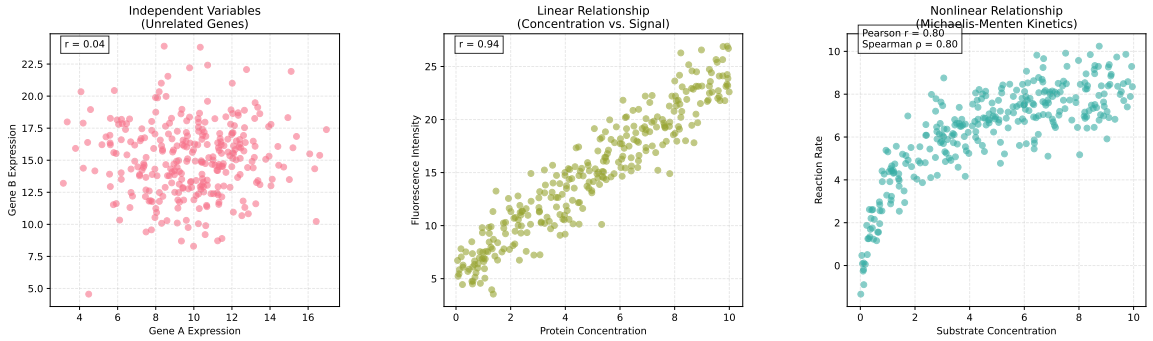
Lecture 3 - **Joint and Bivariate Distributions**

BIOENG-210 Course Notes
Prof. Gioele La Manno

March 2024

1 Joint Probability Distributions

In biological systems, variables rarely are independent from each other. Gene expression levels influence each other, protein concentrations depend on multiple factors, and cellular behaviors are determined by numerous interacting components. To understand these complex relationships, we need to move beyond single random variables to study how multiple random variables interact and relate to each other.



(a) Independent variables (b) Linearly related variables (c) Nonlinearly related variables

Figure 1: **Different types of relationships between variables observed in data**

In biological data analysis, we quickly realize that studying variables in isolation is insufficient. Biological processes are inherently multivariate, with components interacting in complex ways. This setting require us to introduce the concept of joint probability distribution as the appropriate mathematical object that allow us to model and analyze multiple variables simultaneously. We will begin by focusing on the simplest case: distributions involving just two variables, known as *bivariate distributions*.

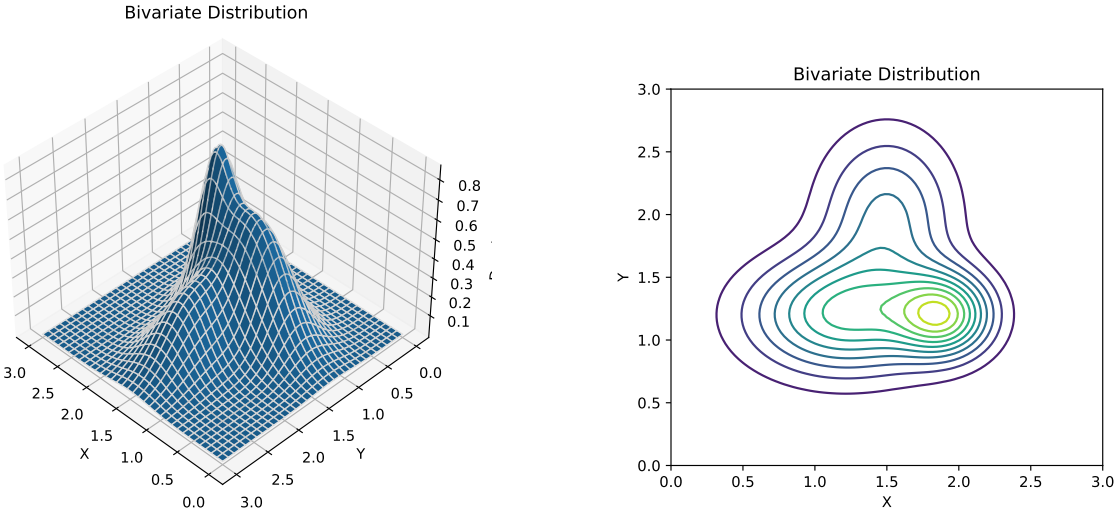
An important reminder is that often in biological systems, the correlation of several variable despite correlated are not linked by cause effect relationships but rather might be linked by common causes or even completelly causually independent and only linked spuriouslly by some confounder.

1.1 From Single Variables to Multiple Dimensions

While univariate distributions help us understand individual characteristics, joint distributions allow us to capture relationships between variables. This shift from one dimension to multiple dimensions brings both new challenges and new insights. Let us see how this transition unfolds in practice.

Let's first obtain a visual intuition of the joint distribution, if to show a pdf plot of a scalar we needed a 2d plot, for a bivariate distribution we need a 3d plot. Here it is a nontrivial bivariate

distribution. One can also represent the joint distribution using contour plots or density heat maps, which are great because a sheet of paper is 2d and we cannot really do 3d.



(a) Complex bivariate normal distribution

(b) Contour plot

Figure 2: **An example of Bivariate distribution**

Different types of plots can effectively represent the joint distribution of two variables.

The 3D plot provides a direct representation of the joint distribution, showing how probability density varies across the entire space of possible values. While comprehensive, these plots can be challenging to interpret on a 2D page.

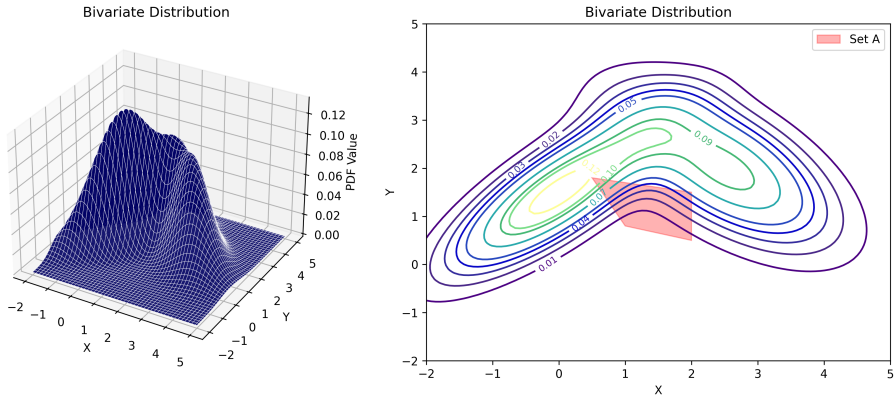
Contour plots offer a more accessible alternative by projecting the distribution onto a 2D plane.

After this visual intuition, let's define the joint probability distribution. Remember that the definition of a univariate pdf considered an interval, for a bivariate pdf we consider a region in the plane.

Definition 1.1 (Joint Probability Distribution)**.** For continuous random variables X and Y, their joint probability density function $f_{X,Y}(x, y)$ satisfies:

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

for any measurable set A in the plane.



(a) Set A over contour plot of bivariate distribution

Figure 3: **Definition of joint probability distribution**

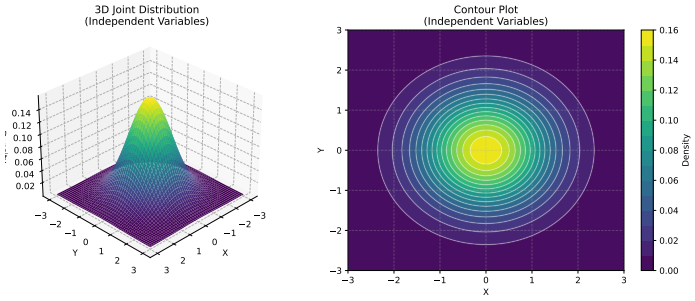
The joint distribution must satisfy the key properties:

$$f_{X,Y}(x, y) \geq 0 \text{ for all } x, y$$

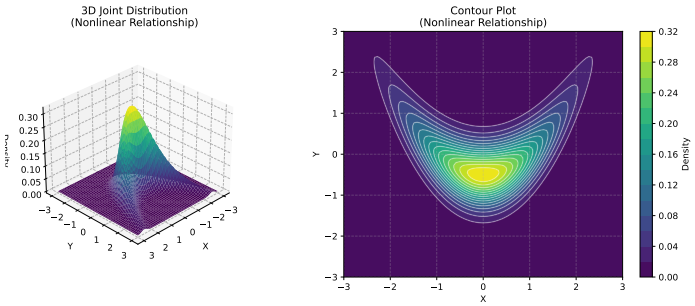
$$\iint_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$$

These properties ensure that our joint distribution is a valid probability distribution, but they tell us nothing about the relationship between the variables. The structure of this relationship is encoded in the shape of the joint distribution.

Let's compare two examples:



(a) Independent variables



(b) Nonlinearly related variables

Figure 4: **Different joint distribution structures**

These plots show in white level sets of the distribution - curves along which the probability density remains constant. Mathematically, a level set at value c is defined as:

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

Level sets are powerful tools for understanding distribution structure because they reveal: (a) The shape and orientation of the distribution, (b) Regions of high and low probability, (c) The presence of correlations or dependencies, (d) The spread and concentration of probability mass

1.2 Marginal Distributions

The term "marginal" has a historical origin that provides intuitive insight into its meaning. In early statistical practice, analysts would arrange bivariate frequency data in contingency tables, with totals for each variable calculated in the margins of the table. These "margin sums" became known as marginal distributions because they appeared literally at the margins of the table.

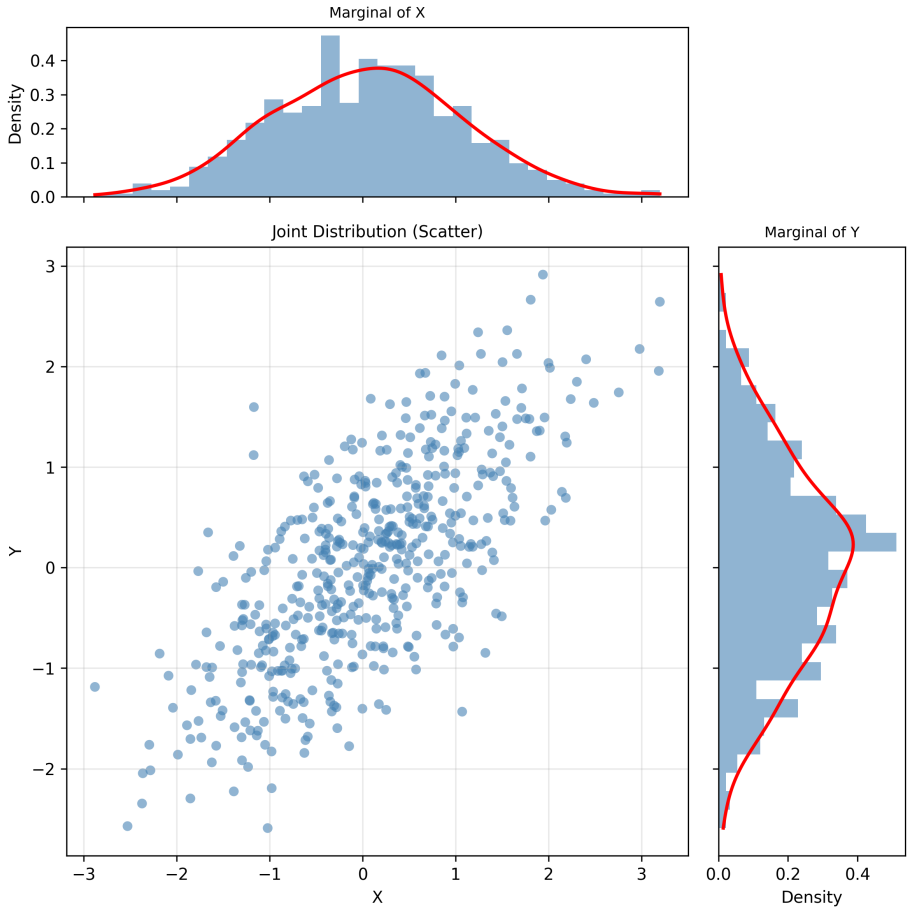


Figure 5: **Visualization of how marginal distributions relate to joint distributions**

Marginalization has an intuitive interpretation. Its relation to data is more immediate to understand than that of a joint distribution. If we have a dataset $\{(x_i, y_i)\}_{i=1}^n$, the marginal distribution of X is simply the distribution of X alone, ignoring the values of Y, and just looking at the histogram of X.

For example, if we have joint measurements of gene expression for two genes but are only interested in the distribution of one gene's expression, we would be examining its marginal distribution.

When working with empirical data, we can understand marginals by considering a scatter plot of bivariate data. If we were to simply "drop" one of the axes and project all points onto the remaining axis, we would get a univariate view corresponding to the marginal distribution of that variable. This is visually equivalent to looking at just one variable's histogram while ignoring the other. Note what is happening to the other variable: for each bin of the histogram we are summing up all the occurrences regardless of the other variable's values - this is the marginalization process.

This process leads us to formalize the concept of marginal distributions. Often, we need to recover information about an individual variable from a joint distribution without making assumptions about the value of the others.

Definition 1.2 (Marginal Distribution)**.** For joint density $f_{X,Y}(x, y)$, the marginal density of X is:

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

and similarly for Y:

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

When working with probability density functions, the integration operation corresponds to "summing up" the probability density over all possible values of the variable we're eliminating. What was a simple operation on data becomes more complex and not necessarily tractable analytically when working with probability density functions. Mathematically, for any set A where X might take values:

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

This expression means we're considering all possible combinations of (X, Y) where X falls within set A, regardless of what value Y takes. Informally, we can say that we have "integrated out" or "marginalizing out" the variable Y from the joint distribution to obtain the marginal distribution of X, which is a function of X alone.

1.3 Conditional Probability and Dependencies

One of the most powerful aspects of joint distributions is their ability to reveal dependencies between variables. All the information about how two variables interact is contained in their joint distribution, allowing us to do the following: Let's consider bivariate case, we can fix one variable and examine how the distribution of the other variable changes based on this fixed value. This approach reveals how knowledge of one variable informs our understanding of the other, leading us to the crucial concept of conditional distributions.

Definition 1.3 (Conditional Distribution)**.** The conditional density of Y given X = x is:

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

when $f_X(x) > 0$.

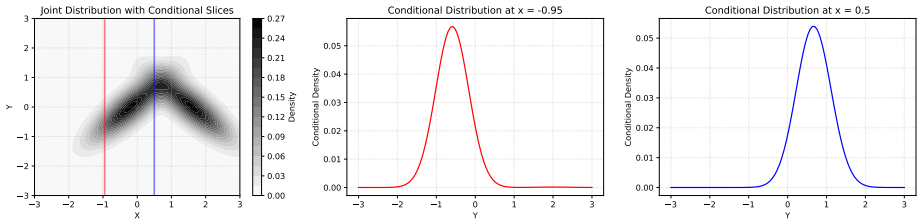
This definition gives rise to one of the most fundamental equations in probability theory:

Theorem 1.4 (Chain Rule of Probability)**.** For any two random variables X and Y:

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

The chain rule is particularly important in biological applications. Consider, for example, how gene expression levels change with cell size. The joint distribution of expression and size can be decomposed into:

- The marginal distribution of cell size ($f_X(x)$)
- The conditional distribution of expression given size ($f_{Y|X}(y|x)$)



(a) Joint and conditional distributions example

Figure 6: **Joint and conditional distributions example**

1.4 Independence: when marginals are sufficient

A special case of particular importance occurs when two variables are independent - when knowing the value of one variable tells us nothing about the other.

Definition 1.5 (Statistical Independence)**.** Random variables X and Y are independent if and only if:

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

for all x and y.

Independence is a strong assumption, and in biological systems, it is often more of an approximation than a reality. However, understanding independence helps us:

- Simplify complex probability calculations
- Identify truly interacting components in biological systems
- Develop null models against which to test for relationships

1.5 A first example: The Bivariate Normal Distribution

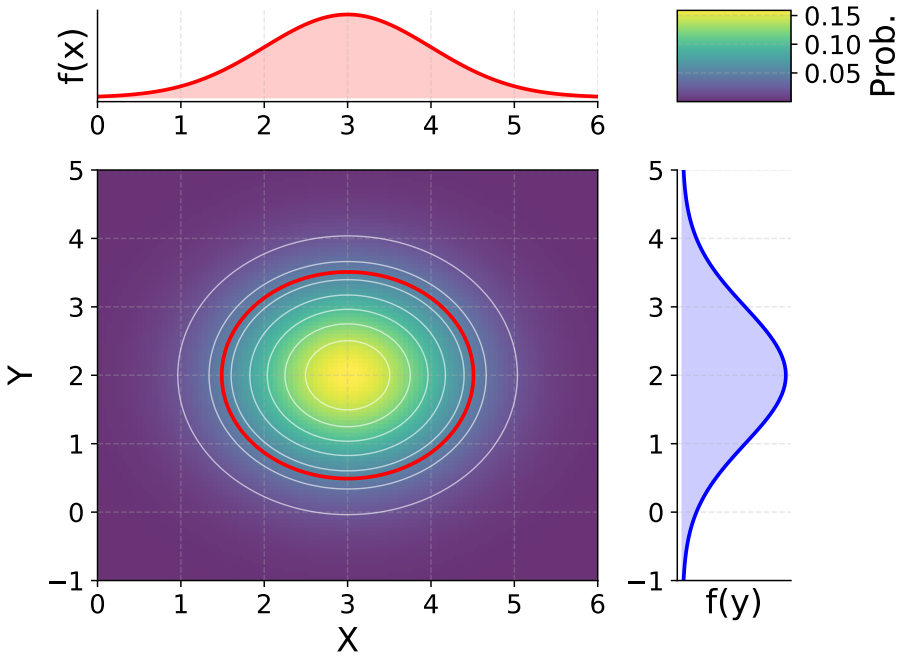
The bivariate normal distribution is a natural extension of the univariate normal distribution to two dimensions. Let's build our understanding step by step, starting with the simplest case.

Imagine we have two independent normally distributed random variables, $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. When variables are independent, their joint distribution is simply the product of their individual distributions:

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

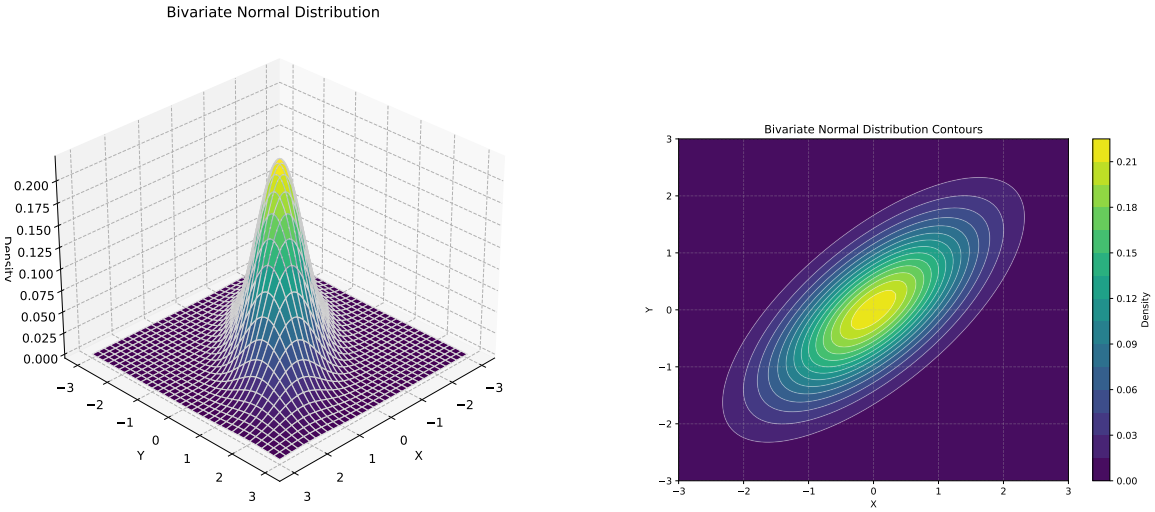
In this case, the level sets of the joint distribution form concentric circles (when $\sigma_1 = \sigma_2$) or ellipses with axes aligned with the coordinate axes (when $\sigma_1 \neq \sigma_2$). This reflects the fact that the value of one variable gives us no information about the other.

However, in many biological systems, variables tend to be correlated. The bivariate normal distribution generalizes to include this correlation through a parameter ρ, which measures the strength and direction of the linear relationship between the variables.



(a) Joint distribution of independent normal variables

Figure 7: **Joint distribution of independent normal variables with their marginals**



(a) 3D plot of bivariate normal distribution

(b) Contour plot of bivariate normal distribution

Figure 8: **Bivariate normal distribution**

The general form of the bivariate normal density function is:

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

This formula might seem complex, but it captures several key properties:

- The means μ_1, μ_2 determine the center of the distribution
- The variances σ_1^2, σ_2^2 control the spread along each axis
- The correlation coefficient ρ governs the shape of the distribution
- The normalizing constant ensures the distribution integrates to 1
- When $\rho = 0$, the distribution reduces to the product of two independent normal distributions

Note how level sets are now elliptical, reflecting the correlation between variables. The orientation, eccentricity, and size of these ellipses provide valuable insights into the relationship between X and Y.

1.6 Geometric Interpretation: Level Sets as Ellipsoids

We already introduced the concept of level sets as curves along which the probability density remains constant. An easy way to understand the pdf formula is to consider the level sets of the bivariate normal distribution, which form ellipses in the plane. This geometric interpretation will serve also later as we go in many more dimensions, serving as a foundation for understanding relationships between variables.

Let's recall the canonical form of an ellipse centered at (c_x, c_y) with semi-major axis a, semi-minor axis b, and rotated by angle α:

$$\frac{((x-c_x)\cos\alpha + (y-c_y)\sin\alpha)^2}{a^2} + \frac{((y-c_y)\cos\alpha - (x-c_x)\sin\alpha)^2}{b^2} = 1$$

Now, the exponent in our bivariate normal density function contains a quadratic form:

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

Setting this equal to a constant $c(1-\rho^2)$ defines the level sets of the distribution:

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

By comparing these forms, we can directly identify the properties of the elliptical level sets:

- **Center**: The ellipse is centered at $(c_x, c_y) = (\mu_1, \mu_2)$, the mean vector of the distribution
- **Rotation**: The correlation coefficient ρ determines the rotation angle α:

$$\tan(2\alpha) = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

When $\sigma_1 = \sigma_2$, the angle simplifies to $\alpha = \frac{\pi}{4}\text{sign}(\rho)$

- **Semi-axes lengths**: The general case is a bit complicated in this notation. Let's just consider now the case when $\sigma_1 = \sigma_2 = \sigma$ (equal variances), the semi-axes simplify to:

$$a = \sqrt{c} \cdot \sigma \cdot \frac{1}{\sqrt{1 - |\rho|}}$$

$$b = \sqrt{c} \cdot \sigma \cdot \frac{1}{\sqrt{1 + |\rho|}}$$

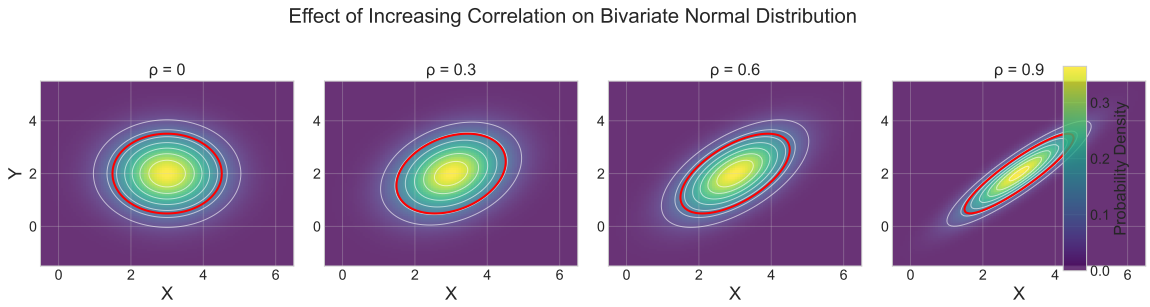


Figure 9: **Elliptical level sets of the bivariate normal distribution with different correlation coefficients**

The correlation coefficient ρ has a clear geometric interpretation:

- $\rho = 0$ means the variables are uncorrelated, and the ellipse axes align with the coordinate axes with lengths determined solely by σ_1 and σ_2
- As $|\rho|$ increases, the ellipse becomes more elongated in the direction of correlation:
  - The major axis increases as $\frac{1}{\sqrt{1-|\rho|}}$
  - The minor axis decreases as $\frac{1}{\sqrt{1+|\rho|}}$
- When $\rho = \pm 1$, the ellipse degenerates into a line (perfect correlation)
- The sign of ρ determines the orientation:
  - $\rho > 0$: The ellipse tilts upward (positive correlation)
  - $\rho < 0$: The ellipse tilts downward (negative correlation)

The constant c acts as a scaling factor that determines which particular level set we're examining:

- $c = 1$ corresponds to the 39% confidence region
- $c = 2$ corresponds to the 63% confidence region
- $c = 3$ corresponds to the 78% confidence region
- $c = 5.991$ corresponds to the 95% confidence region

This geometric interpretation shows how the parameters of the bivariate normal directly control the shape, size, and orientation of these elliptical level sets, providing an intuitive understanding of the distribution's structure.

1.7 A special case: Conditioning and Marginalization in the Bivariate Normal

The bivariate normal distribution provides a concrete example of how conditional and marginal distributions work in practice.

For the bivariate normal distribution:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right)$$

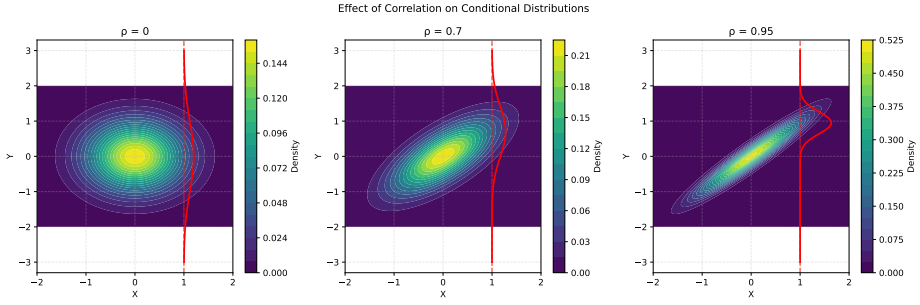
We can compute the conditional distribution of Y given X = x:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_y-\rho\sigma_y(x-\mu_x)/\sigma_x)^2}{\sigma_y^2(1-\rho^2)}\right]\right)$$

This is again a normal distribution with parameters:

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

Note how $\mu_{Y|X=x}$, the mean of Y, depends linearly on x, with a slope determined by the correlation coefficient and the ratio of standard deviations.



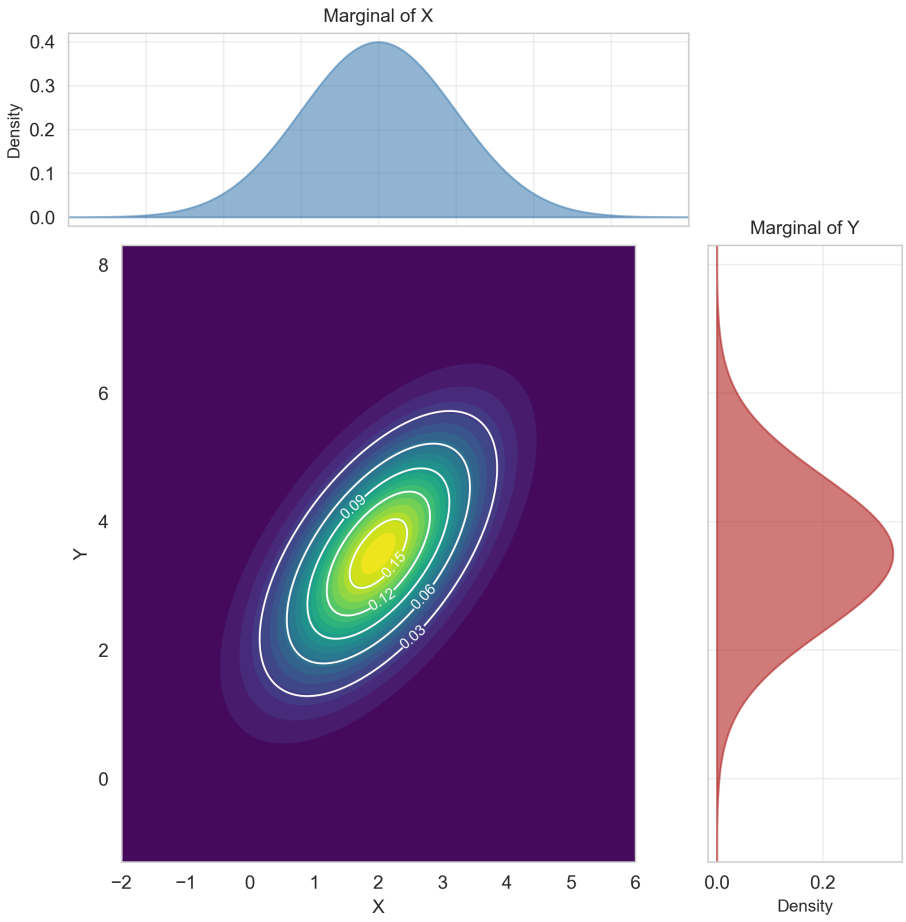
(a) Conditional distribution of bivariate normal

Figure 10: **Effect of correlation on conditional distribution**

Similarly, we can compute the marginal distribution of X:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

This is a normal distribution with mean μ_1 and variance σ_1^2. The same holds for the marginal distribution of Y, which is normal with mean μ_2 and variance σ_2^2. Remarkably, while the joint distribution depends on the correlation ρ, the marginal distributions do not - they retain the same form regardless of how strongly the variables are correlated.



(a) Marginal distribution of X and Y for bivariate normal

Figure 11: **Bivariate normal distribution: marginal distributions**

2 Bivariate estimates

Having established the theoretical framework for bivariate distributions, we now turn to the practical aspects of visualizing and estimating these relationships from data. In biological applications, we often encounter complex patterns that simple correlation measures fail to capture adequately. or instance, gene regulatory networks may exhibit threshold effects, saturations, or other nonlinear dependencies that traditional measures like Pearson's correlation coefficient cannot properly characterize.

Visualization serves as our first tool in understanding bivariate relationships. Through scatter plots, contour plots, and heatmaps, we can gain intuitive insights into the structure of dependencies between variables. However, visualization alone is insufficient for many applications, including statistical tests and simulations.

As we explore these methods, we will build toward introducing mutual information—a powerful, distribution-free measure of association that captures both linear and nonlinear dependencies. Unlike correlation coefficients, which detect specific types of relationships, mutual information quantifies the general statistical dependence between variables, making it particularly valuable when relationships rarely follow simple patterns.

2.1 Histogram 2D

The simplest approach to visualizing and estimating a bivariate distribution is through a 2D histogram. Just as a standard histogram divides the range of a single variable into bins, a bivariate histogram partitions the plane into rectangular bins and counts the number of observations falling into each bin.

Formally, given data points $\{(x_i, y_i)\}_{i=1}^n$, a 2D histogram divides the x-range into bins of width h_x and the y-range into bins of width h_y. The count in each bin is then converted to a probability estimate by dividing by the total number of observations and the area of the bin:

$$\hat{f}(x, y) = \frac{n_{ij}}{n \cdot h_x \cdot h_y}$$

where n_{ij} is the number of observations in the bin containing (x, y).

While conceptually straightforward, bivariate histograms face several challenges:

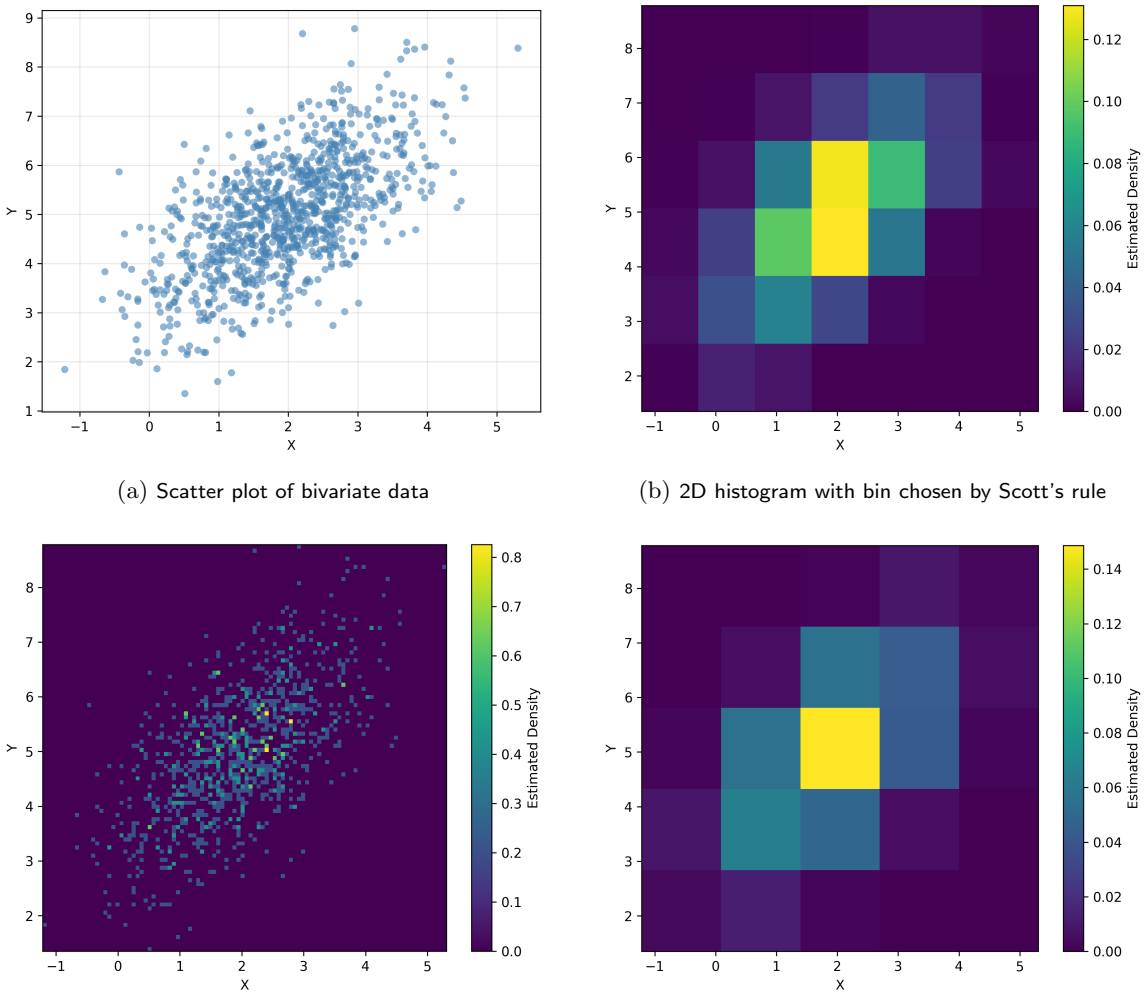
- **Bin width selection:** Selecting appropriate bin widths is crucial. Too narrow, and the histogram becomes noisy with many empty bins; too wide, and important features are smoothed away. Various rules exist, such as extensions of Sturges' rule or Scott's rule to two dimensions, but the optimal choice often depends on the specific structure of the data.
- **Limited resolution:** The histogram's resolution is constrained by bin size, creating a blocky representation that may miss fine details of the underlying distribution.
- **Zero-probability regions:** Areas with no observations are assigned zero probability, even when nearby bins contain data points. This fails to capture the continuity typically expected in real-world distributions.
- **Non-smooth appearance:** The resulting estimate is discontinuous at bin boundaries, creating artificial edges that may not reflect the true distribution.

Some of these problems can be mitigated by choosing an appropriate bin width, such as Scott's rule or the Freedman-Diaconis rule, which aim to balance the trade-off between bias and variance in the histogram estimate.

For bivariate data, extensions of univariate bin width selection rules are typically used:

- **Scott's rule:** $h_x = 3.5 \cdot \sigma_x \cdot n^{-1/6}$ and $h_y = 3.5 \cdot \sigma_y \cdot n^{-1/6}$
- **Freedman-Diaconis rule:** $h_x = 2 \cdot \text{IQR}_x \cdot n^{-1/6}$ and $h_y = 2 \cdot \text{IQR}_y \cdot n^{-1/6}$

where σ is the standard deviation, IQR is the interquartile range, and n is the sample size.



(a) Scatter plot of bivariate data

(b) 2D histogram with bin chosen by Scott's rule

(c) 2D histogram with too small bins

(d) 2D histogram with too large bins

Figure 12: **Effect of bin width on 2D histogram**

Despite these limitations, 2D histograms provide a useful first approximation of the joint distribution when we have sufficient data. They are particularly valuable for initial data exploration and identifying regions of high density or obvious patterns in the relationship between variables.

However, for more refined analysis or when working with limited data, we need to consider more sophisticated estimation techniques that produce smoother, more continuous representations of the underlying distribution.

2.2 Parametric estimation of a joint distribution

When we have a priori knowledge about the form of the joint distribution, parametric estimation provides a powerful approach for characterizing relationships between variables. This involves fitting a specific distribution with parameters that can accommodate the observed pattern in our data.

As discussed in the previous section, we can solve this through maximum likelihood estimation. However, we must ensure the chosen parametric form is capable of capturing the relationship structure evident in our data. The selected model must be flexible enough to represent different correlation patterns - otherwise, the estimation will effectively reduce to independently estimating each marginal distribution, missing the joint structure entirely.

2.2.1 Fitting a Bivariate Normal Distribution

To estimate the parameters of a bivariate normal distribution from observed data, we need to determine five parameters: the means (μ_1, μ_2), standard deviations (σ_1, σ_2), and correlation parameter (ρ). One approach we discussed in the previous lecture would be to use numerical optimization routines to maximize the likelihood function. However, for the bivariate normal distribution, it can be derived a particularly closed-form maximum likelihood estimators through calculus (which for the mean and variance is the same as in univariate normal distribution) and for the correlation coefficient is non-trivial but can be derived.

Theorem 2.1 (Maximum Likelihood Estimators for Bivariate Normal Distribution)**.** Given data points $\{(x_i, y_i)\}_{i=1}^n$, the maximum likelihood estimators for the bivariate normal distribution are:

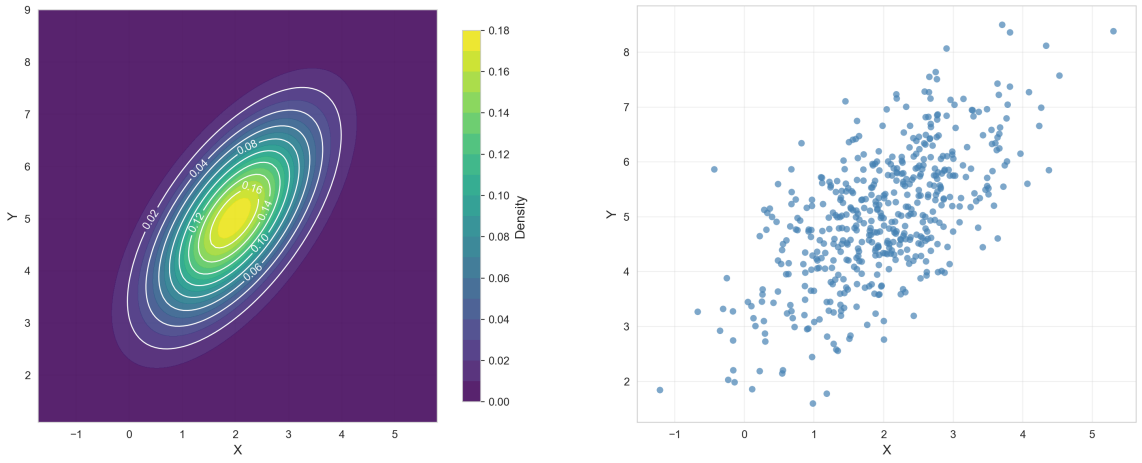
$$\hat{\mu}_1 = \bar{x} \qquad \hat{\mu}_2 = \bar{y}$$

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

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

Thus the maximum likelihood estimators for the bivariate normal distribution are exactly the sample statistics we're familiar with: sample means, sample standard deviations, and the sample correlation coefficient.

In other words, these common summary statistics are sufficient statistics for the bivariate normal distribution. When we calculate these sample statistics, we have extracted all relevant information from our data for estimating the parameters of a bivariate normal distribution.



(a) Bivariate normal distribution fit

(b) Scatter plot of bivariate data

Figure 13: **Fitting a bivariate normal distribution**

2.3 Non parametric estimation of a joint distribution

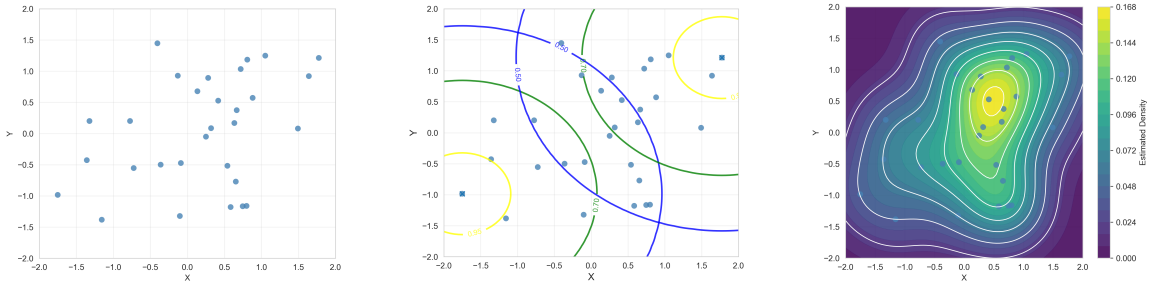
When the underlying structure of bivariate data doesn't conform to a specific parametric form, nonparametric methods offer a flexible alternative. Among these, kernel density estimation (KDE) is very often used as it can create smooth estimates of joint distributions without imposing rigid assumptions on their shape.

2.3.1 Kernel Density Estimation in Two Dimensions

Kernel density estimation (KDE) extends the concept of histograms by placing a probability density function (a kernel) centered at each data point and then adding these functions together.

Given a sample of bivariate observations $\{(x_i, y_i)\}_{i=1}^n$, the kernel density estimate at any point (x, y) is given by:

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



(a) Individual data points (b) Individual kernels (c) Kernel density estimate

Figure 14: **Building a KDE by summing individual kernels**

where:

- $K(\cdot, \cdot)$ is the kernel function, typically a bivariate probability density
- h_x and h_y are bandwidth parameters that control the amount of smoothing in each dimension

Most commonly, a bivariate Gaussian kernel is used:

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

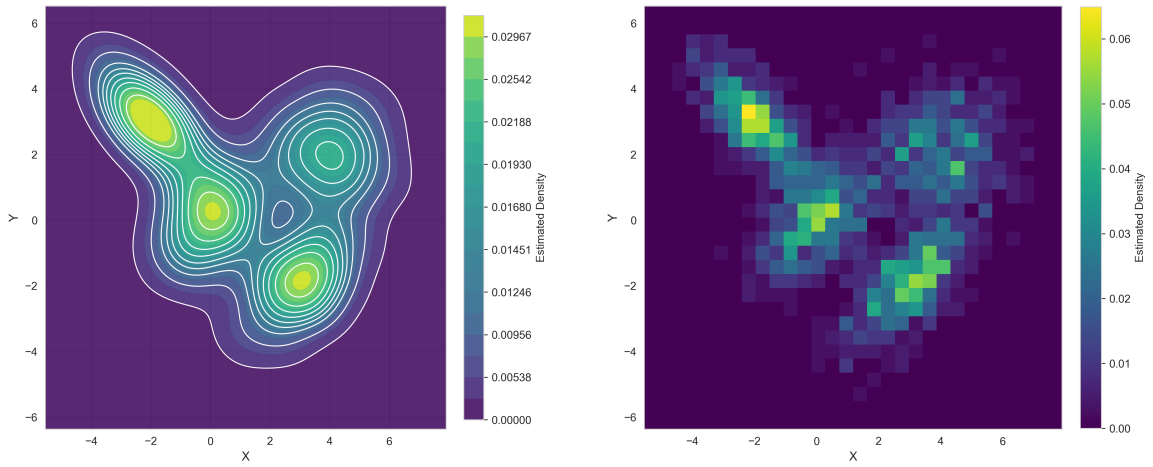
This gives us the final form of the Gaussian kernel density estimator:

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

The KDE approach offers several advantages over histograms:

- **Smoothness:** The resulting density estimate is continuous and differentiable
- **Efficiency:** With appropriate bandwidth selection, KDE can approximate the true density with fewer observations than required by histograms
- **Adaptability:** Different kernel functions can be chosen based on the data characteristics

However, it also has some limitations:



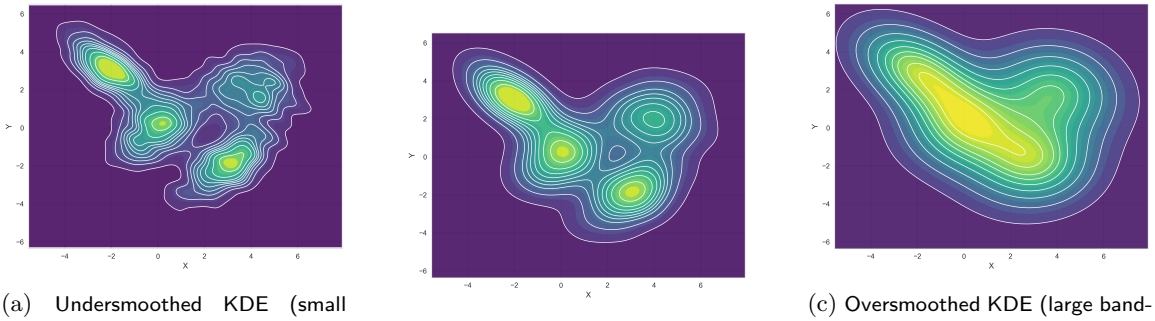
(a) Kernel density estimation example

(b) 2D histogram example

Figure 15: **Kernel density estimation vs. 2D histogram**

- **Boundary effects:** KDE extends probability density to all of $\mathbb{R}^2$, which can be problematic when variables have natural boundaries (e.g., concentrations cannot be negative). This leads to positive probability in impossible regions.
- **Computational complexity:** For large datasets, computing the sum over all observations for each evaluation point becomes expensive.
- **Domain limitations:** Standard KDE is computed within a box that contains the observed data points, potentially missing important characteristics that extend beyond this region.
- **Normalization:** The KDE estimate is not normalized, so one needs to renormalize numerically to obtain a proper density estimate.

The bandwidth parameters (h_x, h_y) control the degree of smoothing and significantly impact the resulting estimate. Too small a bandwidth creates a spiky estimate that overemphasizes random fluctuations in the data, while too large a bandwidth oversmooths and may obscure important features.



(a) Undersmoothed KDE (small bandwidth)

(b) Optimally smoothed KDE

(c) Oversmoothed KDE (large bandwidth)

Figure 16: **Effect of bandwidth on kernel density estimation**

Several methods exist for selecting appropriate bandwidths. A common rule of thumb is the Silverman's rule to two dimensions:

$$h_x = 1.06 \cdot \sigma_x \cdot n^{-1/6} \quad \text{and} \quad h_y = 1.06 \cdot \sigma_y \cdot n^{-1/6}$$

2.3.2 Extensions and Variations

Several extensions to basic KDE exist that can further improve performance in biological applications:

- **Adaptive bandwidth:** Instead of using the same bandwidth for all observations, adjust the bandwidth based on local data density.
- **Boundary correction:** Standard KDE can perform poorly near the boundaries of the support (e.g., when variables cannot be negative). Specialized techniques address this issue.
- **Variable transformation:** For highly skewed data, transforming variables before applying KDE and then back-transforming can improve estimates.

Kernel density estimation provides a flexible framework for understanding complex bivariate relationships in biological data. By creating smooth, continuous estimates of joint distributions without imposing restrictive parametric assumptions, KDE helps reveal subtle patterns that might otherwise remain hidden.

2.4 Mutual Information: Beyond Simple Associations

Now that we have introduced methods for estimating joint distributions, we can extend the discussion to more advanced measures of association between variables.

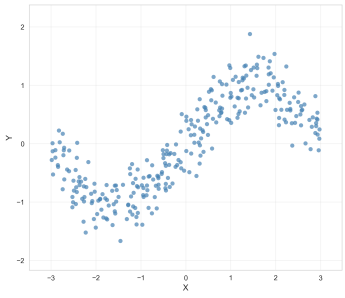
2.4.1 Independence vs. Association

To develop a more comprehensive measure that uses the notions we learnt, let us reflect on what "association" really means. At its core, association is the opposite of independence. When two variables are independent, knowing one tells us nothing about the other. Their joint distribution factors perfectly into the product of their marginals:

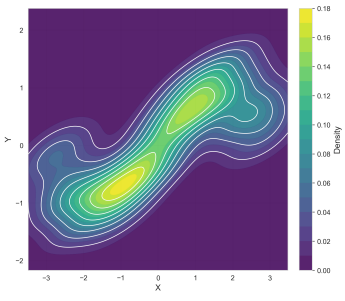
$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

Therefore, a natural way to measure association is to quantify how much a joint distribution differs from what we would expect under independence. Now that we have tools to estimate joint distributions (such as KDE) and marginal distributions, we can construct a meaningful comparison.

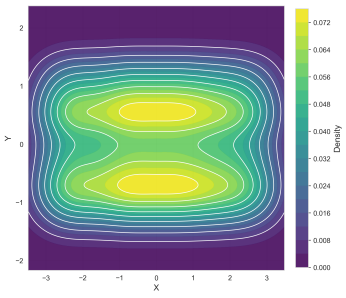
Given a dataset, we can: 1. Estimate the true joint distribution $\hat{f}_{X,Y}(x, y)$, for example using KDE 2. Compute the marginal distributions $\hat{f}_X(x)$ and $\hat{f}_Y(y)$, empirically summing the KDE 3. Construct a "reference" joint distribution assuming independence: $\hat{f}_X(x) \cdot \hat{f}_Y(y)$ 4. Measure the discrepancy between these two distributions



(a) Original data showing nonlinear relationship



(b) Estimated joint distribution



(c) Reference distribution under independence

Figure 17: **Comparing actual joint distribution with independence assumption**

The greater the discrepancy between these distributions, the stronger the association between the variables. But how do we quantify this difference in a meaningful way? Let's invoke some concepts from information theory.

2.4.2 Kullback-Leibler Divergence and Mutual Information

To quantify the difference between two probability distributions, the Kullback-Leibler (KL) divergence provides a principled measure. For two probability distributions $p(x)$ and $q(x)$, the KL divergence is defined as:

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

This measure quantifies the information lost when using $q(x)$ to approximate $p(x)$.

In our context, we want to measure how much information is lost when approximating the true joint distribution with the reference distribution under independence:

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

This expression is precisely the definition of mutual information between X and Y:

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

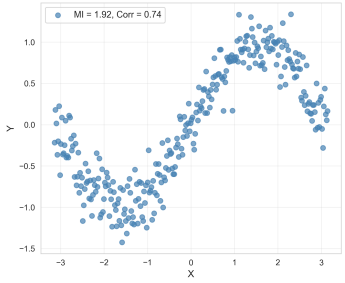
Through this derivation, we see that mutual information has an intuitive interpretation: it measures the "distance" between the actual joint distribution and what we would expect if the variables were independent. When the variables are indeed independent, this distance becomes zero. As the dependency strengthens, the mutual information increases.

Mutual information offers several advantages for analyzing biological data:

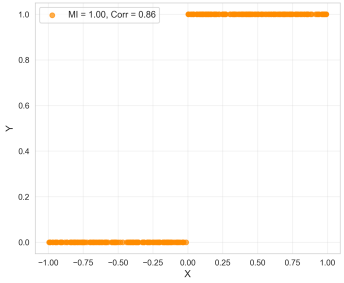
- It captures any form of dependency, not just linear or monotonic relationships
- It is invariant under invertible transformations of the variables
- It provides a principled measure with foundations in information theory

This approach is not without challenges particularly with limited sample sizes. As dimensionality increases the histogram sparsifies, accurate estimation becomes increasingly difficult, requiring larger samples or additional assumptions.

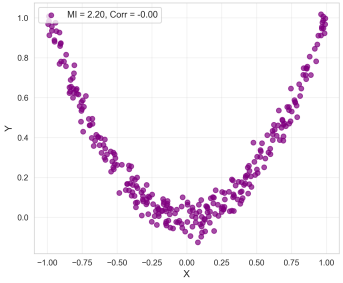
Despite these challenges, mutual information provides a powerful tool for uncovering complex relationships in biological systems, revealing dependencies that might remain hidden when using simpler association measures.



(a) Sine wave relationship: Low correlation, high MI



(b) Threshold relationship: Low correlation, high MI



(c) Complex nonlinear relationship: Low correlation, high MI

Figure 18: **Cases where mutual information detects relationships that correlation misses**