HTF: 1,2,…
DHS: 1
RN: Ch 13

Foundations… Background

(Cmput 466 / 551)

R Greiner
Department of Computing Science
University of Alberta

Foundations

- Notation
- Derivatives
- Vectors
- Probability
  - Variables, Events
  - Conditional Probability
  - Max Likelihood Estimates
  - Gaussians

Notation

- $\forall x \ \ldots$ means "for all x, …"
  - $\forall x \ \ x + 1 = 1 + x$ is true
  - $\forall x \ \ x = 0$ is false
- $\exists y \ \ldots$ means "there exists a y such that …"
  - $\exists y \ y^2 = 1$ is true (even though >1 such y)
  - $\exists y \ y + 1 = y$ is false
- $\underset{z}{\operatorname{argmin}} f(z)$ is the value of z that min's $f(.)$
  - $\underset{z}{\operatorname{argmin}} (z - 7)^2$ is 7 (not 0)

Derivative, Minimum

- Given function $f(x)$,

  derivative is written: $f'(x)$ or $\frac{\partial f}{\partial x}$
- To find (local) minimum of $f(.)$,

  $\min_x f(x)$
  - Find $x^* = \operatorname{argmin}_x f(x)$

    by solving $\frac{\partial f}{\partial x} = 0$
  - (Check 2nd derivative…)
  - Return $f(x^*)$

Vectors

Vectors are COLUMN vectors

- $A = \boxed{\begin{matrix} 3 \\ 4 \\ 2 \\ 13 \\ 9 \end{matrix}}$

- $A^T = \boxed{\begin{matrix} 3 & 4 & 2 & 13 & 9 \end{matrix}}$
  (transpose)

- Dot product: $\langle a, b \rangle = a^T b = \sum_i a_i b_i$
  - projection, …

Terms from Probability Theory

- **Random Variable**:
  Weather $\in$ { Sunny, Rain, Cloudy, Snow }
- **Domain**: Possible values a random variable can take.
  (... finite set, $\Re$, functions…)
- Probability distribution (discrete):
  mapping from domain to values $\in$ [0, 1]
- P(Weather) = $\langle$ 0.7, 0.2, 0.08, 0.02 $\rangle$

  means
$$\left\{ \begin{array}{l} P(\text{ Weather} = \text{Sunny }) = 0.7 \\ P(\text{ Weather} = \text{Rain }) = 0.2 \\ P(\text{ Weather} = \text{Cloudy }) = 0.08 \\ P(\text{ Weather} = \text{Snow }) = 0.02 \end{array} \right\}$$

- Event:
  Each assignment (eg, Weather = Rain) is "event"

General Events

- **Atomic Event**: "Complete specification"
  Conjunction of assignments to EVERY variable [PossibleWorld]
- **Joint Probability Distribution**:
  Probability of every possible atomic event

n binary variables: 2^n entries
 ($2^n - 1$ independent values, as sum = 1)
A huge table!

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

Marginalization

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

- "marginal"

$$P(X_n) = \sum_{x_1,\ldots,x_{n-1}} P(x_1,\ldots,x_{n-1},X_n)$$

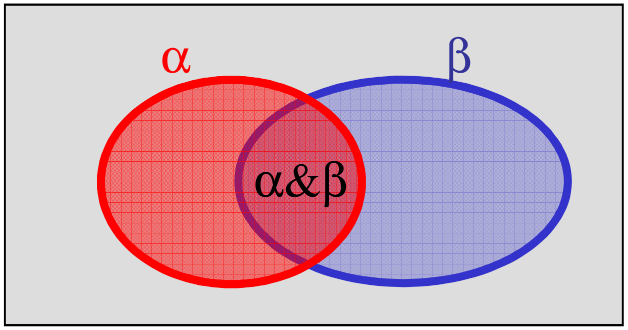
- To compute marginal distribution $P(X_n)$:
  If all binary, 2^{n-1} additions
  - one value for each assignment to $x_1, \ldots, x_{n-1}$
  - One for [0, …, 0, 0], another for [0, …, 0, 1], [0, …, 1, 0], …, [1, …, 1, 1]

Conditional Probabilities

- After learning that β is true,
  how do we feel about α?
- If roll is EVEN, what is chance of rolling 2?
- If have hepatitis, what is chance of jaundice?

β α

P(α | β)



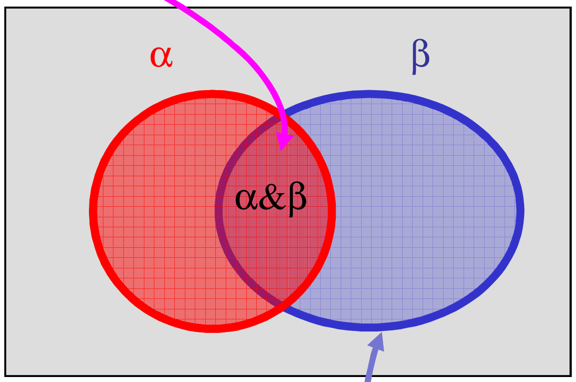
Conditional Probability

- Conditional Probability:
  P(α | β) = Probability of event α,
  given that event β has happened
- P(Jaundice | Hepatitis) = 0.8
  even if P(Jaundice) = 0.01

- In gen'l:

$$P(\alpha \mid \beta) = \frac{P(\alpha \,\&\, \beta)}{P(\beta)}$$

$$P(\alpha \,\&\, \beta) = P(\alpha \mid \beta)\, P(\beta)$$



Independence

- Events α and β are independent *iff*
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\alpha \vee \beta) = 1 - (1 - P(\alpha))\,(1 - P(\beta))$

- Variables independent

  $\Leftrightarrow$ independent for all values

$\forall a, b \quad P(A = a, B = b) = P(A = a)\; P(B = b)$

i.i.d = "independent and identically distributed"

Binomial Distribution

- Model:
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to distribution
  - P(Head) = θ, P(Tail) = $1-\theta$

- $P(H,H,T,T,H) = P(H)\,P(H)\,P(T)\,P(T)\,P(H)$

$$= \theta \quad \theta \quad (1-\theta)\,(1-\theta) \quad \theta$$

$$= \theta^3\,(1-\theta)^2$$

- Sequence D of $\#H$ Heads and $\#T$ Tails:

$$\boxed{P(D \mid \theta) = \theta^{\#H}\,(1-\theta)^{\#T}}$$

Maximum Likelihood Estimation

- **Data:** Observed set *D* of
  #H Heads and #T Tails
- **Hypothesis Space:** Binomial distributions

- Learning “best” θ is an *optimization problem*
  - What’s the objective function?

- **MLE**: Choose θ that maximizes
  the probability of observed data:

$$\hat{\theta} = \arg\max_{\theta} \; P(D|\theta)$$

$$= \arg\max_{\theta} \; \ln P(D \mid \theta)$$

Simple "Learning" Algorithm

$$\hat{\theta} = \arg\max_{\theta} \ \ln P(D \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{h} (1 - \theta)^{t}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\frac{\partial}{\partial\theta}\ln[\theta^{h}(1-\theta)^{t}] = \frac{\partial}{\partial\theta}[h\ln\theta + t\ln(1-\theta)] = \frac{h}{\theta} + \frac{-t}{(1-\theta)}$$

$$\frac{h}{\theta} + \frac{-t}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{\mathrm{h}}{\mathrm{t+h}}$$

So just average!!

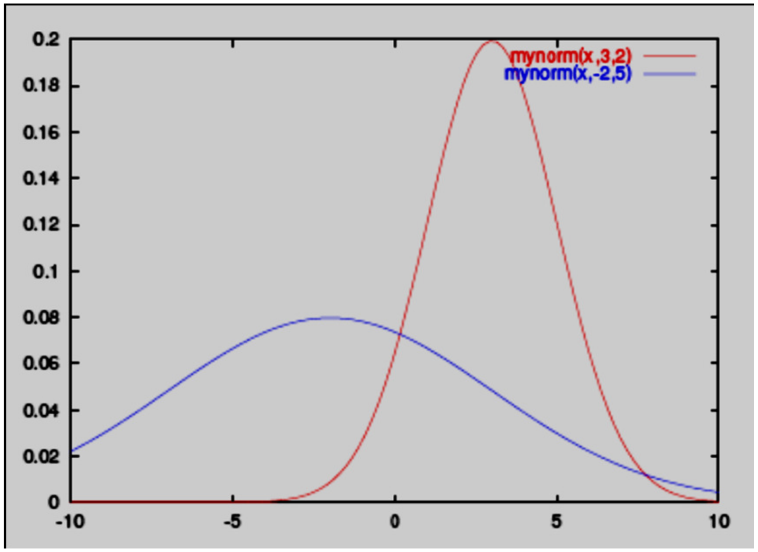
Univariate Gaussian Distributions

- Univariate normal (Gaussian): $N(\mu, \sigma^2)$
  - Mean $\mu = E[x]$

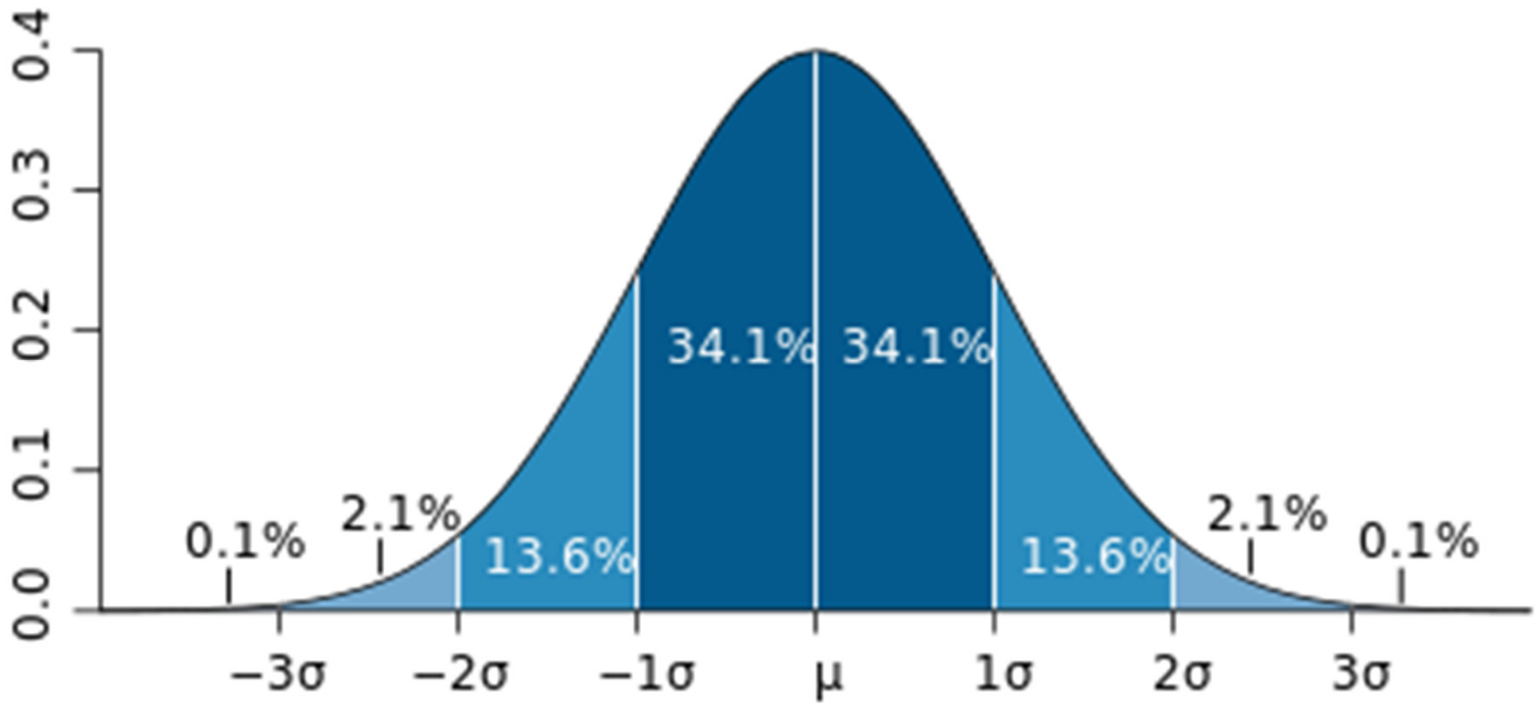
  $= \int x\, p(x;\mu,\ \sigma^2)\, dx \quad \approx \frac{1}{N} \sum_i x_i$
  - Variance $\sigma^2 = E[\,(x-\mu)^2]$

  $= \int (x-\mu)^2 p(x;\mu,\ \sigma^2) dx \quad \approx \frac{1}{N} \sum_i (x_i - \mu)^2$
- PDF (probability distribution fn)

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



Area under the Curve



Useful Properties of Gaussians

- Lots of things can (arguably) be nicely approximated by Gaussians

- Central Limit Theorem:

The sum of IID variables with finite variances will tend towards a Gaussian distribution

- CLT often used as a hand-waving argument to justify using the Gaussian distribution for almost anything

Central Limit Theorem

- Let $X_1, X_2, \ldots$ be an infinite sequence of independent random variables with
  - $E[\, X_i \,] = \mu, \quad E(X_i - \mu)^2 = \sigma^2$
- Define $Z_n = \frac{((X_1 + \ldots + X_n) - n\,\mu)}{\sigma\, n^{\frac{1}{2}}}$
- Then, as $n \to \infty$, Z_n is distributed as $N(0,1)$
- $\approx$ quantities that are the sum of many small effects, tend to become Gaussian

Some Properties of Gaussians

- Affine transformation
  (multiplying by scalar and adding a constant)
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \quad \Rightarrow \quad Y \sim N(a\mu+b,\ a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X, \sigma_X^2)$
  - $Y \sim N(\mu_Y, \sigma_Y^2)$
  - $Z = X+Y \quad \Rightarrow \quad Z \sim N(\mu_X+\mu_Y,\ \sigma_X^2+\sigma_Y^2)$

What if X, Y are independent? Dependent? …

MVG = MultiVariate Gaussian
= Gaussian over many variables…

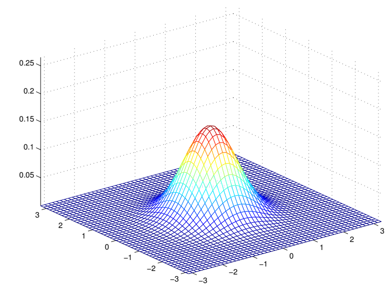
The Multivariate Gaussian

- A 2-dimensional Gaussian is defined by
  - a mean vector $\mu = [\ \mu_1, \mu_2\ ]$
  - a covariance matrix: $\Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{2,1}^2 \\ \sigma_{1,2}^2 & \sigma_{2,2}^2 \end{bmatrix}$
    - $\sigma_{i,j}^2 = E[\ (x_i - \mu_i)\ (x_j - \mu_j)\ ]$ is (co)variance
- Write $\sigma_{i,i}^2$ as σ_i^2 (variance)
- Note Σ is :
  - symmetric
  - "positive semi-definite": $\forall x: \ x^T \Sigma x \geq 0$

MVG = MultiVariate Gaussian
= Gaussian over many variables…

The Multivariate Gaussian

- A 2-dimensional Gaussian is defined by
  - a mean vector $\mu = [\ \mu_1, \mu_2\ ]$
  - a covariance matrix: $\Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{2,1}^2 \\ \sigma_{1,2}^2 & \sigma_{2,2}^2 \end{bmatrix}$
    - $\sigma_{i,j}^2 = E[\ (x_i - \mu_i)\ (x_j - \mu_j)\ ]$ is (co)variance



- PDF:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\ |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

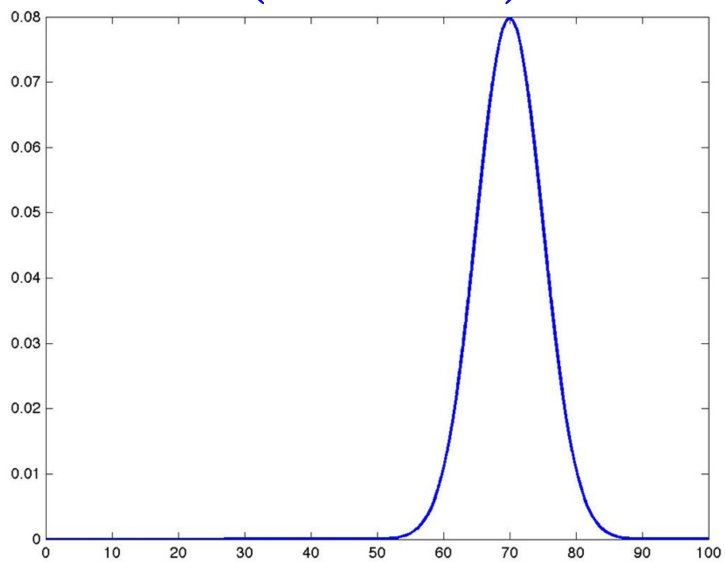
Compare: for n=1:

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

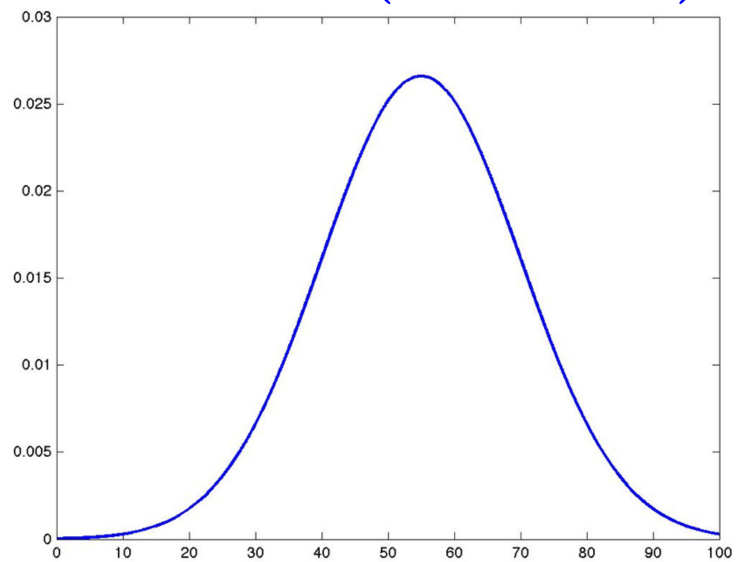
Example: Marks on HW1 & HW2

- Class of m students… 2 HWs…

HW1 ~ $\mathcal{N}(70, 5^2)$



HW2 ~ $\mathcal{N}(55, 15^2)$

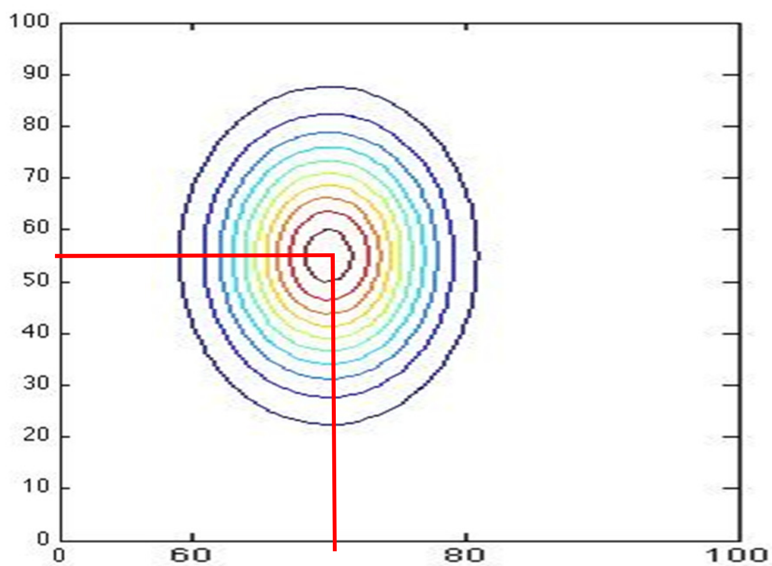


What if these student grades are RELATED to each other?

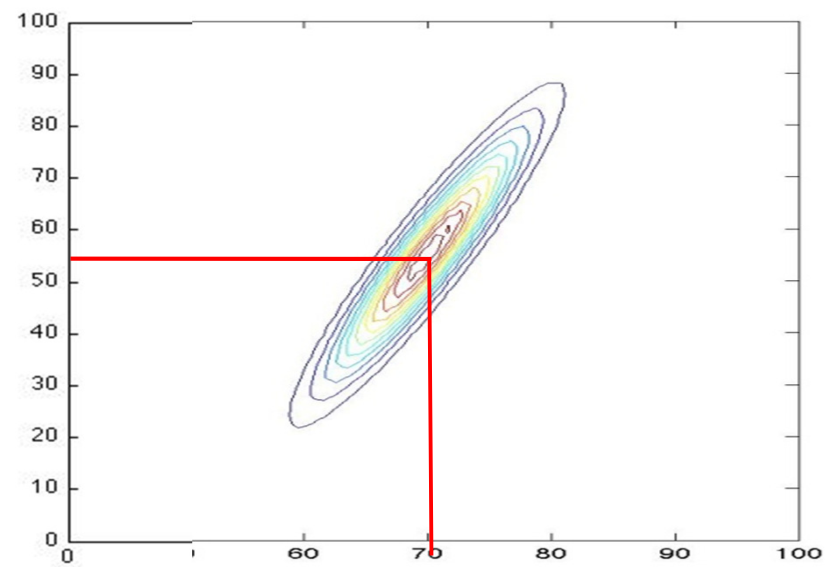
How to model relationship amongst these variables?

Joint Distribution (HW1 & HW2)

Contour plot



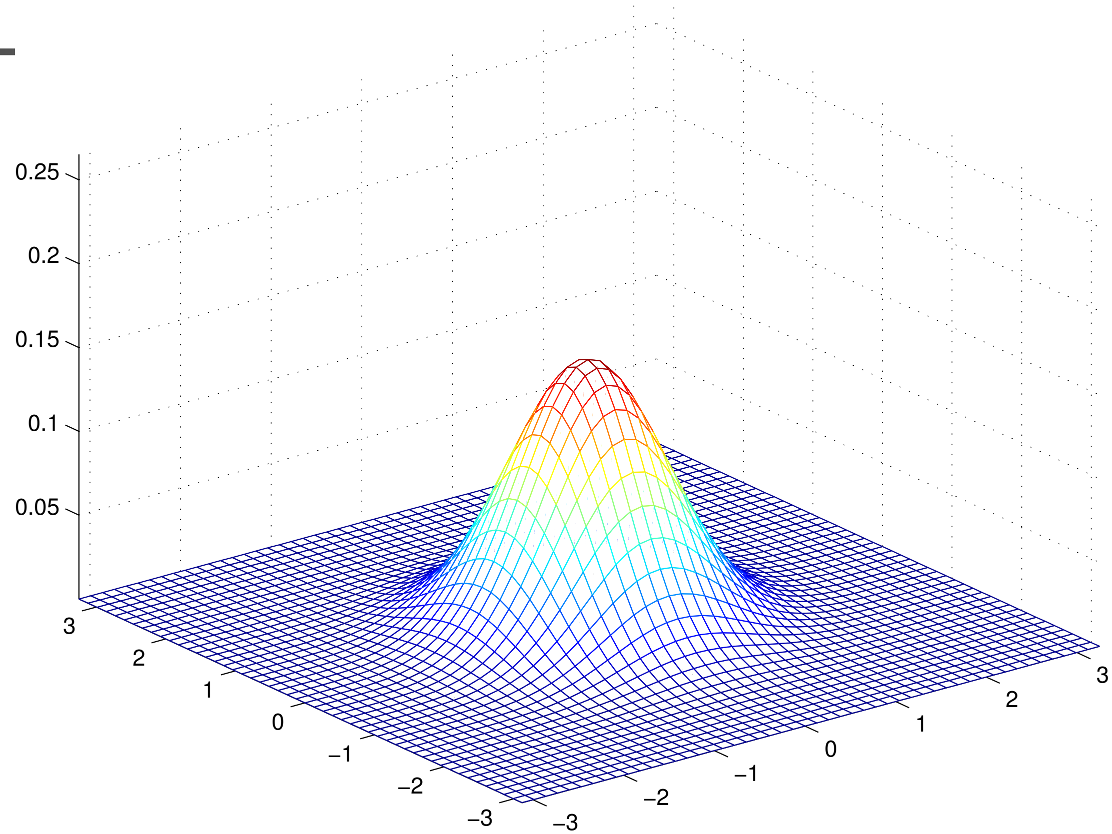
Contour plot



- [x,y] s.t. p([x,y]) = constant

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

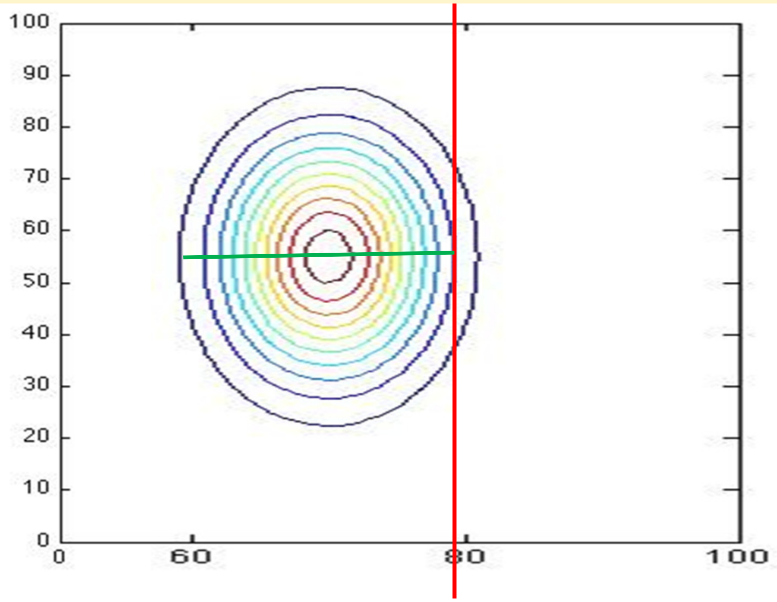
Standard Normal Distribution



- Standard normal for
  - $\mu = (0,0)$
  - Σ = the identity matrix = $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Joint Distribution (HW1 & HW2)

What is mean, variance of HW#2, for values of HW#1 :
P(HW#2 | HW#1=80) ?



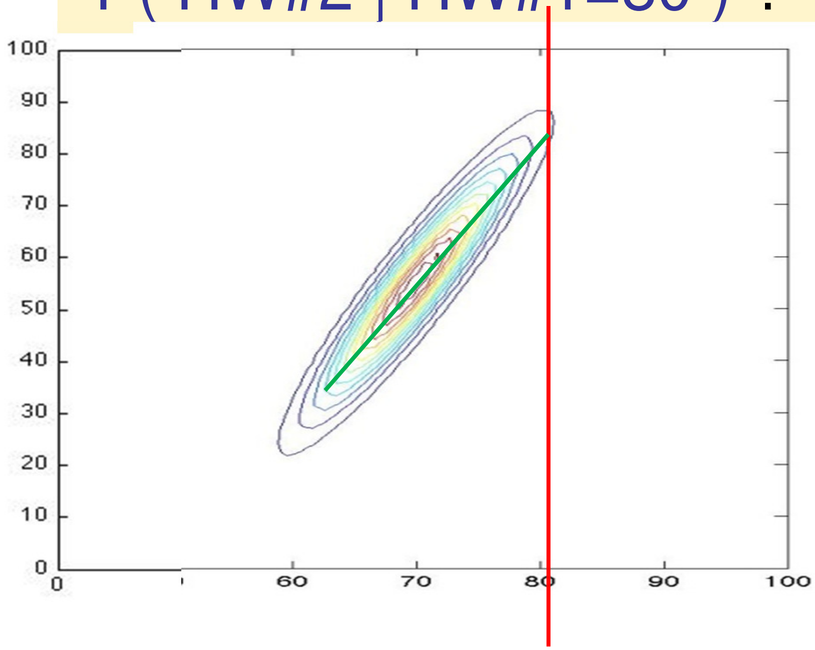
P(HW#2 | HW#1=80) = P(HW#2 | HW#1=60) = …

So

P(HW#2 | HW#1) = P(HW#2)

Joint Distribution (HW1 & HW2)

What is mean, variance of HW#2, for values of HW#1 :
P(HW#2 | HW#1=80) ?

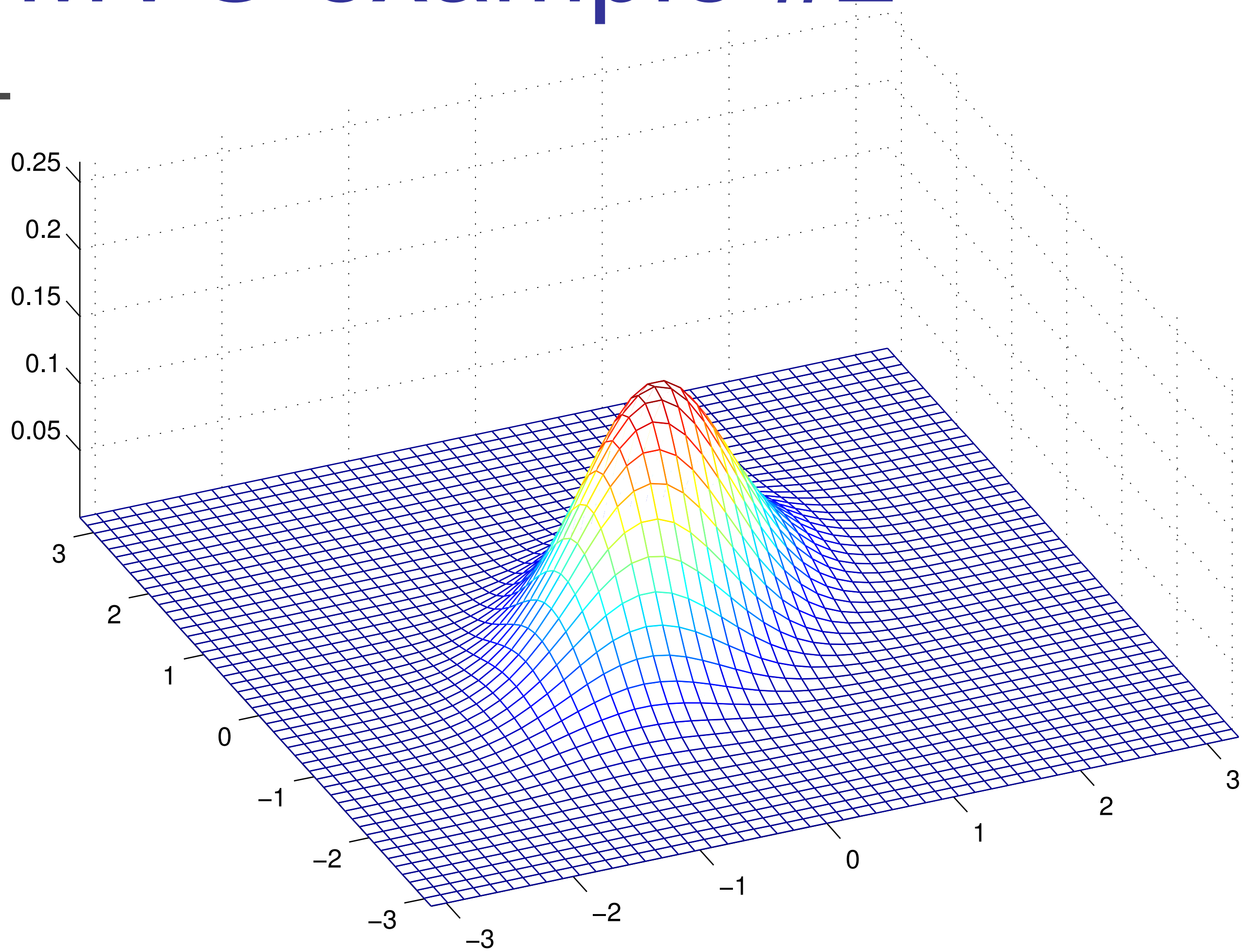


P(HW#2 | HW#1=80) ≠ P(HW#2 | HW#1=60) ≠ …

So

P(HW#2 | HW#1) ≠ P(HW#2)

MVG example #2



$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

Correlation

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- Covariance:

$$\sigma_{1,2}^2 = E[\ (x_i - \mu_i)\ (xj - \mu_j)\ ]$$

- Correlation: $\rho_{1,2} = \frac{\sigma_{1,2}^2}{\sigma_1\sigma_2}$
- $\rho_{1,2} \in [-1, 1]$

  measures linear relationship between variables
  - $\rho_{1,2} = 0$ → independent
  - $\rho_{1,2} = +1$ → identical
  - $\rho_{1,2} = -1$ → opposite sign

Bivariate Gaussian

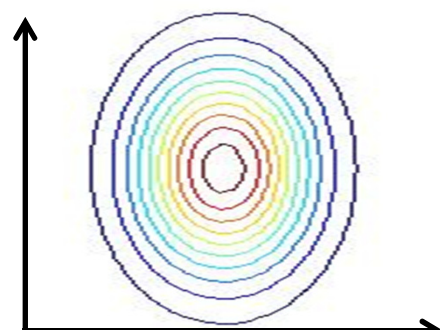
$$p(X) = \frac{\exp\left(-\left(\mathbf{X} - \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}\right)^\top \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}^{-1} \left(\mathbf{X} - \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}\right)\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

Inverting diagonal matrix

- If $\rho = 0$

$$p(X) = \frac{\exp\left(-\left(\mathbf{X} - \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}\right)^\top \begin{bmatrix}\sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2}\end{bmatrix} \left(\mathbf{X} - \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}\right)\right)}{2\pi\sigma_1\sigma_2}$$

- No cross terms, $X_1 X_2$
  $\Rightarrow$ Distribution factors
  ($X_1 \perp X_2$)



Multivariate Distribution

- Covariance form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

n = #dimensions

$$\mathbf{J} = \boldsymbol{\Sigma}^{-1}$$

"Precision Matrix"

- Information form

$$p(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^\top J\mathbf{x} + \mathbf{x}^\top J\mu\right]$$

- Derivation:

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top J(\mathbf{x}-\boldsymbol{\mu})$$

$$= -\frac{1}{2}(\mathbf{x}^\top J\,\mathbf{x} - 2\,\mathbf{x}^\top J\boldsymbol{\mu} + \boldsymbol{\mu}^\top J\boldsymbol{\mu})$$

constant term goes into partition function

Marginalization

- Given

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \right)$$

- Then

$$\boldsymbol{X} \sim N(\mu_X, \quad \Sigma_{XX})$$

Conditioning: p(X | Y=y_0)

- Given

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Gamma_{XX} & \Gamma_{XY} \\ \Gamma_{YX} & \Gamma_{YY} \end{bmatrix} \right)$$

- Then

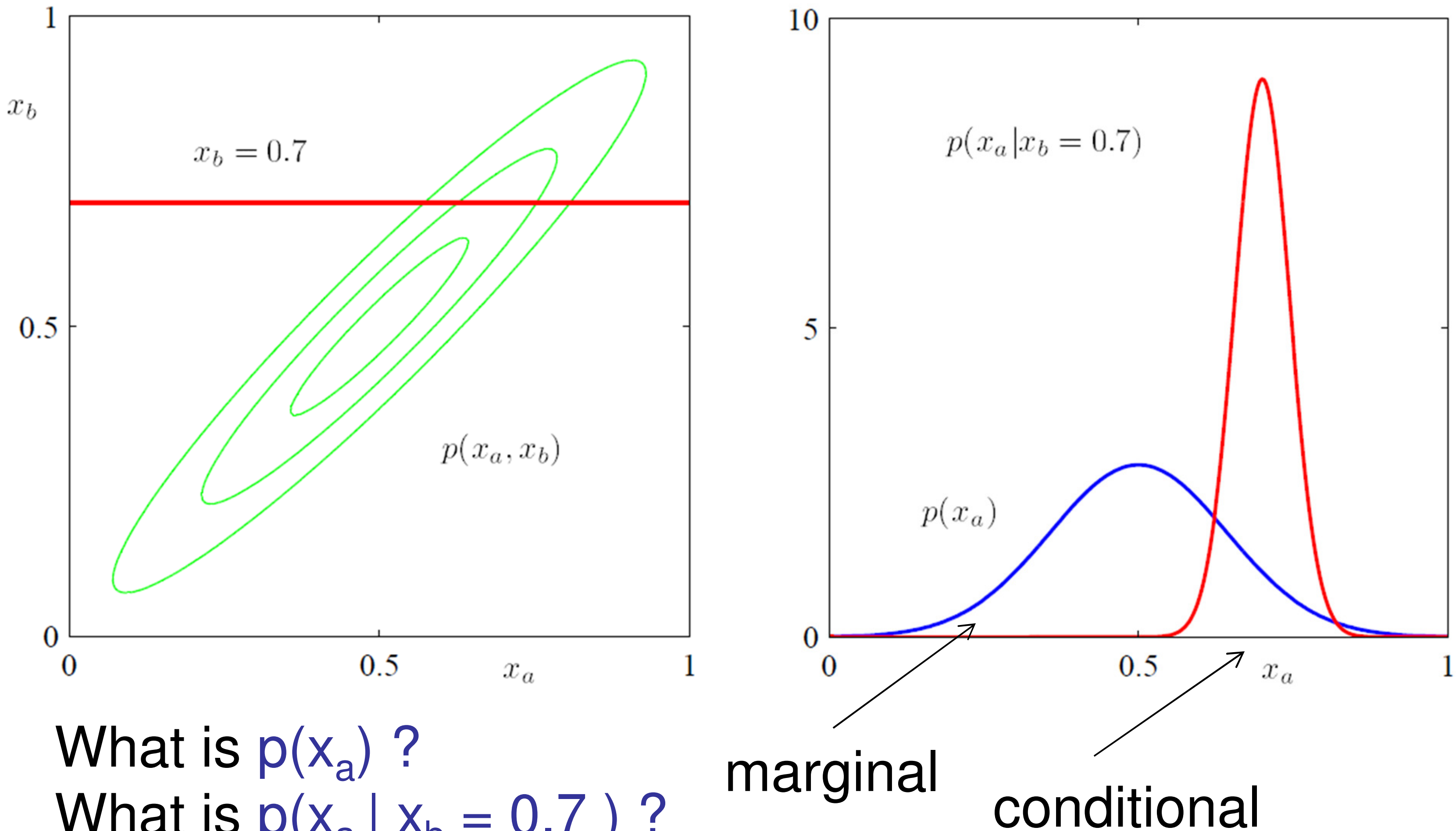
$$X|\, Y = y_0 \;\sim\; N(\, \mu_X - \Gamma_{XX}^{-1}\, \Gamma_{XY}(y_0 - \mu_Y),\; \Gamma_{XX}\,)$$

$$X|\, Y = y_0 \sim N(\, \mu_X - \Sigma_{XX}\, \Sigma_{YY}^{-1}(y_0 - \mu_Y),\; \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\,)$$

- Mean $\mu_{X|Y=y0}$ moved based on correlation and variance of measurement (Y=y_0)
- Covariance $\Sigma_{X|Y=y0}$ does not depend on y_0

http://www.cs.berkeley.edu/~pabbeel/cs287-fa11/slides/Gaussians-v2.pdf

Visualizing Marginalization & Conditioning

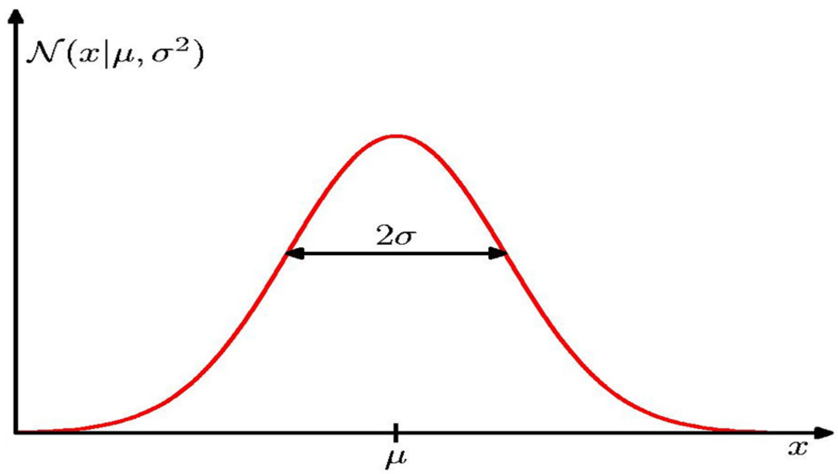


What is $p(x_a)$?
What is $p(x_a \mid x_b = 0.7)$?

Learning a Gaussian

- Collect a set of data, D
  of real-valued i.i.d. instances
  - e.g., exam scores

99
75
82
…
93
:



- Learn parameters
  - Mean, μ
  - Variance, σ

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. instances $D = \{x_1,\ldots,x_N\}$:

$$P(D \mid \mu,\sigma) = \prod_{i=1}^{N} P(x_i \mid \mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N} \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu,\sigma) &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N} \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

MLE for mean of a Gaussian

- What is ML estimate $\hat{\mu}_{MLE}$ for mean μ?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right] = \frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i-\mu) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{N} x_i - N\mu\right]$$

$$\frac{d}{d\mu} \ln P(D \mid \mu, \sigma) = 0 \quad \Rightarrow \quad \left[\sum_{i=1}^{N} x_i - N\mu\right] = 0$$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Just empirical mean!!

Toy Problem

Data, d[i]

A	1	-0.3	-0.8	-1.1	-0.8
B	6.1	5.1	5.4	3.5	2

Estimated Mean

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} d[i]$$

$$= [-0.4, 4.4]$$

Estimated Variance (unbiased)

$$\hat{\Sigma} = \frac{1}{\mathrm{M}-1}\sum_{i=1}^{M}(d[i]-\hat{\mu})(d[i]-\hat{\mu})^{T}$$

$$= \begin{bmatrix} 0.7 & 0.9 \\ 0.9 & 2.7 \end{bmatrix}$$

Estimated Variance (MLE)

$$\hat{\Sigma} = \frac{1}{\mathrm{M}}\sum_{i=1}^{M}(d[i]-\hat{\mu})(d[i]-\hat{\mu})^{T}$$

$$= \begin{bmatrix} 0.6 & 0.7 \\ 0.7 & 2.2 \end{bmatrix}$$

Estimated correlation