Bayesian Methods:
A Refresher

Baback Moghaddam

baback@jpl.nasa.gov

Machine Learning Group

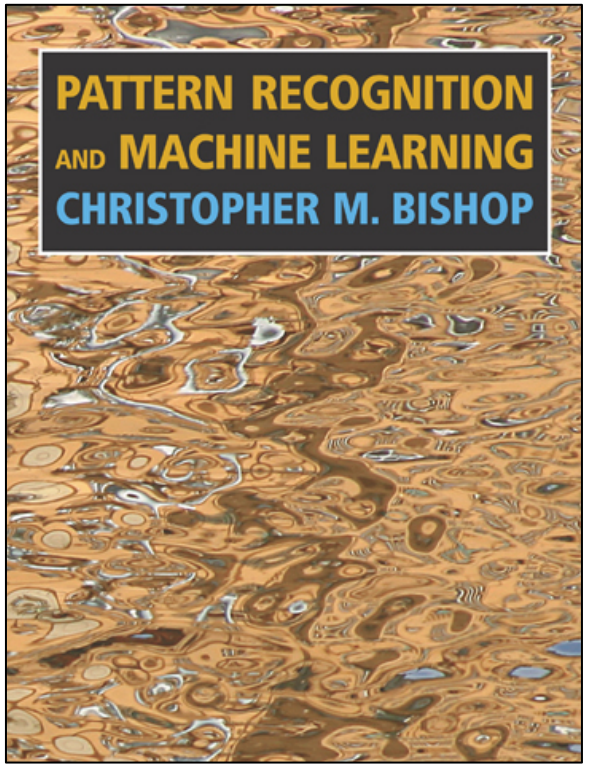
JPL

Acknowledgements

Roughly 80% of the slide material* here is courtesy of

Chris Bishop, Microsoft Research (Cambridge, UK)

and based on his textbook :

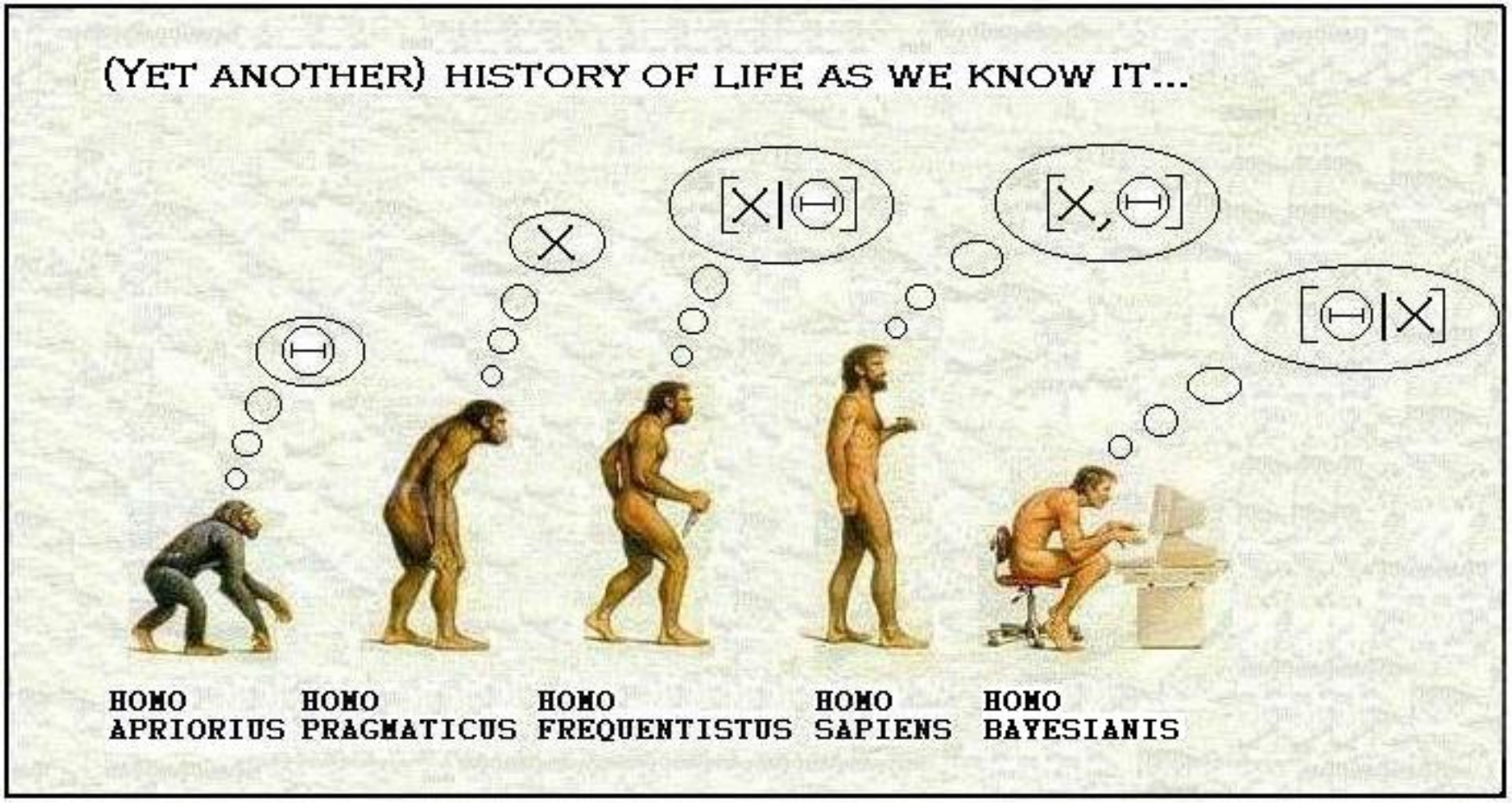


* Some of the original slides have been modified a bit.

Outline

- Bayesian Inference
- Non-Hierarchical Model
- Parametric Distributions
  - Conjugate Priors
  - Exponential Family
  - Non-Informative Priors
- Summary

Evolution of Inference



Bayesian Inference

Posterior *Likelihood* x *Prior*

$$p(\theta \mid X, M) = \frac{p(X \mid \theta, M)\ p(\theta \mid M)}{p(X \mid M)}$$

Evidence

The *evidence* for our model $\boldsymbol{M}$ is also called "Marginal Likelihood'

$$p(X \mid M) = \int p(X \mid \theta, M)\ p(\theta \mid M)\, d\theta$$

Hierarchical Model

Hyperprior

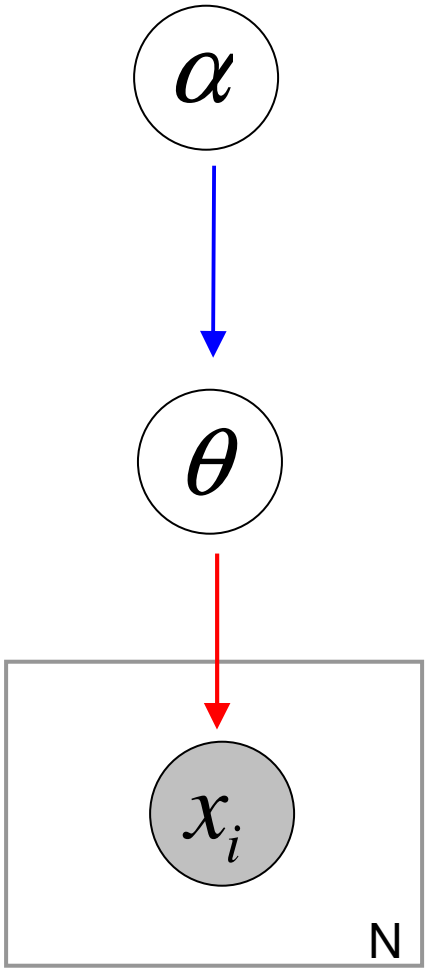
$p(\alpha)$

Prior

$p(\theta \mid \alpha)$

Likelihood

$p(X \mid \theta)$

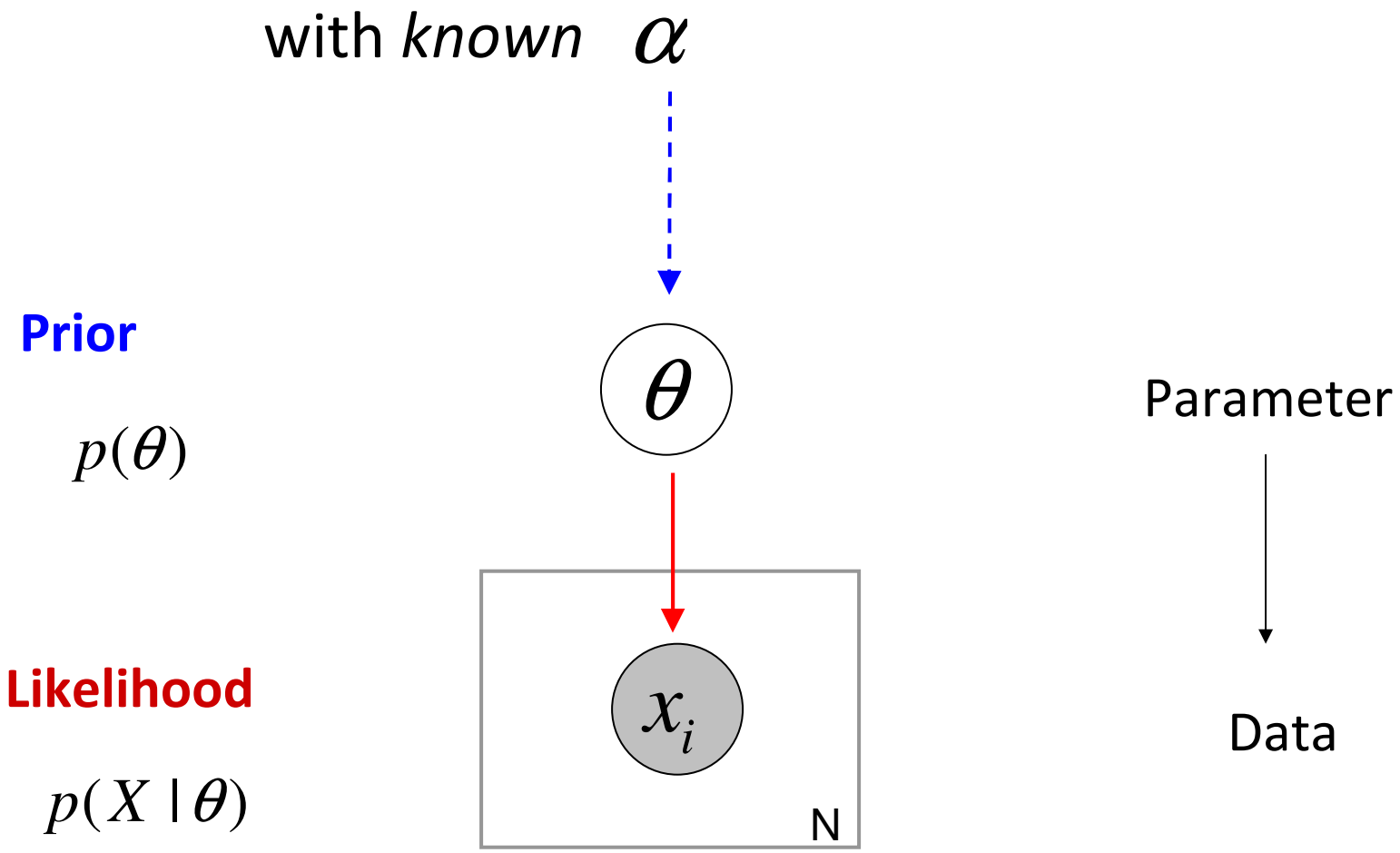


Hyperparameter

Parameter

Data

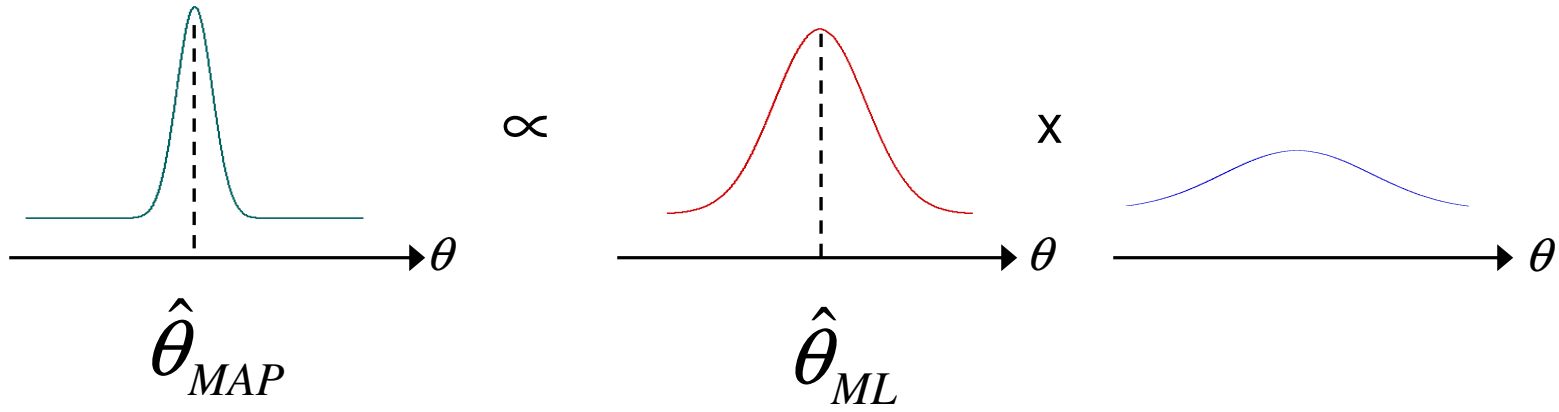
Non-Hierarchical Model



Parametric Inference

Posterior *Likelihood* x *Prior*

$$p(\theta \mid X) \propto p(X \mid \theta)\, p(\theta)$$



Parametric Distributions

Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

Need to "determine" $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Representation: $\hat{\theta}$ or $p(\theta)$?

- point estimate $\hat{\theta}$ (Frequentist)
- distribution $p(\theta)$ (Bayesian)

Parametric Predictions

Model : $p(\mathbf{x}|\boldsymbol{\theta})$ Data : $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Frequentist Prediction :

$$p(\mathrm{x}_{N+1} \mid \mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N) = p(\mathrm{x}_{N+1} \mid \hat{\theta}_{\mathrm{ML}}(\mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N))$$

Bayesian Prediction :

$$p(\mathrm{x}_{N+1} \mid \mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N) = \int p(\mathrm{x}_{N+1} \mid \theta)\, \underbrace{p(\theta \mid \mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N)}_{\text{posterior}}\, d\theta$$

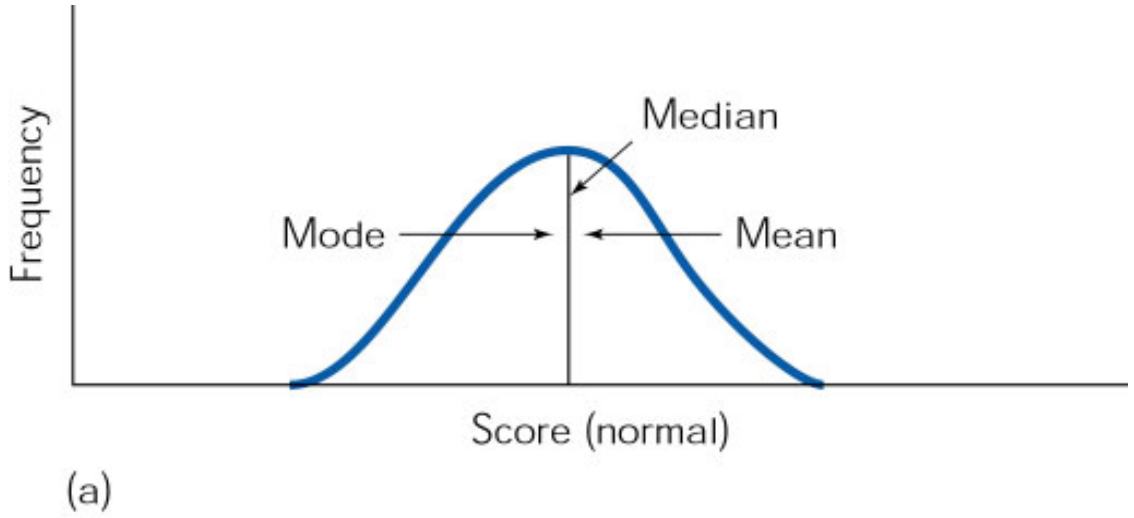
Posterior Point Estimation : $\hat{\theta}$

- Pick an *appropriate* location summary of $p(\theta|D)$
  - The mode is easiest to compute (no integration) but often not representative of the "middle" portion of distribution
  - The mean has the opposite property, tending to chase heavy tails (also not easy due to integration of moment)
  - The median is often best compromise, but is harder to compute given that it's the boundary solution to

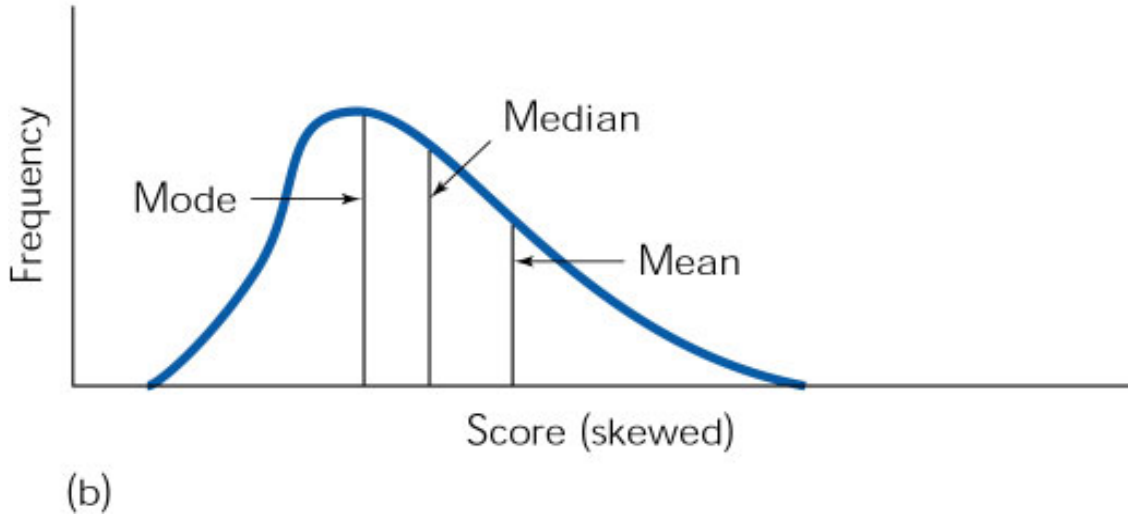
$$\int_{-\infty}^{\hat{\theta}} p(\theta \mid D)\, d\theta = 0.5$$

Posterior Point Estimation : $\hat{\theta}$

symmetric distribution



asymmetric (skewed) distribution



Point Estimation & Decision Theory

- Same as minimizing posterior expected "loss"

$$\hat{\theta} = \operatorname{argmin} \int L(\theta, \theta')\ p(\theta \mid D)\, d\theta$$

- mode minimizes "0-1" loss $\quad L(\theta, \theta') = \begin{cases} 0 & \text{if } \theta' = \theta \\ 1 & \text{otherwise} \end{cases}$
- mean minimizes quadratic loss $\quad L(\theta, \theta') = \| \theta - \theta' \|^2$
- median minimizes absolute loss $\quad L(\theta, \theta') = | \theta - \theta' |$

Why Separate Inference & Decision?

$$\hat{\theta} = \operatorname{argmin} \int L(\theta, \theta')\, p(\theta \mid D)\, d\theta$$

"decision"

inference

- Loss functions may change
- May want a reject option
- Unbalanced class priors

} classification problems

- Need to combine models
- Optimality

Basic Parametric Bayesian Models

- Bayesian Inference for:
  - Bernoulli - Beta
  - Multinomial - Dirichlet
  - Gaussian - Gaussian - Gamma
  - Gaussian - Gaussian - Wishart

  } likelihood - prior

- Conjugate Models
- Exponential Families
- Non-informative Priors

Bernoulli Variables

Coin flipping : Heads = 1 , Tails = 0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

$$\begin{aligned} \text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \text{var}[x] &= \mu(1-\mu) \end{aligned}$$

Binomial Variables

N coin flips: $p(m \text{ heads}|N, \mu)$

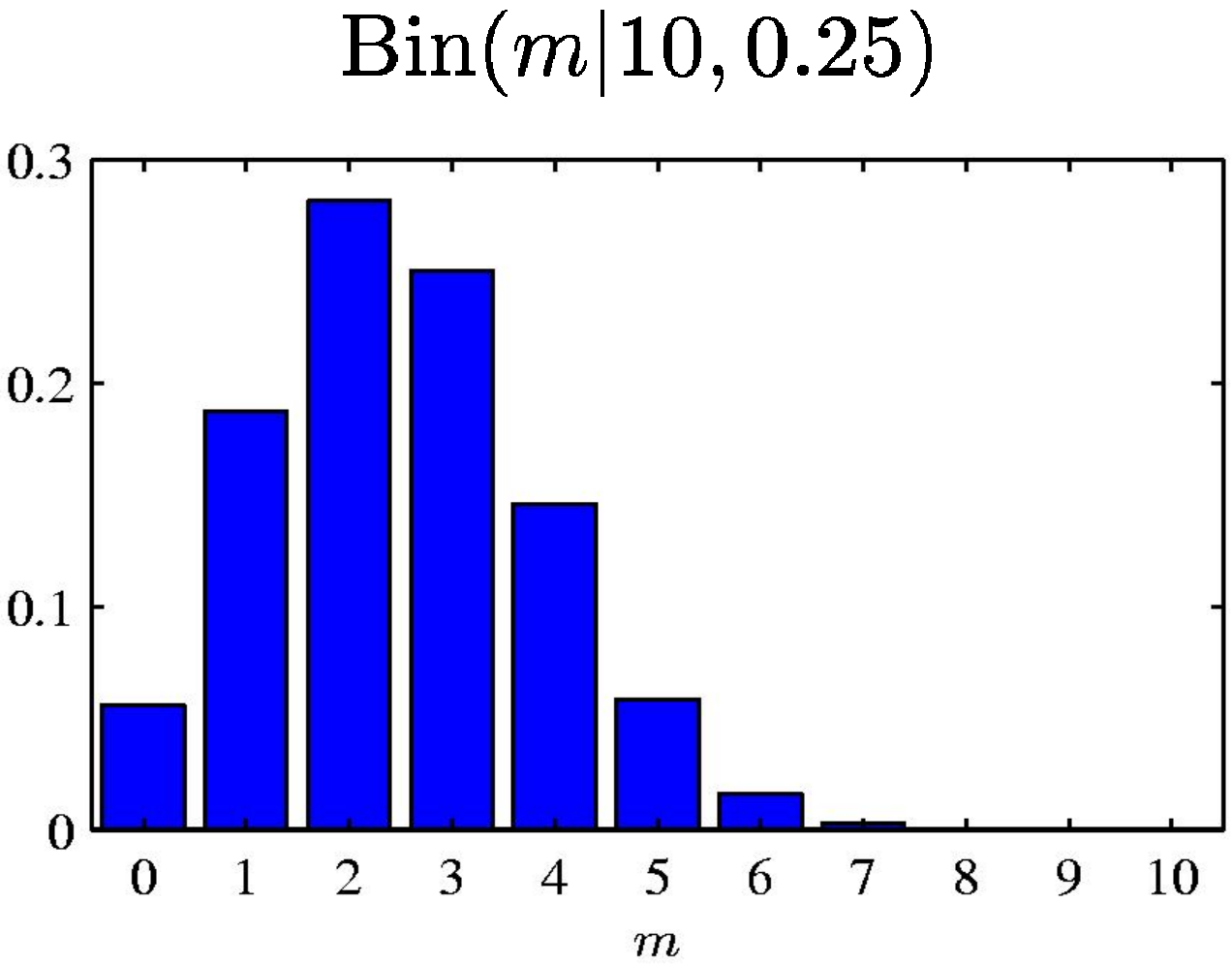
Binomial Distribution = sum of N Bernoulli

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\text{Bin}(m|N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1-\mu)$$

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

Parameter Estimation (2)

Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

Prediction : *all* future tosses will be heads!

We are “overfitting” the ML estimate to this particular (small) observed dataset $\mathcal{D}$

Beta Distribution

Distribution over $\mu \in [0, 1]$

"probability of a probability"

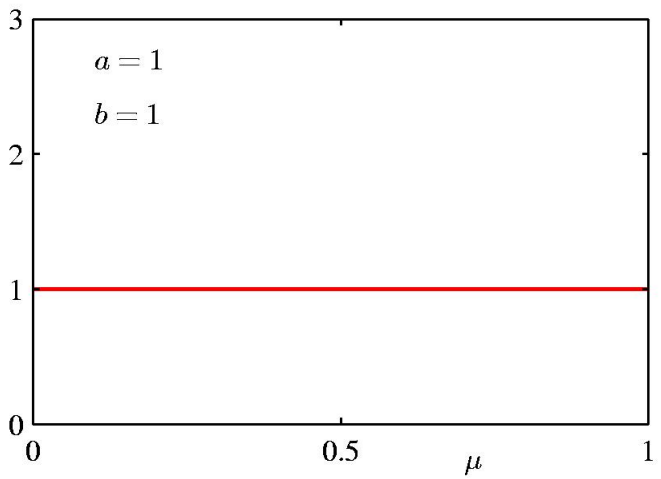
$$\text{Beta}(\mu|a, b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{Beta function}} \mu^{a-1}(1-\mu)^{b-1}$$

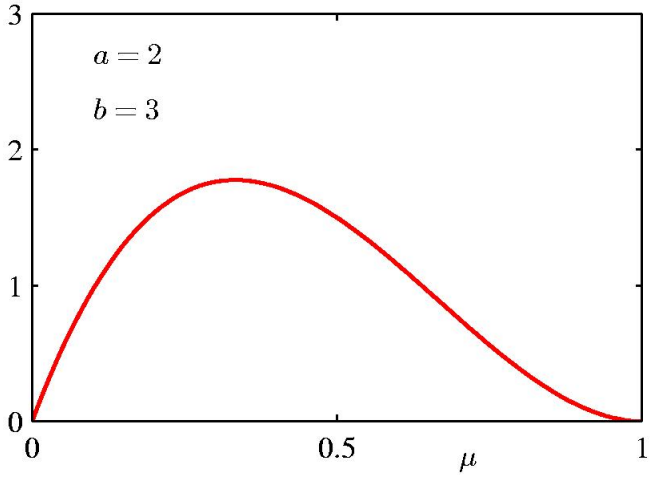
$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

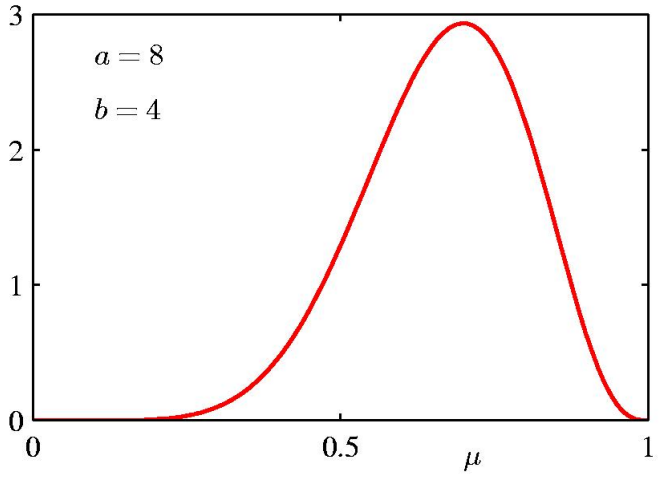
$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

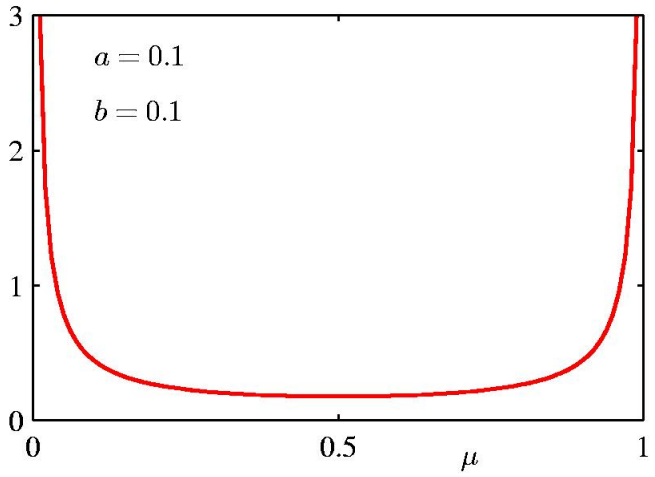
$$\text{mode}[\mu] = \frac{a+b-1}{a+b-2}$$

Beta Distribution









Bayesian Bernoulli

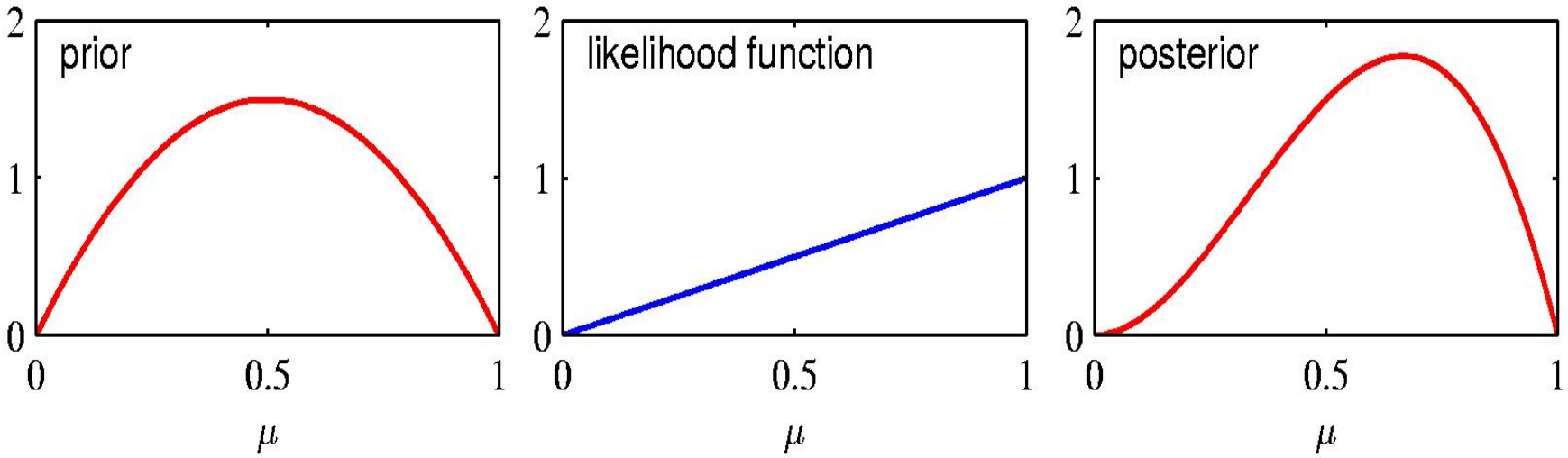
$$
\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}
$$

$$
a_N = a_0 + m \qquad b_N = b_0 + (N - m)
$$

Beta is the *conjugate* prior for Bernoulli likelihood

Prior · Likelihood $\propto$ Posterior

Beta(2,2) H = 1 T = 0 Beta(3,2)

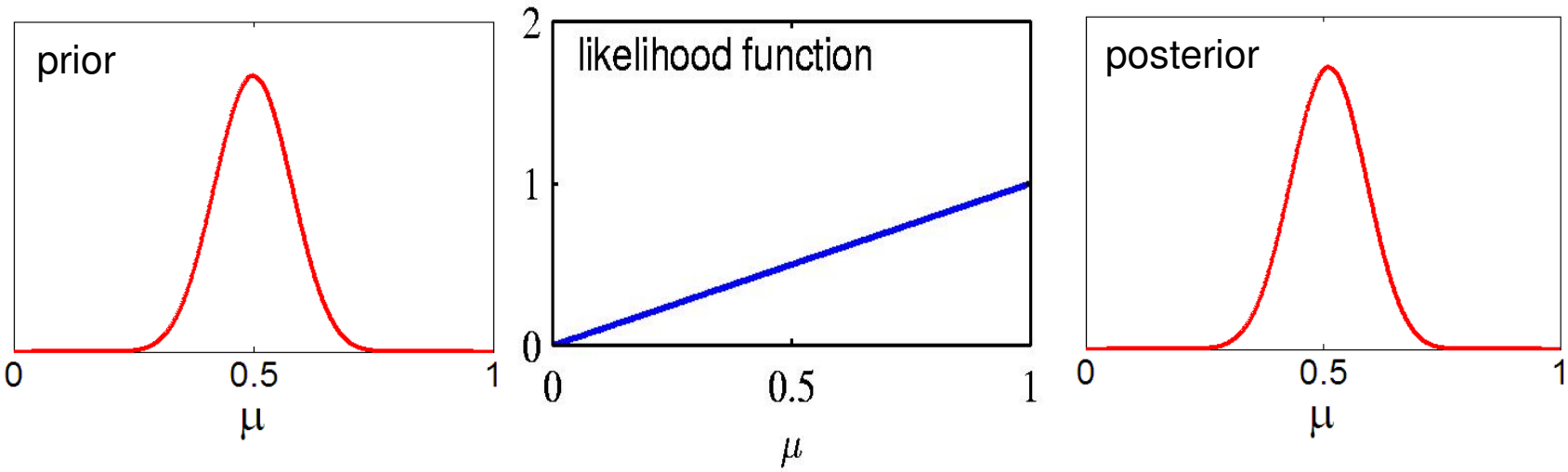


mean = 0.50
mode = 0.50
std dev = 0.22

mean = 0.60
mode = 0.66
std dev = 0.20

Prior · Likelihood $\propto$ Posterior

Beta(20,20) H = 1 T = 0 Beta(21,20)



mean = 0.50
mode = 0.50
std dev = 0.0781

mean = 0.5122
mode = 0.5128
std dev = 0.0771

Properties of the Posterior

As the size of the data set, N, increases

$$\begin{aligned} a_N &\rightarrow m \\ b_N &\rightarrow N - m \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\ \mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0 \end{aligned}$$

Prediction under the Posterior

Probability that the next coin toss will be heads ?

$$
\begin{aligned}
p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] \\
&= \frac{a_N}{a_N+b_N} = \frac{a_0+m}{a_0+b_0+N}
\end{aligned}
$$

Categorical Variables

1-of-K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

ML Parameter Estimation

Given: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}$$

Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier, λ.

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \boxed{\mu_k^{\mathrm{ML}} = \frac{m_k}{N}}$$

Multinomial Distribution

Sum of N categorical variables of dimension K

$$m_k = \sum_{i=1}^{N} x_k^{(i)}$$

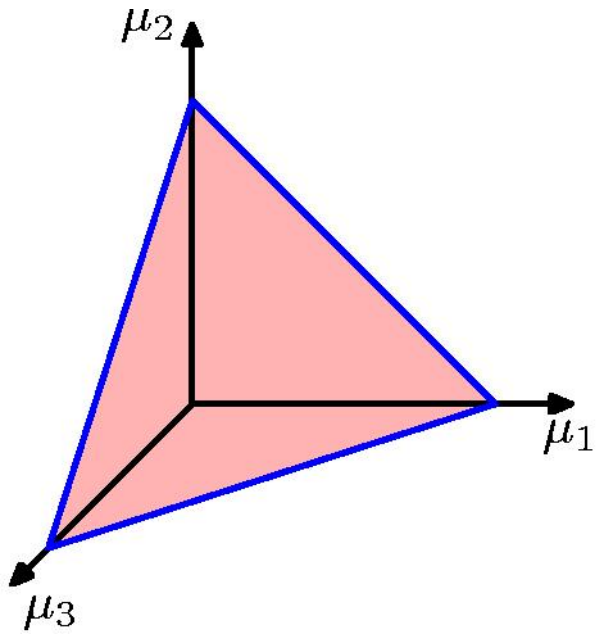
$$\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}$$

Dirichlet Distribution

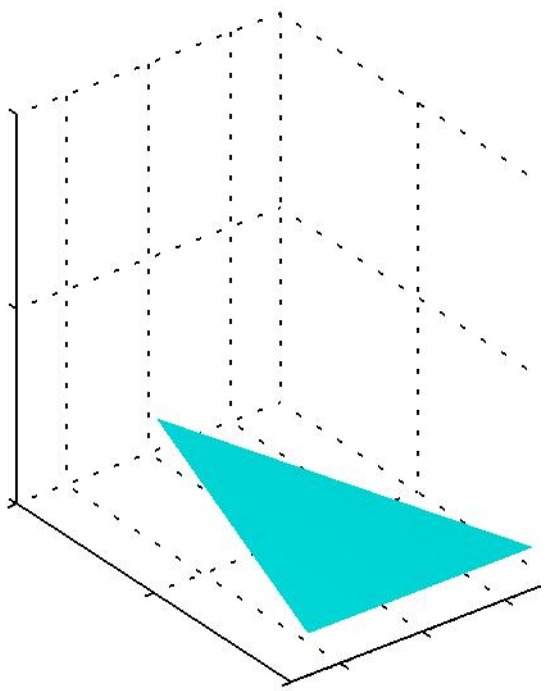
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

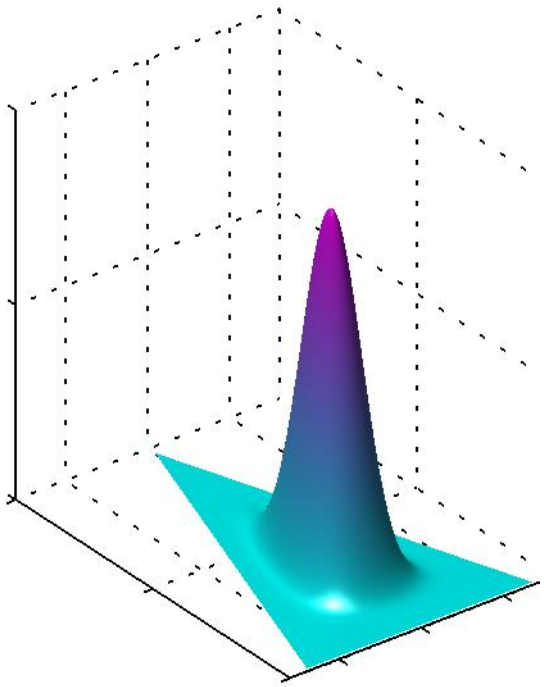
Conjugate prior for the multinomial distribution.



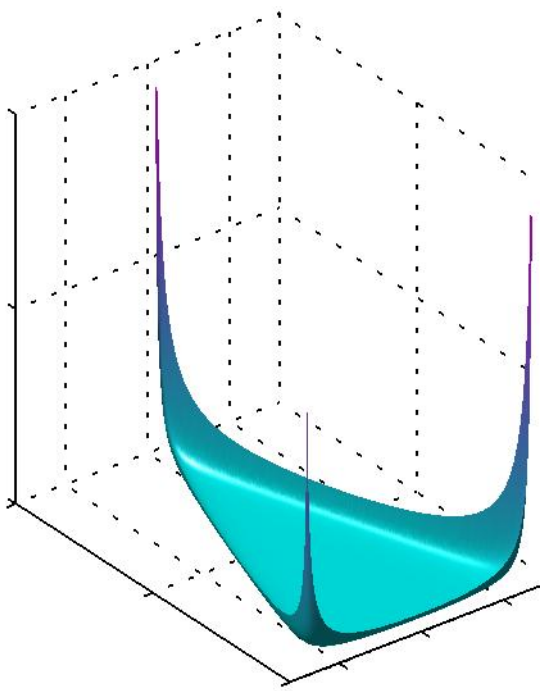
Dirichlet Prior



$\alpha_k = 10^0$



$\alpha_k = 10^1$



$\alpha_k = 10^{-1}$

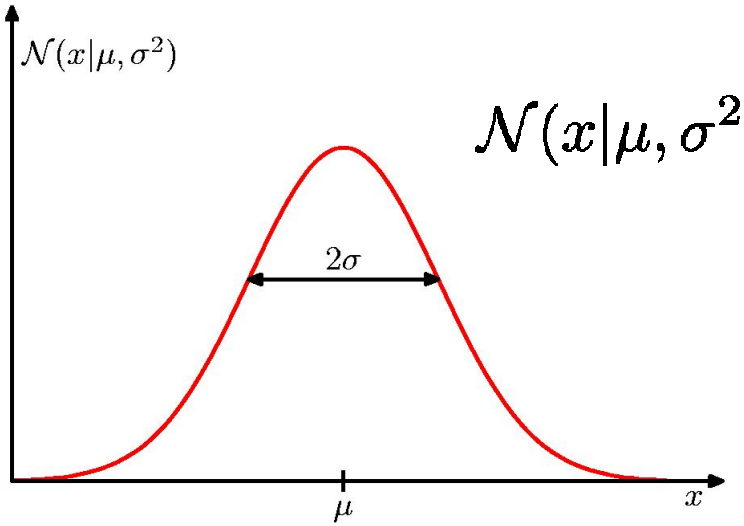
Bayesian Multinomial

Posterior distribution

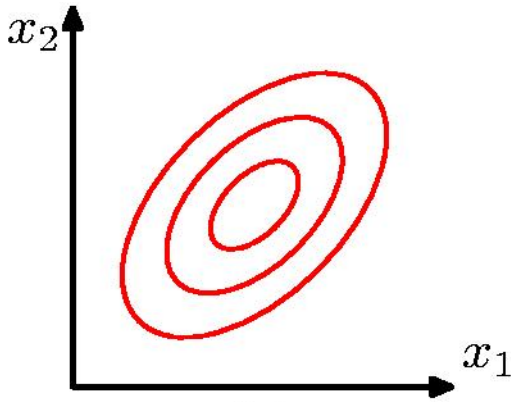
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

Gaussian Distribution



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

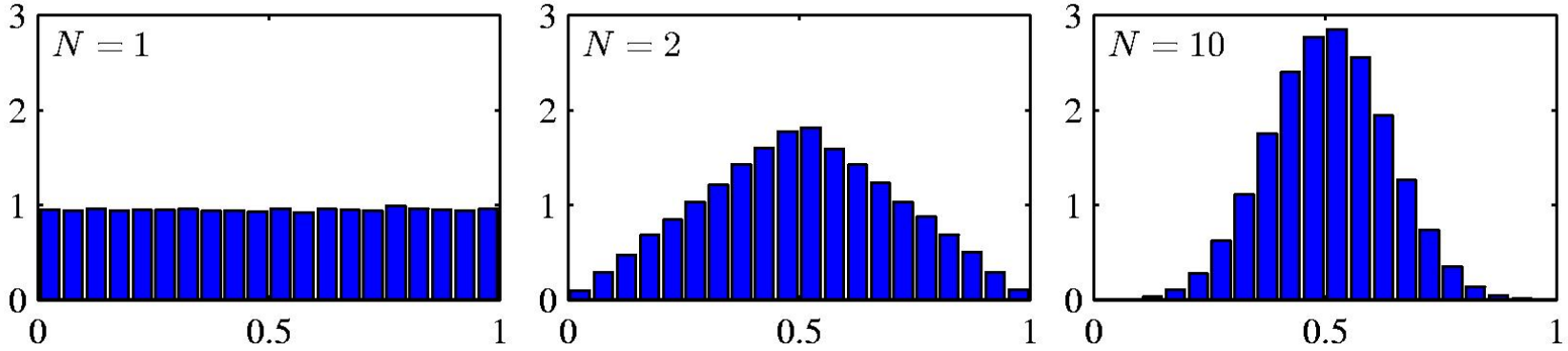


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Central Limit Theorem

The distribution of the sum (or average) of N i.i.d. random variables becomes more and more Gaussian as N grows.

Example : mean of N uniform $[0,1]$ variables



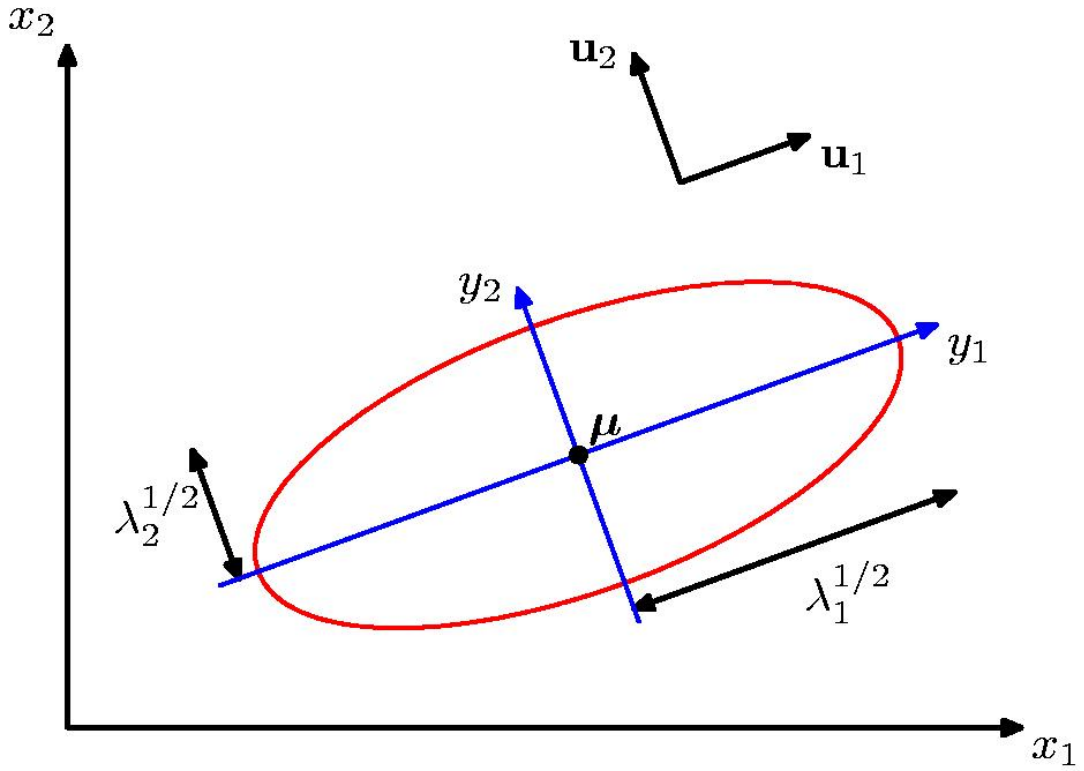
Geometry of the Multivariate Gaussian

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$

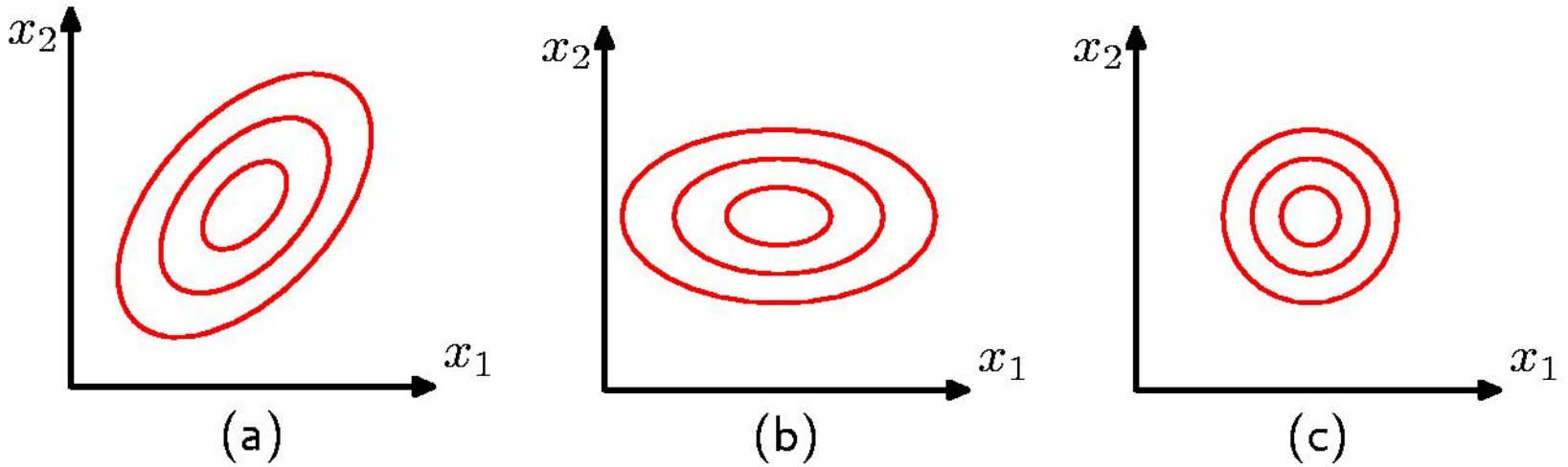


Moments of the Multivariate Gaussian

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

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

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



Partitioned Gaussian Distributions

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

precision $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

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

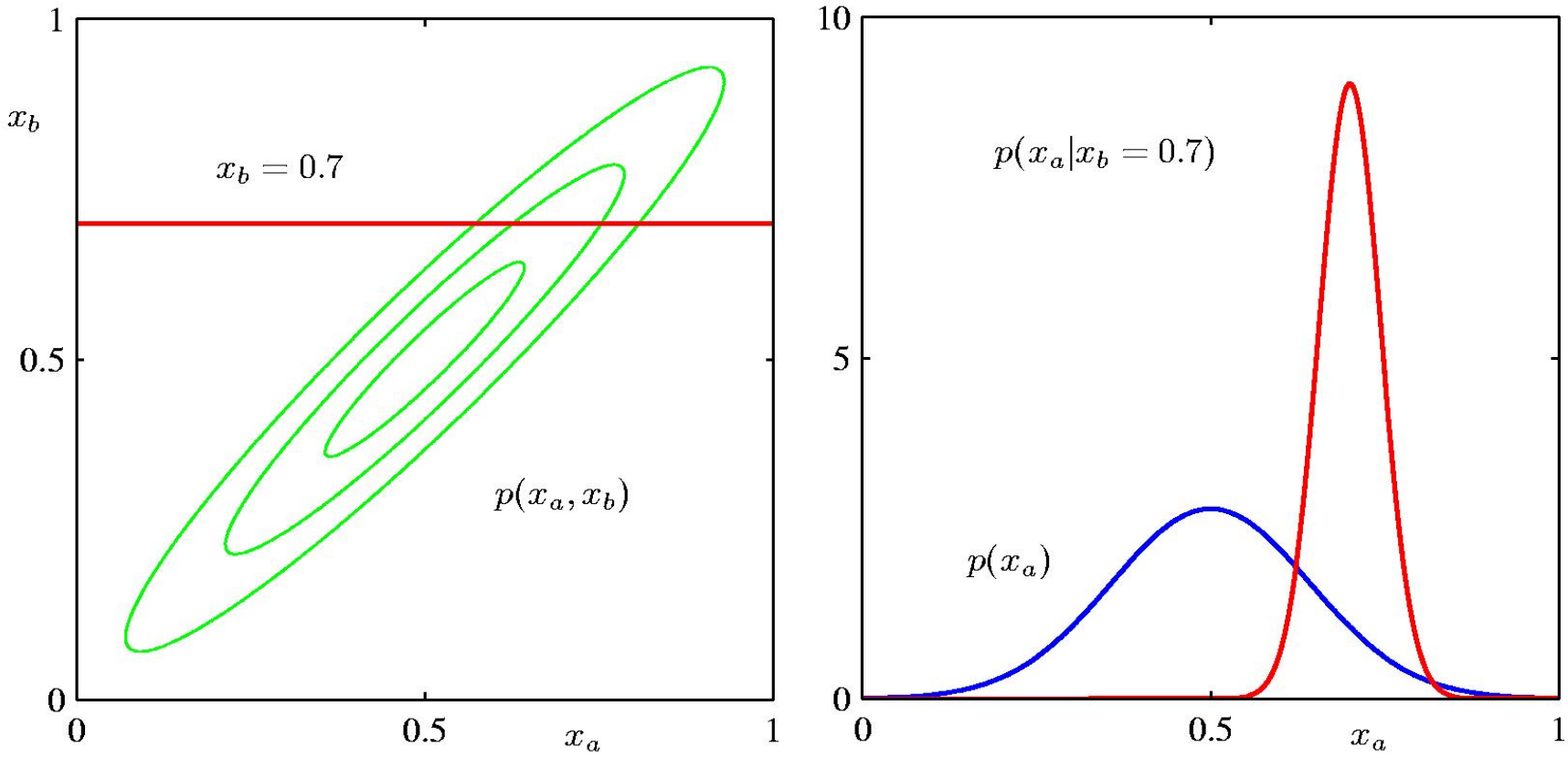
Partitioned Conditionals and Marginals

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

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

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

Partitioned Conditionals and Marginals



Maximum Likelihood for the Gaussian (1)

Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, 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 :

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

Maximum Likelihood for the Gaussian (2)

Set derivative of the log likelihood function to zero

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

solve to obtain

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

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

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

biased !

Hence define

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

Bayesian Inference for Gaussian (1)

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

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

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

Bayesian Inference for 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 Gaussian (3a)

Posterior mean

$$\mu_N = \frac{(\mu_0 / \sigma_0^2) \; + \; (\mu_{\mathrm{ML}} / \sigma_{\mathrm{ML}}^2)}{(\sigma_{\mathrm{ML}}^2 \; + \; \sigma_0^2) \; / \; (\sigma_{\mathrm{ML}}^2 \, \sigma_0^2)}$$

ML estimates

$$\mu_{\mathrm{ML}} \; = \; \sum \mathrm{x}_i \; / \, N$$

$$\sigma_{\mathrm{ML}}^2 \; = \; \sigma^2 \, / \, N$$

Posterior variance $\quad \sigma_N^2 \; = \; \dfrac{\sigma_{\mathrm{ML}}^2 \, \sigma_0^2}{\sigma_{\mathrm{ML}}^2 \; + \; \sigma_0^2}$

Posterior precision $\quad \dfrac{1}{\sigma_N^2} \; = \; \dfrac{1}{\sigma_0^2} \; + \; \dfrac{1}{\sigma_{\mathrm{ML}}^2}$

Bayesian Inference for Gaussian (3b)

$$\mu_N = \lambda \mu_0 + (1-\lambda)\mu_{\mathrm{ML}}$$

$$\sigma_N^2 = \lambda \sigma_0^2 = (1-\lambda)\sigma_{\mathrm{ML}}^2$$

"shrinkage" factor λ :

$$\lambda = \frac{\sigma_{\mathrm{ML}}^2}{\sigma_{\mathrm{ML}}^2 + \sigma_0^2}$$

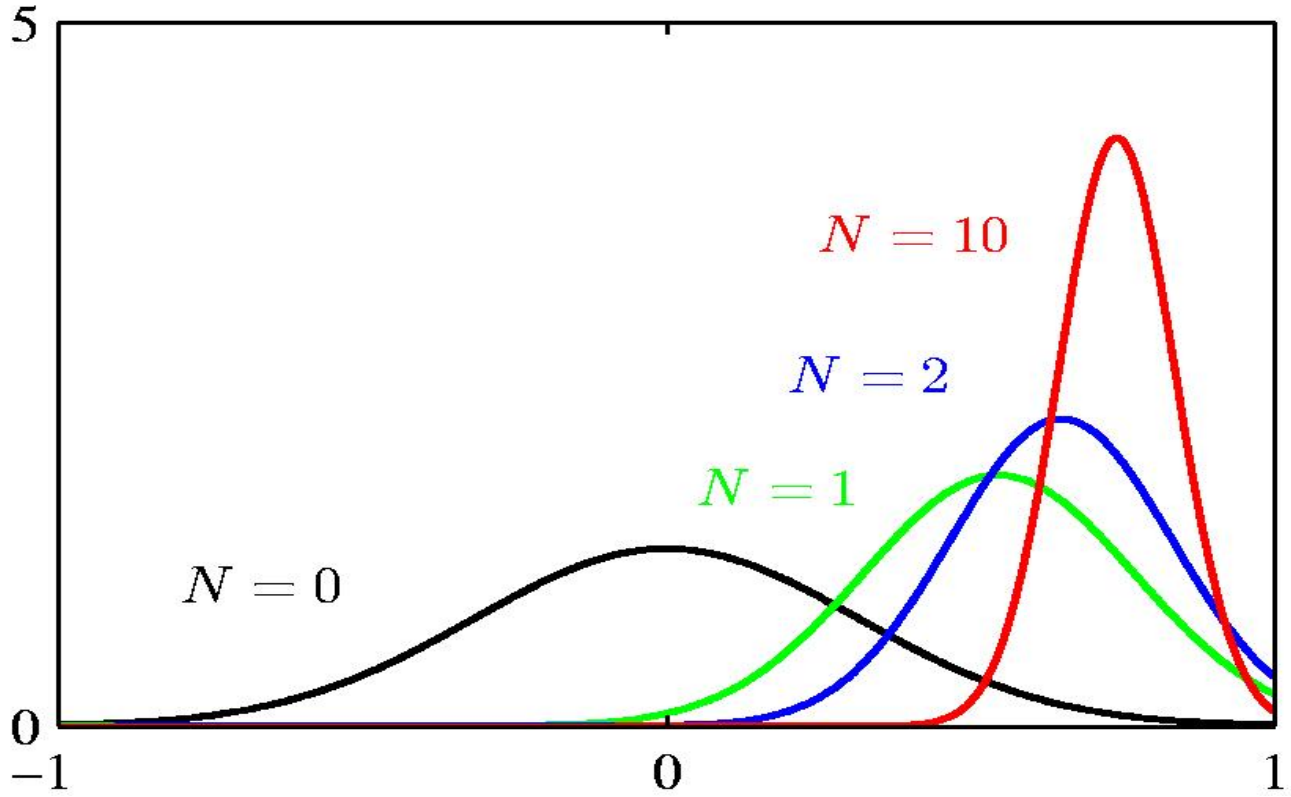
$$\lambda \in (0,1)$$

Note:

	$N=0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0
λ	1	0

Bayesian Inference for Gaussian (4)

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



Multivariate Normal Model *

Prior $\quad p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$

Likelihood $\quad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)$

Posterior of $\mathbf{x}$

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

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

Marginal of $\mathbf{y}$

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

Bayesian Inference for Gaussian (5)

Sequential Estimation

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

posterior obtained after observing $N-1$ data points becomes the prior from observing the N^{th} data point.

Bayesian Inference for Gaussian (6)

Now assume μ is known (fixed constant).

The likelihood function for $\lambda = 1/\sigma^2$ is

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

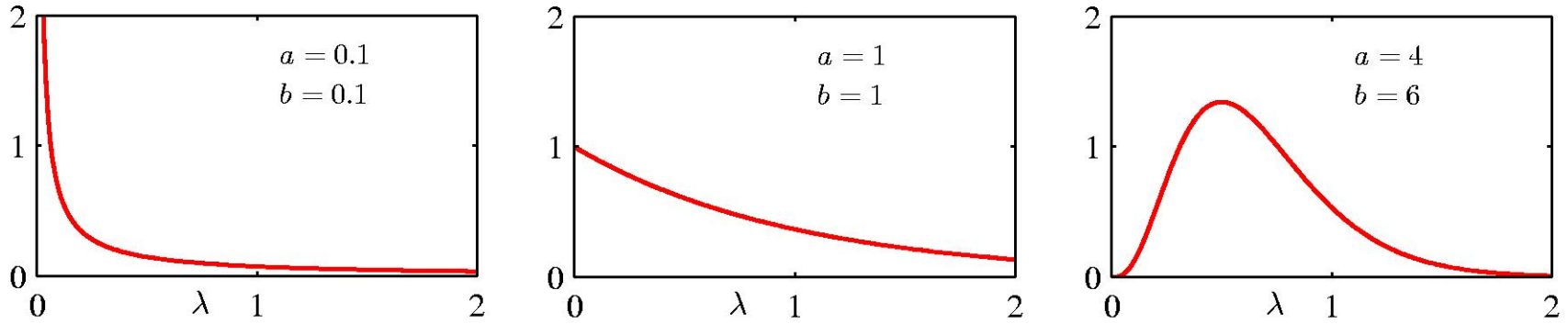
This has a Gamma shape as a function of λ.

Bayesian Inference for 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 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

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}. \end{aligned}$$

Bayesian Inference for 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 Gaussian (11)

The Gaussian-Gamma distribution

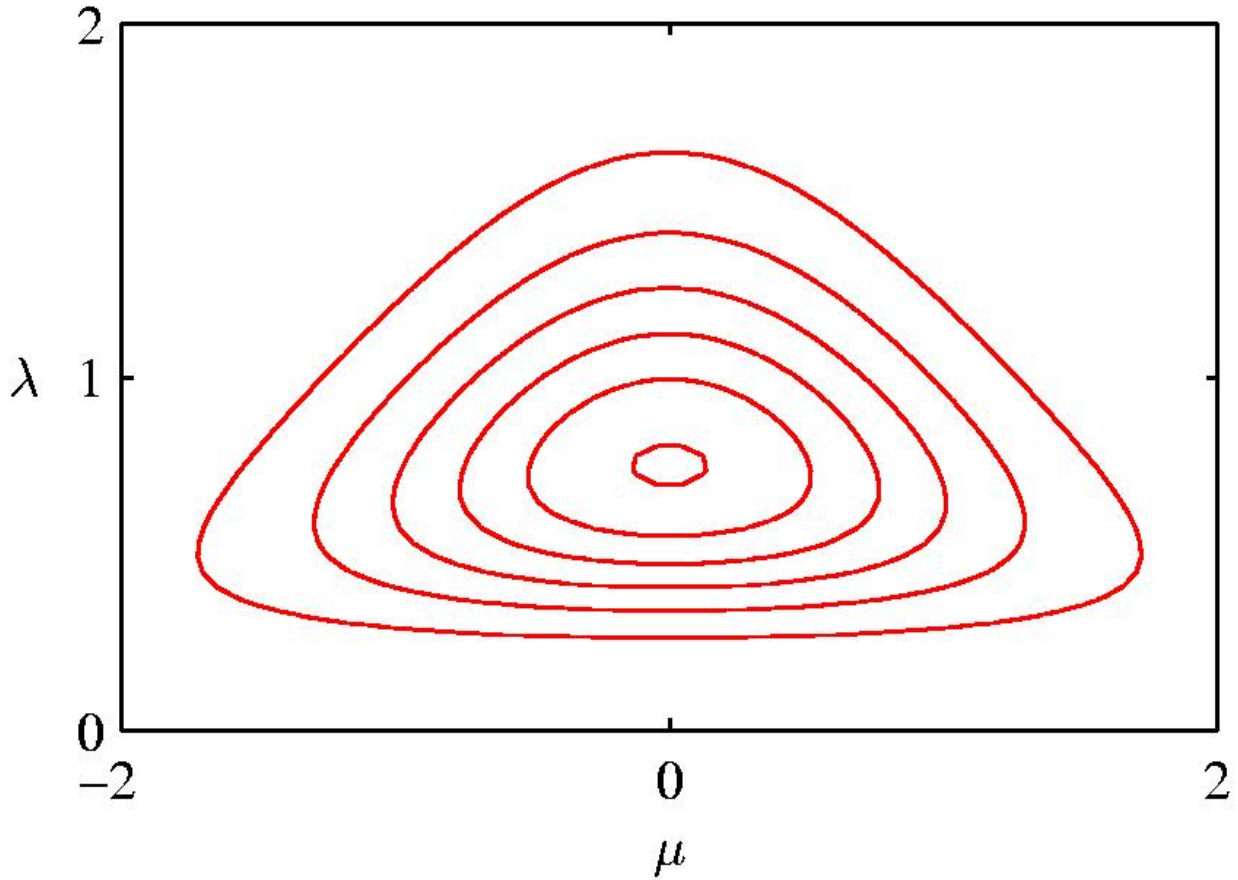
$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

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

- Quadratic in μ
- Linear in λ
- Gamma distribution over λ
- Independent of μ

Bayesian Inference for Gaussian (12)

The Gaussian-Gamma distribution



Bayesian Inference for Gaussian (10)

Multivariate conjugate priors

- $\boldsymbol{\mu}$ unknown, $\boldsymbol{\Lambda}$ known : $p(\boldsymbol{\mu})$ Gaussian
- $\boldsymbol{\Lambda}$ unknown, $\boldsymbol{\mu}$ known : $p(\boldsymbol{\Lambda})$ Wishart

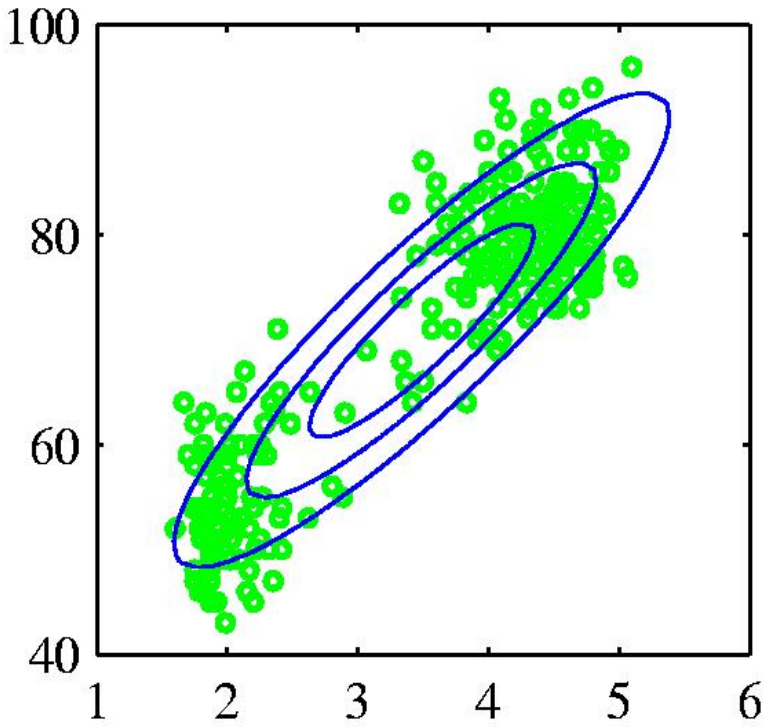
$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

- $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ unknown : $p(\boldsymbol{\mu},\boldsymbol{\Lambda})$ Gaussian-Wishart

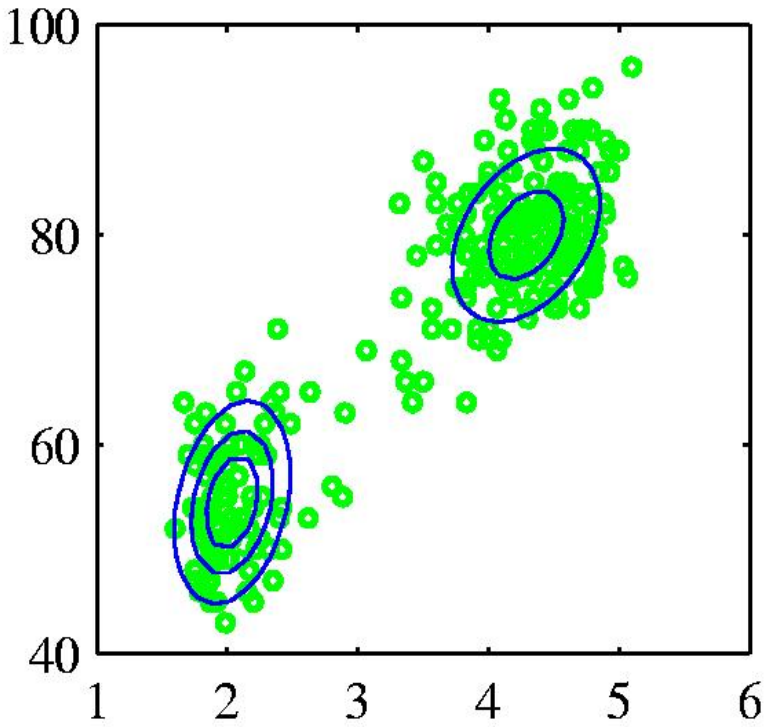
$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\boldsymbol{\Lambda})^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



Mixture of two Gaussians

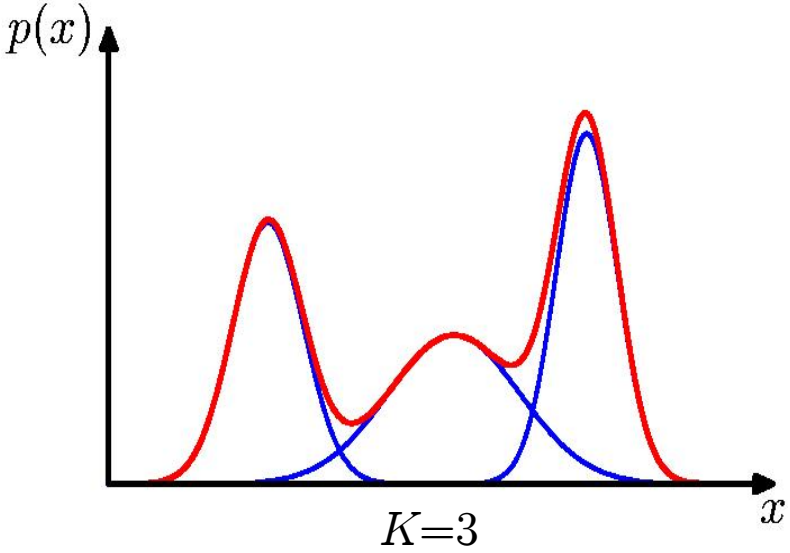
Mixtures of Gaussians (2)

Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

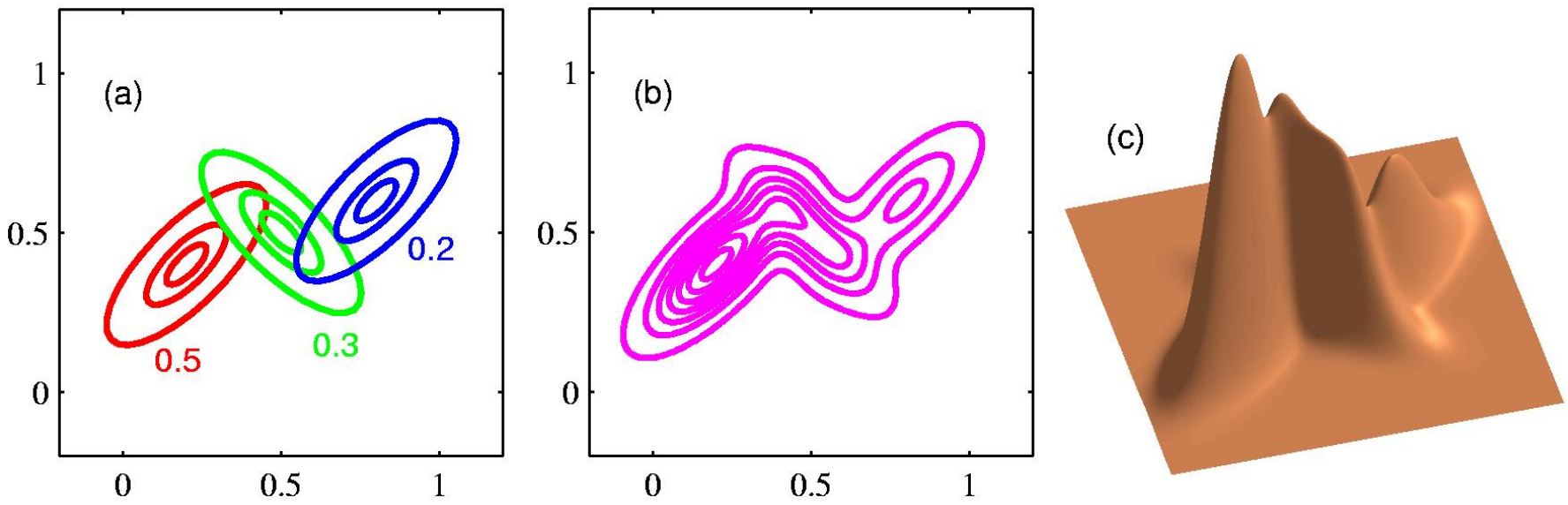
Component

Mixing coefficient



$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Gaussians (3)



Mixtures of Gaussians (4)

Determining parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\pi}$ using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Log of a sum; no closed form maximum.

Solutions: use standard, iterative, numeric optimization methods or the *Expectation Maximization* (EM) algorithm

Mixture of Conjugate Priors

- Example: build a more complex prior for Bernoulli

$$p(\theta) = \pi_1 \,\mathrm{Beta}(20,20) + \pi_2 \,\mathrm{Beta}(30,10)$$

$$D = \{\ 20 \text{ heads, } 10 \text{ tails } \}$$

Recall that a Beta prior's parameters (a,b) are *additively* updated by the observed counts (n_H, n_T) to become a Beta posterior

Posterior is also a conjugate mixture

$$p(\theta \mid D) = w_1 \,\mathrm{Beta}(40,30) + w_2 \,\mathrm{Beta}(50,20)$$

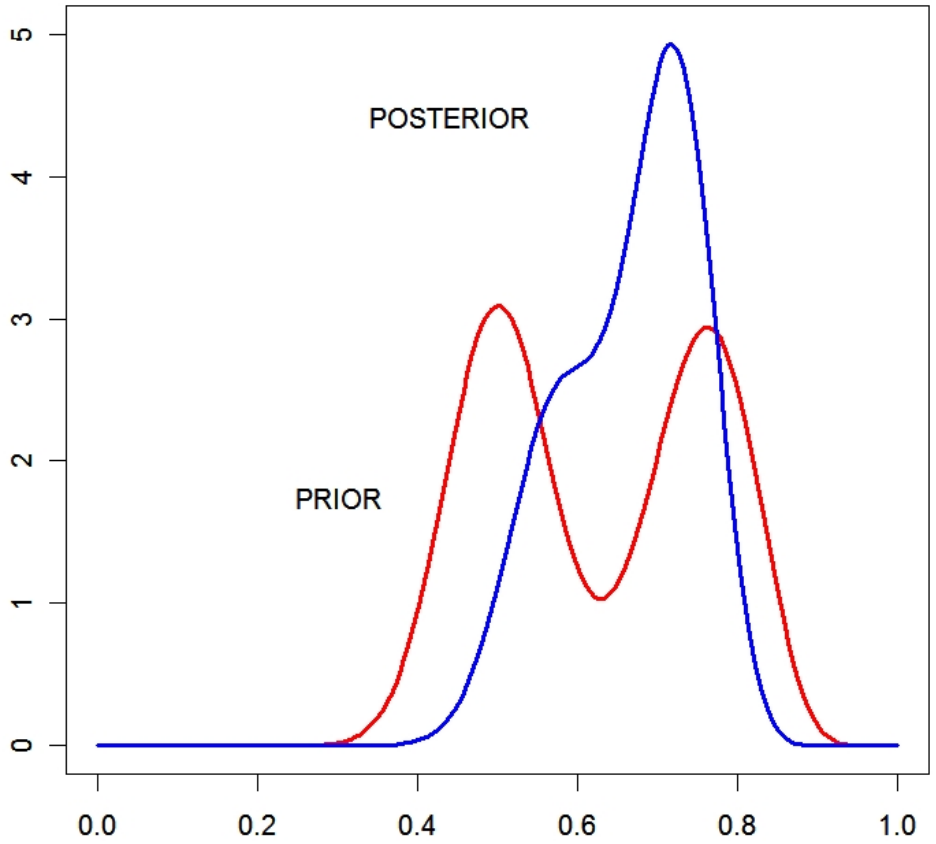
Mixture of Conjugate Priors

$$p(\theta) = \pi_1 \,\text{Beta}(20,20) + \pi_2 \,\text{Beta}(30,10)$$

$$p(\theta \mid D) = w_1 \,\text{Beta}(40,30) + w_2 \,\text{Beta}(50,20)$$

$\pi_1 = 0.5$

$\pi_2 = 0.5$



$w_1 = 0.35$

$w_2 = 0.65$

The Exponential Family (1)

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

where $\boldsymbol{\eta}$ is the *natural parameter* and

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

hence $g(\boldsymbol{\eta})$ is a normalization coefficient.

The Exponential Family (2)

Bernoulli Distribution

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

Comparing with general form we see that

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \quad \text{and so} \quad \mu = \sigma(\eta) = \underbrace{\frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}.$$

The Exponential Family (3)

The Bernoulli distribution can be written as

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1 - \sigma(\eta) = \sigma(-\eta). \end{aligned}$$

The Exponential Family (4)

Multinomial Distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right)$$

where $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$ and

$$\begin{aligned} \eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1. \end{aligned}$$

NOTE: The η_k parameters are not independent since the corresponding μ_k must satisfy

$$\sum_{k=1}^{M} \mu_k = 1.$$

The Exponential Family (5)

Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. This leads to

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{“Softmax” function}}.$$

Here the η_k parameters are independent.

Note that $0 \leqslant \mu_k \leqslant 1$ $\sum_{k=1}^{M-1} \mu_k \leqslant 1.$

The Exponential Family (6)

So multinomial distribution can be written as

$$p(\mathbf{x}|\boldsymbol{\mu}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

where

$$\begin{aligned}
\boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{M-1}, 0)^{\mathrm{T}} \\
\mathbf{u}(\mathbf{x}) &= \mathbf{x} \\
h(\mathbf{x}) &= 1 \\
g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}.
\end{aligned}$$

The Exponential Family (7)

Gaussian Distribution

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}\mu^2\right\} \\ &= h(x)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(x)\right\} \end{aligned}$$

where

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right).$$

ML for the Exponential Family (1)

Starting with $$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x} = 1$$

differentiate it

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \, \mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

to obtain

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

ML for the Exponential Family (2)

Give a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the likelihood is

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp \left\{ \boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \right\}.$$

Thus

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \underbrace{\frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}_{\text{Sufficient statistic}}$$

Exponential Family Conjugate Priors

For any member of the EF there exists a prior

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}.$$

Combining with likelihood, we get

$$p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{\nu+N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)+\nu\boldsymbol{\chi}\right)\right\}.$$

Prior corresponds to ν virtual observations of value $\boldsymbol{\chi}$

Non-Informative Priors (1)

Without *any* information pick a non-informative prior

- λ discrete, K-nomial : $p(\lambda) = 1/K.$
- $\lambda \in [a,b]$ real and bounded: $p(\lambda) = 1/b - a.$
- λ real and unbounded: improper!

A constant prior may no longer be constant after a change of variable; consider $p(\lambda)$ constant and $\lambda = \eta^2$

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{\mathrm{d}\lambda}{\mathrm{d}\eta} \right| = p_\lambda(\eta^2) 2\eta \propto \eta$$

Non-Informative Priors (2)

Translation invariant priors. Consider

$$p(x|\mu) = f(x-\mu) = f((x+c)-(\mu+c)) = f(\widehat{x}-\widehat{\mu}) = p(\widehat{x}|\widehat{\mu}).$$

For a corresponding prior over μ, we have

$$\int_A^B p(\mu)\,\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\,\mathrm{d}\mu = \int_A^B p(\mu-c)\,\mathrm{d}\mu$$

for any A and B.

Thus $p(\mu) = p(\mu - c)$ and $p(\mu)$ must be constant

Non-Informative Priors (3)

Example: For the mean of a Gaussian, μ
the conjugate prior is also a Gaussian

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

As $\sigma_0^2 \to \infty$, this will become constant over μ.

Noninformative Priors (4)

For scale invariant priors, consider $p(x|\sigma) = (1/\sigma)f(x/\sigma)$

and make the change of variable $\widehat{x} = cx$

$$p_{\widehat{x}}(\widehat{x}) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}\widehat{x}}\right| = p_x\left(\frac{\widehat{x}}{c}\right)\frac{1}{c} = \frac{1}{c\sigma}f\left(\frac{\widehat{x}}{c\sigma}\right) = p_x(\widehat{x}|\widehat{\sigma}).$$

For a corresponding prior over σ, we have

$$\int_A^B p(\sigma)\,\mathrm{d}\sigma = \int_{A/c}^{B/c} p(\sigma)\,\mathrm{d}\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right)\frac{1}{c}\,\mathrm{d}\sigma$$

Thus $p(\sigma) \propto 1/\sigma$ and so this prior is improper too. Note that this also corresponds to $p(\ln\sigma)$ being constant.

Noninformative Priors (5)

Example: For the variance of a Gaussian, σ^2, we have

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

If $\lambda = 1/\sigma^2$ and $p(\sigma) \propto 1/\sigma$, then $p(\lambda) \propto 1/\lambda$.

We know that the conjugate distribution for λ is Gamma

$$\mathrm{Gam}(\lambda|a_0,b_0) \propto \lambda^{a_0-1} \exp(-b_0\lambda).$$

Noninformative prior is obtained for $a_0 = 0$ and $b_0 = 0$.

Summary

- Parametric Distributions
- Exponential Family
- Conjugate Priors
  - equivalent to a previous virtual dataset
  - posterior form remains the same as the prior
  - parameters updated by adding sufficient statistics of data
  - conjugate *mixtures* are very useful for greater modeling power
- Non-Informative Priors
  - limit of increasingly diffuse priors (nearly flat)
  - are often viewed (by some) as being “objective”

Some things to remember

"Probability does not exist"

– Bruno de Finetti

"All models are wrong. But some are useful"

– George E. P. Box

(son in law of Ronald Fisher)