Multivariate Statistical Models

Sridhar Mahadevan
mahadeva@cs.umass.edu

University of Massachusetts
Outline

- “Learning from data”: being a detective with numbers!
- Review of (multivariate) statistical modeling
- The interplay between geometry and statistics
- Spectral analysis and statistics
- Example from medical data analysis
### TABLE 1

*Diabetes study: 442 diabetes patients were measured on 10 baseline variables; a prediction model was desired for the response variable, a measure of disease progression one year after baseline*

<table>
<thead>
<tr>
<th>Patient</th>
<th>AGE</th>
<th>SEX</th>
<th>BMI</th>
<th>BP</th>
<th>Serum measurements</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(x_1)</td>
<td>(x_2)</td>
<td>(x_3)</td>
<td>(x_4)</td>
<td>(x_5)</td>
<td>(x_6)</td>
</tr>
<tr>
<td>1</td>
<td>59</td>
<td>2</td>
<td>32.1</td>
<td>101</td>
<td>157</td>
<td>93.2</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>1</td>
<td>21.6</td>
<td>87</td>
<td>183</td>
<td>103.2</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>2</td>
<td>30.5</td>
<td>93</td>
<td>156</td>
<td>93.6</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>1</td>
<td>25.3</td>
<td>84</td>
<td>198</td>
<td>131.4</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>1</td>
<td>23.0</td>
<td>101</td>
<td>192</td>
<td>125.4</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>1</td>
<td>22.6</td>
<td>89</td>
<td>139</td>
<td>64.8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>441</td>
<td>36</td>
<td>1</td>
<td>30.0</td>
<td>95</td>
<td>201</td>
<td>125.2</td>
</tr>
<tr>
<td>442</td>
<td>36</td>
<td>1</td>
<td>19.6</td>
<td>71</td>
<td>250</td>
<td>133.2</td>
</tr>
</tbody>
</table>

Visualizing Data

The image shows a scatter plot with BMI on the x-axis and BP on the y-axis, indicating a relationship between body mass index and blood pressure.
Visualizing Data
Clustering using k-Means

- Initialize $K$ cluster means to random values.
- Repeat

1. **Assignment**: Associate each data point $n$ to the nearest cluster mean. If point $n$ is assigned to cluster $k$, record that by setting an indicator variable $r^k_n = 1$.

2. **Update**: Update the cluster means

$$m^k = \frac{\sum_n r^k_n x^n}{\sum_n r^k_n}$$
k-Means Example
k-Means Example
k-Means Example
k-Means Example

(d)
How to Measure "Length"?

Examples:

1. **Euclidean distance**: 
   \[ d(x, x^m) = \sum_i (x_i - x^m_i)^2 \]

2. **Mahalanobis distance**: 
   \[ d(x, x^m) = (x - x^m)^T \Sigma^{-1}(x - x^m) \]

3. **KL divergence**: 
   \[ d(p, q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \]  
   (relative entropy)

All of these are special cases of **Bregman divergences**
Machine learning exploits the deep connections between statistics and geometry:

- Geometrical notions: distance, length, angle, norm, similarity, neighborhood.
- Statistical notions: random variables, (co)variance, mean, regression, sufficiency.
Each data instance is modeled as a "vector" (e.g., an n-tuple of real-valued measurements, or attributes or features).

A dataset is represented as a “design" matrix $X$ (rows $\rightarrow$ instances, columns $\rightarrow$ features)

Simple model: rows and columns are vectors in Euclidean space ($\mathbb{R}^n$).

Sophisticated models: data lies in an inner product (Hilbert) space, defined by an abstract notion of "length".
Statistical Perspective

- Every data point is viewed as a multidimensional “random” variable.

- A data set is viewed as a sample from a pre-defined distribution with a fixed set of parameters (partially known).

- Formally, if we denote the data set as a random variable $X$, then $x \sim D(\theta)$.

- “Learning” $\rightarrow$ estimation of the parameters $\theta$ of the distribution $D$. 
Random Variables

- A *random variable* $X$ is a function from a sample space $S$ into the real numbers $\mathcal{R}$.
- We denote the value of the variable by $X(s)$ for element $s \in S$ of the sample space.
- A random variable $X$ induces a probability function $P_X : S \rightarrow \mathcal{X}$ where $\mathcal{X}$ is the range of $X$.

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$

- Example: 10 random variables in the diabetes data set: age, sex, BMI, BP, serum measurements.
Expectation and Variance

- The *expected value* (or “mean”) $\mu_X$ of a random variable $X$ is

$$\mu_X = E(X) = \begin{cases} \int_{-\infty}^{\infty} x f_X(x) \, dx & : \quad X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} x P(X = x) & : \quad X \text{ is discrete} \end{cases}$$

- The *variance* (or “average squared deviation from mean”) of a random variable $X$ is

$$Var(X) = E(X - \mu_X)^2$$

- The positive square root of the variance is defined as the *standard deviation* $\sigma_X$. 
Medical Data Analysis

Means in Diabetes Data

1 2 3 4 5 6 7 8 9 10
Expectation: Properties

- **Linearity:**
  \[ E(a_1 X_1 + \ldots + a_n X_n) = a_1 E(X_1) + \ldots + a_n E(X_n) \]

- **Nested expectation:**
  \[ E(E(X)) = E(X) \]

- **Expected deviation around mean is 0:**
  \[ E(X - \mu_X) = 0 \]

- **Exercise:** Show \( var(X) = E(X^2) - \mu_X^2 \)

- **Given a set of independent random variables**
  \( X_1, \ldots, X_n, \)

  \[ Var(a_1 X_1 + \ldots + a_n X_n) = a_1^2 Var(X_1) + \ldots + a_n^2 Var(X_n) \]
Sample vs. Population Statistics

- **Sample** statistics refers to properties computed from the data.

- **Population** statistics refers to the properties of the underlying distribution.

Given a random variable $X$, where $n$ samples $x_1, \ldots, x_n$ are given.

- Sample mean $= \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Sample variance $= s_{xx} = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$
Covariance

- Given two random variables $X$ and $Y$, the covariance between them is defined as

$$cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

- The covariance can also be written as (show this!)

$$cov(X, Y) = E(XY) - \mu_X \mu_Y$$

- This makes it clear that variance is simply the covariance of a random variable with itself.

- When the random variable $X$ is a vector, the covariance of $X$ becomes a matrix.
Correlation

The correlation between two random variables $X$ and $Y$ is defined as

$$
\rho_{XY} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sqrt{\sigma_{XX}} \sqrt{\sigma_{YY}}}
$$

Intuitively, the correlation between two variables measures whether increasing one variable causes the other to increase (positive, closer to 1) or decrease (negative, closer to $-1$).

If the correlation is 0, changing one variable has no effect on the other.
Predictor Response Correlations

Correlations in Diabetes Data
Variance-Covariance Matrix

- The covariance of a $p$-dimensional random variable is defined as $\Sigma_X = E((X - \mu_X)(X - \mu_X)^T)$.

- The diagonal entries are the variances, and the off-diagonal entries are covariances.

\[
\Sigma_X = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}
\end{bmatrix}
\]
Statistical Distance
If each dimension $x_i$ of a $p$-dimensional random variable $x$ has different variance, a new measure of distance is needed:

$$d^2(x, O) = \frac{x_1^2}{\sigma_{11}} + \frac{x_2^2}{\sigma_{22}} = c^2$$

This defines an *ellipse*, with center at $(0, 0)$, semi-major axis $c\sqrt{\sigma_{11}}$ and semi-minor axis $c\sqrt{\sigma_{11}}$.

Points at distance $c$ away from $x$ define a *hyper ellipsoid*:

$$d^2(x, y) = \sum_{i=1}^{p} \frac{x_i^2 - y_i^2}{\sigma_{ii}} = c^2$$
Sample Covariance Matrix

- The sample covariance $s_{ik} = \frac{1}{n} \sum_j (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$

- The diagonal entries are the sample variances, and the off-diagonal entries are sample covariances.

\[
S_X = \begin{bmatrix}
s_{11} & s_{12} & \cdots & s_{1p} \\
s_{21} & s_{22} & \cdots & s_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
&s_{p1} & s_{p2} & \cdots & s_{pp}
\end{bmatrix}
\]
Sample Covariances in Diabetes Study

Covariance Matrix for Diabetes Data

- p. 27/48
Correlation Matrix

- The correlation matrix $\rho_X = V^{-\frac{1}{2}} \Sigma_X V^{-\frac{1}{2}}$

- Here, $V^{-\frac{1}{2}}$ is a diagonal matrix whose entries are the inverse of the standard deviations $\frac{1}{\sqrt{\sigma_{ii}}}$.

- The $i^{th}k^{th}$ entry $\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}} \sqrt{\sigma_{kk}}}$

\[
\rho_X = \begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1p} \\
\rho_{21} & 1 & \cdots & \rho_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p1} & \rho_{p2} & \cdots & 1
\end{pmatrix}
\]
Diabetes Sample Correlation Matrix

Correlation Matrix of Diabetes Data
Covariance matrices are symmetric, hence all their eigenvalues are real.

In particular, the spectral decomposition applies, whereby

\[ \Sigma = V \Lambda V^T \]

Here, \( V \) is a real-valued matrix of orthonormal eigenvectors, so that \( V^T V = I \).

\( \Lambda \) is a diagonal matrix of real eigenvalues.
The eigenvalues reveal a great deal of information about the data.
Structure of Covariance Matrices

- Sum of Projection Matrices: $\Sigma = \sum_{i=1}^{p} \lambda_i v_i v_i^T$
  - Note each “outer product” $v_i v_i^T$ is a $p$ by $p$ matrix of rank 1.
  - By sorting the eigenvalues by their size, a “low-rank” approximation can be constructed.

- Inverse covariance matrix: $\Sigma^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} v_i v_i^T$

- “Square root representation”: $\Sigma = (V \Lambda^{\frac{1}{2}} V^T)(V \Lambda^{\frac{1}{2}} V^T)$
Positive Definite Matrices

- A symmetric matrix $A$ is positive definite if and only if for all nonzero vectors $x$, $x^T A x > 0$. Equivalently, all its eigenvalues are real and positive.

- Positive definite matrices define a distance metric:

$$d^2(x, 0) = x^T A x = x^T \left( \sum_{i=1}^{p} \lambda_i v_i v_i^T \right) x$$

$$= \sum_{i=1}^{p} \lambda_i (x^T v_i)(v_i^T x) = c^2$$

- Hyper-ellipsoid: axes defined by $\pm \frac{c}{\sqrt{\lambda_i}} v_i$. 
Multivariate Gaussian

A general distance measure defined as

\[ d^2(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu) = c^2 \]

This defines a hyper-ellipsoid with center at \( \mu \), and axes defined as \( \pm c \sqrt{\lambda_i} v_i \).

Multivariate Gaussian

\[ p_\theta(x) = \frac{1}{(2\pi)^{p/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)} \]
Multivariate Gaussian PDF

\[ \mu = [0, 0]^T, \Sigma = \begin{bmatrix} .25 & .3 \\ .3 & 1 \end{bmatrix} \]
Multivariate Gaussian CDF
PCA

- How can we construct a “reduced" representation of the data, which eliminates irrelevant variables.

- Principal Components Analysis is a dimensionality reduction method that finds a compact representation of the data $X$. 
PCA

Problem: find a linear combination \( Y = \sum_{i=1}^{p} \alpha_i x_i \) of the predictor variables \( x_i \) that maximizes the variance \( Var(\alpha X) \).

Exercise: Show that \( Var(Y) = Var(\alpha^T X) = \alpha^T \Sigma \alpha \).

Since \( Var(\alpha^T X) \) is unbounded for arbitrary \( \alpha \), we restrict our attention to linear combinations such that \( \alpha^T \alpha = 1 \).
PCA

Problem: Find $\alpha \in \mathbb{R}^p$ such that $\text{var}(\alpha^T X)$ is maximized subject to $\alpha^T \alpha = 1$.

Solution: Using Lagrange multipliers, we can formulate this optimization problem as the unconstrained problem of maximizing:

$$L(\lambda, \alpha) = \alpha^T \Sigma \alpha - \lambda (\alpha^T \alpha - 1)$$

Solving for the gradient $\frac{\partial L}{\partial \alpha} = 2 \Sigma \alpha - 2\lambda \alpha$

Setting the gradient to 0 gives us $\Sigma \alpha = \lambda \alpha$.

In other words, $\alpha$ must be an eigenvector of the covariance matrix $\Sigma$. 
PCA

- In fact, the first principal component must be the eigenvector associated with the largest eigenvalue.

\[ \text{Var}(Y) = \text{Var}(\alpha^T X) = \alpha^T \Sigma \alpha = \lambda \alpha^T \alpha = \lambda \]

- So, we see that the largest eigenvalue is exactly the variance \( \text{Var}(Y) \) of the new derived variable.

**Generalized formulation:** Find a set of \( m \) new variables \( Y_1, \ldots Y_m \) such that \( Y_i = \alpha_i^T X \) and \( \text{Var}(Y_i) \) is maximized, where \( \alpha_i^T \alpha_i = 1 \) and for all \( 1 \leq j \leq i - 1 \), \( \alpha_i^T \alpha_j = 0 \).

**Solution:** The principal components are the eigenvectors of \( \Sigma \).
PCA and Variance

- **Theorem:** If \( X = (X_1, \ldots, X_p) \), then
  \[
  \sum_i \text{var}(X_i) = \sum_{i=1}^p \sigma_{ii} = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{Var}(Y_i)
  \]

- **Proof:** Given spectral decomposition \( \Sigma = V \Lambda V^T \), we have
  \[
  \sum_i \text{Var}(X_i) = \text{tr}(\Sigma) = \text{tr}(V \Lambda V^T) = \text{tr}(V^T V \Lambda) = \text{tr}(\Lambda) = \sum_{i=1}^p \lambda_i
  \]
  Proportion of population variance due to \( k^{th} \) principal component is given by
  \[
  \frac{\lambda_k}{\sum_{i=1}^p \lambda_i}
  \]
PCA on Diabetes Data

![PCA for Diabetes Data](chart)
Factor Analysis

- Factor analysis aims to uncover a small number of "hidden" variables whose interaction explains the variance in the data.

- **Factor model:** \( X - \mu_X = LF + \epsilon \)
  
  where \( X \) is a \( p \)-dimensional variable, \( F \) is a set of \( m < p \) common factors, \( L \) is a \( p \) by \( m \) matrix of "loadings", and \( \epsilon \) is a set of \( p \) individual factors.

- **Assumptions:**
  - \( F \) and \( \epsilon \) are independent
  - \( E(F) = 0, \text{Cov}(F) = I, E(\epsilon) = 0, \text{Cov}(\epsilon) = \Psi \), a diagonal matrix