ECE285/SIO209, Machine learning for physical applications, Spring 2017

Peter Gerstoft, 534-7768, gerstoft@ucsd.edu

We meet Wednesday from 5 to 6:20pm in Spies Hall 330

Text Bishop

Grading A or maybe S

Classes

First 4 weeks: Focus on theory/implementation

Middle 3 weeks: 50% Applications plus 50% theory

Last 3 weeks: 30% Final Project, 30% Applications plus 50% theory

<u>Applications</u>

Graph theory for source localization: Gerstoft

Source tracking in ocean acoustics: Grad Student Emma Reeves

Aramco Research: Weichang Li, Group leader

Seismic network using mobile phones: Berkeley

Eric Orenstein: identifying plankton

Plus more

- Dictionary learning

-

Homework: Both matlab/python will be used, Just email the code to me (I dont need anything else).

Homework is due 11am on Wednesday. That way we can discuss in class.

Hw 1:

matlab/ python/ ipython/ jupyter ?

Tritoned? https://tritoned.ucsd.edu/

May 31 - June 2 Big Data and The Earth Sciences: Grand Challenges Workshop

by John Graham — published Feb 07, 2017 06:31 PM, last modified Mar 28, 2017 08:14 PM

Big Data and The Earth Sciences: Grand Challenges Workshop May 31 - June 2

The goal of the The Big Data and Earth Sciences: Grand Challenges Workshop is to bring thought leaders in Big Data and Earth Sciences together for a three day, intensive workshop to discuss what is needed to advance our understanding and predictability of the Earth systems and to highlight key technological advances and methods that are readily available (or will be soon) to assist this advancement. With the ever growing quantity and quality of hyper-dimensional earth science data (satellite and ground based observations and cutting–edge Numerical Weather Prediction (NWP) models), the advancements in machine learning (e.g. supervised, unsupervised and semi-supervised learning techniques), and the progress made in the application of Graphical Processing Units (GPUs) and GPU clusters, we now have an unprecedented opportunity and challenge to engage these computational advances to improve our understanding of the complex nature of the interactions between various earth science events, their variables and their impacts on society (flooding, drought, agriculture, etc.).

When	May 31, 2017 08:00 AM to Jun 02, 2017 06:00 PM
Where	Qualcomm Institute, Calit2 - University of California, (UCSD) Atkinson Hall Auditorium 9500 Gilman La Jolla, CA 92093
Contact Name	Scott L. Sellars
Add event to calendar	vCal iCal

Grand Challenges Lectures (CONFIRMED):

Dr. Larry Smarr, Founding Director of the California Institute for Telecommunications and Information Technology (Calit2), a UC San Diego/UC Irvine partnership, holds the Harry E. Gruber professorship in Computer Science and Engineering (CSE) at UC San Diego's Jacobs School.

Dr. Vipin Kumar, Regents Professor at the University of Minnesota, holds the William Norris Endowed Chair in the Department of Computer Science and Engineering, University of Minnesota.

Dr. Padhraic Smyth, Professor, Director, UCI Data Science Initiative and Associate Director, Center for Machine Learning and Intelligent Systems, UC Irvine.

Dr. Michael Wehner, Senior Staff Scientists, Computational Research Division at the Lawrence Berkeley National Laboratory.

Hotel accommodations:

- **La Jolla Sheraton** (nice, economical, close by): http://www.sheratonlajolla.com/
- **Estancia** (Closest location and most beautiful): http://meritagecollection.com/estancialajolla/
- **La Jolla Shores** (beach front property - farther away): http://www.ljshoreshotel.com/?src=ppc_google_ljshores_brand_expanded&NCK=ppc_google_ljshores_brand&gclid=CNTF8JTmqdICFQmlfgodfTMP7A

Please send abstracts to scottsellars@ucsd.edu

Please register here: Workshop Registration Form

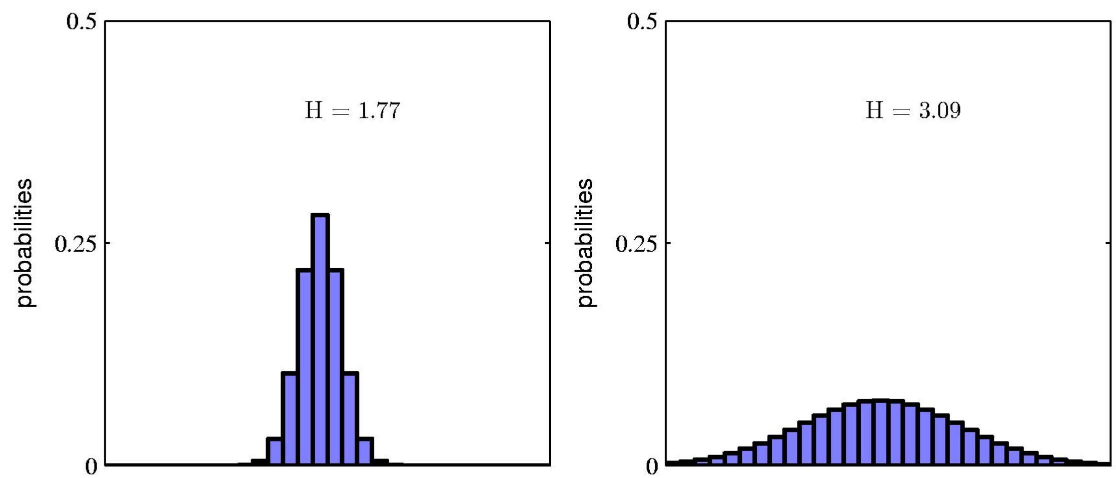
Download the call for papers HERE

Entropy

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning

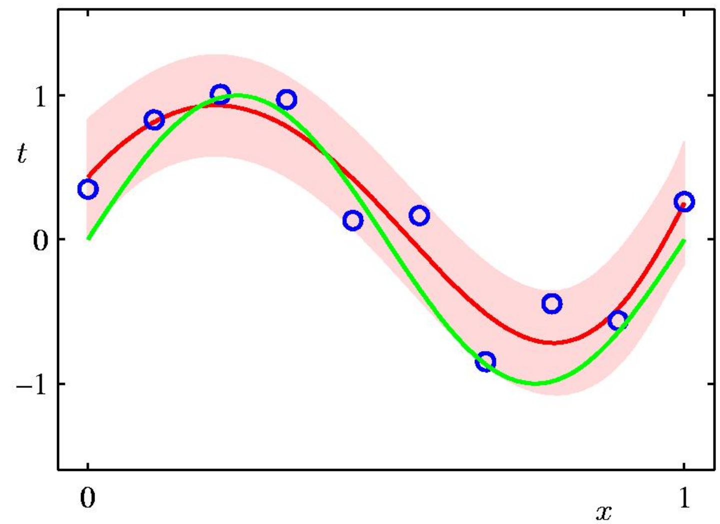


Parametric Distributions

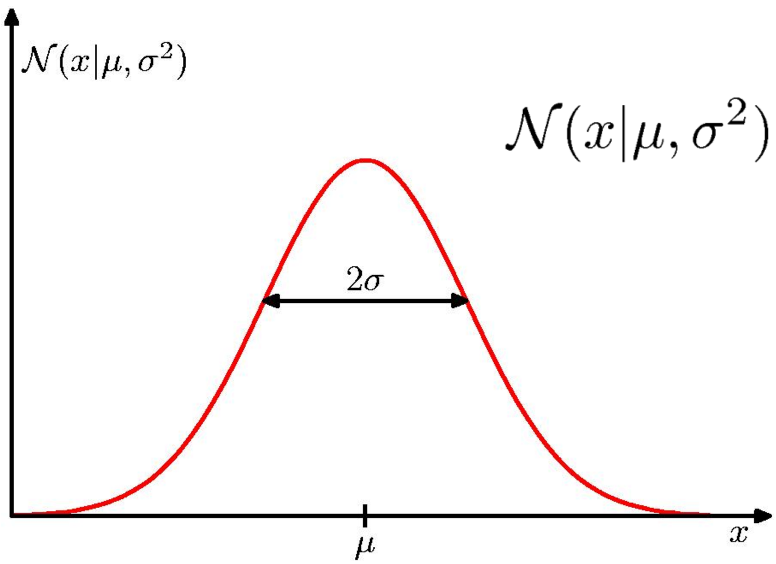
- Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$
- Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Representation: $\boldsymbol{\theta}^\star$ or $p(\boldsymbol{\theta})$

- Recall Curve Fitting

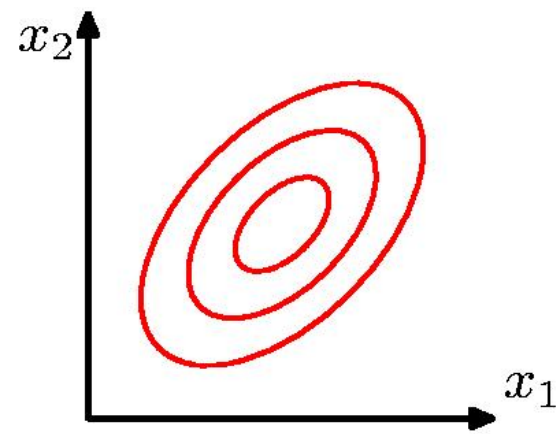
$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w}$$



The Gaussian Distribution



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

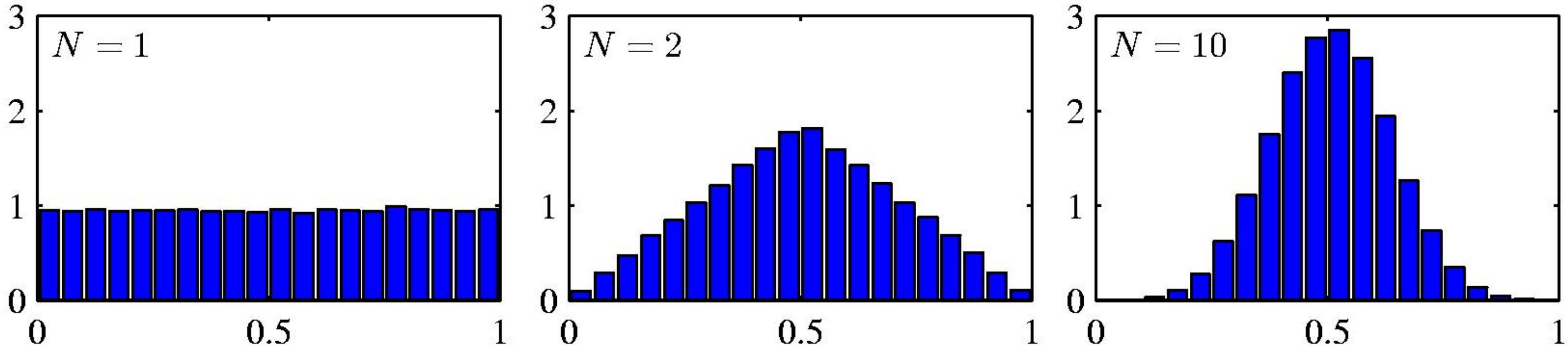


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Central Limit Theorem

•The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.

•Example: N uniform $[0,1]$ random variables.



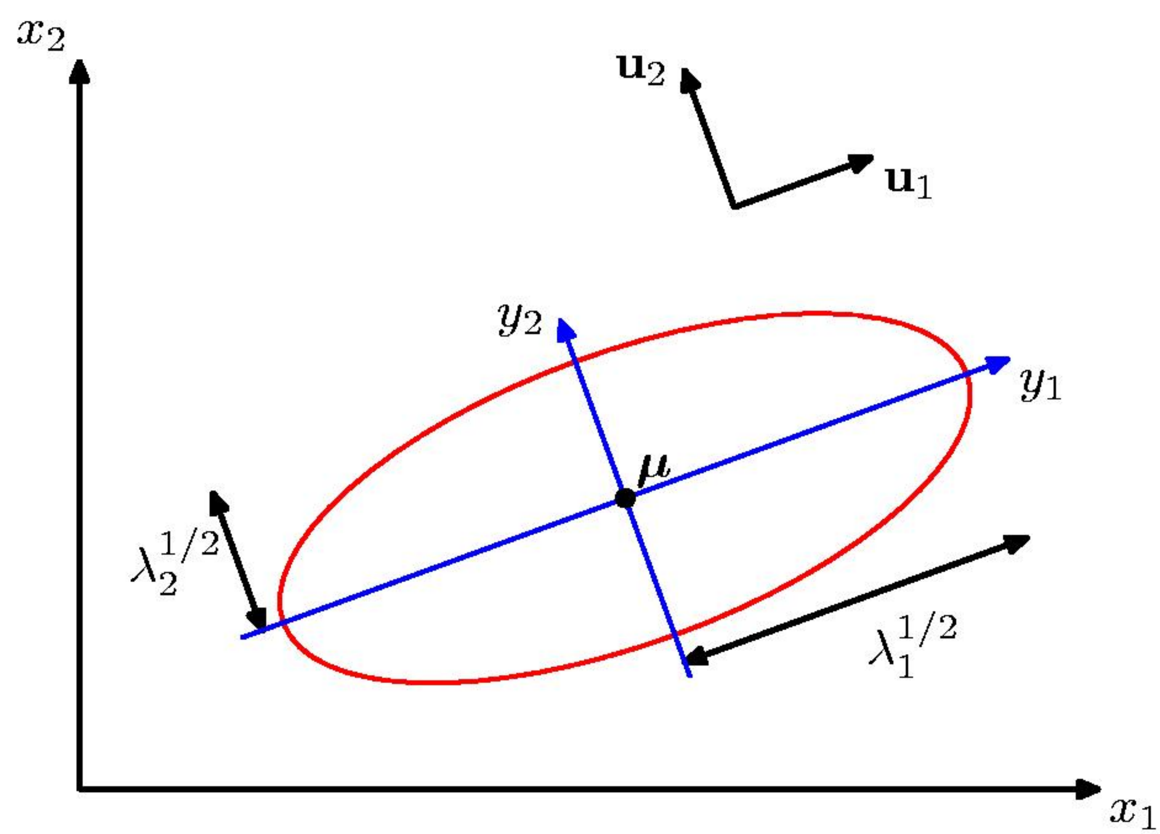
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

thanks to anti-symmetry of z

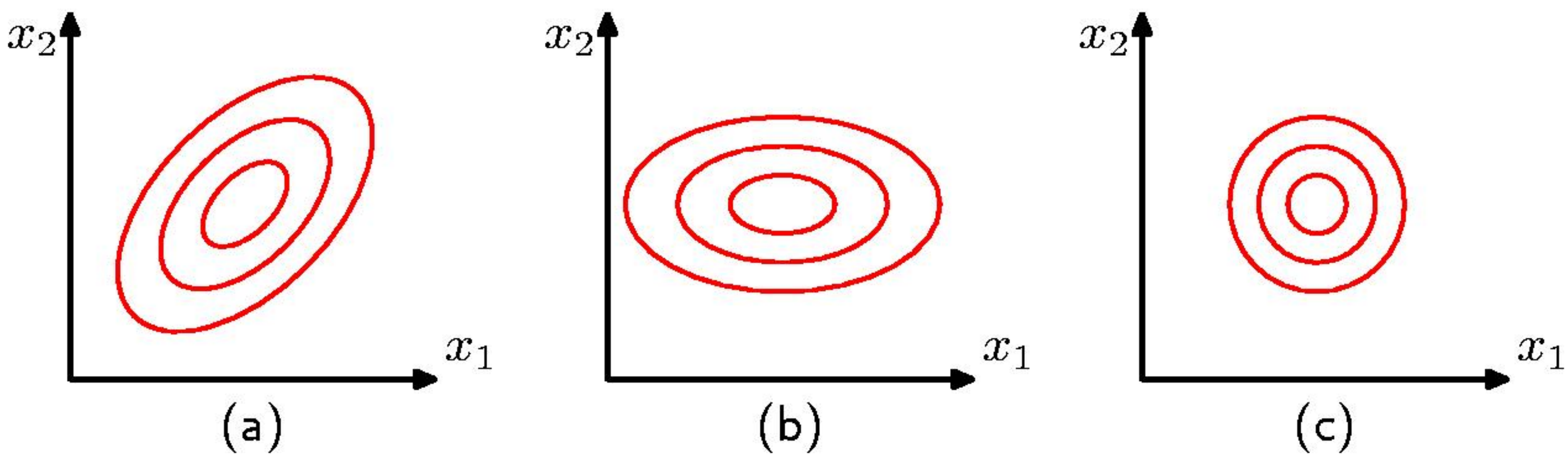
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Multivariate Gaussian (2)

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

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

A Gaussian requires D*(D-1)/2 +D parameters.
Often we use D +D or
Just D+1 parameters.



Partitioned Gaussian Distributions

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

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

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

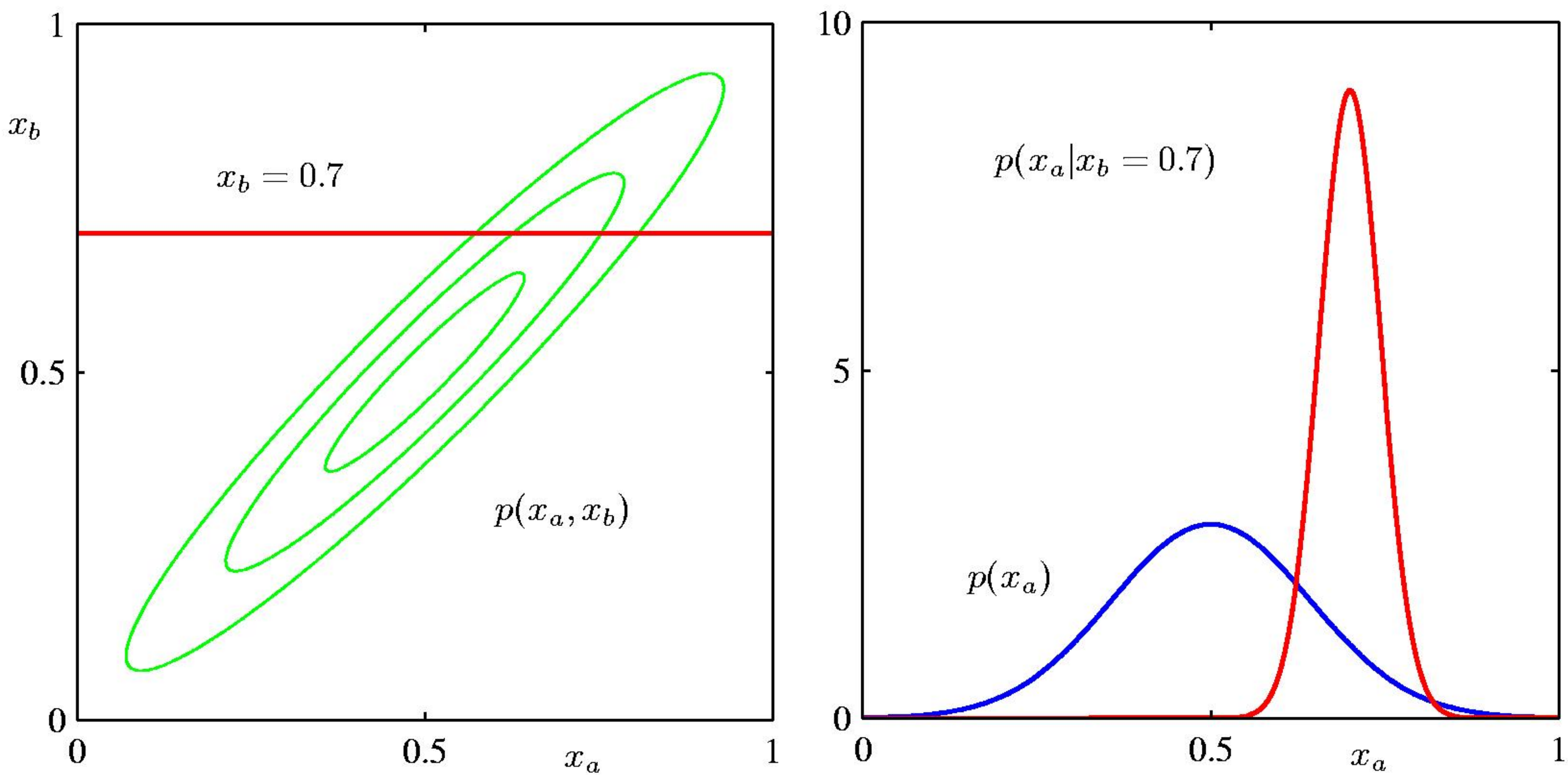
Partitioned Conditionals and Marginals

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

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

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

Partitioned Conditionals and Marginals



Bayes' Theorem for Gaussian Variables

- Given

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

- we have

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

- where

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

Maximum Likelihood for the Gaussian (1)

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, the log likeli-hood 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 the derivative of the log likelihood function to zero,

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

- and solve to obtain

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

- Similarly

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

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

Hence define

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

Sequential Estimation

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

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

Bayesian Inference for the Gaussian (1)

- 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

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

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

Bayesian Inference for 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

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

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. \end{aligned}$$

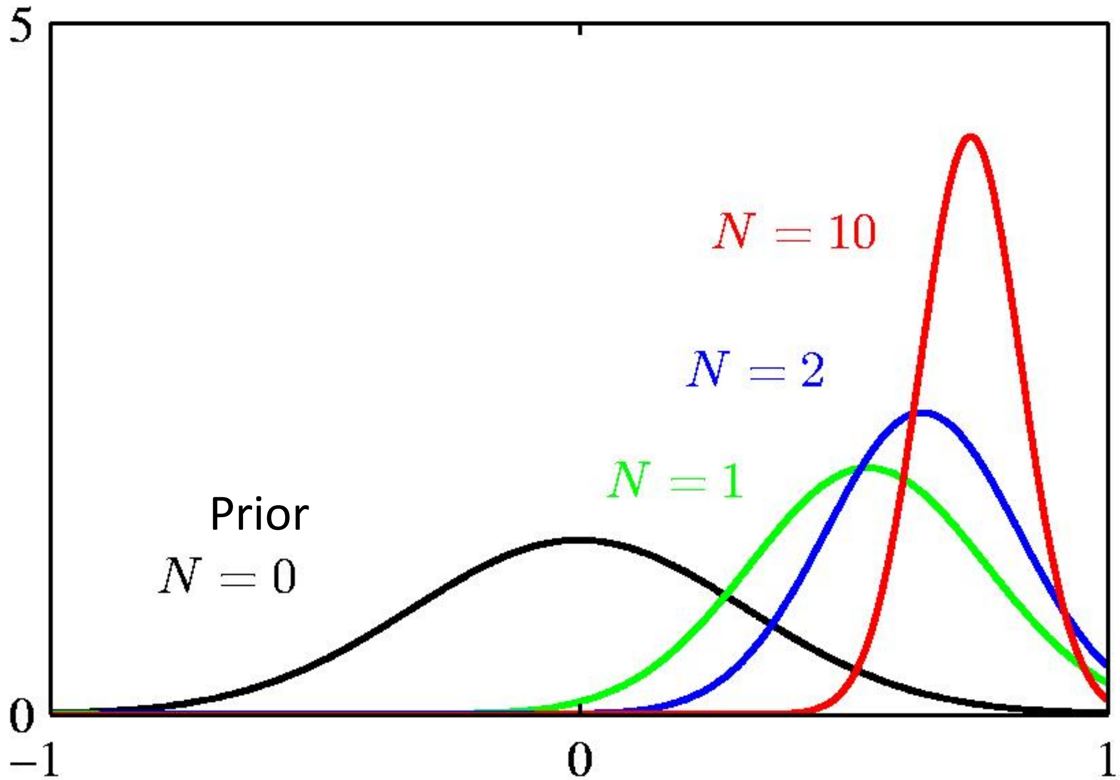
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Bayesian Inference for the Gaussian (4)

- Example: for N = 0, 1, 2 and 10.

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



Bayesian Inference for the Gaussian (5)

- Sequential Estimation

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

- The posterior obtained after observing $\mathrm{N} \{ 1$ data points becomes the prior when we observe the N^{th} data point.

Bayesian Inference for the Gaussian (6)

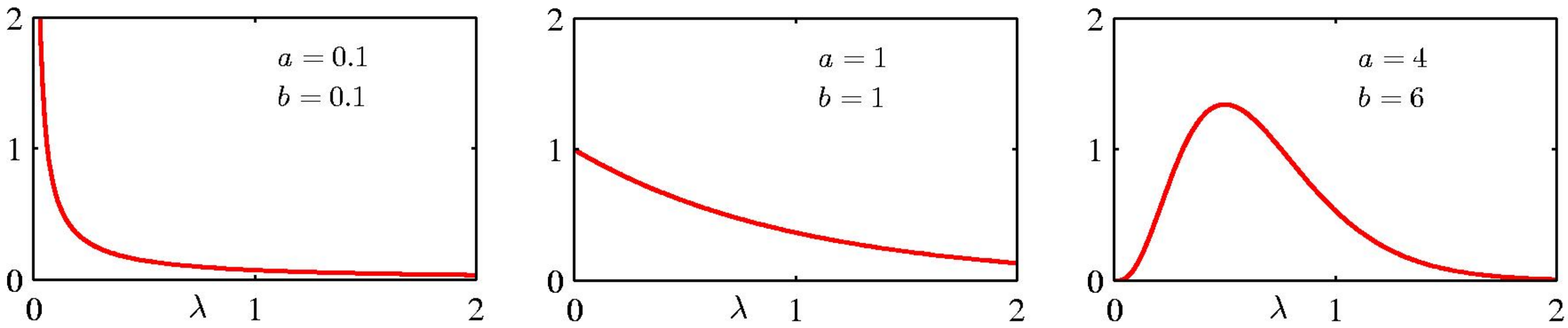
- Now assume $\boldsymbol{\mu}$ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

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

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

- The Gamma distribution:

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \qquad \mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



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

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

Bayesian Inference for the Gaussian (9)

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

$$
\begin{aligned}
p(\mathbf{x}|\mu,\lambda) &= \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\} \\
&\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.
\end{aligned}
$$

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

Bayesian Inference for the Gaussian (10)

- The Gaussian-gamma distribution

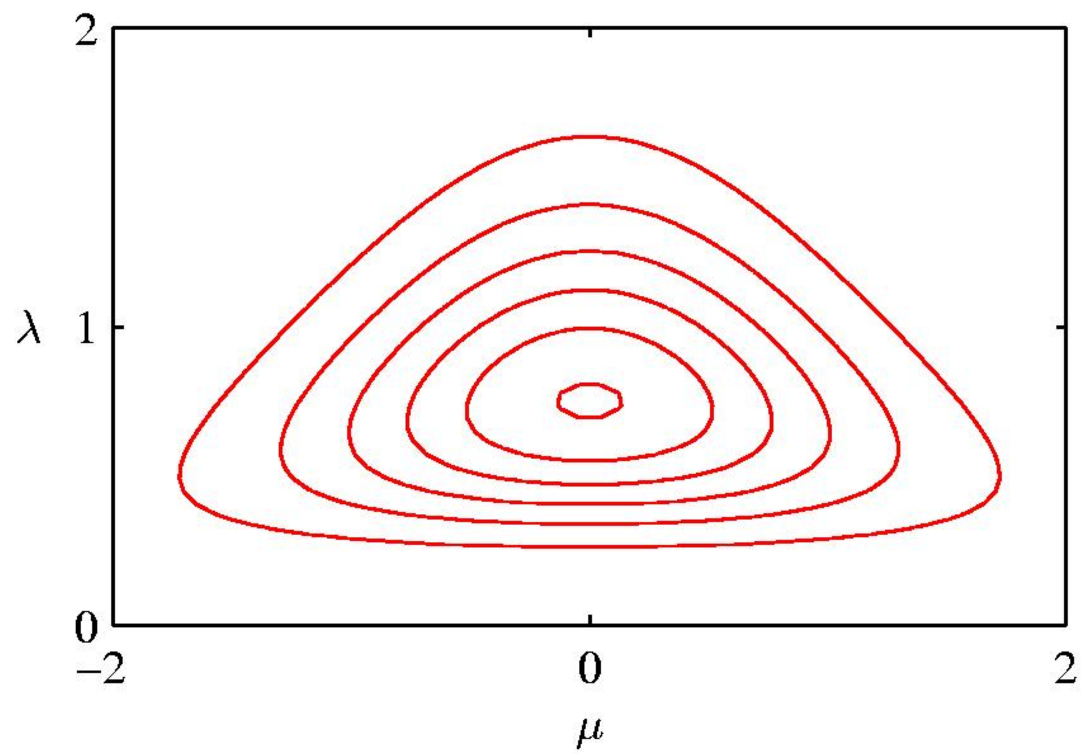
$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

$$\propto \quad \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}}_{\text{• Quadratic in } \mu. \text{ • Linear in } \lambda.} \underbrace{\lambda^{a-1}\exp\{-b\lambda\}}_{\text{• Gamma distribution over } \lambda. \text{ • Independent of } \mu.}$$

• Quadratic in μ.
• Linear in λ.

• Gamma distribution over λ.
• Independent of μ.

$\mu_0=0$, $\beta=2$, a=5, b=6



Bayesian Inference for the Gaussian (12)

- Multivariate conjugate priors
- $\boldsymbol{\mu}$ unknown, $\boldsymbol{\Lambda}$ known: $\mathrm{p}(\mu)$ Gaussian.
- $\boldsymbol{\Lambda}$ unknown, μ known: $\mathrm{p}(\boldsymbol{\Lambda})$ Wishart,

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

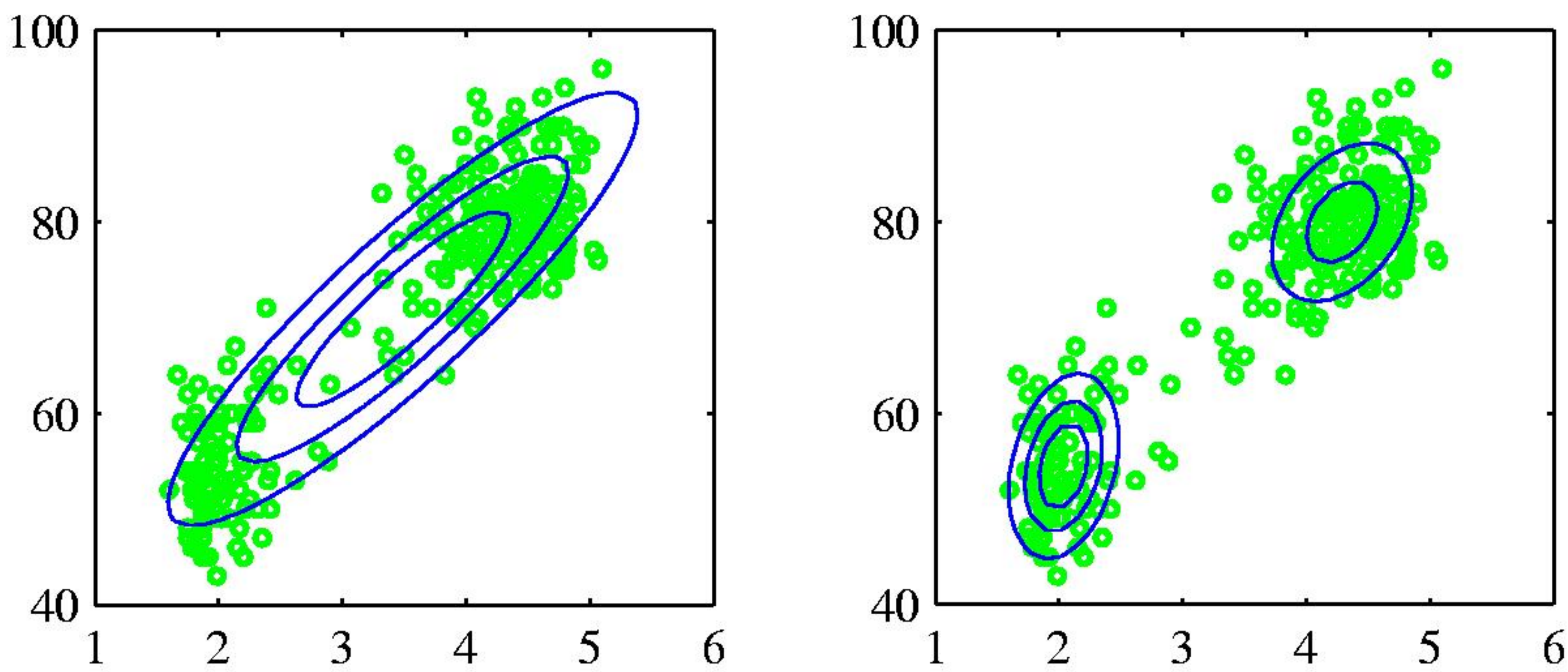
- $\boldsymbol{\Lambda}$ and μ unknown: $\mathrm{p}(\mu, \boldsymbol{\Lambda})$ Gaussian-Wishart,

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

Mixtures of Gaussians (1)

Old Faithful geyser:

The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than 2 ½ minutes, or 91 minutes after an eruption lasting more than 2 ½ minutes.



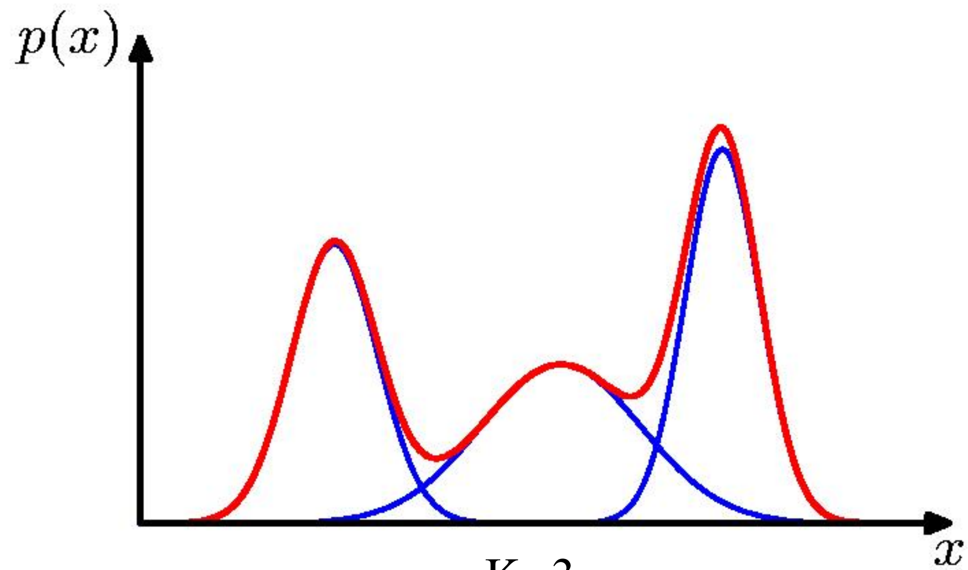
Single Gaussian

Mixture of two Gaussians

Mixtures of Gaussians (2)

- Combine simple models into a complex model:

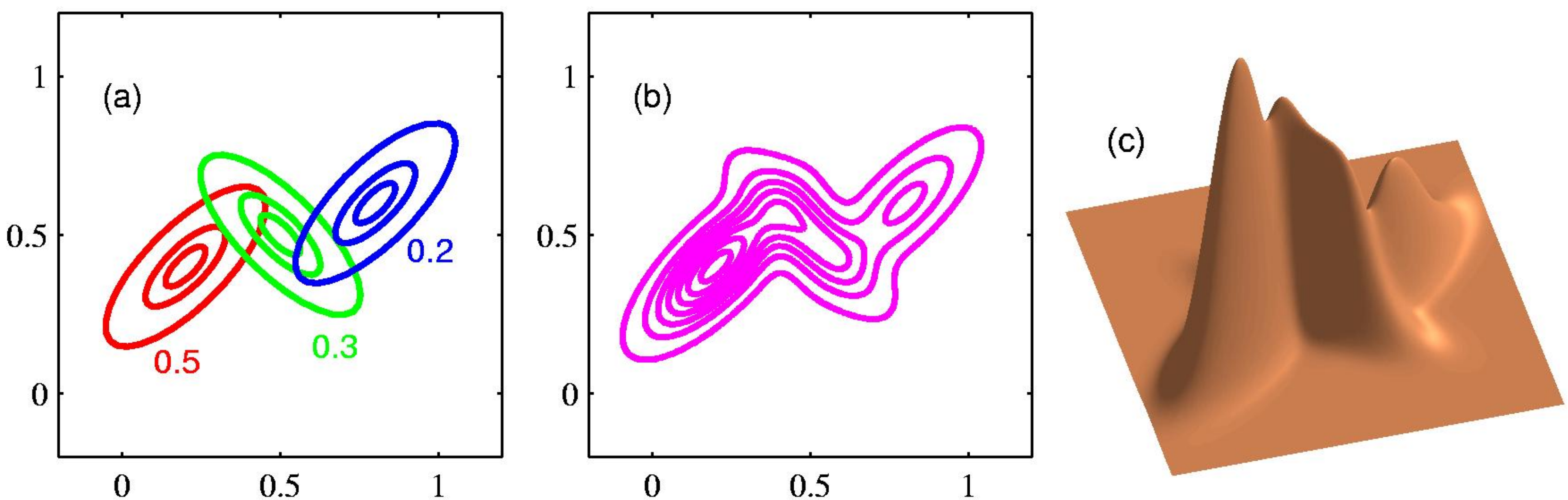
$$p(\mathbf{x}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{Mixing coefficient}} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$



K=3

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

Mixtures of Gaussians (3)



Mixtures of Gaussians (4)

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

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}}_{\text{Log of a sum; no closed form maximum.}}$$

- Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).