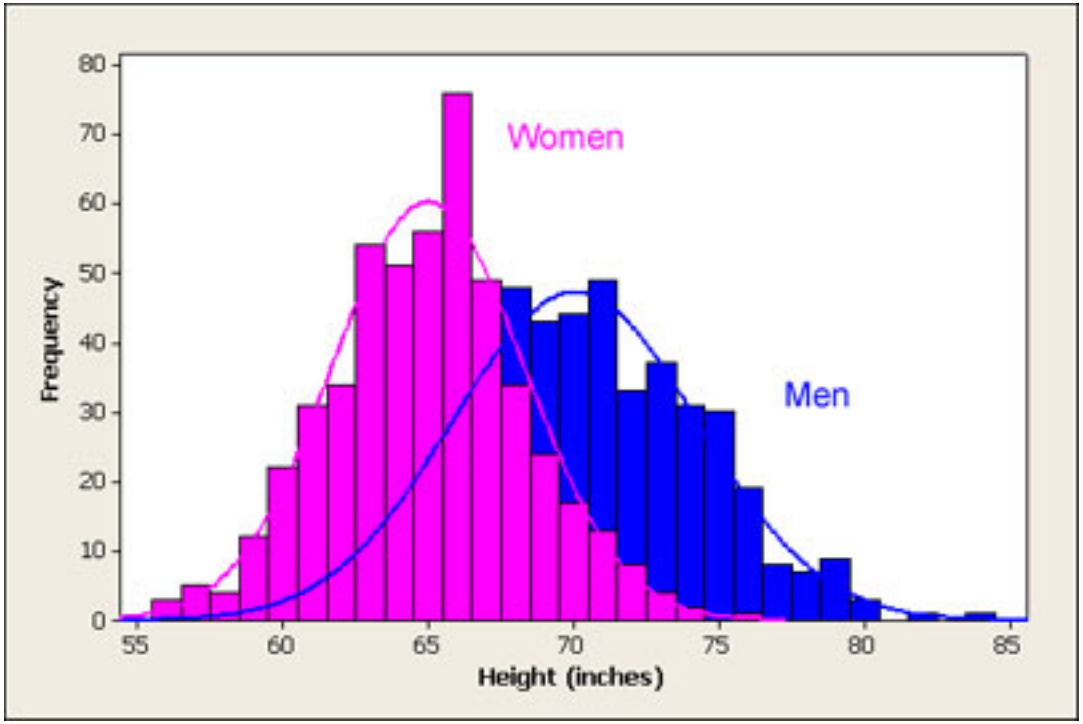
Density Estimation

ECE57000: Artificial Intelligence

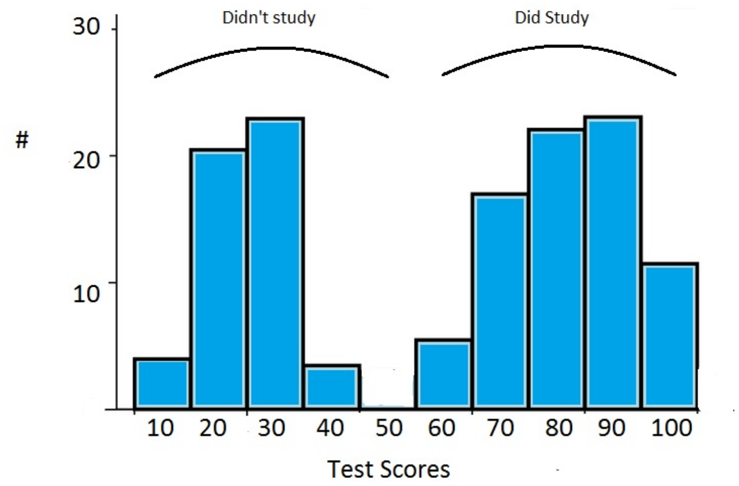
David I. Inouye

Density estimation finds a density (PDF/PMF) that represents the data (or empirical distribution) well

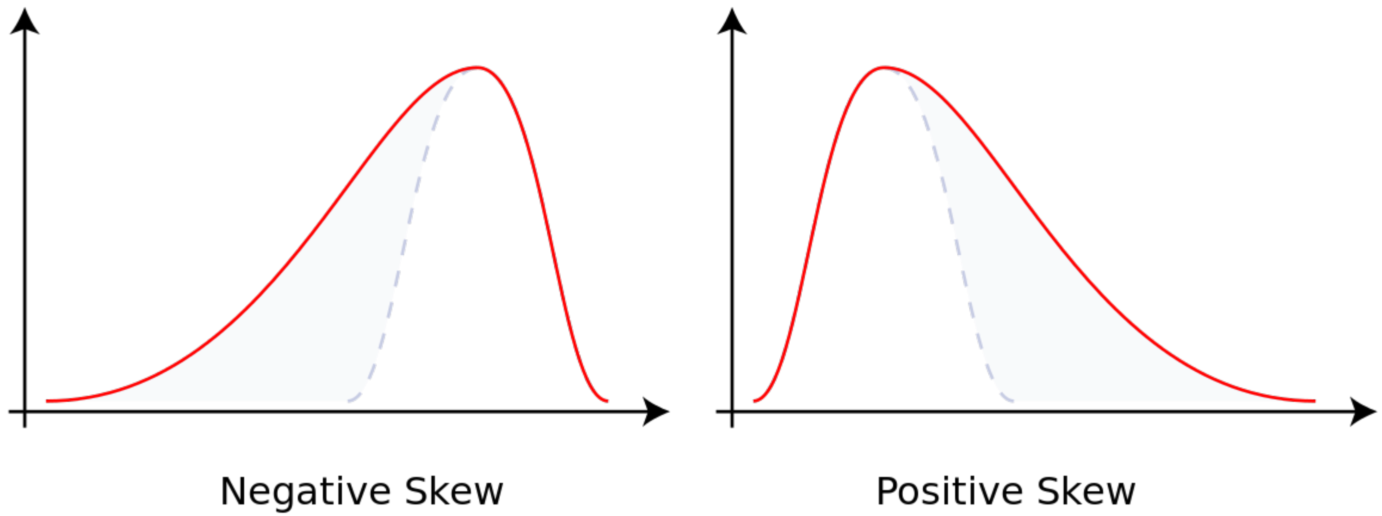


Motivation: Density estimation can be used to uncover underlying structure

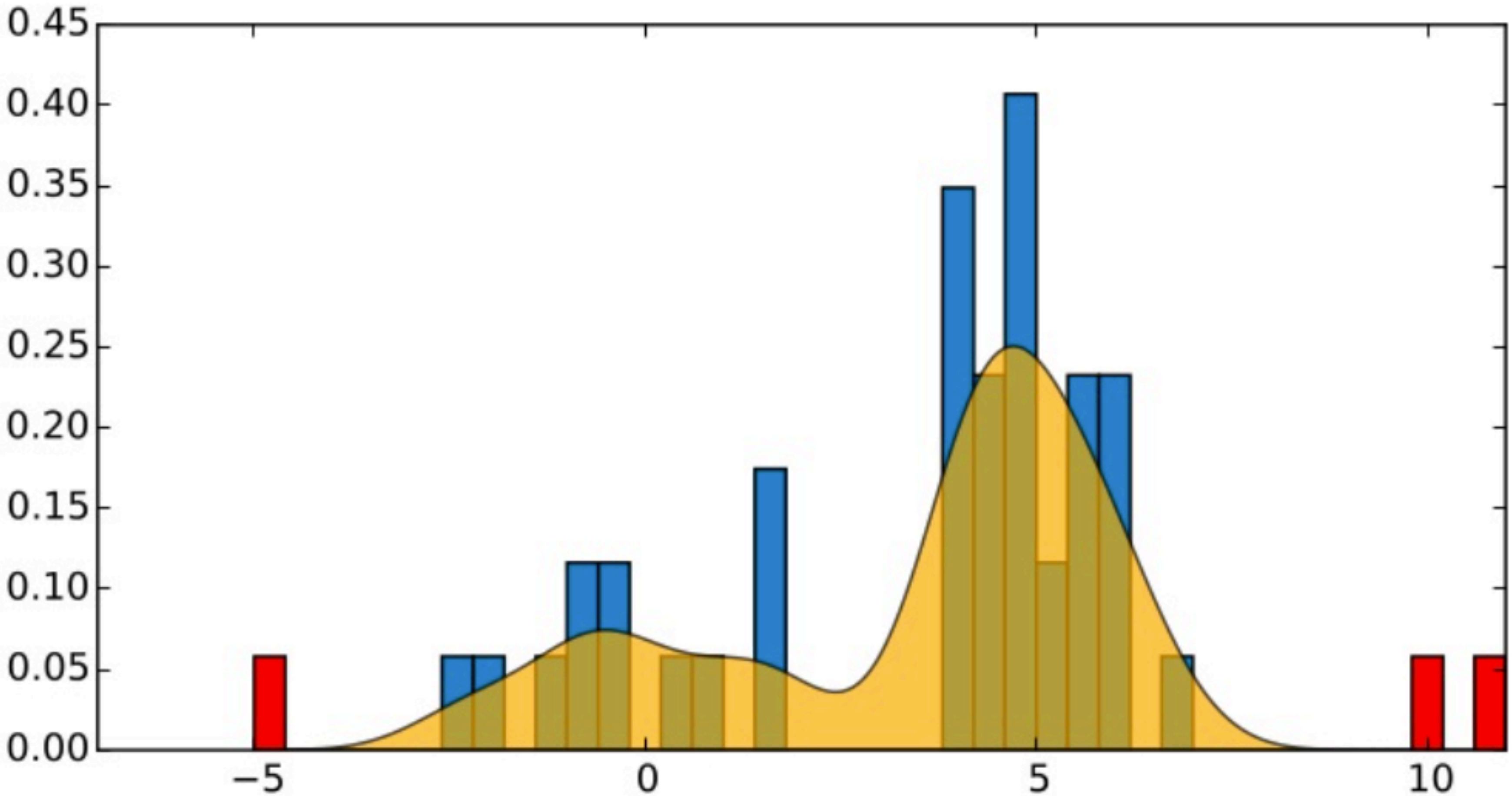
- Uncover multi-modal structure



- Uncover skewness



Motivation: Density estimation can be used for anomaly detection



https://www.slideshare.net/agramfort/anomalynovelty-detection-with-scikitlearn

Parametric density estimation assumes a density model class parameterized by θ

- Assumption: Bernoulli density
$$\theta = [p], \qquad p \in [0,1]$$
- Assumption: Exponential density
$$\theta = [\lambda], \qquad \lambda \in \mathbb{R}_{++}$$
- Assumption: Gaussian density
$$\theta = [\mu, \sigma^2], \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{++}$$
- Assumption: DNN-based model
$$\theta = [\text{“}all\ neural\ network\ parameters\text{”}]$$

How do we determine which model in the model class is the best?

- Classically, people have turned to information theoretic quantities
  - Entropy
  - Kullback Liebler (KL) Divergence
  - Maximum likelihood estimation (MLE)

- However, there other estimators particularly for **robust estimation**
  - Regularized estimation
  - Robust estimation

Informally, **entropy** measures the “amount of randomness/disorder” of a distribution

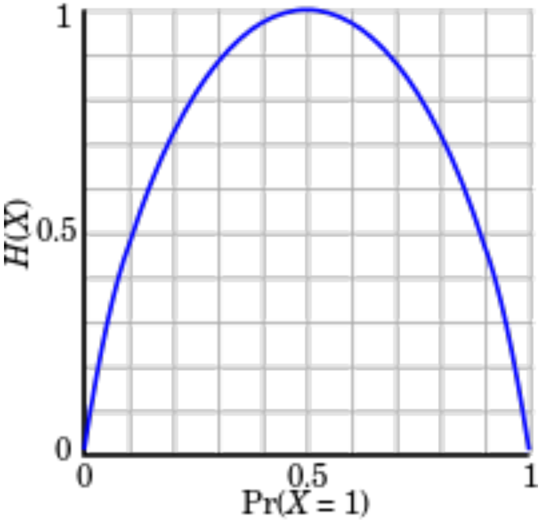
- Formally, **entropy** for discrete variables

$$H\big(P(x)\big) = \mathbb{E}[-\log P(x)] = \sum_x -P(x)\log P(x)$$

- Formally, **differential entropy** for continuous variables

$$H\big(p(x)\big) = \mathbb{E}[-\log p(x)]$$
$$= \int_x -p(x)\log p(x)\,dx$$

- Consider fair coin vs coin where both sides are heads



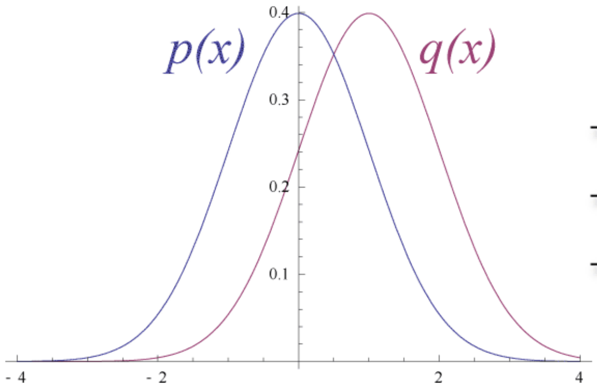
Informally, Kullback-Leibler Divergence (KL) measures the distance between distributions

- Formally, **KL divergence** for discrete variables

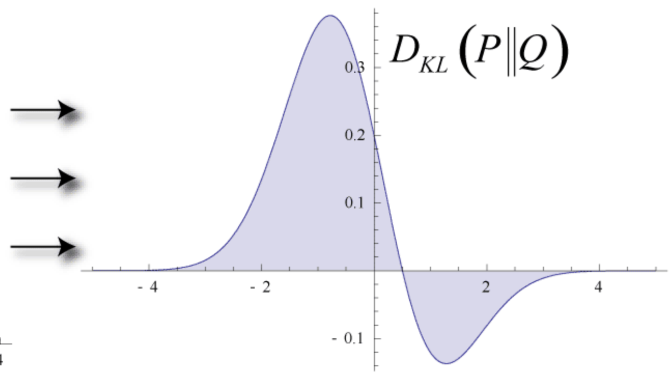
$$KL(P(x), Q(x)) = \mathbb{E}_{x\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

- Formally, **KL divergence** for continuous variables

$$KL(p(x), q(x)) = \mathbb{E}_{X\sim p}\left[\log\frac{p(x)}{q(x)}\right] = \int_x p(x)\log\frac{p(x)}{q(x)}dx$$



Original Gaussian PDF's



KL Area to be Integrated

Informally, Kullback-Leibler Divergence (KL) measures the distance between distributions

$$KL(p(x), q(x)) = \mathbb{E}_{X\sim p}\left[\log \frac{p(x)}{q(x)}\right] = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

- Not symmetric!

$$KL\big(p(x), q(x)\big) \neq KL\big(q(x), p(x)\big)$$

- Non-negative property

$$KL\big(p(x), q(x)\big) \geq 0$$

- Equal distribution property:

$$KL\big(p(x), q(x)\big) = 0 \Leftrightarrow p(x) = q(x)$$

One use of KL divergence is to estimate distribution parameters only from samples

- Let $p(x)$ denote the **real/true** distribution of the data
  - $p(x)$ is ***unknown***
  - We only have samples $\{x_i\}_{i=1}^{n}$ from $p(x)$
- Let $\hat{q}(x;\theta)$ denote an **<u>estimate</u>** of the true distribution
  - Parametrized by θ
- We want to find $\hat{q}(x;\theta)$ that is closest to $p(x)$

$$\theta^* = \arg\min_{\theta} \mathrm{KL}(p(x), \hat{q}(x;\theta))$$

One use of KL divergence is to estimate distribution parameters only from samples

- We want to find $\hat{q}(x;\theta)$ that is closest to $p(x)$

$$\theta^* = \arg\min_{\theta} \mathrm{KL}(\,p(x), \hat{q}(x;\theta))$$

- Wait, but we don't know $p(x)$, how do we do this?

- Two main ideas for simplification
  - Constants with respect to (w.r.t.) θ can be ignored
  - Full expectation replaced by empirical expectation

Derivation of minimum KL divergence with samples

<u>Maximum likelihood estimation (MLE)</u> is another way to estimate distribution parameters from samples

- **<u>Likelihood function</u>** how likely (or probable) a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ is under a distribution with parameters θ
$$\mathcal{L}(\theta; \mathcal{D}) = p(x_1, x_2, \dots, x_n; \theta)$$
- If we ***assume*** samples (or observations) of dataset are **<u>independent and identically distributed (iid)</u>**, then
$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{n} p(x_i; \theta)$$
- Often simplified to the **<u>log-likelihood function</u>**

$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log p(x_i; \theta)$$

Maximum likelihood (MLE) is another way to estimate distribution parameters from samples

- Optimize the following

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta; \mathcal{D})$$

- Equivalent to

$$\theta^* = \arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i; \theta)$$

- Wait, doesn't that look familiar?
- **MLE equivalent to minimum KL divergence!**

MLE is not the only way or necessarily the best distribution estimator

- Corrupt/noisy samples (related to **robustness**)
  - Cashiers using 1111 for birth year: 908 years old
  - One star ratings

- Finite (sometimes small) number of samples
  - One or two coin flips, Bernoulli
  - 1D with one sample, Gaussian
  - 2D with two samples, multivariate Gaussian

- Examples: Median or regularized MLE

Multivariate Gaussian

- Definition
- Properties and intuitions
- MLE estimator for multivariate Gaussian

The most ubiquitous multivariate distribution is the **multivariate Gaussian/normal distribution**

- Compare univariate to multivariate:
  - μ is mean and Σ is covariance

$$p(x) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

$$p(x_1, \dots, x_d) = \frac{1}{\left(\sqrt{2\pi}\right)^d \sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

- $\Theta = \Sigma^{-1}$ is called the **precision matrix** (or **inverse covariance**)
- Σ (and Θ) must be positive definite $\Sigma > 0$
- (Suppose $\Sigma = I$, suppose $\mu = 0$)

Multivariate Gaussian is independent "spherical" Gaussian that is rotated and scaled

$$\Sigma = U\Lambda U^T = \left(U\Lambda^{\frac{1}{2}}\right)\left(\Lambda^{\frac{1}{2}}U^T\right) = \left(U\Lambda^{\frac{1}{2}}\right)\left(U\Lambda^{\frac{1}{2}}\right)^T$$

$$x^T\left(U\Lambda^{-\frac{1}{2}}\right)\left(U\Lambda^{-\frac{1}{2}}\right)^T x = \left(\Lambda^{-\frac{1}{2}}Ux\right)^T\left(\Lambda^{-\frac{1}{2}}Ux\right) = z^T z$$

$$\Rightarrow x = U^T\Lambda^{\frac{1}{2}}z$$

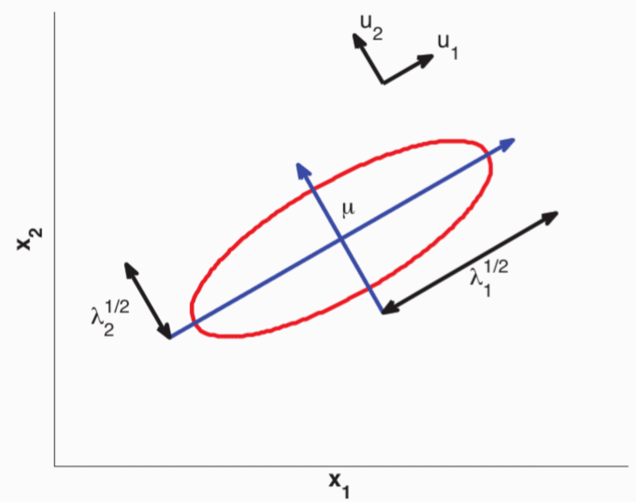


Figure 4.1 Visualization of a 2 dimensional Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely $\mathbf{u}_1$ and $\mathbf{u}_2$. Based on Figure 2.7 of (Bishop 2006a).

Machine Learning, Murphy, 2012.

Marginal and conditional distributions are Gaussian and can be computed in closed-form

- 2D case:

$$\boldsymbol{x} = [x_1, x_2] \sim \mathcal{N}\left(\mu = [\mu_1, \mu_2], \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}\right)$$

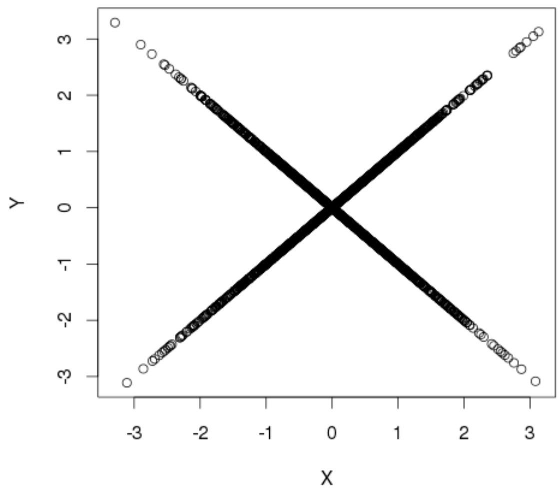
- Marginal distributions:

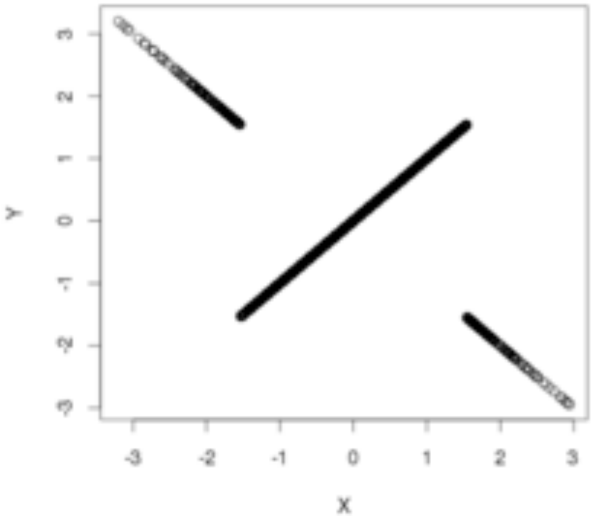
$$x_1 \sim \mathcal{N}(\mu = \mu_1, \sigma^2 = \sigma_1^2)$$
$$x_2 \sim \mathcal{N}(\mu = \mu_2, \sigma^2 = \sigma_2^2)$$

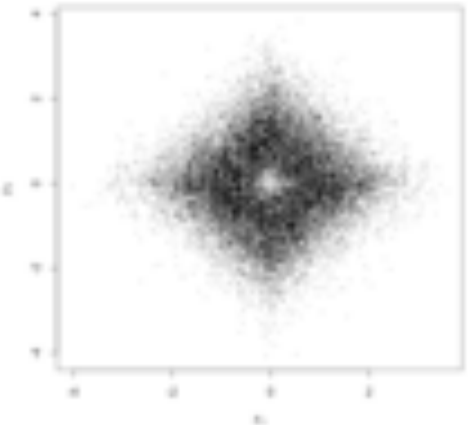
- Conditional distributions:

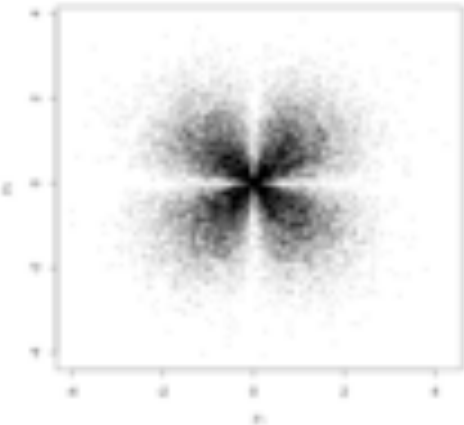
$$x_1 | x_2 = a$$
$$\sim \mathcal{N}\left(\mu = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(\mathrm{a} - \mu_2), \sigma^2 = \sigma_1^2 - \frac{\sigma_{21}^2}{\sigma_2^2}\right)$$

Gaussian marginals does <u>NOT</u> imply jointly multivariate Gaussian (converse <u>NOT</u> generally true)









<u>Affine transformations</u> of multivariate Gaussian vector are also multivariate Gaussian

- If $x \sim \mathcal{N}(\mu, \Sigma)$ and $y = Ax + b$, then
$$y \sim \mathcal{N}\left(A\mu + b, A\Sigma \mathrm{A}^{\mathrm{T}}\right).$$
- Special case: Marginal distribution when A is:
$$A_i = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases}$$
$$\text{then } y = x_k \sim p(x_k).$$
- Key point: Marginals, conditionals and affine functions known in **<u>closed-form</u>**.
- Consequence 1: Easy to manipulate.
- Consequence 2: Gaussians and linear ideas play nicely with each other.

MLE of multivariate Gaussian can be computed via empirical mean and covariance matrix

- Log-likelihood of multivariate Gaussian ($\mu = 0$)

$$-\frac{1}{2}\log|\Sigma| - \frac{1}{2n}\sum_{i=1}^{n} x_i^T \Sigma^{-1} x_i + const$$

- Three main identities:
  - $\frac{\partial \log|\mathrm{A}|}{\partial \mathrm{A}} = A^{-T}$
  - $\mathrm{Tr}(x^T A x) = \mathrm{Tr}(A x x^T)$
  - $\frac{\partial \mathrm{Tr}(AX)}{\partial \mathrm{X}} = A$
- Hint: Do derivative with respect to Σ^{-1}

Simplification and derivation of MLE for multivariate Gaussian

$$L(\Sigma; \mathcal{D}) = \frac{n}{2} \log|\Sigma^{-1}| - \frac{1}{2} \mathrm{Tr}\left(\Sigma^{-1} \left(\sum_i x_i x_i^T \right) \right)$$

$$\frac{\partial L}{\partial \Sigma^{-1}} = \frac{n}{2} \Sigma - \frac{1}{2} \sum_i x_i x_i^T$$

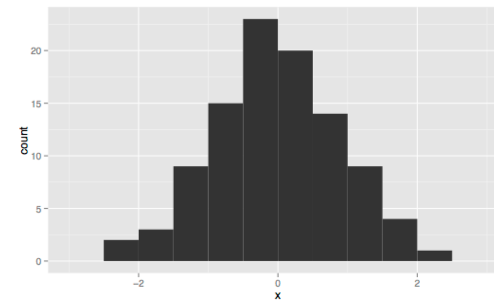
$$\Sigma = \frac{1}{n} \sum_i x_i x_i^T$$

Non-parametric density estimation

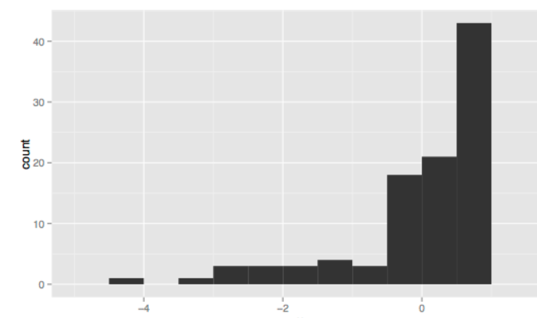
- Motivation
- Histograms
  - Choosing k
  - Choosing bin edges
- Kernel density
  - Choosing bandwidth
  - Curse of dimensionality again

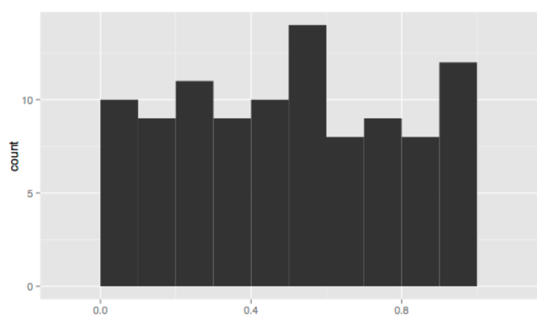
Why non-parametric density estimates?

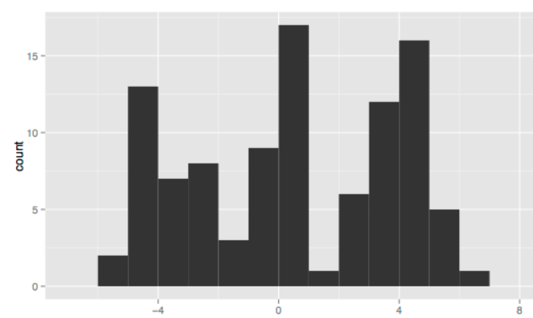
- Parametric densities are excellent if the assumptions are correct (e.g., Gaussian)

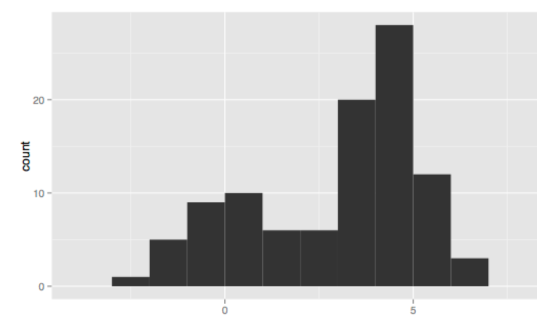


- However, the distributions may not align with the assumptions



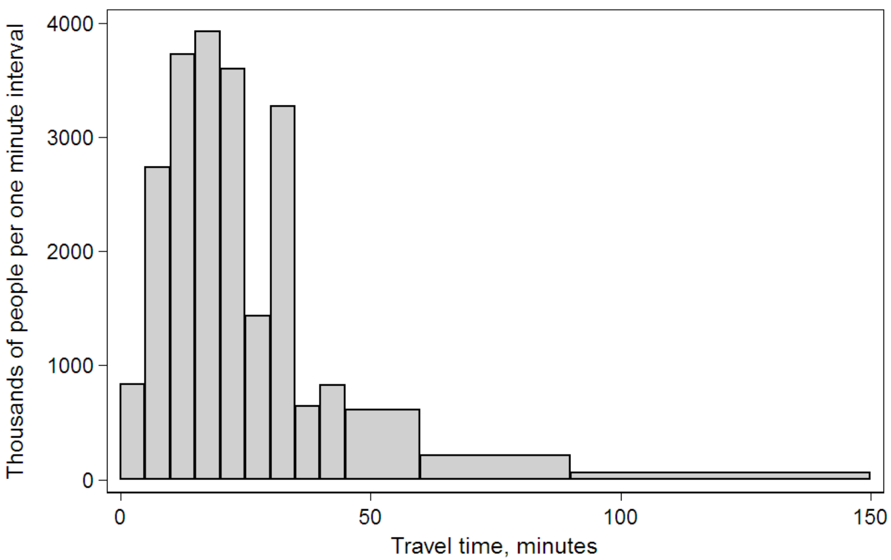


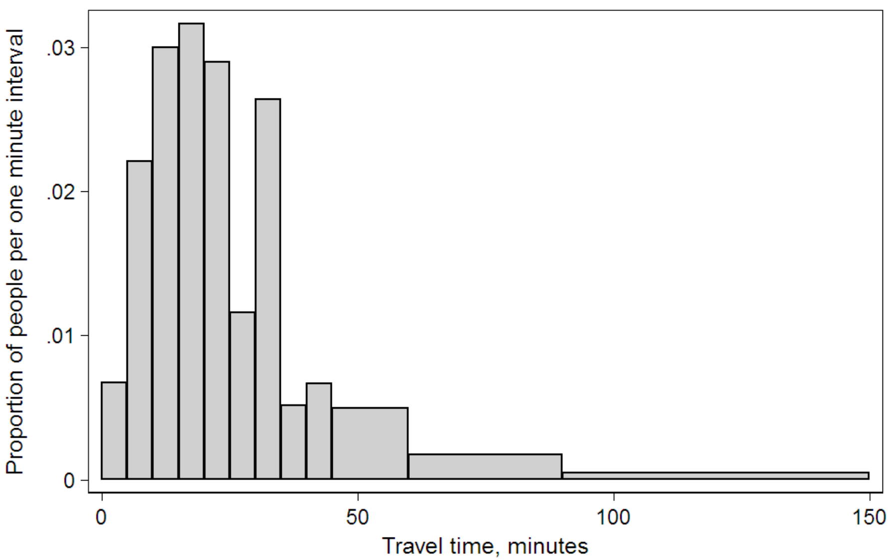




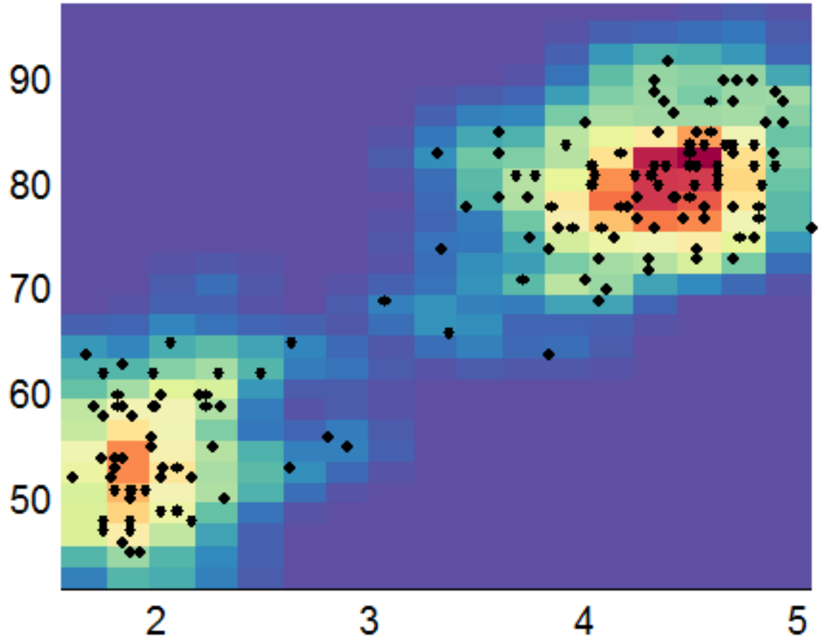
Histograms are the simplest density estimators

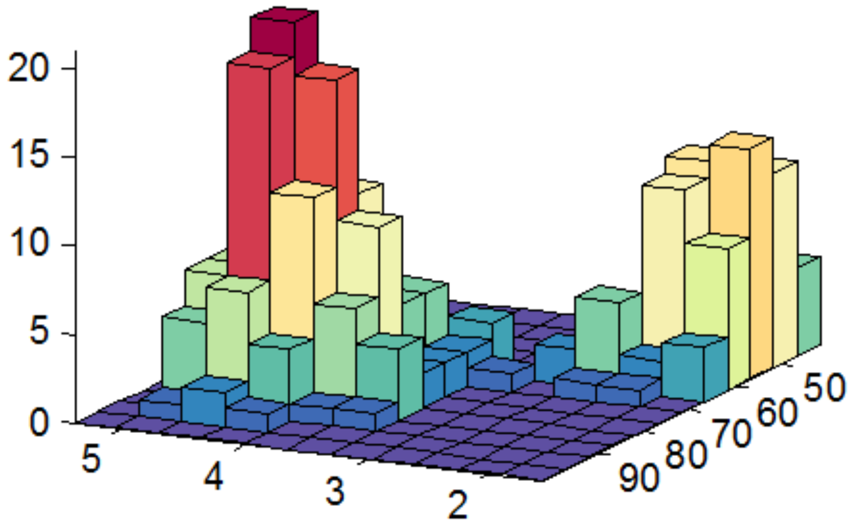
- Setup bin locations
- Count number of samples that fall in each bin
- Normalize to be a density





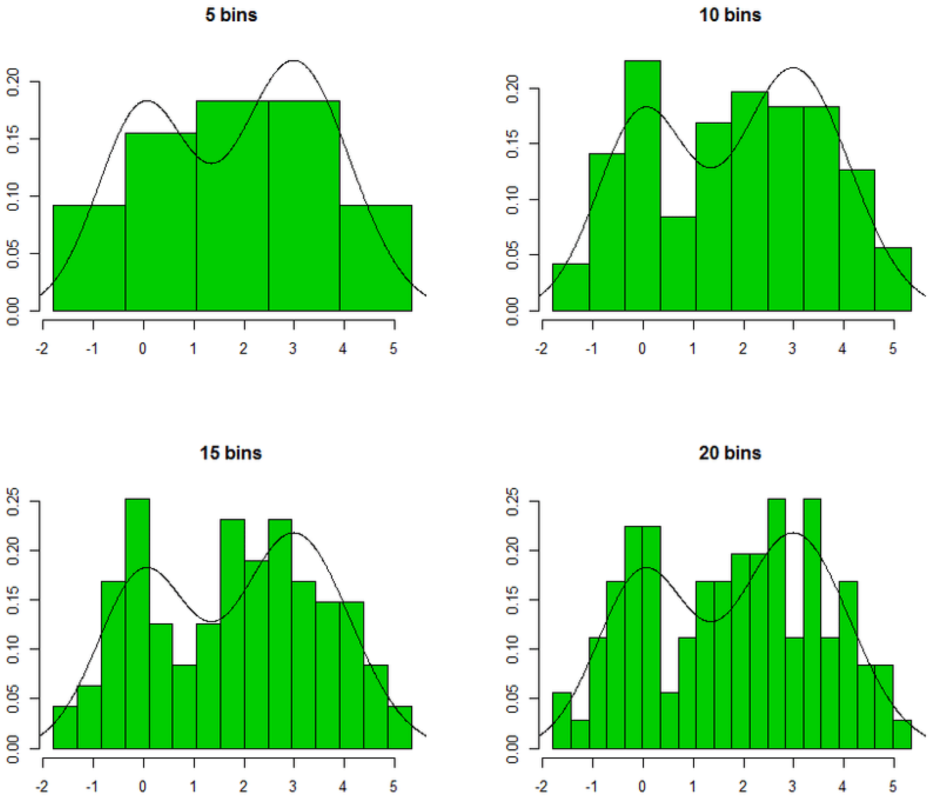
2D Histograms can be created





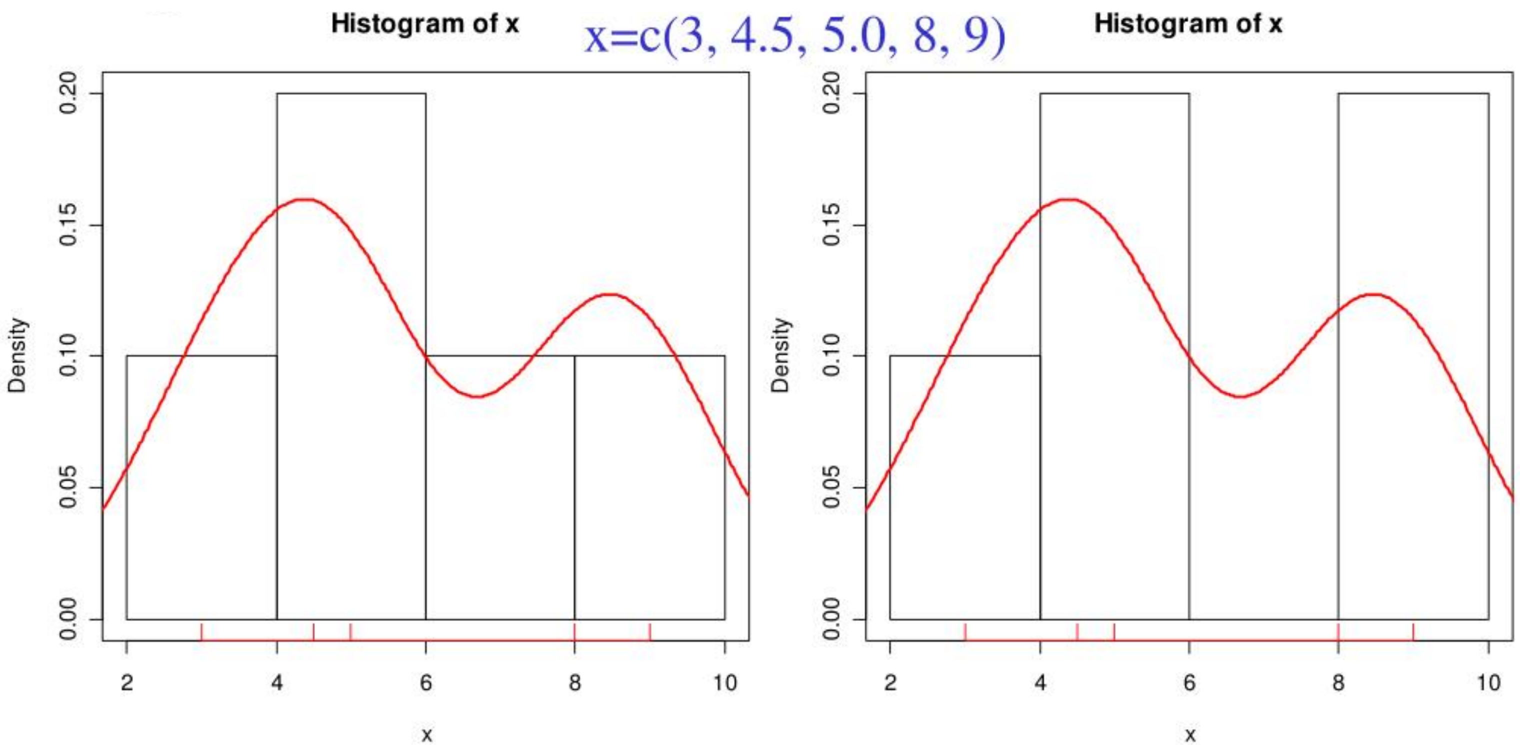
How to select the number of bins (usually denoted k)?

- Too few bins will underfit
- Too many bins will overfit
- ML approach: **<u>CV/Test</u>** log likelihood



Drawbacks: Histograms can depend on bin edges and are not smooth

x=c(3, 4.5, 5.0, 8, 9)



- hist(x,right=**T**,freq=F), R-default
- (a,b] right closed (left-open)

- hist(x,right=**F**,freq=F)
- [a,b) left closed (right-open)

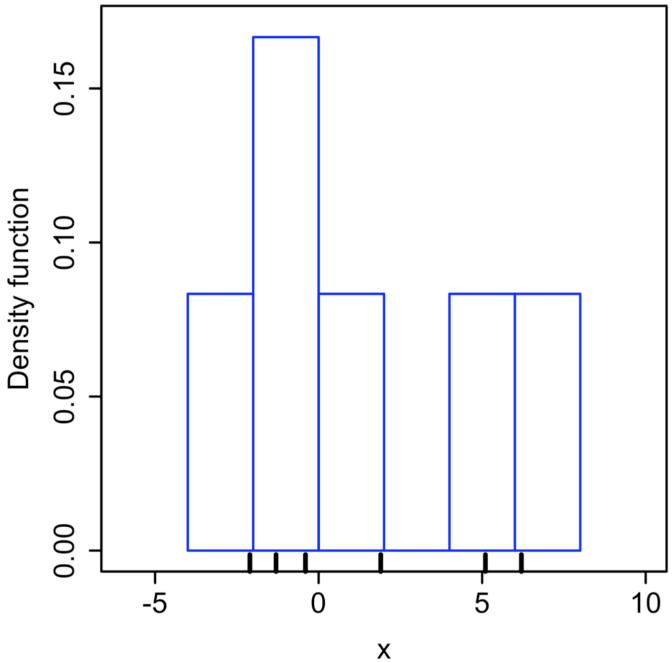
Area=1

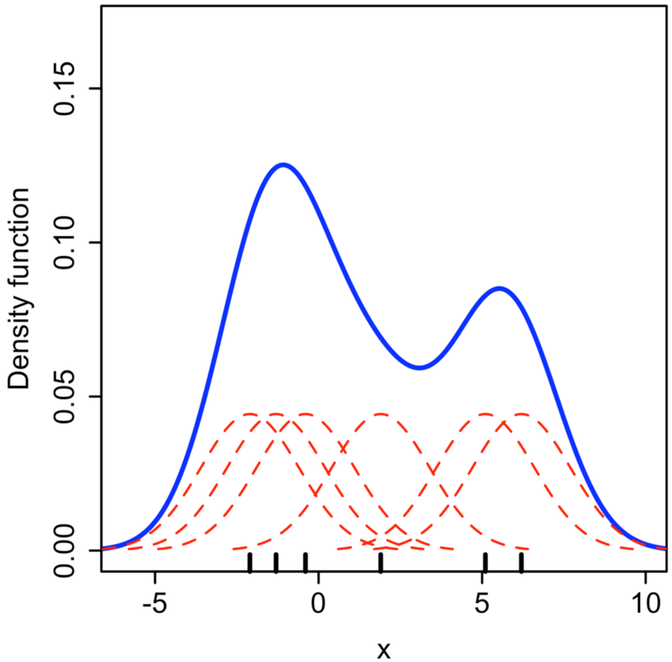
https://www.slideserve.com/geona/introduction-to-non-parametric-statistics-kernel-density-estimation

Kernel densities overcome this drawback by placing a Gaussian density at each point

- Kernel density has the following form:

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} p_{\text{base}}(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(x - x_i, \sigma)$$





Similar to number of bins, the key parameter for kernel densities is the "bandwidth" or σ parameter

- Bandwidth can be selected via CV/Test log likelihood (similar to number of histogram bins)