

## 2. Probability Distribution

## 2.0. Introduction

- 1) Chapter Scope
  - A. Examples of probability distributions
  - B. Their properties
- 2) Purpose of Introducing Distributions
  - A. a building blocks for more complex models
  - B. a recipe to discuss some essential statistical concept, e.g., Bayesian inference
  - C. to model the probability distribution  $p(\mathbf{x})$ , i.e., density estimation
    - \* Model Selection becomes an issue since density estimation is fundamentally ill-posed problem in that infinitely many distributions can fit the observed data set.
- 3) Parametric distribution vs. Non-Parametric distribution
  - A. Parametric distribution
    - i. binomial distribution, multinomial distribution, Gaussian distribution (continuous R.V.)
    - ii. For density estimation, the parameters shall be determined with an observed data set.
      1. Frequentist: specific values for parameters (earned by optimizing some criterion, e.g., likelihood function)
      2. Bayesian: estimate posterior distribution with introduced prior distributions over the parameters as well as the observed data
    - iii. Conjugate Priors: To simplify the Bayesian analysis, use conjugate prior which let posterior distribution be in the same form of prior distribution.
      1. Exponential family of distributions is presented as it possesses a number of important properties.
  - B. Non-Parametric distribution
    - i. Distribution form is not forced by a user but typically depends on the size of the data set
    - ii. Still has the parameters but they do not determine the distribution form but the complexity
    - iii. Histogram, nearest-neighbors, kernels

**Table. 1 Conjugate prior with posterior distribution in exponential family**

Conjugate Prior	Posterior Distribution
Dirichlet distribution	Multinomial distribution
Gaussian distribution	Gaussian distribution

## 2.1. Binary Variables

### 2.1.1. Bernoulli distribution

#### 2.1.1.1. Definition

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \text{ where } 0 \leq \mu < 1 \text{ and } x \in \{0, 1\} \quad (2.1)$$

#### 2.1.1.2. Properties

$$\mathbb{E}[x] = \mu \quad (2.2)$$

$$\text{var}[x] = \mu(1 - \mu) \quad (2.3)$$

### 2.1.2. Density estimation

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

#### 2.1.2.1. Frequentist

- 1) Estimate  $\mu$  by maximizing the likelihood function, i.e., maximize the log of likelihood

$$\ln(p(\mathcal{D}|\mu)) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\} \quad (2.4)$$

- 2) The above log likelihood function depends on the N observations only through their sum, i.e., *sufficient statistics*:  $\sum_n x_n$ .

- 3)  $\mu_{ML} = \frac{m}{N} = \text{sample mean}$

#### 2.1.2.2. Bayesian

Flip a coin 3 times resulting all heads  $\rightarrow$  what is the reasonable prediction? (overfitting)

##### 2.1.2.2.1. Binomial distribution

- 1) **Definition**

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m} \quad (2.5)$$

- 2) **Properties**

For independent events, 1) the means of the sum is the sum of the mean and 2) the variance of the sum is the sum of the variance

$$\mathbb{E}[m] = N\mu \quad (2.6)$$

$$\text{var}[m] = N\mu(1 - \mu) \quad (2.7)$$

#### 2.1.2.2.2. The beta distribution (conjugate prior for the binomial distribution)

##### 1) Motivation for the conjugate prior distribution

- A. Prior distribution is required in order to develop a Bayesian treatment.
- B. Make posterior distribution have the same functional form as the prior (conjugacy).

##### 2) Definition

$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$ , gamma coefficient for the normalization purpose

$$\mathbb{E}[\mu] = \frac{a}{a+b} \quad (2.8)$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \quad (2.9)$$

##### 3) Gamma function

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du \quad (2.10)$$

$$\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1, \Gamma(x+1) = x! \quad (2.11)$$

a and b controls the distribution of the parameter  $\mu$ , and thus called *hyperparameters*

##### 4) Posterior distribution

The posterior distribution of  $\mu$ : prior distribution(beta)  $\times$  likelihood function(binomial)

→ Normalization

$$p(\mu|m, l, a, b) = \text{Beta}(\mu|a, b) \times \text{Bin}(m|N, \mu) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1} \quad (2.12)$$

$$l = N - m = \# \text{ of tails}$$

$$m = \# \text{ of heads}$$

##### A. Sequential nature

- i. a and b (in the prior): *effective number of observations* of  $x = 1$  and  $x = 0$ , respectively
- ii. The posterior can act as the prior if subsequent observation is followed. If following observation is  $x=1(x=0)$ ,  $a(b)$  will increase by 1.
- iii. Bayesian viewpoint raises such sequential approach of learning

##### B. Prediction of the future outcome

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}] \quad (2.13)$$

$$p(x = 1|\mathcal{D}) = \frac{m + a}{m + a + l + b} \quad (2.14)$$

→ Total fraction of an effective number for  $x = 1$  (including both real&fictitious)

### C. Big Data

- i. Bayesian result converges to ML (general phenomenon):

$$p(x = 1|\mathcal{D}) = \mathbb{E}[\mu|\mathcal{D}] = \mu_{ML} = \text{sample mean} = \frac{m}{N} \quad (2.15)$$

- ii.  $\text{var}[\mu|\mathcal{D}]$  is approaching zero (Eq 2.9):

In general, the posterior variance is smaller than the prior variance on average (not for every observation)

$$\mathbb{E}_{\mathcal{D}}[\text{var}_{\theta}[\theta|\mathcal{D}]] = \text{var}_{\theta}[\theta] - \text{var}_{\mathcal{D}}[\mathbb{E}_{\theta}[\theta|\mathcal{D}]] \quad (2.16)$$

## 2.2. Multinomial Variables

### 2.2.1. Multinomial Distribution

#### 2.2.1.1. Definition

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1, m_2, \dots, m_K} \prod_{k=1}^K \mu_k^{m_k} \quad (2.17)$$

$$m_k = \sum_n x_{nk}$$

It is the generalization of the Bernoulli distribution to more than two outcomes(states).

#### 2.2.1.2. Properties

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

### 2.2.2. Density estimation

### 2.2.2.1. Frequentist

$$\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$

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

Sufficient statistic for this distribution =  $m_k = \sum_n x_{nk}$

$$\mu_k^{ML} = \frac{m_k}{N}$$

### 2.2.2.2. Bayesian (Dirichlet distribution)

$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1} \quad (2.18)$$

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

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto \text{likelihood} \times \text{prior} = p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m})$$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_K + m_K)} \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1}$$

\*  $\alpha_k = \text{effective number of observations of } x_k = 1$

## 2.3. Gaussian Distribution

Gaussian is a widely used model for continuous variable whereas Bernoulli, Binomial, Multinomial are for discrete.

### 2.3.1. Definition

#### 2.3.1.1. Single variable

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.19)$$

#### 2.3.1.2. Multivariate

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}} \quad (2.20)$$

### 2.3.2. Motivation

- 1) Gaussian maximizes the entropy for the single, continuous, and real variable.
- 2) Central Limit Theorem: the sum of a set of random variables becomes more Gaussian as the # of

variable increases (under certain mild condition), e.g., binomial becomes Gaussian as  $N \rightarrow \infty$ .

### 2.3.3. Geometrical form (Transformation)

- 1) Mahalanobis distance:

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (2.21)$$

- 2)  $\boldsymbol{\Sigma}$  can be symmetric = its eigenvalue is real & eigenvectors can be an orthonormal set

#### 2.3.3.1. Transform

$$\boldsymbol{\Sigma} = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^T \quad (2.22)$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^D \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i^T \quad (2.23)$$

Therefore, equation 2.21 becomes,

$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i} \quad (2.24)$$

$$y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu}) \quad (2.25)$$

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu}) \quad (2.26)$$

#### 2.3.3.2. Requirements for normalization

- 1) Positive definite: Properly normalized with elliptical shape
  - A. Center: @  $\boldsymbol{\mu}$
  - B. Axes: along  $\mathbf{u}$
  - C. Scaling factor:  $\lambda_i^{1/2}$
- 2) Positive semi-definite: subspace of lower dimensionality (singular)
- 3) Negative eigenvalue: Probability cannot be defined

#### 2.3.3.3. Normalization

$$|\mathbf{j}|^2 = 1$$

By transforming (shift, rotate) and using a new coordinate system, Multivariate Gaussian becomes the product of D independent univariate Gaussian distribution:  $\prod_{j=1}^D 1 = 1 \therefore$  Normalized

### 2.3.3.4. Properties

#### 2.3.3.4.1. First moment

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

#### 2.3.3.4.2. Second moment (Covariance)

$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \boldsymbol{\Sigma}$$

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

### 2.3.4. Limitation and Alternatives

#### 2.3.4.1. Too many parameters

Total of  $D(D+3)/2$  parameters determining Gaussian distribution (grows quadratically with  $D$ )

- 1) Just use Gaussian:  $D(D+3)/2$
- 2)  $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}_i^2)$ : 2D (axis-aligned ellipsoid)
- 3)  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ :  $D+1$  (isotropic covariance)

#### 2.3.4.2. Unimodal

Introduce latent variable as a solution

- 1) Discrete latent variable
  - A. Mixture of Gaussian
- 2) Continuous latent variable
  - A. Markov random field: to model pixel intensities of an image considering spatial organization
  - B. Linear dynamical system: to model time series data for applications such tracking
  - C. Probabilistic graphical model

### 2.3.5. Conditional & Marginal Gaussian distribution

If two sets of distributions are jointly Gaussian,



$$\mathbf{x} = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

### 2.3.5.1. Conditional distribution

$$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Lambda_{aa}^{-1}) \quad (2.27)$$

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_b) \quad (2.28)$$

\*The mean of the conditional distribution, given by (2.27), is a linear function of  $x_b$  and that the covariance is independent of  $x_b$ .

### 2.3.5.2. Marginal distribution

$$p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa}) \quad (2.29)$$

\*the mean and covariance of a marginal distribution are most simply expressed in terms of the partitioned covariance matrix whereas those of conditional distribution are well expressed by the partitioned precision matrix.

## 2.3.6. Bayes' theorem for Gaussian variables

### 2.3.6.1. Given distributions

- 1) Linear Gaussian model, i.e.,  $\mathbb{E}(y|x)$  is linear function of  $x$ .
- 2) Gaussian  $p(x)$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1}) \quad (2.30)$$

- 3) Gaussian  $p(y|x)$

$$p(y|x) = \mathcal{N}(y|Ax + b, L^{-1}) \quad (2.31)$$

### 2.3.6.2. Target distributions

- 1)  $p(y)$
- 2)  $p(x|y)$

### 2.3.6.3. Note

- 1) We are given a prior and likelihood instead of the joint distribution as in chapter 2.3.5. We will derive the equation for the target distribution with the prior and likelihood.
- 2) Find the mean, covariance, and the precision matrix for the joint distribution  $\mathbf{z}$ .

$$\mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}$$

3) Derive the target equations using 2.28 and 2.29.

#### 2.3.6.4. Results

1) Marginal (normalization term)

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.32)$$

2) Posterior distribution

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | (\mathbf{I} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1} \{ \mathbf{A}^T \mathbf{L} (\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu} \}, (\mathbf{I} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1}) \quad (2.33)$$

### 2.3.7. Maximum likelihood for the Gaussian

#### 2.3.7.1. Given observation

A data set  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$

#### 2.3.7.2. Log likelihood function

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

#### 2.3.7.3. Sufficient statistics

$$\sum_{n=1}^N \mathbf{x}_n, \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T$$

#### 2.3.7.4. ML solution for mean and variance

1) Differentiate the log likelihood by  $\boldsymbol{\mu}$  and then by  $\boldsymbol{\Sigma}$

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

$$\text{ii. } \boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{ML}) (\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T$$

2) Evaluate ML solutions under the true distribution

$$\text{i. } \mathbb{E}[\boldsymbol{\mu}_{ML}] = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$$

$$\text{ii. } \mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N} \boldsymbol{\Sigma}$$

3) Correct the bias for unbiased estimator of variance

$$\text{i. } \tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{ML}) (\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T$$

## 2.3.8. Sequential estimation

### 2.3.8.1. Particular version: mean estimator

Maximum likelihood estimator of the mean based on N observations, i.e.  $\mu_{ML}^{(N)}$

Is obtained by moving the old estimate a small amount, proportional to  $1/N$ , in the direction of the 'error' signal, i.e.  $X_N - \mu_{ML}^{(N-1)}$

$$\mu_{ML}^{(N)} = \mu_{ML}^{(N-1)} + \frac{1}{N}(X_N - \mu_{ML}^{(N-1)}) \quad (2.34)$$

### 2.3.8.2. General version: Robbins-Monro algorithm

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

Where  $z$  is an output of a function that takes  $\theta$  as its argument.

Three conditions should be satisfied:

- 1) To converge the process to a limiting value:

$$\lim_{N \rightarrow \infty} a_N = 0$$

- 2) So that the convergence does not stop earlier:

$$\sum_{N=1}^{\infty} a_N = \infty$$

- 3) so that the accumulated noise has finite variance and thus does not spoil convergence

$$\sum_{N=1}^{\infty} a_N^2 < \infty$$

General ML solution for N observations using log likelihood and finding a stationary point:

$$\frac{\partial}{\partial \theta} \left\{ -\frac{1}{N} \sum_{n=1}^N \ln p(x_n | \theta) \right\} = \mathbb{E}_x \left[ -\frac{\partial}{\partial \theta} \ln p(x | \theta) \right] = 0$$

Therefore, ML solution equals the root of a derivative of regression function. As a result,

$$\mu^{(N)} = \mu^{(N-1)} + a_{N-1} \frac{1}{\sigma^2} (x_N - \mu_{ML}^{(N-1)}) \quad (2.35)$$

If we choose  $a_{N-1}$  as  $\frac{\sigma^2}{N}$ , then 2.34 becomes equal to 2.35.

\*Regression function is

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

## 2.3.9. Bayesian inference for the Gaussian

### 2.3.9.1. Derivation

#### 2.3.9.1.1. Likelihood function

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

Mean (ML solution):  $\mu_{ML} = \sum_{n=1}^N x_n$

Variance:  $\sigma^2$

#### 2.3.9.1.2. Prior distribution

Considering the form of the likelihood choose prior distribution as a Gaussian (conjugate prior)

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

#### 2.3.9.1.3. Posterior distribution

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

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_{posterior}, \sigma_{posterior}^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML}$$

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

#### 2.3.9.1.4. Properties

- 1) Posterior mean is a compromise between the prior mean and the ML solution for the mean

- 2)  $N \rightarrow 0, \mu_{posterior} \rightarrow \mu_0$
- 3) Take a look at the form of  $\frac{1}{\sigma_{posterior}^2}$ , precision will provide more simplicity including additivity
- 4)  $N \rightarrow \infty, \mu_{posterior} \rightarrow \mu_{ML}$
- 5)  $N \rightarrow \infty, \sigma_{posterior}^2 \rightarrow 0$
- 6)  $N \rightarrow \infty$ , posterior distribution becomes infinitely peaked around the  $\mu_{ML}$
- 7) For finite  $N$  &  $\sigma_0^2 \rightarrow \infty$ ,  $\mu_{posterior} \rightarrow \mu_{ML}$  &  $\sigma_{posterior}^2 = \frac{\sigma^2}{N}$
- 8) The sequential view of Bayesian inference is very general and applies to any problem in which the observed data are assumed to be i.i.d.
- 9)