Probabilistic principal component analysis

Роберт Сафиуллин

Москва, 2022

Стандартный Principal Component Analysis

Выборка $t_n, n \in 1..N$

$$Sw_j = \lambda_j w_j$$

Главные оси: $w_j, j \in 1..q$
Ковариационная матрица: $S = \sum_n (t_n - \bar{t})(t_n - \bar{t})^T / N$
Тогда q гланых компонент вектора из исходного пространства:

$$x_n = W^T (t_n - \bar{t})$$

Что не нравится?

Отстутствие вероятностной модели для исходных данных

Хотим:

То же самое через плотность распределения

Еще мотивация

- Вероятностная модель предлагает более широкий функционал для PCA (можно комбинировать модели)
- Можно эффективно посчитать ковариацию исходя из максимума правдоподобия

Вероятностный PCA

Хотим получить генеративную модель:

$$t = Wx + \mu + \epsilon$$

Возьмем $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$ и запишем условную вероятность:

$$t|x \sim \mathbf{N}(Wx + \mu, \sigma^2\mathbf{I})$$

Распределение выборки:

$$t \sim \mathbf{N}(\mu, \mathbf{C})$$
$$\text{где } \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Соответствующий логарифм правдоподобия:

$$\mathbf{L} = -\frac{N}{2}[dln(2\pi) + ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\mathbf{S})]$$
$$\text{где } \mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)^T$$

Хорошие свойства

При следующем условном распределении:

$$x|t \sim \mathbf{N}(\mathbf{M}^{-1}\mathbf{W}^{\mathbf{T}}(t-\mu), \sigma^2\mathbf{M}^{-1})$$
$$\mathbf{M} = \mathbf{W}^{\mathbf{T}}\mathbf{W} + \sigma^2\mathbf{I}$$

Максимальное правдоподобие

$$\mathbf{W_{opt}} = \mathbf{U}_q(\Lambda_q - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R}$$

$\mathbf{U_q}$ - матрица собственных векторов $\mathbf{S}$

Λ_q - диагональная матрица собственных значений

$\mathbf{R}$ - матрица поворота

Соответствующая дисперсия

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j$$

Снижение размерности смесью распределений

$$\langle x_n|t_n\rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathbf{T}}_{\mathbf{opt}}(t_n - \mu)$$

Пропущенные значения

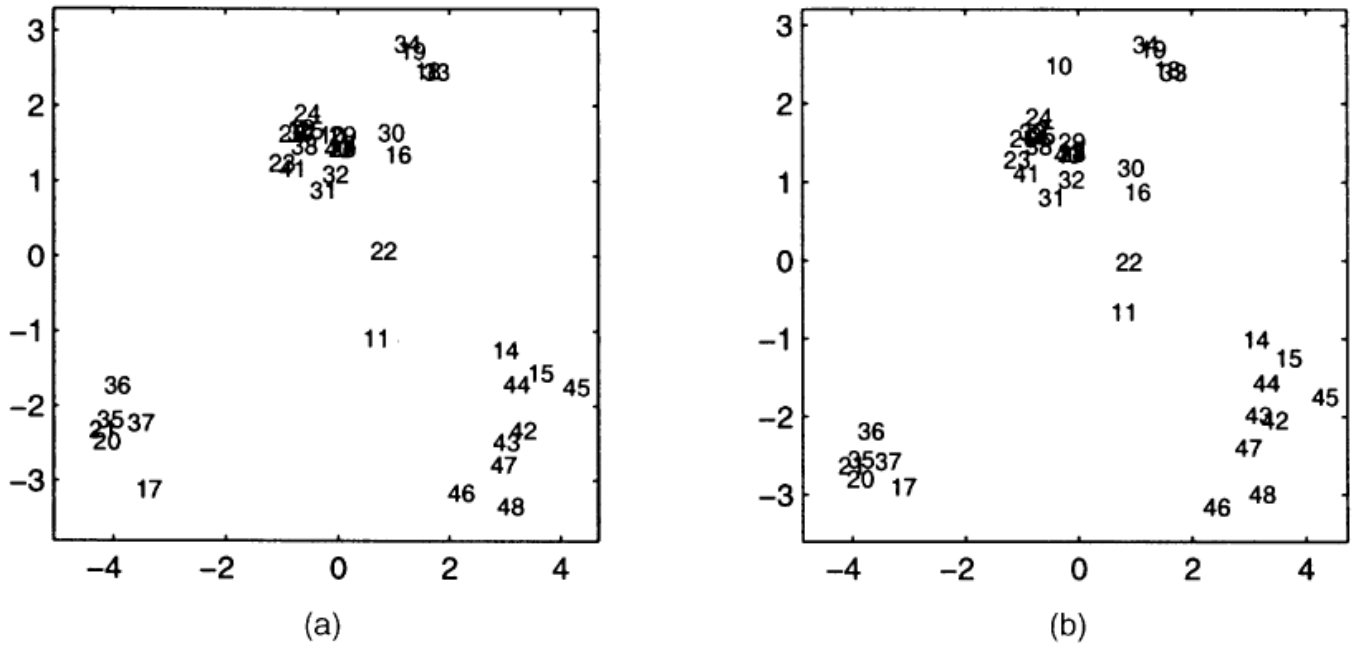

Fig. 1. Projections of the *Tobamovirus* data by using (a) PCA on the full data set and (b) PPCA with 136 missing values

Использованная литература

PPCA