Probability Distributions

Chapter 2

Prof. Reza Azadeh

University of Massachusetts Lowell

Binary Variables

Consider a single binary random variable $x \in \{0, 1\}$. For example x might describe the outcome of flipping a coin with heads=1, tails=0. The probability of landing heads is

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

where $0 \leq \mu \leq 1$.
We can then write

$$p(x = 0|\mu) = 1 - \mu$$

Bernoulli Distribution

The probability distribution over x can be described using the Bernoulli distribution as

$$\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

The mean and the variance are given by

$$\mathbb{E}[x] = \mu$$
$$\mathbb{V}[x] = \mu(1-\mu)$$

Binomial Distribution

For N coin flips, we can write the probability of observing m heads as

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

This can be described using the Binomial distribution as

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

where

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

Binomial Distribution (2)

The mean and the variance of the Binomial distribution are given by

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

$$\mathbb{V}[m] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2\mathrm{Bin}(m|N\mu) = N\mu(1-\mu).$$

$\rightarrow$ Note that Binomial distribution is used to describe an event with a binary outcome (success/failure) when we perform N independent trials and each trial has the same probability of success, μ.

Binomial Distribution - Example

Suppose we roll a dice 6 times. What is the probability of rolling a 6 three times? In this example, number of trials $N = 6$, and the probability of success, $\mu = 1/6$ and $m = 3$. We can calculate the probability of the event using the definition of PMF:

$$P(X = m) = \binom{N}{m} \mu^m (1-\mu)^{N-m} = \binom{6}{3} (\frac{1}{6})^3 (\frac{5}{6})^3 = 0.053$$

Binomial Distribution - Example

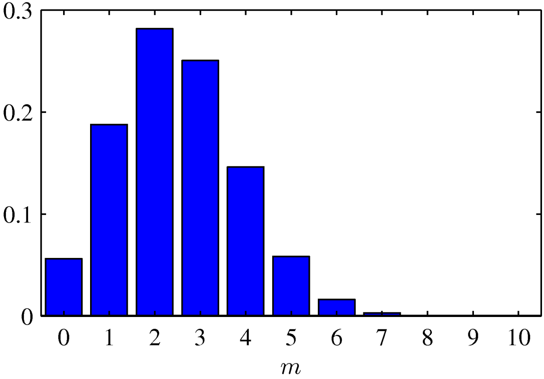


Figure: Histogram plot of the binomial distribution as a function of m for $N = 10$ and $\mu = 0.25$.

Parameter Estimation

Now suppose we have observed N coin flips and have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$. We can write the likelihood function on the assumption that the observations are drawn independently from $p(x|\mu)$

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

The log-likelihood then can be written as

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

Parameter Estimation (2)

If we set the derivative of $\ln p(\mathcal{D}|\mu)$ w.r.t μ equal to zero, we obtain

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

If we have observed m heads in our experiment, then we have $\mu_{ML} = \frac{m}{N}$, which is basically the fraction of observing heads over the total number of samples.
Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{ML} = \frac{3}{3} = 1$. Prediction: all future tosses will land heads up. This is the **overfitting** issue of the maximum likelihood to the data set.

Beta Distribution

The Beta distribution of $\mu \in [0,1]$ is given by

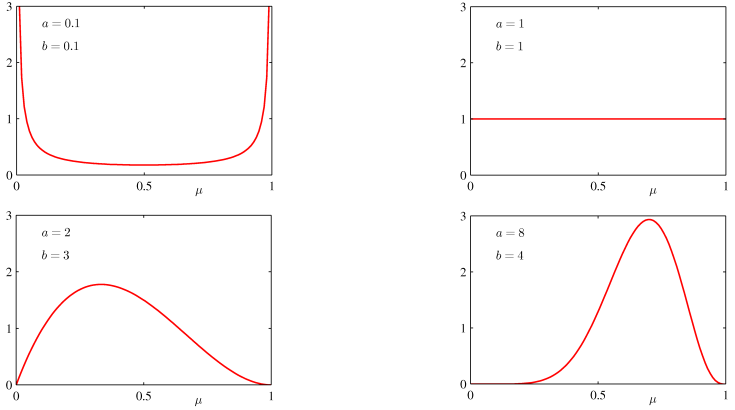
$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1},$$

where $\Gamma(x)$ is the gamma function. The mean and variance of the Beta distribution are given by

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

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

Beta Distribution (2)



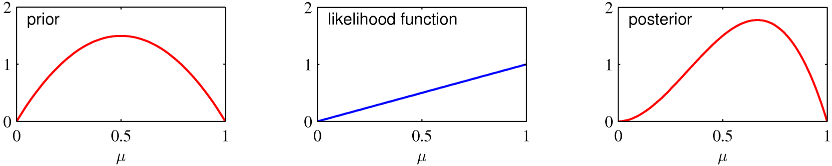
Bayesian Bernoulli

Instead of performing maximum likelihood, we can go one step further and develop a Bayesian formulation

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

with $a_N = a_0 + m$ and $b_N = b_0 + (N-m)$.
The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

Bayesian Inference



One step of sequential Bayesian inference. The prior is a Beta distribution with $a = 2$ and $b = 2$, and the likelihood is a Binomial distribution with $N = m = 1$, and applying Bayes' formula corresponds to a posterior as a Beta distribution with $a = 3$ and $b = 2$.

Properties of the Posterior

As the size of the data set, N, increases

$$\begin{aligned}
a_N &\to m \\
b_N &\to N - m \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \to \frac{m}{N} = \mu_{ML} \\
\mathbb{V}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \to 0
\end{aligned}$$

Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

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

Multinomial Variables

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

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

$$\forall\, k : \mu_k \geq 0 \text{ and } \sum_{k=1}^{K} \mu_k = 1$$

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

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

ML Parameter Estimation

Given $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ the likelihood can be written as

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

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

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

We can see that $\mu_k = -m_k/\lambda$ and so $\mu_k^{\mathrm{ML}} = m_k/N$.

Multinomial Distribution

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

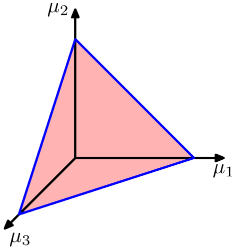
$$\begin{aligned} \mathbb{E}[m_k] &= N\mu_k \\ \mathbb{V}[m_k] &= N\mu_k(1-\mu_k) \\ \mathrm{Cov}[m_j m_k] &= -N\mu_j\mu_k \end{aligned}$$

Dirichlet Distribution

The Dirichlet distribution is conjugate prior for the multinomial distribution.

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

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



Bayesian Multinomial (1)

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

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

Bayesian Multinomial (2)

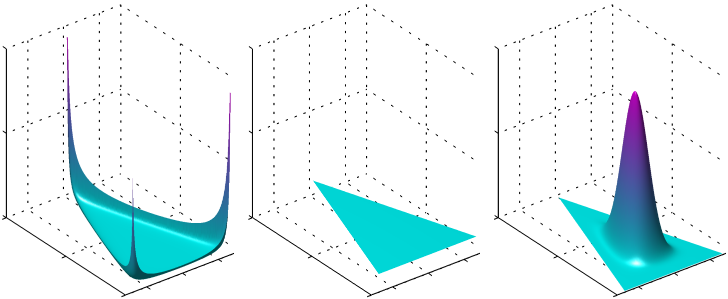
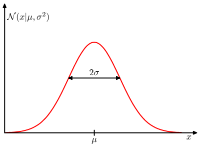


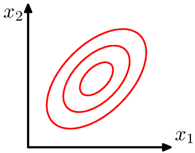
Figure: (left) $\alpha_k = 10^{-1}$, (middle) $\alpha_k = 10^{0}$, (right) $\alpha_k = 10^{1}$

Gaussian Distribution

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

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





(a)

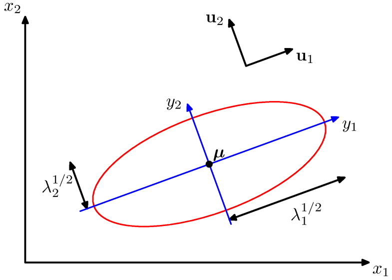
Geometry of Multivariate Gaussian

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

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

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

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



Moments of Multivariate Gaussian

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

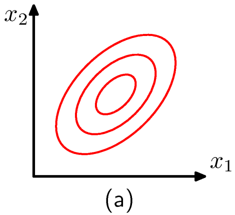
Thanks to anti-symmetry of $\mathbf{z}$

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

Moments of Multivariate Gaussian

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

$$\mathrm{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top] = \Sigma$$



Partitioned Gaussian Distributions

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

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

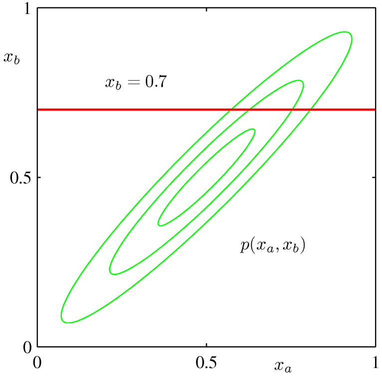
$$\Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

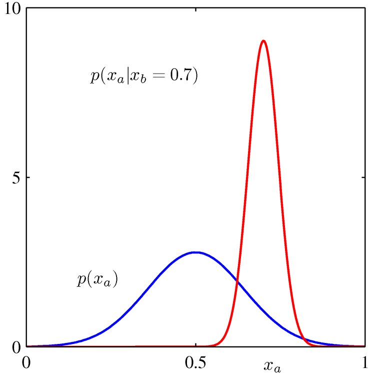
Partitioned Conditional and Marginals

$$
\begin{aligned}
p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \Sigma_{a|b}) \\
\Sigma_{a|b} &= \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \\
\boldsymbol{\mu}_{a|b} &= \Sigma_{a|b}\{\Lambda_{aa}\boldsymbol{\mu}_a - \Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\
&= \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}
$$

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

Partitioned Conditionals and Marginals





Bayes' Theorem for Gaussian Variables

Given

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

We have

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

where

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

Maximum Likelihood for the Gaussian (1)

Given i.i.d data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$, the log likelihood function is given by

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

Sufficient statistics

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

Maximum Likelihood for the Gaussian (2)

Set the derivative of the log likelihood function to zero

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

and solve to obtain

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

Similarly

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$
$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\Sigma$$

To make it unbiased, we define

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

Sequential Estimation

Contribution of the N^{th} data point, $\mathbf{x}_N$

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

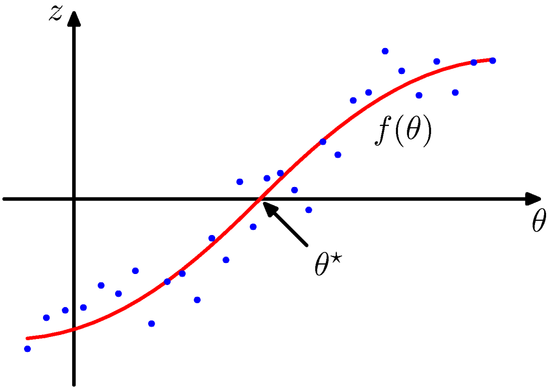
The Robbins-Monro Algorithm (1)

Consider θ and z governed by $p(z, \theta)$ and define the regression function

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int z p(z|\theta) dz$$

Seek θ^* such that $f(\theta^*) = 0$.

The Robbins-Monro Algorithm (2)



Assume we are given samples from $p(z, \theta)$, one at the time.

The Robbins-Monro Algorithm (2)

Successive estimates of θ^* are then given by

$$\theta^N = \theta^{N-1} - a_{N-1} z(\theta^{N-1})$$

Conditions on a_N for convergence:

$$\lim_{N \to \infty} = 0, \quad \sum_{N=1}^{\infty} a_N = \infty, \quad \sum_{N=1}^{\infty} a_N^2 < \infty$$

Robbins-Monro for Maximum Likelihood (1)

Regarding

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta)=\mathbb{E}_x[-\frac{\partial}{\partial\theta}\ln p(x_n|\theta)]$$

as a regression function, finding its root is equivalent to finding the maximum likelihood solution θ_{ML}. Thus

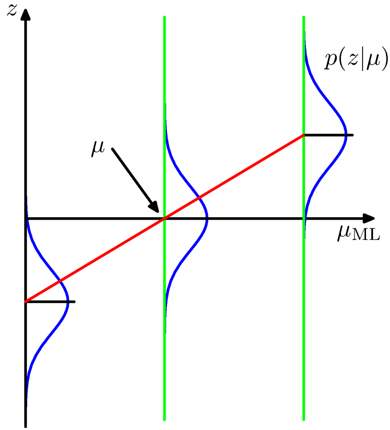
$$\theta^{(N)}=\theta^{(N-1)}-a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}[-\ln p(x_N|\theta^{(N-1)})].$$

Robbins-Monro for Maximum Likelihood (2)

Example: estimate the mean of a Gaussian.

$$z = \frac{\partial}{\partial \mu_{\mathrm{ML}}}[-\ln p(x|\mu_{\mathrm{ML}}, \sigma^2)]$$
$$= -\frac{1}{\sigma^2}(x - \mu_{\mathrm{ML}})$$

The distribution of $z$is Gaussian with mean $\mu - \mu_{\mathrm{ML}}$. For the Robbins-Monro update equation, $a_N = \sigma^2/N$.



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

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

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,

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

this gives the posterior

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

Completing the square over μ, we see that

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

Bayesian Inference for the Gaussian (3)

... where

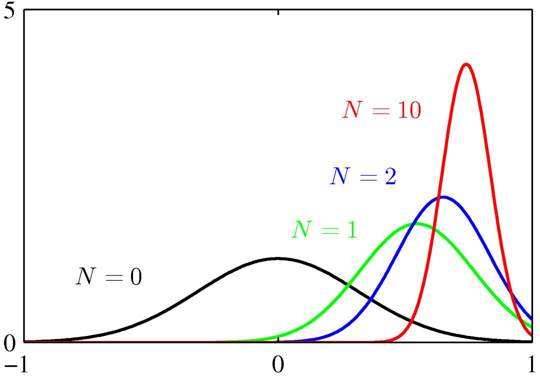
$$
\begin{aligned}
\mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \\
\mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\
\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma_N^2}
\end{aligned}
$$

Note that

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

Bayesian Inference for the Gaussian (4)

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



Bayesian Inference for the Gaussian (5)

Sequential estimation

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

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

Bayesian Inference for the Gaussian (6)

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

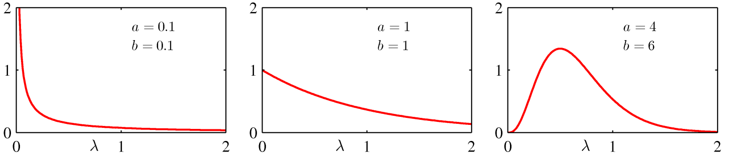
$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \{-\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2\}.$$

This has a Gamma shape as a function of λ.

The Gamma distribution

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

$$\mathbb{E}[\lambda] = \frac{a}{b} \quad \mathbb{V}[\lambda] = \frac{a}{b^2}$$



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

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

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

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

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

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

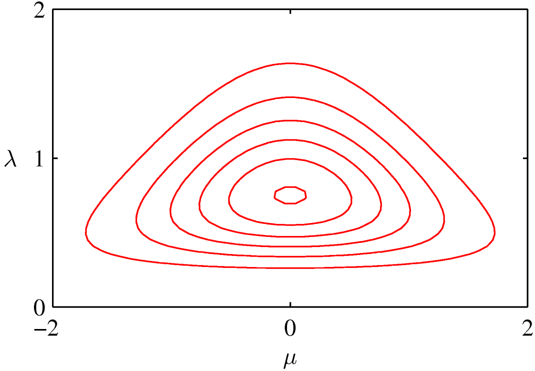
Bayesian Inference for the Gaussian (10)

The Gaussian-Gamma distribution

$$
\begin{aligned}
p(\mu, \lambda) &= \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b) \\
&\propto \exp\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\}\lambda^{a-1}\exp\{b\lambda\}
\end{aligned}
$$

The left term (inside the exp is Quadratic in μ and linear in λ. The right term is a Gamma distribution over λ and independent of μ.

The Gaussian-Gamma distribution



Bayesian Inference for the Gaussian (12)

Multivariate conjugate priors

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

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

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

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

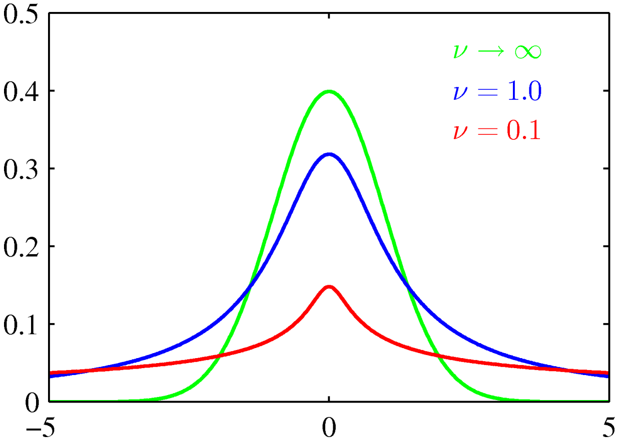
Student's t-Distribution

$$
\begin{aligned}
p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)d\tau \\
&= \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1})\mathrm{Gam}(\eta|\nu/2, \nu/2)d\eta \\
&= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}(\frac{\lambda}{\pi\ nu})^{1/2}[1 + \frac{\lambda(x-\mu)^2}{\nu}]^{-\nu/2-1/2} \\
&= \mathrm{St}(x|\mu, \lambda, \nu)
\end{aligned}
$$

where $\lambda = a/b$, $\eta = \tau b/a$, and $\nu = 2a$.

Note that the integral is in fact infinite mixture of Gaussians.

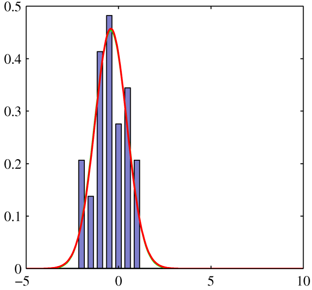
Student's t-Distribution



	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x\|\mu, \lambda^{-1})$

Student's t-Distribution

Robustness to outliers: Gaussian (green) vs. t-distribution (red)



(a) (b)

Student's t-Distribution

The D-variate case:

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta/2,\nu/2)d\eta \\ &= \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}[1+\frac{\Delta^2}{\nu}]^{-D/2-\nu/2}\end{aligned}$$

where $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$. Properties:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, \quad \text{if } \ \nu > 1 \\ \mathrm{Cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, \quad \text{if } \ \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

Perdiodic variables

Examples: calendar time, direction, ... We require

$$p(\theta) \geq 0$$
$$\int_0^{2\pi} p(\theta) d\theta = 1$$
$$p(\theta + 2\pi) = p(\theta)$$

von Mises Distribution (1)

This requirement is statisfied by

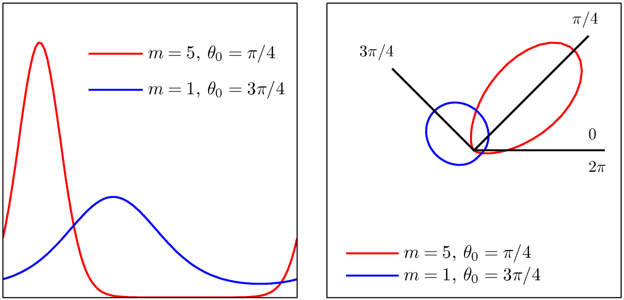
$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m \cos(\theta - \theta_0)\}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{m \cos\theta\} d\theta$$

is the 0^{th} order modified Bessel function of the 1^{st} kind.

von Mises Distribution (2)



Maximum Likelihood for von Mises

Given a data set, $\mathcal{D} = \{\theta_1, \ldots, \theta_N\}$, the log likelihood function is given by

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0)$$

Maximizing w.r.t θ_0 we directly obtain

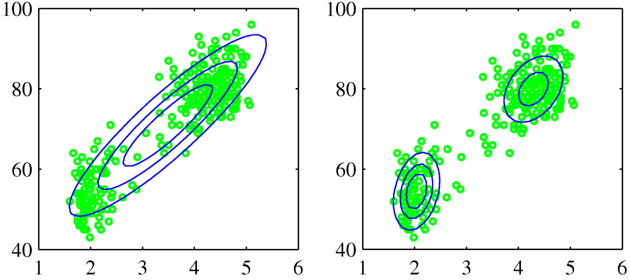
$$\theta_0^{\mathrm{ML}} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}$$

Similarly, maximizing w.r.t m we get

$$\frac{I_1(m_{\mathrm{ML}})}{I_0(m_{\mathrm{ML}})} = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{\mathrm{ML}})$$

which can be solved numerically for m_{ML}.

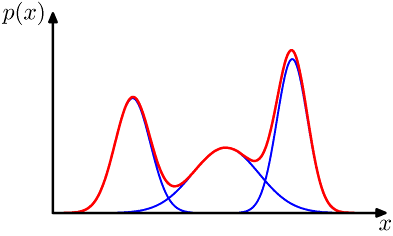
Mixture of Gaussians (1)



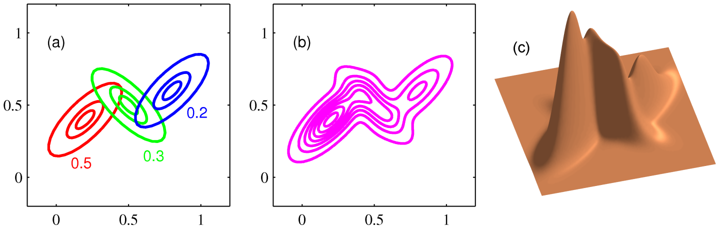
Combine simple models into a complex model

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

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



Mixture of Gaussians (3)



Mixture of Gaussians (4)

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

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

Solution: use standard iterative numeric optimization methods or the *Expectation-Maximization* algorithm.

The Exponential Family (1)

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\eta)\exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\}$$

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

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} d\mathbf{x} = 1$$

so $g(\boldsymbol{\eta})$ can be interpreted as a normalization coefficient.

The Exponential Family (2.1)

The Bernoulli distribution

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

Comparing with the general form we see that

$$
\begin{aligned}
\eta &= \ln(\frac{\mu}{1-\mu}) \\
\mu &= \sigma(\eta) = \frac{1}{1+\exp(-\eta)} \quad \text{(logistic sigmoid)}
\end{aligned}
$$

The Exponential Family (2.2)

The Bernoulli distribution can be then written as

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

where $u(x) = x$, $h(x) = 1$, and $g(\eta) = 1 - \sigma(\eta) = \sigma(-\eta)$.

The Exponential Family (3.1)

The Multinomial Distribution

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

where $\mathbf{x} = (x_1, \ldots, x_M)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^\top$ 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 (3.2)

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

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \text{ and}$$
$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}$$

Here the η_k parameters are independent. Note that $0 \leq \mu_k \leq 1$ and $\sum_{k=1}^{M-1} \mu_k \leq 1$.

The Exponential Family (3.3)

The Multinomial distribution can be written as

$$p(\mathbf{x}|\boldsymbol{\mu}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x}))$$

where

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

The Exponential Family (4)

The Gaussian distribution

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

where

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

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

ML for Exponential Family (1)

From the definition of $g(\boldsymbol{\eta})$ we get

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} d\mathbf{x} = 0$$

The right term is $\mathbb{E}[\mathbf{u}(\mathbf{x})]$ and the left integral is $1/g(\boldsymbol{\eta})$ Thus

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

ML for Exponential Family (2)

Given a data set $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_N$ the likelihood function is given by

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

Thus we have

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

Conjugate priors

For any member of the exponential family, there exists a prior

$$p(\boldsymbol{\eta}|\mathcal{X}, \nu) = f(\mathcal{X}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\{\nu \boldsymbol{\eta}^{\top} \mathcal{X}\}$$

Combining with the likelihood function, we get

$$p(\boldsymbol{\eta}|\mathbf{X}, \mathcal{X}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\{\boldsymbol{\eta}^{\top} (\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu \mathcal{X})\}$$

Prior corresponds to ν pseudo-observations with value $\mathcal{X}$.

Noninformative priors (1)

With a little or no information available a priori, we might choose 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 $\lambda = \eta^2$:

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

Noninformative priors (2)

Translation invariant priors. Consider

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

For a corresponding prior over μ, we have

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

for any A and B. Thus $p(\mu) = p(\mu - c)$ and $p(\mu)$ must be constant.

Noninformative priors (3)

Example: 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)

Scale invariant priors. Consider $p(x|\sigma) = (1/\sigma)f(x/\sigma)$ and make the change of variable $\hat{x} = cx$

$$p_{\hat{x}}(\hat{x}) = p_x(x)|\frac{dx}{d\hat{x}}| = p_x(\frac{\hat{x}}{c})\frac{1}{c} = \frac{1}{c\sigma}f(\frac{\hat{x}}{c\sigma}) = p_x(\hat{x}|\hat{\sigma})$$

For a corresponding prior over σ, we have

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

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

Noninformative priors (5)

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

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

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 the Gamma distribution,

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

A noninformative prior is obtained when $a_0 = 0$ and $b_0 = 0$.

Nonparametric methods (1)

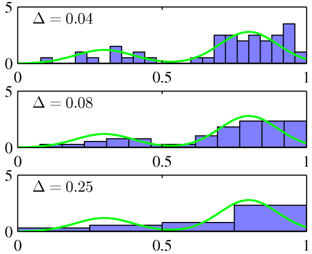
Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modeling a multimodal distribution with a single unimodal model.

Nonparametric approaches make few assumptions about the overall shape of the distribution being modeled.

Nonparametric methods (2)

Histogram methods - partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$



- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.
- In a D-dimensional space using M bins in each dimension will require M^D bins!

Nonparametric methods (3)

Assume observations drawn from a density $p(\mathbf{x})$ and consider a small region $\mathcal{R}$ containing $\mathbf{x}$ such that

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}.$$

The probability that K out of N observations lie inside $\mathcal{R}$ is $\mathrm{Bin}(K|N, P)$ and if N is large

$$K \approx NP.$$

If the volume of $\mathcal{R}$, V, is sufficiently small, $p(\mathbf{x})$ is approximately constant over $\mathcal{R}$ and $P \approx p(\mathbf{x})V$, thus

$$p(\mathbf{x}) = \frac{K}{NV}.$$

V small, yet $K > 0$, therefore N large?

Nonparametric methods (4)

Kernel Density Estimation - fix V, estimate K from the data. Let $\mathcal{R}$ be a hypercube centered on $\mathbf{x}$ and define the kernel function (Parzen window)

$$k(\frac{\mathbf{x}-\mathbf{x}_n}{h}) = \begin{cases} 1, & |\frac{x_i - x_{ni}}{h}| \leq \frac{1}{2}, \quad i = 1, \ldots, D, \\ 0, & \text{otherwise} \end{cases}$$

It follows that
$K = \sum_{n=1}^{N} k(\frac{\mathbf{x}-\mathbf{x}_n}{h})$ and hence $p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k(\frac{\mathbf{x}-\mathbf{x}_n}{h})$.

Nonparametric methods (5)

To avoid discontinuities in $p(x)$, use a smooth kernel, e.g. a Gaussian

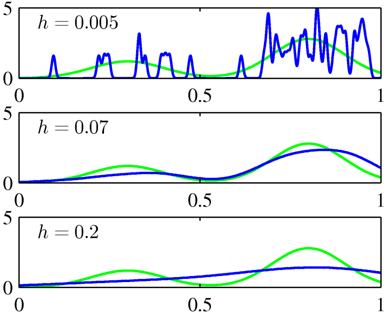
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \exp\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\}$$

Any kernel such that

$$k(\mathbf{u}) \geq 0$$

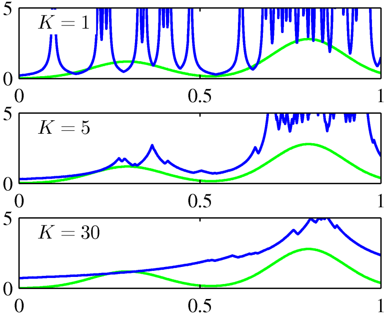
$$\int k(\mathbf{u})d\mathbf{u} = 1$$

will work.



Nearest Neighbor Density Estimation - fix K, estimate V from data. Consider a hypersphere centered on $\mathbf{x}$ and let it grow to a volume, V^*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \approx \frac{K}{NV^*}$$



K acts as a smoother

Nonparametric methods (7)

Nonparametric models (not histograms) require storing and computing with the entire data set.
Parametric models, once fitted, are much more efficient in terms of storage and computation.

K-Nearest-Neighbor for Classification (1)

Given a data set with N_k data points from class $\mathcal{C}_k$ and $\sum_k N_k = N$, we have

$$p(\mathbf{x}) = \frac{K}{NV}$$

and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}$$

Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x}} = \frac{K_k}{K}$$

K-Nearest-Neighbor for Classification (2)

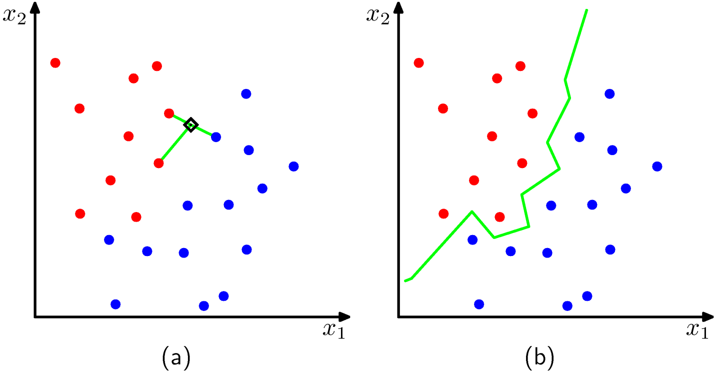
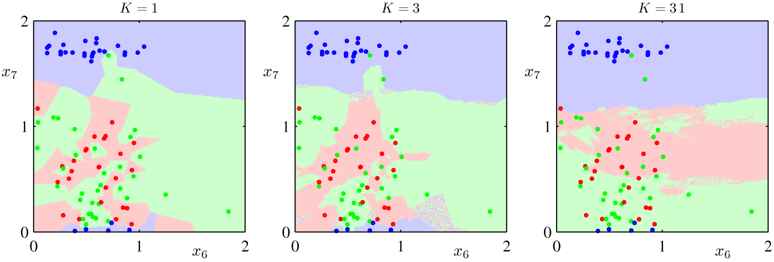


Figure: (left) $K = 3$, (right) $K = 1$

K-Nearest-Neighbor for Classification (3)



- K acts as a smoother
- for $N \to \infty$, the error rate of the 1-nearest neighbor classifier is never more than twice the optimal error (obtained from the true conditional class distributions).