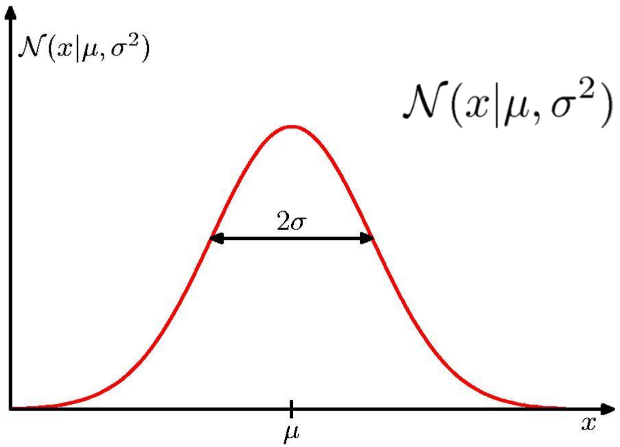
One-parameter Models
Part2 - continuous observation and Two parameter model

(chapter 5, chapter 7.1)

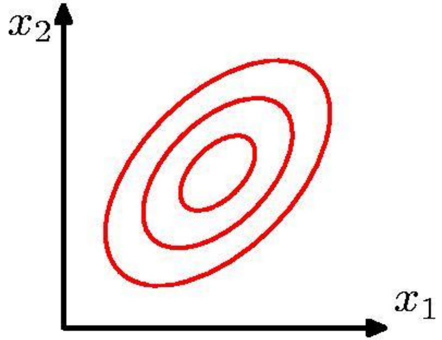
Outline

- I). Univariate & Multivariate Gaussian distribution
- II). MLE of Gaussian
- III). Bayesian inference for Gaussian
- IV). Two-parameter model - Gaussian

I). Univariate & Multivariate Gaussian distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

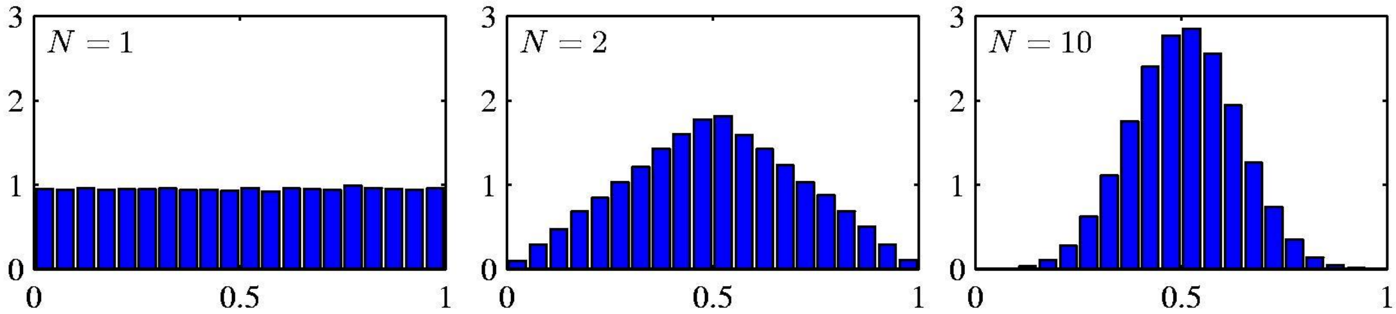


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Dimention

Why Gaussian - Central Limit Theorem

- The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.
- Example: N uniform [0,1] random variables.



Moments of the Multivariate Gaussian

let $x = z\mu$

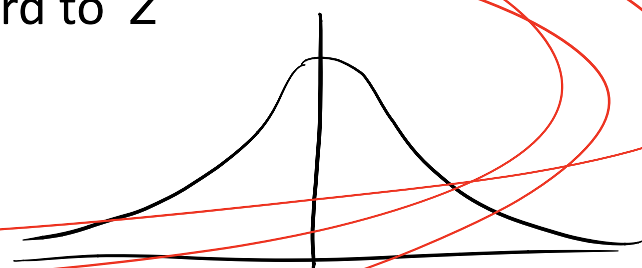
$$\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\} \mathbf{x} \, \mathrm{d}\mathbf{x}$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{ -\frac{1}{2}\mathbf{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{z} \right\} (\mathbf{z}+\boldsymbol{\mu}) \, \mathrm{d}\mathbf{z}$$

??

thanks to symmetry of the density function with regard to z

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

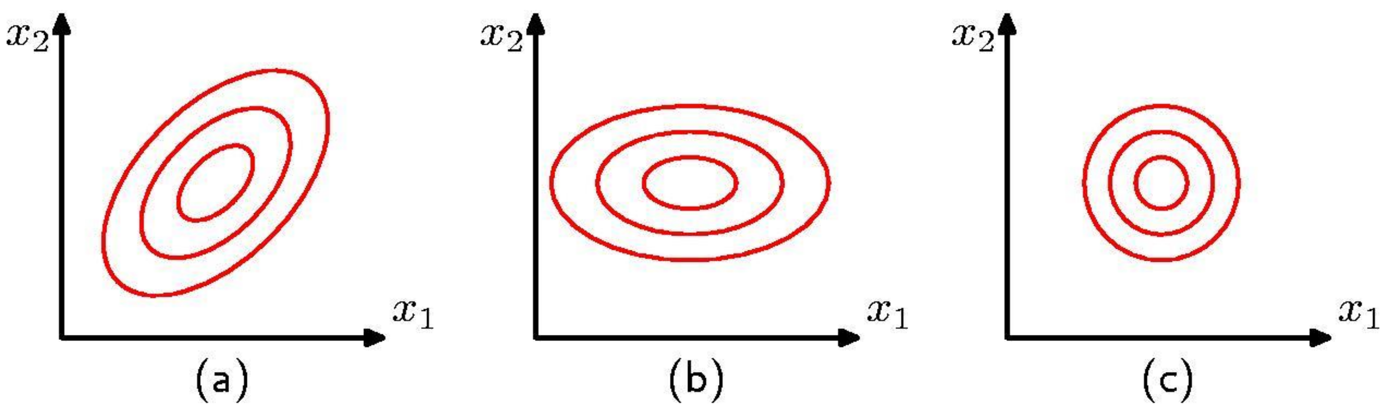


Moments of the Multivariate Gaussian

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} A & B & C \end{pmatrix} = \begin{bmatrix} aA & aB & aC \\ bA & bB & bC \\ cA & cB & cC \end{bmatrix}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$



Partitioned Gaussian Distributions

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

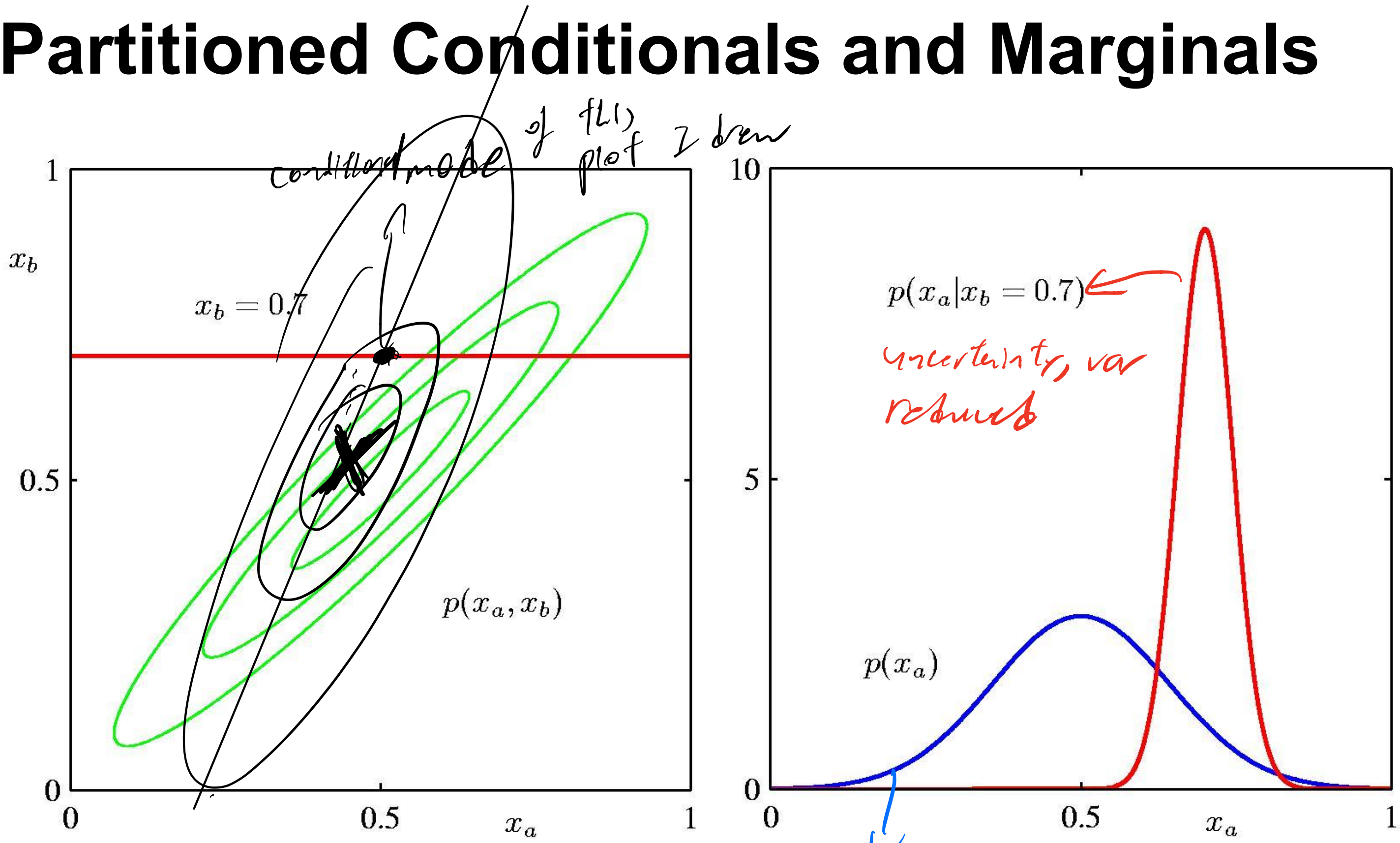
Partitioned Conditionals and Marginals

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Partitioned Conditionals and Marginals



marginal dist of x_a

Bayes' Theorem for Gaussian Variables

$p(x) \cdot p(y|x) = p(x,y)$
$\shortparallel$
$p(y) \cdot p(x|y)$

prior

- Given

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

likelihood

- we have

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})
\end{aligned}
$$

posterior

- where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}$$

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \overset{+ve}{0} & & \\ & +ve & \\ & & +ve \end{pmatrix}$$

If value $= 0$, $\Lambda_{12} = 0$

X_1 and x_2 ind $| X_3$

Frequentist

II). MLE of Gaussian

Maximum Likelihood for the Gaussian (1)

D dimention data

N 個 D dimention vector

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ (N×D), the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Sufficient statistics

vector

n×n matrix

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \sum_{n=1}^{N}\boxed{\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}}$$

covariance

Maximum Likelihood for the Gaussian (2)

- Set the derivative of the log likelihood function to zero,

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \overbrace{\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})}^{\text{vector}} = \underbrace{0}_{}$$

zero vector

set
$$\begin{bmatrix} \frac{\partial L}{\partial \mu_1} = 0 \\ \vdots \\ \frac{\partial L}{\partial \mu_N} = 0 \end{bmatrix}$$

- and solve to obtain

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

- Similarly

$$\boxed{\boldsymbol{\Sigma}_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

sample covariance

Maximum Likelihood for the Gaussian (3)

Under the true distribution

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}.
\end{aligned}$$

Bias term

Hence define

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

unbiased estimator of population covariance matrix

Sequential Estimation

Contribution of the N^{th} data point, x_N

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})
\end{aligned}
$$

current data

correction given x_N

correction weight

old estimate

III). Bayesian Inference for Gaussian

Bayesian

- Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

1 dimention

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

Distribution of $\mathbf{x}$
not dist. of μ

Bayesian Inference for the Gaussian (2)

- Combined with a Gaussian prior over μ,

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

- this gives the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

- Completing the square over μ, we see that

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

Bayesian Inference for the Gaussian (3)

- … where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \boxed{\frac{1}{\sigma_0^2}} + \boxed{\frac{N}{\sigma^2}}.$$

new precision

old precision

data precision

as $N \to \infty$

$\frac{1}{\sigma_n^2} \to$ Big $\Rightarrow \sigma_n^2$ small

- Note:

	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

If changing Informative Prior to non-informative prior

no prior idea what μ can be !!

Let $\sigma_0^2 \to \infty$, then p(μ)∝ 1 → looks like uniform dist between $-\infty$ to ∞

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

→ even $\sigma_0^2 \to \infty$, new data still speaks and give reasonable σ^2

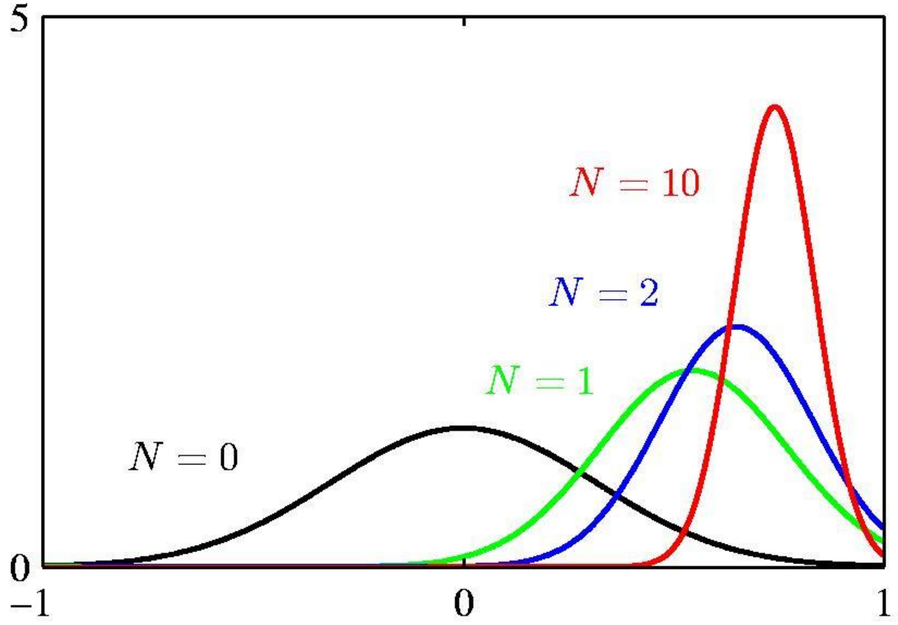
An improper prior

=> Same as MLE for the mean of Gaussian

But posterior can still be proper

Bayesian Inference for the Gaussian (4)

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.



Bayesian Inference for the Gaussian (5)

- Sequential Estimation

Prior

likelihood

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \left[p(\mu) \prod_{n=1}^{N-1} p(x_n|\mu) \right] p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1}, \sigma_{N-1}^2\right) p(x_N|\mu)
\end{aligned}
$$

normal

normal

- The posterior obtained after observing N-1 data points becomes the prior when we observe the Nth data point.

1 - Dimensional

Bayesian Inference for the Gaussian (6)

- Now assume μ is known. The likelihood function for $\lambda=1/\sigma^2$ is given by

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

- This has a Gamma shape as a function of λ.

estimate precision rather than variance

easier

Prior

$$p(\lambda) \propto \underline{\lambda^{c} e^{-\lambda d}}$$

$$p(x|\lambda) \propto \lambda^{a} e^{-\lambda b}$$

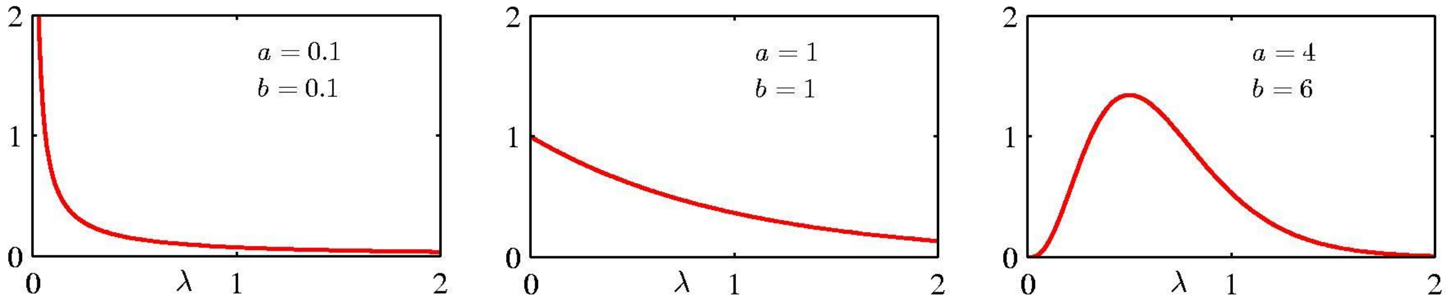
$$p(\lambda|x) \propto \lambda^{a+c} e^{-\lambda(a+b)}$$

Bayesian Inference for the Gaussian (7)

- The Gamma distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



Bayesian Inference for the Gaussian (8)

- Now we combine a Gamma prior, $\mathrm{Gam}(\lambda|a_0, b_0)$, with the likelihood function for λ to obtain

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- which we recognize as $\mathrm{Gam}(\lambda|a_N, b_N)$ with

prior sample size

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}.$$

prior sample var

If changing Informative Prior to non-informative prior

when absolutely no prior knowledge

Let $a_0 = b_0 = 0$, then $p(\lambda) \propto 1/\lambda$

$$p(\lambda|\mathbf{x}) \propto \lambda^{N/2-1} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$$=\mathrm{Gam}\left(\frac{N}{2}, \frac{N}{2}\sigma^2_{\mathrm{ML}}\right)$$

Note:

Improper priors can lead to proper posteriors

IV). Two-parameter model - Gaussian
Bayesian Inference for the Gaussian (9)

- If both μ and λ are unknown, the joint likelihood function is given by

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

- We need a prior with the same functional dependence on μ and λ

Bayesian Inference for the Gaussian (10)

Prior, when both μ and λ unknown

- The Gaussian-gamma distribution

normal gamma

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

The two dist. not indep.

$$\propto \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}}\ \underbrace{\lambda^{a-1}\exp\{-b\lambda\}}$$

- Quadratic in μ.
- Linear in λ.

- Gamma distribution over λ.
- Independent of μ.

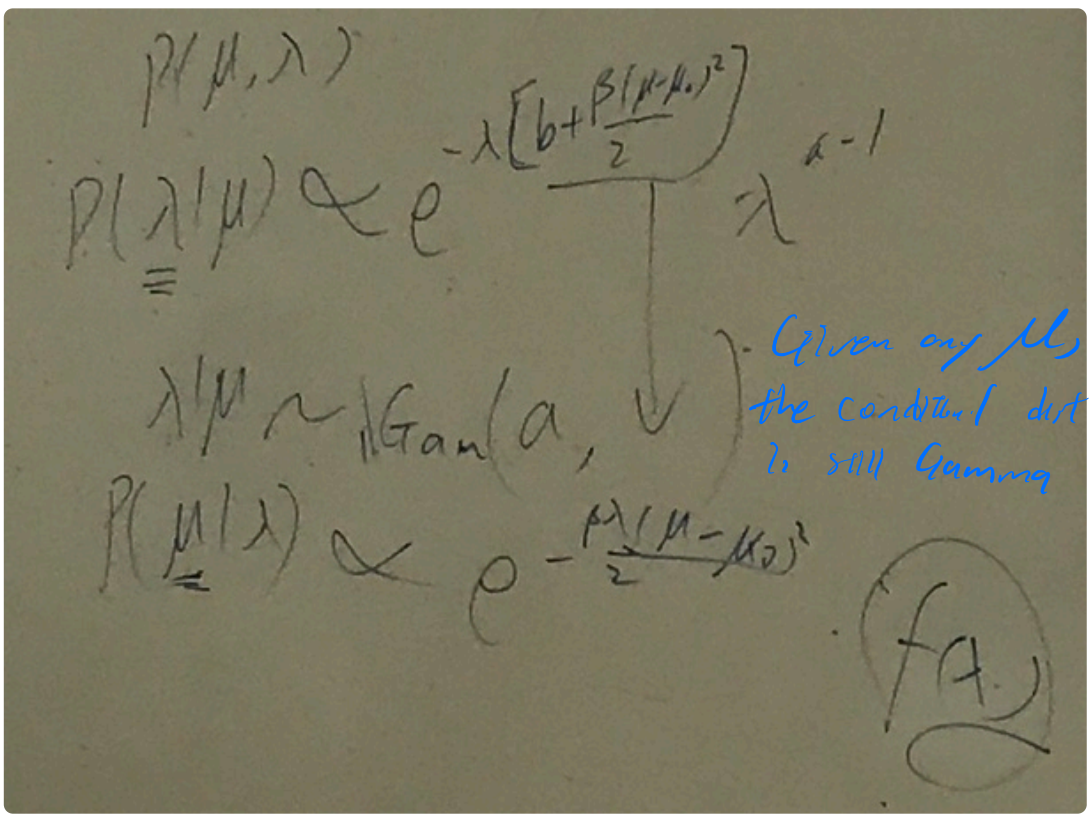
- If we integrate μ, only $\mathrm{Gam}(\lambda|a,b)$ left
- If we condition on any λ, only $\mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})$ left, Normal!!!

Any joint distribution

$$P(a,b,c) = P(a) \cdot P(b|a) \cdot P(c|a,b)$$

However, if we know $c \perp\!\!\!\perp b \mid a$

$$\Rightarrow P(c|a,b) = P(c|a)$$



Given any μ, the conditional dist of λ is still Gamma

Just treat the given parameter as constant let $p(a,b) = C \cdot f(a) \cdot g(b)$

$$p(a|b) = \frac{p(a,b)}{p(b)} = \frac{p(a,b)}{\int p(a,b)\,da}$$

$$= \frac{C \cdot f(a) \cdot f(b)}{\int C \cdot f(a) \cdot f(b)\,da}$$

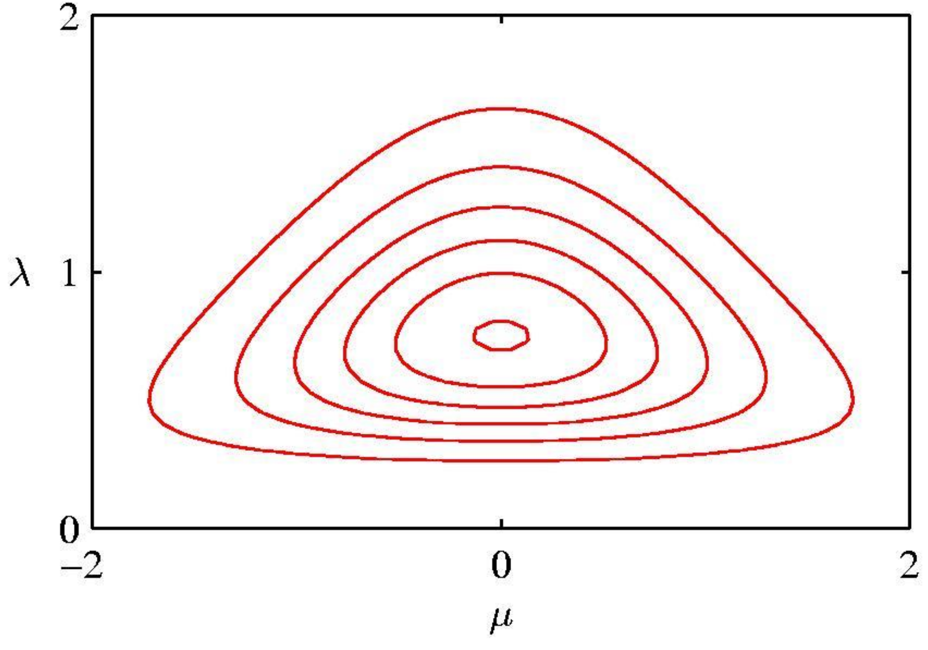
$$= \frac{\cancel{C} \cdot f(a) \cdot \cancel{f(b)}}{\cancel{C} \cdot \cancel{f(b)} \cdot \int f(a)\,da}$$

$$= f(a) \Big/ \int f(a)\,da$$

$$\therefore \underline{\underline{p(a|b) \propto f(a)}}$$

Bayesian Inference for the Gaussian (11)

- The Gaussian-gamma distribution



General Discussion for Two-parameter model

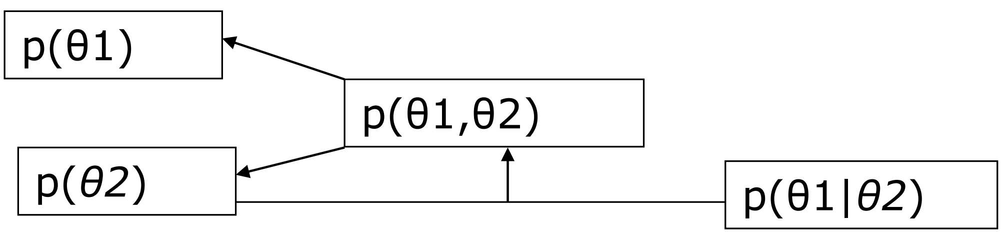
Joint, Marginal, Conditional Prior

$$P(\theta 1)=\int p(\theta 1,\theta 2)d\theta 2$$

$$P(\theta 2)=\int p(\theta 1,\theta 2)d\theta 1$$

$$P(\theta 1,\theta 2)=p(\theta 1|\theta 2)p(\theta 2)$$
$$=p(\theta 2|\theta 1)p(\theta 1)$$

Any joint dist can be specified by a marginal and a conditional



let's say we have this prior

$$p(x_1, x_2) \propto p(x_1) \cdot p(x_2)$$

Indep

It is possible that the posterior

$$p(x_1, x_2 | data) \propto N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix} \right)$$

not indept.

If x_1 and x_2 are indeed related, they will speak in your data

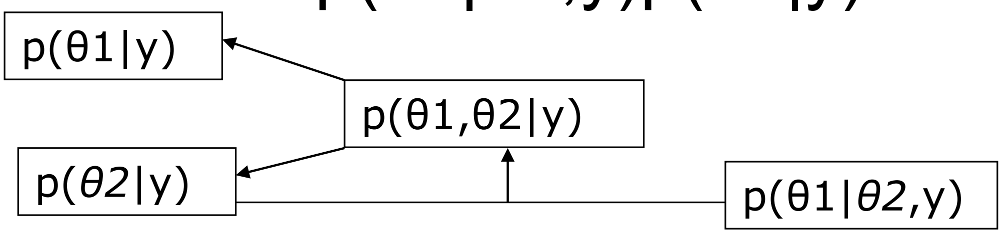
Joint, Marginal, Conditional Posterior

$$P(\theta1|y)=\int p(\theta1,\theta2|y)d\theta2$$

$$P(\theta2|y)=\int p(\theta1,\theta2|y)d\theta1$$

$$P(\theta1,\theta2|y)=p(\theta1|\theta2,y)p(\theta2|y)$$

$$=p(\theta2|\theta1,y)p(\theta1|y)$$



Strategy for Multi-Parameter Inference

Just assume indep prior, if parameters are indeed correlated, data will speak

1. Specify a joint prior p(θ1, *θ2*)

2. Write down likelihood
   1. L(θ1, *θ2*)=P(y|θ1, *θ2*)

3. Derive the joint posterior
   1. P(θ1, *θ2* |y)∝ p(θ1, *θ2*) p(y|θ1, *θ2*)

4. Then make joint or marginal inference

Strategy for Multi-Parameter Inference for Normal Data

1. Specify a joint prior $p(\mu, \sigma^2)$

   $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$
2. Write down likelihood

   $L(\mu, \sigma^2) = P(y|\mu, \sigma^2)$
3. Derive the joint posterior

   $P(\mu, \sigma^2|y) \propto p(\mu, \sigma^2)\, p(y|\mu, \sigma^2)$
4. Then make marginal inference
   - $P(\mu|y) = \boxed{\int p(\mu, \sigma^2|y)d\sigma^2}$
   - $P(\sigma^2|y) = \int p(\mu, \sigma^2|y)d\mu$

Normal model, $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$

Non-informative prior

- joint posterior

$$p(\mu, \sigma^2 \mid y) \propto \sigma^{-(n+2)}$$
$$\exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right)$$

- conditional posterior distribution of μ

$p(\mu \mid \sigma^2, y) \propto e^{-n(\bar{y}-\mu)^2}$

$p(\sigma^2 \mid \mu, y) \propto \sigma^{2(-\frac{n+2}{2})} \cdot e^{-\frac{A}{2\sigma^2}}$

$$\mu \mid \sigma^2, y \sim N(\mu \mid \bar{y}, \frac{\sigma^2}{n})$$

- marginal posterior distribution of σ^2

$$\sigma^2 \mid y \sim IG(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Marginal posterior distribution of μ

$$\mu \mid y \quad \sim t_{n-1}(\bar{y}, \frac{s^2}{n})$$

IG = Inverse Gamma

Numerical Integration

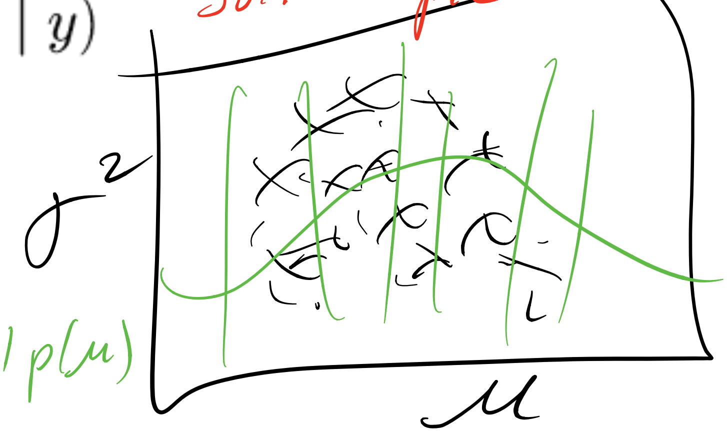
a posterior

$$p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y)p(\sigma^2 \mid y)$$

1. • draw $\boxed{\sigma^{2(m)}} \sim IG(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$ → set as marginal
2. • draw $\mu^{(m)} \sim N(\mu \mid \bar{y}, \sigma^{2(m)}/n)$ → set as conditional $\mu \mid \sigma^2$

• $\mu^{(m)}, \sigma^{2(m)}$ $m = 1, \ldots, M$ is a sample from the joint posterior distribution $p(\mu, \sigma^2 \mid y)$

Joint Sample



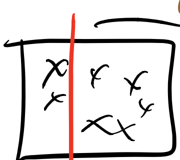
To get $p(\mu)$, $\int d\sigma^2$

but require large sample size

marginal $p(\mu)$

→ no data

$p(b|a)$ b



Fail....

better do an interval

Normal with Conjugate Prior

$$p(y \mid \mu, \sigma^2) \sim (\tfrac{1}{\sigma})^n \exp\left(-\tfrac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right)$$

$$p(\mu, \sigma^2) \sim (\tfrac{1}{\sigma})(\tfrac{1}{\sigma})^{\nu_0+2} \exp\left(-\tfrac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + k_0(\mu_0-\mu)^2\right]\right)$$

Inverse gamma

- μ and σ^2 are dependent in their joint conjugate prior density

$$\mu \mid \sigma^2 \sim N(\mu_0, \sigma^2/k_0)$$

$$\sigma^2 \sim IG\left(\frac{\vartheta_0}{2}, \frac{\vartheta_0\sigma_0^2}{2}\right)$$

- $$\mu \mid \sigma^2, y \sim N(\mu_n, \sigma^2/k_n)$$
  $$\sigma^2 \mid y \sim IG(\tfrac{\nu_n}{2}, \tfrac{\nu_n\sigma_n^2}{2})$$

$$\begin{cases} \mu_n = \frac{k_0}{k_0+n}\mu_0 + \frac{n}{k_0+n}\bar{y} \\ k_n = k_0 + n \\ \nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_0+n}(\bar{y}-\mu_0)^2 \\ \nu_n = \nu_0 + n \end{cases}$$

Normal with Semi-Conjugate Prior

- μ and σ^2 are independent now μ not dependent on σ^2

$$\mu \mid \sigma^2 \sim N(\mu_0, \tau_0^2)$$

$$\sigma^2 \sim IG\left(\frac{\vartheta_0}{2}, \frac{\vartheta_0 \sigma_0^2}{2}\right)$$

- $\mu_n = \frac{\mu_0/\tau_0^2 + n/\sigma^2}{1/\tau_0^2 + n/\sigma^2}$
- $\tau_n^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}$

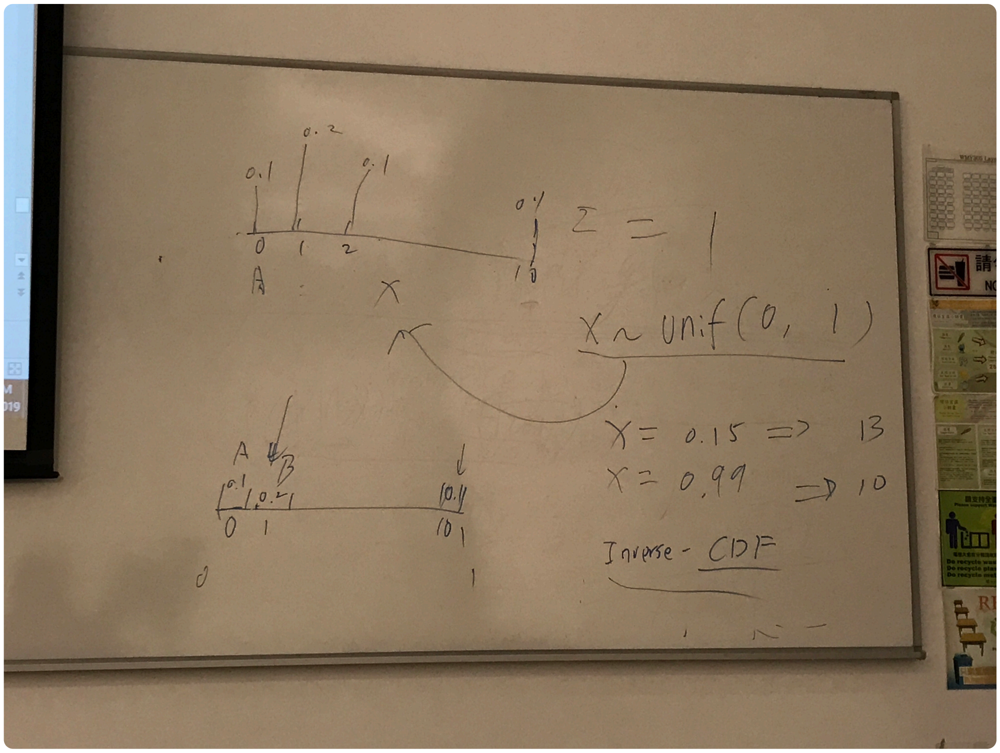
only μ maintains prior form

$$\mu \mid \sigma^2, y \sim N(\mu_n, \tau_n^2)$$

$$\sigma^2 \mid y \quad \sim \text{not in closed form}$$

obtain numerically

sampling discrete distribution



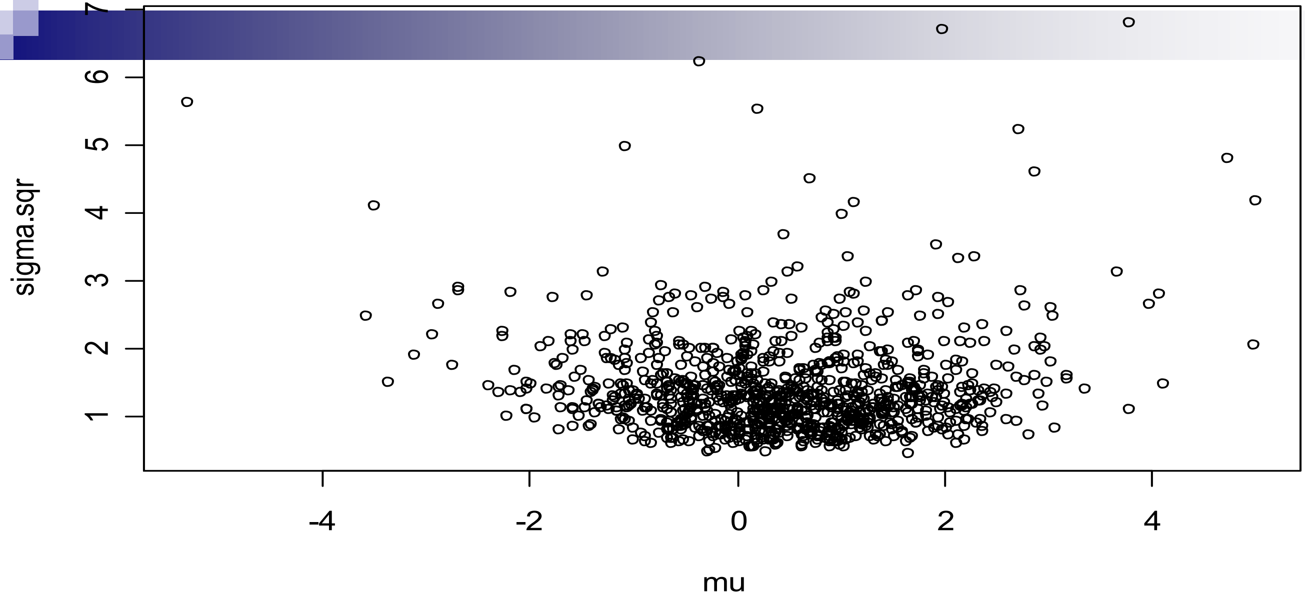
Simulating from Posterior

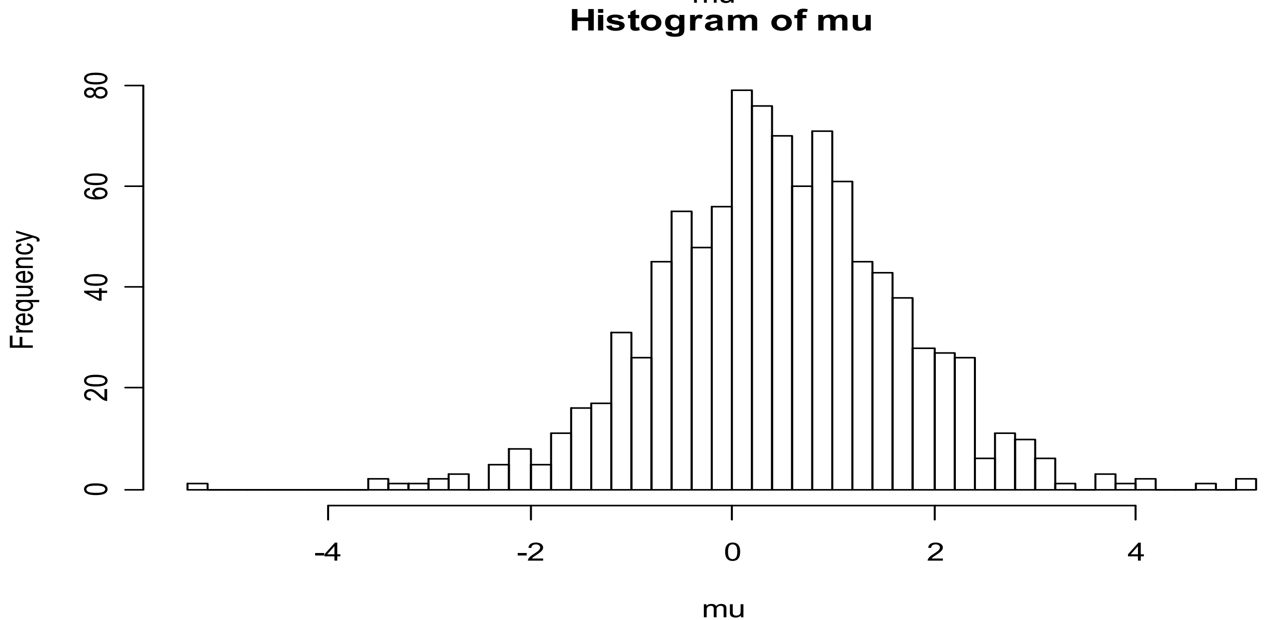
- draw $\theta_2^{(m)} \sim p(\theta_2 \mid y)$ ←σ²
- draw $\theta_1^{(m)} \sim p(\theta_1 \mid \theta_2^{(m)}, y)$ ←μ|σ2

Analytical

$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y) d\theta_2$$

$$= \int p(\theta_1|\theta_2,y)p(\theta_2|y)d\theta_2$$


sigma.sqr
1
2
3
4
5
6
7
-4
-2
0
2
4
mu


Histogram of mu
Frequency
0
20
40
60
80
-4
-2
0
2
4
mu

Simulation vs Exact

Difficulty with Exact

- Some posteriors have no close form
- Some likelihoods are too complex, high dimensional, etc.
- Needs strong math skills

Limitations of Simulation

- Simulation has stochastic error
- Iterative algorithms may not converge
- Sometimes too easy … beware
- Requires programming