

# Density Estimation

Seungjin Choi

Department of Computer Science  
POSTECH, Korea  
seungjin@postech.ac.kr

1

## Probability Theory vs Statistics

- Denote by  $\mathcal{M}$  model, by  $\theta$  parameter, and by  $\mathcal{X}$  observed data. Then we have

$$\begin{aligned} (\mathcal{M}, \theta) &\longrightarrow \mathcal{X} && \text{(probability)} \\ \mathcal{X} &\longrightarrow (\mathcal{M}, \theta) && \text{(statistics)}. \end{aligned}$$

- Probability:** Given a particular choice of graphical model and a set of local conditional probabilities or potentials, probability theory involves probabilistic inference, i.e., inferring the probabilities of events of interest such as the marginal or conditional probability.
- Statistics:** In a statistical setting, given random variables, we are interested in inferring the model from the data.

3

# Outline

- ▶ Bayesians vs frequentists
- ▶ Statistical problems
  - Density estimation
  - Regression (later)
  - Classification (later)
- ▶ Parameter estimation
  - Maximum likelihood estimation
  - Maximum *a posteriori* (MAP) estimation
  - Bayesian estimation
- ▶ Model selection and model averaging

2

## Bayesians vs Frequentists

- ▶ Bayesian statistics is at some level an attempt to deny any fundamental distinction between probability theory and statistics.
- ▶ Frequentist view  $p(x|\theta)$  as a conditional probability distribution, i.e., the assignment of probability mass to the unknown values of  $X$ , given a fixed value  $\theta$ .
- ▶ Bayesians view  $X$  as known values (observed realization  $x$ ) and view  $\theta$  as unknown. They would like to be able to invert the relationship between  $x$  and  $\theta$ , using Bayes rule:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

- ▶ Bayesian believe in considering all possible estimators of parameters to estimate the best one while frequentists believe in finding the best estimator of the parameter based on a loss function.

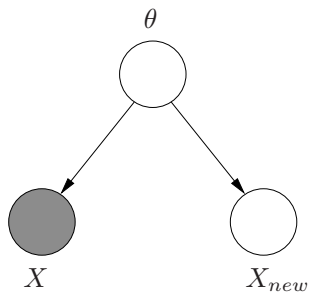
4

## Density Estimation

- The density estimation aims at inducing a probability density (mass) function  $p(x)$ , given a finite number of data points,  $\{x_t\}_{t=1}^N$ .
- Where to use it?
  - Assessment: fault detection, outlier detection
  - Clustering
  - Dimensionality reduction
  - Prediction
- Approaches to density estimation
  - Parametric estimation
  - Nonparametric estimation
  - Semiparametric estimation

5

## Bayesian Prediction



### Bayesian

$$\begin{aligned}
 p(x_{new}|x) &= \int p(x_{new}, \theta|x) d\theta \\
 &= \int p(x_{new}|\theta, x) p(\theta|x) d\theta \\
 &= \int p(x_{new}|\theta) p(\theta|x) d\theta
 \end{aligned}$$

### Frequentist

$$p(x_{new}|\hat{\theta}_{ML})$$

7

## Estimators

### MLE

$$\hat{\theta}_{ML} = \arg \max_{\theta} p(x|\theta)$$

### MAP

$$\begin{aligned}
 \hat{\theta}_{MAP} &= \arg \max_{\theta} p(\theta | x) \\
 &= \arg \max_{\theta} p(x | \theta) p(\theta)
 \end{aligned}$$

### Bayes estimator

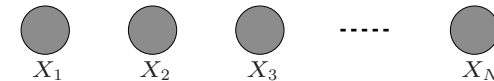
$$\hat{\theta}_{Bayes} = \langle \theta | x \rangle = \int \theta p(\theta|x) d\theta$$

### Penalized MLE

$$\hat{\theta}_{MAP} = \arg \max_{\theta} [\log p(x|\theta) + \log p(\theta)]$$

6

## Univariate Gaussian Density Estimation: MLE



$$\begin{aligned}
 p(x|\theta) &= \prod_{n=1}^N p(x_n|\theta) = \prod_{n=1}^N \frac{1}{(2\pi\sigma^2)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x_n-\mu)^2\right\}} \\
 &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n-\mu)^2\right\} \\
 l(\theta;x) &= \log p(x|\theta) \\
 \frac{\partial l(\theta;x)}{\partial \mu} &= \frac{\partial}{\partial \mu} \left( -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n-\mu)^2 \right) = \frac{1}{\sigma^2} \sum_{n=1}^N (x_n-\mu) = 0 \\
 \frac{\partial l(\theta;x)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \left( -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n-\mu)^2 \right) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^N (x_n-\mu)^2 = 0 \\
 \Rightarrow \hat{\mu}_{ML} &= \frac{1}{N} \sum_{n=1}^N x_n, \quad \hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu}_{ML})^2
 \end{aligned}$$

8

## Univariate Gaussian Density Estimation: MAP

Assume  $x \sim \mathcal{N}(\mu, 1)$ . Use a prior  $p(\mu) \sim \mathcal{N}(0, \alpha^2)$ .

Then we have

$$\begin{aligned} \mathcal{L} &= \log p(\mathcal{X}|\theta) + \log p(\theta) \\ &\propto -\frac{1}{2} \sum_{t=1}^N (x_t - \mu)^2 - \frac{1}{2\alpha^2} \mu^2. \end{aligned}$$

It follows from  $\frac{\partial \mathcal{L}}{\partial \mu} = 0$  that

$$\hat{\mu}_{MAP} = \frac{1}{(N + \frac{1}{\alpha^2})} \sum_{t=1}^N x_t.$$

9

## Univariate Gaussian Density Estimation: MAP (Cont'd)

- For  $N \gg \frac{1}{\alpha^2}$  (the influence of the prior is negligible), we have

$$\hat{\mu}_{MAP} \rightarrow \hat{\mu}_{ML} = \frac{1}{N} \sum_{t=1}^N x_t$$

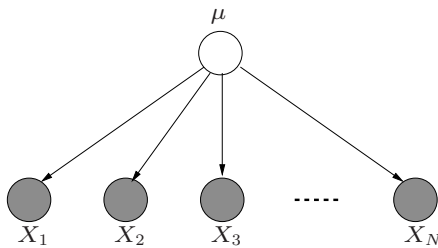
- For very strong belief in the prior, i.e.,  $\frac{1}{\alpha^2} \gg N$ , we have

$$\hat{\mu}_{MAP} \rightarrow 0.$$

If few data points are available, the prior will bias the estimate towards the priori expected value.

10

## Bayesian Univariate Gaussian Density Estimation



$$p(\mu) = \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}$$

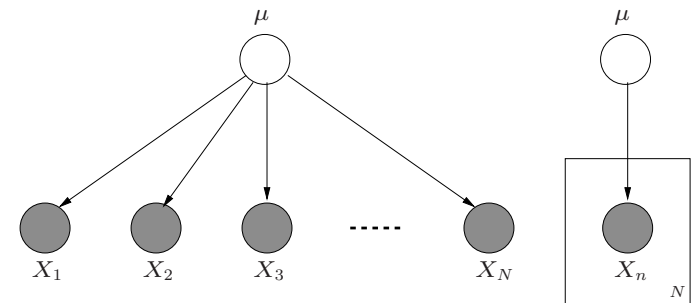
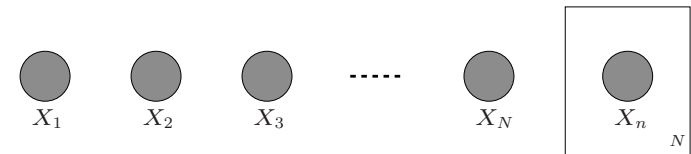
$$p(x, \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2\right\} \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}$$

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

$$\Rightarrow \tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\tau^2} \bar{x} + \frac{1/\tau^2}{N/\sigma^2 + 1/\tau^2} \mu_0, \quad \tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

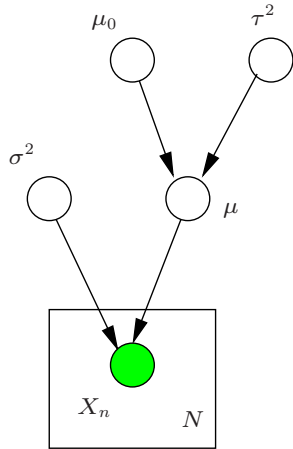
11

## Plates



12

## Gaussian Mean: Graphical Model



13

## Density Estimation for Discrete Data: MLE

Suppose that random variables  $X_n$  can take on one of  $M$  values. Let the range of  $X_n$  be the set of binary  $M$ -component vectors with one component equal to one and the other components equal to zero.  $X_n^k$  refers to the  $k$ th component of the variable  $X_n$ .

Let  $\theta_k$  represent the probability that  $X_n$  takes on its  $k$ th value, i.e.,  $\theta_k = p(x_n^k = 1)$ . Then we have

$$p(x_n|\theta) = \theta_1^{x_n^1} \theta_2^{x_n^2} \dots \theta_M^{x_n^M}. \quad (\text{multinomial distribution})$$

The likelihood is given by

$$p(x|\theta) = \prod_{n=1}^N \theta_1^{x_n^1} \theta_2^{x_n^2} \dots \theta_M^{x_n^M}.$$

The log-likelihood is given by

$$l(\theta; x) = \sum_{n=1}^N \sum_{k=1}^M x_n^k \log \theta_k.$$

14

## Density Estimation for Discrete Data: MLE (Cont'd)

In order to incorporate with a constraint,  $\sum_k \theta_k = 1$ , we form the Lagrangian:

$$\tilde{l}(\theta; x) = \sum_{n=1}^N \sum_{k=1}^M x_n^k \log \theta_k + \lambda \left( 1 - \sum_{k=1}^M \theta_k \right).$$

Take derivatives w.r.t.  $\theta_k$ :  $\frac{\partial \tilde{l}(\theta; x)}{\partial \theta_k} = \frac{\sum_{n=1}^N x_n^k}{\theta_k} - \lambda$ , and set equal to zero:  $\lambda = \frac{\sum_{n=1}^N x_n^k}{\hat{\theta}_{k,ML}}$ .

Incorporating with the constraint  $\sum_k \theta_k = 1$ , leads to

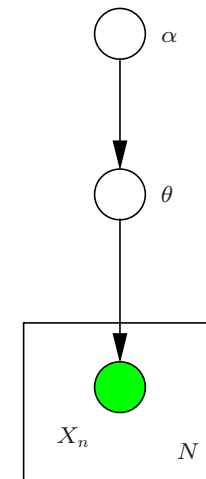
$$\lambda = \sum_{n=1}^N \sum_{k=1}^M x_n^k = N.$$

Thus, the MLE for  $\theta$  is given by

$$\hat{\theta}_{k,ML} = \frac{1}{N} \sum_{n=1}^N x_n^k.$$

15

## A Graphical Model for Bayesian Density Estimation: Discrete Data



16

## Bayesian Density Estimation for Discrete Data

We use the **Dirichlet prior** (which has the same functional form as the multinomial distribution but  $\theta_i$  are random variables in the Dirichlet distribution and parameters in the multinomial distribution):

$$p(\theta) = C(\alpha)\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\dots\theta_M^{\alpha_M-1}, \quad C(\alpha) = \frac{\Gamma\left(\sum_{i=1}^M \alpha_i\right)}{\prod_{i=1}^M \Gamma(\alpha_i)}.$$

We calculate the posterior probability:

$$\begin{aligned} p(\theta|x) &\propto \left[ \prod_{n=1}^N \theta_1^{x_n^1} \theta_2^{x_n^2} \dots \theta_M^{x_n^M} \right] \left[ \prod_{k=1}^M \theta_k^{\alpha_k-1} \right] \\ &= \prod_{k=1}^M \theta_k^{\left(\sum_n x_n^k + \alpha_k - 1\right)}, \end{aligned}$$

which is **again Dirichlet distribution (conjugate prior)** with parameters  $\sum_n x_n^k + \alpha_k$ .

17

### A Special Case: $M = 2$

$X_n$  is treated as a binary random variable, i.e,  $x_n \in \{0, 1\}$ .

In such a case, the multinomial distribution reduces to **Bernoulli distribution**:

$$p(x_n|\theta) = \theta^{x_n} (1 - \theta)^{1-x_n},$$

where the parameter  $\theta$  encodes the probability that  $X_n$  takes the value one.

In the case,  $M = 2$ , the Dirichlet distribution specializes to the **Beta distribution**:

$$p(\theta) = C(\alpha)\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}.$$

See Fig. 5.7 in the textbook!

19

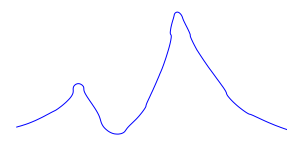
## Bayes Estimator: Discrete Data

$$\begin{aligned} \hat{\theta}_{k, Bayes} &= \langle \theta_k | x \rangle \\ &= \frac{\sum_n x_n^k + \alpha_k}{\sum_{k=1}^M [\sum_n x_n^k + \alpha_k]} \\ &= \frac{\sum_n x_n^k + \alpha_k}{\sum_{k=1}^M \sum_n x_n^k + \sum_{k=1}^M \alpha_k} \\ &= \frac{N}{N + \|\alpha\|_1} \bar{x}_k + \frac{\|\alpha\|_1}{N + \|\alpha\|_1} \frac{\alpha_k}{\|\alpha\|_1} \\ &= \beta \underbrace{\bar{x}_k}_{MLE} + (1 - \beta) \underbrace{\frac{\alpha_k}{\|\alpha\|_1}}_{\text{prior belief}} \end{aligned}$$

**Note:** Adding a prior belief in the form of  $\alpha$ , produces an estimate that is a **weighted sum of prior belief and the maximum likelihood estimate**. This phenomenon is general; it happens with all members of the **exponential family**.

18

## Mixture Models



Multimodal distributions reflect the presence of **subpopulations** or **clusters**.

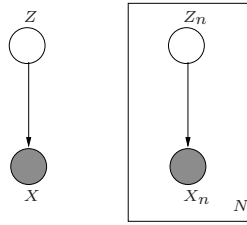
**Divide and Conquer**  $\Rightarrow$  **mixture models!**

$$p(x|\theta) = \sum_{k=1}^K \alpha_k f_k(x|\theta_k),$$

where parameters  $\alpha_k$  are referred to as **mixing proportions** and the densities  $f_k(x|\theta_k)$  are referred to as **mixture components**.

20

## Mixture Models: Graphical Representations



We introduce a multinomial random variable  $Z$ . Define  $\alpha_k \triangleq p(z^k = 1)$  and  $\theta \triangleq (\alpha_1, \dots, \alpha_K, \theta_1, \dots, \theta_K)$ .

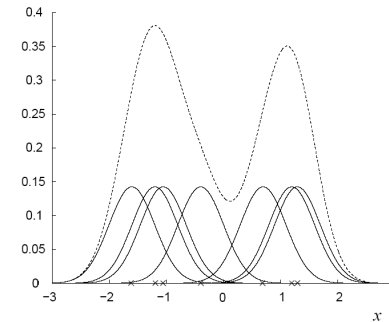
$$p(x, z^k = 1 | \theta) = p(x | z^k = 1, \theta) p(z^k = 1 | \theta) = \alpha_k f_k(x | \theta_k)$$

$$p(x | \theta) = \sum_{k=1}^K p(x, z^k = 1 | \theta) = \sum_{k=1}^K \alpha_k f_k(x | \theta_k)$$

$$p(z^k = 1 | x, \theta) = \frac{p(x | z^k = 1, \theta_k) p(z^k = 1)}{\sum_j p(x | z^j = 1, \theta_j) p(z^j = 1)} = \frac{\alpha_k f_k(x | \theta_k)}{\sum_j \alpha_j f_j(x | \theta_j)}$$

21

## Nonparametric Density Estimation



Let  $k(x, x_n, \lambda)$  be a **kernel function** - a nonnegative function integrating to one (with respect to  $x$ ).

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^N k(x, x_n, \lambda)$$

22