# Probability Data Model

Sanjay Lall and Stephen Boyd

EE104
Stanford University

## Probability data models

- data model: $x$ *comes from a distribution with density* $p(x;\theta)$
- $\theta$ parametrizes the density function
- implausibility function is negative log density $\ell(x;\theta) = -\log p(x;\theta)$

- example: exponential density $p(x;\theta) = \theta e^{-\theta x}$, for $x \geq 0$, $\theta > 0$
- loss is $\ell(x;\theta) = \theta x - \log\theta$

## Fitting a probability model

- for probability model with $\ell(x;\theta) = -\log p(x;\theta)$, empirical loss is

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\log p(x^i;\theta)$$

  the *average negative log likelihood function*

- we should choose the probability density parameter $\theta$ to mimimize the average negative log likelihood
- called *maximum likelihood fitting* (of a parametrized density to a data set)

## Imputing missing entries with a probability model

- given vector $x$ with missing entries, and a probability model
- we minimize $\ell(\hat{x};\theta)$ over $\hat{x}$, subject to $\hat{x}_i = x_i$ for $i \in \mathcal{K}$
- called *maximum likelihood imputation*

## Gaussian data model

- data model: $x \sim \mathcal{N}(\mu, \Sigma)$
- this means $x$ is a sample from a Gaussian distribution with density

$$p(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

- parameter $\theta$ contains the entries of $\mu$ and $\Sigma$ (in some order)
- $\mu = \mathsf{E}\, x$ is the *mean* of $x$
- $\Sigma = \mathsf{E}(x-\mu)(x-\mu)^T$ is the $d \times d$ *covariance matrix* of $x$
- implausibility function is negative log density

$$\ell(x;\theta) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \frac{1}{2}\left(d \log 2\pi + \log \det \Sigma\right)$$

- second term does not depend on $x$

## Fitting a Gaussian model

- choose model parameters $\mu$ and $\Sigma$ to minimize the average negative log likelihood

$$\frac{1}{2n}\sum_{i=1}^{n}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)+\frac{1}{2}\left(d\log 2\pi+\log\det\Sigma\right)$$

- solution is

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x^i, \qquad \Sigma=\frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^T$$

  the *empirical mean and covariance*

## Derivation of Gaussian model fitting

- to find optimal $\mu$, set gradient of loss w.r.t. $mu$ to zero:

$$0 = \nabla_\mu \ell = \frac{1}{n} \sum_{i=1}^{n} \Sigma^{-1}(x^i - \mu)$$

- multiply by $\Sigma$ to get $\mu = \frac{1}{n} \sum_{i=1}^{n} x^i$
- let $R = \Sigma^{-1}$ (the precision matrix) so loss is

$$\frac{1}{2n} \sum_{i=1}^{n} (x^i - \mu)^T R (x^i - \mu) + \frac{1}{2} \left( d \log 2\pi - \log \det R \right)$$

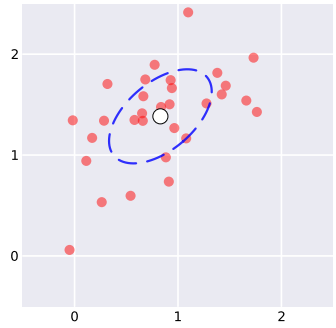
- set gradient w.r.t. $R$ to zero to get (using $\nabla \log \det R = R^{-1}$)

$$\frac{1}{2n} \sum_{i=1}^{n} (x^i - \mu)(x^i - \mu)^T - \frac{1}{2} R^{-1} = 0$$

- $\Sigma = R^{-1} = \frac{1}{n} \sum_{i=1}^{n} (x^i - \mu)(x^i - \mu)^T$

## Example

- $n = 30$ points in $\mathbb{R}^2$
- $\mu = (0.83, 1.38)$, $\Sigma = \begin{bmatrix} 0.25 & 0.12 \\ 0.12 & 0.22 \end{bmatrix}$
- ellipse shows $\{x \mid (x-\mu)^T \Sigma^{-1} (x-\mu)/2 = 1\}$



## Imputing missing entries with a Gaussian data model

- minimize
$$\ell(\hat{x};\theta) = \frac{1}{2}(\hat{x}-\mu)^T\Sigma^{-1}(\hat{x}-\mu) + \frac{1}{2}\left(d\log 2\pi + \log\det\Sigma\right)$$
subject to $\hat{x}_i = x_i$ for $i \in \mathcal{K}$
- minimize $(\hat{x}-\mu)^T\Sigma^{-1}(\hat{x}-\mu)$ subject to $\hat{x}_i = x_i$ for $i \in \mathcal{K}$
- $\hat{x}$ is the conditional mean of $x$, given the known entries

# Example

- same 30 points in $\mathbb{R}^2$ from pervious example
- impute $x_2$ from $x = (-0.17, ?)$