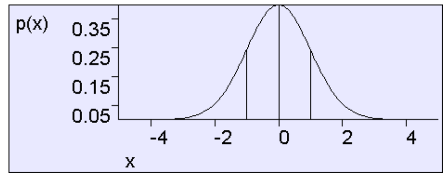
Gaussians

Andrew W. Moore
Associate Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

 Aug 29, 2001

Unit variance Gaussian

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



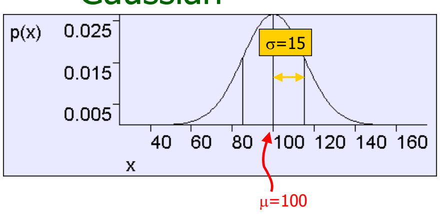
$$E[X] = 0$$

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

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

General Gaussian

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



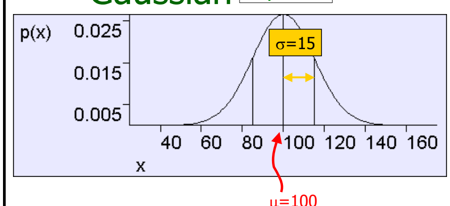
$$E[X] = \mu$$

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

General Gaussian

Also known as the normal distribution or Bell-shaped curve

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



$$E[X] = \mu$$

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

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 Central Limit Theorem

- If (X_1,X_2, … 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

Bivariate Gaussians

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

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

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 Σ is symmetric non-negative definite

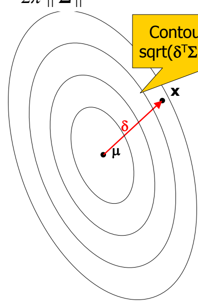
It turns out that E[X] = μ and Cov[X] = Σ. (Note that this is a resulting property of Gaussians, not a definition)*

*This note rates 7.4 on the pedanticness scale

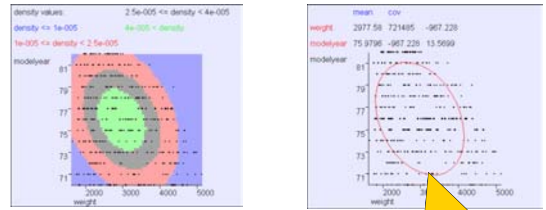
Contour Map

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

Contours defined by sqrt($\delta^T\Sigma^{-1}\delta$) = constant



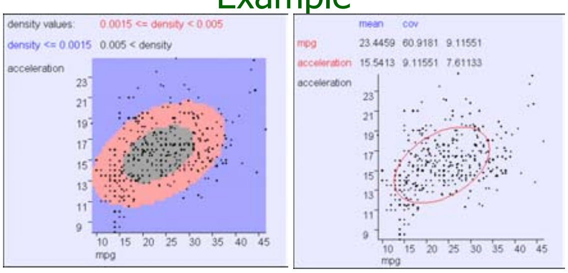
Example



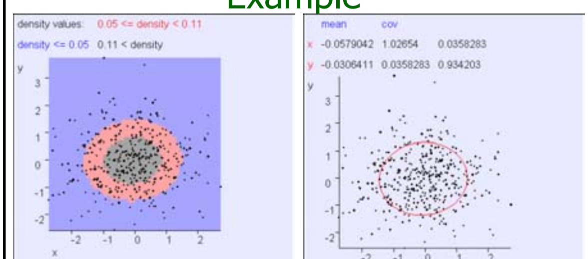
Observe: Mean, Principal axes, implication of off-diagonal covariance term, max gradient zone of p(x)

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

Example

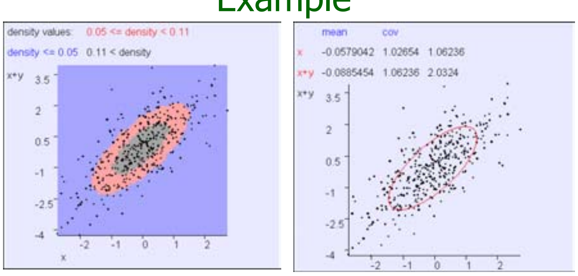


Example



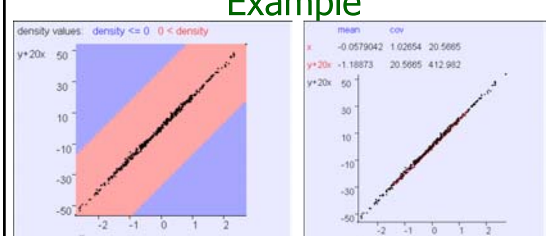
In this example, x and y are almost independent

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

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

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

Where the Gaussian's parameters have...

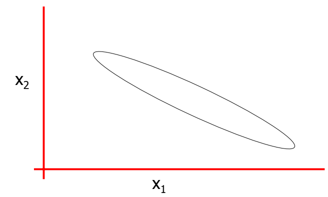
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots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

Where we insist that Σ is symmetric non-negative definite

Again, E[X] = μ and Cov[X] = Σ. (Note that this is a resulting property of Gaussians, not a definition)

General Gaussians

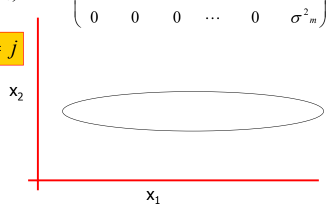
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots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



Axis-Aligned Gaussians

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

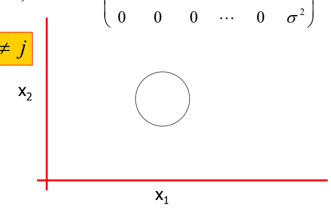
$X_i \perp X_i$ for $i \neq j$



Spherical Gaussians

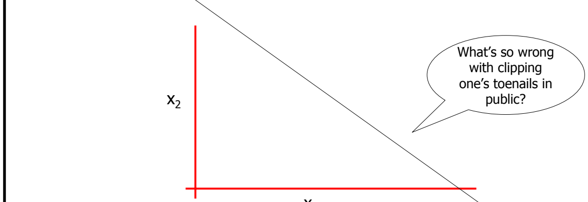
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0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

$X_i \perp X_i$ 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

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

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

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

THEN U is also distributed as a Gaussian

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

Gaussian Marginals are Gaussian

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

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

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

THEN U is also distributed as a Gaussian

This fact is not immediately obvious

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

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

Gaussian Marginals are Gaussian

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

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

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

THEN U is also distributed as a Gaussian

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

How would you prove this?

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

Linear Transforms remain Gaussian

$\mathbf{X}$ → Multiply (Matrix *A*) → $\mathbf{AX}$

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 (note $p \leq m$):

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

…where A is a p x m matrix. Then…

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

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

Adding samples of 2 independent Gaussians is Gaussian

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

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

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

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

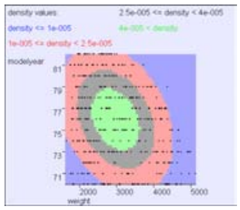
$\begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix} \rightarrow$ Condition-alize $\rightarrow \mathbf{U}\mid\mathbf{V}$

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

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

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



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

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

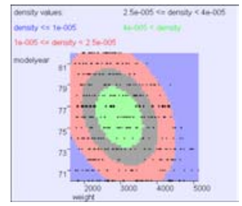
$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{uv}$

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

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

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



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

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

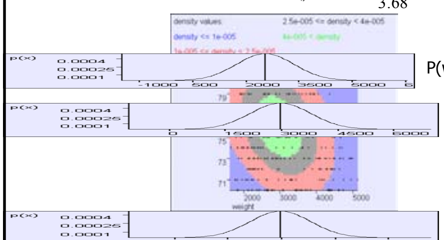
$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{uv}$

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

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

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



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

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

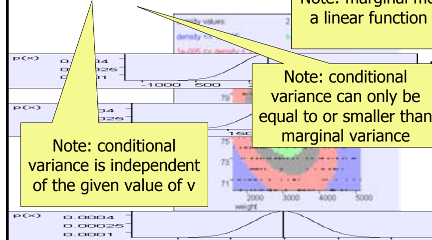
$\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{uv}$

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

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

Let A be a constant matrix

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

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

Available Gaussian tools

$\begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix} \rightarrow$ Margin-alize $\rightarrow \mathbf{U}$ — 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 ***A***

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

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

$\begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix} \rightarrow$ Condition-alize $\rightarrow \mathbf{U}\mid\mathbf{V}$ — 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 $\mathbf{U}\mid\mathbf{V} \sim N(\boldsymbol{\mu}_{u|v}, \boldsymbol{\Sigma}_{u|v})$ where $\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{vv}^{-1}(\mathbf{V}-\boldsymbol{\mu}_v)$ $\boldsymbol{\Sigma}_{u|v} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{uv}$

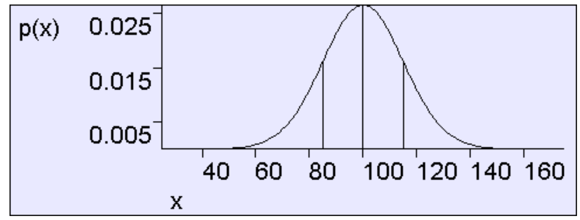
$\mathbf{U}\mid\mathbf{V} \rightarrow$, $\mathbf{V} \rightarrow$ Chain Rule $\rightarrow \begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix}$ — IF $\mathbf{U}\mid\mathbf{V} \sim N(\mathbf{AV}, \boldsymbol{\Sigma}_{u|v})$ and $\mathbf{V} \sim N(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{vv})$ THEN $\begin{pmatrix}\mathbf{U}\\\mathbf{V}\end{pmatrix}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

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 $N(100,15^2)$



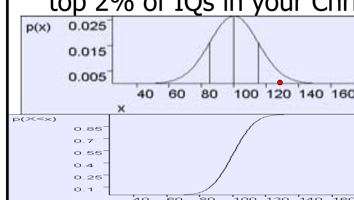
IQ tests

- 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.



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

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

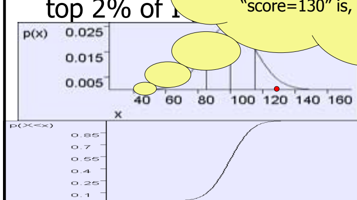
$erf(2) = 0.977$

Assume…

- You drag your k…
- She gets a sc…
- "Yippee" y… to casually … top 2% of I… …ter.

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.



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

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

$erf(2) = 0.977$

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(100,15²)
- S|IQ ~ N(IQ, 10²)
- 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~N(100,15²)
- S|IQ ~ N(IQ, 10²)
- S=130

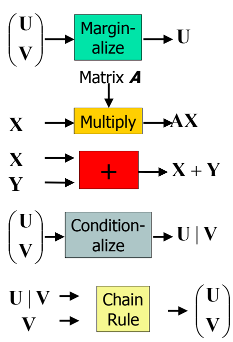
- Question: What is IQ | (S=130)?

Called the Posterior Distribution of IQ

Which tool or tools?

- IQ~N(100,15²)
- S|IQ ~ N(IQ, 10²)
- S=130

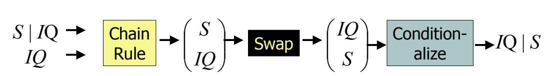
- Question: What is IQ | (S=130)?



Plan

- IQ~N(100,15²)
- S|IQ ~ N(IQ, 10²)
- S=130

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



Working…

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

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

IQ~N(100,15²)
S|IQ ~ N(IQ, 10²)
S=130

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

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

$$\boldsymbol{\mu}_{IQ|S} = \boldsymbol{\mu}_{IQ} + \boldsymbol{\Sigma}_{IQ}(\boldsymbol{\Sigma}_{IQ} + \boldsymbol{\Sigma}_{S|IQ})^{-1}(S - \boldsymbol{\mu}_{IQ})$$

$$\boldsymbol{\Sigma}_{IQ|S} = \boldsymbol{\Sigma}_{IQ} + \boldsymbol{\Sigma}_{IQ}(\boldsymbol{\Sigma}_{IQ} + \boldsymbol{\Sigma}_{S|IQ})^{-1}\boldsymbol{\Sigma}_{IQ}$$

What happens when $\Sigma_{S|IQ} = 0$? $\Sigma_{S|IQ} = \infty$?

That was an important result! It explains how to combine noisy measurements (sensor fusion) So I will do it again in 1D

Combining Measurements: 1D

- True value x
- Measurements m_1, m2: $E(m_1-x) = 0$, $Var(m_1) = \sigma_1^2$, $E(m_2-x) = 0$, $Var(m_2) = \sigma_2^2$, independent
- Linear estimate $\hat{x} = k_1 m_1 + k_2 m_2$
- Unbiased estimate means $k_2 = 1 - k_1$ so $E(\hat{x}) = x$
- Minimize $Var(\hat{x}) = k_1^2 \sigma_1^2 + (1 - k_1)^2 \sigma_2^2$
- So $\partial Var(\hat{x})/\partial k_1 = 0$ -> $2k_1(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 = 0$
- So $k_1 = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$, $k_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$
- So $Var(\hat{x}) = \sigma_1^2 \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$
- What happens when $\sigma_2^2 = 0$? $\sigma_2^2 =$ infinity?
- BLUE: Best Linear Unbiased Estimator

Your pride and joy's posterior IQ

- If you did the working, you now have p(IQ|S=130)
- This is a density, not a number!
- If you have to give the most likely IQ given the score you should give

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

- This is the mean for a Gaussian
- 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
- Understand the Bayesian approach to combining information