Tutorial on Estimation and Multivariate Gaussians

STAT 27725/CMSC 25400: Machine Learning

Shubhendu Trivedi - shubhendu@uchicago.edu

Toyota Technological Institute

October 2015

- Things we will look at today
  - Maximum Likelihood Estimation
  - ML for Bernoulli Random Variables
  - Maximizing a Multinomial Likelihood: Lagrange Multipliers
  - Multivariate Gaussians
  - Properties of Multivariate Gaussians
  - Maximum Likelihood for Multivariate Gaussians
  - (Time permitting) Mixture Models

The Principle of Maximum Likelihood

- Suppose we have N data points $X = \{x_1, x_2, \ldots, x_N\}$ (or $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$)
- Suppose we know the probability distribution function that describes the data $p(x; \theta)$ (or $p(y|x; \theta)$)
- Suppose we want to determine the parameter(s) θ
- Pick θ so as to *explain* your data best
- What does this mean?
- Suppose we had two parameter values (or vectors) θ_1 and θ_2.
- Now suppose you were to *pretend* that θ_1 was really the true value parameterizing p. What would be the probability that you would get the dataset that you have? Call this $P1$
- If $P1$ is very small, it means that such a dataset is very unlikely to occur, thus perhaps θ_1 was not a good guess

The Principle of Maximum Likelihood

- We want to pick θ_{ML} i.e. the best value of θ that explains the data you *have*
- The plausibility of given data is measured by the "likelihood function" $p(x;\theta)$
- Maximum Likelihood principle thus suggests we pick θ that maximizes the likelihood function
- The procedure:
  - Write the log likelihood function: $\log p(x;\theta)$ (we'll see later why log)
  - Want to maximize - So differentiate $\log p(x;\theta)$ w.r.t θ and set to zero
  - Solve for θ that satisfies the equation. This is θ_{ML}

The Principle of Maximum Likelihood

- As an aside: Sometimes we have an initial guess for θ BEFORE seeing the data
- We then use the data to *refine* our guess of θ using Bayes Theorem
- This is called MAP (Maximum a posteriori) estimation (we'll see an example)
- Advantages of ML Estimation:
  - Cookbook, "turn the crank" method
  - "Optimal" for large data sizes
- Disadvantages of ML Estimation
  - Not optimal for small sample sizes
  - Can be computationally challenging (numerical methods)

A Gentle Introduction: Coin Tossing

Problem: estimating bias in coin toss

- A single coin toss produces H or T.
- A sequence of n coin tosses produces a sequence of values;
  $n = 4$
  T,H,T,H
  H,H,T,T
  T,T,T,H
- A probabilistic model allows us to model the uncertainly inherent in the process (randomness in tossing a coin), as well as our uncertainty about the properties of the source (fairness of the coin).

Probabilistic model

- First, for convenience, convert $H \to 1$, $T \to 0$.
  - We have a *random variable* X taking values in $\{0, 1\}$
- Bernoulli distribution with parameter μ:

$$\Pr(X = 1;\, \mu) \;=\; \mu.$$

- We will write for simplicity $p(x)$ or $p(x; \mu)$ instead of $\Pr(X = x; \mu)$
- The parameter $\mu \in [0, 1]$ specifies the bias of the coin
  - Coin is fair if $\mu = \frac{1}{2}$

Reminder: probability distributions

- Discrete random variable X taking values in set $\mathcal{X} = \{x_1, x_2, \ldots\}$
- Probability mass function $p : \mathcal{X} \to [0, 1]$ satisfies the law of total probability:

$$\sum_{x \in \mathcal{X}} p(X = x) \; = \; 1$$

- Hence, for Bernoulli distribution we know

$$p(0) \; = \; 1 - p(1;\; \mu) \; = \; 1 - \mu.$$

Sequence probability

- Now consider two tosses of the same coin, $\langle X_1, X_2 \rangle$
- We can consider a number of probability distributions:

  Joint distribution $p(X_1, X_2)$
  Conditional distributions $p(X_1 \mid X_2)$, $p(X_2 \mid X_1)$,
  Marginal distributions $p(X_1)$, $p(X_2)$

- We already know the marginal distributions:
  $p(X_1 = 1; \mu) \equiv p(X_2 = 1; \mu) = \mu$
- What about the conditional?

Sequence probability (contd)

- We will assume the sequence is i.i.d. - *independently identically distributed*.
- Independence, by definition, means

$$p(X_1 \,|\, X_2) = p(X_1), \qquad p(X_2 \,|\, X_1) = p(X_2)$$

  i.e., the conditional is the same as marginal - knowing that X_2 was H does not tell us anything about X_1.
- Finally, we can compute the joint distribution, using chain rule of probability:

$$p(X_1, X_2) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2)$$

Sequence probability (contd)

$$p(X_1, X_2) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2)$$

- More generally, for i.i.d. sequence of n tosses,

$$p(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} p(x_i; \mu).$$

- Example: $\mu = \frac{1}{3}$. Then,

$$p(H, T, H;\, \mu) = p(H;\, \mu)^2 p(T;\, \mu) = \left(\frac{1}{3}\right)^2 \cdot \frac{2}{3} = \frac{2}{27}.$$

Note: the order of outcomes does not matter, only the number of Hs and Ts.

The parameter estimation problem

- Given a sequence of n coin tosses $x_1, \ldots, x_n \in \{0,1\}^n$, we want to estimate the bias μ.
- Consider two coins, each tossed 6 times:
  coin 1 H,H,T,H,H,H
  coin 2 T,H,T,T,H,H
- What do you believe about μ_1 vs. μ_2?
- Need to convert this intuition into a precise procedure

Maximum Likelihood estimator

- We have considered $p(x; \mu)$ as a function of x, parametrized by μ.
- We can also view it as a function of μ. This is called the *likelihood* function.
- Idea for estimator: choose a value of μ that maximizes the likelihood given the observed data.

ML for Bernoulli

- Likelihood of an i.i.d. sequence $\mathbf{X} = [x_1, \ldots, x_n]$:

$$L(\mu) = p(\mathbf{X}; \mu) = \prod_{i=1}^{n} p(x_i; \mu) = \prod_{i=1}^{n} \mu^{x_i} (1-\mu)^{1-x_i}$$

- log-likelihood:

$$l(\mu) = \log p(\mathbf{X}; \mu) = \sum_{i=1}^{n} [x_i \log \mu + (1-x_i) \log(1-\mu)]$$

- Due to monotonicity of $\log$, we have

$$\operatorname*{argmax}_{\mu} p(\mathbf{X}; \mu) = \operatorname*{argmax}_{\mu} \log p(\mathbf{X}; \mu)$$

- We will usually work with log-likelihood (why?)

ML for Bernoulli (contd)

- ML estimate is
  $\widehat{\mu}_{ML} = \text{argmax}_\mu \{\sum_{i=1}^n [x_i \log \mu + (1 - x_i) \log(1 - \mu)]\}$
- To find it, set the derivative to zero:

$$\frac{\partial}{\partial \mu} \log p(\mathbf{X}; \mu) = \frac{1}{\mu} \sum_{i=1}^n x_i - \frac{1}{1-\mu} \sum_{j=1}^n (1 - x_j) = 0$$

$$\frac{1-\mu}{\mu} = \frac{\sum_{j=1}^n (1 - x_j)}{\sum_{i=1}^n x_i}$$

$$\widehat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ML estimate is simply the fraction of times that H came up.

Are we done?

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Example: H,T,H,T $\rightarrow \widehat{\mu}_{ML} = \frac{1}{2}$
- How about: $H\ H\ H\ H$? $\rightarrow \widehat{\mu}_{ML} = 1$
  Does this make sense?
- Suppose we record a very large number of 4-toss sequences for a coin with true $\mu = \frac{1}{2}$.
  We can expect to see H,H,H,H about 1/16 of all sequences!
- A more extreme case: consider a single toss.
  $\widehat{\mu}_{ML}$ will be either 0 or 1.

Bayes rule

- To proceed, we will need to use Bayes rule
- We can write the joint probability of two RV in two ways, using chain rule:

$$p(X, Y) = p(X)p(Y|X) = p(Y)p(X|Y).$$

- From here we get the *Bayes rule*:

$$p(X|Y) = \frac{p(X)p(Y|X)}{p(Y)}$$

Bayes rule and estimation

- Now consider μ to be a RV. We have

$$p(\mu \,|\, \mathbf{X}) \;=\; \frac{p(\mathbf{X} \,|\, \mu) p(\mu)}{p(\mathbf{X})}$$

- Bayes rule converts *prior* probability $p(\mu)$ (our belief about μ prior to seeing any data) to *posterior* $p(\mu|\mathbf{X})$, using the likelihood $p(\mathbf{X}|\mu)$.

MAP estimation

$$p(\mu \,|\, \mathbf{X}) \;=\; \frac{p(\mathbf{X} \,|\, \mu) p(\mu)}{p(\mathbf{X})}$$

- The *maximum a-posteriori* (MAP) estimate is defined as

$$\widehat{\mu}_{MAP} \;=\; \underset{\mu}{\operatorname{argmax}}\, p(\mu|\mathbf{X})$$

- Note: $p(\mathbf{X})$ does not depend on μ, so if we only care about finding the MAP estimate, we can write

$$p(\mu|\mathbf{X}) \;\propto\; p(\mathbf{X}|\mu)p(\mu)$$

- What's $p(\mu)$?

Choice of prior

- Bayesian approach: try to reflect our *belief* about μ
- Utilitarian approach: choose a prior which is computationally convenient
  - Later in class: *regularization* - choose a prior that leads to better prediction performance
- One possibility: uniform $p(\mu) \equiv 1$ for all $\mu \in [0, 1]$.
  "Uninformative" prior: MAP is the same as ML estimate

Constrained Optimization: A Multinomial Likelihood

Problem: estimating biases in Dice

- A dice is rolled n times: A single roll produces one of $\{1, 2, 3, 4, 5, 6\}$
- Let $n_1, n_2, \ldots n_6$ count the outcomes for each value
- This is a multinomial distribution with parameters $\theta_1, \theta_2, \ldots, \theta_6$
- The joint distribution for $n_1, n_2, \ldots, n_6$ is given by

$$p(n_1, n_2, \ldots, n_6; n, \theta_1, \theta_2, \ldots, \theta_6) = \left(\frac{n!}{n_1! n_2! n_3! n_4! n_5! n_6!} \right) \prod_{i=1}^{6} \theta_i^{n_i}$$

- Subject to $\sum_i \theta_i = 1$ and $\sum_i n_i = n$

A False Start

- The likelihood is

$$L(\theta_1, \theta_2, \ldots, \theta_6) = \left(\frac{n!}{n_1! n_2! n_3! n_4! n_5! n_6!} \right) \prod_{i=1}^{6} \theta_i^{n_i}$$

- The Log-Likelihood is

$$l(\theta_1, \theta_2, \ldots, \theta_6) = \left(\log \frac{n!}{n_1! n_2! n_3! n_4! n_5! n_6!} \right) + \sum_{i=1}^{6} n_i \log \theta_i$$

- Optimize by taking derivative and setting to zero:

$$\frac{\partial l}{\partial \theta_1} = \frac{n_1}{\theta_1} = 0$$

- Therefore: $\theta_1 = \infty$
- What went wrong?

A Possible Solution

- We forgot that $\sum_{i=1}^{6} \theta_i = 1$
- We could use this constraint to eliminate one of the variables:

$$\theta_6 = 1 - \sum_{i=1}^{5} \theta_i$$

- and then solve the equations

$$\frac{\partial l}{\partial \theta_i} = \frac{n_1}{\theta_i} - \frac{n_6}{1 - \sum_{i=1}^{5} \theta_i} = 0$$

- Gets messy

A More Elegant Solution: Lagrange Multipliers

- General constrained optimization problem:

$$\max_{\theta} f(\theta) \text{ subject to } g(\theta) - c = 0$$

- We can then define the Lagrangian

$$\mathcal{L}(\theta, \lambda) = f(\theta) - \lambda(g(\theta) - c)$$

- Is equal to f when the constraint is satisfied
- Now do unconstrained optimization over θ and λ:
- Optimizing the Lagrange multiplier λ enforces constraint
- More constraints, more multipliers

Back to Rolling Dice

- Recall

$$l(\theta_1, \theta_2, \ldots, \theta_6) = \left(\log \frac{n!}{n_1! n_2! n_3! n_4! n_5! n_6!} \right) + \sum_{i=1}^{6} n_i \log \theta_i$$

- The Lagrangian may be defined as:

$$\mathcal{L} = \log \frac{n!}{\prod_i n_i!} + \sum_{i=1}^{6} n_i \log \theta_i - \lambda \Big(\sum_{i=1}^{6} \theta_i - 1 \Big)$$

Back to Rolling Dice

- Taking derivative with respect to θ_i and setting to 0

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0$$

- Let optimal $\theta_i = \theta_i^*$

$$\frac{n_i}{\theta_i^*} - \lambda^* = 0 \implies \frac{n_i}{\lambda^*} = \theta_i^*$$

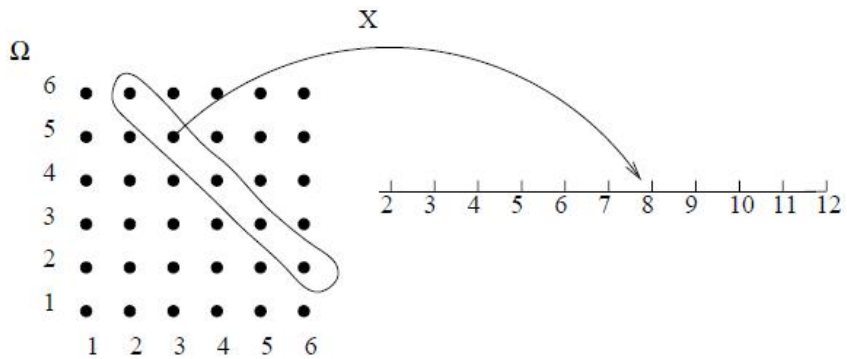
$$\sum_{i=1}^{6} \frac{n_i}{\lambda^*} = \sum_{i=1}^{6} \theta_i^* = 1$$

$$\lambda^* = \sum_{i=1}^{6} n_i \implies \theta_i^* = \frac{n_i}{\sum_{i=1}^{6} n_i}$$

Multivariate Gaussians

Quick Review: Discrete/Continuous Random Variables

- A Random Variable is a function $X : \Omega \mapsto \mathbb{R}$
- The set of all possible values a random variable X can take is called its **range**
- **Discrete** random variables can only take isolated values (probability of a random variable taking a particular value reduces to counting)
- Discrete Example: Sum of two fair dice



- Continuous Example: Speed of a car

Discrete Distributions

- Assume X is a discrete random variable. We would like to specify probabilities of events $\{X = x\}$
- If we can specify the probabilities involving X, we can say that we have specified the probability distribution of X
- For a countable set of values $x_1, x_2, \ldots x_n$, we have $\mathbb{P}(X = x_i) > 0, i = 1, 2, \ldots, n$ and $\sum_i \mathbb{P}(X = x_i) = 1$
- We can then define the **probability mass function** f of X by $f(X) = \mathbb{P}(X = x)$
- Sometimes write as f_X

Probability Mass Function

- Example: Toss a die and let X be its face value. X is discrete with range $\{1, 2, 3, 4, 5, 6\}$. The pmf is

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

- Another example: Toss two dice and let X be the largest face value. The pmf is

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

Probability Density Functions

- A random variable X taking values in set $\mathcal{X}$ is said to have a continuous distribution if $\mathbb{P}(X = x) = 0$ for all $x \in \mathcal{X}$
- The probability density function of a continuous random variable X satisfies
  - $f(x) \geq \ \forall\ x$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx\ \forall\ a, b$
- Probabilities correspond to areas under the curve $f(x)$
- Reminder: No longer need to have $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \leq 1$ but must have $\int_{-\infty}^{\infty} f(x)dx = 1$

Why Gaussians?

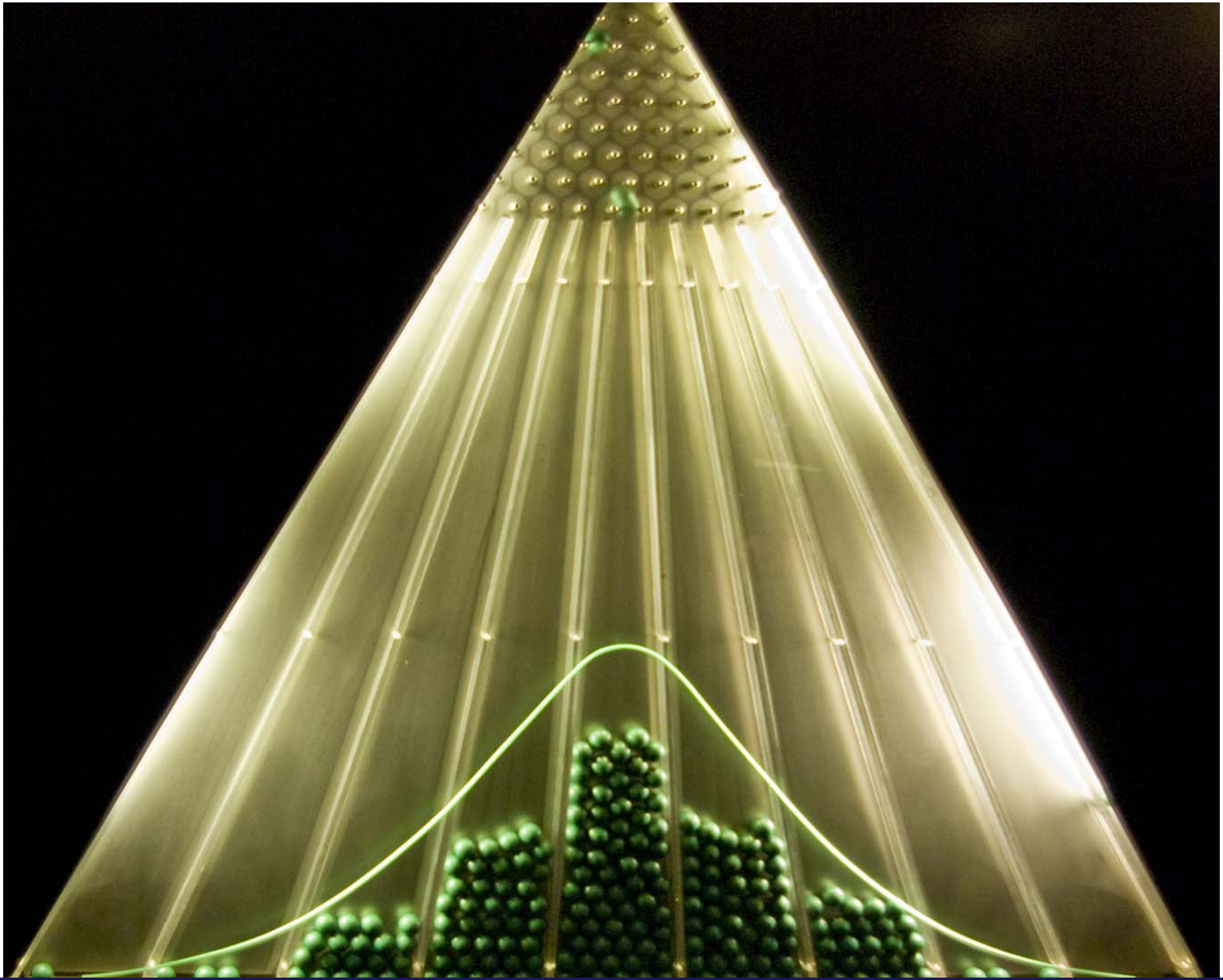
- Gaussian distributions are widely used in machine learning:
  - Central Limit Theorem!

$$\bar{X}_n = X_1 + X_2 + \cdots + X_n$$

$$\sqrt{n}\bar{X}_n \xrightarrow{d} \mathcal{N}\left(x;\, \mu, \sigma^2\right)$$

  - Actually, there are a set of "Central Limit Theorems" (e.g. corresponding to p-Stable Distributions)

Why Gaussians?



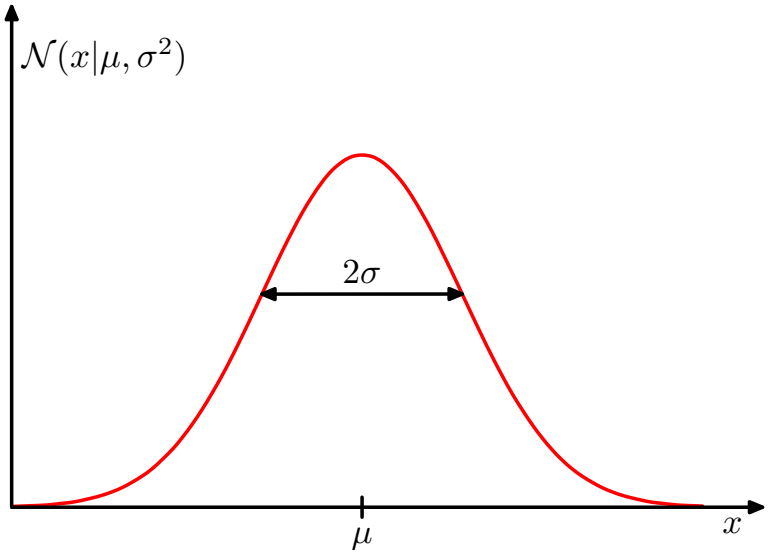
Why Gaussians?

- Gaussian distributions are widely used in machine learning:
  - Central Limit Theorem!
  - Gaussians are convenient computationally;
  - Mixtures of Gaussians (just covered in class) are sufficient to approximate a wide range of distributions;
  - Closely related to squared loss (have seen earlier in class), an important error measure in statistics.

Reminder: univariate Gaussian distribution

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

- mean μ determines location
- variance σ^2;
  standard deviation $\sqrt{\sigma^2}$
  determines the spread
  around μ



Moments

- Reminder: expectation of a RV x is $E[x] \triangleq \int xp(x)dx$, so

$$E[x] = \int_{-\infty}^{\infty} x\mathcal{N}(x;\mu,\sigma^2)dx = \mu$$

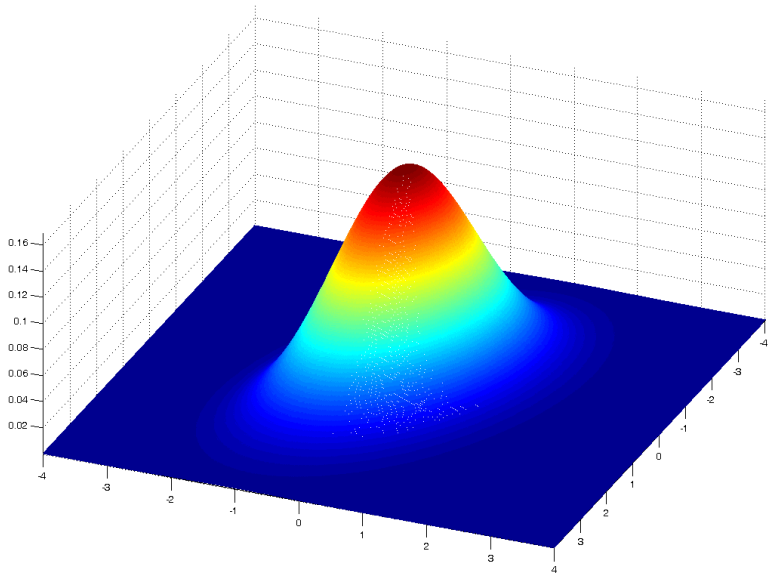
- Variance of x is $\operatorname{var} x \triangleq E\left[(x - E[x])^2\right]$, and

$$\operatorname{var} x = \int_{-\infty}^{\infty} (x-\mu)^2\mathcal{N}(x;\mu,\sigma^2)dx = \sigma^2$$

Multivariate Gaussian

- Gaussian distribution of a random vector $\mathbf{x}$ in $\mathbb{R}^d$:

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



- The $\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}$ factor ensures it's a pdf (integrates to one).

Matrix notation

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

- Boldfaced lowercase vectors $\mathbf{x}$, uppercase matrices $\boldsymbol{\Sigma}$.
- Determinant $|\boldsymbol{\Sigma}|$
- Matrix inverse $\boldsymbol{\Sigma}^{-1}$
- Transpose $\mathbf{x}^T, \boldsymbol{\Sigma}^T$

Mean of the Gaussian

- By definition,

$$E[\mathbf{x}] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \mathbf{x}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dx_1 \dots dx_d$$

- Solving this we indeed get

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

Covariance

- Variance of a RV x with mean μ: $\sigma_x^2 = E\left[(x-\mu)^2\right]$
- Generalization to two variables: *covariance*

$$\mathrm{Cov}_{x_1,x_2} \triangleq E\left[(x_1-\mu_1)(x_2-\mu_2)\right]$$

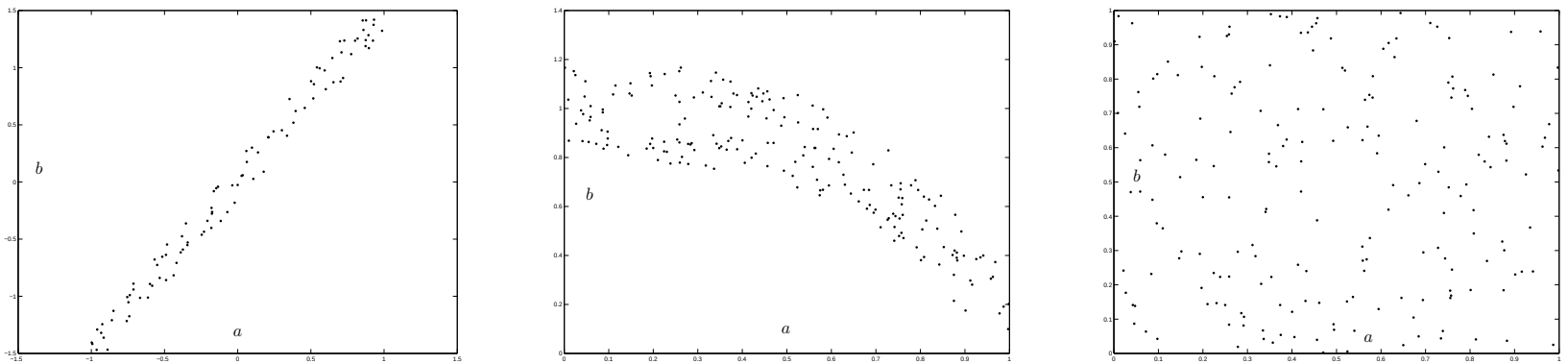
- Measures how the two variables deviate together from their means ("co-vary").
- Note: $\mathrm{Cov}_{x,x} \equiv \mathrm{var}(x) = \sigma_x^2$

Correlation vs. covariance

- Correlation:

$$\mathrm{cor}(a,b) \triangleq \frac{\mathrm{Cov}_{a,b}}{\sigma_a \sigma_b}.$$

$\mathrm{cor} \approx 1$ $\quad -1 < \mathrm{cor} < 0$ $\quad \mathrm{cor} \approx 0$



- $\mathrm{cor}(a,b)$ measures the linear relationship between a and b.
- $-1 \leq \mathrm{cor}(a,b) \leq +1$; $+1$ *or* -1 means a is a linear function of b.

Covariance matrix

- For a random vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ with mean $\boldsymbol{\mu}$,

$$\mathrm{Cov}_{\mathbf{x}} \triangleq \begin{bmatrix} \sigma^2_{x_1} & \mathrm{Cov}_{x_1,x_2} & \ldots\ldots & \mathrm{Cov}_{x_1,x_d} \\ \mathrm{Cov}_{x_2,x_1} & \sigma^2_{x_2} & \ldots\ldots & \mathrm{Cov}_{x_2,x_d} \\ \ddots & & \ddots & \ddots \\ \mathrm{Cov}_{x_d,x_1} & \mathrm{Cov}_{x_d,x_2} & \ldots\ldots & \sigma^2_{x_d} \end{bmatrix}.$$

- Square, symmetric, non-negative main diagonal–why? variances ≥ 0, and $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$ by definition
- One can show (directly from definition):

$$\mathrm{Cov}_{\mathbf{x}} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right]$$

  i.e. expectation of the *outer product* of $\mathbf{x} - E[\mathbf{x}]$ with itself.
- Note: so far nothing Gaussian-specific!

Covariance of the Gaussian

- We need to calculate $E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]$
- With a bit of algebra, we get

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

- Now, we already have $E[\mathbf{x}] = \boldsymbol{\mu}$, and

$$\begin{aligned} E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] &= E\left[\mathbf{x}\mathbf{x}^T - \boldsymbol{\mu}\mathbf{x}^T - \mathbf{x}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right] \\ &= E\left[\mathbf{x}\mathbf{x}^T\right] - \underbrace{\left\{\boldsymbol{\mu}(E[\mathbf{x}])^T + E[\mathbf{x}]\,\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T\right\}}_{=\boldsymbol{\mu}\boldsymbol{\mu}^T} \\ &= E\left[\mathbf{x}\mathbf{x}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \boldsymbol{\Sigma} \end{aligned}$$

Properties of the covariance

- Consider the eigenvector equation: $\mathbf{\Sigma u} = \lambda \mathbf{u}$
- As a covariance matrix, $\mathbf{\Sigma}$ is symmetric $d \times d$ matrix. Therefore, we have d solutions $\{\lambda_i, \mathbf{u}_i\}_{i=1}^{d}$ where the *eigenvalues* λ_i are real, and the eigenvectors $\mathbf{u}_i$ are orthonormal, i.e., inner product

$$\mathbf{u}_j^T \mathbf{u}_i = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

- The covariance matrix Σ then may be written as: $\Sigma = \sum_i \lambda_i u_i u_i^T$
- Thus, the inverse covariance may be written as: $\Sigma^{-1} = \sum_i \frac{1}{\lambda_i} u_i u_i^T$

Continued..

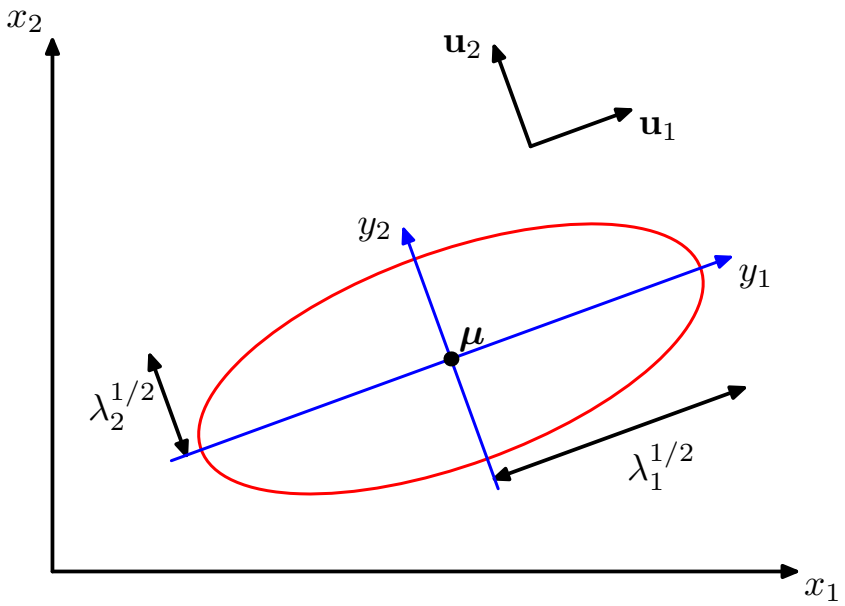
- The quadratic form $(x-\mu)^T\Sigma^{-1}(x-\mu)$ becomes:

$$\sum_i \frac{y_i^2}{\lambda_i}$$

  where $y_i = u_i^T(x-\mu)$
- $\{y_i\}$ may be interpreted as a new coordinate system defined by the orthonormal vectors u_i that are shifted and rotated with respect to the original coordinate system
- Stack the d transposed orthonormal eigenvectors of $\mathbf{\Sigma}$ into $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1^T \\ \cdots \\ \mathbf{u}_d^T \end{bmatrix}$. Then, $\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$ defines rotation (and possibly reflection) of $\mathbf{x}$, shifted so that $\boldsymbol{\mu}$ becomes origin.

Geometry of the Gaussian

- $\sqrt{\lambda_i}$ gives scaling along $\mathbf{u}_i$
- Example in 2D:



Geometry Continued ...

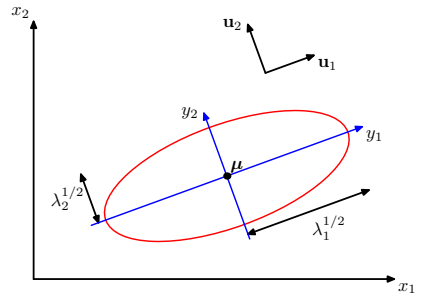
- The determinant of the covariance matrix may be written as the product of its eigenvalues i.e. $|\Sigma|^{\frac{1}{2}} = \prod_j \lambda_j^{\frac{1}{2}}$
- Thus, in the y_i coordinate system, the Gaussian distribution takes the form:

$$p(y) = \prod_j \frac{1}{(2\pi\lambda_j)^{\frac{1}{2}}} \exp\left(-\frac{y_j^2}{2\lambda_j}\right)$$

- which is the product of d independent univariate Gaussians
- The eigenvectors thus define a new set of shifted and rotated coordinates w.r.t which the joint probability distribution factorizes into a product of independent distributions

Density contours

- What are the constant density contours?



$$\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) = \text{const}$$

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{const}$$

- This is a quadratic form, whose solution is an ellipsoid (in 2D, simply an ellipse)

Density Contours are Ellipsoids

- We saw that: $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \text{const}^2$
- Recall that $\Sigma^{-1} = \sum_i \frac{1}{\lambda_i} u_i u_i^T$
- Thus we have:

$$\sum_i \frac{y_i^2}{\lambda_i} = \text{const}^2$$

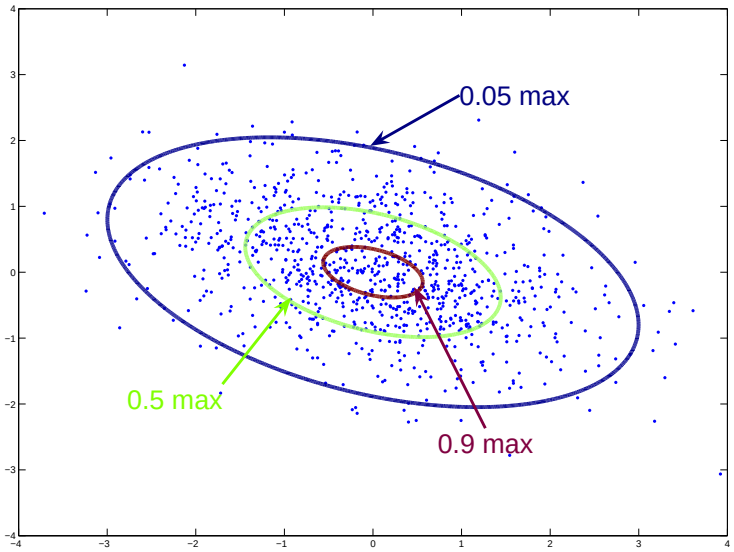
  where $y_i = u_i^T (x - \mu)$

- Recall the expression for an ellipse in $2D$: $\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$

Intuition so far

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

- Falls off exponentially as a function of (squared) Euclidean distance to the mean $\|\mathbf{x}-\boldsymbol{\mu}\|^2$;
- the *covariance matrix* $\boldsymbol{\Sigma}$ determines the shape of the density;



- Determinant $|\boldsymbol{\Sigma}|$ measures the "spread" (analogous to σ^2).
- $\mathcal{N}$ is the joint density of coordinates $x_1, \ldots, x_d$.

Linear functions of a Gaussian RV

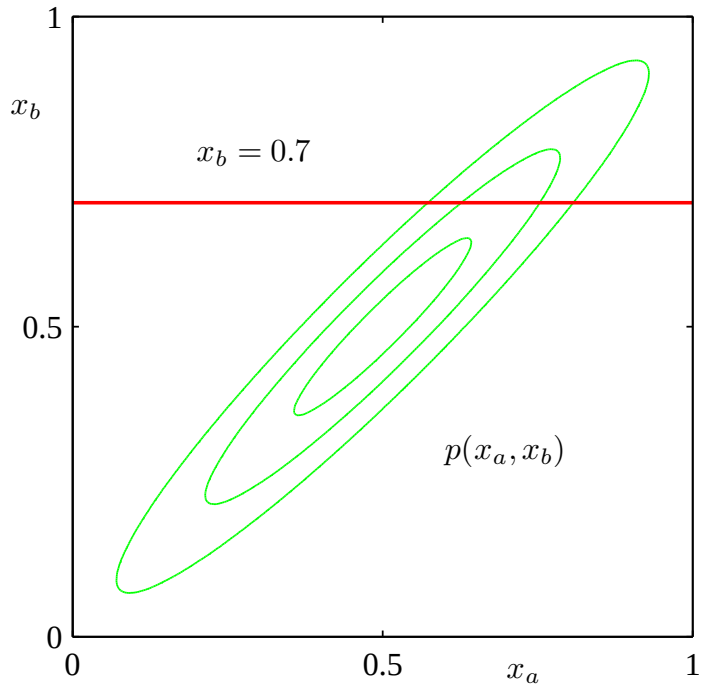
- For any RV $\mathbf{x}$, and for any $\mathbf{A}$ and $\mathbf{b}$,

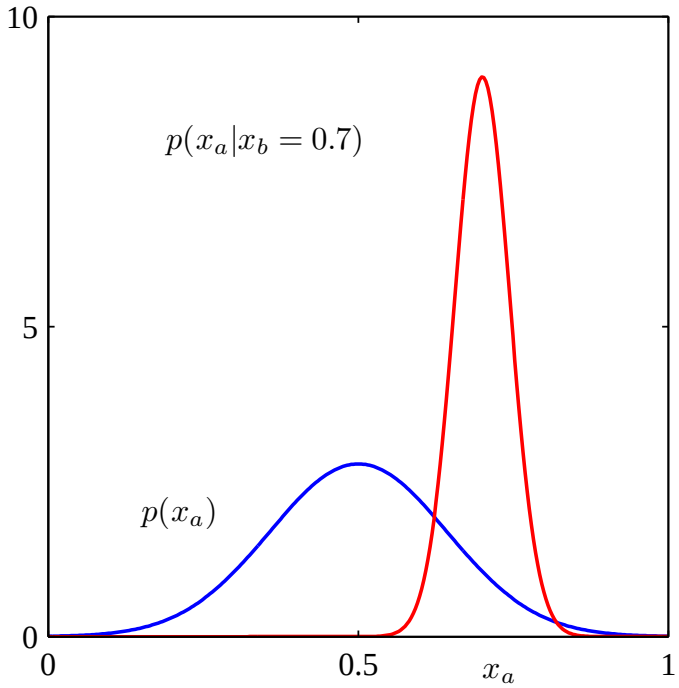
$$E[\mathbf{A}\mathbf{x}+\mathbf{b}] = \mathbf{A}E[\mathbf{x}]+\mathbf{b}, \qquad \mathrm{Cov}(\mathbf{A}\mathbf{x}+\mathbf{b}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{x})\mathbf{A}^T.$$

- Let $\mathbf{x} \sim \mathcal{N}(\cdot;\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$; then $p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z};\, \mathbf{A}\boldsymbol{\mu}+\mathbf{b},\, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right)$.
- Consider a row vector $\mathbf{a}^T$ that "selects" a single component from $\mathbf{x}$, i.e., $a_k = 1$ and $a_j = 0$ if $j \neq k$. Then, $z = \mathbf{a}^T\mathbf{x}$ is simply the coordinate x_k.
- We have: $E[z] = \mathbf{a}^T\boldsymbol{\mu} = \mu_k$, and $\mathrm{Cov}(z) = \mathrm{var}(z) = \boldsymbol{\Sigma}_{k,k}$. i.e., marginal of a Gaussian is also a Gaussian

Conditional and marginal

- Marginal ("projection" of the Gaussian on a subset of coordinates) is Gaussian
- Conditional ("slice" through Gaussian at fixed values for a subset of coordinates) is Gaussian





Log-likelihood

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

- Take the log, for a single example $\mathbf{x}$:

$$\log \mathcal{N}(\mathbf{x};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

- Can ignore terms independent of parameters:

$$\log \mathcal{N}(\mathbf{x};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \text{const}$$

Log-likelihood (contd)

$$\log \mathcal{N}(\mathbf{x};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; -\frac{1}{2}\log|\boldsymbol{\Sigma}| \;-\; \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \;+\; \text{const}$$

- Given a set $\mathbf{X}$ of n i.i.d. vectors, we have

$$\log \mathcal{N}(\mathbf{X};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) + \text{const}$$

- We are now ready to compute ML estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

ML for parameters

$$\log \mathcal{N}\left(\mathbf{X};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \;=\; -\frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \text{const}$$

- To find ML estimate, we use the rule

$$\frac{\partial}{\partial \mathbf{a}}\mathbf{a}^T\mathbf{b} \;=\; \frac{\partial}{\partial \mathbf{a}}\mathbf{b}^T\mathbf{a} \;=\; \mathbf{b},$$

and set derivative w.r.t. $\boldsymbol{\mu}$ to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}}\log \mathcal{N}\left(\mathbf{X};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \;=\; \sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \;=\; 0,$$

which yields $\widehat{\boldsymbol{\mu}}_{ML} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

ML for parameters (contd)

- A somewhat lengthier derivation produces ML estimate for the covariance:

$$\widehat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

- Note: the $\boldsymbol{\mu}$ above is the ML estimate $\widehat{\boldsymbol{\mu}}_{ML}$.
- Thus ML estimates for the mean is the *sample mean* of the data, and ML estimate for the covariance is the *sample covariance* of the data.

Mixture Models and Expected Log Likelihood

Mixture Models

- Assumptions:
  - k underlying types (clusters/components)
  - y_i is the identity of the component "responsible" for x_i
  - y_i is a *hidden* (latent) variable: never observed
- A mixture model:

$$p(x;\pi) = \sum_{c=1}^{k} p(y=c)p(x|y=c)$$

- π_c are called mixing probabilities
- The component densities $p(x|y=c)$ needs to be parameterized

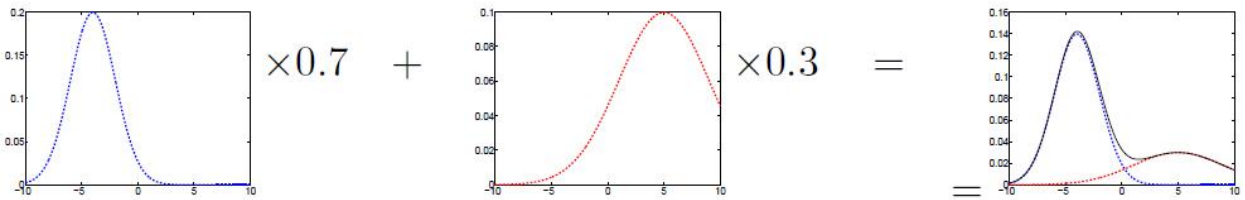
Next few slides adapted from TTIC 31020 by Gregory Shakhnarovich

Parametric Mixtures

- Suppose the parameters of the c-th component are θ_c. Then we can denote $\theta = [\theta_1, \ldots, \theta_k]$ and write

$$p(x; \theta, \pi) = \sum_{c=1}^{k} \pi_c p(x, \theta_c)$$

- Any valid setting of θ and π, such that $\sum_{c=1}^{k} \pi_c = 1$ produces a valid pdf
- Example: Mixture of Gaussians



Generative Model for a Mixture

- The generative process with a k-component mixture:
  - The parameters θ_c for each component are fixed
  - Draw $y_i \sim [\pi_1, \ldots, \pi_k]$
  - Given y_i, draw $x_i \sim p(x|y_i; \theta_{y_i})$
- The entire generative model for x and y

$$p(x, y; \theta, \pi) = p(y; \pi)p(x|y; \theta_y)$$

- What does this mean? Any data point x_i could have been generated in k ways
- If the c-th component is Gaussian i.e. $p(x|y = c) = \mathcal{N}(x; \mu_c, \Sigma_c)$

$$p(x; \theta, \pi) = \sum_{c=1}^{k} \pi_c \mathcal{N}(x; \mu_c, \Sigma_c)$$

where $\theta = [\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k]$

Likelihood of a Mixture Model

- Usual Idea: Estimate set of parameters that maximize likelihood given observed data
- The log-likelihood of π, θ for $X = \{x_1, \ldots, x_N\}$:

$$\log p(X; \pi, \theta) = \sum_{i=1}^{N} \log \sum_{c=1}^{k} \pi_c \mathcal{N}(x_i; \mu_c, \Sigma_c)$$

- No closed form solution because of sum inside log
- How will we estimate parameters?

Scenario 1: Known Labels. Mixture Density Estimation

- Suppose that we do observe $y_i \in \{1, \ldots, k\}$ for each $i = 1, \ldots, N$
- Let us introduce a set of binary indicator variables $\mathbf{z_i} = [z_{i1}, \ldots, z_{ik}]$, where:

$$z_{ic} = \begin{cases} 1 & \text{if } y_i = c \\ 0 & \text{otherwise} \end{cases}$$

- The count of examples from c-th component

$$N_c = \sum_{i=1}^{N} z_{ic}$$

Scenario 1: Known Labels. Mixture Density Estimation

- If we know z_i, the ML estimates of the Gaussian components are simply (as we have seen earlier)

$$\hat{\pi}_c = \frac{N_c}{N}$$

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N} z_{ic} x_i,$$

$$\hat{\Sigma}_c = \frac{1}{N_c} \sum_{i=1}^{N} z_{ic} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c{}^T$$

Scenario 2: Credit Assignment

- When we *don't know* y, we face a credit assignment problem: Which component is responsible for x_i?
- Suppose for a moment that we do know the component parameters $\theta = [\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k]$ and mixing probabilities $\pi = [\pi_1, \ldots, \pi_k]$
- Then, we can compute the posterior of each label using Bayes' theorem:

$$\gamma_{ic} = \hat{p}(y = c|x; \theta, \pi) = \frac{\pi_c p(x; \mu_c, \Sigma_c)}{\sum_{l=1}^{k} \pi_l p(x; \mu_l, \Sigma_l)}$$

- We call γ_{ic} the *responsibility* of the c-th component for x

Expected Likelihood

- The "complete data" likelihood (when $\mathbf{z}$ are known):

$$p(X, Z; \pi, \theta) = \propto \prod_{i=1}^{N} \prod_{c=1}^{k} (\pi_c \mathcal{N}(x_i; \mu_c, \Sigma_c))^{z_{ic}}$$

and the log

$$p(X, Z; \pi, \theta) = \text{ const } + \sum_{i=1}^{N} \sum_{c=1}^{k} z_{ic} (\log \pi_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c))$$

- We can't compute it (why?), but can take the expectation w.r.t the posterior of z, which is just γ_{ic} i.e. $\mathbb{E}[z_{ic}] = \gamma_{ic}$
- The expected likelihood of the data:

$$\mathbb{E}[\log p(X, Z; \pi, \theta)] = \text{ const } + \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} (\log \pi_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c))$$

Expectation Maximization

- The expected likelihood of the data:

$$\mathbb{E}[\log p(X, Z; \pi, \theta)] = \text{ const } + \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} (\log \pi_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c))$$

- We can find π, θ that maximizes this *expected* likelihood - by setting derivatives to zero and for π, using Lagrange Multipliers to enforce $\sum_c \pi_c = 1$

Expectation Maximization

- If we know the parameters and indicators (assignments) we are done
- If we know the indicators but not the parameters, we can do ML estimation of the parameters - and we are done
- If we know the parameters but not the indicators, we can compute the posteriors of the indicators. With known posteriors, we can estimate parameters that maximize the expected likelihood - and then we are done
- In reality, we know neither the parameters nor the indicators

Expectation Maximization for Mixture Models

- General Mixture Models: $p(x) = \sum_{c=1}^{k} \pi_c p(x; \theta_c)$
- Initialize π, θ^{old}, and iterate until convergence:
  - E-Step: Compute responsibilities:

$$\gamma_{ic} = \frac{\pi_c^{old} p(x_i; \theta_c^{old})}{\sum_{l=1}^{k} \pi_l^{old} p(x_i; \theta^{old})}$$

  - M-Step: Re-estimate mixture parameters:

$$\pi^{old}, \theta^{new} = \arg\max_{\theta,\pi} \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} (\log \pi_c + \log p(x_i; \theta_c))$$