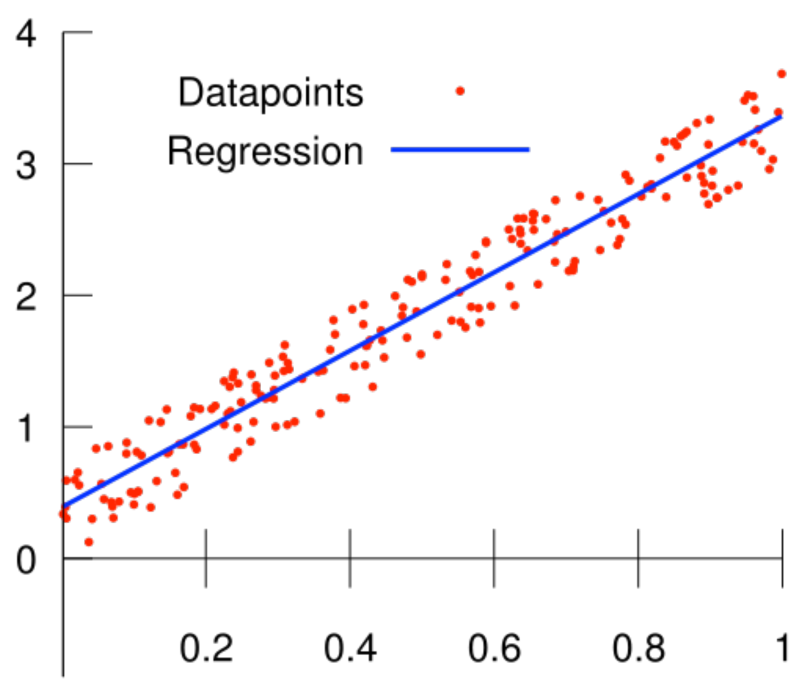
CSI 436/536
Introduction to Machine Learning

Dimension reduction and total LLSE

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

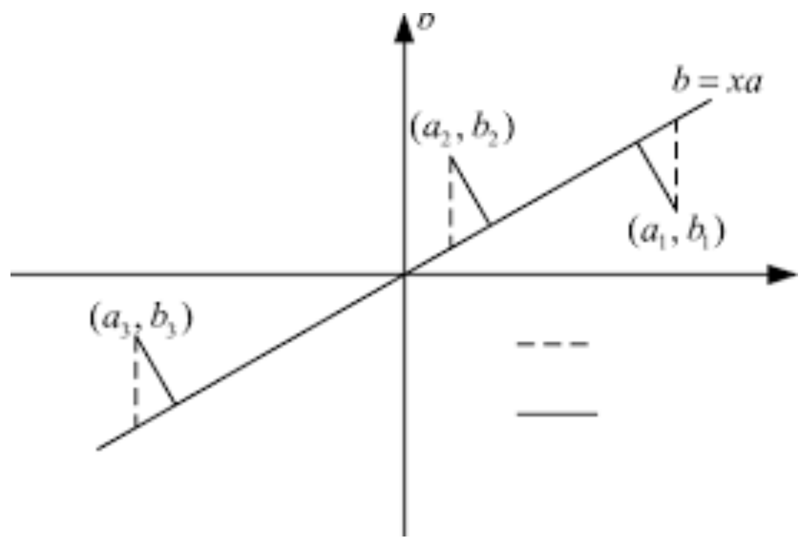
Linear least squares

- Fitting a line (linear model) by minimizing prediction error
  - The error is on the y-axis only



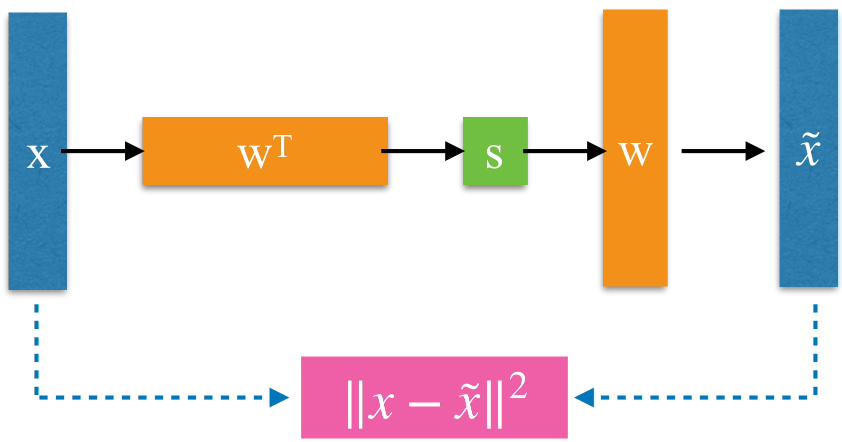
Total least squares estimation

- Fitting a line (linear model) by minimizing the total error
  - The error is for both x- and y- coordinates



TLSE as encoder-decoder

- The TLSE model can be understood with an encoder-decoder model



  - The encoder takes the input and reduces it to a code
  - The decoder takes the code and reconstructs it to an output
  - The low dimensional code is the “information bottleneck”
- Learning is achieved through “self-supervision”, i.e., reducing the error between the input and the reconstructed output

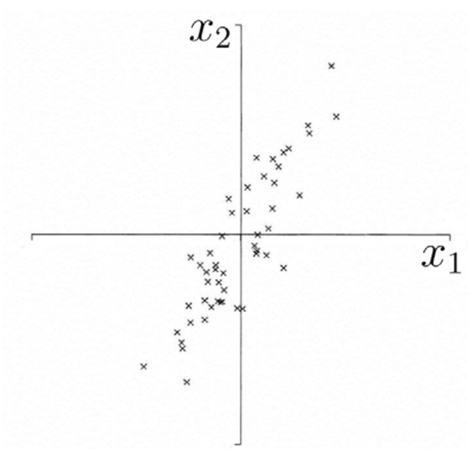
TLSE as data compression

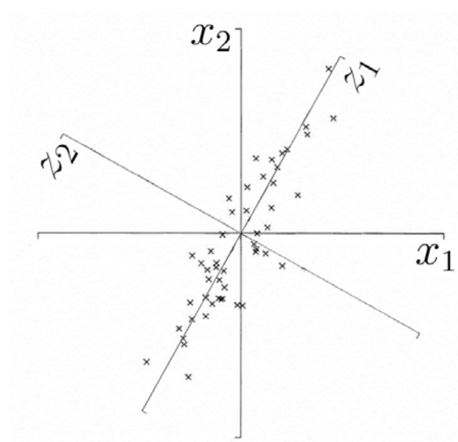
- The encoder-decoder interpretation of TLSE also suggests that TLSE can be viewed as a data compression procedure
  - Input and output have dimension d
  - The code has dimension 1
- Compression

Total least squares

- given m data vector of dimensions $X = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$.
- assumption: data are centered, i.e., $\sum_i \mathbf{x}_i = 0$.
- find the best one-dimensional approximation to X minimizing ℓ_2 errors.
- specifically, find a unit vector $\mathbf{v}$ (why?), and scaling factors $(s_1, , s_m)$, s.t.,

$$\min_{\mathbf{v}:\|\mathbf{v}\|_2=1, s_1,,s_m} \sum_{i=1}^{m} \|x_i - s_i\mathbf{v}\|_2^2.$$





solution

First, given $\mathbf{v}$, find optimal solution to s_i.

$$
\begin{aligned}
& \frac{\partial}{\partial s_i} \sum_{i=1}^{m} \|x_i - s_i \mathbf{v}\|_2^2 = \frac{\partial}{\partial s_i} \|x_i - s_i \mathbf{v}\|_2^2 = 0 \\
\Rightarrow \quad & \frac{\partial}{\partial s_i} (x_i - s_i \mathbf{v})^T (x_i - s_i \mathbf{v}) = 0 \\
\Rightarrow \quad & \frac{\partial}{\partial s_i} (s_i^2 \mathbf{v}^T \mathbf{v} - 2 s_i \mathbf{v}^T \mathbf{x}_i + \mathbf{x}_i^T \mathbf{x}_i) = 0 \\
\Rightarrow \quad & s_i = \mathbf{x}_i^T \mathbf{v}.
\end{aligned}
$$

$$
\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{i=1}^{m} \|\mathbf{x}_i - (\mathbf{x}_i^T \mathbf{v})\mathbf{v}\|_2^2.
$$

solution (continued)

$$\begin{aligned}\sum_{i=1}^{m} \|\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{v})\mathbf{v}\|_2^2 &= \sum_{i=1}^{m}(\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{v})\mathbf{v})^T(\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{v})\mathbf{v}) \\ = \quad & \sum_{i=1}^{m}(\mathbf{x}_i^T\mathbf{x}_i - 2(\mathbf{x}_i^T\mathbf{v})\mathbf{x}_i^T\mathbf{v} + (\mathbf{x}_i^T\mathbf{v})^2\mathbf{v}^T\mathbf{v}) \\ = \quad & \sum_{i=1}^{m}(\mathbf{x}_i^T\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{v})^2) = \sum_{i=1}^{m}(\mathbf{x}_i^T\mathbf{x}_i - \mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v})\end{aligned}$$

Total LLSE

Furthermore

$$\sum_{i=1}^{m} \mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v} = \mathbf{v}^T \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{v} = \mathbf{v}^T (XX^T) \mathbf{v}.$$

Recall XX^T is the covariance matrix of data matrix X

- because X is centered.

Equivalently, in PCA, we seek

$$\begin{array}{ll} \max_{\mathbf{v}} & \mathbf{v}^T (XX^T) \mathbf{v} \\ s.t. & \|\mathbf{v}\|_2^2 - 1 = 0 \end{array}$$

Total LLSE

Constrained optimization:

$$\begin{array}{ll} \max_{\mathbf{v}} & \mathbf{v}^T(XX^T)\mathbf{v} \\ s.t. & \mathbf{v}^T\mathbf{v} - 1 = 0 \end{array}$$

Lagrangian

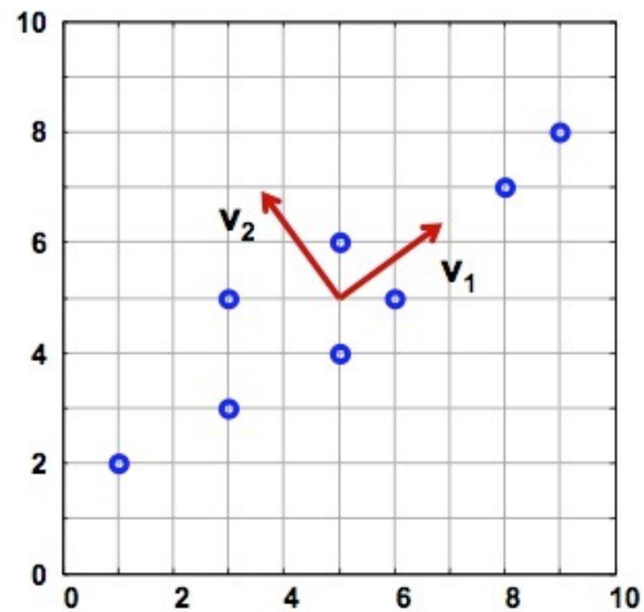
$$L(\mathbf{v}, \lambda) = \mathbf{v}^T(XX^T)\mathbf{v} - \lambda(\mathbf{v}^T\mathbf{v} - 1)$$

Derivative w.r.t. **x** sets to zero

$$(XX^T)\mathbf{v} = \lambda\mathbf{v}.$$

Example

- data
  - X={(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)}
- centering
- covariance
- EVD
  - $\lambda 1 = 9.34$
  - $\lambda 2 = 0.41$
  - v1 = [0.81 0.59], v2 = [0.81 -0.59],



Dimension reduction

- Total LLSE fits a 1D line to a set of multi-dimensional vectors with minimum distortion
- This can be equivalently viewed as finding a low dimensional approximation (in this case 1D) of a high-dimensional data point
- The procedure is known as dimension reduction, and it is behind image compression algorithms
- We will talk about the more general version of dimension reduction known as principal component analysis (PCA) later