

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials>. Comments and corrections gratefully received.

Gaussians

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm

awm@cs.cmu.edu

412-268-7599

Copyright © Andrew W. Moore

Slide 1

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

Copyright © Andrew W. Moore

Slide 2

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
- ...

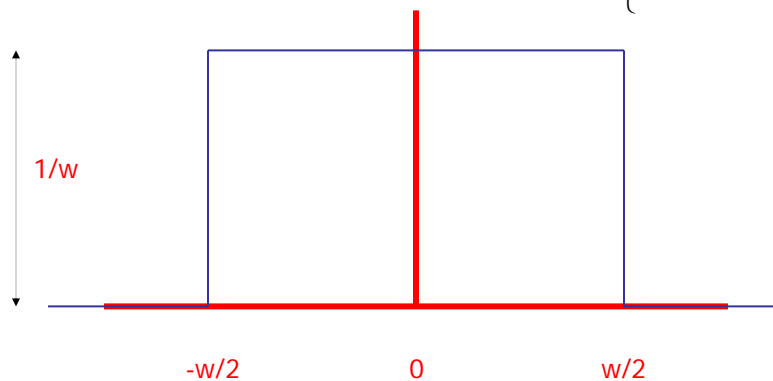
(You get the idea)

Copyright © Andrew W. Moore

Slide 3

The "box" distribution

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

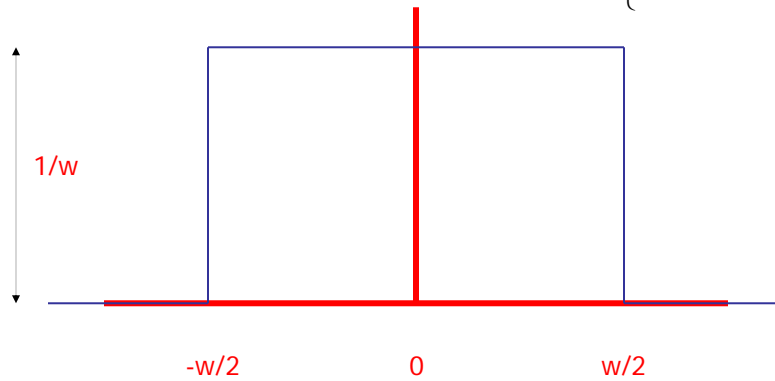


Copyright © Andrew W. Moore

Slide 4

The "box" distribution

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



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

Copyright © Andrew W. Moore

Slide 5

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...

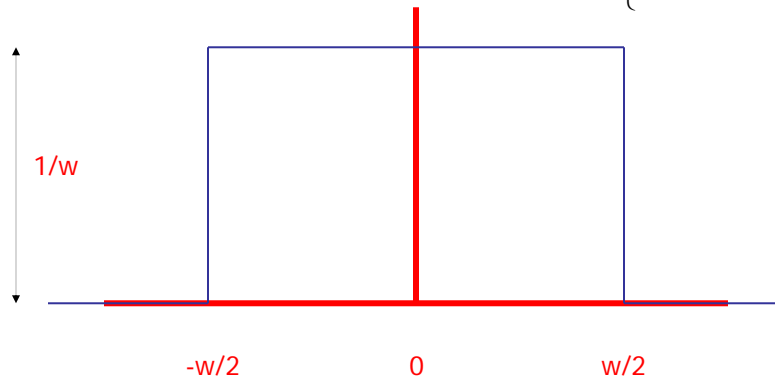
- ...the harder it is to predict
- ...the harder it is to compress it
- ...the less spiky the distribution

Copyright © Andrew W. Moore

Slide 6

The "box" distribution

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



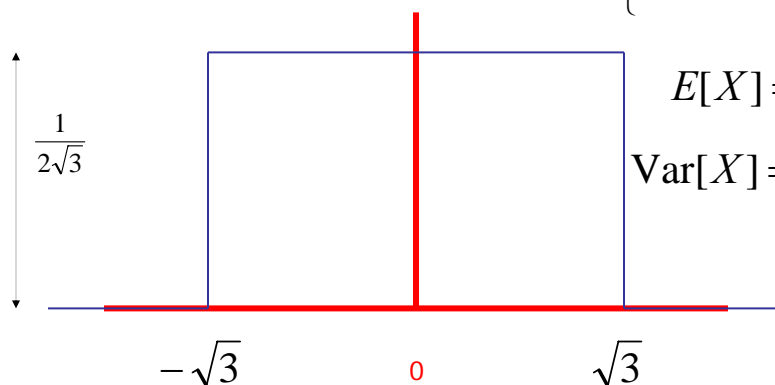
$$H[X] = - \int_{-\infty}^{\infty} p(x) \log p(x) dx = - \int_{-w/2}^{w/2} \frac{1}{w} \log \frac{1}{w} dx = - \frac{1}{w} \log \frac{1}{w} \int_{-w/2}^{w/2} dx = \log w$$

Copyright © Andrew W. Moore

Slide 7

Unit variance box distribution

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



$$E[X] = 0$$

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

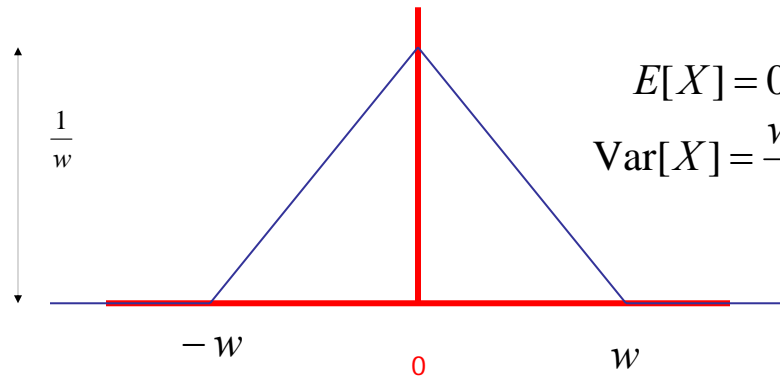
if $w = 2\sqrt{3}$ then $\text{Var}[X] = 1$ and $H[X] = 1.242$

Copyright © Andrew W. Moore

Slide 8

The Hat distribution

$$p(x) = \begin{cases} w - |x| & \text{if } |x| \leq w \\ 0 & \text{if } |x| > w \end{cases}$$



$$E[X] = 0$$

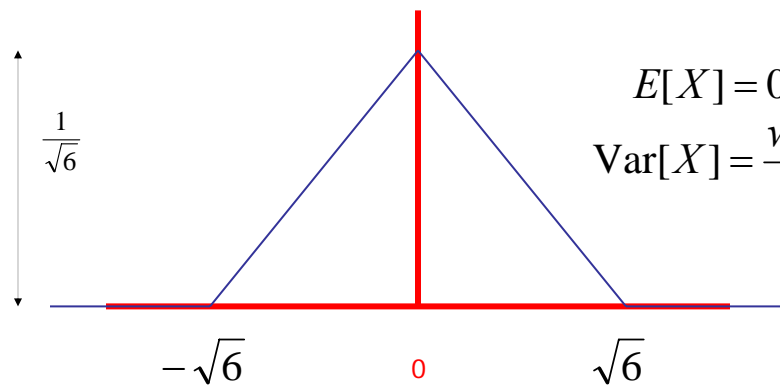
$$\text{Var}[X] = \frac{w^2}{6}$$

Copyright © Andrew W. Moore

Slide 9

Unit variance hat distribution

$$p(x) = \begin{cases} w - |x| & \text{if } |x| \leq w \\ 0 & \text{if } |x| > w \end{cases}$$



$$E[X] = 0$$

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

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

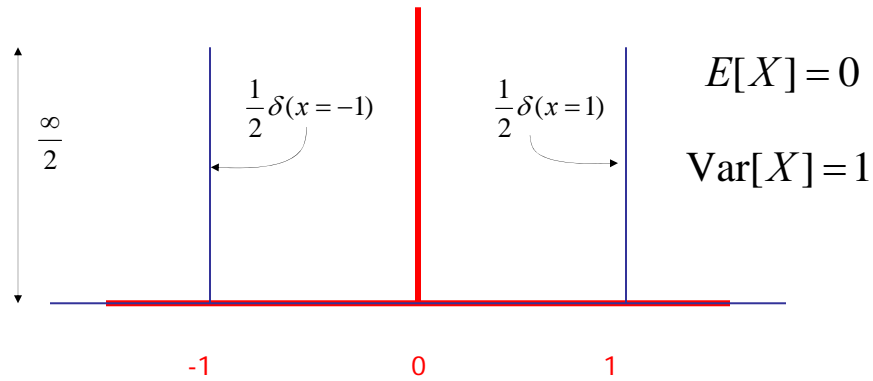
Copyright © Andrew W. Moore

Slide 10

The "2 spikes" distribution

Dirac Delta

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



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

Copyright © Andrew W. Moore

Slide 11

Entropies of unit-variance distributions

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

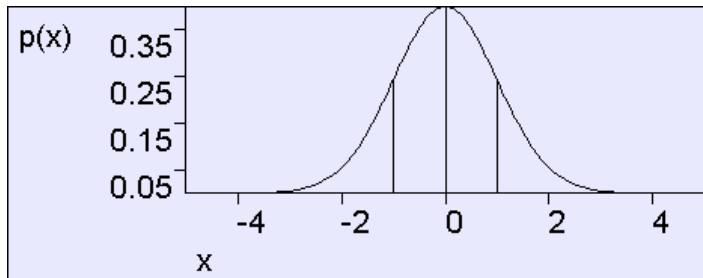
Largest possible entropy of any unit-variance distribution

Copyright © Andrew W. Moore

Slide 12

Unit variance Gaussian

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



$$E[X] = 0$$

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

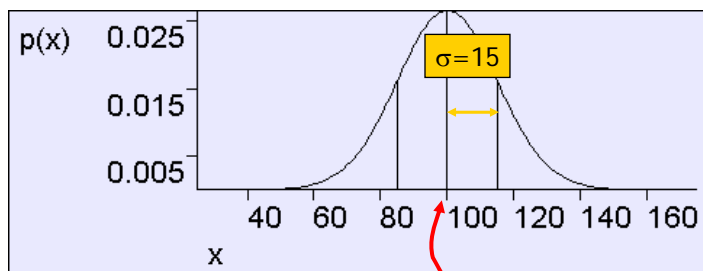
$$H[X] = - \int_{x=-\infty}^{\infty} p(x) \log p(x) dx = 1.4189$$

Copyright © Andrew W. Moore

Slide 13

General Gaussian

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



$$E[X] = \mu$$

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

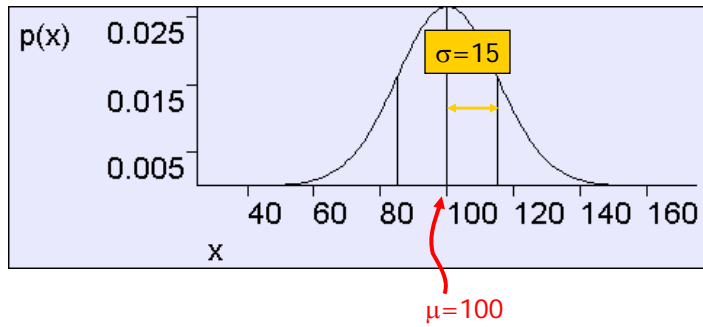
Copyright © Andrew W. Moore

Slide 14

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 $X \sim N(\mu, \sigma^2)$ to mean "X is distributed as a Gaussian with parameters μ and σ^2 ".

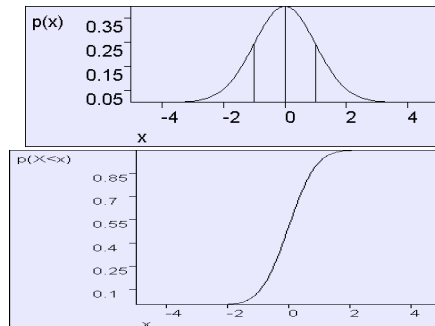
In the above figure, $X \sim N(100, 15^2)$

The Error Function

Assume $X \sim N(0,1)$

Define $\text{ERF}(x) = P(X < x) = \text{Cumulative Distribution of } X$

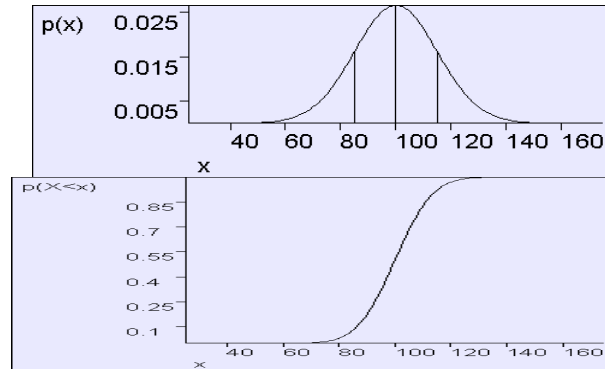
$$\begin{aligned} \text{ERF}(x) &= \int_{z=-\infty}^x p(z) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{z=-\infty}^x \exp\left(-\frac{z^2}{2}\right) dz \end{aligned}$$



Using The Error Function

Assume $X \sim N(\mu, \sigma^2)$

$$P(X < x | \mu, \sigma^2) = \text{ERF}\left(\frac{x - \mu}{\sigma}\right)$$



Copyright © Andrew W. Moore

Slide 17

The Central Limit Theorem

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

Somewhat of a justification for assuming
Gaussian noise is common

Copyright © Andrew W. Moore

Slide 18

Other amazing facts about Gaussians

- Wouldn't you like to know?
- We will not examine them until we need to.

Copyright © Andrew W. Moore

Slide 19

Bivariate Gaussians

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

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

Where the Gaussian's parameters are...

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

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

Copyright © Andrew W. Moore

Slide 20

Bivariate Gaussians

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

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

Where the Gaussian's parameters are...

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

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

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

*This note rates 7.4 on the pedanticness scale

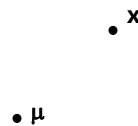
Copyright © Andrew W. Moore

Slide 21

Evaluating $p(\mathbf{x})$: Step 1

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

1. Begin with vector \mathbf{x}



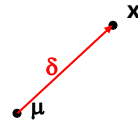
Copyright © Andrew W. Moore

Slide 22

Evaluating $p(\mathbf{x})$: Step 2

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

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



Copyright © Andrew W. Moore

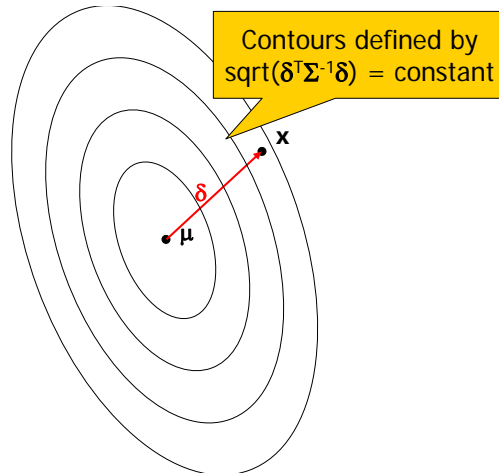
Slide 23

Evaluating $p(\mathbf{x})$: Step 3

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

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

$D = \text{this count} = \text{sqrt}(\boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta})$
 = Mahalanobis Distance
 between \mathbf{x} and $\boldsymbol{\mu}$



Copyright © Andrew W. Moore

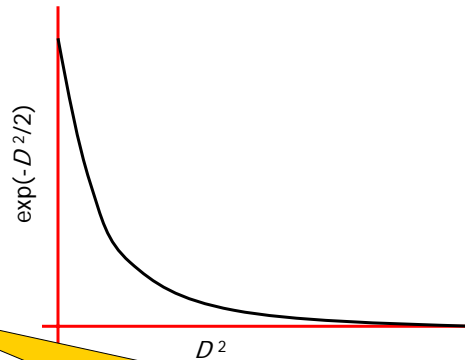
Slide 24

Evaluating $p(\mathbf{x})$: Step 4

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

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

 $D = \text{this count} = \text{sqrt}(\boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta})$
 $= \text{Mahalanobis Distance}$
 between \mathbf{x} and $\boldsymbol{\mu}$
4. Define $w = \exp(-D^2/2)$



\mathbf{x} close to $\boldsymbol{\mu}$ in squared Mahalanobis space gets a large weight. Far away gets a tiny weight

Copyright © Andrew W. Moore

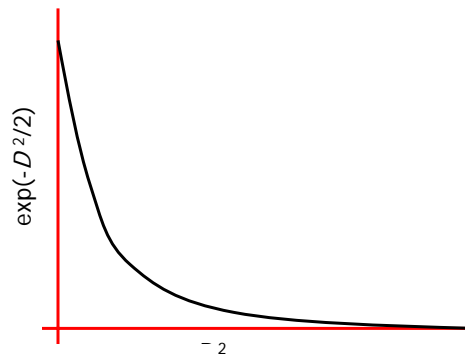
Slide 25

Evaluating $p(\mathbf{x})$: Step 5

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

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

 $D = \text{this count} = \text{sqrt}(\boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta})$
 $= \text{Mahalanobis Distance}$
 between \mathbf{x} and $\boldsymbol{\mu}$
4. Define $w = \exp(-D^2/2)$

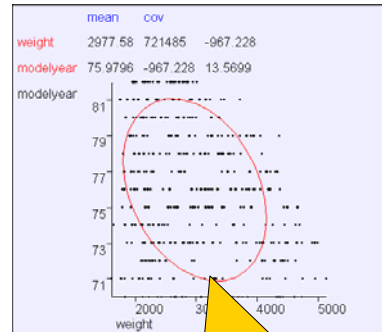
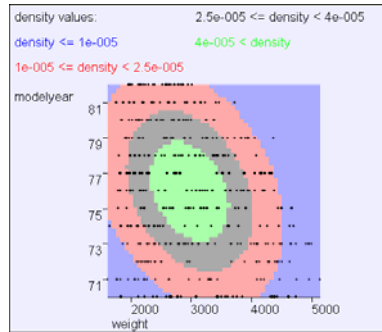


5. Multiply w by $\frac{1}{\sqrt{2\pi} \|\Sigma\|^{1/2}}$ to ensure $\int p(\mathbf{x}) d\mathbf{x} = 1$

Copyright © Andrew W. Moore

Slide 26

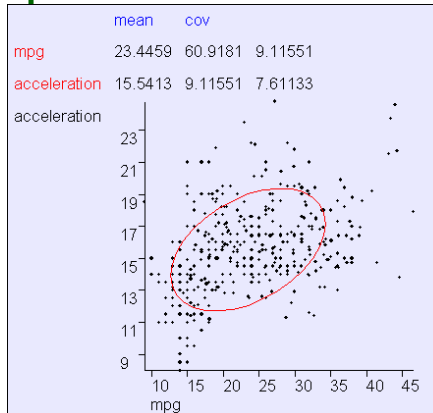
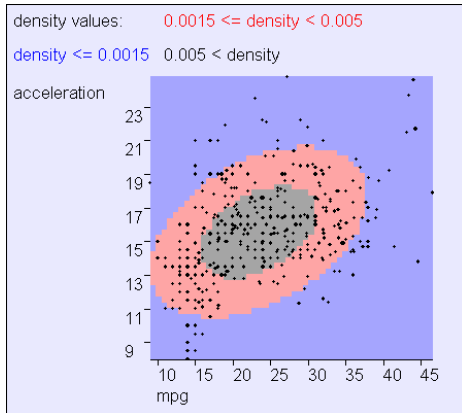
Example



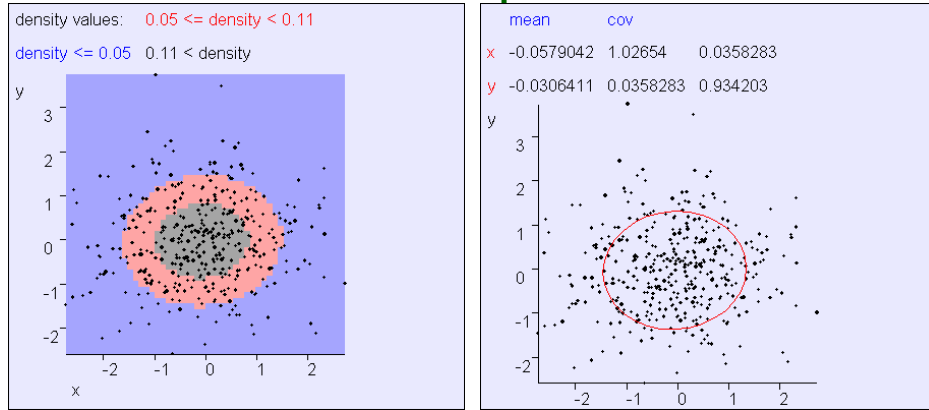
Observe: Mean, Principal axes, implication of off-diagonal covariance term, max gradient zone of $p(x)$

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

Example



Example

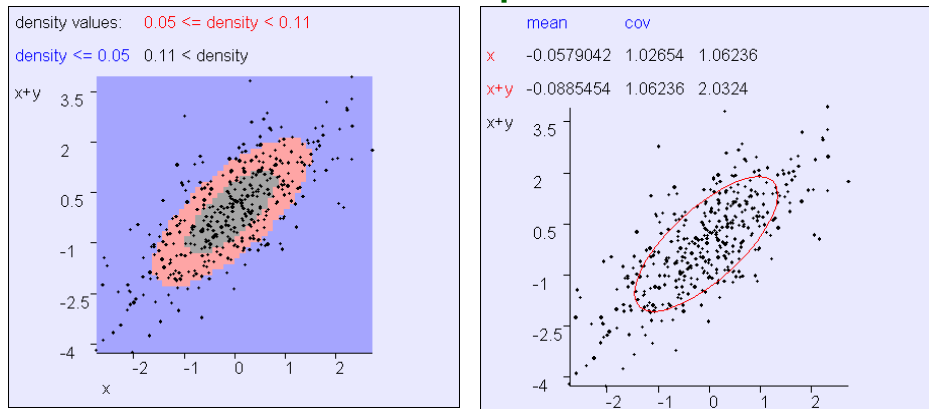


In this example, x and y are almost independent

Copyright © Andrew W. Moore

Slide 29

Example

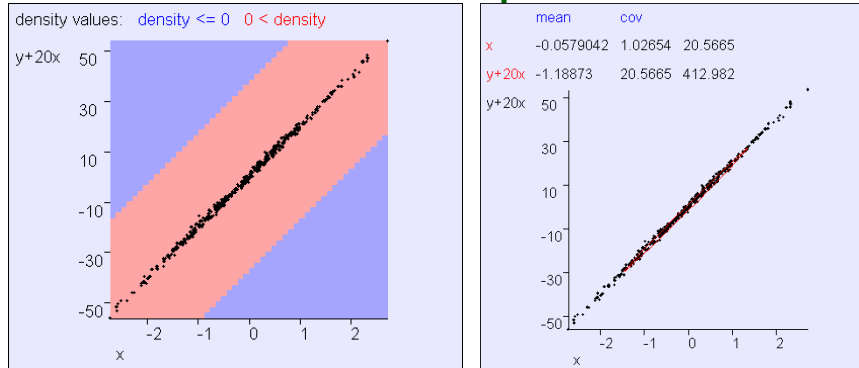


In this example, x and "x+y" are clearly not independent

Copyright © Andrew W. Moore

Slide 30

Example



In this example, x and " $20x+y$ " are clearly not independent

Copyright © Andrew W. Moore

Slide 31

Multivariate Gaussians

Write r.v. $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$ Then define $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mean

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

Where the Gaussian's parameters have...

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_{mm}^2 \end{pmatrix}$$

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

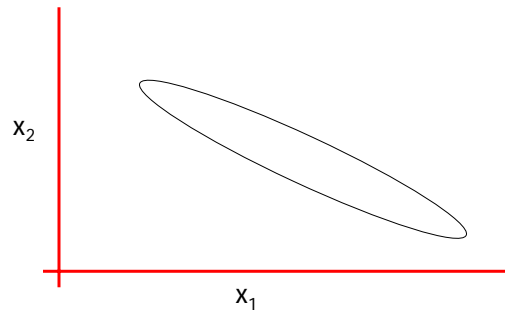
Again, $E[X] = \boldsymbol{\mu}$ and $\text{Cov}[X] = \boldsymbol{\Sigma}$. (Note that this is a resulting property of Gaussians, not a definition)

Copyright © Andrew W. Moore

Slide 32

General Gaussians

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



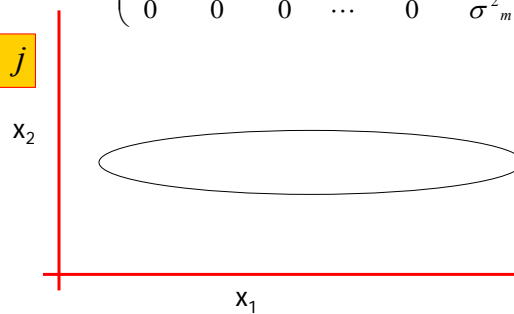
Copyright © Andrew W. Moore

Slide 33

Axis-Aligned Gaussians

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

$X_i \perp X_j$ for $i \neq j$



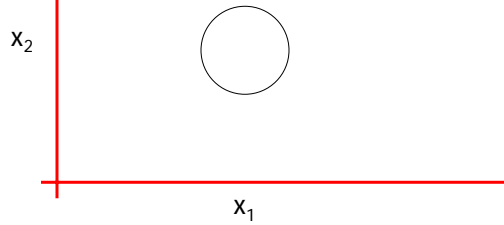
Copyright © Andrew W. Moore

Slide 34

Spherical Gaussians

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

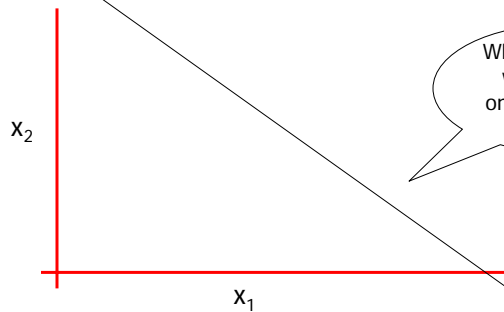
$X_i \perp X_j$ for $i \neq j$



Degenerate Gaussians

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}$$

$$\|\boldsymbol{\Sigma}\| = 0$$



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

Copyright © Andrew W. Moore

Slide 37

Subsets of variables

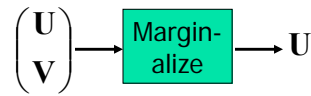
$$\text{Write } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \text{ as } \mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \text{ where } \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ X_{m(u)} \end{pmatrix}$$
$$\mathbf{V} = \begin{pmatrix} X_{m(u)+1} \\ \vdots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

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

Copyright © Andrew W. Moore

Slide 38

Gaussian Marginals are Gaussian



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

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

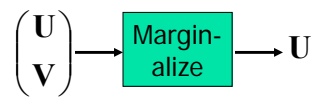
THEN \mathbf{U} is also distributed as a Gaussian

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

Copyright © Andrew W. Moore

Slide 39

Gaussian Marginals are Gaussian



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

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

THEN \mathbf{U} is also distributed as a Gaussian

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

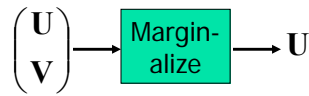
This fact is not immediately obvious

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

Copyright © Andrew W. Moore

Slide 40

Gaussian Marginals are Gaussian



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

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

THEN \mathbf{U} is also distributed as a Gaussian

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

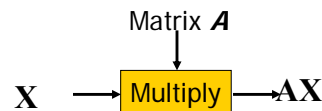
How would you prove this?

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

Copyright © Andrew W. Moore

Slide 41

Linear Transforms remain Gaussian



Assume \mathbf{X} is an m -dimensional Gaussian r.v.

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

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

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

...where \mathbf{A} is a $p \times m$ matrix. Then...

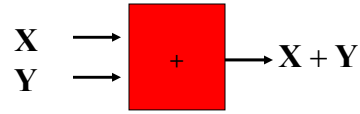
$$\mathbf{Y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

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

Copyright © Andrew W. Moore

Slide 42

Adding samples of 2 independent Gaussians is Gaussian



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

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

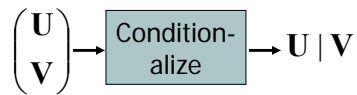
Why doesn't this hold if X and Y are dependent?

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

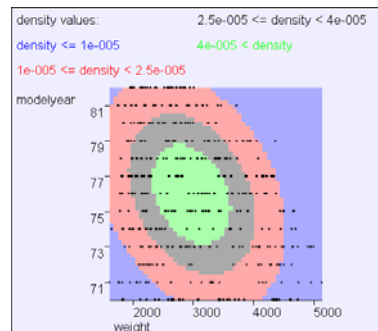


IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}\right)$

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

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

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



$$\text{IF } \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}\right) \quad \text{IF } \begin{pmatrix} w \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 2977 \\ 76 \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ -967 & 3.68^2 \end{pmatrix}\right)$$

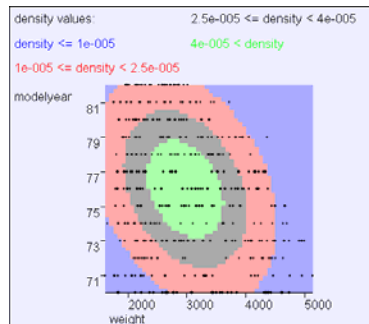
THEN $\mathbf{U} | \mathbf{V} \sim N(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$ where THEN $w | y \sim N(\boldsymbol{\mu}_{w|y}, \boldsymbol{\Sigma}_{w|y})$ where

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

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

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

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



Copyright © Andrew W. Moore

Slide 45

$$\text{IF } \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}\right) \quad \text{IF } \begin{pmatrix} w \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 2977 \\ 76 \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ -967 & 3.68^2 \end{pmatrix}\right)$$

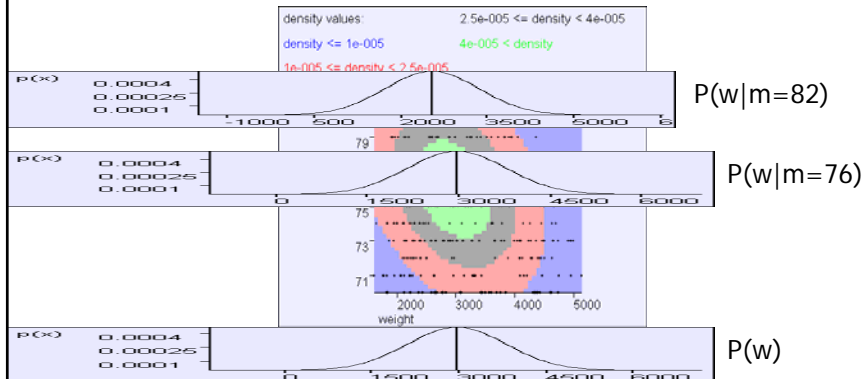
THEN $\mathbf{U} | \mathbf{V} \sim N(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$ where THEN $w | y \sim N(\boldsymbol{\mu}_{w|y}, \boldsymbol{\Sigma}_{w|y})$ where

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

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

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

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



Copyright © Andrew W. Moore

Slide 46

IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}\right)$ IF $\begin{pmatrix} w \\ m \end{pmatrix} \sim N\left(\begin{pmatrix} 2977 \\ 849 \end{pmatrix}, \begin{pmatrix} 849^2 & -967 \\ -967 & 3.68^2 \end{pmatrix}\right)$

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

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

$\boldsymbol{\mu}_{w|m} = 2977 - \frac{3.68^2}{3.68^2} m$

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 is independent of the given value of v

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

Copyright © Andrew W. Moore Slide 47

Gaussians and the chain rule

$\mathbf{U} | \mathbf{V} \rightarrow$ Chain Rule $\rightarrow \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$

$\mathbf{V} \rightarrow$

Let A be a constant matrix

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

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

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

Copyright © Andrew W. Moore Slide 48

Available Gaussian tools

$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$

Margin-
alize

→

\mathbf{U}

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

Matrix \mathbf{A}

↓

\mathbf{X}

Multiply

→

\mathbf{AX}

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

\mathbf{X}

→

+

→

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

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

$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$

Condition-
alize

→

$\mathbf{U} | \mathbf{V}$

IF $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{vv} \end{pmatrix}\right)$ THEN $\mathbf{U} | \mathbf{V} \sim N(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$
 where $\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$ $\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{uv}$

$\mathbf{U} | \mathbf{V}$
 \mathbf{V}

Chain
Rule

→

$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$

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

Copyright © Andrew W. Moore
Slide 49

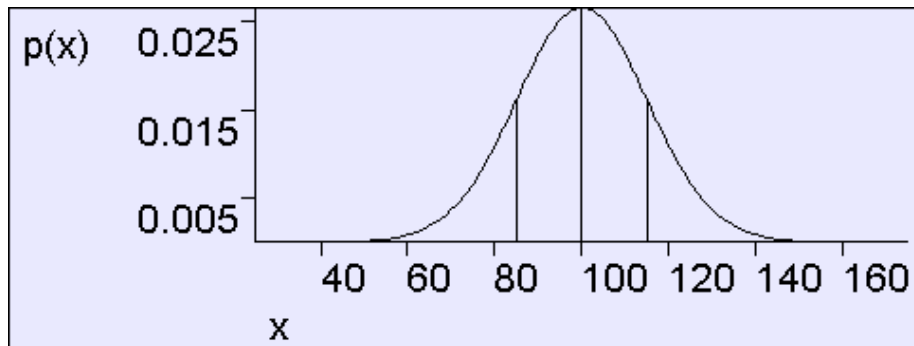
Assume...

- You are an intellectual snob
- You have a child

Copyright © Andrew W. Moore
Slide 50

Intellectual snobs with children

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



Copyright © Andrew W. Moore

Slide 51

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.

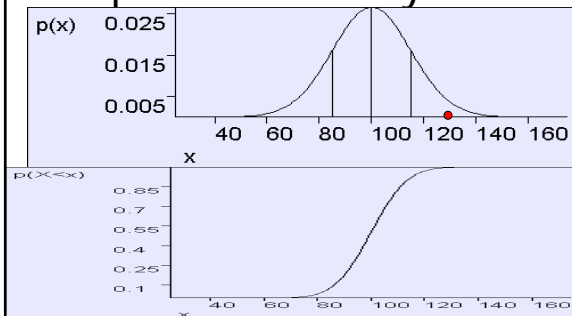
$$\text{SCORE} \mid \text{IQ} \sim N(\text{IQ}, 10^2)$$

Copyright © Andrew W. Moore

Slide 52

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.



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

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

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

Copyright © Andrew W. Moore

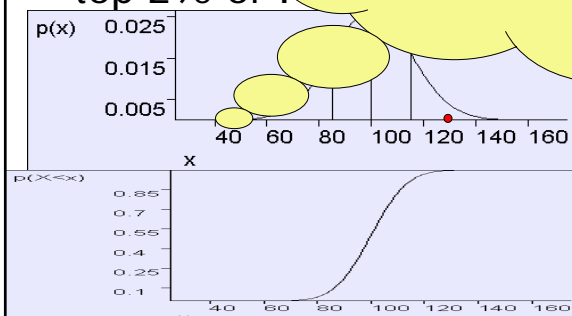
Slide 53

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.

You are thinking:

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



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

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

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

Can we trust this reasoning?

Copyright © Andrew W. Moore

Slide 54

Maximum Likelihood IQ

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

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

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

$$IQ^{mle} = 130$$

Copyright © Andrew W. Moore

Slide 55

BUT....

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

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

- 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"

Copyright © Andrew W. Moore

Slide 56

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

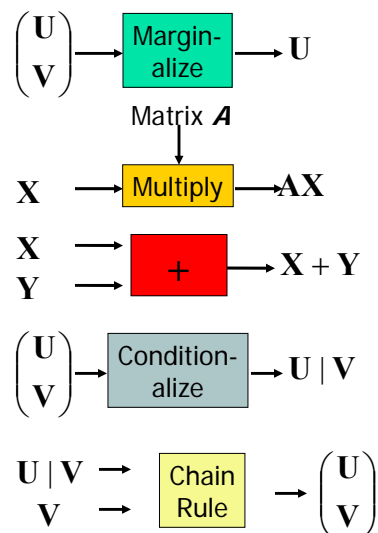
Copyright © Andrew W. Moore

Slide 57

Which tool or tools?

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

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



Copyright © Andrew W. Moore

Slide 58

Plan

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

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



Copyright © Andrew W. Moore

Slide 59

Working...

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

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

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

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

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

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

Copyright © Andrew W. Moore

Slide 60

Your pride and joy's posterior IQ

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

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

- where MAP means "Maximum A-posteriori"

Copyright © Andrew W. Moore

Slide 61

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

Copyright © Andrew W. Moore

Slide 62