Probability Distributions

CPS 271
Ron Parr

Some Figures courtesy Andrew Ng and Chris Bishop and © original authors.
Thanks to Lise Getoor for some slides

Bernouli Distribution

- What is P(x=I(heads)=1)?
- P(x)=μ
- E(x)= μ
- Var(x)= $\mu(1-\mu)$

- Empirical mean = Sample mean = maximum likelihood = μ_{ML}

Is The Empirical Mean Reasonable?

- ML solution is presented as frequentist solution
- We know:
  - $E(\mu_{ML}) = \mu$
  - μ_{ML} converges to μ

- What about small numbers of samples?

Binomial Distribution

- Probability of getting m heads in N flips?
- Add up different ways this can happen

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

$$E(m) = N\mu$$
$$Var(m) = N(1-\mu)\mu$$

Conjugate Priors

- We know μ_{ML} maximizes P(D|H)
- For small data sets, this seems unreliable
- Can we maximize P(H|D)=P(D|H)P(H)/P(D)?
- Questions:
  - What form should P(H) take?
  - If H is in some class (binomial, Bernouli), we want P(D|H)P(H)=P(HD) to generate answers that are also in this class
- In general, if P(D|H)P(H) is in the same class as P(H), we say that P(H) is conjugate for P(D|H)

Background: Gamma Function

- For discrete variables:

$$\Gamma(x+1) = x!$$
$$\Gamma(x+1) = x\Gamma(x)$$

- For continuous variables, continuous generalization of factorial:

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$
$$\Gamma(x+1) = x\Gamma(x)$$

Beta Distribution

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

$$E(\mu) = \frac{a}{a+b}$$

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

Observation: Beta has very similar form to binomial

Posterior with Beta Prior

- Want P(D|H)P(H)
- P(D|H) = Binomial
- P(H)=Beta

$$P(D \mid H)P(H) \propto \mu^m(1-\mu)^{N-m}\mu^{a-1}(1-\mu)^{b-1} = \mu^{m+a-1}(1-\mu)^{N-m+b-1}$$

$$P(H \mid D) = \frac{\Gamma(m+a+N-m+b)}{\Gamma(m+a)\Gamma(N-m+b)} \mu^{m+a-1}(1-\mu)^{N-m+b-1}$$

$$= beta(\mu \mid m+a, N-m+b)$$

$$\approx Bin(m+a \mid \mu, N+a+b)$$

Interpreting the Beta Prior

$$P(D \mid H)P(H) \propto \mu^m(1-\mu)^{N-m}\mu^{a-1}(1-\mu)^{b-1} = \mu^{m+a-1}(1-\mu)^{N-m+b-1}$$

$$P(H \mid D) = \frac{\Gamma(m+a+N-m+b)}{\Gamma(m+a)\Gamma(N-m+b)} \mu^{m+a-1}(1-\mu)^{N-m+b-1}$$

$$= beta(\mu \mid m+a, N-m+b)$$

$$\approx Bin(m+a \mid \mu, N+a+b)$$

- A beta prior with parameters a,b is like having "imagined" a previous heads, b previous tails
- Examples:
  - a=b=1000 implies strong prior towards fairness
  - a=b=1 implies weak prior towards fairness
  - a=1000, b=1 implies strong prior towards heads bias
  - a=1, b=1000 implies weak prior towards head bias

Multinomial

- Multinomial generalizes binomial to >2 outcomes

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

- Dirichlet is conjugate

$$dir(\boldsymbol{\mu}, \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

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

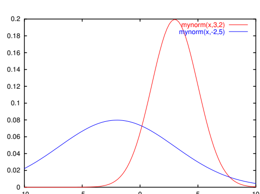
- α parameters correspond to phantom observations

Multivariate Gaussian Distribution

- also called multivariate normal
- First, recall the univariate Gaussian distribution:

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

- where μ is the mean and σ^2 is the variance



Multivariate Gaussian Distribution

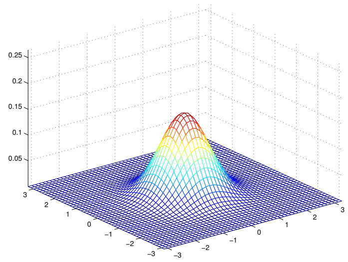
- A 2-dimensional Gaussian is defined by a mean vector $\mu = (\mu_1,\mu_2)$ and a covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 \end{bmatrix}$$

- where $\sigma_{i,j}^2 = E[(x_i - \mu_i)(x_j - \mu_j)]$
  - is the variance if $x_i = x_j$
  - covariance if $x_i \neq x_j$

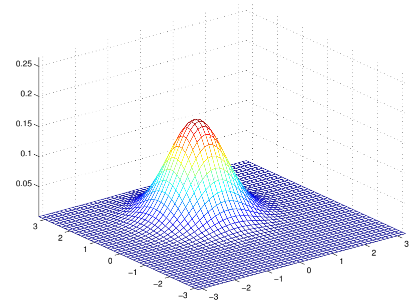
$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)\,|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Standard normal distribution



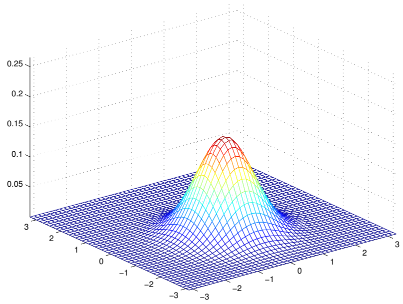
- We get the standard normal for Σ = the identity matrix $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mu = (0,0)$

MVG examples



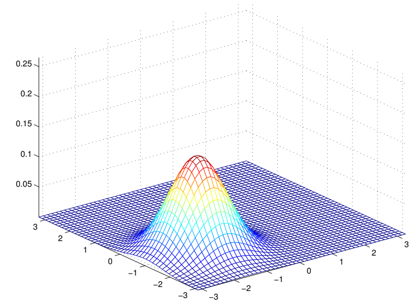
$\mu = (1, 0)$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

MVG examples



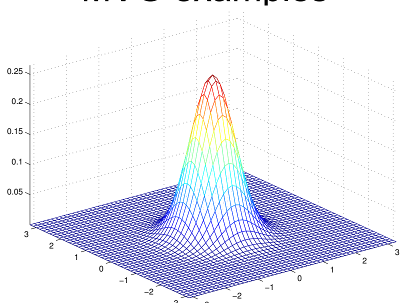
$\mu = (-0.5, 0)$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

MVG examples



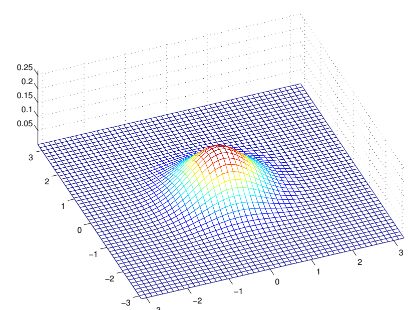
$\mu = (-1,-1.5)$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

MVG examples



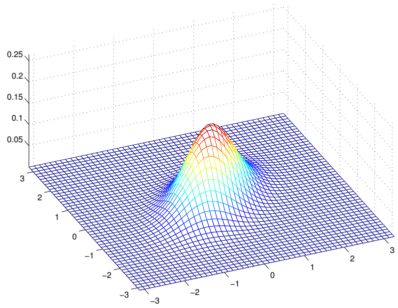
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$

MVG examples



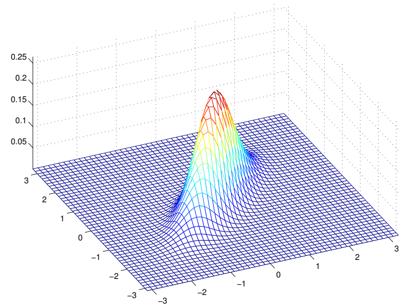
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

MVG examples



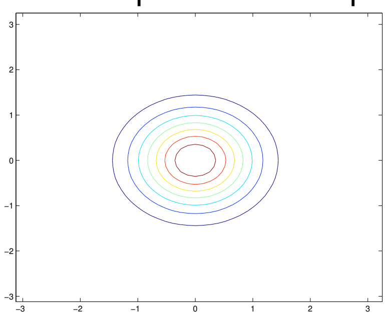
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

MVG examples



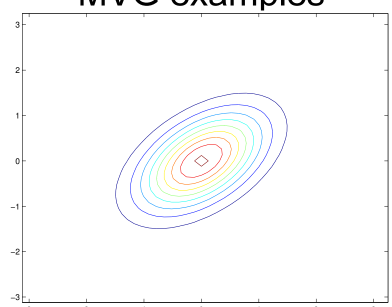
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

MVG examples – contour plots



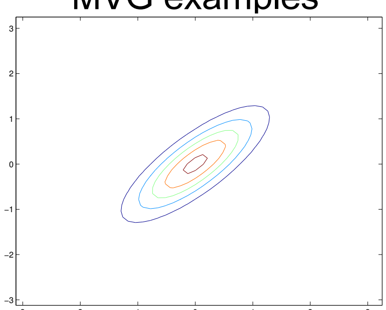
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

MVG examples



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

MVG examples



$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

Multivariate normal distribution

- We can generalize this to n dimensions
- parameters
  - mean vector $\mu \in \Re^n$
  - a covariance matrix $\Sigma \in \Re^{n \times n}$, where $\Sigma \geq 0$ is symmetric and positive semi-definite
- Written $N(\mu, \Sigma)$, density is

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

- where $|\Sigma|$ is the determinant of the matrix Σ

- For $X \sim N(\mu, \Sigma)$
  - $E[X] = \int_x x\ p(x;\mu,\Sigma)\ dx = \mu$
  - $Cov(X) = E[XX^T] - (E[X])(E[X])^T = \Sigma$

A note about covariances

- By construction, the covariance matrix is
  - Symmetric
  - Positive semi-definite
- Diagonal covariance matrices:
  - Can be expressed as a product of I and a vector of variances
  - Imply independence between variables

Useful Properties of Gaussians I

- Surfaces of equal probability for standard (mean 0, I covariance) Gaussians are spheroids
- Surfaces of equal probability for general Gaussians are ellipsoids
- Every general Gaussian can be viewed as a standard Gaussian that has undergone an affine transformation

Useful Properties of Gaussians II

- A Gaussian distribution is completely specific by the a vector of means and covariance matrix
- Requires $O(n^2)$ space
- Requires $O(n^3)$ time to manipulate
- If these seem bad, recall that a joint distribution over n binary variables requires $O(2^n)$ space

Useful Properties of Gaussians III

- Marginals of Gaussians are Gaussian
- Given: $x = (x_a, x_b), \mu = (\mu_a, \mu_b)$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

- Marginal Distribution:

$$p(x_a) = N(x_a \mid \mu_a, \Sigma_{aa})$$

- (Marginalize by ignoring)

Useful Properties of Gaussians IV

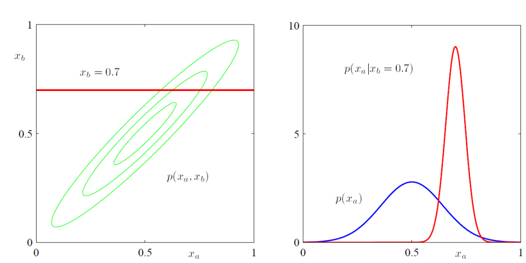
- Conditionals of Gaussians are Gaussian
- Notation:

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

- Conditional Distribution:

$$p(x_a \mid x_b) = N(x_a \mid \mu_{a|b}, \Lambda_{aa}^{-1})$$
$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_a)$$

Visualizing Marginalization & Conditioning



Useful Properties of Gaussians V

- Affine transformations of Gaussian variables are Gaussian
  - Suppose x is Gaussian
  - y=Ax+b is Gaussian
- Uses:
  - Compute distribution on Y from distribution on x
  - Compute posterior on x after observing y

Useful Properties of Gaussians

- Lots of things can (arguably) be approximated well by Gaussians
- The central limit theorem: The sum of IID variables with finite variances will tend towards a Gaussian distribution

- Note: This is often used a hand-waving argument to justify using the Gaussian distribution for almost anything

Limitations of Gaussians

- Gaussians are unimodal (single peak at mean)
- $O(n^2)$ and $O(n^3)$ can get expensive
- Definite integrals of Gaussian distributions do not have a closed form solution (somewhat inconvenient)
  - Must approximate, use lookup tables, etc.
  - Sampling from Gaussian is inelegant

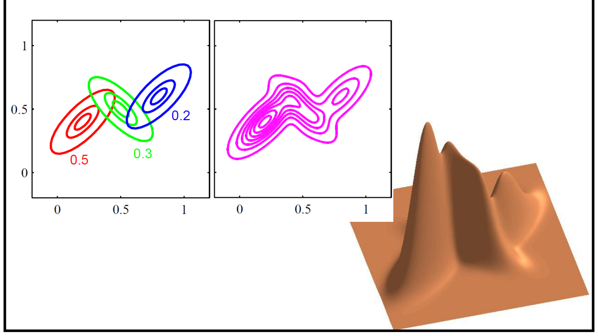
Mixtures of Gaussians

- Want to approximate distribution that is not unimodal?
- Density is weighted combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$
$$\sum_{k=1}^{K} \pi_k = 1$$

- Idea: Flip coin (roll dice) to select Gaussian, then sample from the Gaussian
- Can be arbitrarily expressive with enough Gaussians

Mixture of Gaussians Example



Fitting Gaussians

- Maximum Likelihood
- Mean:

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

- Covariance:

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

Bayesian Fits with Known Variance

- Can use a Gaussian prior:

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

- Posterior:

$$p(\mu \mid X) = N(\mu \mid \mu_N, \sigma_N{}^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0{}^2 + \sigma^2} + \frac{N\sigma_0{}^2}{N\sigma_0{}^2 + \sigma^2}\mu_{ML}$$

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

Bayesian Fit with Unknown Variance, Known Mean

- For single variable, gamma distribution is conjugate
- For multiple variables, Wishart is conjugate
- No conjugate for unknown mean & variance