Probabilistic modelling - Multivariate Gaussian and PCA

Peter Tino
School of Computer Science
University of Birmingham
UK
Consider a **random variable** $X$ taking on values in $\mathbb{R}$. $x \in \mathbb{R}$ are **realisations of** $X$.

$N$ repeated **i.i.d. draws from** $X$:
Imagine $N$ independent and identically distributed random variables $X^1, X^2, ..., X^N$.
$x^i$ is a realisation of $X^i$, $i = 1, 2, ..., N$.

**Continuous RV:**
Realisations are from a continuous subset $A$ of $\mathbb{R}$.
Probability density $p(x)$: $\int_A p(x) \, dx = 1$.

**Discrete RV:**
Realisations are from a discrete subset $A$ of $\mathbb{R}$.
Probability distribution $P(x)$: $\sum_{x \in A} P(x) = 1$. 
Characterising random variables

**Mean of RV** \( X \): Center of gravity around which realisations of \( X \) happen. First central moment.

\[
E[X] = \sum_{x \in A} x \cdot P(X = x) \quad \text{or} \quad E[X] = \int_A x \cdot p(x) \, dx
\]

**Variance of RV** \( X \): (Squared) fluctuations of realisations \( x \) around the center of gravity \( E[X] \). Second central moment.

\[
\text{Var}[X] = E[(X - E[X])^2] = \sum_{x \in A} (x - E[X])^2 \cdot P(X = x),
\]

or

\[
\text{Var}[X] = E[(X - E[X])^2] = \int_A (x - E[X])^2 \cdot p(x) \, dx
\]
Estimating central moments of $X$

$N$ i.i.d. realisations of $X$:
$x^1, x^2, ..., x^N \in \mathbb{R}$.

$$E[X] \approx \hat{E}[X] = \frac{1}{N} \sum_{i=1}^{N} x^i$$

$$\text{Var}[X] \approx \hat{\text{Var}}[X]_{ML} = \frac{1}{N} \sum_{i=1}^{N} \left(x^i - \hat{E}[X]\right)^2$$

Unbiased estimation of variance:

$$\frac{1}{N-1} \sum_{i=1}^{N} \left(x^i - \hat{E}[X]\right)^2$$
Several random variables

Consider 2 RVs $X$ and $Y$

Still can compute central moments of individual RVs, i.e. $E[X]$, $E[Y]$ and $Var[X]$, $Var[Y]$.

In addition we can ask whether $X$ and $Y$ are ‘statistically tight together’ in some way

**Covariance of RVs $X$ and $Y$**

Co-fluctuations around the means:

Introduce a new random variable $Z = (X - E[X]) \cdot (Y - E[Y])$

$$\text{Cov}[X, Y] = E[Z] = E[(X - E[X]) \cdot (Y - E[Y])]$$
Estimating covariance of $X$ and $Y$

$N$ i.i.d. realisations of $(X, Y)$:
$(x^1, y^1), (x^2, y^2), ..., (x^N, y^N) \in \mathbb{R}^2$.

$$\text{Cov}[X, Y] \approx \hat{\text{Cov}}[X, Y] = \frac{1}{N} \sum_{i=1}^{N} (x^i - \hat{E}[X]) \cdot (y^i - \hat{E}[Y])$$

For centred RVs (means are 0), we have

$$\text{Cov}[X, Y] = \frac{1}{N} \sum_{i=1}^{N} x^i y^i$$
Covariance matrix of $X$, $Y$

Note that formally
\[ \text{Var}[X] = \text{Cov}[X, X] \]
and
\[ \hat{\text{Var}}[X] = \hat{\text{Cov}}[X, X]. \]

Covariance matrix summarises the variance/covariance structure in $(X, Y)$:

\[
\begin{bmatrix}
\text{Var}[X] & \text{Cov}[X, Y] \\
\text{Cov}[Y, X] & \text{Var}[Y]
\end{bmatrix}
\]
Covariance matrix of a vector RV

Consider a vector random variable

\[ \mathbf{X} = (X_1, X_2, \ldots, X_d)^T \]

Covariance matrix of \( \mathbf{X} \) is

\[
\text{Cov}[\mathbf{X}] = \begin{bmatrix}
\text{Var}[X_1] & \text{Cov}[X_1, X_2] & \text{Cov}[X_1, X_3] & \cdots & \text{Cov}[X_1, X_d] \\
\text{Cov}[X_2, X_1] & \text{Var}[X_2] & \text{Cov}[X_2, X_3] & \cdots & \text{Cov}[X_2, X_d] \\
\text{Cov}[X_3, X_1] & \text{Cov}[X_3, X_2] & \text{Var}[X_3] & \cdots & \text{Cov}[X_3, X_d] \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{Cov}[X_d, X_1] & \text{Cov}[X_d, X_2] & \text{Cov}[X_d, X_3] & \cdots & \text{Var}[X_d]
\end{bmatrix}
\]

Note that \( \text{Cov}[\mathbf{X}] \) is square and symmetric.
Estimating $\text{Cov}[\mathbf{X}]$

$N$ i.i.d. realisations of the vector RV $\mathbf{X} = (X_1, X_2, ..., X_d)^T$:

$\mathbf{x}^1 = (x_1^1, x_2^1, ..., x_d^1)^T, \mathbf{x}^2 = (x_1^2, x_2^2, ..., x_d^2)^T, ..., \mathbf{x}^N = (x_1^N, x_2^N, ..., x_d^N)^T.$

Collect the realisations $\mathbf{x}^i$ of $\mathbf{X}$ as columns of the design matrix

$\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N].$

Assume the RV $\mathbf{X}$ is centred ($E[X_i] = 0, i = 1, 2, ..., d$).

Then

$$\text{Cov}[\mathbf{X}] \approx \widehat{\text{Cov}}[\mathbf{X}] = \frac{1}{N} \mathbf{X} \mathbf{X}^T.$$
Aligning Co-ordinate Axes with the Data

Rotate our axes so that the data is “more aligned” with the new axes than with the original ones.

In the new axes system the co-variances vanish. It is then easy to identify directions where most of data variation happens and the ‘minor’ ones where the fluctuations are small.

\[ \tilde{X} = (\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d)^T \] – random variable \( X \) expressed in the new axes.

\[
\text{Cov}[\tilde{X}] =
\begin{bmatrix}
\text{Var}[\tilde{X}_1] & 0 & 0 & \ldots & 0 \\
0 & \text{Var}[\tilde{X}_2] & 0 & \ldots & 0 \\
0 & 0 & \text{Var}[\tilde{X}_3] & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \text{Var}[\tilde{X}_d]
\end{bmatrix}
\]
(canonical) dot product of vectors: \( \mathbf{a} = [a_1, a_2, \ldots, a_d]^T, \)
\( \mathbf{b} = [b_1, b_2, \ldots, b_d]^T \) vectors in \( \mathbb{R}^d \). Let \( \alpha \) be the angle between \( \mathbf{a} \) and \( \mathbf{b} \).

The *dot product* between \( \mathbf{a} \) and \( \mathbf{b} \) is

\[
\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = \sum_{i=1}^{d} a_i \cdot b_i = \|\mathbf{a}\| \cdot \|\mathbf{b}\| \cos \alpha,
\]

where \( \|\mathbf{a}\| \) is the \((L_2)\) norm (or length) of \( \mathbf{a} \),

\[
\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \ldots + a_d^2}.
\]

Note that \( \|\mathbf{a}\|^2 = \mathbf{a}^T \mathbf{a} \).
Rotating Co-ordinate Axes

Collect normal mutually orthogonal directional vectors of the new axes in matrix \( \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d] \), \( \|\mathbf{v}_i\| = 1 \) and for \( i \neq j \), \( \mathbf{v}_i^T \mathbf{v}_j = 0 \).

Expressing point \( \mathbf{x}^i \) in the new co-ordinates:

\[
\tilde{\mathbf{x}}^i = \mathbf{V}^T \cdot \mathbf{x}^i
\]

Moving back into the original co-ordinate system:

\[
\mathbf{x}^i = \mathbf{V} \cdot \tilde{\mathbf{x}}^i
\]

Note: For orthogonal matrices \( \mathbf{V}^{-1} = \mathbf{V}^T \).
Rotated Covariance Matrix

Design matrix in the new co-ordinate system:

$$\tilde{\mathbf{x}} = \mathbf{V}^T \mathbf{x}$$

We can transform back as

$$\mathbf{x} = \mathbf{V} \tilde{\mathbf{x}}.$$
Rotated Covariance Matrix

\[
\mathbf{C} = \frac{1}{N} \mathbf{x} \mathbf{x}^T \\
= \frac{1}{N} \mathbf{V} \mathbf{\tilde{x}} (\mathbf{V} \mathbf{\tilde{x}})^T \\
= \frac{1}{N} \mathbf{V} \mathbf{\tilde{x}} \mathbf{\tilde{x}}^T \mathbf{V}^T \\
= \mathbf{V} \left[ \frac{1}{N} \mathbf{\tilde{x}} \mathbf{\tilde{x}}^T \right] \mathbf{V}^T \\
= \mathbf{V} \tilde{\mathbf{C}} \mathbf{V}^T
\]

This holds in general. If we want the new axis \( \mathbf{V} \) such that \( \tilde{\mathbf{C}} \) is diagonal, \( \mathbf{V} \) should be eigenvectors of \( \mathbf{C} \).
A vector $\mathbf{v} \in \mathbb{R}^d$ is called an **eigenvector** of $\mathbf{A} \in \mathbb{R}^{d \times d}$ if there exist a non-zero constant $\lambda$, such that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$ 

The scalar $\lambda \in \mathbb{R}$ is called **eigenvalue** of $\mathbf{A}$ corresponding to the eigenvector $\mathbf{v}$.

A $d \times d$ **symmetric matrix** will have exactly $d$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$, with **real** $\lambda_1, \lambda_2, \ldots, \lambda_d$ as the corresponding eigenvalues (strictly speaking this is true only if all eigenvalues are non-zero and different).
Covariance

Eigen-decomposition of square symmetric matrices

For any square matrix \( A \), we can decompose it into a special canonical form, known as matrix diagonalisation:

\[
A = VDV^T,
\]

were \( V = [v_1, v_2, \ldots, v_d] \) is a matrix formed by the eigenvectors of \( A \) (as columns), and \( D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) \) is a diagonal matrix with non-zero diagonal entries equal to the eigenvalues of \( A \).

Hence, if \( V \) contains be eigenvectors of \( C \), the eigenvalues of \( C \) represent variances along the new co-ordinate axis (diagonal \( \tilde{C} \))!
Understanding multi-variate Gaussian

\[
p(x|\mu, C) = \frac{1}{(2\pi)^{d/2} |C|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu) \right\}
\]

Shifting the origin to the mean \(\mu\): \(x' = x - \mu\)

\[
C^{-1} = \left( \tilde{V} \tilde{C} V^T \right)^{-1} = \left( V^T \right)^{-1} \tilde{C}^{-1} V^{-1} = \tilde{V} \tilde{C}^{-1} V^T
\]
Understanding multi-variate Gaussian

\[ x'^T C^{-1} x' = x'^T V \tilde{C}^{-1} V^T x' \]
\[ = (V^T x')^T \tilde{C}^{-1} (V^T x') \]
\[ = \tilde{x}'^T \tilde{C}^{-1} \tilde{x}' \]
\[ = \|\tilde{x}'\|^2_{\tilde{C}^{-1}} \]

Weighted norm of \( \tilde{x}' \) - **Mahalanobis norm**

\[ \|\tilde{x}'\|^2_{\tilde{C}^{-1}} \] is the **Mahalanobis distance** from \( x \) to the mean \( \mu \).
Example - 2d Gaussian

\[ \mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad C = \begin{bmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{bmatrix} \]
PCA for dimensionality reduction

1. Given \( N \) centred data points \( \mathbf{x}^i = (x_{1i}, x_{2i}, \ldots, x_{ni})^T \in \mathbb{R}^n \), \( i = 1, 2, \ldots, N \), construct the design matrix \( \mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N] \).

2. Estimate the covariance matrix: \( \mathbf{C} = \frac{1}{N} \mathbf{X} \mathbf{X}^T \).

3. Compute eigen-decomposition of \( \mathbf{C} \).
   All eigenvectors \( \mathbf{v}_j \) are normalized to unit length.

4. Select only the eigenvectors \( \mathbf{v}_j, j = 1, 2, \ldots, k < n \), with large enough eigenvalues \( \lambda_j \).

5. Project the data points \( \mathbf{x}^i \) to the hyperplane defined by the span of
   the selected eigenvectors \( \mathbf{v}_j \): \( \tilde{x}^i = \mathbf{v}_j^T \mathbf{x}^i \)

Amount of variance explained in the projections \( \tilde{x}^i \): \( \sum_{\ell=1}^{k} \frac{\lambda_\ell}{\sum_{\ell=1}^{n} \lambda_\ell} \)
Principal Component Analysis

Data visualization using PCA

Select the 2 eigenvectors \( \mathbf{v}_1 \) and \( \mathbf{v}_2 \) with the largest eigenvalues

\[
\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n
\]

Represent points \( \mathbf{x}^i \in \mathbb{R}^n \) by two-dimensional projections

\[
\tilde{\mathbf{x}}^i = (\tilde{x}_1^i, \tilde{x}_2^i)^T,
\]

where

\[
\tilde{x}_j^i = \mathbf{v}_j^T \mathbf{x}^i, \quad j = 1, 2
\]

Plot the projections \( \tilde{x}_j^i \) on the computer screen.

You may use other eigenvectors \( \mathbf{v}_j \) with large enough eigenvalues \( \lambda_j \).
Boston Housing Dataset

Information collected by the U.S Census Service concerning housing in the area of Boston Mass.

Obtained from the StatLib archive
http://lib.stat.cmu.edu/datasets/boston.

506 cases.

Two prototasks associated with this data set:
For a given neighborhood, predict

1. nitrous oxide level
2. median value of a home
Boston Housing Dataset - attributes

There are 14 attributes in each data item of the dataset:

1. **CRIM** - per capita crime rate by town
2. **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS** - proportion of non-retail business acres per town.
4. **CHAS** - Charles River dummy - 1 if tract bounds river; 0 otherwise
5. **NOX** - nitric oxides concentration (parts per 10 million)
6. **RM** - average number of rooms per dwelling
7. **AGE** - proportion of owner-occupied units built prior to 1940
8. **DIS** - weighted distances to five Boston employment centres
9. **RAD** - index of accessibility to radial highways
10. **TAX** - full-value property-tax rate per USD 10,000
11. **PTRATIO** - pupil-teacher ratio by town
12. **B** - $1000(Bk − 0.63)^2$
13. **LSTAT** - % lower status of the population
14. **MEDV** - Median value of owner-occupied homes in USD 1000’s
How well-posed is the median house price prediction?

Gain more insight about this data set.

One may ask, for example, how well-posed is the task No. 2 of predicting the median value of a home based on the remaining 13 attributes (features) that vaguely characterise the neighborhood.
Prepare the data

From the original data set construct two data sets:
- column No. 14 (house prices) only
- the remaining columns No. 1-13.

View histogram of possible prices to see what the price distribution looks like.

It makes sense to discretize the house prices into:
"Low" - 1
"Medium" - 2
"High" - 3
"Very High" - 4

Most prices are in the Medium range, there are few extremely expensive houses.
Histogram of house prices

Boston data - histogram of prices
Prepare the data

Label the 13-dimensional data points (original data without the price attribute) based on where the corresponding house price falls.

"Low" - black star
"Medium" - blue circle
"High" - green cross
"Very High" - red square

We will use the label information to set markers for data projections on the visualization plots.
Eigenvalues of covariance matrix
Principal Component Analysis - Example

Cummulative variance explained

Cumulative sum of eigenvalues
PCA projection
Be careful...

Depends what you are after...

In this case, we are interested in the overall inherent dependency structure of the data, regardless of (arbitrary) measurement units/scales in individual dimensions!

Rescale the data to std deviation $\sigma = 1$ in each dimension - equivalent to calculating correlation coefficients instead of covariances.
Normalize the data to $\sigma = 1$
Cummulative variance explained
PCA projection
Understanding the plot

To understand the axis, you must understand the leading eigenvectors of the covariance matrix, i.e. the eigenvectors with the largest eigenvalues.

What are the dominant (large in absolute value) coordinates? What dominant co-ordinates share the same sign? (What does it mean?)

The eigenvectors stand for co-varying groups of original co-ordinates. See if you can assign reasonable names to them.