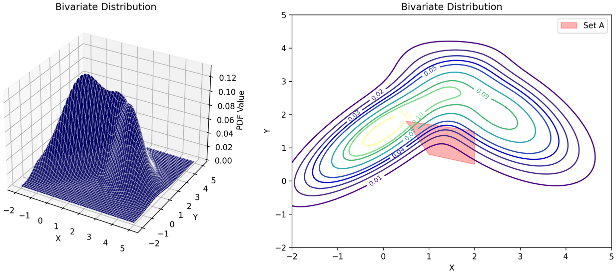


Defining Joint Probability Distributions

Definition (Joint Probability Distribution)

For continuous random variables X and Y, their joint probability density function $f_{X,Y}(x, y)$ satisfies for any measurable set A in the plane:

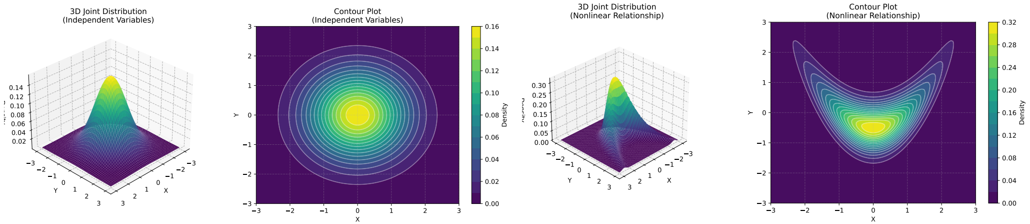
$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y)dxdy$$



Level Sets: Windows into Distribution Structure

Level sets are curves along which the probability density remains constant:

$$L_c = \{(x, y) \in \mathbb{R}^2 : f_{X,Y}(x, y) = c\}$$



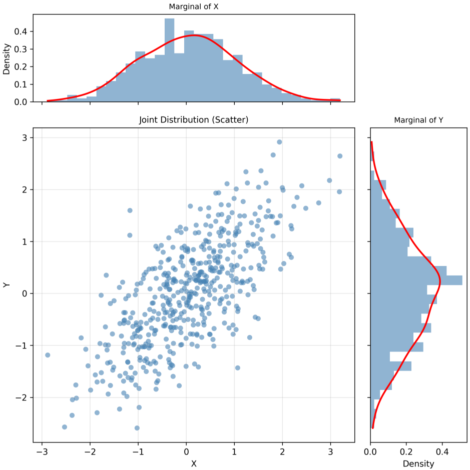
Marginal Distributions: where does the name come from?

The term "marginal" has a historical origin:

Early statisticians arranged bivariate data in contingency tables with totals calculated in the table margins.

These "margin sums" became known as marginal distributions.

Marginals tell us about a single variable when we disregard information about the other variable.

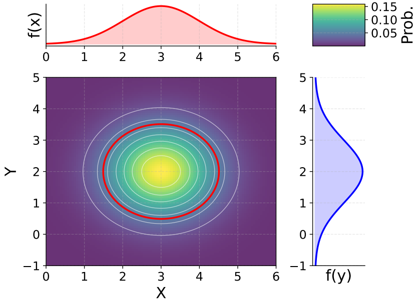


Formal Definition of Marginal Distributions

Definition (Marginal Distribution)

For joint density $f_{X,Y}(x, y)$, the marginal density of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$



"Integrating out"

This corresponds to "integrating out" or "marginalizing out" the other variable.

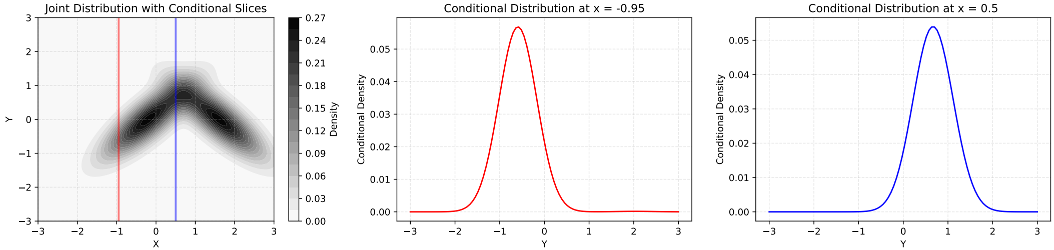
For any set A where X might take values:

$$P(X \in A) = \int_A f_X(x)dx = \int_A \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx$$

Conditional Distributions: Fixing One Variable

One of the most powerful aspects of joint distributions is revealing dependencies between variables.

The key question: How does the distribution of one variable change when we fix the value of the other?



Conditional Distributions: Formal definition

Definition (Conditional Distribution)

The conditional density of Y given X = x is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Theorem (Chain Rule of Probability)

For any two random variables X and Y:

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$$

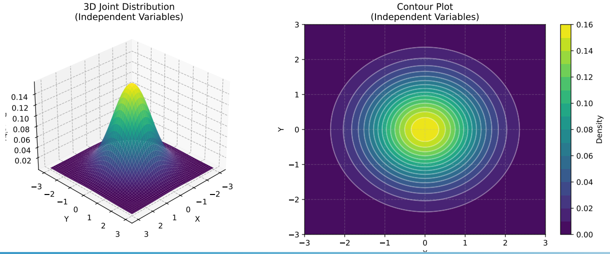
Independence: When Marginals Are Sufficient

A special case occurs when variables are independent - when knowing one variable tells us nothing about the other.

Definition (Statistical Independence)

Random variables X and Y are independent if and only if:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$



Independence: Implications and Importance

Independence is a strong assumption, rarely perfectly satisfied in biological systems, but it:

- Simplifies probability calculations
- Helps identify truly interacting components
- Provides null models for testing relationships

In practical terms, independence means:

- Knowing one variable gives no information about the other
- Joint distribution factorizes into product of marginals
- Conditional distribution equals the marginal distribution

The Bivariate Normal Distribution: A Fundamental Example

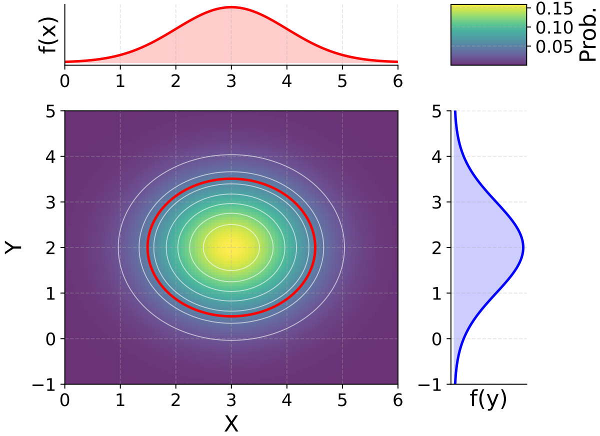
The bivariate normal extends the univariate normal to two dimensions:

For independent normal variables $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) =$$

$$\frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right)$$

The Bivariate Normal Distribution: A Fundamental Example



Level sets form concentric ellipses aligned with the coordinate axes.

The General Bivariate Normal Distribution

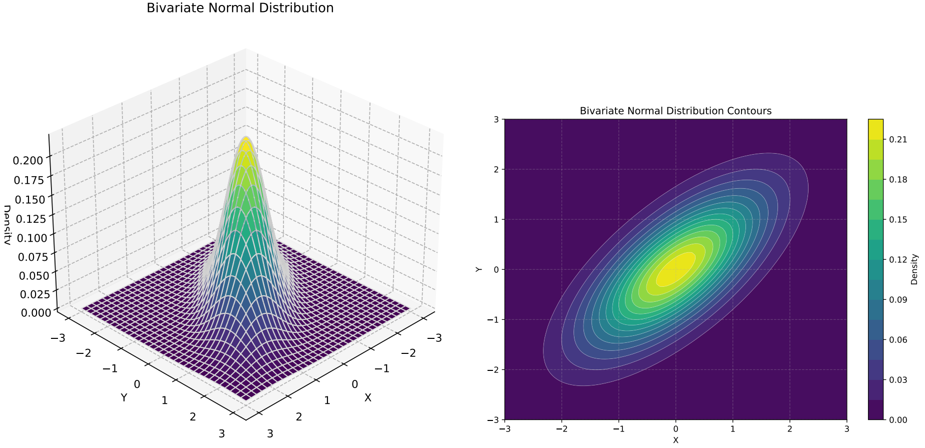
The general form includes correlation through parameter ρ:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right)$$

Key parameters:

- μ_1, μ_2: centers the distribution
- σ_1^2, σ_2^2: control spread along each axis
- ρ: correlation coefficient $(-1 \leq \rho \leq 1)$

The General Bivariate Normal Distribution: Visualization



Geometric Interpretation: Level Sets as Ellipses

The level sets of the bivariate normal form ellipses in the plane:

Setting the quadratic form in the exponent equal to a constant:

$$Q(x, y) = c(1 - \rho^2)$$

$$Q(x, y) = \frac{(x - \mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x - \mu_1)(y - \mu_2)}{\sigma_1\sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2}$$

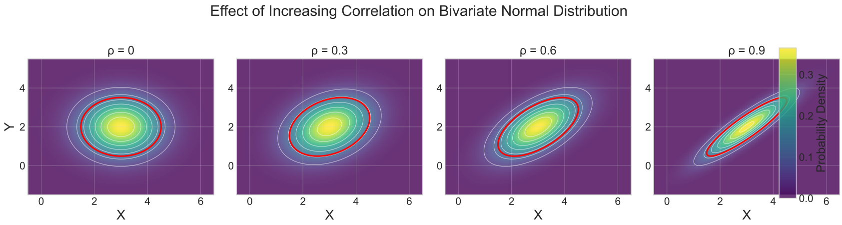
Comparing with the standard form of an ellipse reveals the geometric structure of the bivariate normal distribution.

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)}\left[\frac{(x - \mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x - \mu_1)(y - \mu_2)}{\sigma_1\sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2}\right]\right)$$

Geometric Interpretation: Properties of the Ellipses

The elliptical level sets of the bivariate normal distribution have the following properties:

- **Center**: Located at (μ_1, μ_2) - the mean of the distribution
- **Rotation**: Determined by the correlation coefficient ρ
- **Semi-axes**: Determined by σ_1, σ_2, and ρ



Marginals of the Bivariate Normal Distribution

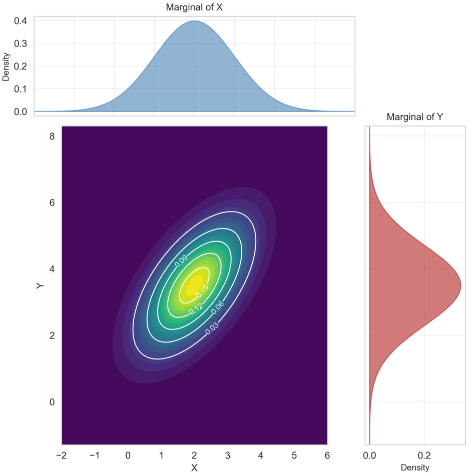
For a bivariate normal distribution with parameters μ_1, μ_2, σ_1, σ_2, and ρ:

The marginal distribution of X and Y are:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right)$$

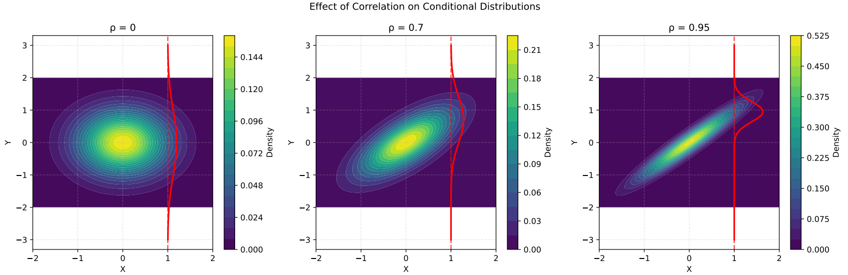
Do not depend on the correlation ρ !



Conditional Distributions of the Bivariate Normal

Applying the definition $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ one can find: The conditional distribution of Y given $X = x$ is a normal distribution with

- $\mu_{Y|X=x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$
- $\sigma^2_{Y|X=x} = \sigma^2_y(1 - \rho^2)$



Properties of Conditional Bivariate Normal

Key observations about the conditional distribution:

- Mean depends linearly on x with slope determined by ρ
- Variance is reduced by a factor of $(1 - \rho^2)$
- Stronger correlation means greater reduction in variance

This explains why knowing one variable in a correlated pair improves our prediction of the other variable - we have both a more precise mean estimate and reduced uncertainty.

Bivariate Estimates

From Theory to Practice: Estimating Bivariate Distributions

Having established the theoretical framework for bivariate distributions, we now move to practical estimation:

Biological systems often exhibit relationships that simple metrics like correlation cannot fully capture:

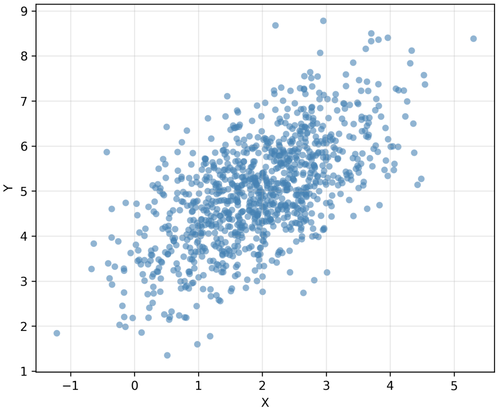
- Threshold effects in gene regulatory networks
- Saturation in enzyme kinetics
- Non-monotonic dose-response relationships

From Theory to Practice: Estimating Bivariate Distributions

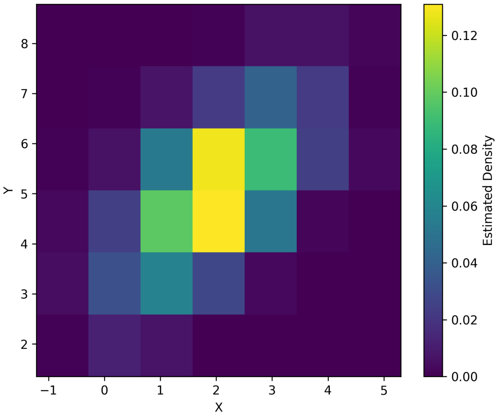
Our toolkit for bivariate estimation includes:

- Visualization tools: Scatter plots, heatmaps, contours
- Parametric approaches: Fitting specific distribution families
- Non-parametric methods: Letting the data speak for itself
- Information-theoretic measures: Quantifying general dependencies

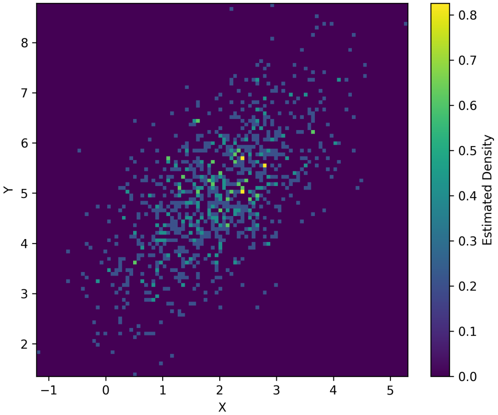
2D Histogram: From Data to Density Estimation



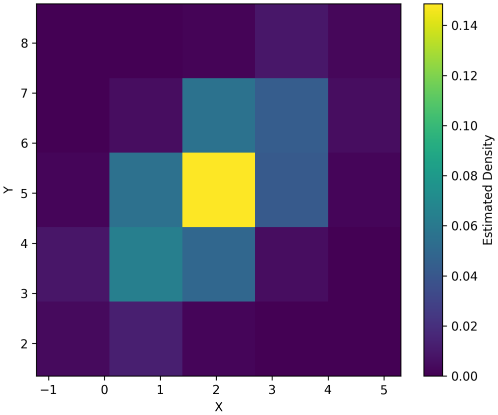
Raw data



Histogram 2d



Too small bins (noisy)

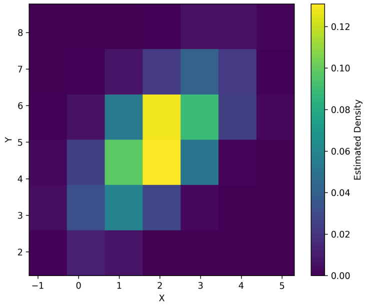


Too large bins (oversmoothed)

Limitations of Bivariate Histograms

While conceptually straightforward, 2D histograms face several challenges:

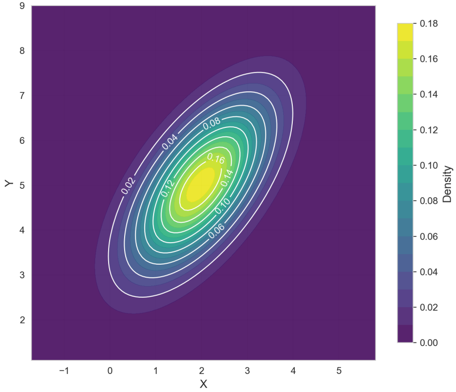
- **Limited resolution**: Constrained by bin size, creating blocky representations
- **Zero-probability regions**: Areas with no observations are assigned zero probability
- **Non-smooth appearance**: Discontinuous at bin boundaries, creating artificial edges
- **Curse of dimensionality**: Bins become increasingly sparse as dimensions increase



Despite limitations, histograms provide valuable initial insights for exploratory analysis.

Parametric Estimation of Joint Distributions

When we have prior knowledge about the distribution form, parametric estimation offers advantages:



Maximum Likelihood for the Bivariate Normal

For the bivariate normal distribution, we need to estimate five parameters

Theorem (MLE for Bivariate Normal)

Given data $\{(x_i, y_i)\}_{i=1}^n$*, the maximum likelihood estimators are:*

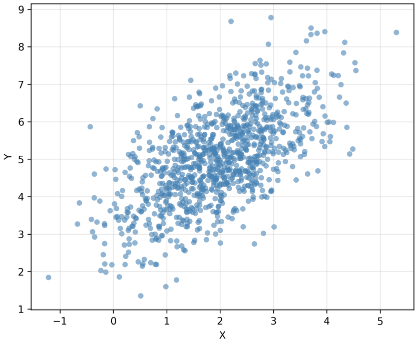
$$\hat{\mu}_1 = \bar{x} \quad \hat{\mu}_2 = \bar{y}$$

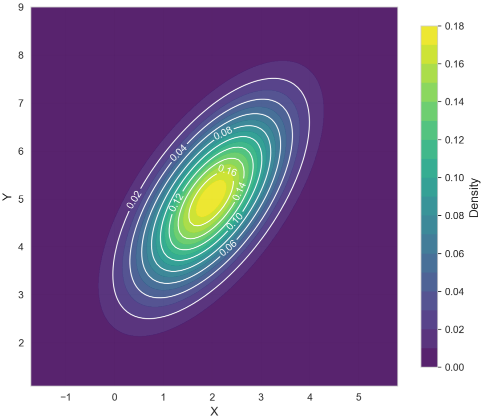
$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \quad \hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Remarkably, these are exactly the familiar sample statistics!

Maximum Likelihood for the Bivariate Normal: Visualization





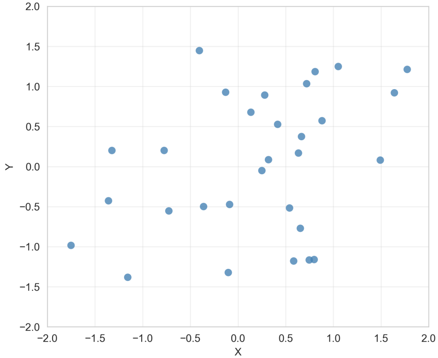
Non-parametric Estimation: Beyond Model Assumptions

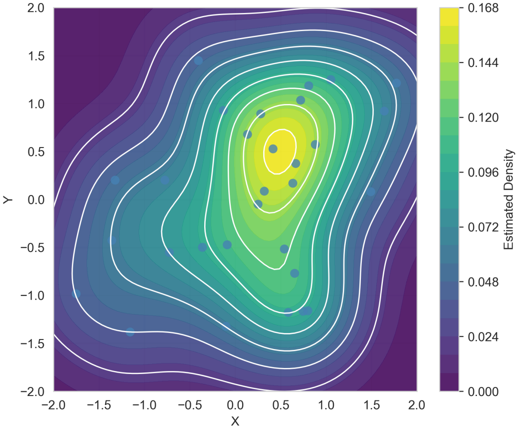
When the underlying structure doesn't conform to a specific parametric form, non-parametric methods offer flexibility.

Kernel Density Estimation (KDE) extends histograms by:

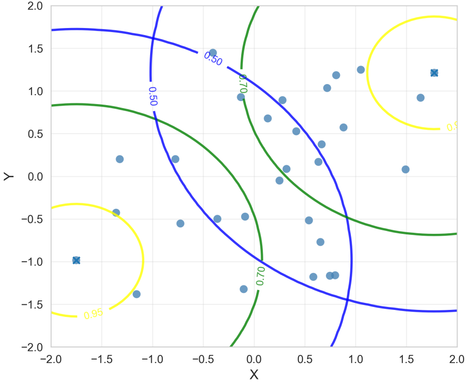
- Placing a probability density (kernel) at each data point
- Summing these kernels to create a smooth estimate
- Avoiding the "blocky" appearance of histograms

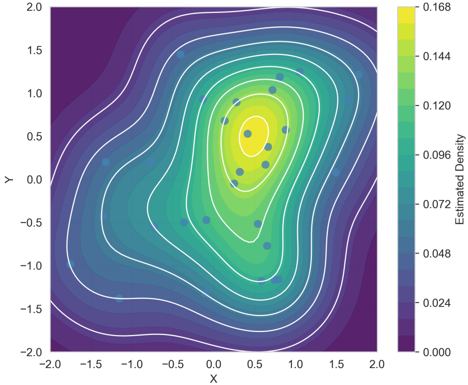
Non-parametric Estimation: Beyond Model Assumptions





Kernel Density Estimation in Two Dimensions





Kernel Density Estimation: Mathematical Formulation

Given bivariate observations $\{(x_i, y_i)\}_{i=1}^n$, the kernel density estimate is:

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_x h_y} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right)$$

With a Gaussian kernel:

$$K(u, v) = \frac{1}{2\pi} \exp\left(-\frac{u^2 + v^2}{2}\right)$$

This yields:

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi h_x h_y} \exp\left(-\frac{1}{2}\left[\left(\frac{x - x_i}{h_x}\right)^2 + \left(\frac{y - y_i}{h_y}\right)^2\right]\right)$$

The Crucial Role of Bandwidth Selection

Bandwidth parameters (h_x, h_y) control the degree of smoothing:

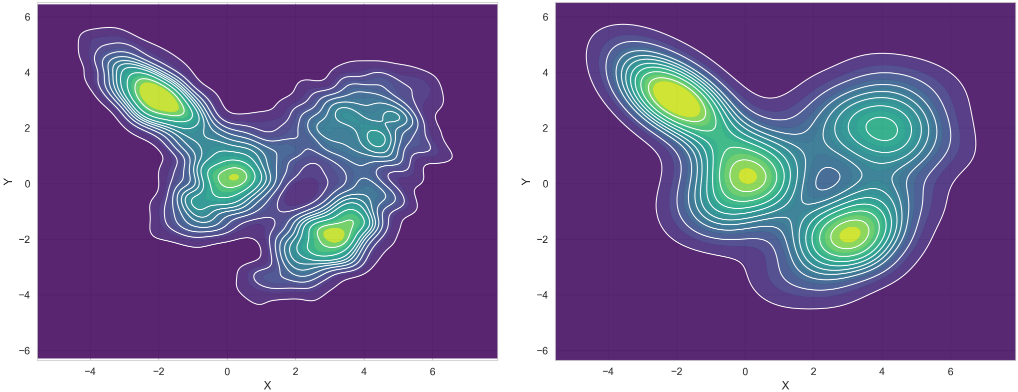
- Too small: Spiky estimate that overemphasizes random fluctuations
- Too large: Oversmoothed estimate that obscures important features

Silverman's rule of thumb for two dimensions:

$$h_x = 1.06 \cdot \sigma_x \cdot n^{-1/6}$$

$$h_y = 1.06 \cdot \sigma_y \cdot n^{-1/6}$$

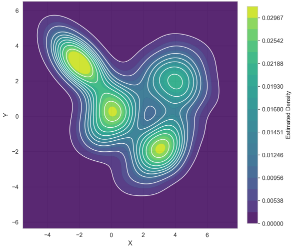
The Effect of Bandwidth Selection

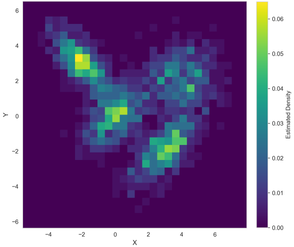


Advantages and Extensions of KDE

KDE offers several advantages over histograms:

- **Smoothness**
- **Efficiency**
- **Adaptability**



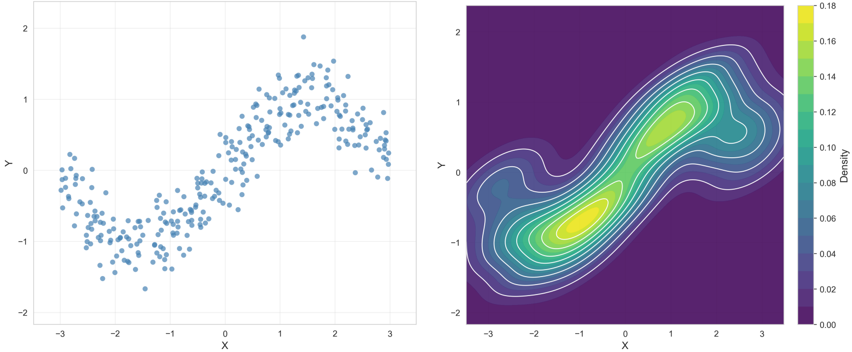


Mutual Information: Beyond Simple Associations

We can now use our distribution estimation tools to develop more comprehensive association measures.

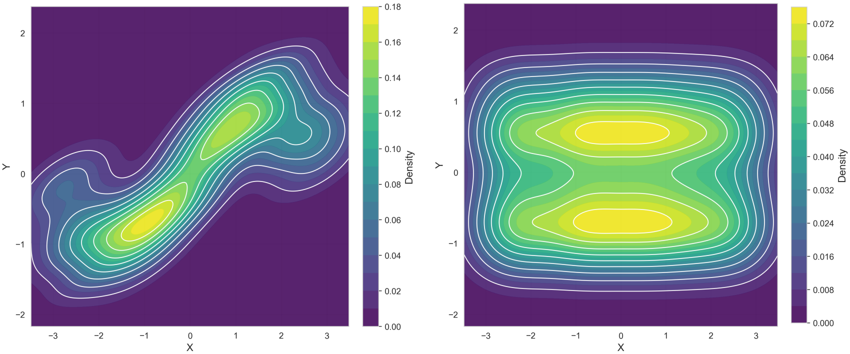
At its core, association is the opposite of independence:

Independence: $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$



Mutual Information: fundamental comparison

The main idea is comparing the joint distribution with the product of the marginals (independence case)



Mutual Information: Beyond Simple Associations

A natural approach to measure association:

1. Estimate true joint distribution $\hat{f}_{X,Y}(x, y)$
2. Compute marginals $\hat{f}_X(x)$ and $\hat{f}_Y(y)$
3. Create reference distribution $\hat{f}_X(x) \cdot \hat{f}_Y(y)$
4. Measure discrepancy between these distributions

From Kullback-Leibler Divergence to Mutual Information

To quantify the difference between distributions, we use the Kullback-Leibler (KL) divergence:

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Applying this to our joint and product distributions:

$$D_{KL}(f_{X,Y}||f_X \cdot f_Y) = \iint f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy$$

From Kullback-Leibler Divergence to Mutual Information

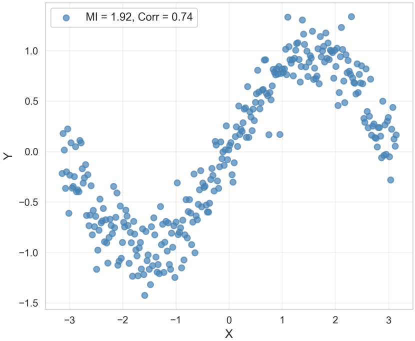
This KL divergence between the joint distribution and the product of marginals is precisely the definition of mutual information:

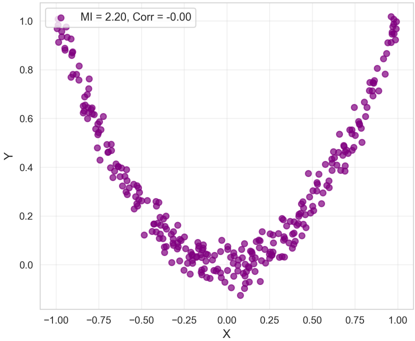
$$I(X;Y) = \iint f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy$$

The Power of Mutual Information

Mutual information offers several advantages:

- Captures any form of dependency, not just linear or monotonic
- Valuable in biological data where relationships often follow complex patterns.
- Invariant under invertible transformations of variables





Question 1: Marginal Distributions

Which of the following correctly defines the marginal distribution of X from a joint distribution $f_{X,Y}(x, y)$?

1. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$
2. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$
3. $f_X(x) = f_{X,Y}(x, y)/f_Y(y)$
4. $f_X(x) = \max_y f_{X,Y}(x, y)$

Question 2: Independence in Gene Expression

A researcher measures expression levels of two genes across hundreds of cells and finds that the joint probability density function can be expressed as $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$. What does this imply about these genes?

1. They are co-regulated by the same transcription factor
2. They are statistically independent of each other
3. They must have equal mean expression levels
4. The genes are located on the same chromosome

Question 3: Conditional Distributions

In a bivariate normal distribution with correlation coefficient ρ, what happens to the conditional distribution of Y given X=x as $|\rho|$ increases?

1. The variance of the conditional distribution increases
2. The mean of the conditional distribution becomes more dependent on the value of x
3. The conditional distribution approaches a uniform distribution
4. The conditional distribution becomes independent of the value of x

Question 4: Bivariate Normal Properties

For a bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and correlation $\rho = 0.5$, what is the probability density at the point (1, 1) relative to the density at the origin (0, 0)?

1. 0.223 (about 22.3% of the density at the origin)
2. 0.368 (about 36.8% of the density at the origin)
3. 0.472 (about 47.2% of the density at the origin)
4. 0.607 (about 60.7% of the density at the origin)

Question 5: Understanding KDE Visualization

Based on the KDE contour plot shown, what can we conclude about the underlying bivariate distribution?

1. It shows independent variables with no correlation
2. It exhibits a single mode with roughly circular level sets
3. It shows a strongly bimodal distribution with two distinct clusters
4. It displays a complex distribution with multiple local maxima and a curved ridge structure