

第四章 习题课

第1章：随机事件与概率

第2章：随机变量及其分布

第3章：多维随机变量及其分布

第4章：随机变量的数字特征

第四章 习题课

- 第四章：随机变量的数字特征
 - §1 数学期望
 - §2 方差
 - §3 协方差及相关系数
 - §4 矩、协方差矩阵

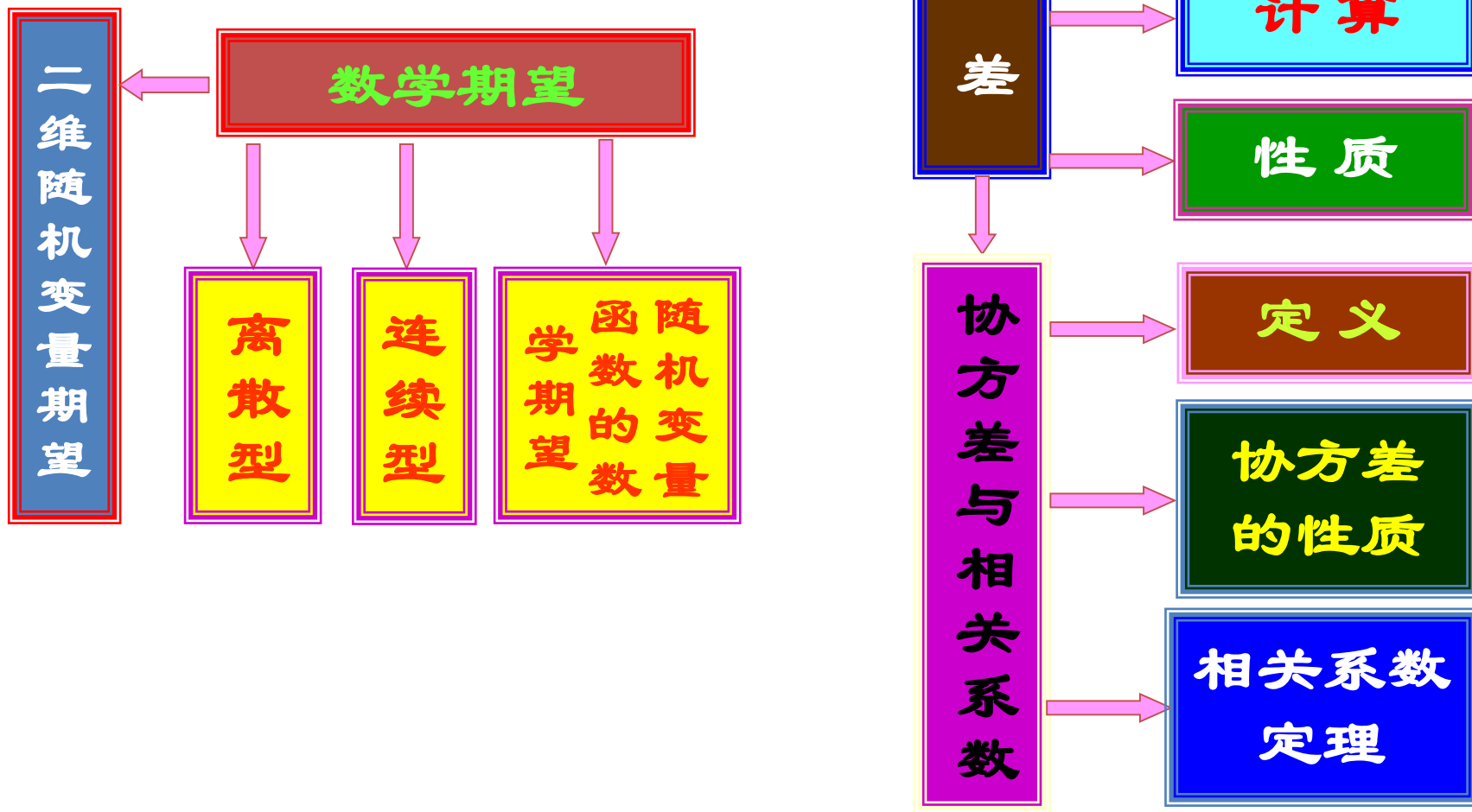
一、重点与难点

二、主要内容

三、典型例题

四、拓展：主成分分析

二、主要内容



§1 数学期望 - 离散型随机变量

设离散型随机变量 X 的分布律为

$$P\{X = x_k\} = p_k, \quad k = 1, 2, \dots$$

若级数 $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛,

则称级数 $\sum_{k=1}^{\infty} x_k p_k$ 为随机变量 X 的数学期望,

记为 $E(X)$, 即 $E(X) = \sum_{k=1}^{\infty} x_k p_k$.

§1 数学期望 - 连续型随机变量

X 是连续型随机变量, 它的概率密度为 $f(x)$,

若积分 $\int_{-\infty}^{+\infty} xf(x) dx$ 绝对收敛,

则称此积分值为连续型随机变量 X 的数学期望, 记为 $E(X)$,

即 $E(X) = \int_{-\infty}^{+\infty} xf(x) dx$.

§1 数学期望 - 随机变量函数的数学期望

离散型随机变量函数的数学期望为

若 $Y = g(X)$, 且 $P\{X = x_k\} = p_k$, ($k = 1, 2, \dots$)

则有 $E(g(X)) = \sum_{k=1}^{\infty} g(x_k) p_k.$

若 X 是连续型的, 它的分布密度为 $f(x)$,

则有 $E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx.$

§1 数学期望 - 性质

1. 设 C 是常数, 则有 $E(C) = C$.
2. 设 X 是一个随机变量, C 是常数, 则有

$$E(CX) = CE(X).$$

3. 设 X, Y 是两个随机变量, 则有

$$E(X + Y) = E(X) + E(Y).$$

4. 设 X, Y 是相互独立的随机变量, 则有

$$E(XY) = E(X)E(Y).$$

§1 数学期望 - 二维随机变量的数学期望

设 (X, Y) 为二维随机变量, 若 $E(X), E(Y)$ 都存在, 则其期望值定义为

$$E(X) = \begin{cases} \sum_i \sum_j x_i p_{ij}, & (X, Y) \text{ 的概率分布为 } p_{ij}; \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f(x, y) dx dy, & (X, Y) \text{ 的密度为 } f(x, y). \end{cases}$$

同理可得

$$E(Y) = \begin{cases} \sum_i \sum_j y_i p_{ij}, & (X, Y) \text{ 的概率分布为 } p_{ij}; \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x, y) dx dy, & (X, Y) \text{ 的密度为 } f(x, y). \end{cases}$$

第四章 习题课

1. 若 X, Y 为离散型随机变量, $g(x, y)$ 是二元函数,

$$\text{则 } E[g(X, Y)] = \sum_i \sum_j g(x_i, y_j) p_{ij},$$

当 (X, Y) 的联合概率分布为 p_{ij} .

2. 若 X, Y 为连续型随机变量, $g(x, y)$ 是二元函数,

$$\text{则 } E[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy,$$

当 (X, Y) 的联合分布密度为 $f(x, y)$.

§2 方差 - 定义

设 X 是一个随机变量 若 $E\{[X - E(X)]^2\}$ 存在, 则称 $E\{[X - E(X)]^2\}$ 是 X 的方差, 记作

$$D(X) \quad \text{或} \quad \text{Var}(X),$$

即 $D(X) = \text{Var}(X) = E\{[X - E(X)]^2\}$,
称 $\sqrt{D(X)}$ 为标准差或均方差, 记为 $\sigma(X)$.

§2 方差 - 计算

$$D(X) = E(X^2) - [E(X)]^2.$$

离散型随机变量的方差

$$D(X) = \sum_{k=1}^{\infty} [x_k - E(X)]^2 p_k,$$

其中 $P\{X = x_k\} = p_k$, $k = 1, 2, \dots$ 是 X 的分布律.

连续型随机变量的方差

$$D(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x) dx,$$

其中 $f(x)$ 为概率密度.

§2 方差 - 性质

1. 设 C 是常数, 则有 $D(C) = 0$.

2. 设 X 是一个随机变量, C 是常数, 则有

$$D(CX) = C^2 D(X).$$

3. 设 X, Y 相互独立, $D(X), D(Y)$ 存在, 则

$$D(X \pm Y) = D(X) + D(Y).$$

4. $D(X) = 0$ 的充要条件是 X 以概率 1 取常数 C , 即

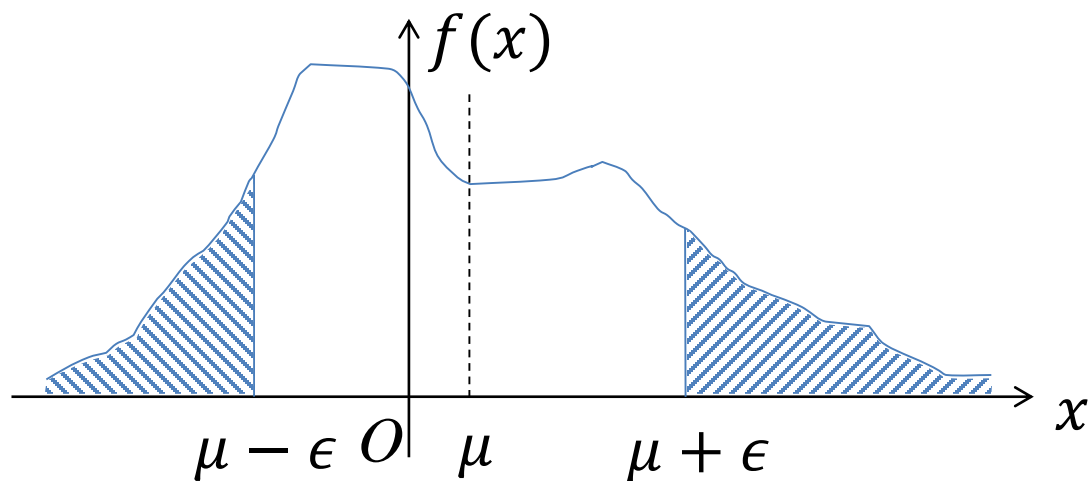
$$P\{X = C\} = 1.$$

§2 方差 - 性质

定理(切比雪夫(Chebyshev)不等式)

设随机变量 X 具有数学期望 $E(X) = \mu$ ，方差 $D(X) = \sigma^2$ ，则对于任意正数 ϵ ，则有

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \quad , \quad \Leftrightarrow P\{|X - \mu| < \epsilon\} \geq 1 - \frac{\sigma^2}{\epsilon^2}$$



§3 协方差与相关系数 - 定义

$E\{[X - E(X)][Y - E(Y)]\}$ 称为随机变量 X 与 Y 的协方差, 记为 $\text{Cov}(X, Y)$,

$$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}.$$

称 $\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \cdot \sqrt{D(Y)}}$ 为随机变量 X 与 Y 的相关系数.

§3 协方差与相关系数 - 性质

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X).$

2. $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$ (a, b 为常数)

3. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y).$

§3 协方差与相关系数 - 定理

(1) $|\rho_{XY}| \leq 1.$

(2) $|\rho_{XY}| = 1$ 的充要条件是：存在常数 a, b 使

$$P\{Y = a + bX\} = 1.$$

正态分布

如果连续型随机变量 X 的密度函数为

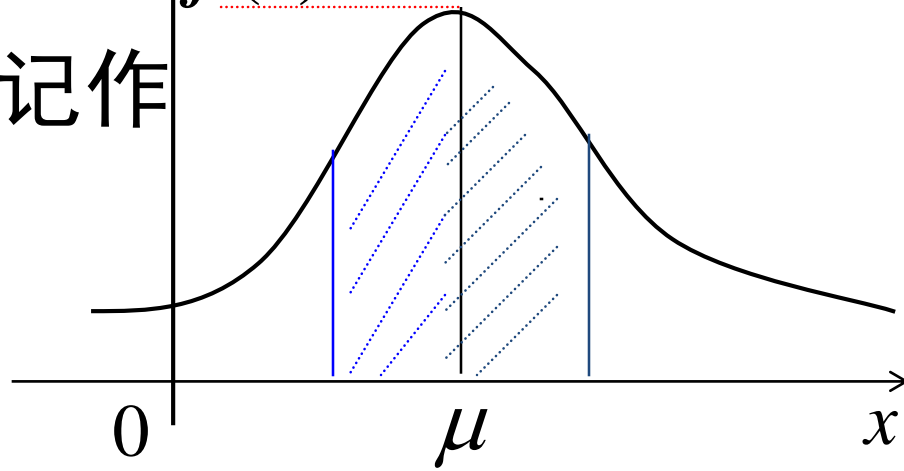
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

(其中 $-\infty < \mu < +\infty$, $\sigma > 0$ 为参数),

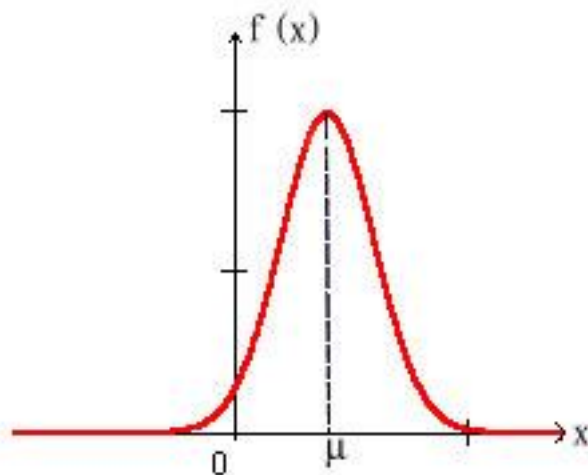
则称随机变量 X 服从参数为 $f(x)$

(μ, σ^2) 的正态分布. 记作

$$X \sim N(\mu, \sigma^2)$$



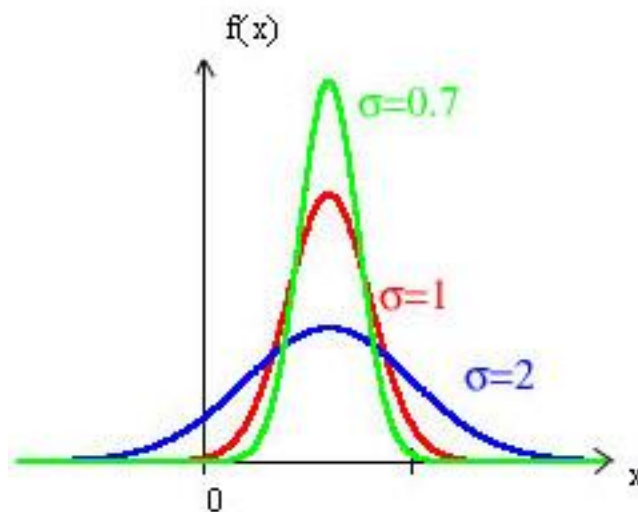
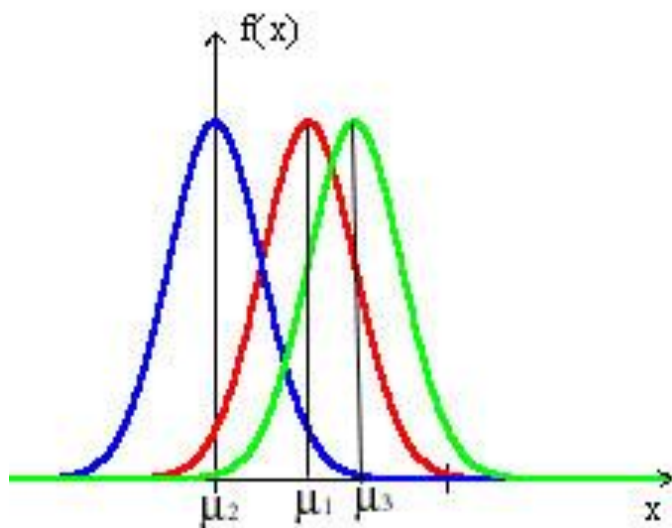
正态分布 $N(\mu, \sigma^2)$ 的图形特点



正态分布的密度曲线是一条关于 μ 对称的钟形曲线.

特点是“两头小，中间大，左右对称”.

正态分布 $N(\mu, \sigma^2)$ 的图形特点



μ 决定了图形的中心位置， σ 决定了图形中峰的陡峭程度。

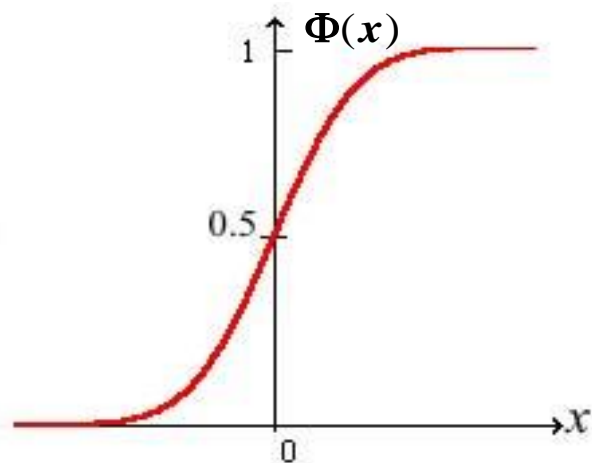
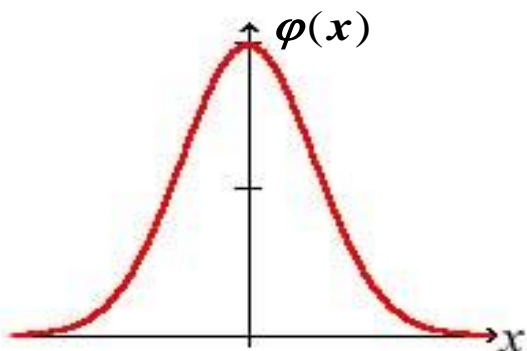
标准正态分布

$\mu = 0, \sigma = 1$ 的正态分布称为标准正态分布.

其密度函数和分布函数常用 $\varphi(x)$ 和 $\Phi(x)$ 表示:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$



3 σ 准则

$$P\{|X - \mu| < \epsilon\} \geq 1 - \frac{\sigma^2}{\epsilon^2}$$

由标准正态分布的查表计算可以求得，
当 $X \sim N(0, 1)$ 时，

切比雪夫界

$$P(|X| \leq 1) = 2\Phi(1) - 1 = 0.6826$$

0

$$P(|X| \leq 2) = 2\Phi(2) - 1 = 0.9544$$

0.75

$$P(|X| \leq 3) = 2\Phi(3) - 1 = 0.9974$$

0.89

这说明， X 的取值几乎全部集中在 $[-3, 3]$ 区间内，超出这个范围的可能性仅占不到 0.3%.

第四章 习题课

| 分 布 | 参 数 | 数学期望 | 方差 |
|------|----------------------------|-----------|--------------|
| 两点分布 | $0 < p < 1$ | p | $p(1-p)$ |
| 二项分布 | $n \geq 1,$ $0 < p < 1$ | np | $np(1-p)$ |
| 泊松分布 | $\lambda > 0$ | λ | λ |
| 均匀分布 | $a < b$ | $(a+b)/2$ | $(b-a)^2/12$ |
| 指数分布 | $\theta > 0$ | θ | θ^2 |
| 正态分布 | $\mu, \sigma > 0$ | μ | σ^2 |

三、典型例题

例1 设 X 服从几何分布, 它的分布律为

$$P\{X = k\} = (1 - p)^{k-1}p, \quad k = 1, 2, \dots$$

求 $E(X)$ 和 $D(X)$.

解

$$E(X) = \sum_{k=1}^{\infty} k \cdot q^{k-1} p \quad (\text{其中 } q = 1 - p)$$

$$= p \sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{p}{(1-q)^2} = \frac{1}{p},$$

乘以q相减

三、典型例题

例1 设 X 服从几何分布, 它的分布律为

$$P\{X = k\} = (1 - p)^{k-1}p, \quad k = 1, 2, \dots$$

求 $E(X)$ 和 $D(X)$.

解

$$E(X) = \sum_{k=1}^{\infty} k \cdot q^{k-1}p = \frac{1}{p},$$

$$E(X^2) = \sum_{k=1}^{\infty} k^2 \cdot q^{k-1}p = p \sum_{k=1}^{\infty} k^2 \cdot q^{k-1} = \frac{p(1+q)}{(1-q)^3} = \frac{1+q}{p^2}$$

$$D(X) = E(X^2) - [E(X)]^2 = \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.$$

第四章 习题课

例2 从数字 $0, 1, 2, \dots, n$ 中任取两个不同的数字, 求这两个数字之差的绝对值的数学期望.

解 设 X 为所选的两个数字之差的绝对值,
则 X 的所有可能取值为 $1, 2, 3, \dots, n$,

$$P\{X = 1\} = n/C_{n+1}^2, P\{X = 2\} = (n-1)/C_{n+1}^2$$

一般的
$$P\{X = k\} = \frac{(n-k+1)}{C_{n+1}^2}, k = 1, 2, \dots, n.$$

$$E(X) = \sum_{k=1}^n kP\{X = k\} = \sum_{k=1}^n k \cdot (n-k+1)/C_{n+1}^2 = \frac{n+2}{3}$$

$(n+1)^3$ 展
开累加

第四章 习题课

例3 设随机变量 X 取非负整数值 $n \geq 0$ 的概率为 $p_n = \frac{AB^n}{n!}$, 已知 $E(X) = a$, 求 A 与 B 的值.

解 因为 p_n 是 X 的分布列,
$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$\sum_{n=0}^{\infty} P\{X = n\} = \sum_{n=0}^{\infty} A \cdot \frac{B^n}{n!} = Ae^B = 1, \quad \text{得 } A = e^{-B},$$

$$E(X) = \sum_{n=0}^{\infty} nA \cdot \frac{B^n}{n!} = \sum_{n=1}^{\infty} \frac{A \cdot B^n}{(n-1)!} = AB e^B = a,$$

因此 $A = e^{-a}$, $B = a$.

第四章 习题课

例4 (Laplace配对) 将 n 只球 (1~ n 号) 随机地放入 n 个盒子 (1~ n 号) , 一个盒子装一只球, 若一只球装入与球同号的盒子中, 则称为一个配对. 记 X 为总的配对数, 求其期望?

解 引入随机变量 X_i ,

$$X_i = \begin{cases} 1, & \text{第 } i \text{ 个球放入第 } i \text{ 个盒子,} \\ 0, & \text{其他,} \end{cases} \quad i = 1, 2, \dots, n.$$

则
$$X = X_1 + X_2 + \dots + X_n.$$

第四章 习题课

$$\text{则有 } P\{X_i = 1\} = \frac{1}{n}, \quad P\{X_i = 0\} = 1 - \frac{1}{n},$$

$$\text{由此 } E(X_i) = \frac{1}{n}, \quad i = 1, 2, \dots.$$

$$\begin{aligned} \text{得 } E(X) &= E(X_1 + X_2 + \dots + X_n) \\ &= E(X_1) + E(X_2) + \dots + E(X_n) \\ &= n \times \frac{1}{n} = 1 \end{aligned}$$

第四章 习题课

例5 设随机变量 X 的密度函数为

$$f(x) = \begin{cases} c(1-x^2)^\alpha, & -1 < x < 1, \\ 0, & \text{其他.} \end{cases} \quad (\alpha < 0)$$

求 $E(X)$ 和 $D(X)$.

解 因为 $f(x)$ 是偶函数,

$$\text{所以 } E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-1}^1 cx(1-x^2)^\alpha dx = 0,$$

$$D(X) = E(X^2) - [E(X)]^2 = E(X^2)$$

第四章 习题课

$$= \int_{-1}^1 c x^2 (1-x^2)^\alpha dx$$

$$= -\frac{c}{2(\alpha+1)} x(1-x^2)^{\alpha+1} \Big|_{-1}^1 + \frac{c}{2(\alpha+1)} \int_{-1}^1 (1-x^2)^{\alpha+1} dx$$

$$= \frac{1}{2(\alpha+1)} \int_{-1}^1 c(1-x^2)^\alpha dx - \frac{1}{2(\alpha+1)} \int_{-1}^1 c x^2 (1-x^2)^\alpha dx$$

$$= \int_{-\infty}^{+\infty} f(x) dx = 1 \quad = \int_{-\infty}^{+\infty} x^2 f(x) dx = D(X)$$

$$\text{于是 } D(X) = \frac{1}{2(\alpha+1)} - \frac{1}{2(\alpha+1)} D(X),$$

$$\text{故 } D(X) = \frac{1}{2\alpha+3}.$$

第四章 习题课

例6 设随机变量 X 的概率密度 $f(x) = \frac{1}{\pi(1+x^2)}$,

求 $E[\min(|X|, 1)]$.

$$\begin{aligned}\text{解} \quad E[\min(|X|, 1)] &= \int_{-\infty}^{+\infty} \min(|x|, 1) f(x) dx \\&= \int_{|x| < 1} |x| f(x) dx + \int_{|x| \geq 1} f(x) dx \\&= \frac{1}{\pi} \int_{-1}^1 \frac{|x|}{1+x^2} dx + \frac{1}{\pi} \int_{|x| \geq 1} \frac{1}{1+x^2} dx \\&= \frac{2}{\pi} \int_0^1 \frac{x}{1+x^2} dx + \frac{2}{\pi} \int_1^{+\infty} \frac{1}{1+x^2} dx = \frac{1}{\pi} \ln 2 + \frac{1}{2}.\end{aligned}$$

第四章 习题课

例7 设二维连续型随机变量 (X, Y) 的联合密度

$$\text{函数为 } f(x, y) = \begin{cases} \frac{1}{2} \sin(x + y), & 0 \leq x \leq \frac{\pi}{2}, 0 \leq y \leq \frac{\pi}{2}, \\ 0, & \text{其他} \end{cases}$$

且 $Z = \cos(X + Y)$, 求 $E(Z)$ 和 $D(Z)$.

解 $E(Z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cos(x + y) f(x, y) dx dy$

$$\sin \alpha \cos \beta = \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)]$$

$$\cos \alpha \sin \beta = \frac{1}{2} [\sin(\alpha + \beta) - \sin(\alpha - \beta)]$$

$$\cos \alpha \cos \beta = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$$

$$\sin \alpha \sin \beta = -\frac{1}{2} [\cos(\alpha + \beta) - \cos(\alpha - \beta)]$$

$$= \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \frac{1}{2} \cos(x + y) \sin(x + y) dx dy$$

$$\int \sin x = -\cos x + C \quad = \int_0^{\frac{\pi}{2}} \frac{1}{2} [\cos 2x - \cos(\pi + 2x)] dx = 0,$$

第四章 习题课

$$D(Z) = E(Z^2)$$

$$= \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \frac{1}{2} \cos^2(x+y) \sin(x+y) dx dy$$

$$= \frac{1}{6} \int_0^{\frac{\pi}{2}} \left[\cos^3 x - \cos^3 \left(x + \frac{\pi}{2} \right) \right] dx$$

$$= \frac{2}{9}.$$

例8 设二维连续型随机变量 (X, Y) 的联合密度

$$\text{函数为 } f(x, y) = \begin{cases} \frac{6}{7}(x^2 + \frac{1}{2}xy), & 0 < x < 1, 0 < y < 2, \\ 0, & \text{其他} \end{cases}$$

求 (X, Y) 的协方差矩阵及相关系数.

$$\begin{aligned} \text{解 } E(X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f(x, y) dx dy \\ &= \int_0^1 \int_0^2 \frac{6}{7} x (x^2 + \frac{1}{2}xy) dy dx = \int_0^1 \left(\frac{12}{7} x^3 + \frac{6}{7} x^2 \right) dx \\ &= \frac{5}{7}, \end{aligned}$$

第四章 习题课

$$E(X^2) = \int_0^1 \int_0^2 \frac{6}{7} x^2 (x^2 + \frac{1}{2} xy) dx dy = \frac{39}{70},$$

$$\text{故 } D(X) = \frac{39}{70} - \left(\frac{5}{7}\right)^2 = \frac{23}{490},$$

$$\text{因为 } E(Y) = \int_0^1 \int_0^2 \frac{6}{7} y (x^2 + \frac{1}{2} xy) dy dx = \frac{8}{7},$$

$$E(Y^2) = \int_0^1 \int_0^2 \frac{6}{7} y^2 (x^2 + \frac{1}{2} xy) dy dx = \frac{34}{21},$$

$$\text{故 } D(Y) = \frac{34}{21} - \left(\frac{8}{7}\right)^2 = \frac{46}{147},$$

第四章 习题课

$$E(XY) = \int_0^1 \int_0^2 \frac{6}{7} xy(x^2 + \frac{1}{2}xy) dy dx = \frac{17}{21},$$

$$\text{故 } \text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$= \frac{17}{21} - \frac{5}{7} \times \frac{8}{7} = -\frac{1}{147},$$

$$\text{于是 } (X, Y) \text{ 的协方差矩阵为 } \begin{pmatrix} \frac{23}{490} & -\frac{1}{147} \\ -\frac{1}{147} & \frac{46}{147} \end{pmatrix}.$$

$$X \text{ 与 } Y \text{ 的相关系数 } \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = -\frac{\sqrt{15}}{69}.$$

例9 假设一批种子的良种率为 $\frac{1}{6}$ ，从中任意选出600粒，试用切比雪夫 (Chebyshev) 不等式估计：这600粒种子中良种所占比例与 $\frac{1}{6}$ 之差的绝对值不超过0.02的概率。

解： 设 X 表示600粒种子中的良种数则 $X \sim B(600, \frac{1}{6})$.

$$EX = 600 \times \frac{1}{6} = 100, \quad DX = 600 \times \frac{1}{6} \times \frac{5}{6} = \frac{250}{3}.$$

$$P\left\{\left|\frac{X}{600} - \frac{1}{6}\right| \leq 0.02\right\} = P\left\{\left|\frac{X - 100}{600}\right| \leq 0.02\right\} = P\{|X - 100| \leq 12\}.$$

由切比雪夫不等式有

$$P\{|X - 100| \leq 12\} \geq 1 - \frac{DX}{12^2} = 1 - \frac{\frac{250}{3}}{144} = 0.4213.$$

四、拓展：主成分分析

Principal Component Analysis

回顾：协方差矩阵

设 n 维随机变量 (X_1, X_2, \dots, X_n) 的二阶混合中心矩

$$c_{ij} = \text{Cov}(X_i, X_j) = E\{[X_i - E(X_i)][X_j - E(X_j)]\}$$
$$i, j = 1, 2, \dots, n$$

都存在, 则称矩阵

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

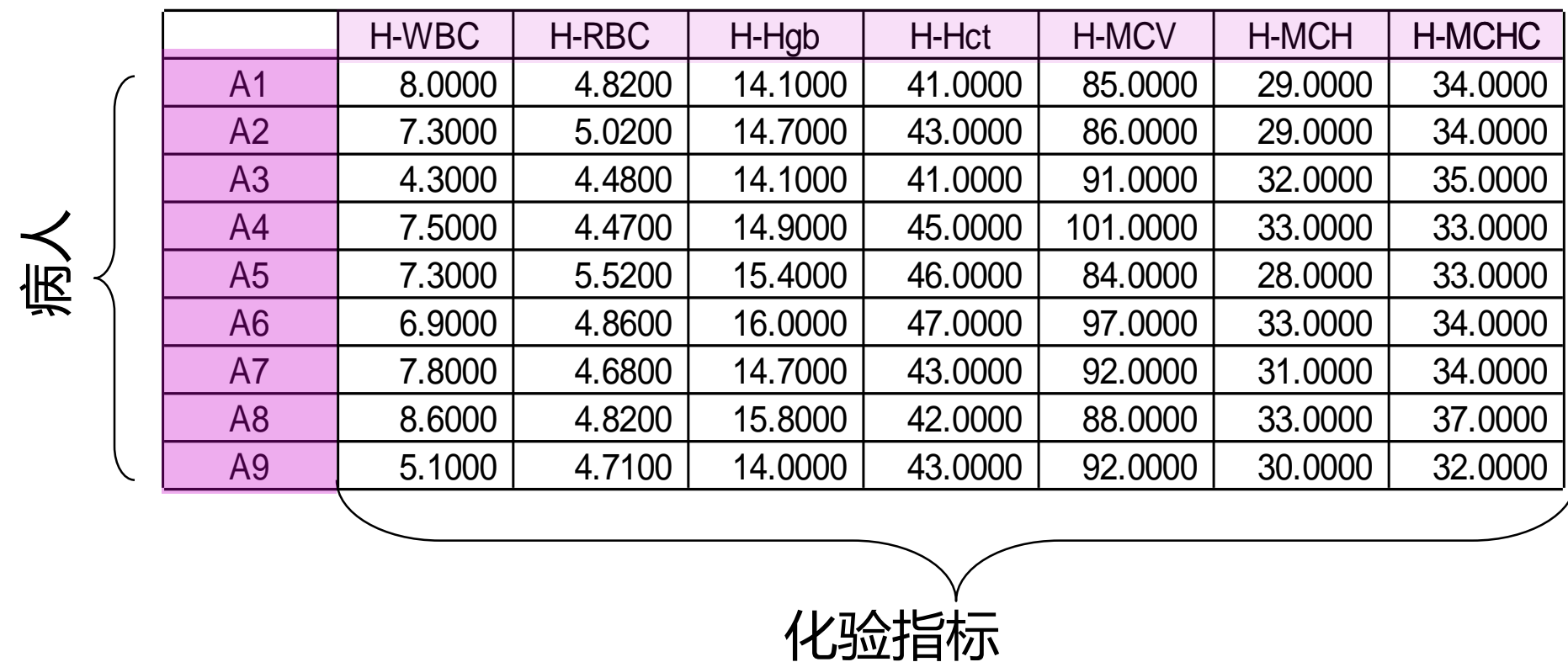
为 n 维随机变量的协方差矩阵

主成分分析-应用背景

- 数据可视化
- 数据压缩
- 噪声去除/抑制
- 数据分类
- ...

数据可视化的例子

- 给定65个病人的53种化验指标，矩阵格式表示 (65x53)

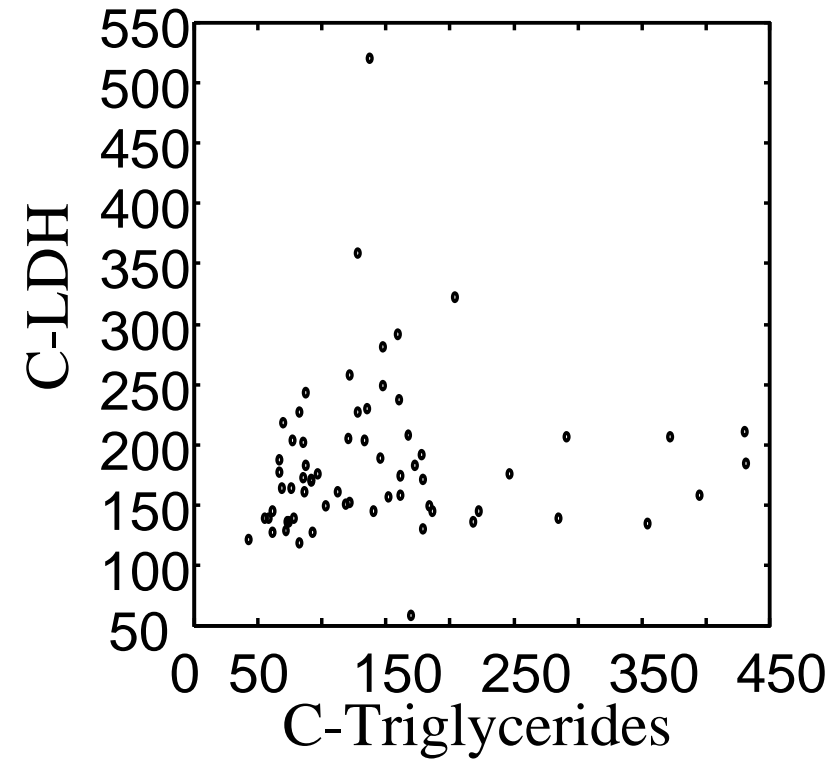


| | H-WBC | H-RBC | H-Hgb | H-Hct | H-MCV | H-MCH | H-MCHC |
|----|--------|--------|---------|---------|----------|---------|---------|
| A1 | 8.0000 | 4.8200 | 14.1000 | 41.0000 | 85.0000 | 29.0000 | 34.0000 |
| A2 | 7.3000 | 5.0200 | 14.7000 | 43.0000 | 86.0000 | 29.0000 | 34.0000 |
| A3 | 4.3000 | 4.4800 | 14.1000 | 41.0000 | 91.0000 | 32.0000 | 35.0000 |
| A4 | 7.5000 | 4.4700 | 14.9000 | 45.0000 | 101.0000 | 33.0000 | 33.0000 |
| A5 | 7.3000 | 5.5200 | 15.4000 | 46.0000 | 84.0000 | 28.0000 | 33.0000 |
| A6 | 6.9000 | 4.8600 | 16.0000 | 47.0000 | 97.0000 | 33.0000 | 34.0000 |
| A7 | 7.8000 | 4.6800 | 14.7000 | 43.0000 | 92.0000 | 31.0000 | 34.0000 |
| A8 | 8.6000 | 4.8200 | 15.8000 | 42.0000 | 88.0000 | 33.0000 | 37.0000 |
| A9 | 5.1000 | 4.7100 | 14.0000 | 43.0000 | 92.0000 | 30.0000 | 32.0000 |

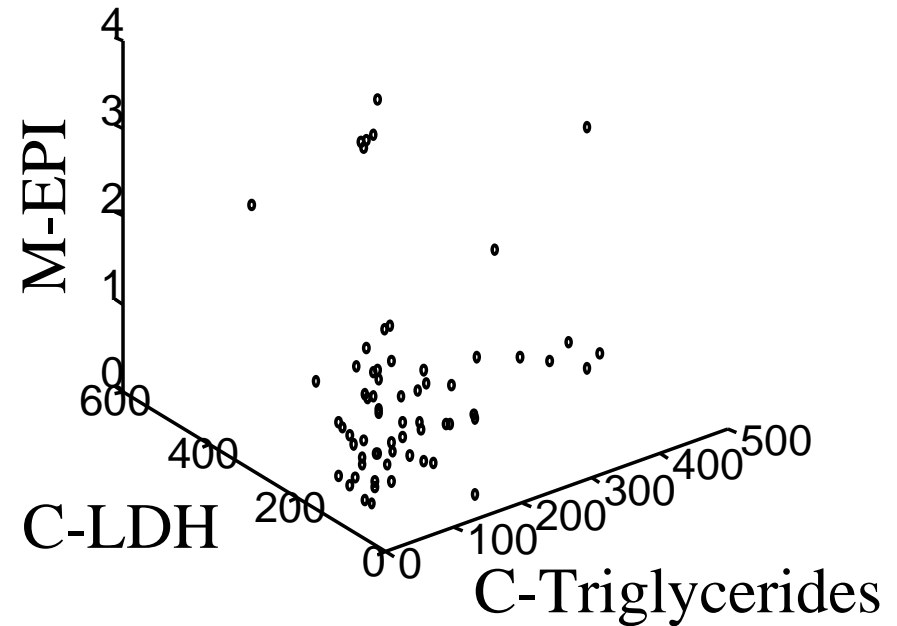
难以从中发现各项指标之间的相关性...

数据可视化的例子

2个变量观察空间



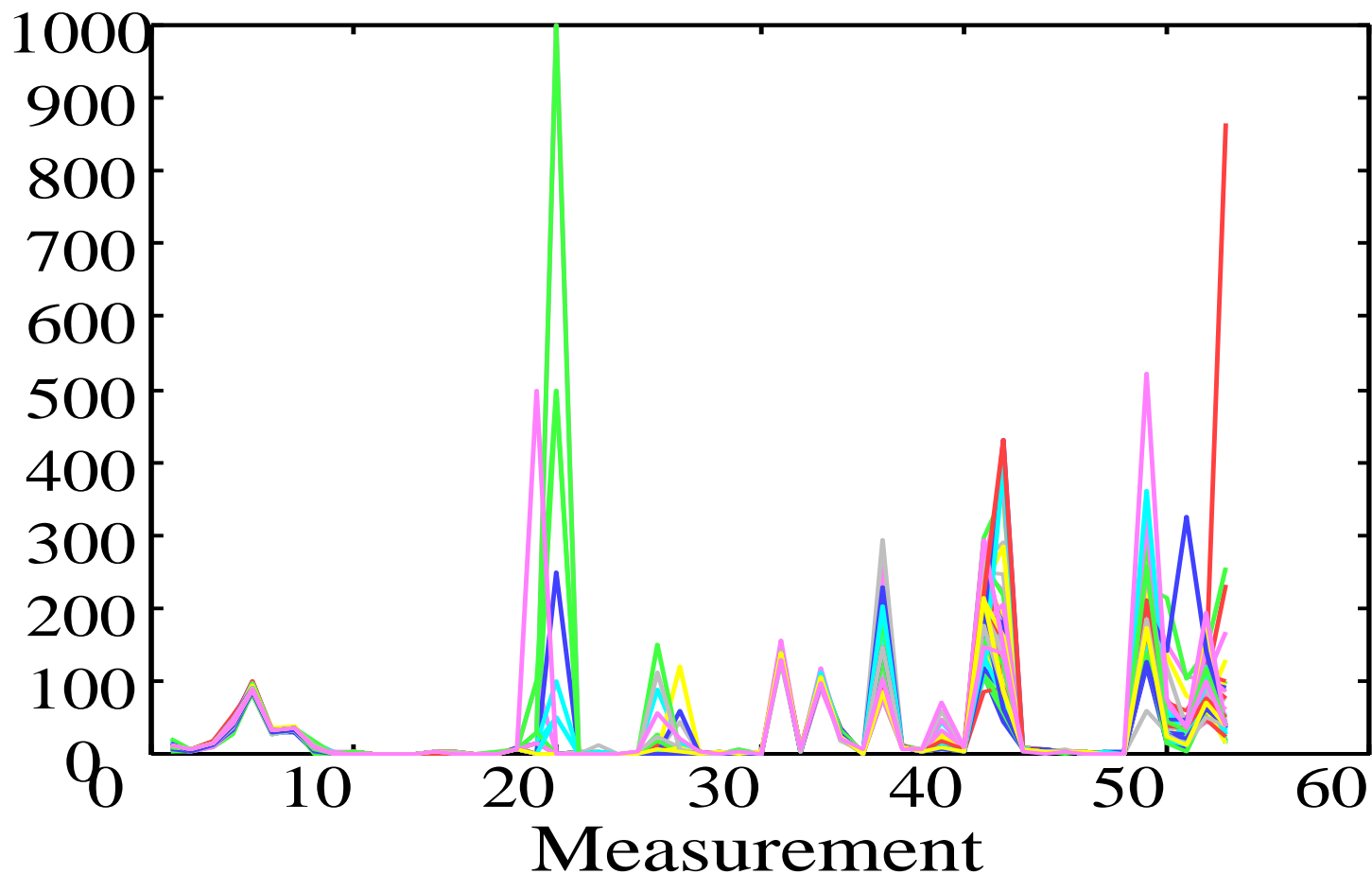
3个变量观察空间



如何对其它更多变量进行可视化分析？
... 难以观察4维及以上的更高维空间...

数据可视化的例子

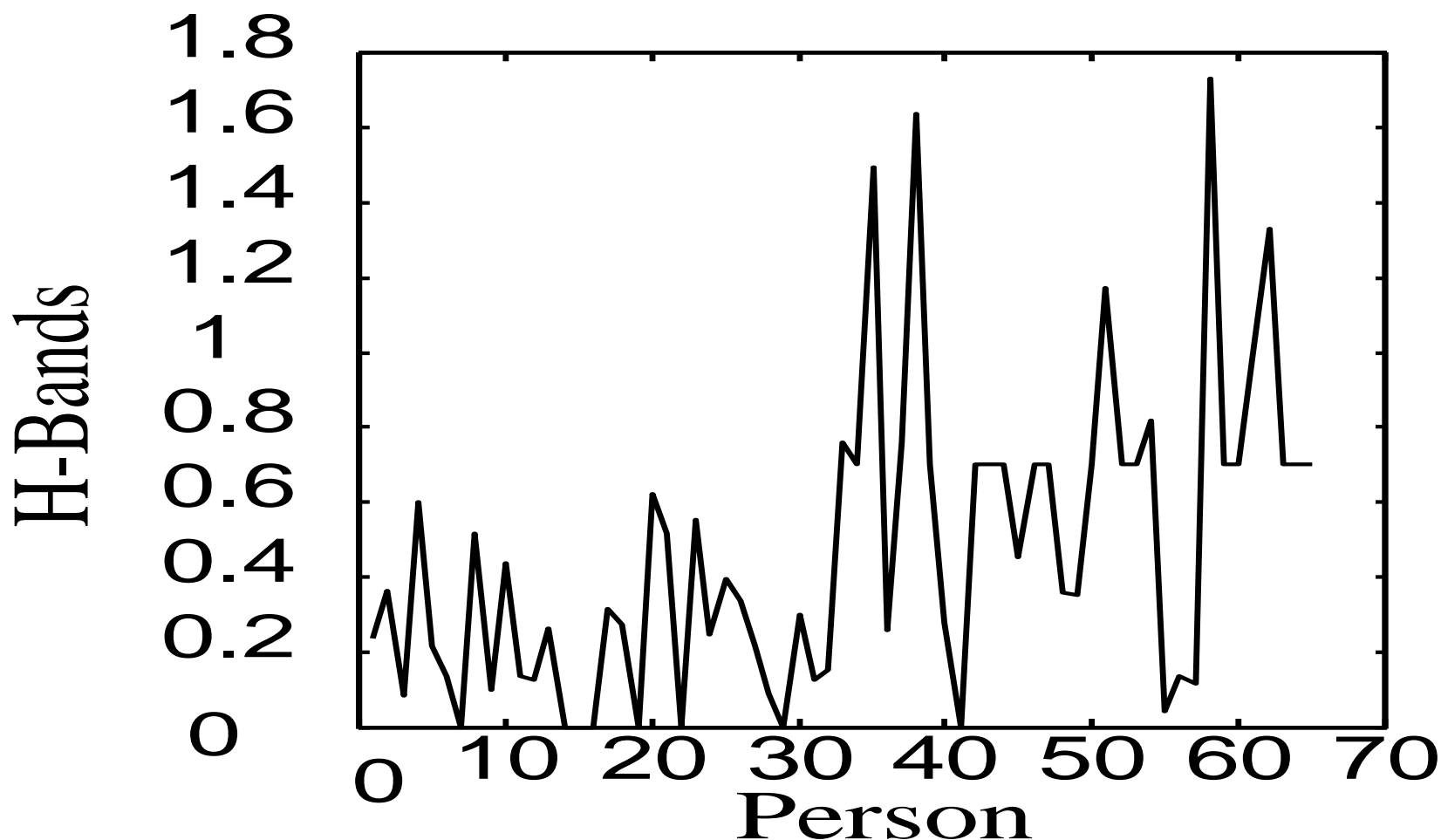
- 谱图形式表示？ (65幅图，每人1幅)



难以对不同病人进行横向比较

数据可视化的例子

- 谱图形式表示？ (53幅图，每个指标1幅)

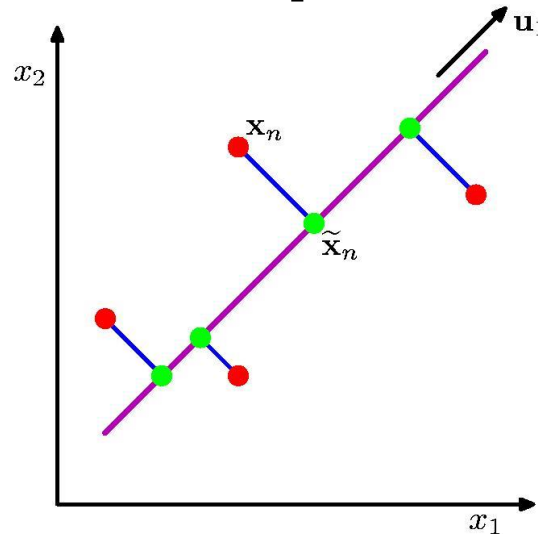


难以发现不同指标之间的关系

数据可视化的例子

- 是否存在一种比已有坐标空间更好的表示？
- 是否确有必要观察所有53项指标？
 - 如果某些指标之间有很强的相关性，会如何？
- 我们如何发现
这样的一个最小维度的子空间（包含在原始的53维空间中），其可以最大程度地保留原始数据中的最多信息？
- 解决方案：主成分分析（Principal Component Analysis，简称PCA）

Principle Component Analysis



PCA:

将数据正交投影，变换到较低维度的线性空间中，使得：

- 最大化投影后数据的方差 (图中粉色直线所示)
- 最小化原始数据点与投影后的数据点之间的均方距离 (mean squared distance) data point and
 - (图中蓝色线段累加之和)

Principle Components Analysis

基本思想:

- 给定d维观察空间的数据点, 将其投影到更低维度的空间, 同时保持尽可能多的信息
 - 例1: 找到3维数据的最佳近似平面
 - 例2: 找到 10^4 维度的原始数据的最好的12维近似表示
- 特别地, 选择的投影变换应该最小化重构原始空间数据的均方误差(*mean squared error*)

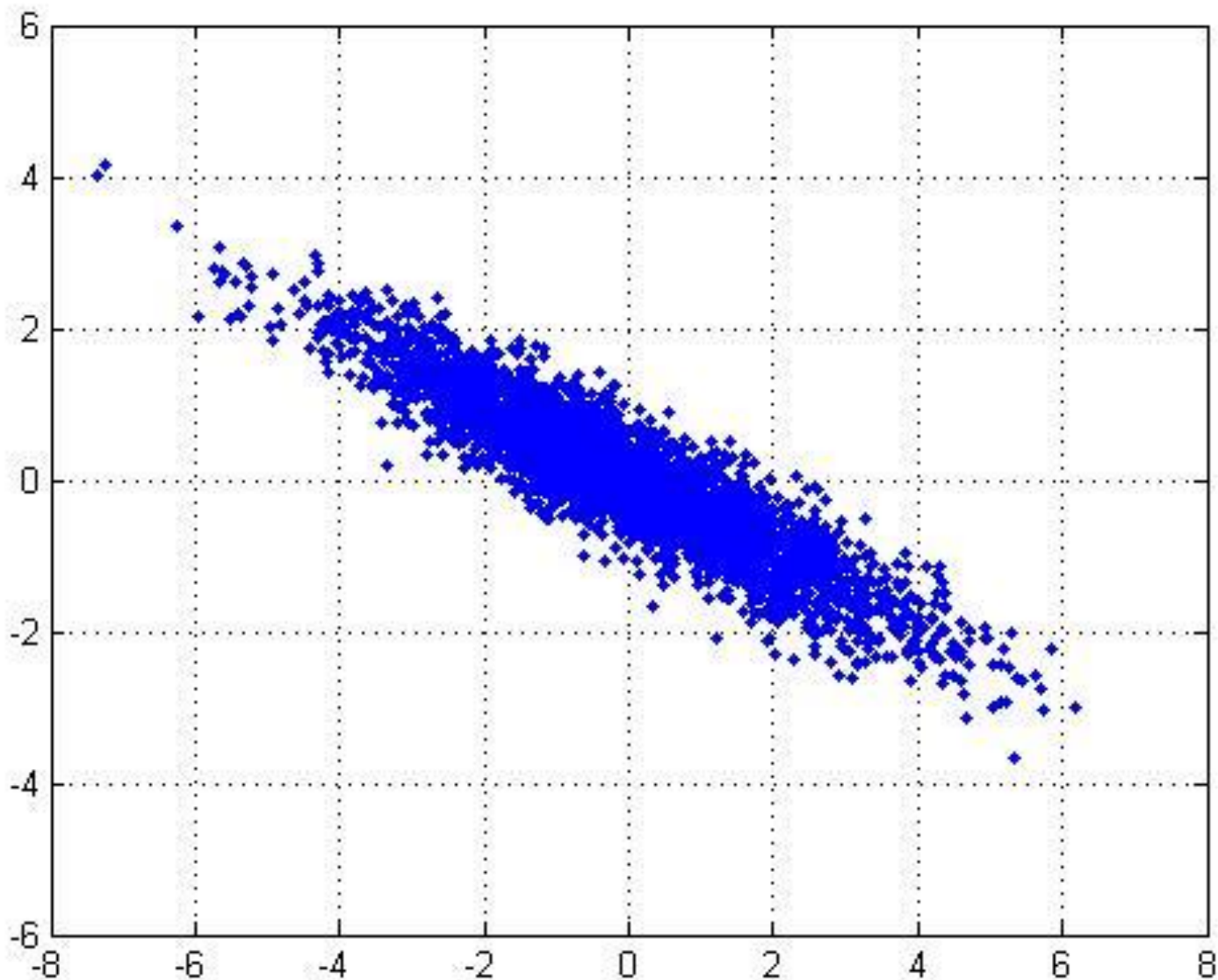
$$\min \|X - W * W^T X\|_2^2$$

从方差的角度理解

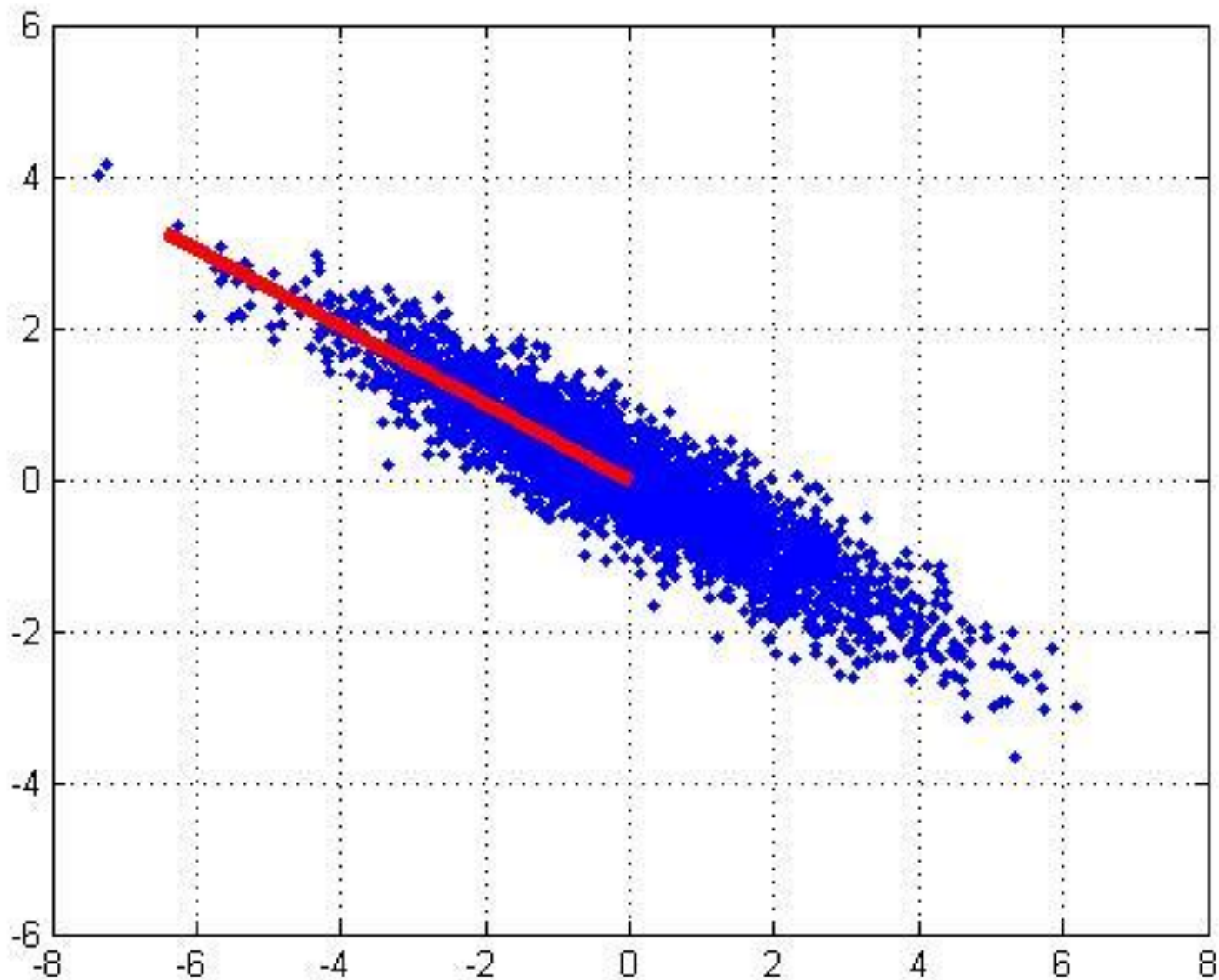
我们所要寻找的这样一些投影方向（即主成分向量，The Principal Components）

- 它们应该是以数据分布最集中区域的中心为原点向外发射的一组向量
- 所找到的第一个投影方向(Principal component #1)应该对应于数据分布的方差最大的方向
- 后续的第二、三...投影方向
 - 应与之前的投影方向相互正交
 - 对应于剩余补空间中的方差最大的方向

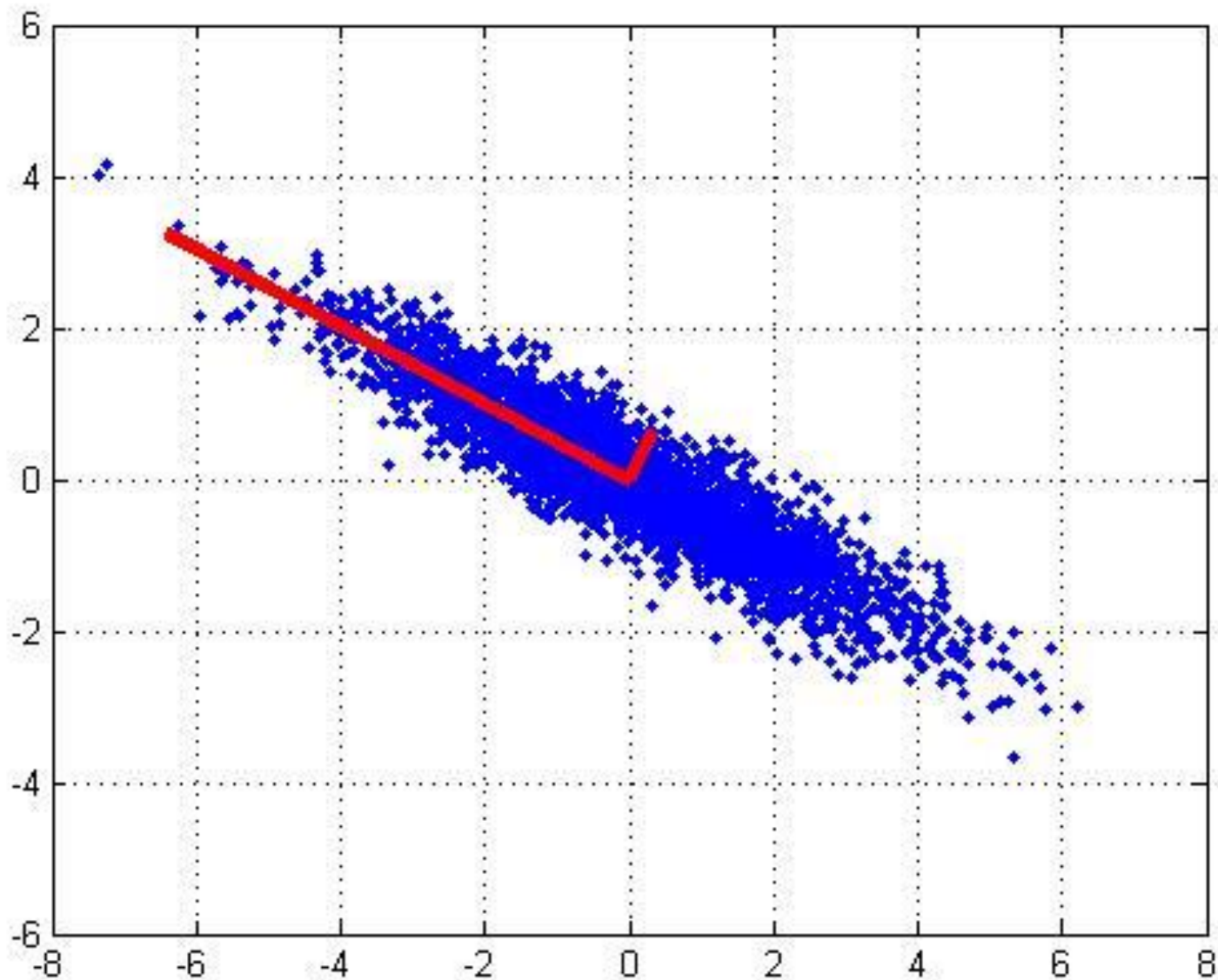
2D 高斯分布数据集



1st PCA axis—投影方向



2nd PCA axis—投影方向



求解算法—采用协方差矩阵

- 给定m个数据点 $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, 计算协方差矩阵 Σ

$$\Sigma = \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

where

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i$$

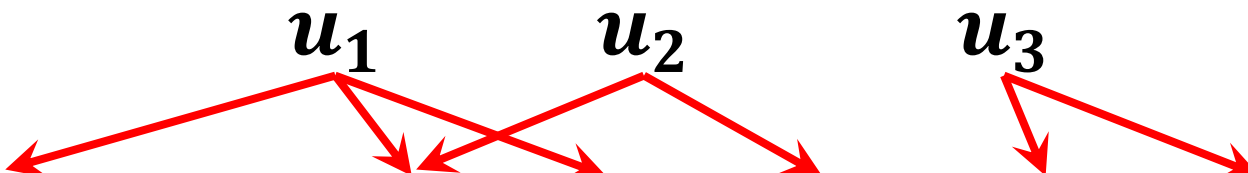
- PCA 投影方向向量 = 协方差矩阵 Σ 的特征向量
- 特征值 (必定非负) 越大 \Rightarrow 携带的信息越多

$$\min \|X - W * W^T X\|_2^2 \Leftrightarrow \text{svd}(\Sigma)$$

Principle Components Analysis

基本思想:

- 给定d维观察空间的数据点, 将其投影到更低维度的空间, 同时保持尽可能多的信息



The diagram illustrates the concept of Principal Component Analysis (PCA). It shows three principal components, u_1 , u_2 , and u_3 , represented by red arrows. These arrows originate from the top and point towards the first three columns of the data table below, indicating that the data is being projected onto these components. The table contains 9 data points (A1 to A9) across 8 variables (H-WBC, H-RBC, H-Hgb, H-Hct, H-MCV, H-MCH, H-MCHC).

| | H-WBC | H-RBC | H-Hgb | H-Hct | H-MCV | H-MCH | H-MCHC |
|----|--------|--------|---------|---------|----------|---------|---------|
| A1 | 8.0000 | 4.8200 | 14.1000 | 41.0000 | 85.0000 | 29.0000 | 34.0000 |
| A2 | 7.3000 | 5.0200 | 14.7000 | 43.0000 | 86.0000 | 29.0000 | 34.0000 |
| A3 | 4.3000 | 4.4800 | 14.1000 | 41.0000 | 91.0000 | 32.0000 | 35.0000 |
| A4 | 7.5000 | 4.4700 | 14.9000 | 45.0000 | 101.0000 | 33.0000 | 33.0000 |
| A5 | 7.3000 | 5.5200 | 15.4000 | 46.0000 | 84.0000 | 28.0000 | 33.0000 |
| A6 | 6.9000 | 4.8600 | 16.0000 | 47.0000 | 97.0000 | 33.0000 | 34.0000 |
| A7 | 7.8000 | 4.6800 | 14.7000 | 43.0000 | 92.0000 | 31.0000 | 34.0000 |
| A8 | 8.6000 | 4.8200 | 15.8000 | 42.0000 | 88.0000 | 33.0000 | 37.0000 |
| A9 | 5.1000 | 4.7100 | 14.0000 | 43.0000 | 92.0000 | 30.0000 | 32.0000 |

人脸识别的例子

- 给定一些人的照片，如何识别其身份
 - 需要考虑到对眼镜、不同光照等外部影响因素的稳定性
- ⇒ 不能仅仅使用原始输入的 256×256 大小图像中的像素信号

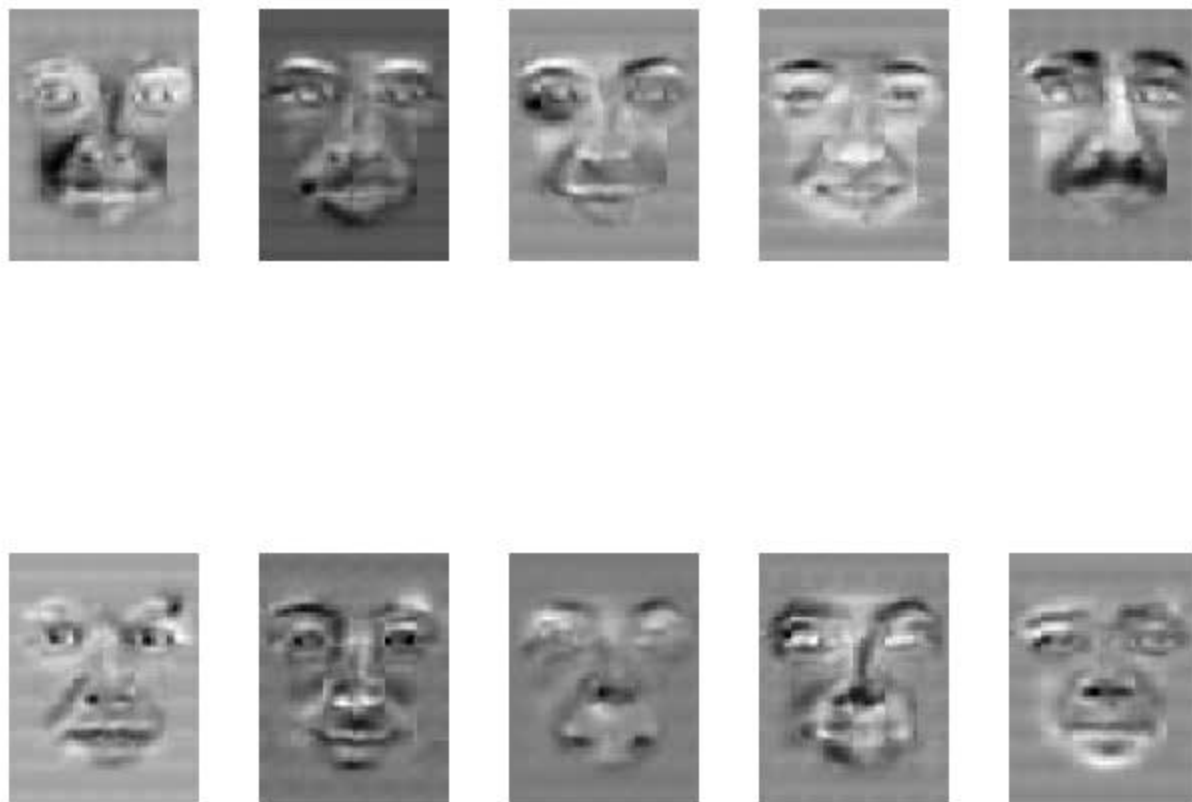


采用PCA计算得到的特征脸 (又称鬼脸, eigenface)

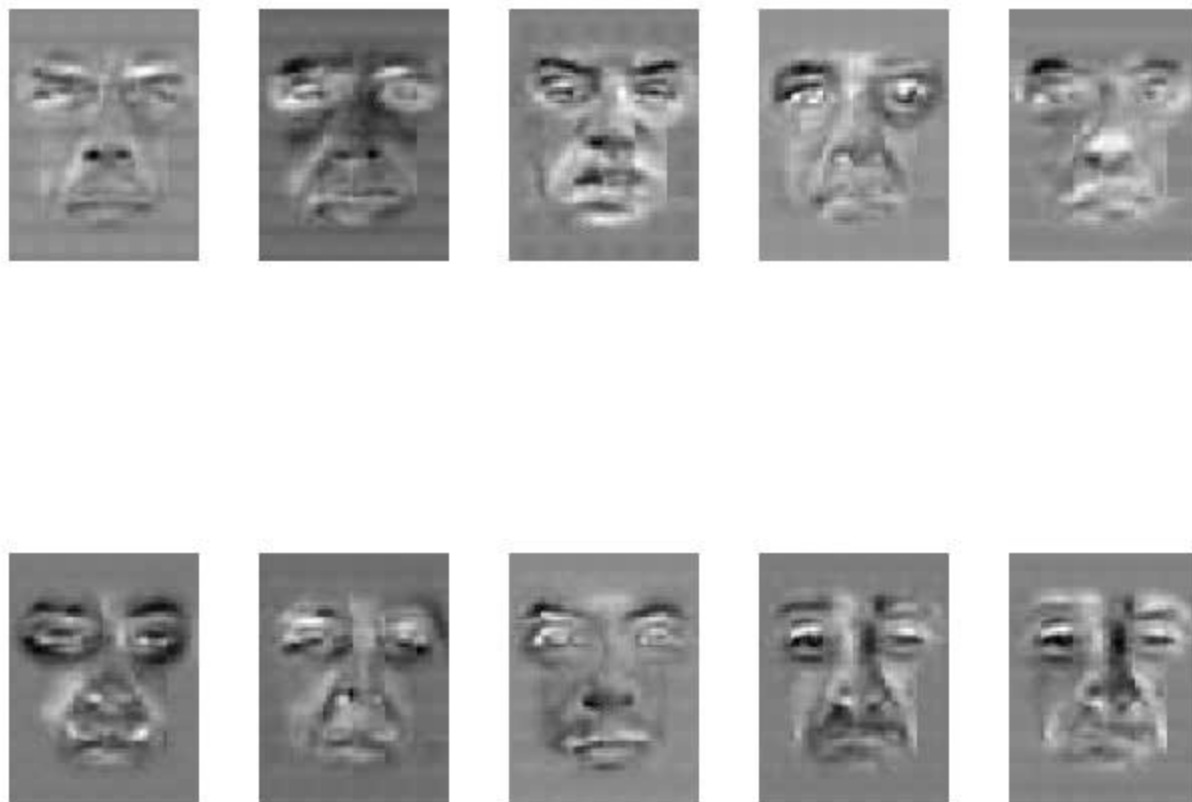


上面每一幅图像对应一个PCA投影方向 (想象一下, 图像→
矩阵→向量)

鬼脸中的“笑脸”子空间



鬼脸中的“气脸”子空间



重构得到的人脸（用以近似原始图像）

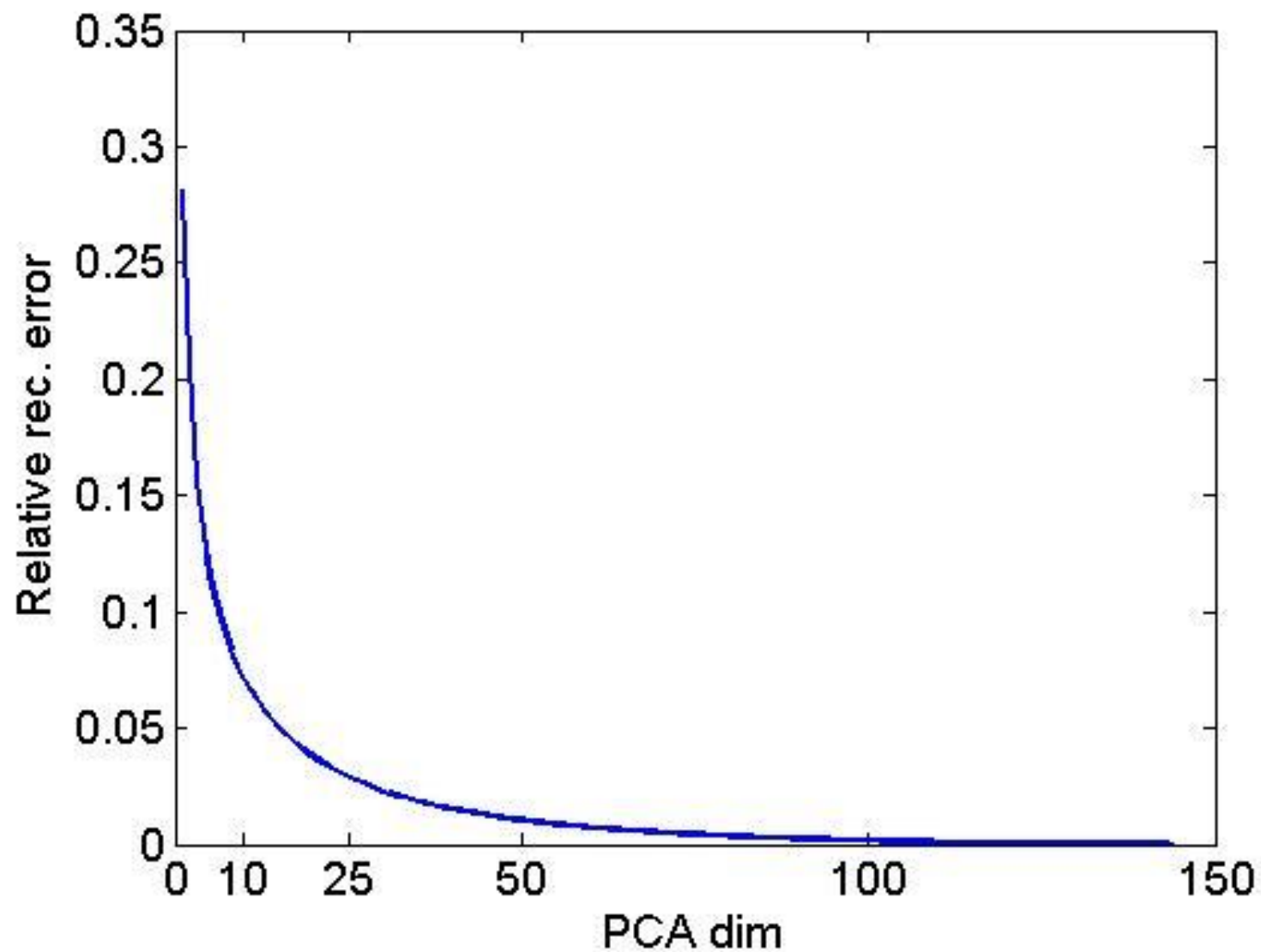


Original Image



- Divide the original 372x492 image into patches:
 - Each patch is an instance that contains 12x12 pixels on a grid
- View each as a 144-D vector

L_2 error and PCA dim



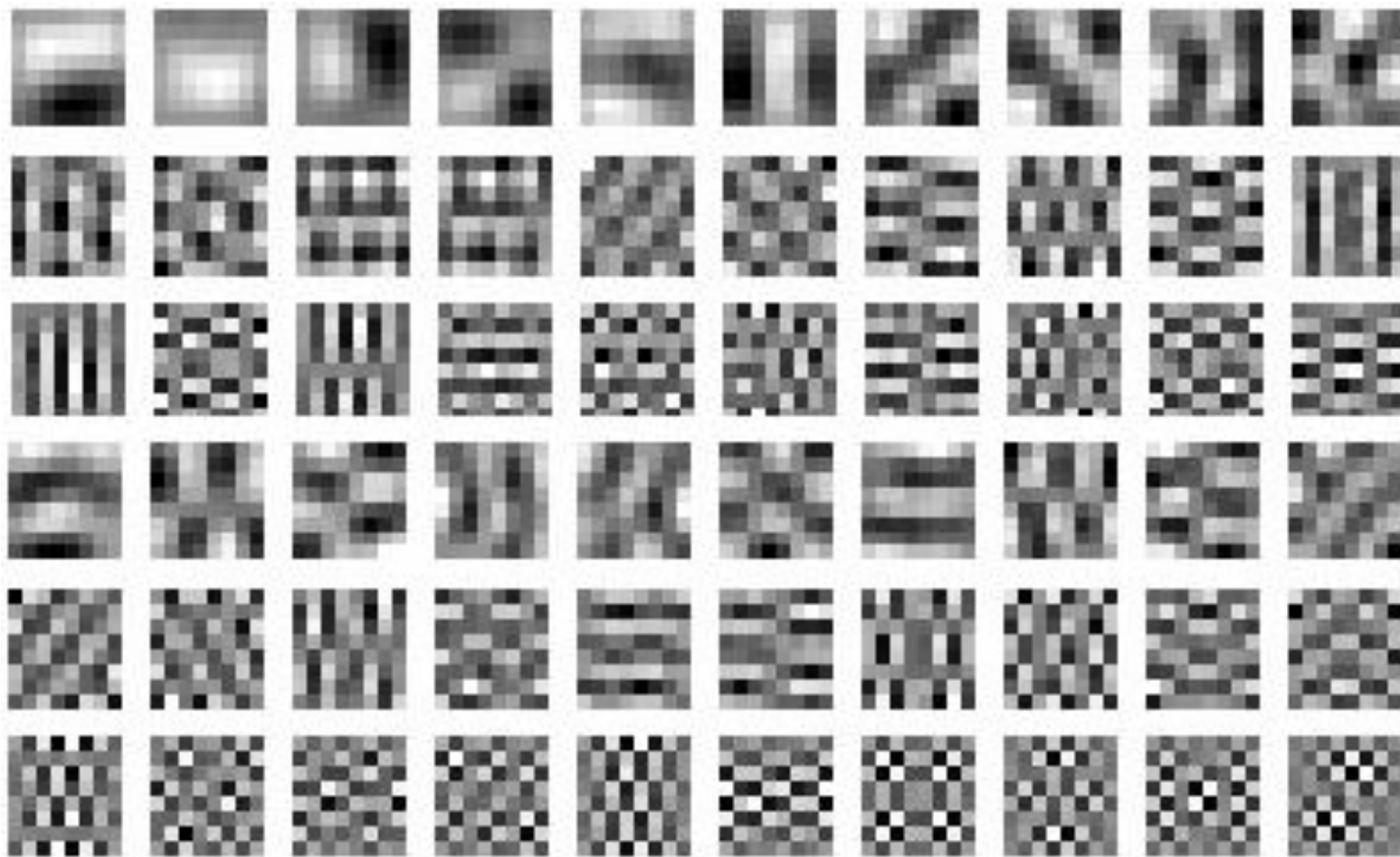
PCA compression: 144D



PCA compression: 144D -> 60D

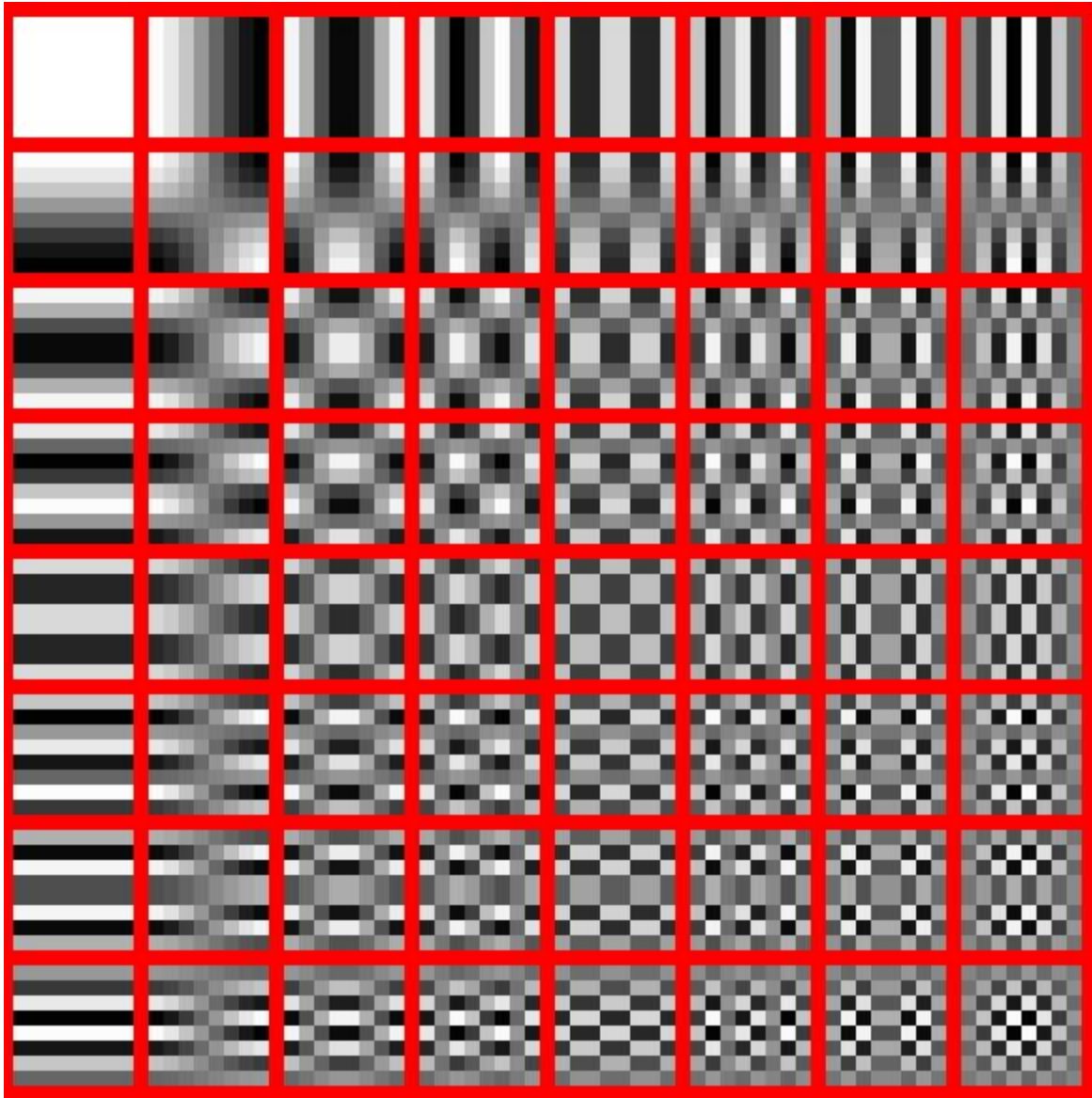


60 most important eigenvectors



Looks like the discrete cosine bases of JPG!...

2D Discrete Cosine Basis

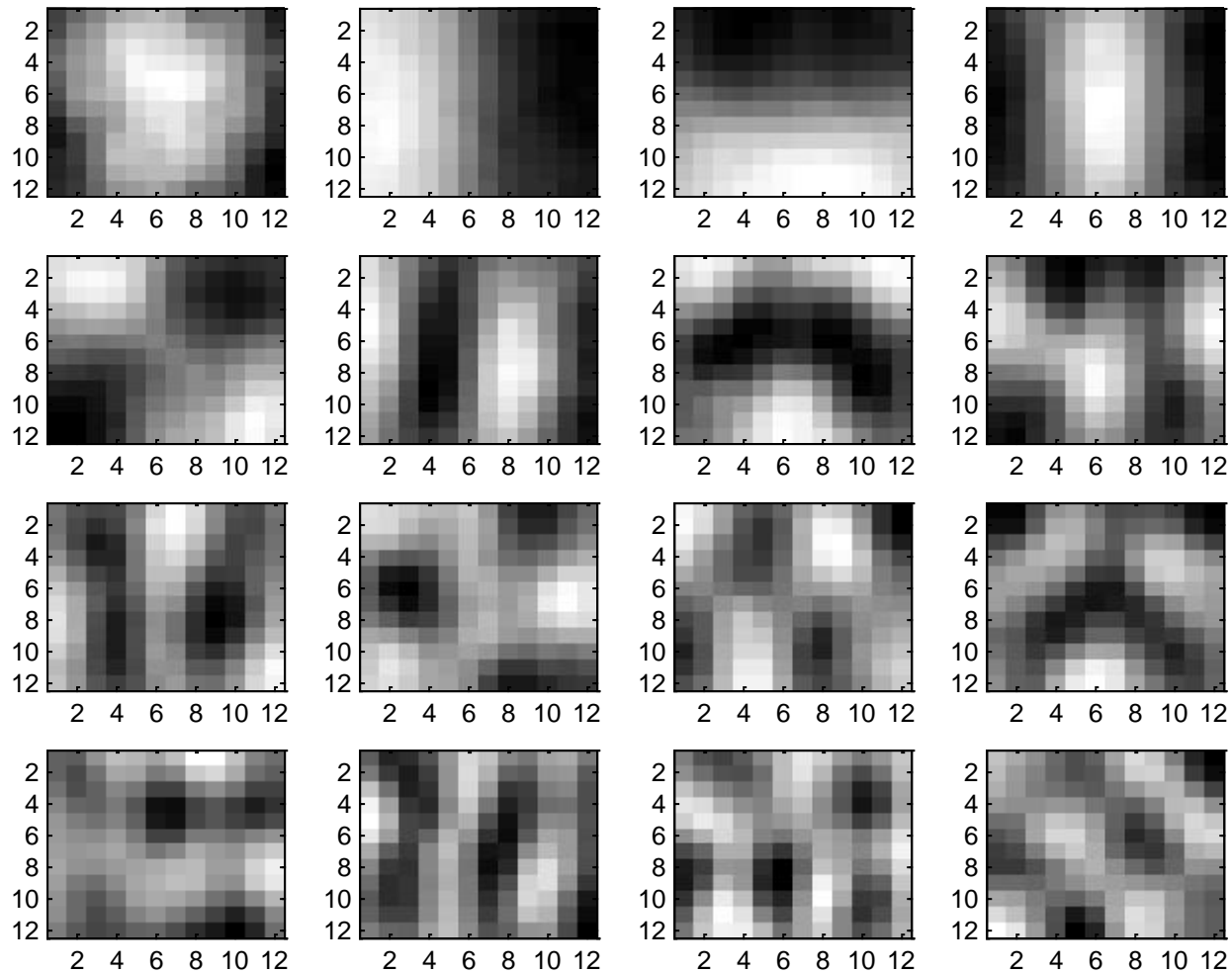


http://en.wikipedia.org/wiki/Discrete_cosine_transform

PCA compression: 144D -> 16D



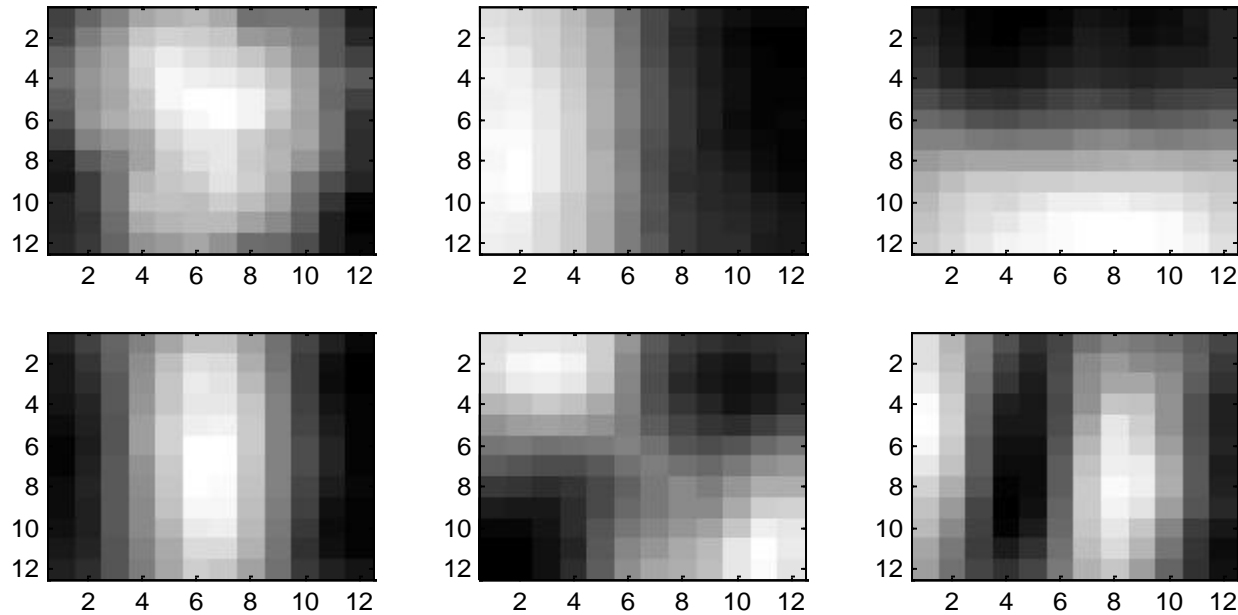
16 most important eigenvectors



PCA compression: 144D -> 6D



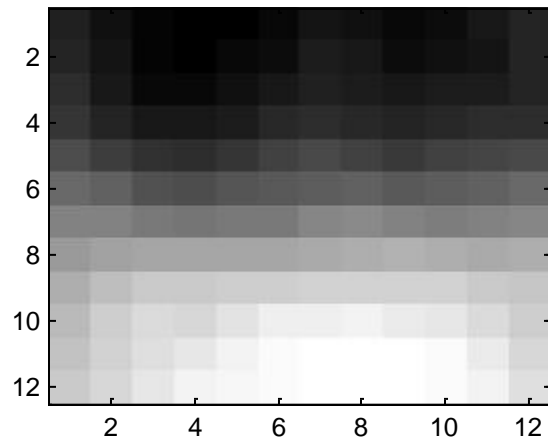
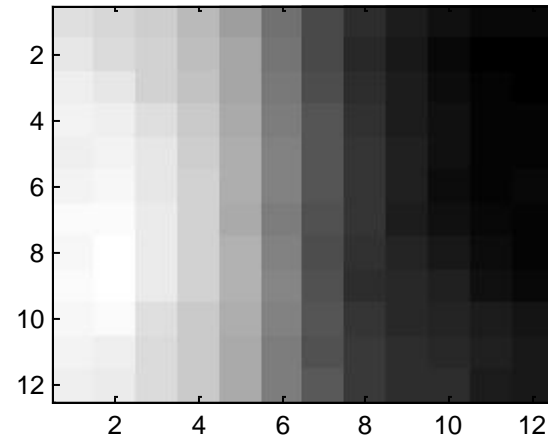
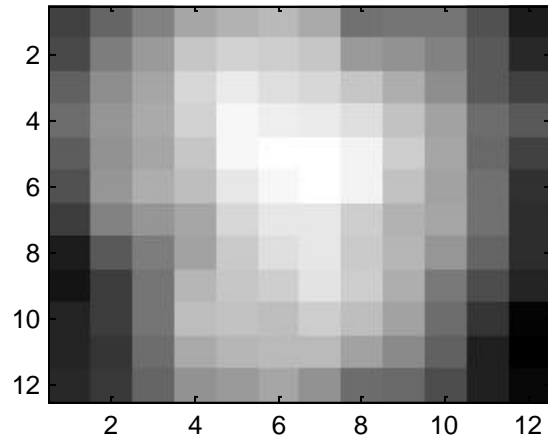
6 most important eigenvectors



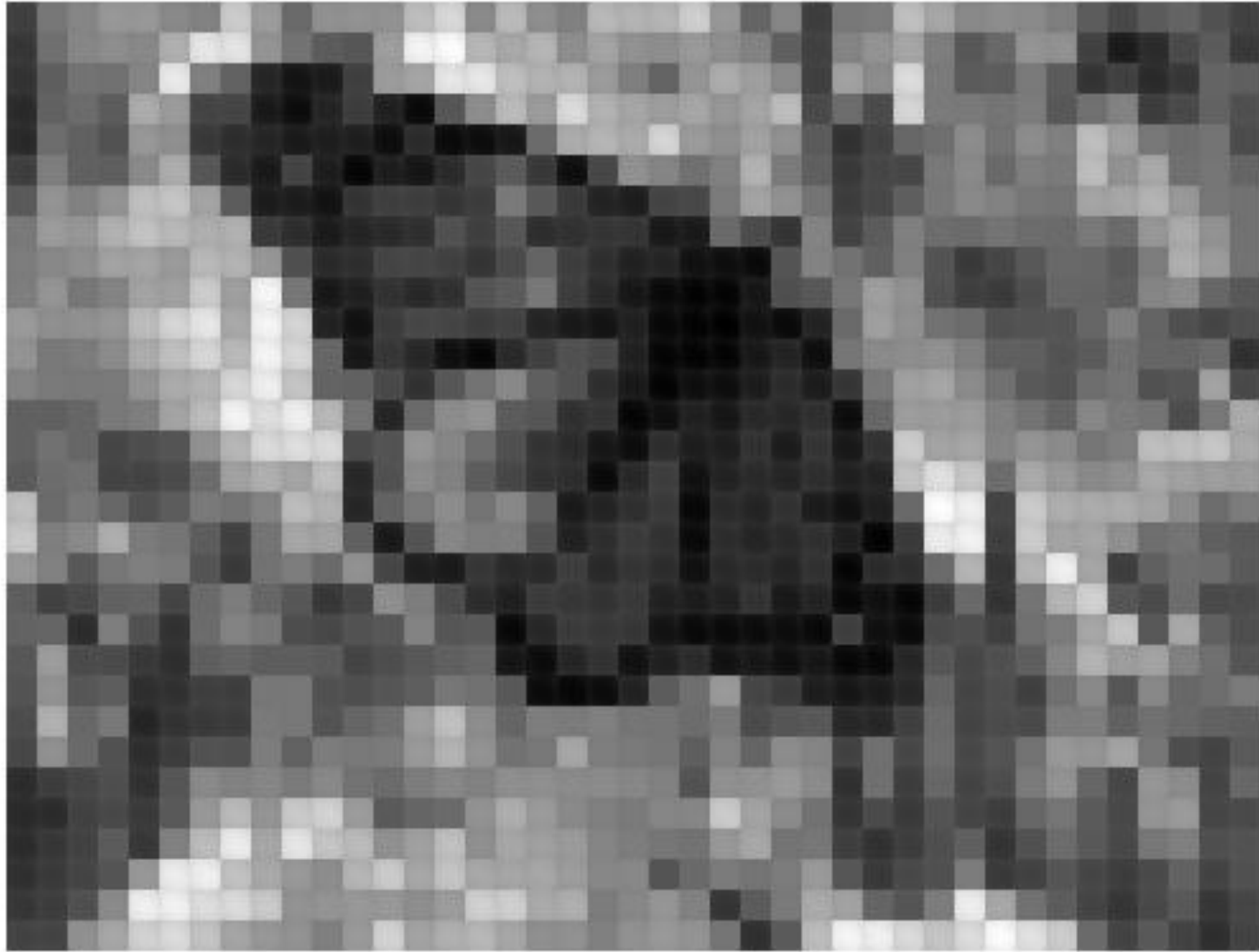
PCA compression: 144D -> 3D



3 most important eigenvectors

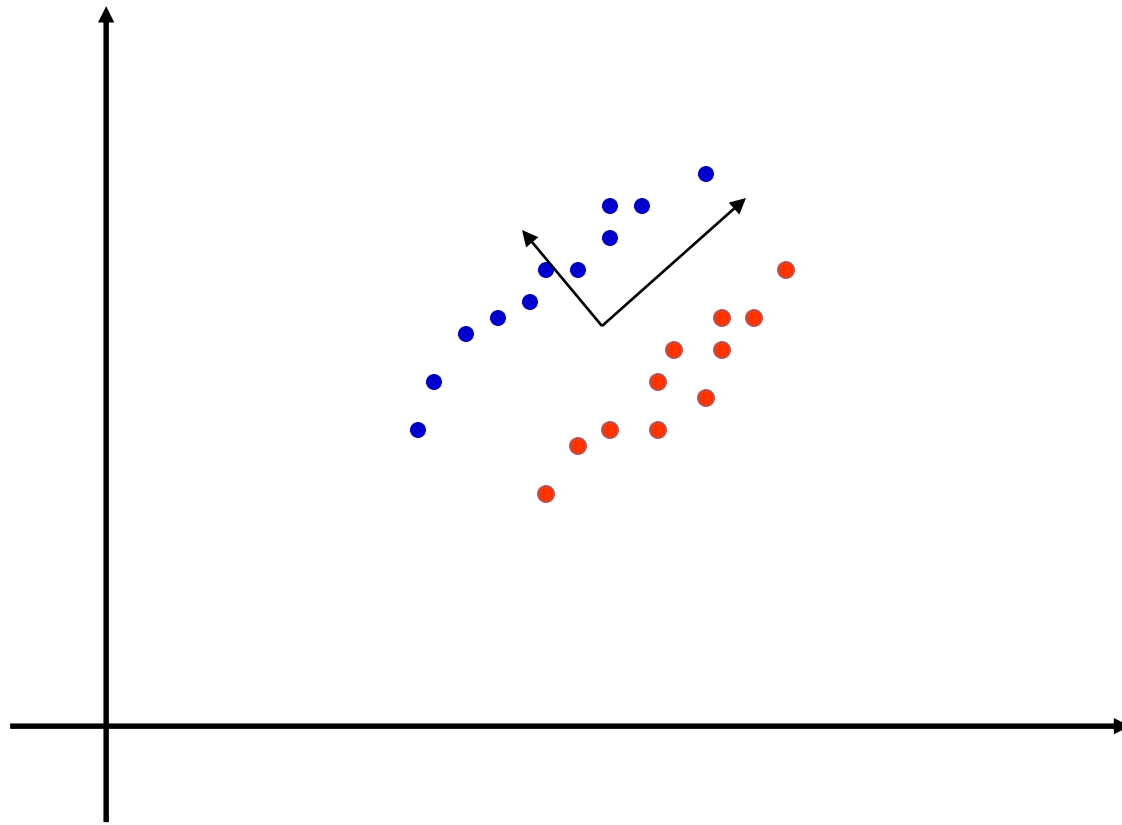


PCA compression: 144D -> 1D



PCA Shortcomings

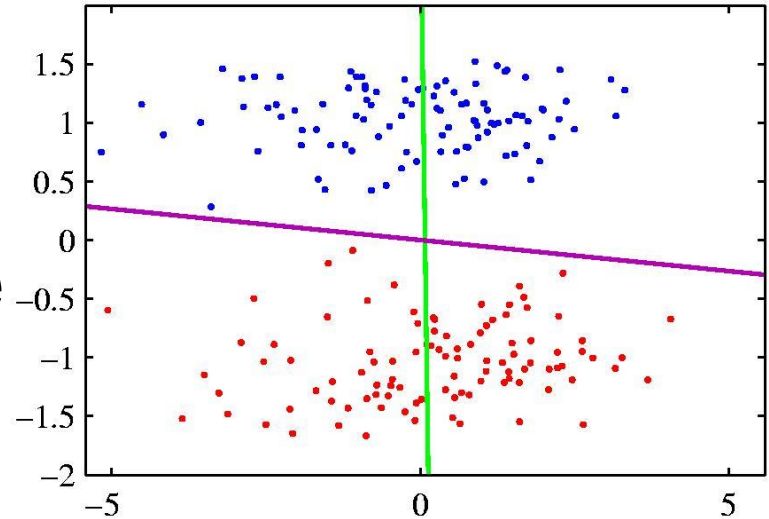
PCA, a Problematic Data Set



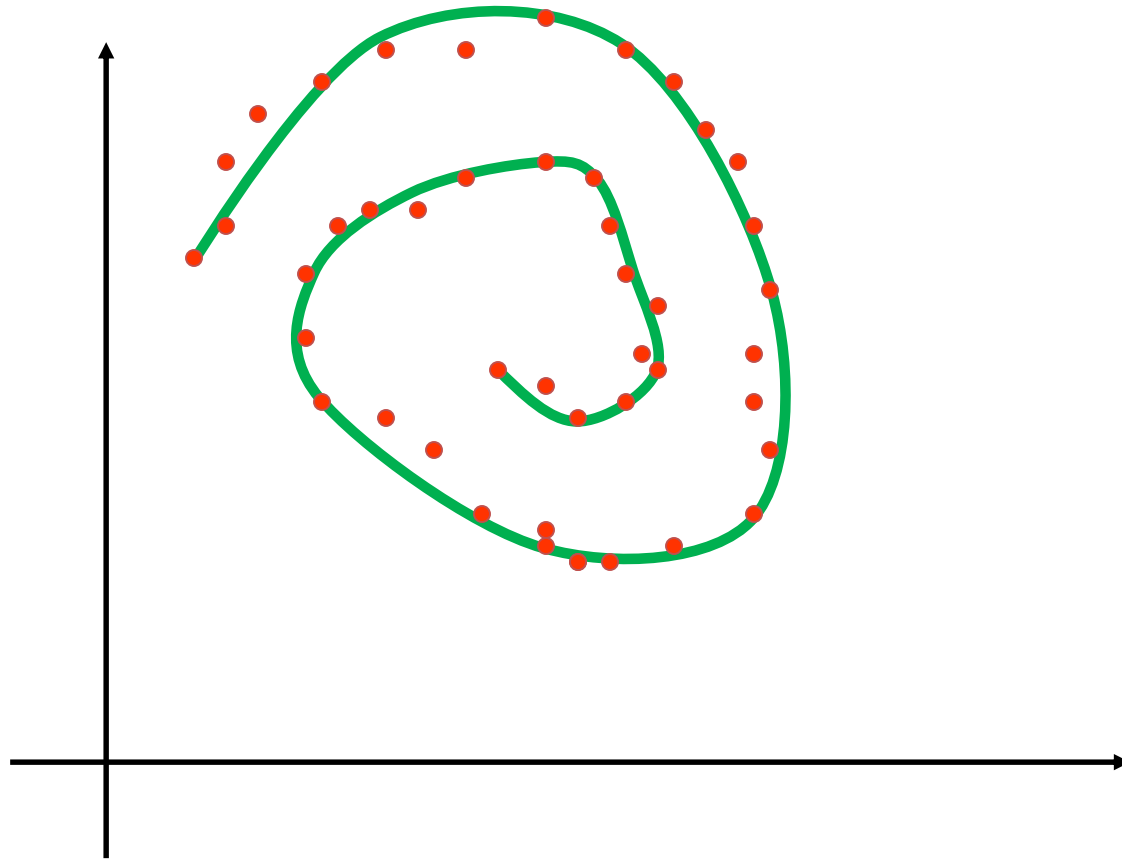
PCA doesn't know labels!

PCA vs Fisher Linear Discriminant

- PCA maximizes variance *independent of class*
⇒ magenta
- FLD attempts to separate classes
⇒ green line



PCA, a Problematic Data Set



PCA cannot capture NON-LINEAR structure!