Gaussian Model

Jiaqi Hu, Jiaqi Xia

University of Science and Technology of China

September 29, 2021

Outline

Statistical Model

A statistical model is a set of probability distributions on the sample space δ: $\mathcal{P}(\delta)$. For example, $\delta := (\mathbb{R}^n, \mathbb{B}(\mathbb{R}^n))$ and $\mathcal{P}(\delta)$ contains the probability measures with distributions indexed by $\mu, \sigma^2 \in \mathbb{R}$, which density is

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right].$$

We call this kind of statistical model as the parametric model.

In general, given the data D (assumed to be sampled from δ with unknown law), we specify a parameter space $\Theta \subset \mathbb{R}^p$ corresponding to $\mathcal{P}(\delta)$ via a map $\mathcal{M} : \Theta \to \mathcal{P}(\delta)$, where $\mathcal{M}$ is a surjection and we call $\mathcal{M}$ is a parametric model for D.

Statistical Model

Let $p_\theta \in \mathcal{P}(\delta)$, $\theta \in \Theta$. If $\mathcal{M}$ is an injection, the parametric model would be identifiable, therefore, we can use the maximum likelihood estimator (MLE)

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \log [p_\theta(D)]$$

to extract a estimation of θ, hence the law of D is estimated as $p_{\hat{\theta}}$.

Model specification is a summary and reduction for data in nature.

"we may know by experience what forms are likely to be suitable, and adequacy of our choices may be tested as a posterioi."

Model diagnosis is necessary for the justification of our choice.

Statistical Model

To ease the restriction of parametric model, we introduce the non-parametric model where the model structure is not specified a prior but instead determined from data. The term of "non-parametric" has defined mainly in following two ways:

- Distribution free: the assumption for data generation is not limited to any parametric family of distributions, instead, more flexible "parametric space" (e.g. smoothed function space) is introduced as the opposite of finite-dimensional parameter space.
- Data-dependent model structure: the structure of non-parametric model is not fixed, which complexity (e.g. the order of smoothness) grows in size to accommodate the amount of data.

Statistical Model

While the non-parametric model is more general, the poor estimation efficiency of which would lead to a requirement for larger data to perform well. The dependence within data (e.g. time series, spatial random field) may also undermine the statistical efficiency due to the lack of replication. There are also the semi-parametric models combing the wisdom of parametric and non-parametric modelling.

Statistical Model

Box's memorable statement:

"All the model is wrong, but some are useful."

"It would be very remarkable if any system could be exactly represented by the simple model. However, cunningly chosen parsimonious models (Ockham's razor) often do provide remarkably useful approximations. For such a model, there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No" (non-falsifiability). The only question of interest is "Is the model illuminating and useful?"

Normal distribution

The probability density function (**pdf**) for an normal distribution is defined by:[5]

$$f(\mathbf{x} \mid \mu, \sigma) \triangleq \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}}$$

History:[10]

- 1733: De Moivre and Laplace(later) find the approximated binomial distribution from gambling.
- 1823: Gauss published his monograph "Theoria combinationis observationum erroribus minimis obnoxiae"
- Middle of the 19th century: Maxwell demonstrated that the normal distribution may also occur in natural phenomena.

Multivariate normal Distribution

The pdf for a **Multivariate normal (MVN)** in p dimensions with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is defined by the following:

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

We usually denote as $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Especially, when $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$, it is standard normal distribution. The **pdf** is:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right].$$

Why Gaussian Model?

Simplicity

- Its mean, median and mode are all same.
- The entire distribution can be specified using just mean and variance.

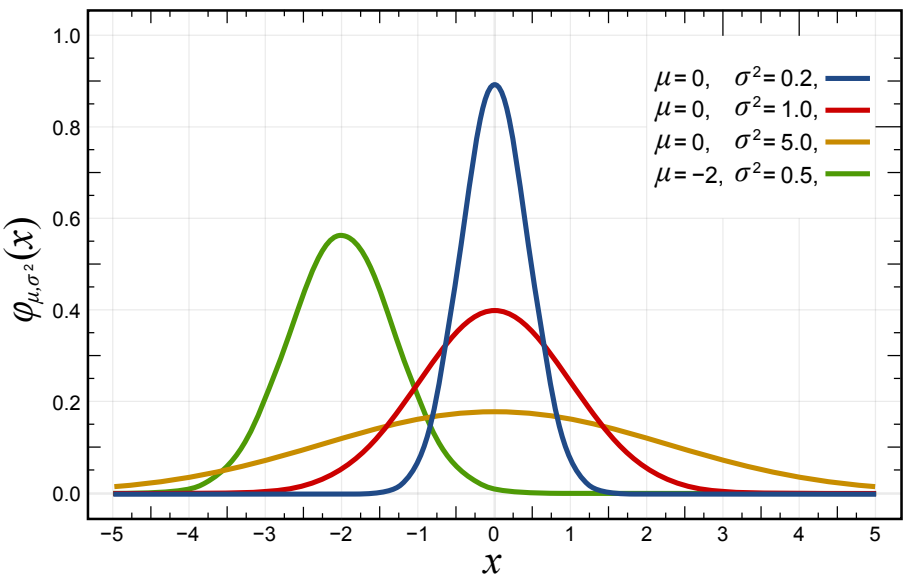


Figure 1: PDF of Normal distribution with different mean and variance

Why Gaussian Model?

Maximum entropy distribution with known mean and variance is Gaussian.

The entropy of a probability density $p(\boldsymbol{x})$ on $\mathbb{R}^n$ is defined by

$$-\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x}$$

Among all densities with a fixed mean vector $\boldsymbol{\mu} = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}$ and variance matrix $\boldsymbol{\Sigma} = \int (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T p(\boldsymbol{x}) d\boldsymbol{x}$, multivariate normal has maximum entropy. It's reasonable to assume the Gaussian assumption if we just want to focus on the mean and variance of a distribution.

Why Gaussian Model?

Ubiquitous in natural phenomena [1]

- Height, blood of pressure of human beings.
- Measurement errors.
- Position of a particle that experiences diffusion.
- The superposition of multi-source noise is usually following the Gaussian distribution (Central Limit theorem).

Why Gaussian Model?

Central Limit Theorem

- **Lindeberg–Lévy CLT**: Suppose $\{X_1, \cdots, X_n\}$ is a sequence of $i.i.d.$ random variables with $E(X_i) = \mu$, $Var(X_i) = \sigma^2 < \infty$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

- **Lyapunov CLT**: Suppose $\{X_1, \cdots, X_n\}$ is a sequence of independent random variables with $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2 < \infty$, define $s_n^2 = \sum_{i=1}^n \sigma_i^2$, if for some $\delta > 0$, Lyapunov's condition $\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0$ is satisfied, then we have

$$\frac{1}{s_n} \sum_{i=1}^n (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1).$$

Why Gaussian Model?

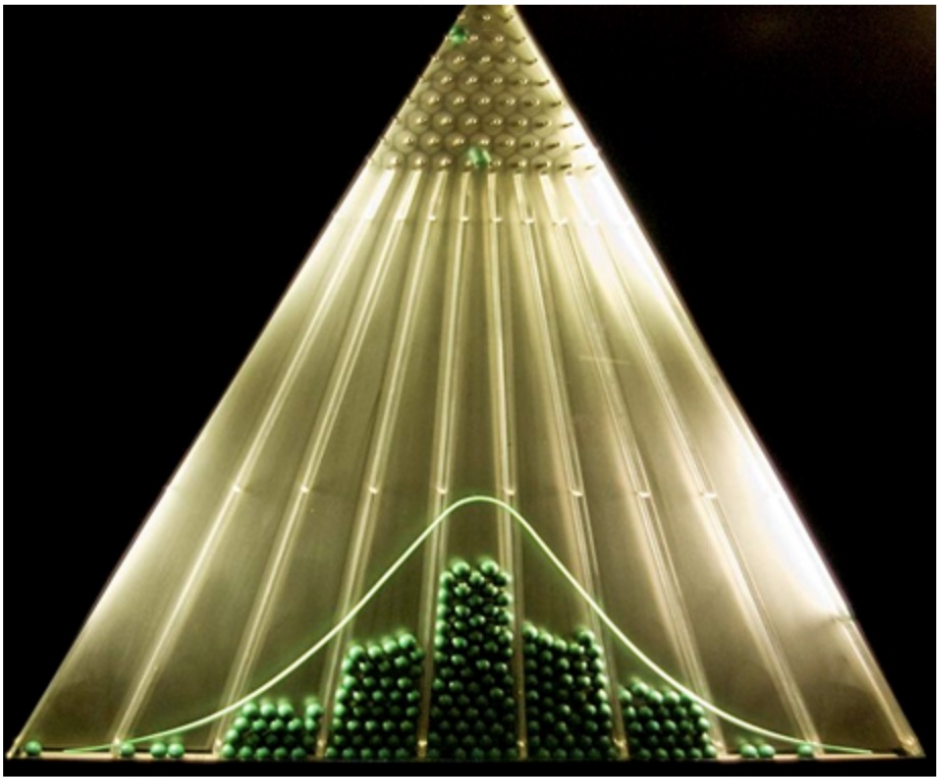
- **Lindeberg CLT**: Suppose $\{X_1, \cdots, X_n\}$ is a sequence of independent random variables with $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2 < \infty$, define $s_n^2 = \sum_{i=1}^n \sigma_i^2$, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{X_i : |X_i - \mu_i| > \varepsilon s_n\}}\right] = 0$$

  then we have

$$\frac{1}{s_n} \sum_{i=1}^n (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1).$$

Why Gaussian Model?



Source: The bean machine is called the first generator of normal random variables

Figure 2: Position of a particle that experiences diffusion.

Why Gaussian Model?

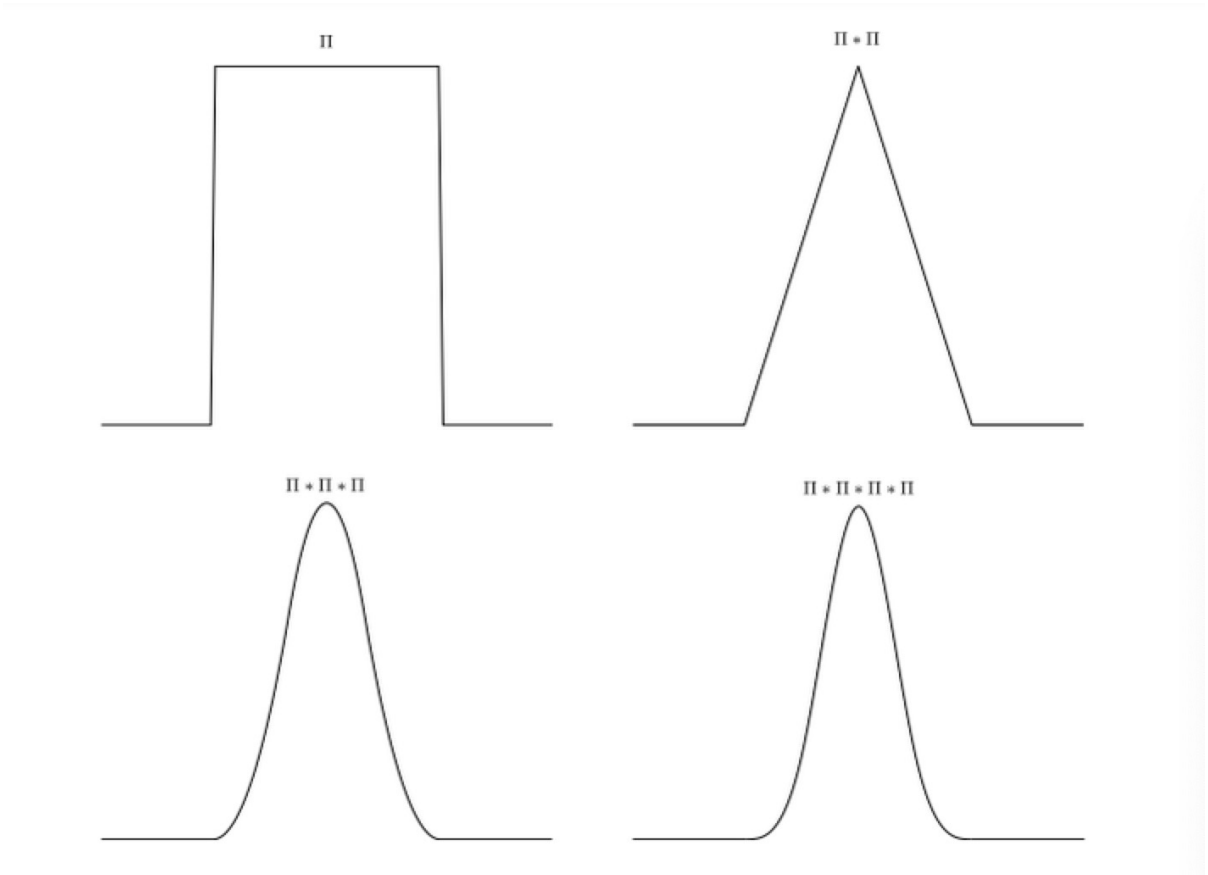


Figure 3: Convolution of independent random variables.

Conditional distribution

Theorem

Let $X \sim N_p(\mu, \Sigma), p \geq 2$. *We represent* X *as*

$$X = \begin{pmatrix} Y_{q\times 1} \\ Z_{(p-q)\times 1} \end{pmatrix}, \mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}.$$

We have $Y|Z \sim \mathcal{N}_q(\mu_{Y|Z}, \Sigma_{Y|Z})$, *where*

$$\mu_{Y|Z} = \mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z), \quad \Sigma_{Y|Z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}.$$

Conditional distribution

Proof.

Notice that two Gaussian vectors are independent iff they are uncorrelated.
Consider a linear transformation of Z via a matrix A that satisfy:
$Cov(Y - AZ, Z) = 0$, we can solve $A = \Sigma_{YZ}\Sigma_{ZZ}^{-1}$. Then
$Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}Z \perp Z$.

$$E(Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}Z) = \mu_Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\mu_Z.$$

$$Var(Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}Y) = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}.$$

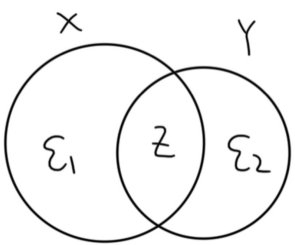
$$Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}Z|Z \sim \mathcal{N}_q(\mu_Y - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\mu_Z, \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}).$$

$$\Rightarrow Y|Z \sim \mathcal{N}_q(\mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z), \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}).$$

Conditional distribution

Some Remarks:

- $\mu_{Y|Z}$ is a linear transformation of Z. If Y is a scaler variable, we have
$$Y = \alpha + \beta^T Z + \varepsilon, \text{ where } \varepsilon \perp Z,$$
indicating that the conditional distribution of Y given Z could be divided into two independent components, one of which is the linear transformation of Z.
- $X \in \mathbb{R}^d, Y \in \mathbb{R}^q, Z \in \mathbb{R}^k$, $X \perp Y|Z$ iff $\varepsilon_1 \perp \varepsilon_2$, where $\varepsilon_1, \varepsilon_2$ are the noises, s.t. $X = E(X|Z) + \varepsilon_1, Y = E(Y|Z) + \varepsilon_2$.



Conditional distribution

Auto-regressive Model:

$$X_t = \alpha + \sum_{g=1}^{p} \beta_g X_{t-g} + \varepsilon_t,$$

where $\varepsilon_t \perp (X_s)_{s<t}$.

Figure 4: Auto-regressive Model

Conditional distribution

Kriging gives the best linear unbiased prediction (BLUP) at unsampled locations based upon the prior covariance.

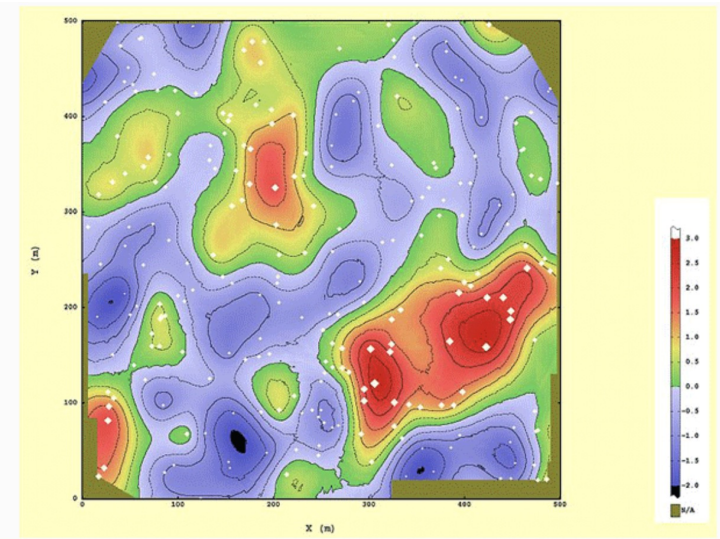


Figure 5: Example of kriging: an linear interpolated surface created by ordinary kriging based on 242 data points.

MLE for MVN

All we need to do is to extract the mean and covariance matrix from raw. Here, we firstly focus on the frequently used framework: we possess the iid samples. Further discussion would be presented when this assumption is violated.

MLE for MVN

Theorem

If we have n iid samples $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE for the the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ *is*

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Proof.

The likelihood function:

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

$$l(\mathbf{x}|\mu, \Sigma) \propto \frac{n}{2}\log|\Phi| - \frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \mu)^T\Phi(\mathbf{x}_i - \mu)\right], \text{ where } \Phi = \Sigma^{-1}$$

MLE for MVN

Proof.

$$\propto \frac{n}{2}\log|\Phi| - \frac{1}{2}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})^T\Phi(\mathbf{x}_i-\bar{\mathbf{x}}) + n(\bar{\mathbf{x}}-\mu)^T\Phi(\bar{\mathbf{x}}-\mu)\right]$$

$$\Longrightarrow \hat{\mu} = \bar{\mathbf{x}}$$

$$l(\mathbf{x}|\hat{\mu},\Sigma) \propto \frac{n}{2}\left[\log|\Phi| - \mathrm{tr}(\mathbf{C}\Phi)\right], \text{ where } \mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$$

$$\propto \frac{n}{2}\left[\log|\mathbf{A}| - \mathrm{tr}(\mathbf{A})\right], \text{ where } \mathbf{A} = \mathbf{C}\Phi, \text{is a positive definite.}$$

All eigenvalues of $\mathbf{A}$ are $\lambda_1, \cdots, \lambda_p > 0$.
$\log|\mathbf{A}| - \mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n}\log(\lambda_i) - \sum_{i=1}^{n}\lambda_i$. When $\lambda_i = 1$, it attains maximum. So $\hat{\mathbf{A}} = \mathbf{I}_p$, $\hat{\Phi} = \mathbf{C}^{-1}$, $\hat{\Sigma} = \hat{\Phi}^{-1} = \mathbf{C}$ □

Covariance Matrix Estimation

When $n \gg p$, we would simply use the sample covariance matrix to estimate the Covariance Matrix under the law of large numbers theorem.

While when $n < p$ or even $n \ll p$, sample covariance matrix based on the observed data is singular. In addition, the aggregation of massive amount of estimation errors can make considerable adverse impacts on the estimation accuracy.[3]

Covariance Matrix Estimation Example

[9]The following figures show the results of simulations for a random ensemble $\Sigma = I_d$, with each $X_i \sim N(0, I_d)$ for $i = 1, \cdots, n$, and calculate the eigenvalues $\gamma(\hat{\Sigma}) \in \mathbb{R}^d$ of the sample covariance matrix $\hat{\Sigma}$.

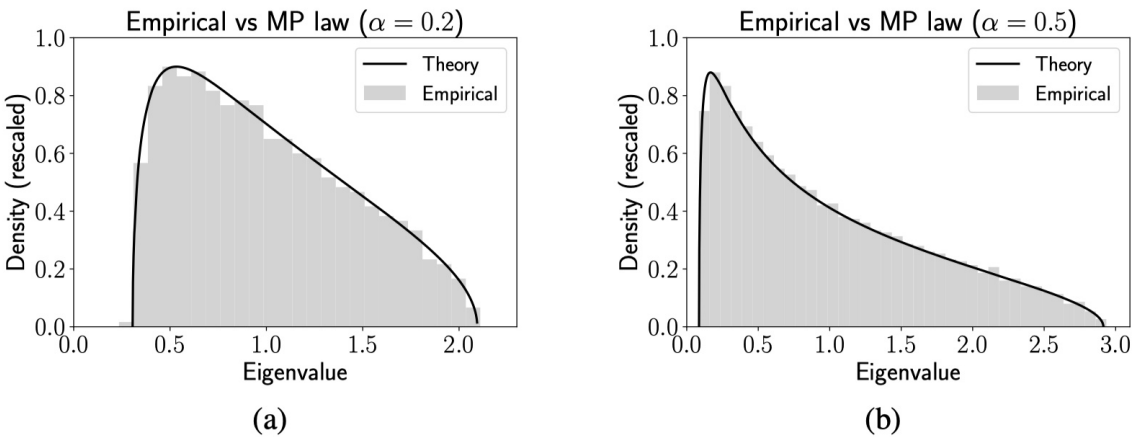


Figure 1.2 Empirical distribution of the eigenvalues of a sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ versus the asymptotic prediction of the Marčenko–Pastur law. It is specified by a density of the form $f_{\mathrm{MP}}(\gamma) \propto \sqrt{\frac{(t_{\max}(\alpha)-\gamma)\,(\gamma-t_{\min}(\alpha))}{\gamma}}$, supported on the interval $[t_{\min}(\alpha), t_{\max}(\alpha)] = [(1-\sqrt{\alpha})^2,\ (1+\sqrt{\alpha})^2]$. (a) Aspect ratio $\alpha = 0.2$ and $(n, d) = (4000, 800)$. (b) Aspect ratio $\alpha = 0.5$ and $(n, d) = (4000, 2000)$. In both cases, the maximum eigenvalue $\gamma_{\max}(\boldsymbol{\Sigma})$ is very close to $(1+\sqrt{\alpha})^2$, consistent with theory.

Discriminant Analysis

Discriminant analysis is a supervised learning technique used to classify observations according to their features.

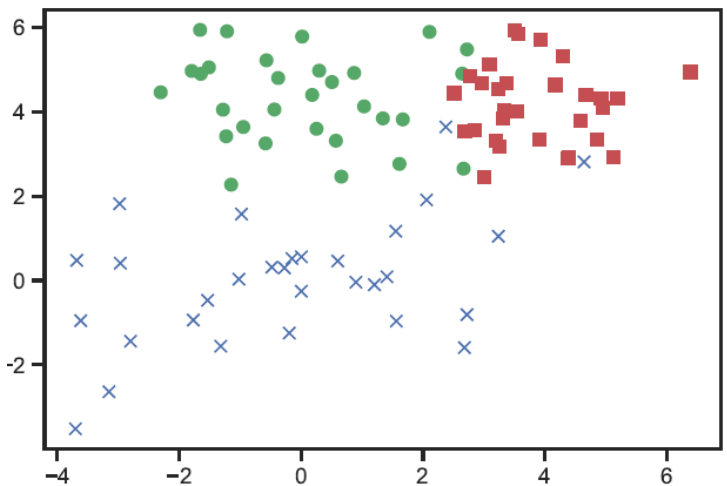


Figure 6: Labeled data points, adopted from [8].

Discriminant Analysis

Precisely, given N independent random samples (training set):

$$(X_1, G_1), \ldots, (X_N, G_N),$$

where for each $i \in \{1, \ldots, N\}$,

- $X_i \in \mathbb{R}^p$ is the feature vector;
- $G_i \in \{1, \ldots, K\}$ is the group label,

we aim to train a classifier $C : \mathbb{R}^p \to \{1, \ldots, K\}$ that can predict the unknown group labels of new observations.

Introduction

Generative model

For $k = 1, \ldots, K$, let $\pi_k := Pr(G = k)$. We suppose that

$$X|G = k \sim f_k,$$

where f_k's are to be specified.

Given a new observation with $X = x$, by the Bayes rule we have

$$Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}.$$

Clearly, it holds that

$$\arg\max_k Pr(G = k|X = x) = \arg\max_k \ f_k(x)\pi_k,$$

by which we define $\delta_k(x) := \log(f_k(x)\pi_k)$ as the **discriminant function**.

Quadratic Discriminant Analysis (QDA)

Setup

In QDA, we assume that

$$X|G = k \sim N(\mu_k, \mathbf{\Sigma}_k),\ k = 1 \ldots, K.$$

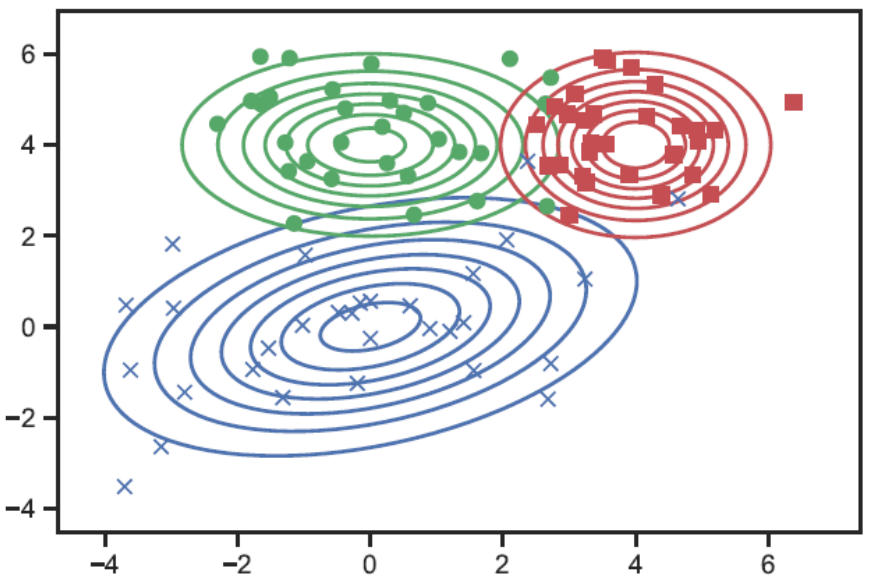


Figure 7: Fitted Gaussian distribution, adopted from [8].

Quadratic Discriminant Analysis (QDA)

Hence, each discriminant function is quadratic:

$$\delta_k(x) = \frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k.$$

Quadratic Discriminant Analysis (QDA)

Estimation

The parameters $\Theta = (\mu_1, \mathbf{\Sigma}_1, \pi_1, \ldots, \mu_K, \mathbf{\Sigma}_K, \pi_K)$ are to be estimated through the maximum likelihood estimation (MLE).

$$L(\Theta; X_1, \ldots, X_N, G_1, \ldots, G_n) = \sum_{k=1}^{K} \sum_{i:G_i=k} \Big(-\frac{1}{p}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(X_i - \mu_k)^T \mathbf{\Sigma}_k^{-1}(X_i - \mu_k) + \log \pi_k \Big).$$

Denote by $N_k = \#\{i : G_i = k\}$, the resultant MLEs are

$$\hat{\pi}_k = N_k/N; \quad \hat{\mu}_k = \frac{1}{N_k} \sum_{i:G_i=k} X_i;$$

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{N_k} \sum_{i:G_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T.$$

Quadratic Discriminant Analysis (QDA)

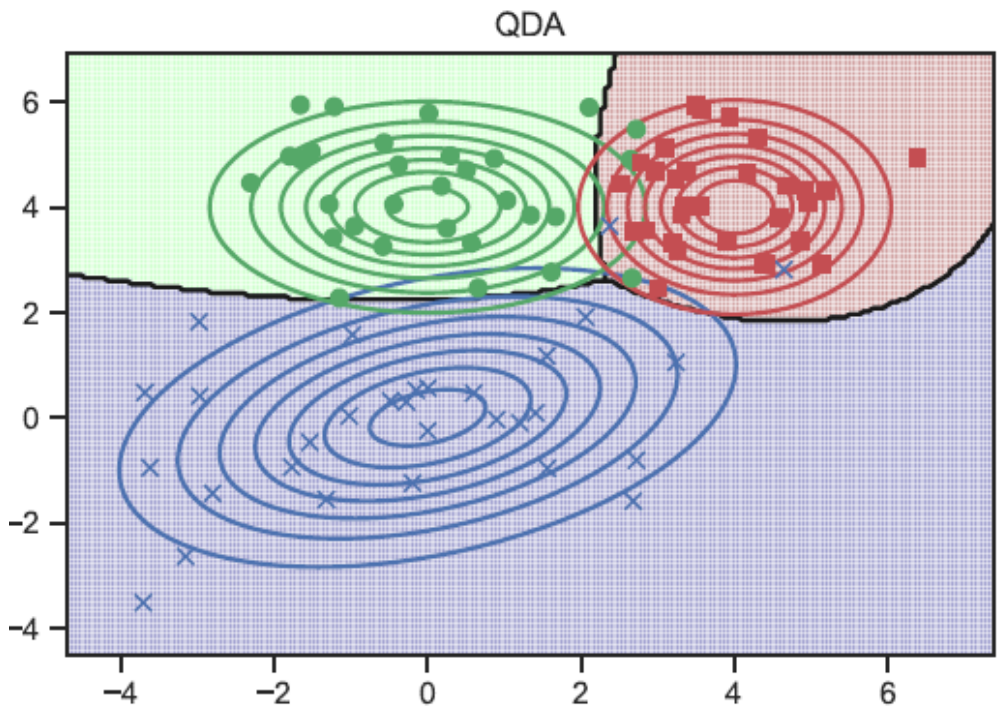


Figure 8: Decision boundaries, adopted from [8].

Linear Discriminant Analysis (LDA)

Setup

There are in total $(K-1)(\frac{p(p+1)}{2}+p+1)$ parameters in QDA. For simplification and ease of computation, LDA additionally assumes that

$$\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma},$$

by which the number of parameters is now $\frac{p(p+1)}{2} + (K-1)(p+1)$.

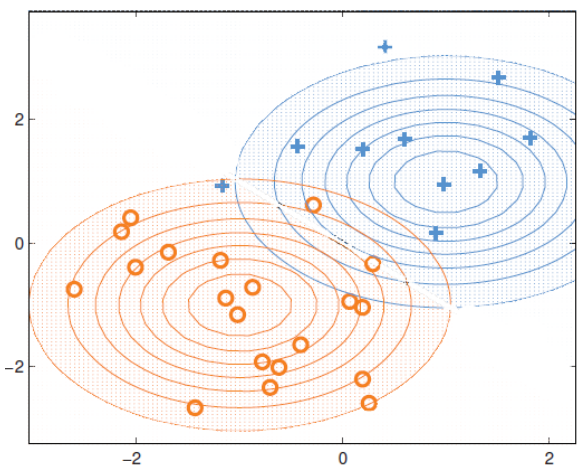


Figure 9: Fitted Gaussian distribution, edited from [7].

Linear Discriminant Analysis (LDA)

The discriminant functions can be simplified to

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k,$$

which now becomes linear.

The decision boundary between each pair of classes k and l is now a hyperplane:

$$\{x : \delta_k(x) = \delta_l(x)\}$$

$$= \{x : x^T \mathbf{\Sigma}^{-1}(\mu_l - \mu_l) - \frac{1}{2}(\mu_k - \mu_l)\mathbf{\Sigma}^{-1}(\mu_k - \mu_l) + \log(\frac{\pi_k}{\pi_l}) = 0\}.$$

Linear Discriminant Analysis (LDA)

Estimation

The log-likelihood for $\Theta = (\mu_1, \pi_1, \ldots, \mu_K, \pi_K, \mathbf{\Sigma})$ is

$$L(\Theta; X_1, \ldots, X_N, G_1, \ldots, G_n) = \sum_{k=1}^{K} \sum_{i:G_i=k} \Big(-\frac{1}{p}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(X_i - \mu_k)^T \mathbf{\Sigma}^{-1}(X_i - \mu_k) + \log \pi_k \Big).$$

The MLEs are

$$\hat{\pi}_k = N_k/N; \quad \hat{\mu}_k = \frac{1}{N_k} \sum_{i:G_i=k} X_i;$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i:G_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T.$$

Linear Discriminant Analysis (LDA)

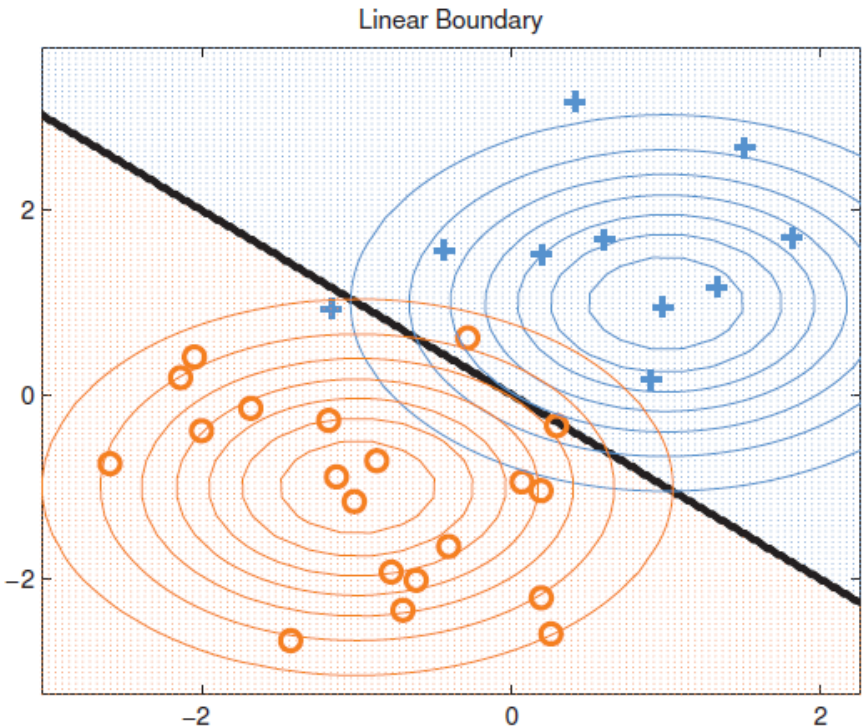


Figure 10: Decision boundary, adopted from [7].

Sufficient Dimension Reduction in LDA

Problems of high dimensionality

When p is relatively large compared to N:

- presence of collinearity, which makes $\hat{\mathbf{\Sigma}}$ ill-conditioned;
- $\hat{\mathbf{\Sigma}}$ is singular when $p > N - K$;
- some features may be noninformative, risk of overfitting;
- veiling the pattern of grouping.

Sufficient Dimension Reduction in LDA

A geometric interpretation of LDA

Look back on the original score function for LDA:

$$\log(f_k(x)\pi_k) = \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(x^T - \mu_k)^T\mathbf{\Sigma}^{-1}(x - \mu_k) + \log\pi_k.$$

Note that $(x^T - \mu_k)^T\mathbf{\Sigma}^{-1}(x - \mu_k)$ is the Mahalanobis distance between x and μ_k, the essence of LDA is two-fold:

- sphere x and calculate its Euclidean distances to the sphered centroids;
- amend the distances with corresponding prior probabilities.

Sufficient Dimension Reduction in LDA

The K centroids in R^p lie in an affine subspace of dimension $\min\{K-1, p\}$.

Comparing the distance in the original space is equivalent to first projecting the features and centroids onto the subspace and comparing the distance there.

We consider finding (at most $K-1$) projected features, which are called the **discrimant variables**.

Sufficient Dimension Reduction in LDA

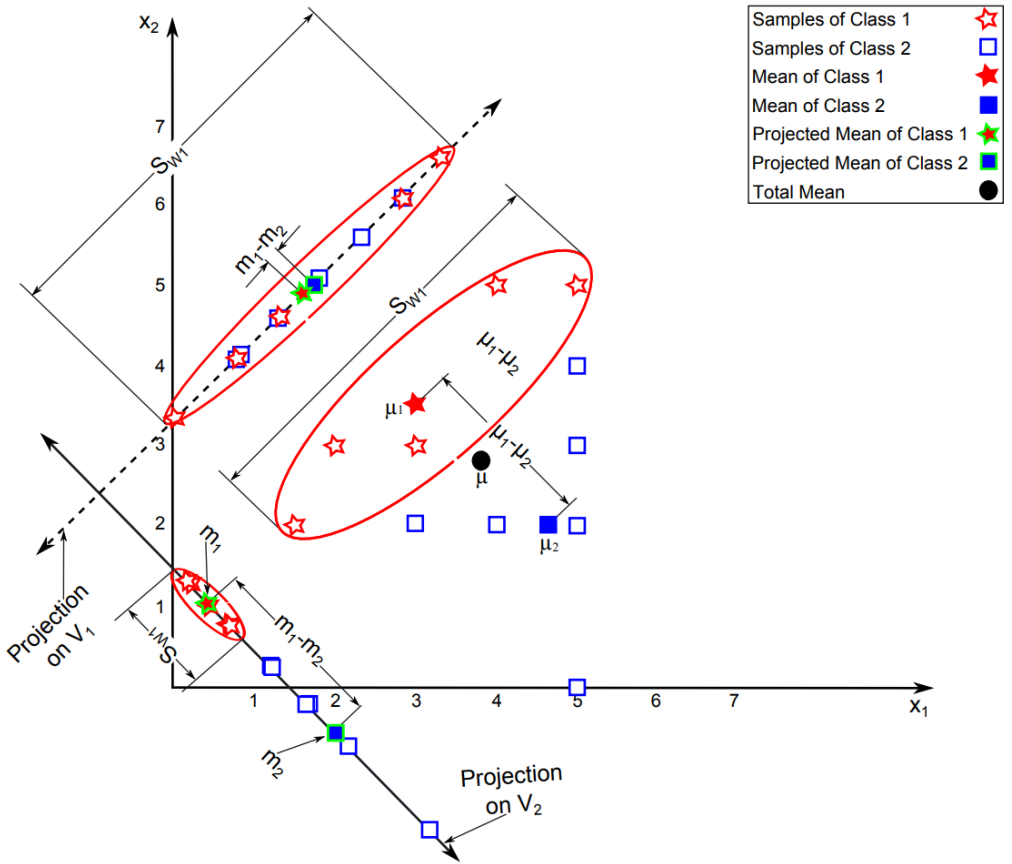


Figure 11: Different projecting directions

Sufficient Dimension Reduction in LDA

Fisher's Linear Discriminant

When $K = 2$, for $a \in \mathbb{R}^p$, let $Z = a^T X$ be the discrimant variables, [4] proposed to obtain a via

$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}.$$

where $\mathbf{B}$ and $\mathbf{W}$ are called, respectively, the between-class scatter matrix and the within-class scatter matrix:

$$\mathbf{B} = \sum_{k=1}^{K} N_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T;$$

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{i:G_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T.$$

Sufficient Dimension Reduction in LDA

That is, we find a through:

- maximize the distance between $a^T\mu_1$ and $a^T\mu_2$;
- minimizing the variance of each group after projection.

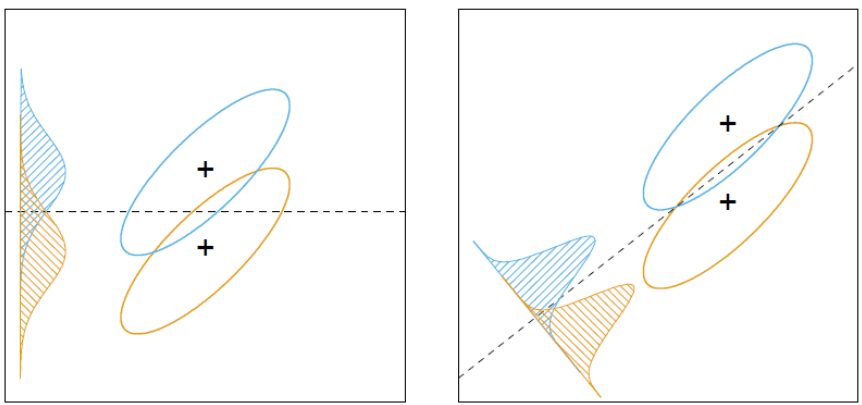


Figure 12: Left figure: maximizing the between-class distance; right figure: maximizing the within-class distance. Adopted from [6].

Hence, a is the solution of the generalized eigenvalue problem $\mathbf{B}a = \lambda\mathbf{W}a$.

Sufficient Dimension Reduction in LDA

Sufficient Dimension Reduction

When $K > 2$, for $a_i \in \mathbb{R}^p,\ i = 1, \ldots, L\ (L \leq K - 1)$, we consider calculating $\mathbf{A} = (a_1, \ldots, a_L)$.

Clearly, if

$$G|\mathbf{A}^T X \sim G|X,$$

then $\mathbf{A}^T X$ is sufficient for classification.

Sufficient Dimension Reduction in LDA

Consider the following decomposition:

$$\tilde{X}_i := \mathbf{\Sigma}^{-1/2} X_i = \nu + \mathbf{\Gamma} f_{G_i} + \varepsilon_i,\ i = 1, \ldots, N,$$

where $\nu = E\tilde{X}$; $\mathbf{\Gamma} \in \mathbb{R}^{p \times L}$, $\mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_L$; $f_{G_i} \in \{f_1, \ldots, f_K : f_i \in \mathbb{R}^L\}$ is a random vector depending on G_i, $\sum_{i=1}^{N} f_{G_i} = 0$; $\varepsilon_i \sim N(\mathbf{0}, \mathbf{I}_p)$ and is independent of G_i.

Sufficient Dimension Reduction in LDA

Then we have $G|\mathbf{\Gamma}^T\tilde{X} \sim G|\tilde{X}$ (factorization theorem).

Derivation of $\mathbf{\Gamma}$: maximizing the likelihood.

$$L(\nu, \mathbf{\Gamma}, f_1, \ldots, f_K) \propto -\frac{1}{2}\sum_{i=1}^{N}(\tilde{X}_i - \nu - \mathbf{\Gamma} f_{G_i})^T(\tilde{X}_i - \nu - \mathbf{\Gamma} f_{G_i}),$$

Fixing $\mathbf{\Gamma}$ yields:

$$\hat{\nu} = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}_i;$$

$$\hat{f}_k = \mathbf{\Gamma}^T(\hat{\nu}_k - \hat{\nu}), \text{ where } \hat{\nu}_k = \frac{1}{N_k}\sum_{i:G_i=k}\tilde{X}_i.$$

Sufficient Dimension Reduction in LDA

$$
\begin{aligned}
& L(\hat{\nu}, \mathbf{\Gamma}, \hat{f}_1, \ldots, \hat{f}_K) \\
\propto & -\sum_{i=1}^{N} \left(\tilde{X}_i - \hat{\nu} - \mathbf{\Gamma}\mathbf{\Gamma}^T(\hat{\nu}_{G_i} - \hat{\nu})\right)^T \left(\tilde{X}_i - \hat{\nu} - \mathbf{\Gamma}\mathbf{\Gamma}^T(\hat{\nu}_{G_i} - \hat{\nu})\right) \\
= & -\sum_{i=1}^{N} \left(\tilde{X}_i - \hat{\nu}_{G_i} + \mathbf{P}(\hat{\nu}_{G_i} - \hat{\nu})\right)^T \left(\tilde{X}_i - \hat{\nu}_{G_i} + \mathbf{P}(\hat{\nu}_{G_i} - \hat{\nu})\right) \\
= & -tr(\mathbf{W}_{\tilde{X}} + \mathbf{P}\mathbf{B}_{\tilde{X}}\mathbf{P}) \propto tr(\mathbf{\Gamma}^T\mathbf{B}_{\tilde{X}}\mathbf{\Gamma}),
\end{aligned}
$$

where $\mathbf{W}_{\tilde{X}}$ and $\mathbf{B}_{\tilde{X}}$ are, respectively, the within-class and between-class scatter matrix of $\tilde{X}$ and $\mathbf{P} = \mathbf{I}_p - \mathbf{\Gamma}\mathbf{\Gamma}^T$.

Sufficient Dimension Reduction in LDA

Note that $\mathbf{A}^T = \mathbf{\Gamma}^T \mathbf{\Sigma}^{1/2}$, $\mathbf{B}_{\tilde{X}} = \mathbf{\Sigma}^{-1/2} \mathbf{B} \mathbf{\Sigma}^{-1/2}$ and $\hat{\mathbf{\Sigma}} = \mathbf{W}$, the induced optimization problem is:

$$\max_{\mathbf{A}} \ tr(\mathbf{A}^T \mathbf{B} \mathbf{A}),$$
$$\text{s.t. } \mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{I}_L.$$

Sufficient Dimension Reduction in LDA

Implementation (when $\mathbf{W}$ is nonsingular)

1.Supervised dimension reduction

- Compute $\mathbf{B}^* = \mathbf{W}^{-\frac{1}{2}}\mathbf{B}\mathbf{W}^{-\frac{1}{2}}$.
- Compute the eigen-decomposition of $\mathbf{B}^*$: $\mathbf{B}^* = \mathbf{V}^*\mathbf{D}_\mathbf{B}(\mathbf{V}^*)^T$.
- Denote by v_l^* the lth column of $\mathbf{V}^*$, then $a_l = \mathbf{W}^{-\frac{1}{2}}v_l^*$.
- The projection matrix is $\mathbf{A} = (a_1, \ldots, a_L)$, $L \leq K-1$.

2.Classification

- Compute $\mathbf{M} = (\hat{\mu}_1, \ldots, \hat{\mu}_K)^T$.
- Compute $\mathbf{M}^* = \mathbf{M}\mathbf{W}^{-\frac{1}{2}}$ using the eigen-decomposition of $\mathbf{W}$.
- Compare the Euclidean distances between $\mathbf{A}^T x$ and the rows of $\mathbf{M}^*$.

Sufficient Dimension Reduction in LDA

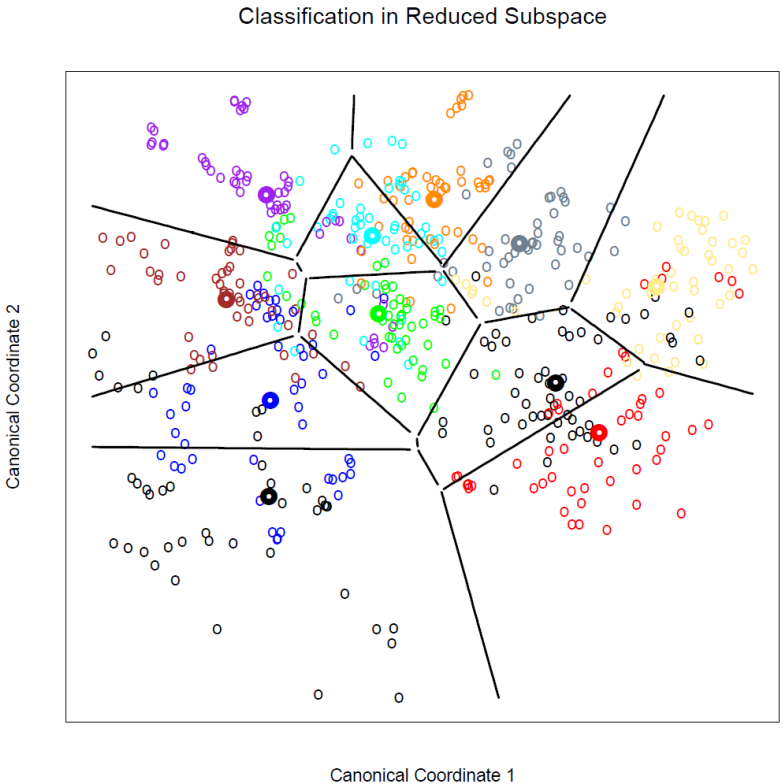


Figure 13: The projected vowel training data with $K = 10$, $p = 10$, and $L = 2$. Adopted from [6].

Sufficient Dimension Reduction in LDA

Relation to Principal Component Analysis

If $N_1 = N_2 = \cdots = N_K = 1$, we have

$$\mathbf{B} = \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T = \hat{\mathbf{\Sigma}};$$

$$\mathbf{W} = 0.$$

The criterion can be modified to

$$\max_{\mathbf{A}} \ tr(\mathbf{A}^T \hat{\mathbf{\Sigma}} \mathbf{A}),$$

$$\text{s.t. } \mathbf{A}^T \mathbf{A} = \mathbf{I}_L,$$

which is equivalent to principal component analysis.

References I

[1] Why data scientists love gaussian?

[2] R. Dennis Cook.
Fisher Lecture: Dimension Reduction in Regression.
Statistical Science, 22(1):1 – 26, 2007.

[3] Jianqing Fan, Yuan Liao, and Han Liu.
An overview of the estimation of large covariance and precision matrices.
The Econometrics Journal, 19(1):C1–C32, 2016.

[4] R. A. FISHER.
The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7(2):179–188, 1936.

[5] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al.
The elements of statistical learning, volume 1.
Springer series in statistics New York, 2001.

References II

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer New York, 2009.

[7] Kevin P Murphy.
Machine learning: a probabilistic perspective.
MIT press, 2012.

[8] Kevin P. Murphy.
Probabilistic Machine Learning: An introduction.
MIT Press, 2022.

[9] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.

References III

[10] Wikipedia contributors.
Normal distribution — Wikipedia, the free encyclopedia, 2021.
[Online; accessed 14-September-2021].

[11] Wikipedia contributors.
Statistical model — Wikipedia, the free encyclopedia, 2021.
[Online; accessed 14-September-2021].