Gaussians

Andrew W. Moore

Professor

School of Computer Science

Carnegie Mellon University

www.cs.cmu.edu/~awm

awm@cs.cmu.edu

412-268-7599

Gaussians in Data Mining

- Why we should care
- The entropy of a PDF
- Univariate Gaussians
- Multivariate Gaussians
- Bayes Rule and Gaussians
- Maximum Likelihood and MAP using Gaussians

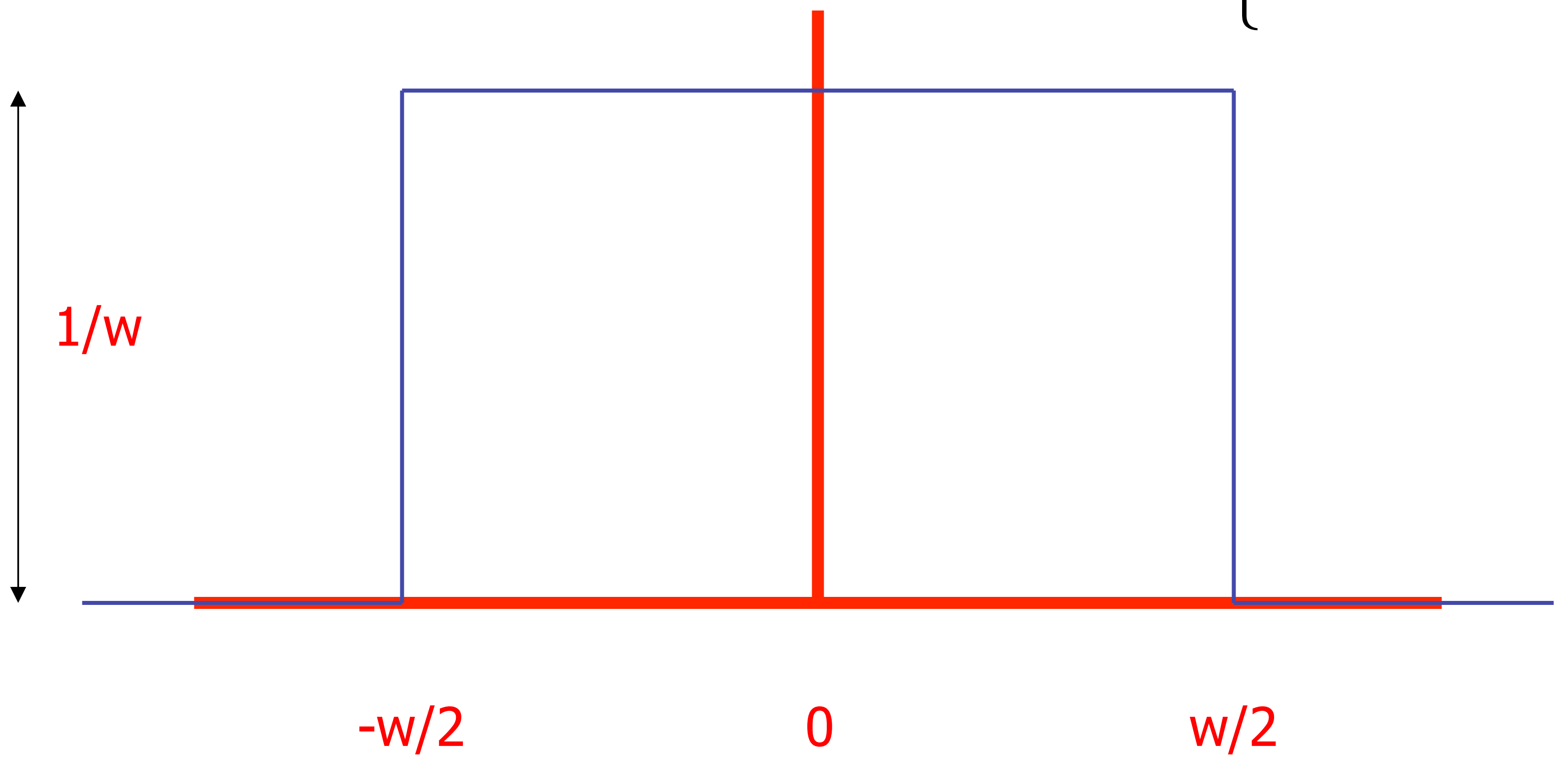
Why we should care

- Gaussians are as natural as Orange Juice and Sunshine
- We need them to understand Bayes Optimal Classifiers
- We need them to understand regression
- We need them to understand neural nets
- We need them to understand mixture models
- …

(You get the idea)

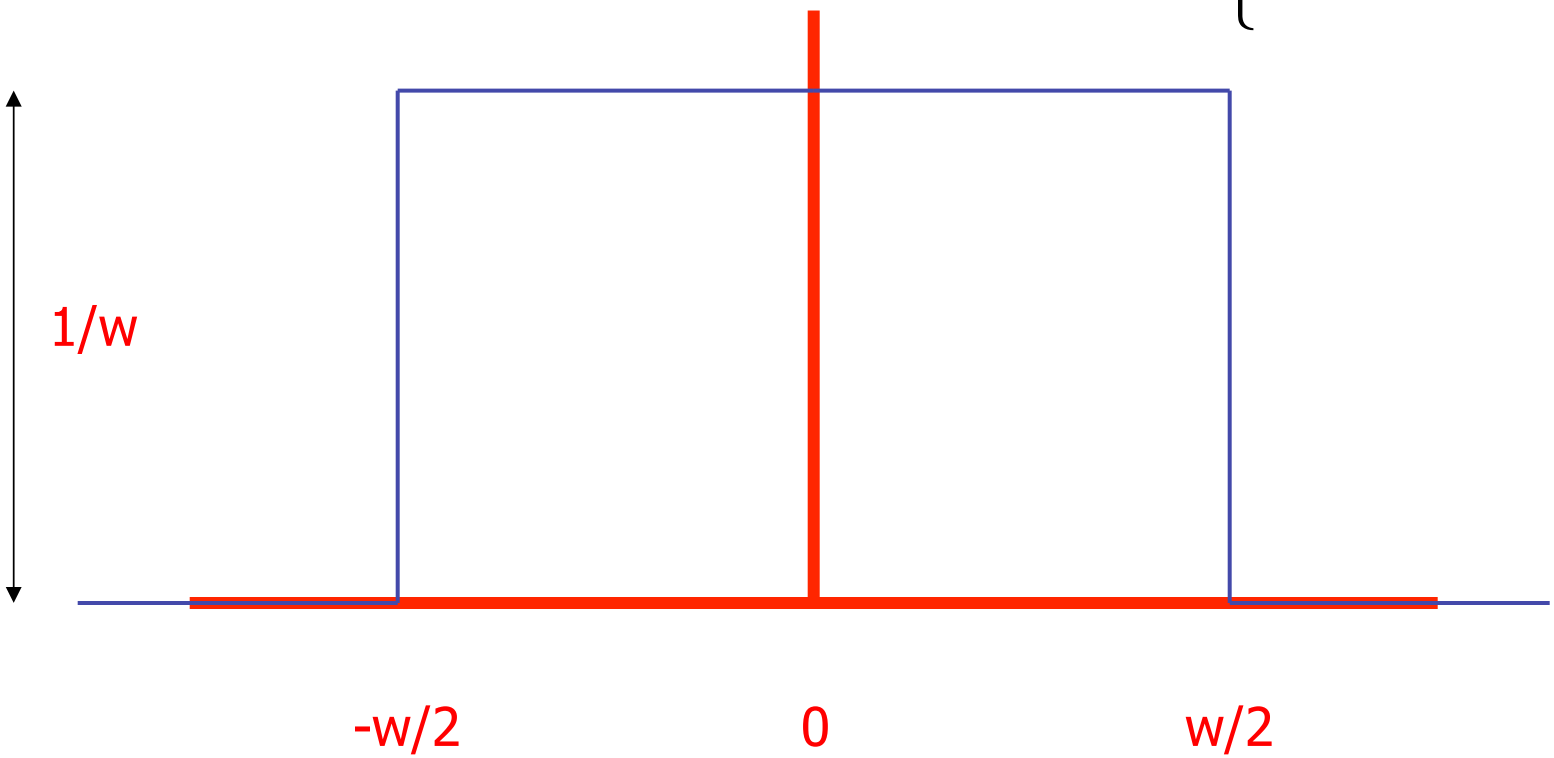
The "box" distribution

$$p(x) = \begin{cases} \frac{1}{w} & \text{if} \quad |x| \le \frac{w}{2} \\ 0 & \text{if} \quad |x| > \frac{w}{2} \end{cases}$$



The "box" distribution

$$p(x) = \begin{cases} \frac{1}{w} & \text{if} \quad |x| \le \frac{w}{2} \\ 0 & \text{if} \quad |x| > \frac{w}{2} \end{cases}$$



$$E[X] = 0 \qquad \text{Var}[X] = \frac{w^2}{12}$$

Entropy of a PDF

$$\text{Entropy of } X = H[X] = -\int_{x=-\infty}^{\infty} p(x)\log p(x)\,dx$$

Natural log (ln or $\log_e$)

The larger the entropy of a distribution…

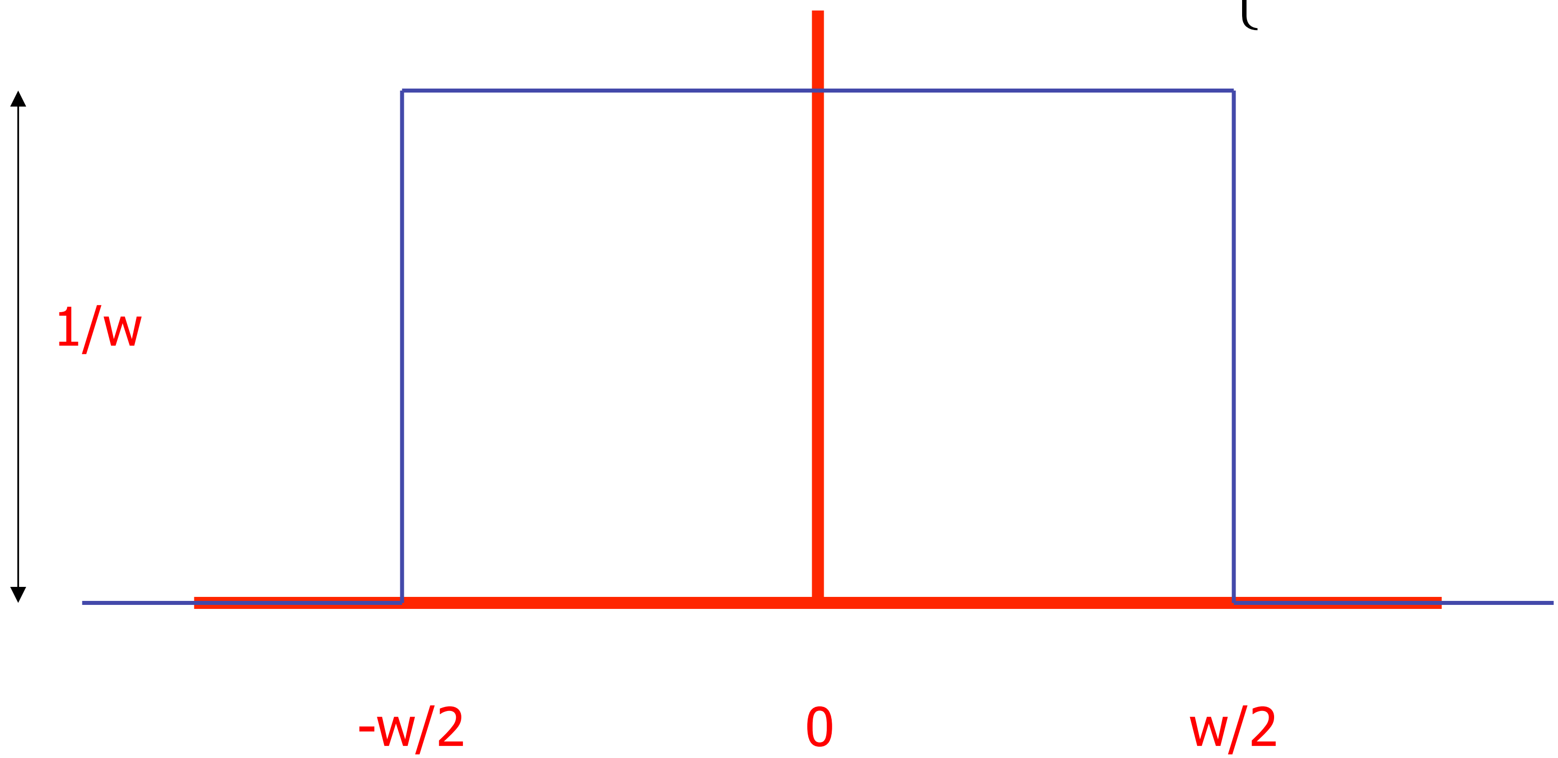
…the harder it is to predict

…the harder it is to compress it

…the less spiky the distribution

The "box" distribution

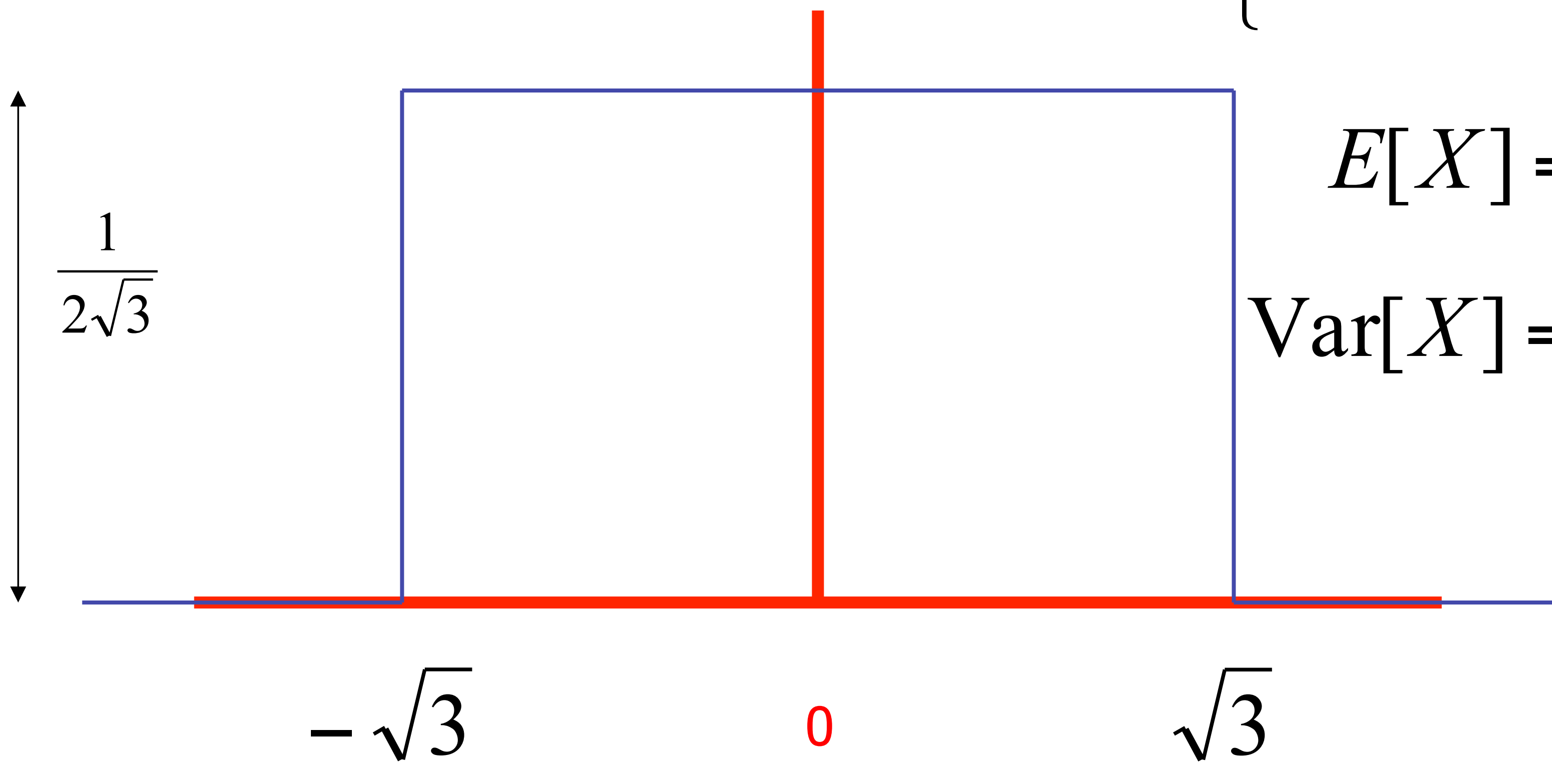
$$p(x) = \begin{cases} \frac{1}{w} & \text{if} \quad |x| \le \frac{w}{2} \\ 0 & \text{if} \quad |x| > \frac{w}{2} \end{cases}$$



$$H[X] = -\int_{x=-\infty}^{\infty} p(x)\log p(x)dx = -\int_{x=-w/2}^{w/2} \frac{1}{w}\log\frac{1}{w}dx = -\frac{1}{w}\log\frac{1}{w}\int_{x=-w/2}^{w/2} dx = \log w$$

Unit variance box distribution

$$p(x) = \begin{cases} \frac{1}{w} & \text{if} \quad |x| \le \frac{w}{2} \\ 0 & \text{if} \quad |x| > \frac{w}{2} \end{cases}$$



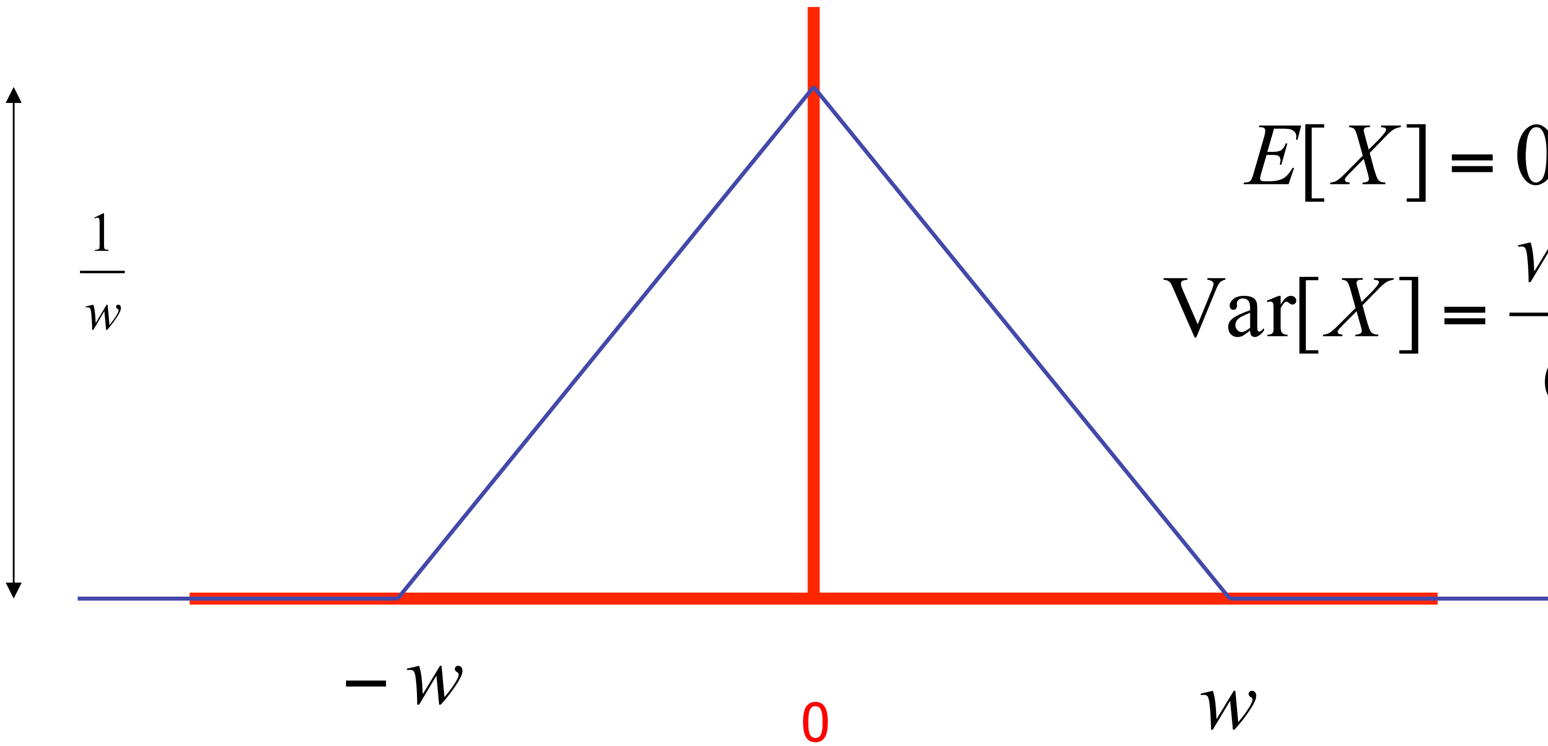
$$E[X] = 0$$

$$\text{Var}[X] = \frac{w^2}{12}$$

$$\text{if } w = 2\sqrt{3} \text{ then } \text{Var}[X] = 1 \text{ and } H[X] = 1.242$$

The Hat distribution

$$p(x) = \begin{cases} \frac{w - |x|}{w^2} & \text{if} \quad |x| \le w \\ 0 & \text{if} \quad |x| > w \end{cases}$$

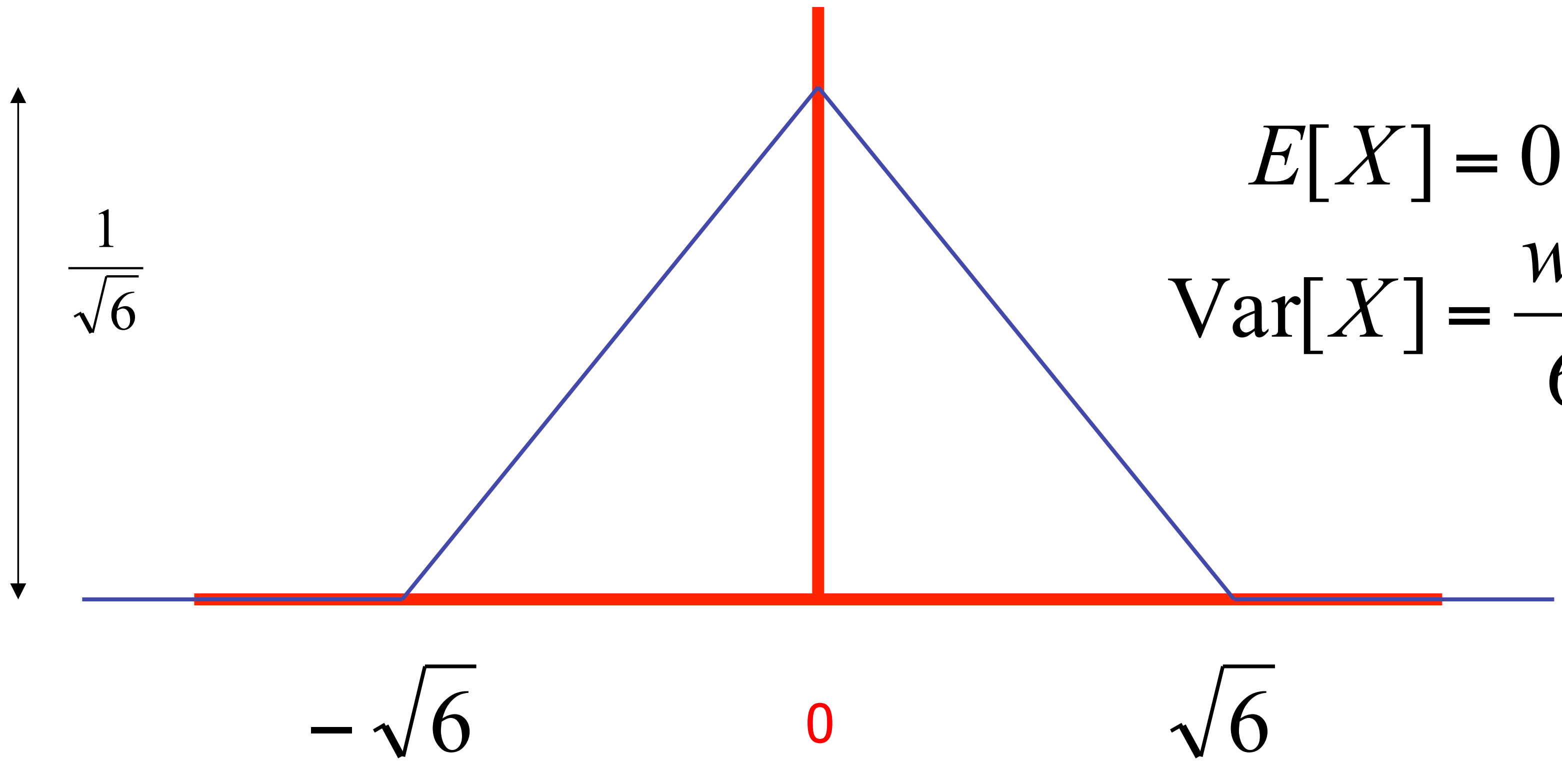


$$E[X] = 0$$

$$\mathrm{Var}[X] = \frac{w^2}{6}$$

Unit variance hat distribution

$$p(x) = \begin{cases} \dfrac{w - |x|}{w^2} & \text{if} \quad |x| \le w \\ 0 & \text{if} \quad |x| > w \end{cases}$$



$$E[X] = 0$$

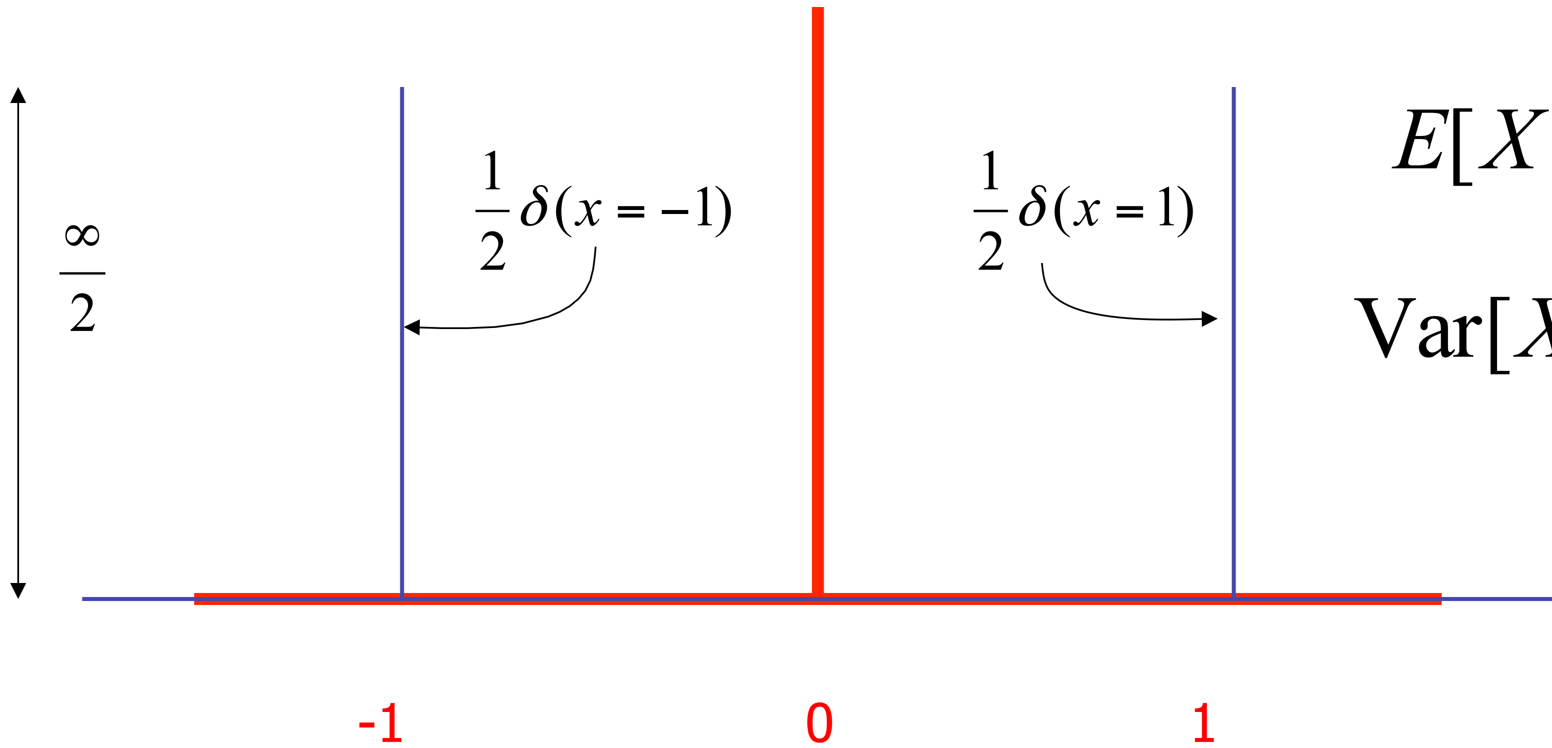
$$\text{Var}[X] = \frac{w^2}{6}$$

$$\text{if } w = \sqrt{6} \text{ then } \text{Var}[X] = 1 \text{ and } H[X] = 1.396$$

The "2 spikes" distribution

Dirac Delta

$$p(x) = \frac{\delta(x=-1) + \delta(x=1)}{2}$$



$$E[X] = 0$$

$$\mathrm{Var}[X] = 1$$

$$H[X] = -\int_{x=-\infty}^{\infty} p(x) \log p(x) dx = -\infty$$

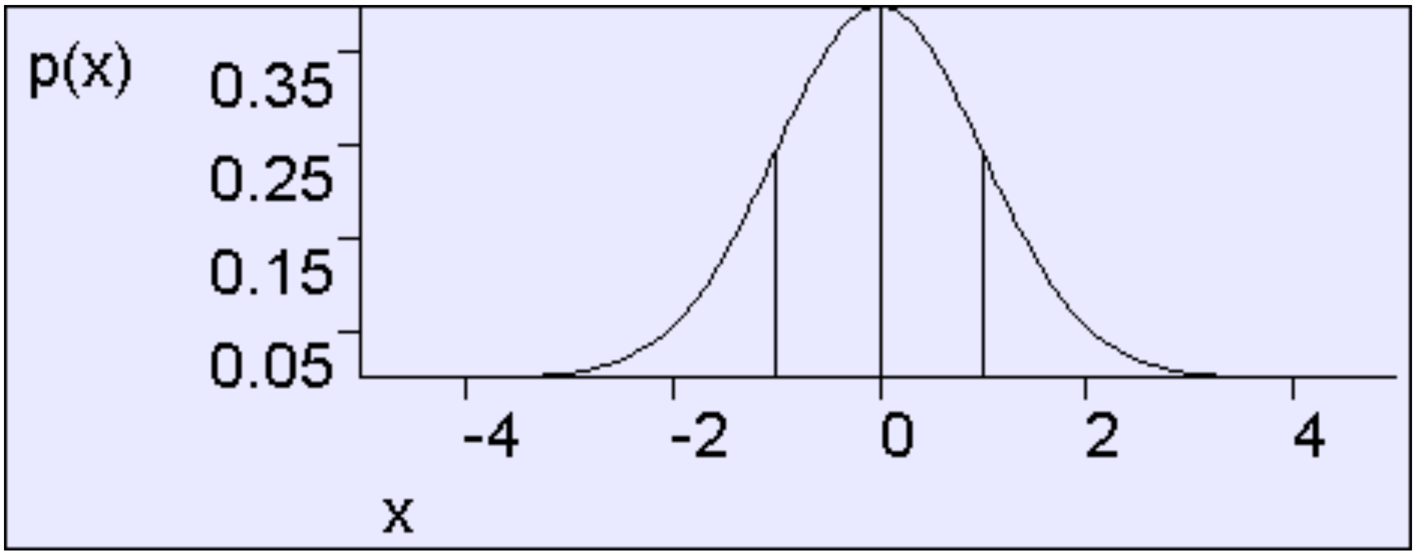
Entropies of unit-variance distributions

Distribution	Entropy
Box	1.242
Hat	1.396
2 spikes	-infinity
???	1.4189

Largest possible entropy of any unit-variance distribution

Unit variance Gaussian

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



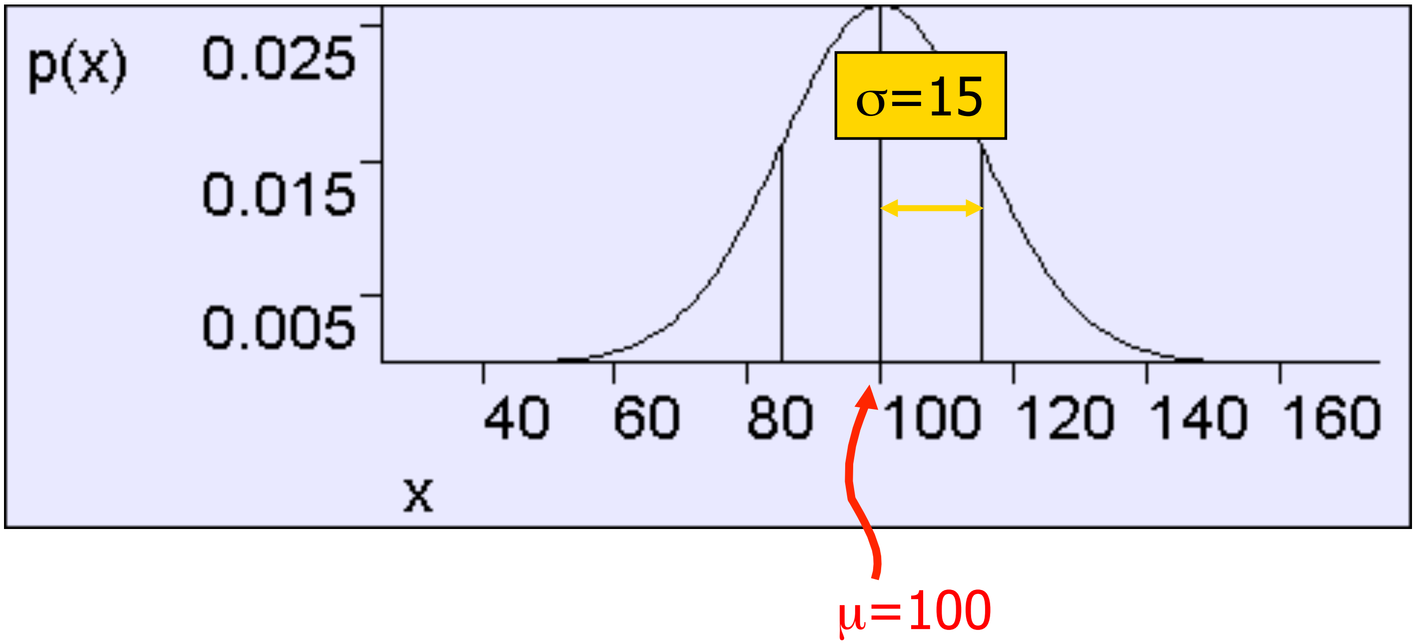
$$E[X] = 0$$

$$\mathrm{Var}[X] = 1$$

$$H[X] = -\int_{x=-\infty}^{\infty} p(x) \log p(x) dx = 1.4189$$

General Gaussian

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



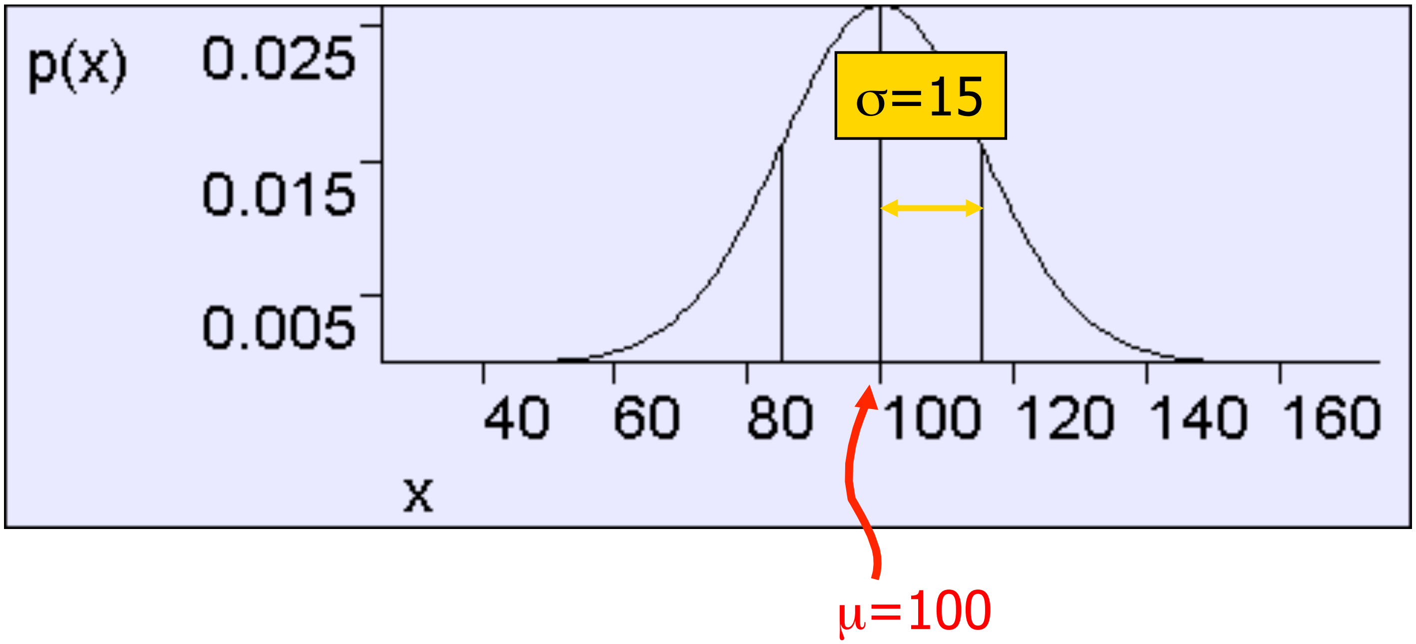
$$E[X] = \mu$$

$$\mathrm{Var}[X] = \sigma^2$$

General Gaussian

Also known as the normal distribution or Bell-shaped curve

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



$$E[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

Shorthand: We say X ~ N(μ,σ^2) to mean "X is distributed as a Gaussian with parameters μ and σ^2".

In the above figure, X ~ N(100,15^2)

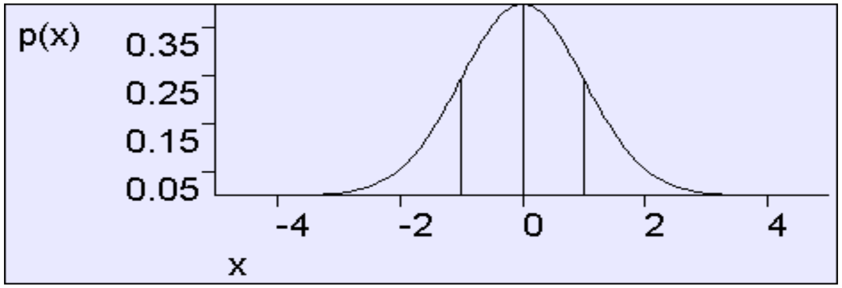
The Error Function

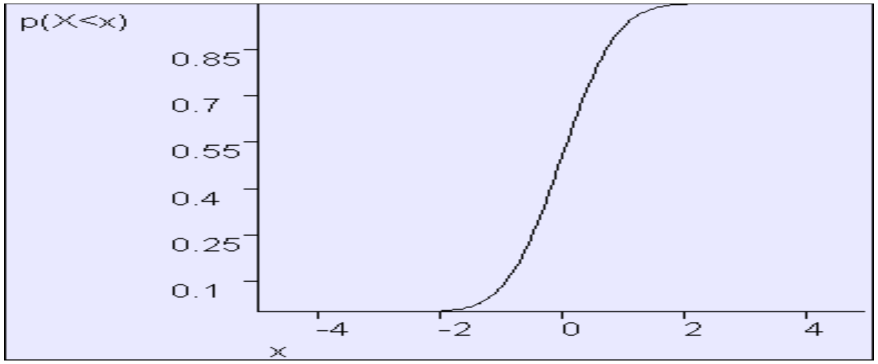
Assume X ~ N(0,1)

Define ERF(x) = P(X<x) = Cumulative Distribution of X

$$ERF(x) = \int_{z=-\infty}^{x} p(z)\,dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{z=-\infty}^{x} \exp\left(-\frac{z^2}{2}\right) dz$$



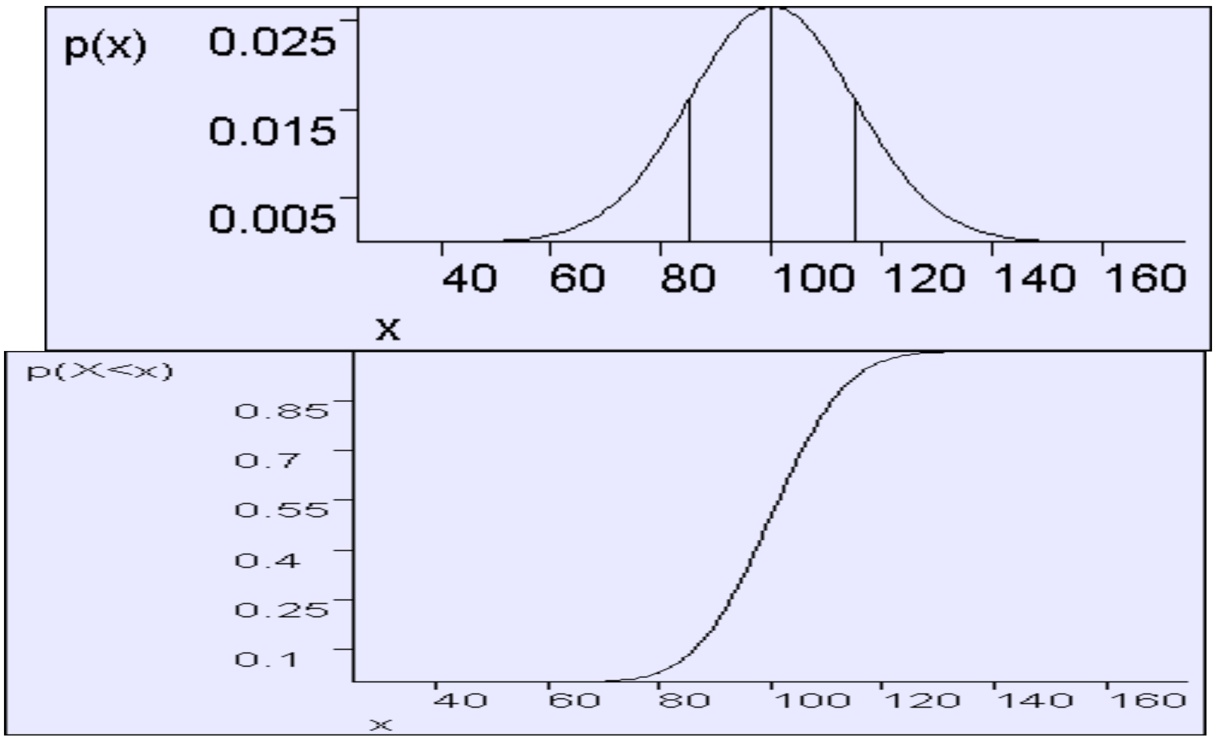


Using The Error Function

Assume X ~ N(μ,σ^2)

$$ERF(\frac{x-\mu}{\sigma^2})$$

P(X<x| μ,σ^2) =



The Central Limit Theorem

- If $(X_1, X_2, \ldots X_n)$ are i.i.d. continuous random variables
- Then define $$z = f(x_1, x_2, \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- As n-->infinity, p(z)--->Gaussian with mean $E[X_i]$ and variance $Var[X_i]$

Somewhat of a justification for assuming Gaussian noise is common

Other amazing facts about Gaussians

- Wouldn't you like to know?

- We will not examine them until we need to.

Bivariate Gaussians

Write r.v. $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ine to mean

$$p(\mathbf{x}) = \frac{1}{2\pi \| \boldsymbol{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Where the Gaussian's parameters are…

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

Where we insist that $\boldsymbol{\Sigma}$ is symmetric non-negative definite

Bivariate Gaussians

$$\text{Write r.v. } \mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ine to mean

$$p(\mathbf{x}) = \frac{1}{2\pi \| \boldsymbol{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \; \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Where the Gaussian's parameters are…

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} {\sigma^2}_x & \sigma_{xy} \\ \sigma_{xy} & {\sigma^2}_y \end{pmatrix}$$

Where we insist that $\boldsymbol{\Sigma}$ is symmetric non-negative definite

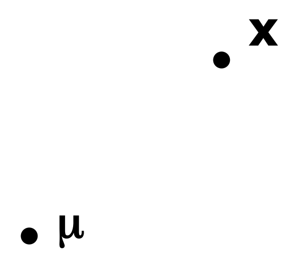
It turns out that E[X] = μ and Cov[X] = $\boldsymbol{\Sigma}$. (Note that this is a resulting property of Gaussians, not a definition)*

*This note rates 7.4 on the pedanticness scale

Evaluating p(x): Step 1

$$p(\mathbf{x}) = \frac{1}{2\pi \| \boldsymbol{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \ \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

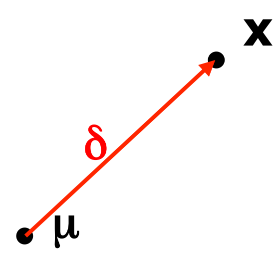
1. Begin with vector **x**



Evaluating p(x): Step 2

$$p(\mathbf{x}) = \frac{1}{2\pi \| \mathbf{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \; \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

1. Begin with vector **x**
2. Define $\delta = \mathbf{x} - \mu$

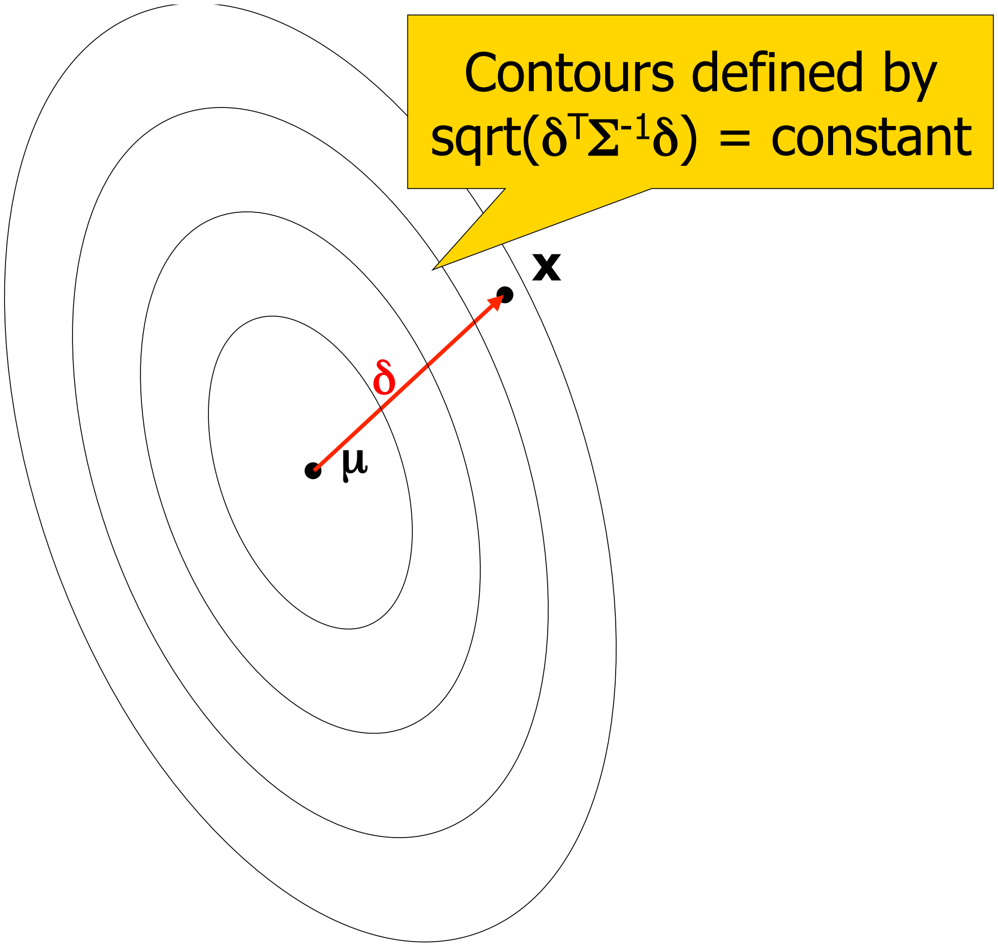


Evaluating p(x): Step 3

$$p(\mathbf{x}) = \frac{1}{2\pi \|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

1. Begin with vector **x**
2. Define δ **= x -** μ
3. Count the number of contours crossed of the ellipsoids formed Σ^{-1}

D = this count = sqrt($\delta^T\Sigma^{-1}\delta$) = Mahalonobis Distance between **x** and μ

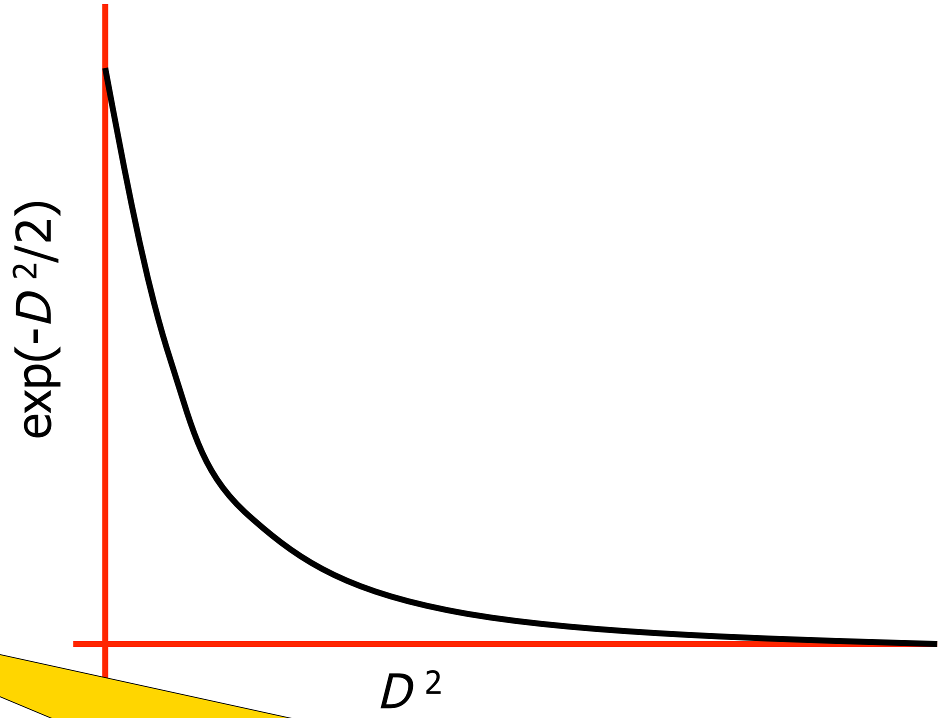


Evaluating p(x): Step 4

$$p(\mathbf{x}) = \frac{1}{2\pi \|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

1. Begin with vector **x**
2. Define δ **= x -** μ
3. Count the number of contours crossed of the ellipsoids formed Σ^{-1}

   D = this count = sqrt($\delta^T \Sigma^{-1} \delta$) = Mahalonobis Distance between **x** and μ
4. Define w = exp(-D^2/2)



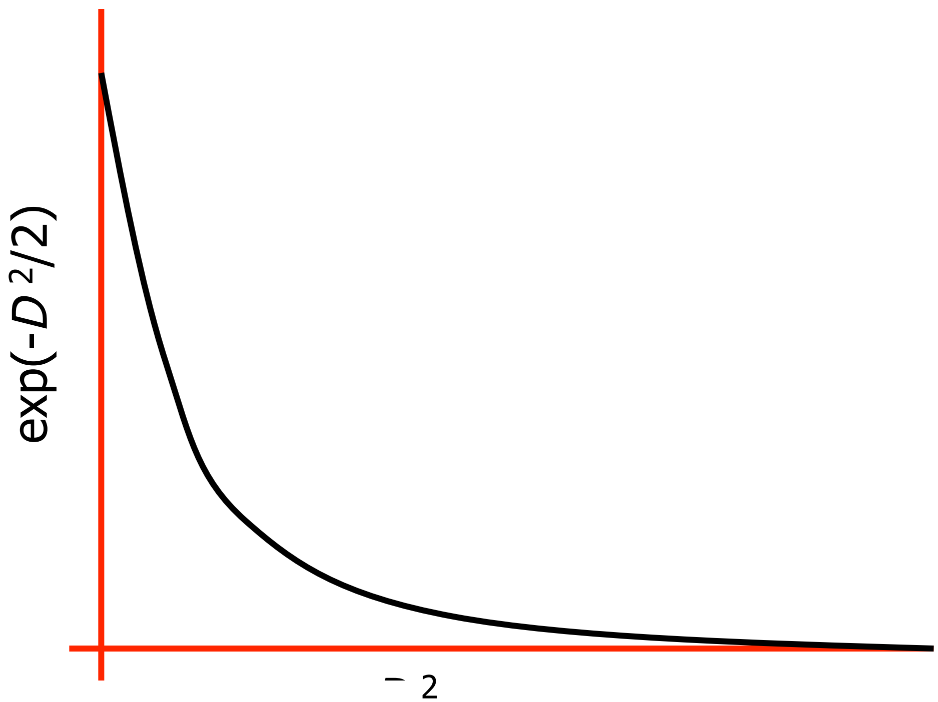
x close to μ in squared Mahalonobis space gets a large weight. Far away gets a tiny weight

Evaluating p(x): Step 5

$$p(\mathbf{x}) = \frac{1}{2\pi \|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

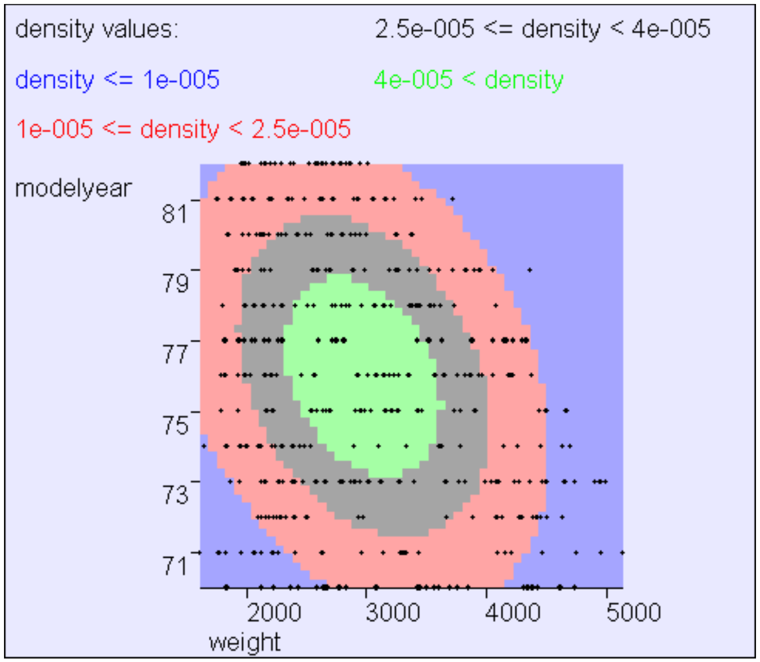
1. Begin with vector **x**
2. Define $\delta = \mathbf{x} - \mu$
3. Count the number of contours crossed of the ellipsoids formed Σ^{-1}

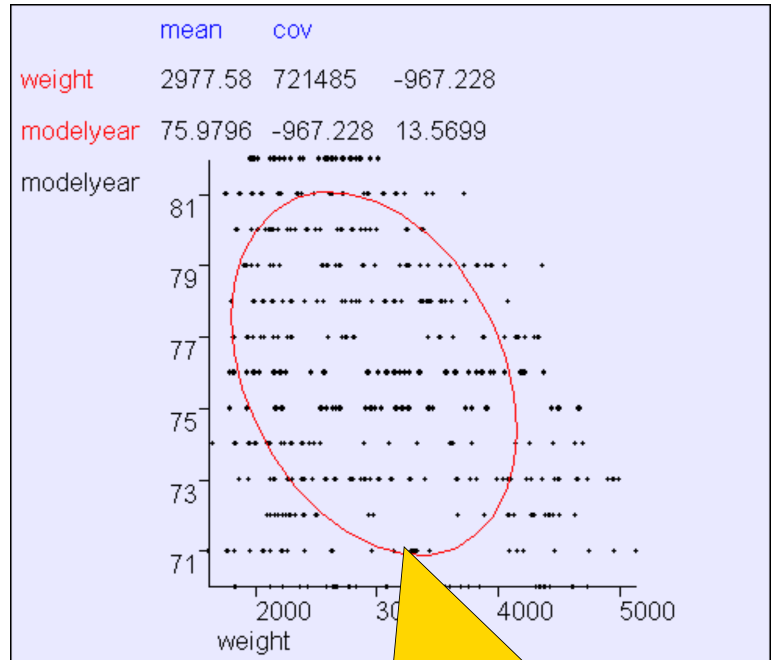
   D = this count = sqrt($\delta^T\Sigma^{-1}\delta$) = Mahalonobis Distance between **x** and μ
4. Define w = exp($-D^2/2$)



Multiply w by $\frac{1}{\sqrt{2\pi} \|\boldsymbol{\Sigma}\|^{1/2}}$ to ensure $\int p(\mathbf{x})d\mathbf{x} = 1$

Example

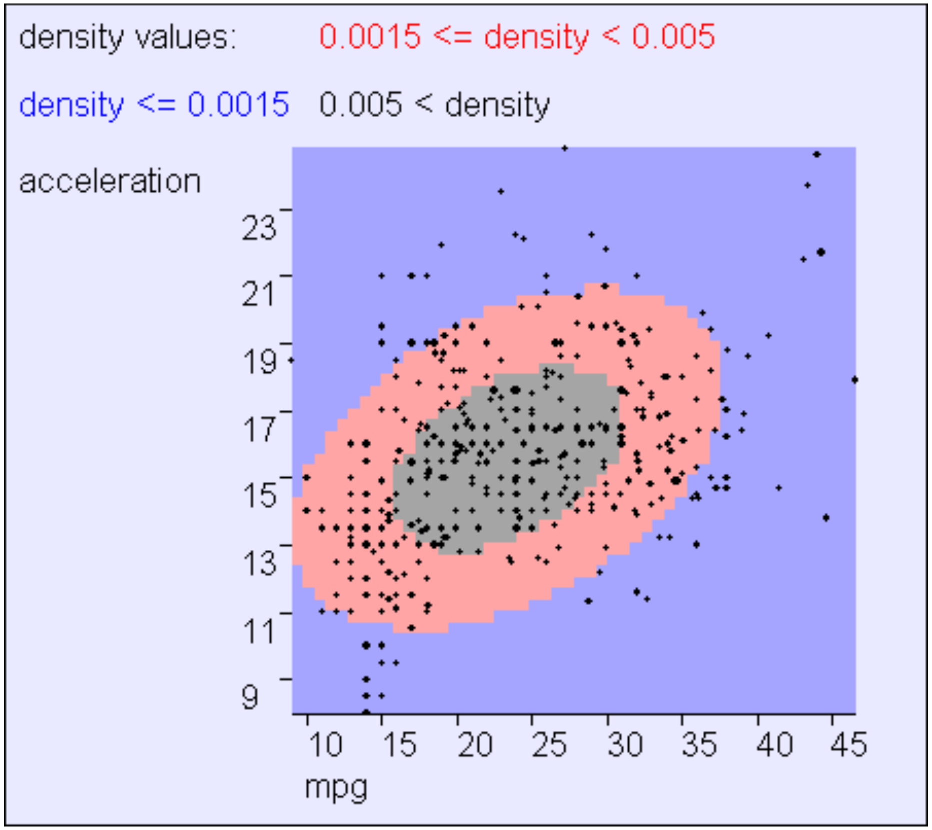


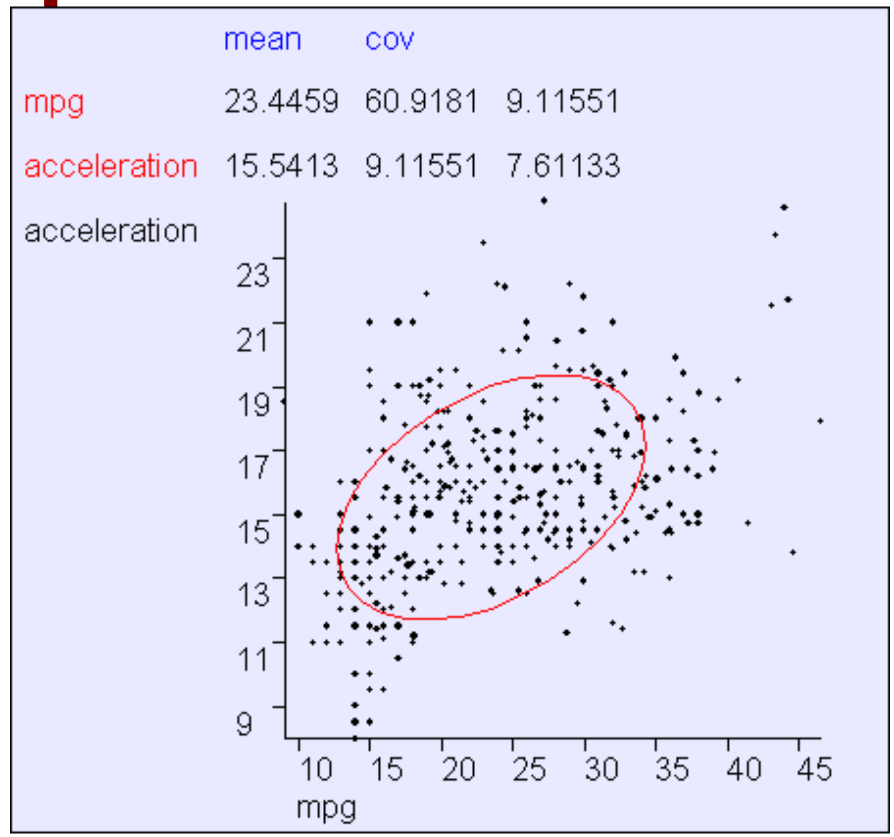


Observe: Mean, Principal axes, implication of off-diagonal covariance term, max gradient zone of p(x)

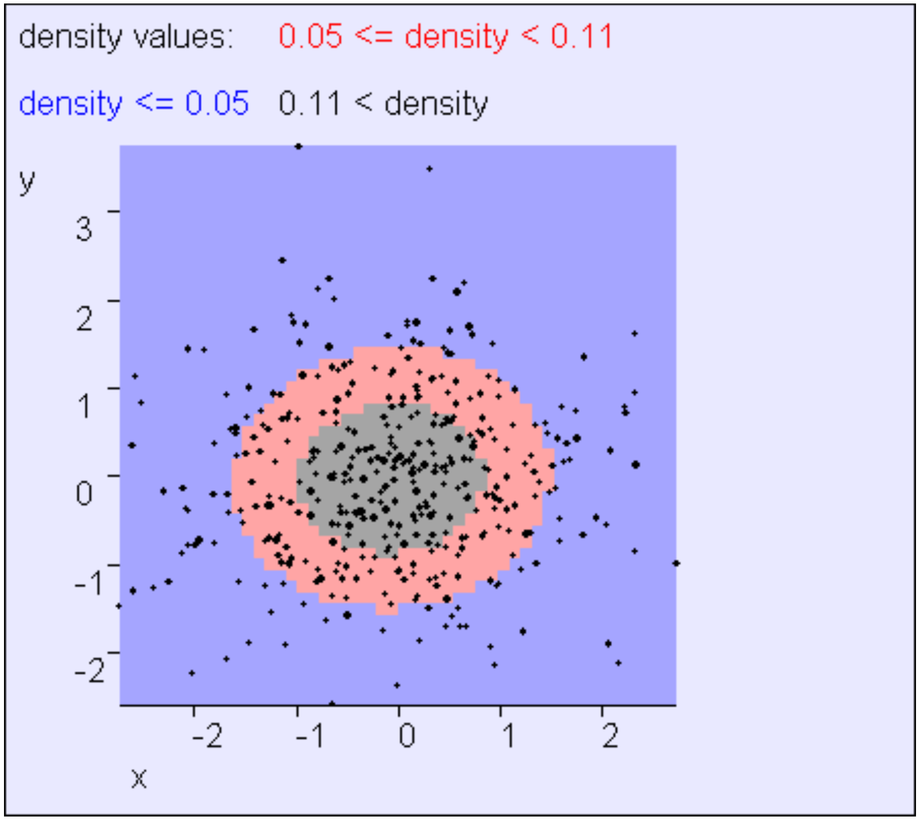
Common convention: show contour corresponding to 2 standard deviations from mean

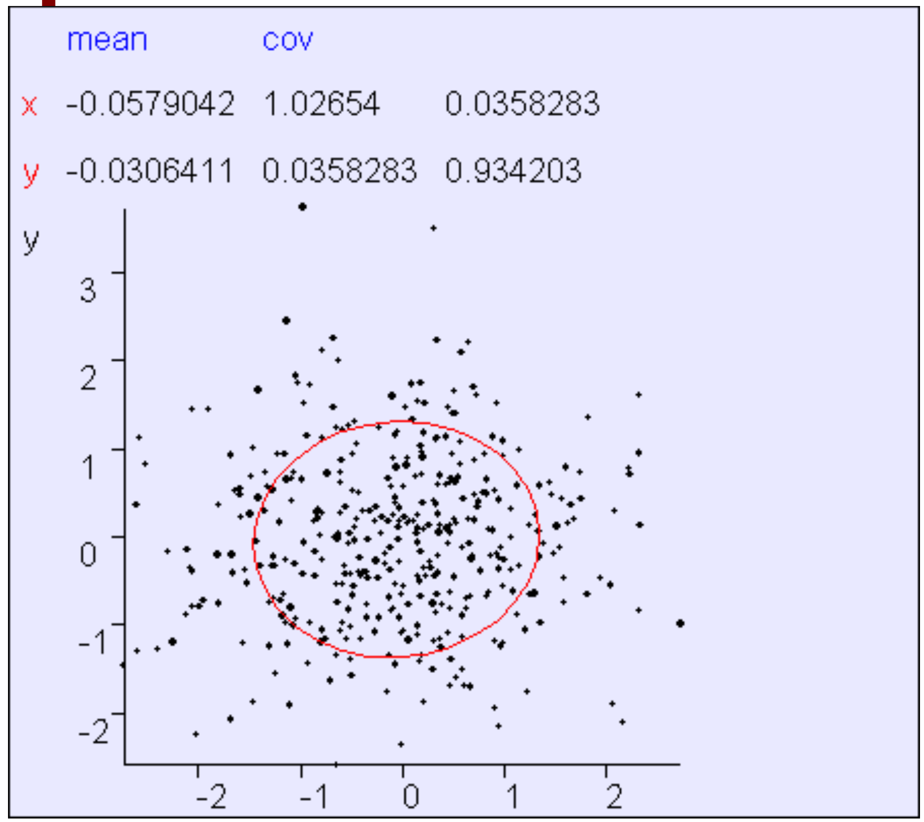
Example


density values:
0.0015 <= density < 0.005
density <= 0.0015
0.005 < density
acceleration
23
21
19
17
15
13
11
9
10
15
20
25
30
35
40
45
mpg


mean
cov
mpg
23.4459
60.9181
9.11551
acceleration
15.5413
9.11551
7.61133
acceleration
23
21
19
17
15
13
11
9
10
15
20
25
30
35
40
45
mpg

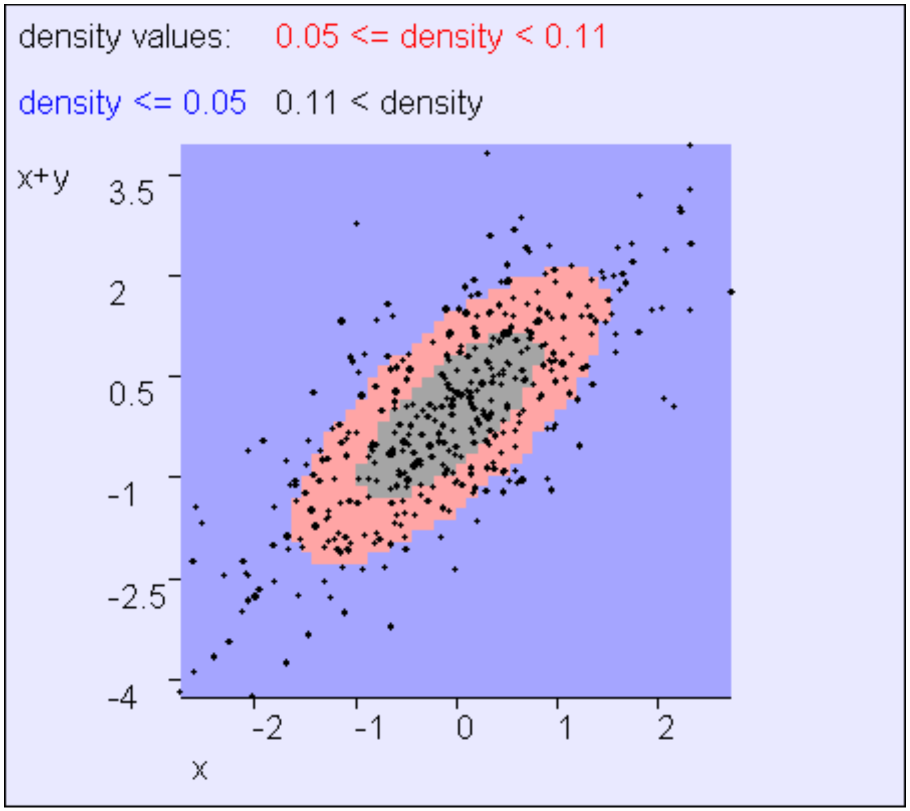
Example

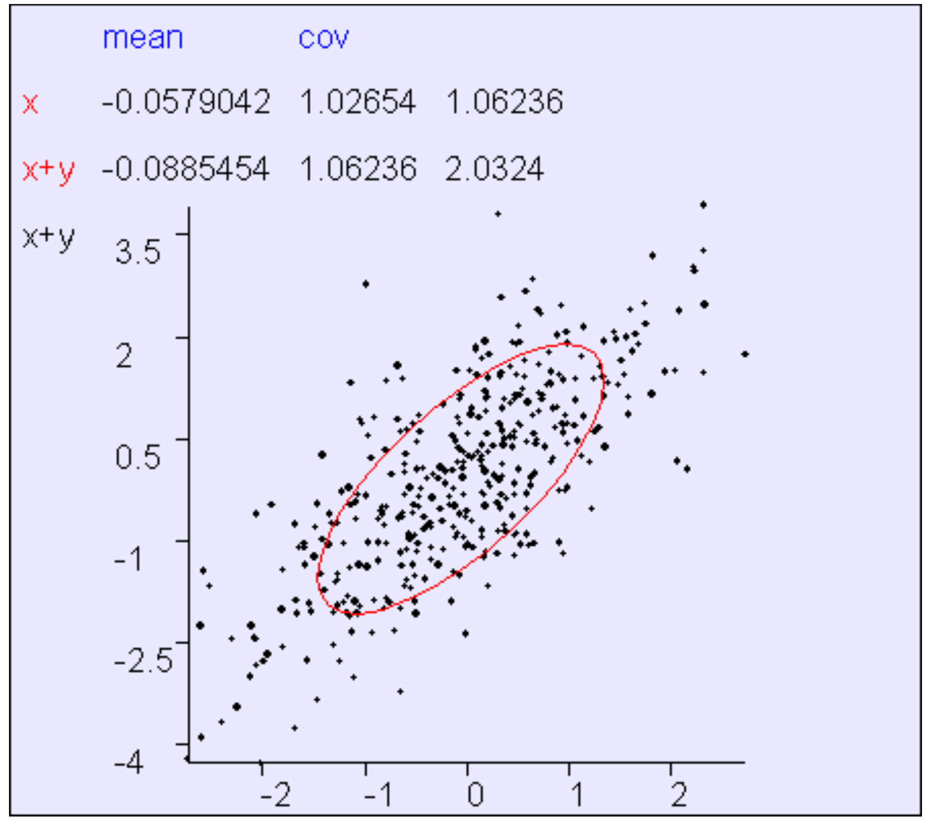




In this example, x and y are almost independent

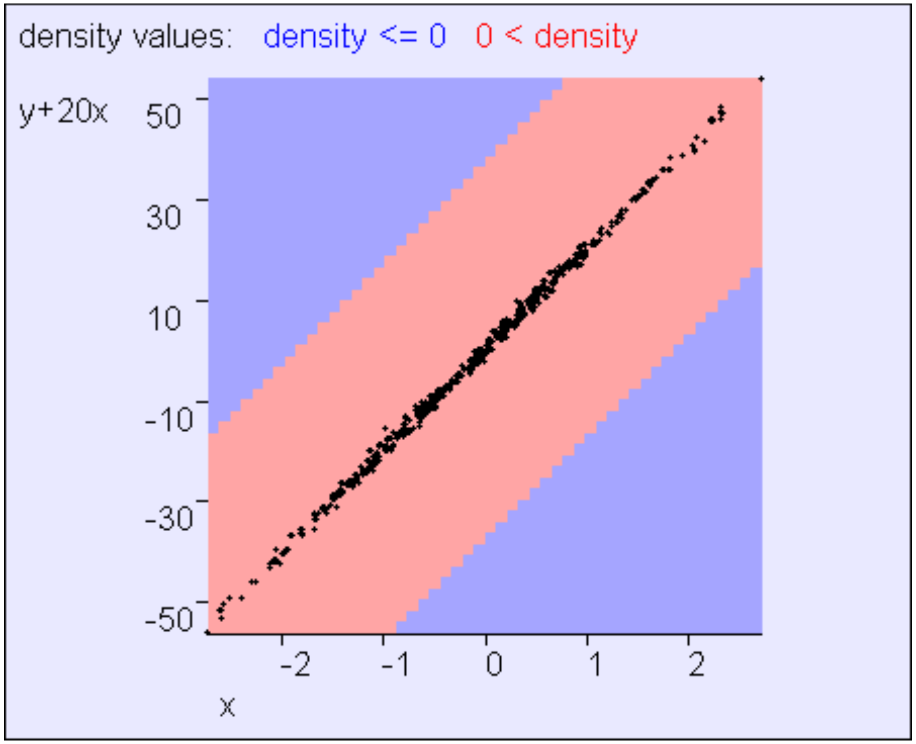
Example

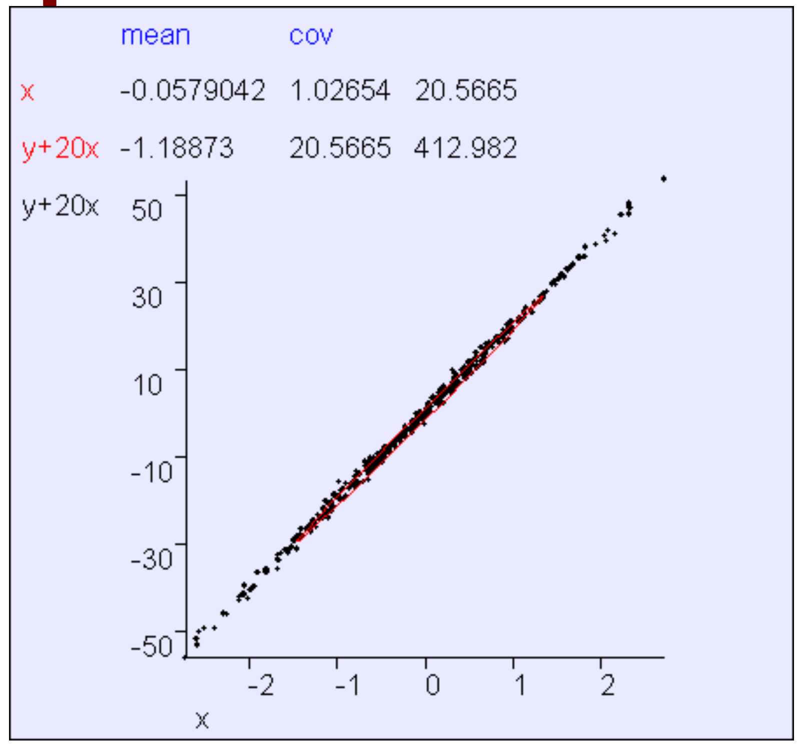




In this example, x and "x+y" are clearly not independent

Example





In this example, x and “20x+y” are clearly not independent

Multivariate Gaussians

$$\text{Write r.v. } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$$

$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ine to mean

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \| \boldsymbol{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Where the Gaussian's parameters have…

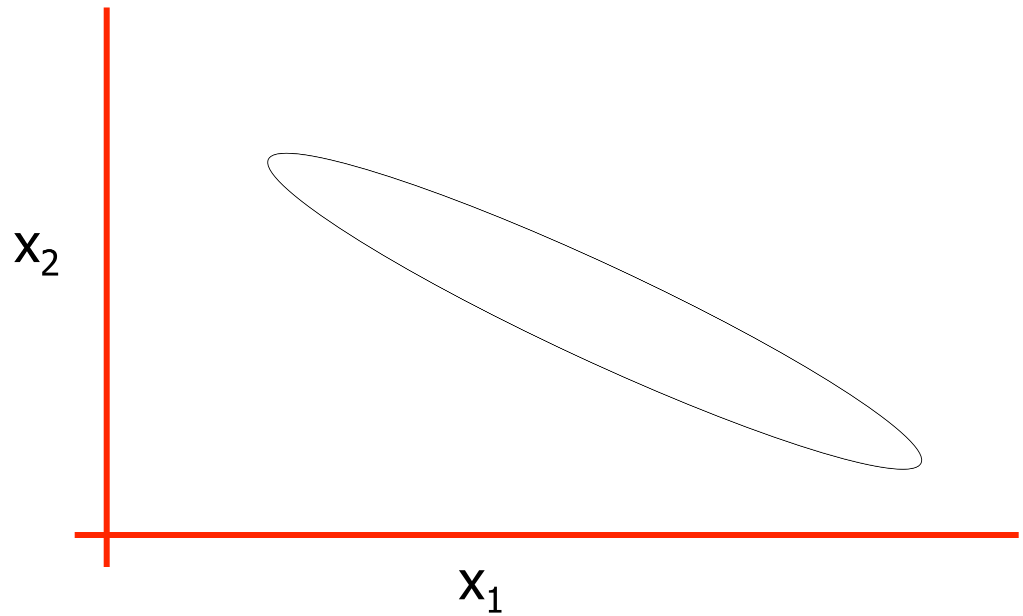
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma^2{}_m \end{pmatrix}$$

Where we insist that $\boldsymbol{\Sigma}$ is symmetric non-negative definite

Again, E[X] = μ and Cov[X] = $\boldsymbol{\Sigma}$. (Note that this is a resulting property of Gaussians, not a definition)

General Gaussians

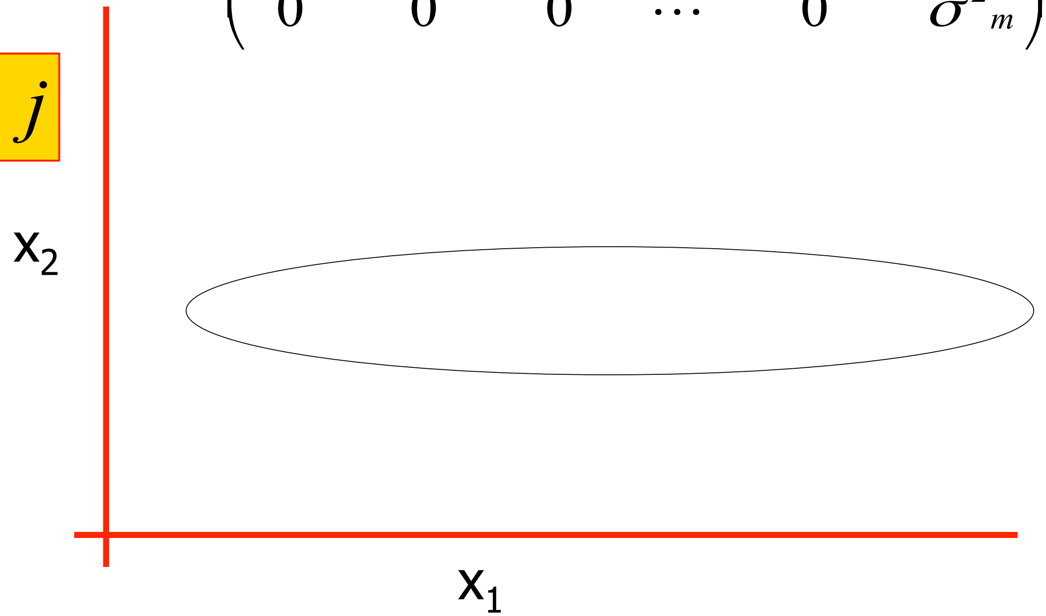
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma^2{}_m \end{pmatrix}$$



Axis-Aligned Gaussians

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2{}_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2{}_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2{}_{m-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2{}_m \end{pmatrix}$$

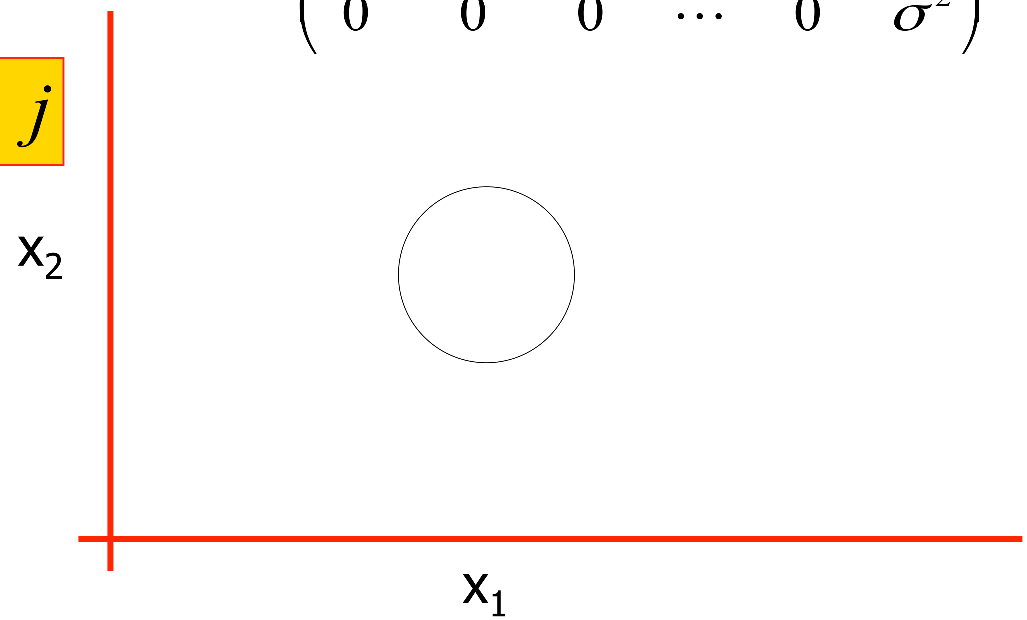
$$X_i \perp X_i \text{ for } i \neq j$$



Spherical Gaussians

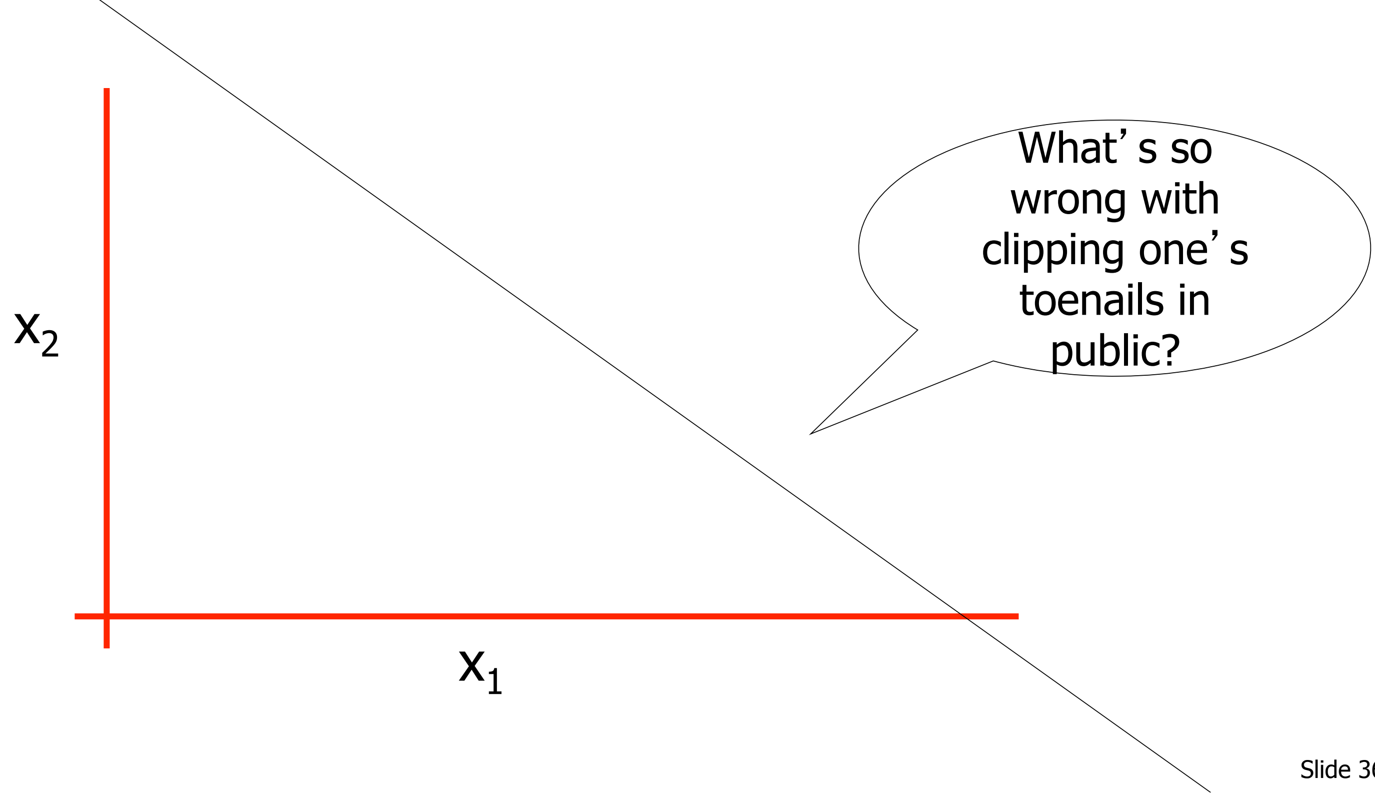
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

$X_i \perp X_i \text{ for } i \neq j$



Degenerate Gaussians

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \| \boldsymbol{\Sigma} \| = 0$$



Where are we now?

- **We've seen the formulae for Gaussians**
- **We have an intuition of how they behave**
- **We have some experience of "reading" a Gaussian's covariance matrix**

- **Coming next:**

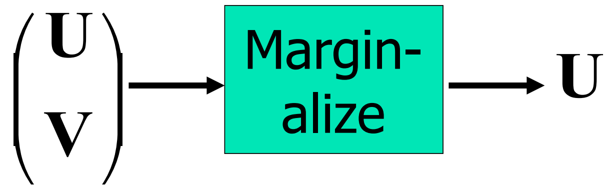
Some useful tricks with Gaussians

Subsets of variables

$$\text{Write } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \text{ as } \mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \text{ where } \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ X_{m(u)} \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} X_{m(u)+1} \\ \vdots \\ X_m \end{pmatrix}$$

This will be our standard notation for breaking an m-dimensional distribution into subsets of variables

Gaussian Marginals are Gaussian



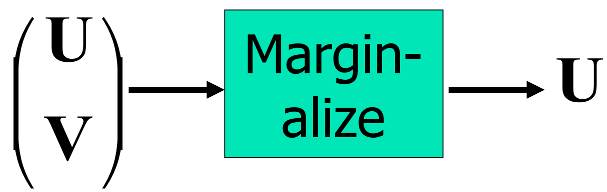
$$\text{Write } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \text{ as } \mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \text{ where } \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ X_{m(u)} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} X_{m(u)+1} \\ \vdots \\ X_m \end{pmatrix}$$

$$\text{IF} \quad \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$$

THEN U is also distributed as a Gaussian

$$\mathbf{U} \sim \mathrm{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu})$$

Gaussian Marginals are Gaussian



$$\text{Write } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \text{ as } \mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \text{ where } \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ X_{m(u)} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} X_{m(u)+1} \\ \vdots \\ X_m \end{pmatrix}$$

$$\text{IF} \quad \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$$

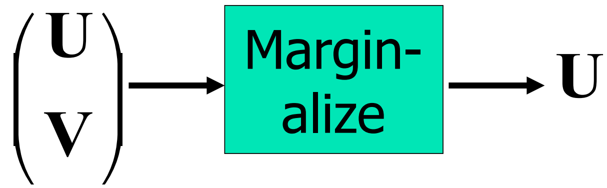
This fact is not immediately obvious

THEN U is also distributed as a Gaussian

$$\mathbf{U} \sim \mathrm{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu})$$

Obvious, once we know it’ s a Gaussian (why?)

Gaussian Marginals are Gaussian



$$\text{Write } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \text{ as } \mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \text{ where } \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ \end{pmatrix}, \mathbf{V} = \begin{pmatrix} X_{m(u)+1} \\ \vdots \\ \end{pmatrix}$$

$$\text{IF } \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$$

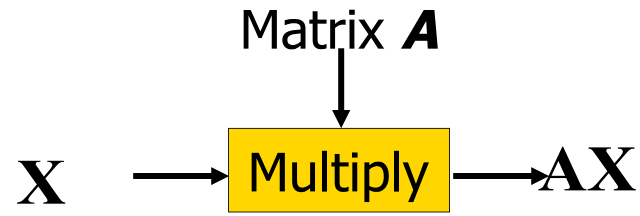
THEN U is also distributed as

$$\mathbf{U} \sim \mathrm{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu})$$

How would you prove this?

$$p(\mathbf{u}) = \int_{\mathbf{v}} p(\mathbf{u}, \mathbf{v}) d\mathbf{v} = (\text{snore...})$$

Linear Transforms remain Gaussian



Assume X is an m-dimensional Gaussian r.v.

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Define Y to be a p-dimensional r. v. thusly ($p \leq m$):

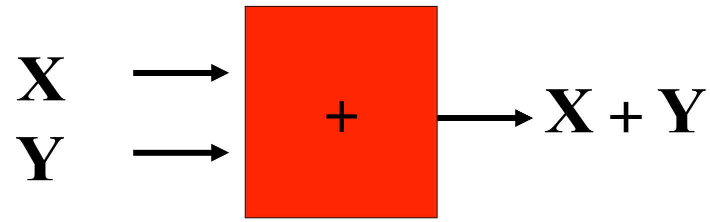
$$\mathbf{Y} = \mathbf{AX}$$

...where A is a p x m matrix. Then...

$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\, \mathbf{A}^{T}\right)$$

Note: the "subset" result is a special case of this result

Adding samples of two independent Gaussians is Gaussian



$$\text{if } \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \text{ and } \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \text{ and } \mathbf{X} \perp \mathbf{Y}$$

$$\text{then } \mathbf{X} + \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$$

Why doesn't this hold if X and Y are dependent?

Which of the below statements is true?

If X and Y are dependent, then X+Y is Gaussian but possibly with some other covariance

If X and Y are dependent, then X+Y might be non-Gaussian

Conditional of Gaussian is Gaussian

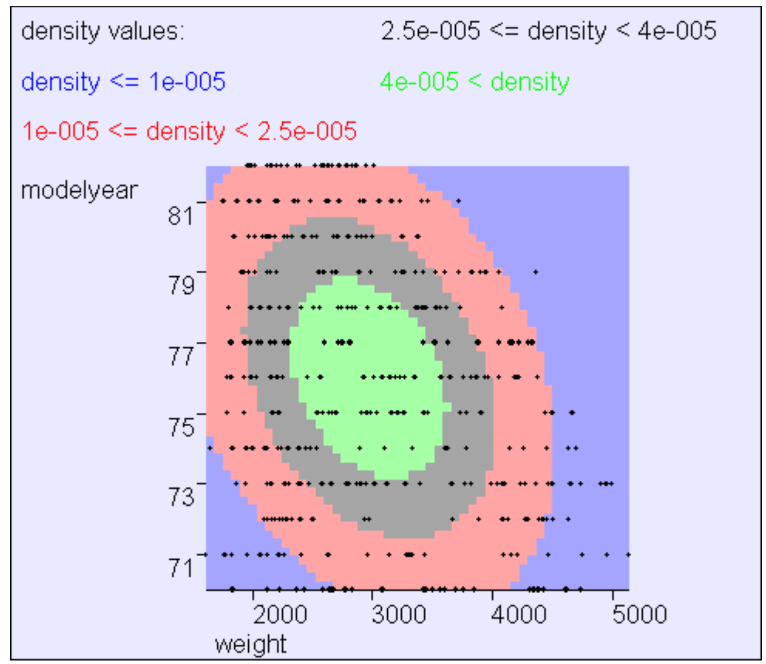
$$\begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix} \rightarrow \boxed{\text{Condition-alize}} \rightarrow \mathbf{U}\,|\,\mathbf{V}$$

$$\text{IF} \quad \begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\mu}_u\\\boldsymbol{\mu}_v\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv}\\\boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv}\end{pmatrix}\right)$$

$$\text{THEN} \quad \mathbf{U}\,|\,\mathbf{V} \sim \mathrm{N}\left(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v}\right) \text{where}$$

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$$

$$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{uv}$$



$$\text{IF} \quad \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$$

$$\text{THEN} \quad \mathbf{U} \mid \mathbf{V} \sim \mathrm{N}\left(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v}\right) \text{where}$$

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$$

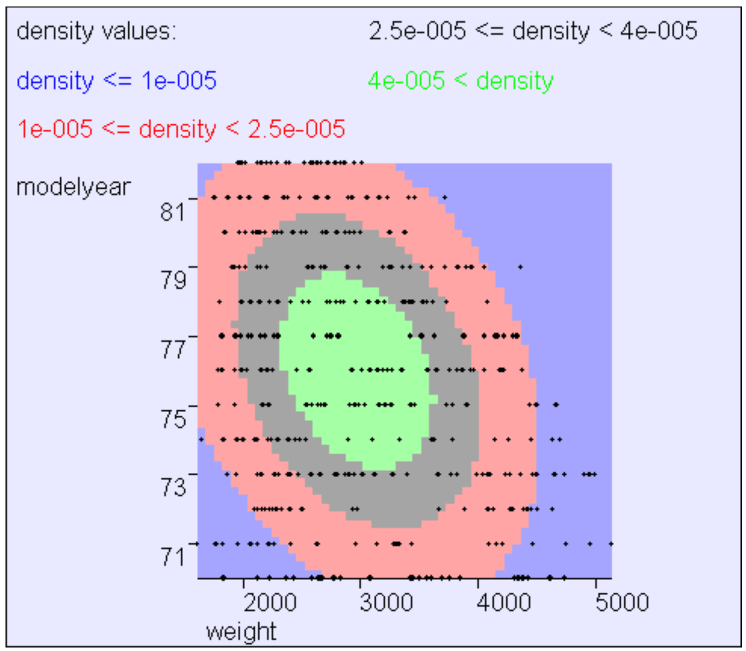
$$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{uv}$$

$$\text{IF} \quad \begin{pmatrix} w \\ y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 2977 \\ 76 \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ -967 & 3.68^2 \end{pmatrix} \right)$$

$$\text{THEN} \quad w \mid y \sim \mathrm{N}\left(\boldsymbol{\mu}_{w|y}, \boldsymbol{\Sigma}_{w|y}\right) \text{where}$$

$$\boldsymbol{\mu}_{w|y} = 2977 - \frac{976(y-76)}{3.68^2}$$

$$\boldsymbol{\Sigma}_{w|y} = 849^2 - \frac{967^2}{3.68^2} = 808^2$$



IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$

THEN $\mathbf{U} \mid \mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$ where

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$$

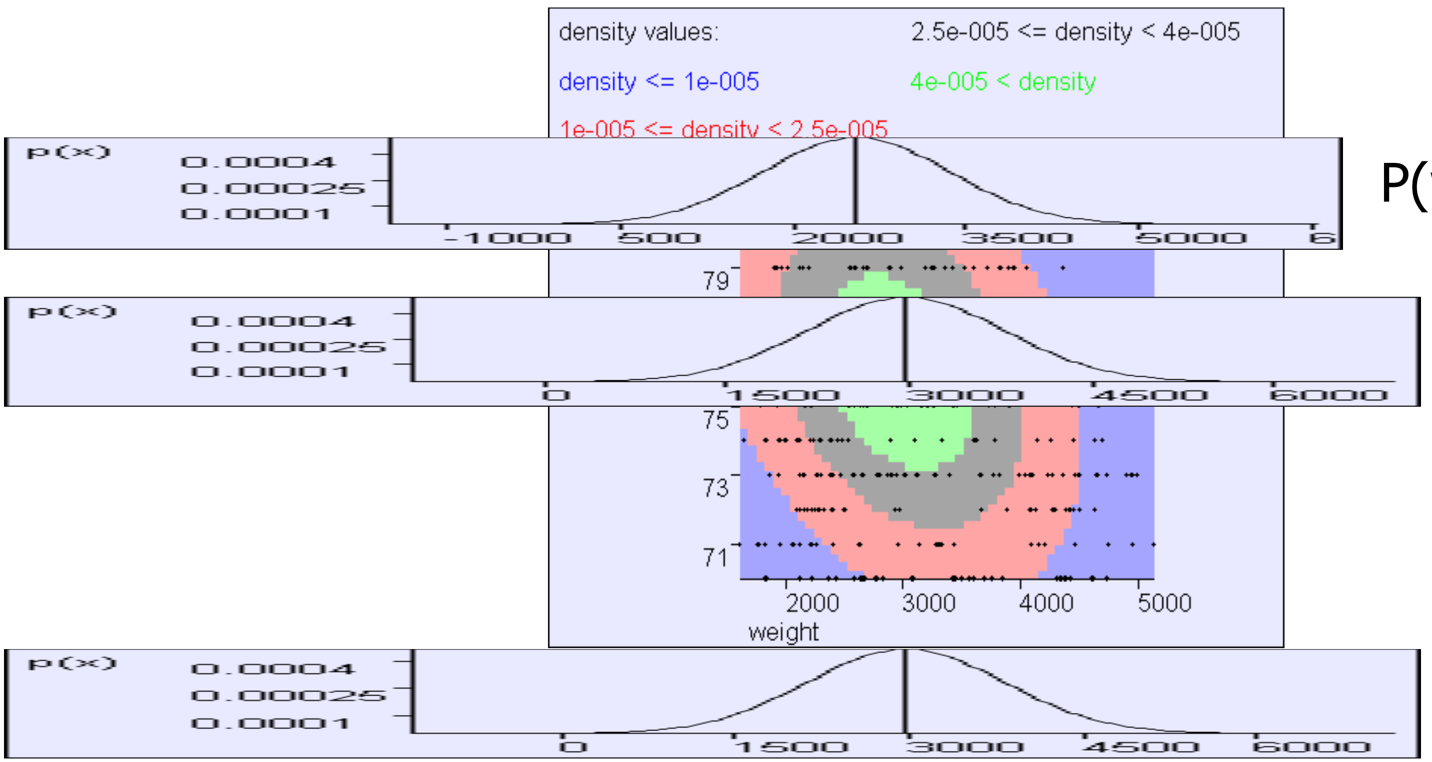
$$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{uv}$$

IF $\begin{pmatrix} w \\ y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 2977 \\ 76 \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ -967 & 3.68^2 \end{pmatrix} \right)$

THEN $w \mid y \sim \mathrm{N}(\boldsymbol{\mu}_{w|y}, \boldsymbol{\Sigma}_{w|y})$ where

$$\boldsymbol{\mu}_{w|y} = 2977 - \frac{976(y - 76)}{3.68^2}$$

$$\boldsymbol{\Sigma}_{w|y} = 849^2 - \frac{967^2}{3.68^2} = 808^2$$



P(w|m=82)

P(w|m=76)

P(w)

$$\text{IF} \quad \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}^T_{uv} & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$$

$$\text{THEN} \quad \mathbf{U} \mid \mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v}) \text{ where}$$

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}^T_{uv} \boldsymbol{\Sigma}^{-1}_{vv} (\mathbf{V} - \boldsymbol{\mu}_v)$$

$$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}^T_{uv} \boldsymbol{\Sigma}^{-1}_{vv} \boldsymbol{\Sigma}_{uv}$$

IF $\begin{pmatrix} w \\ \ldots \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 2977 \\ \ldots \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ \ldots & \ldots^2 \end{pmatrix} \right)$

THEN ...

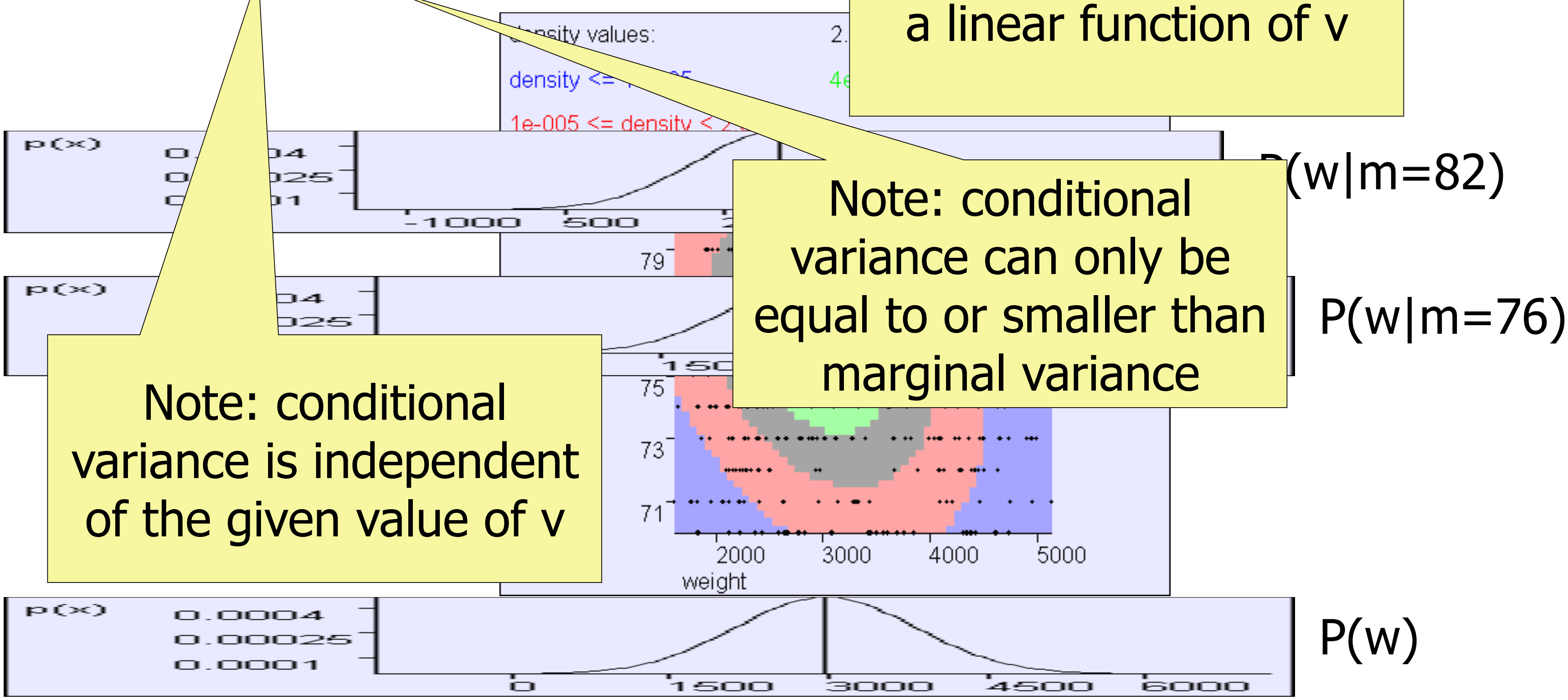
$\boldsymbol{\mu}_{w|y} = 2977 - \dfrac{\ldots}{3.68^2}$

Note: when given value of v is μ_v, the conditional mean of u is μ_u

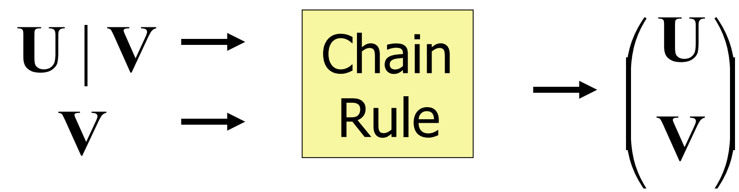
Note: marginal mean is a linear function of v

Note: conditional variance can only be equal to or smaller than marginal variance

Note: conditional variance is independent of the given value of v



Gaussians and the chain rule



Let A be a constant matrix

$$\text{IF} \quad \mathbf{U} \mid \mathbf{V} \sim \mathrm{N}\left(\mathbf{AV}, \mathbf{\Sigma}_{u|v}\right) \text{and} \quad \mathbf{V} \sim \mathrm{N}\left(\boldsymbol{\mu}_v, \mathbf{\Sigma}_{vv}\right)$$

$$\text{THEN} \quad \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma}), \text{ with}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mathbf{A}\boldsymbol{\mu}_v \\ \boldsymbol{\mu}_v \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{A\Sigma}_{vv}\mathbf{A}^T + \mathbf{\Sigma}_{u|v} & \mathbf{A\Sigma}_{vv} \\ (\mathbf{A\Sigma}_{vv})^T & \mathbf{\Sigma}_{vv} \end{pmatrix}$$

Available Gaussian tools

$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$ → Margin-alize → $\mathbf{U}$

IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$ THEN $\mathbf{U} \sim \mathrm{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_{uu})$

Matrix $\boldsymbol{A}$

$\mathbf{X}$ → Multiply → $\mathbf{AX}$

IF $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ AND $\mathbf{Y} = \mathbf{AX}$ THEN $\mathbf{Y} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$

$\mathbf{X}$, $\mathbf{Y}$ → + → $\mathbf{X} + \mathbf{Y}$

if $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ and $\mathbf{X} \perp \mathbf{Y}$
then $\mathbf{X} + \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$

$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$ → Condition-alize → $\mathbf{U} \mid \mathbf{V}$

IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix} \right)$ THEN $\mathbf{U} \mid \mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$

where $\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$ $\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{uv}$

$\mathbf{U} \mid \mathbf{V}$, $\mathbf{V}$ → Chain Rule → $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$

IF $\mathbf{U} \mid \mathbf{V} \sim \mathrm{N}(\mathbf{AV}, \boldsymbol{\Sigma}_{u|v})$ and $\mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{vv})$

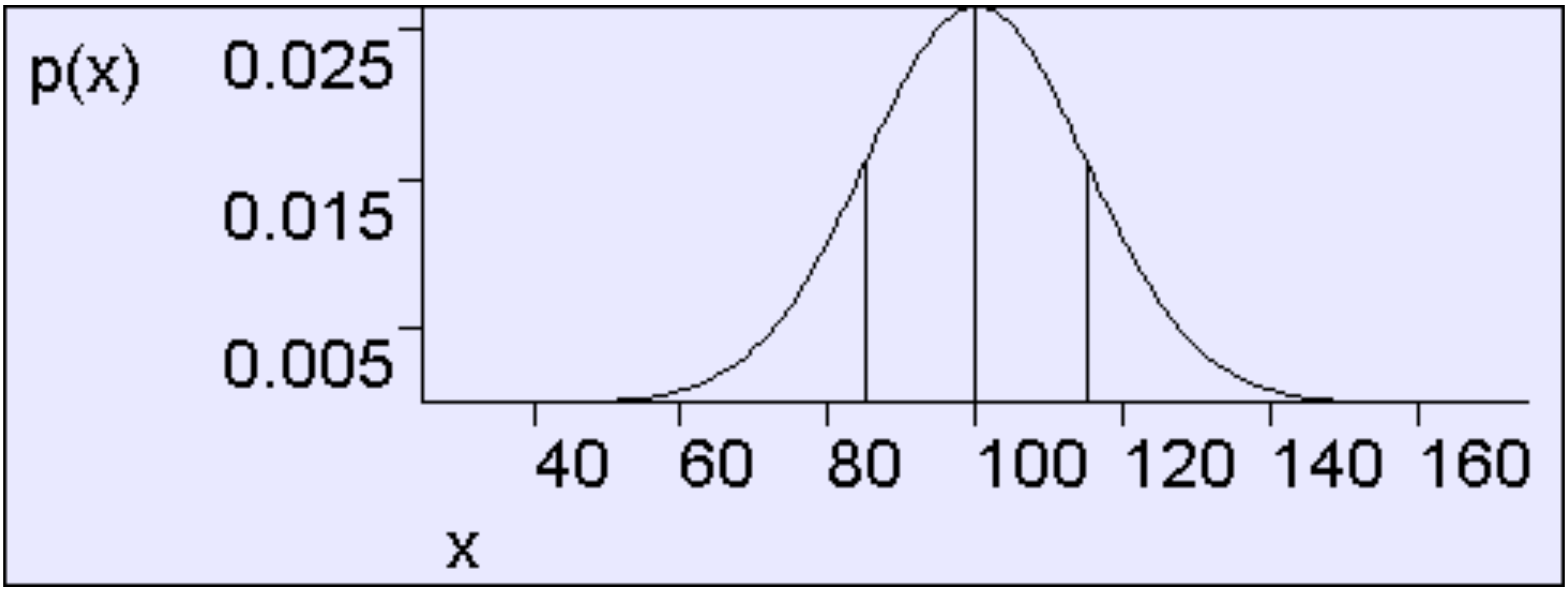
THEN $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{A}\boldsymbol{\Sigma}_{vv}\mathbf{A}^T + \boldsymbol{\Sigma}_{u|v} & \mathbf{A}\boldsymbol{\Sigma}_{vv} \\ (\mathbf{A}\boldsymbol{\Sigma}_{vv})^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}$

Assume…

- You are an intellectual snob
- You have a child

Intellectual snobs with children

- …are obsessed with IQ
- In the world as a whole, IQs are drawn from a Gaussian $N(100,15^2)$



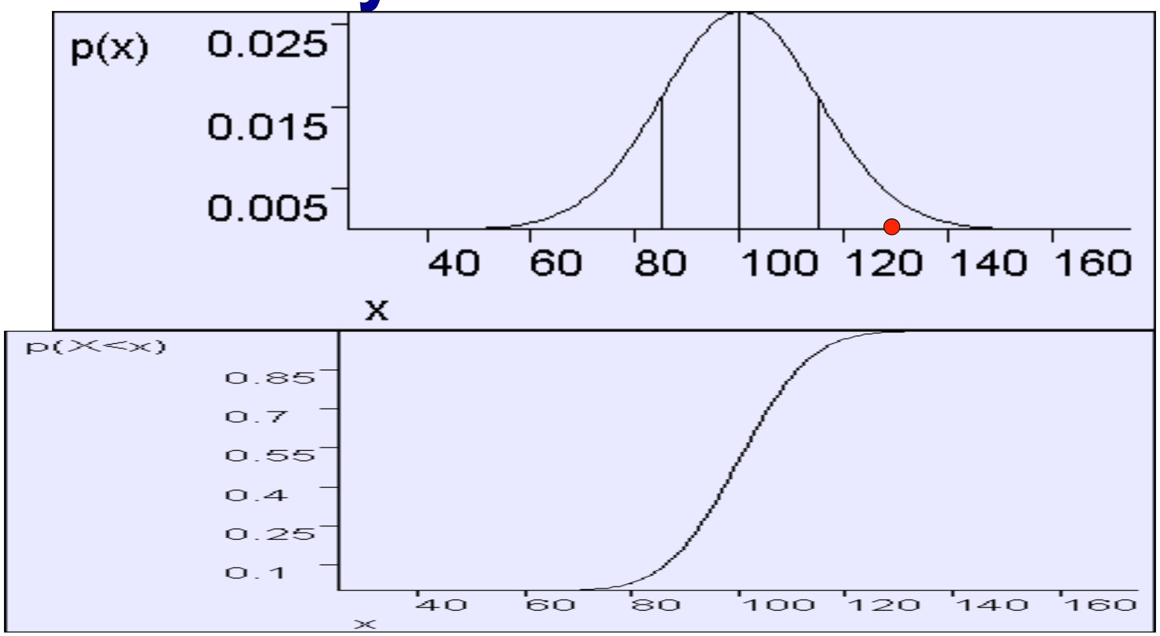
IQ tests

- If you take an IQ test you'll get a score that, on average (over many tests) will be your IQ
- But because of noise on any one test the score will often be a few points lower or higher than your true IQ.

$$SCORE \mid IQ \sim N(IQ, 10^2)$$

Assume…

- You drag your kid off to get tested
- She gets a score of 130
- "Yippee" you screech and start deciding how to casually refer to her membership of the top 2% of IQs in your Christmas newsletter.



$P(X<130|\mu=100,\sigma^2=15^2) =$

$P(X<2|\ \mu=0,\sigma^2=1) =$

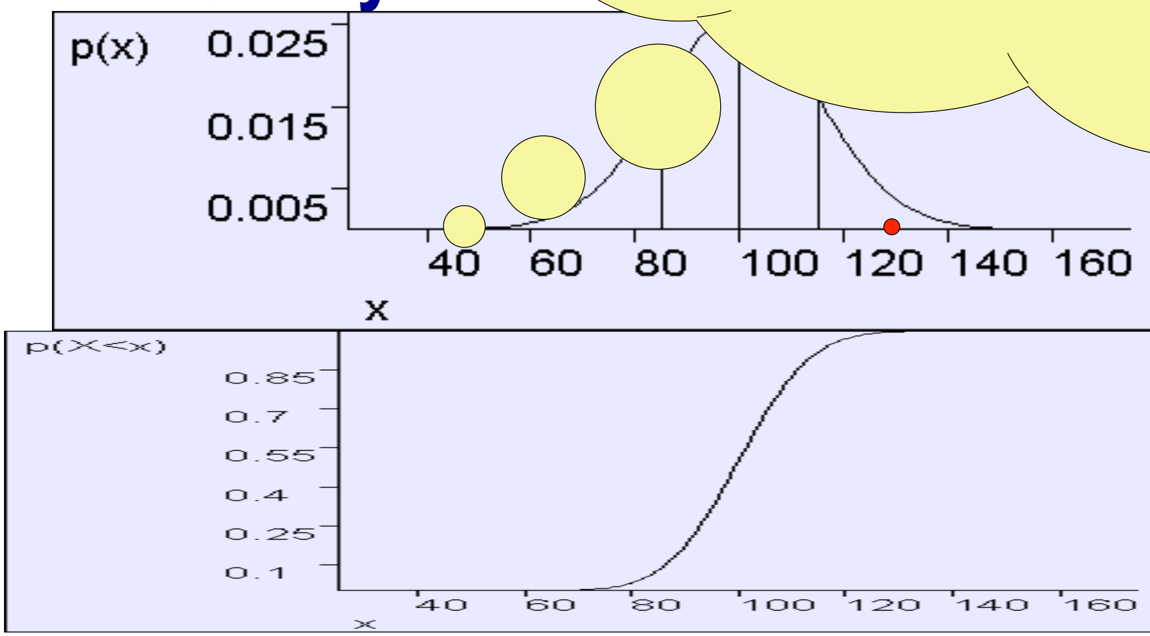
$\text{erf}(2) = 0.977$

Assume…

- You drag your ki
- She gets a sco
- “Yippee” y
  casually refe
  IQs in your C

You are thinking:

Well sure the test isn’t accurate, so she might have an IQ of 120 or she might have an 1Q of 140, but the most likely IQ given the evidence “score=130” is, of course, 130.



P(X<13 |μ=100,σ²=15²) =

P(X< μ=0,σ²=1) =

er) = 0.977

Can we trust this reasoning?

Maximum Likelihood IQ

- $IQ \sim N(100, 15^2)$
- $S|IQ \sim N(IQ, 10^2)$
- $S=130$

$$IQ^{mle} = \arg\max_{iq} p(s = 130 \mid iq)$$

- The MLE is the value of the hidden parameter that makes the observed data most likely
- In this case

$$IQ^{mle} = 130$$

BUT....

- **IQ~N(100,15^2)**
- **S|IQ ~ N(IQ, 10^2)**
- **S=130**

$$IQ^{mle} = \arg\max_{iq} p(s = 130 \mid iq)$$

- **The MLE is the value of the hidden parameter that makes the observed data most likely**
- **In this case**

$$IQ^{mle} = 130$$

This is **not** the same as "The most likely value of the parameter given the observed data"

What we really want:

- $IQ \sim N(100,15^2)$
- $S|IQ \sim N(IQ, 10^2)$
- S=130

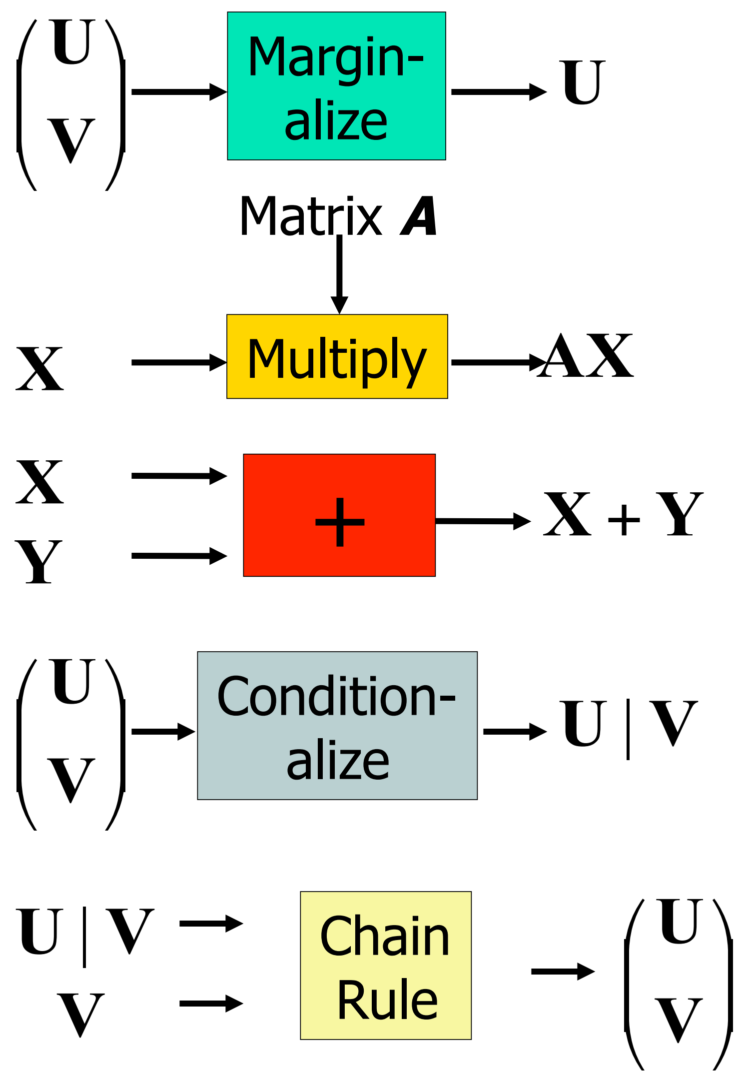
- Question: What is IQ | (S=130)?

Called the Posterior Distribution of IQ

Which tool or tools?

- $IQ \sim N(100, 15^2)$
- $S|IQ \sim N(IQ, 10^2)$
- $S=130$

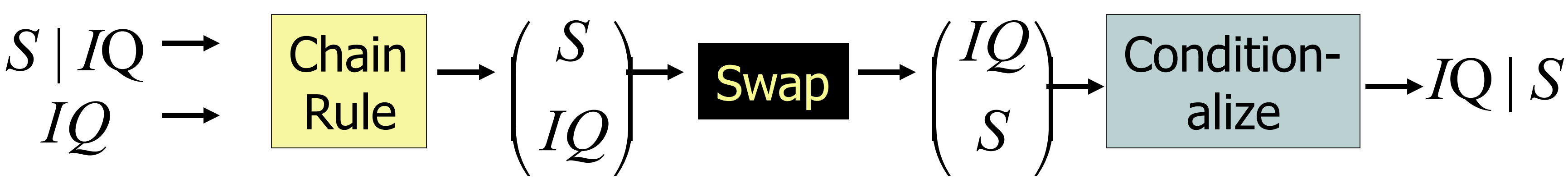
- Question: What is IQ | (S=130)?



Plan

- $IQ \sim N(100,15^2)$
- $S|IQ \sim N(IQ, 10^2)$
- $S=130$

- Question: What is IQ | (S=130)?



Working…

IQ~N(100,15^2)

S|IQ ~ N(IQ, 10^2)

S=130

Question: What is IQ | (S=130)?

IF $\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\mu}_u\\ \boldsymbol{\mu}_v\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv}\\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv}\end{pmatrix}\right)$ THEN

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1}(\mathbf{V} - \boldsymbol{\mu}_v)$$

IF $\mathbf{U} \mid \mathbf{V} \sim \mathrm{N}(\mathbf{AV}, \boldsymbol{\Sigma}_{u|v})$ and $\mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{vv})$

THEN $\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \begin{pmatrix}\mathbf{A}\boldsymbol{\Sigma}_{vv}\mathbf{A}^T + \boldsymbol{\Sigma}_{u|v} & \mathbf{A}\boldsymbol{\Sigma}_{vv}\\ (\mathbf{A}\boldsymbol{\Sigma}_{vv})^T & \boldsymbol{\Sigma}_{vv}\end{pmatrix}$

Your pride and joy's posterior IQ

- If you did the working, you now have p(IQ|S=130)
- If you have to give the most likely IQ given the score you should give

$$IQ^{map} = \arg\max_{iq} p(iq \mid s = 130)$$

- where MAP means "Maximum A-posteriori"

What you should know

- The Gaussian PDF formula off by heart
- Understand the workings of the formula for a Gaussian
- Be able to understand the Gaussian tools described so far
- Have a rough idea of how you could prove them
- Be happy with how you could use them