Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 21:
Principal Components Analysis
Factor Analysis & Probabilistic PCA

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Dimensionality Reduction

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- **Goal:** Infer label/response y given only features x
- **Classical:** Find latent variables y good for *compression* of x
- **Probabilistic learning:** Estimate parameters of joint distribution p(x,y) which *maximize marginal probability* p(x)

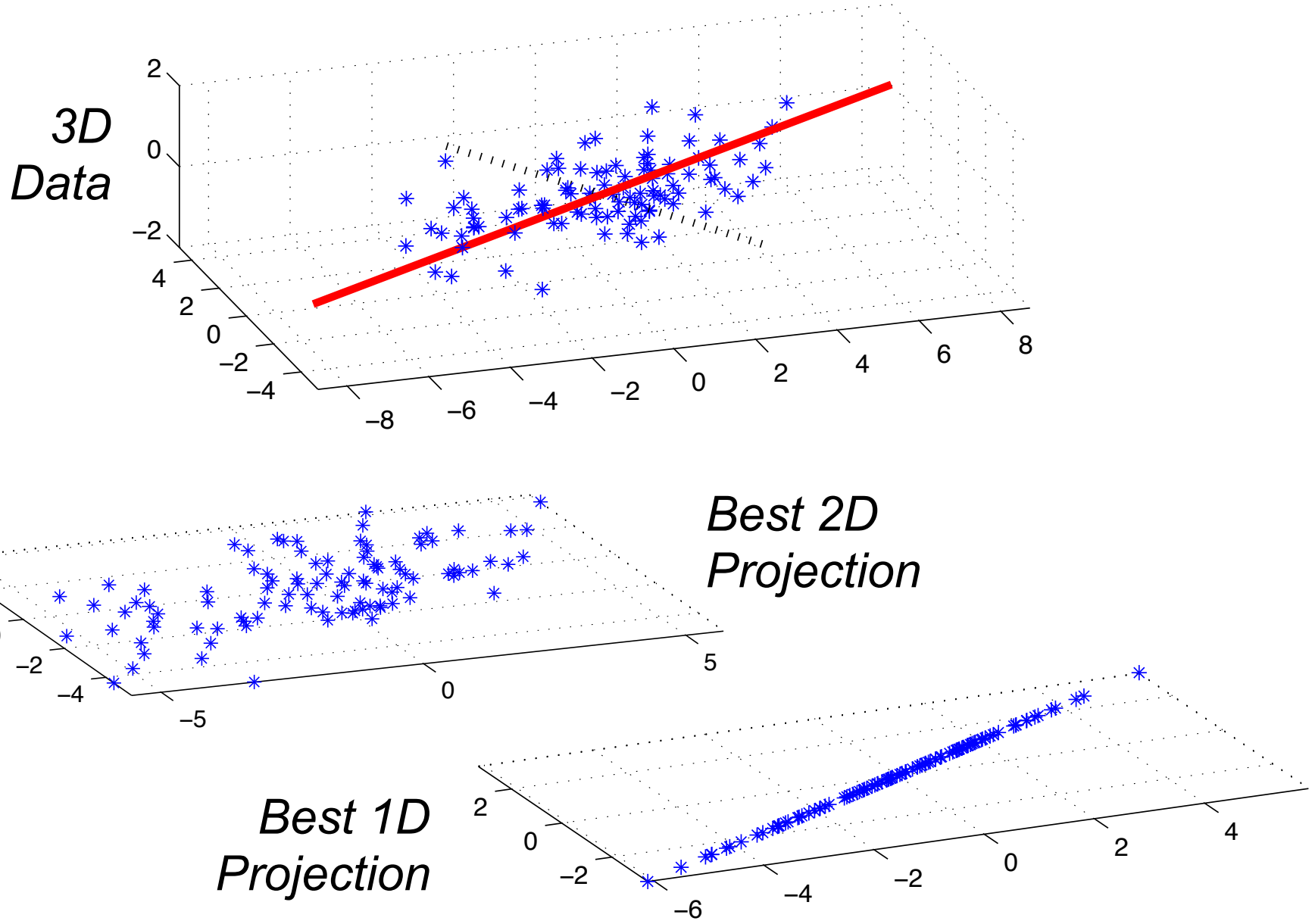
PCA Objective: Compression

- Observed feature vectors: $x_n \in \mathbb{R}^D, \quad n = 1, 2, \ldots, N$
- Hidden manifold coordinates: $z_n \in \mathbb{R}^M, \quad n = 1, 2, \ldots, N$
- Hidden linear mapping: $\tilde{x}_n = W z_n + b$ $\quad W \in \mathbb{R}^{D \times M}$, $b \in \mathbb{R}^{D \times 1}$

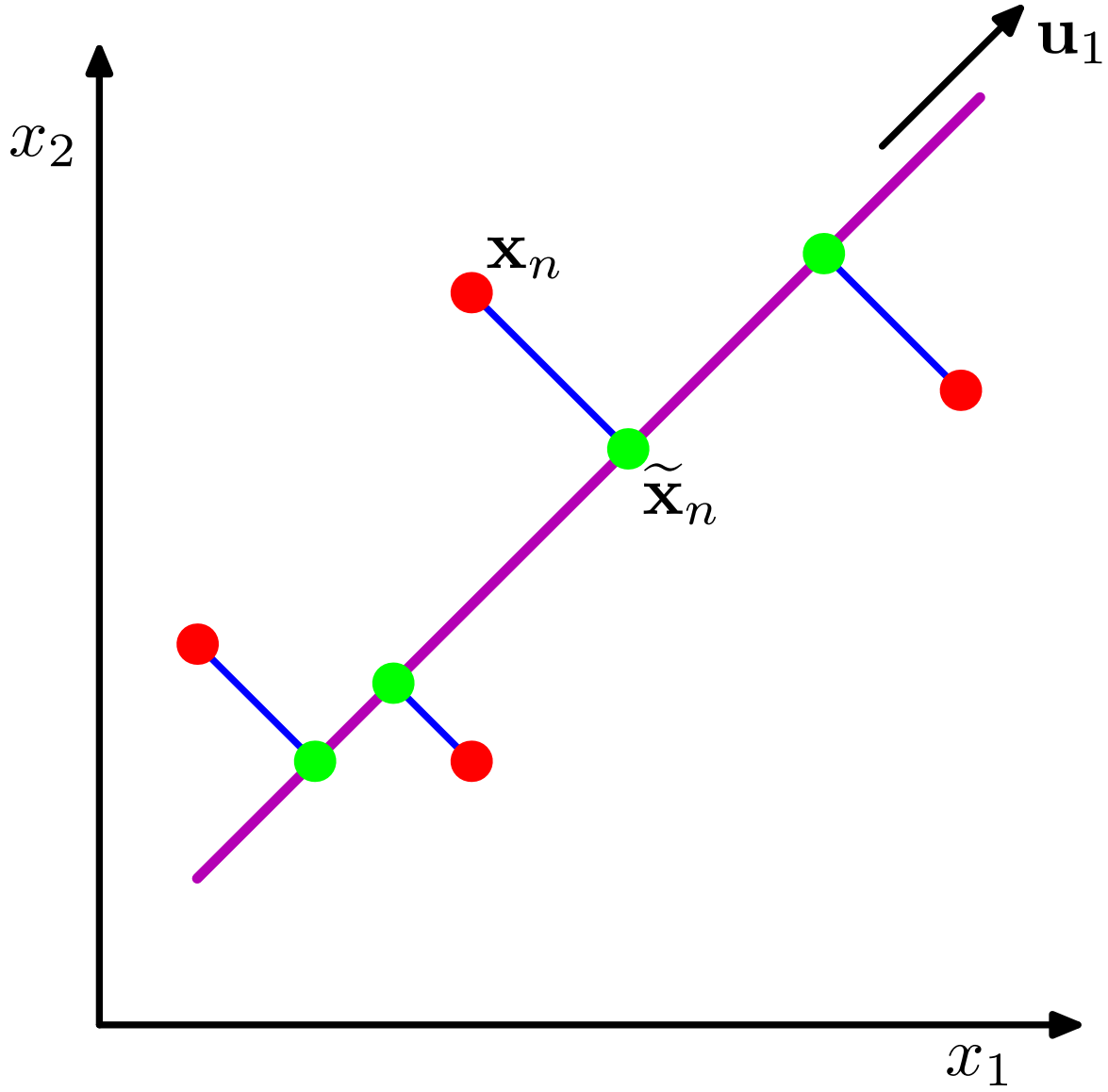
$$J(z, W, b \mid x, M) = \sum_{n=1}^{N} ||x_n - \tilde{x}_n||^2 = \sum_{n=1}^{N} ||x_n - W z_n - b||^2$$

- We can find the *global optimum via an eigendecomposition*
- Contrast with the K-means algorithm for clustering…

Principal Components Analysis (PCA)



Maximizes Variance & Minimizes Error



C. Bishop, Pattern Recognition & Machine Learning

PCA Optimal Solution

$$J(z, W, b \mid x, M) = \sum_{n=1}^{N} ||x_n - \tilde{x}_n||^2 = \sum_{n=1}^{N} ||x_n - W z_n - b||^2$$

$$b = \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad X = [x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_N - \bar{x}]$$

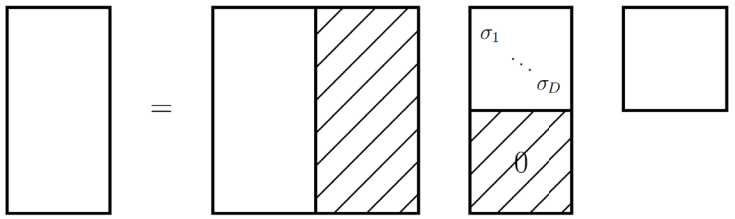
- Option A: Eigendecomposition of sample covariance matrix

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N} X X^T = U \Lambda U^T$$

Construct W from eigenvectors with M largest eigenvalues

- Option B: Singular value decomposition (SVD) of centered data

$$X = U S V^T$$



Construct W from singular vectors with M largest singular values

Probabilistic PCA & Factor Analysis

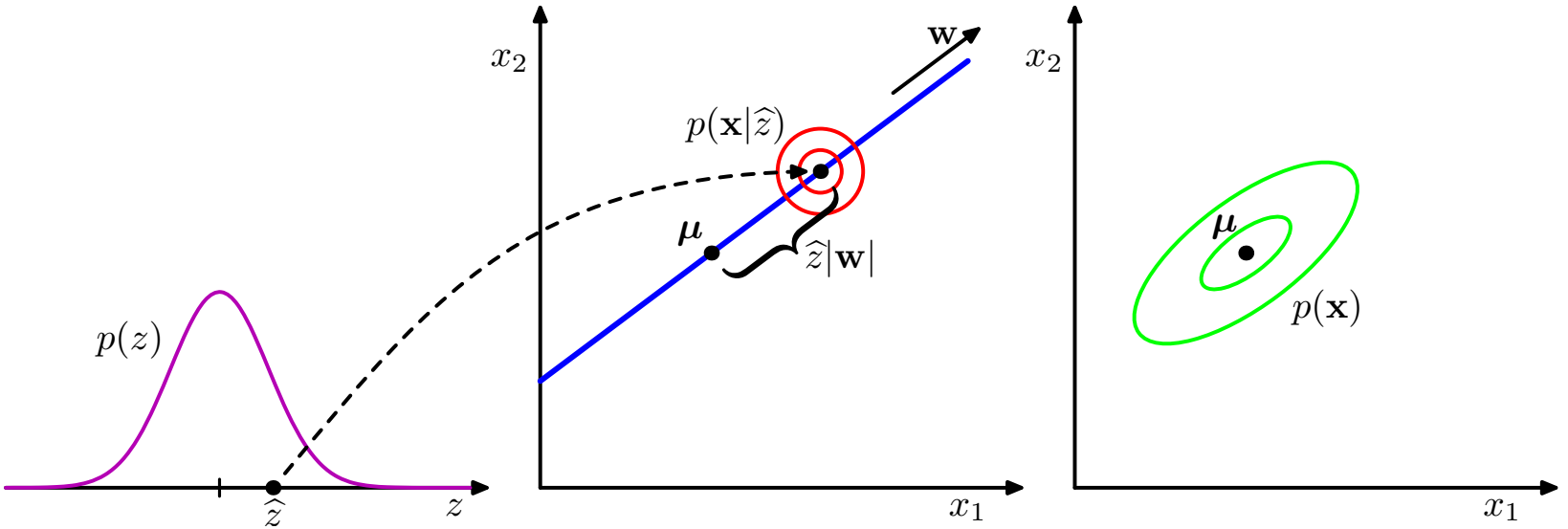
- **Both Models:** Data is a linear function of low-dimensional latent coordinates, plus Gaussian noise

$$p(x_i \mid z_i, \theta) = \mathcal{N}(x_i \mid W z_i + \mu, \Psi) \qquad p(z_i \mid \theta) = \mathcal{N}(z_i \mid 0, I)$$

$$p(x_i \mid \theta) = \mathcal{N}(x_i \mid \mu, W W^T + \Psi)$$

low rank covariance parameterization

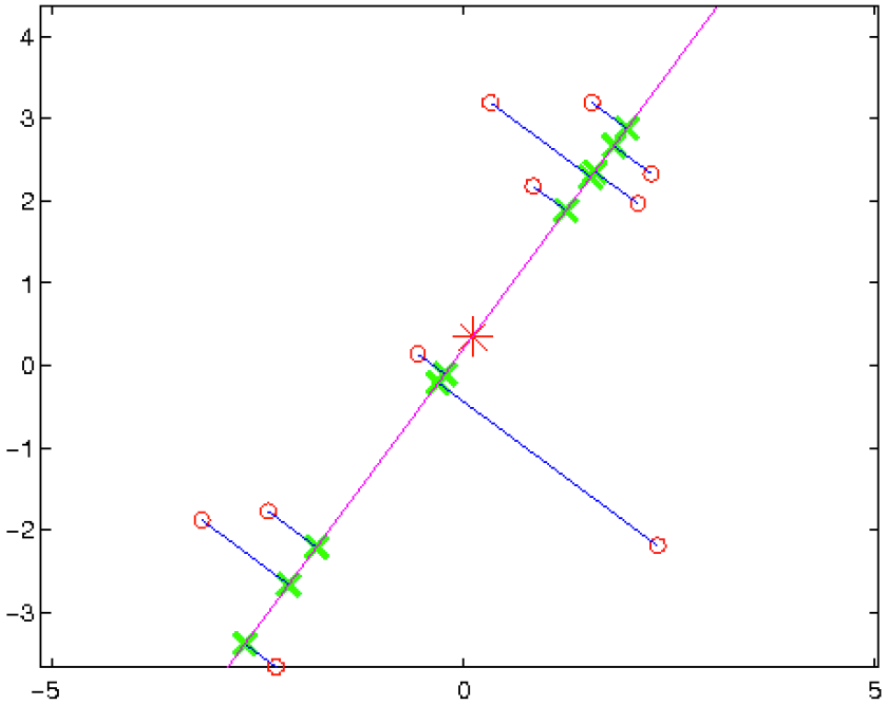
- **Factor analysis:** Ψ is a general diagonal matrix
- **Probabilistic PCA:** $\Psi = \sigma^2 I$ is a multiple of identity matrix



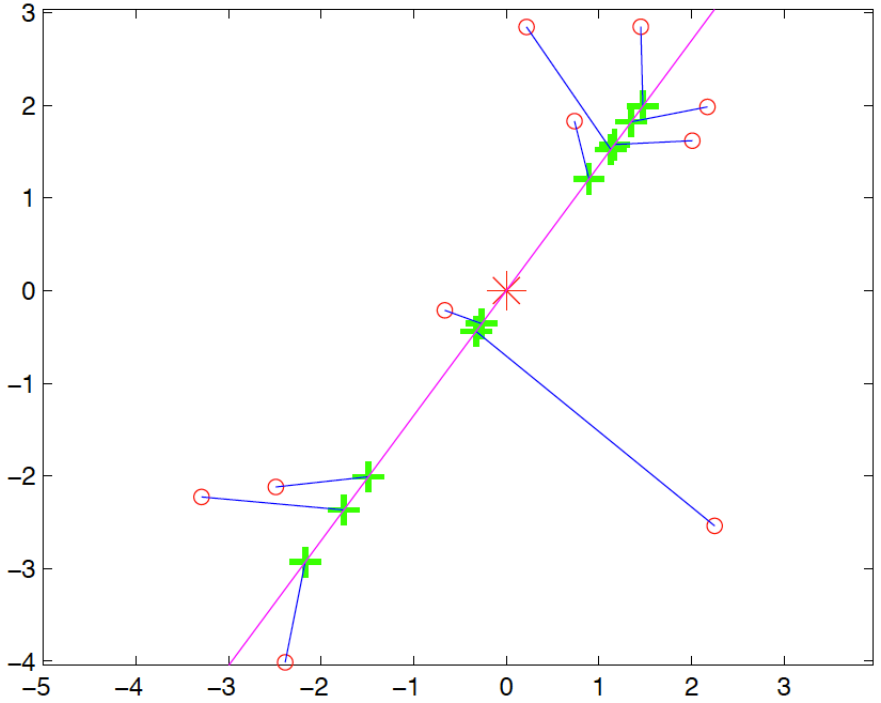
C. Bishop, Pattern Recognition & Machine Learning

PCA versus Probabilistic PCA

$$p(z_i \mid x_i, W, \mu, \Psi) = \mathcal{N}(z_i \mid \Sigma_i W^T \Psi^{-1}(x_i - \mu), \Sigma_i) \qquad \Sigma_i^{-1} = I + W^T \Psi^{-1} W$$



Standard PCA
(orthogonal projection)

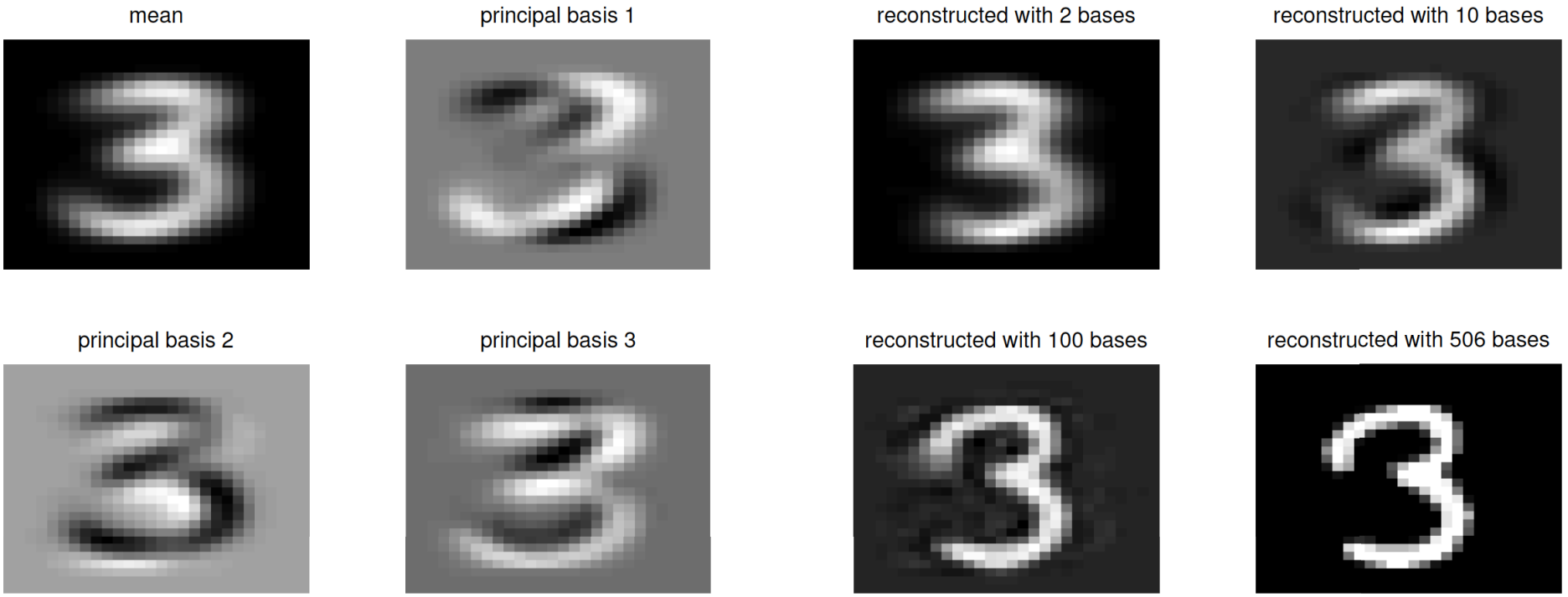


Probabilistic PCA
(shrunk towards mean)

- Maximum likelihood estimates of probabilistic PCA parameters are equal to the classic PCA eigenvector solution
- These optimal parameters are not unique. Why?
- Most likely latent coordinates are biased towards mean (zero)

Principal Components Analysis Example

PCA Analysis of MNIST Images of the Digit 3



- PCA models all translations of data equally well
- PCA models all rotations of data equally well
- Appropriate when modeling quantities over time, space, etc.

Factor Analysis Example

Features of Cars in 2004

Suggested retail price in USD

Price to dealer in USD

Engine size in liters

Num. cylinders

Horsepower

City MPG

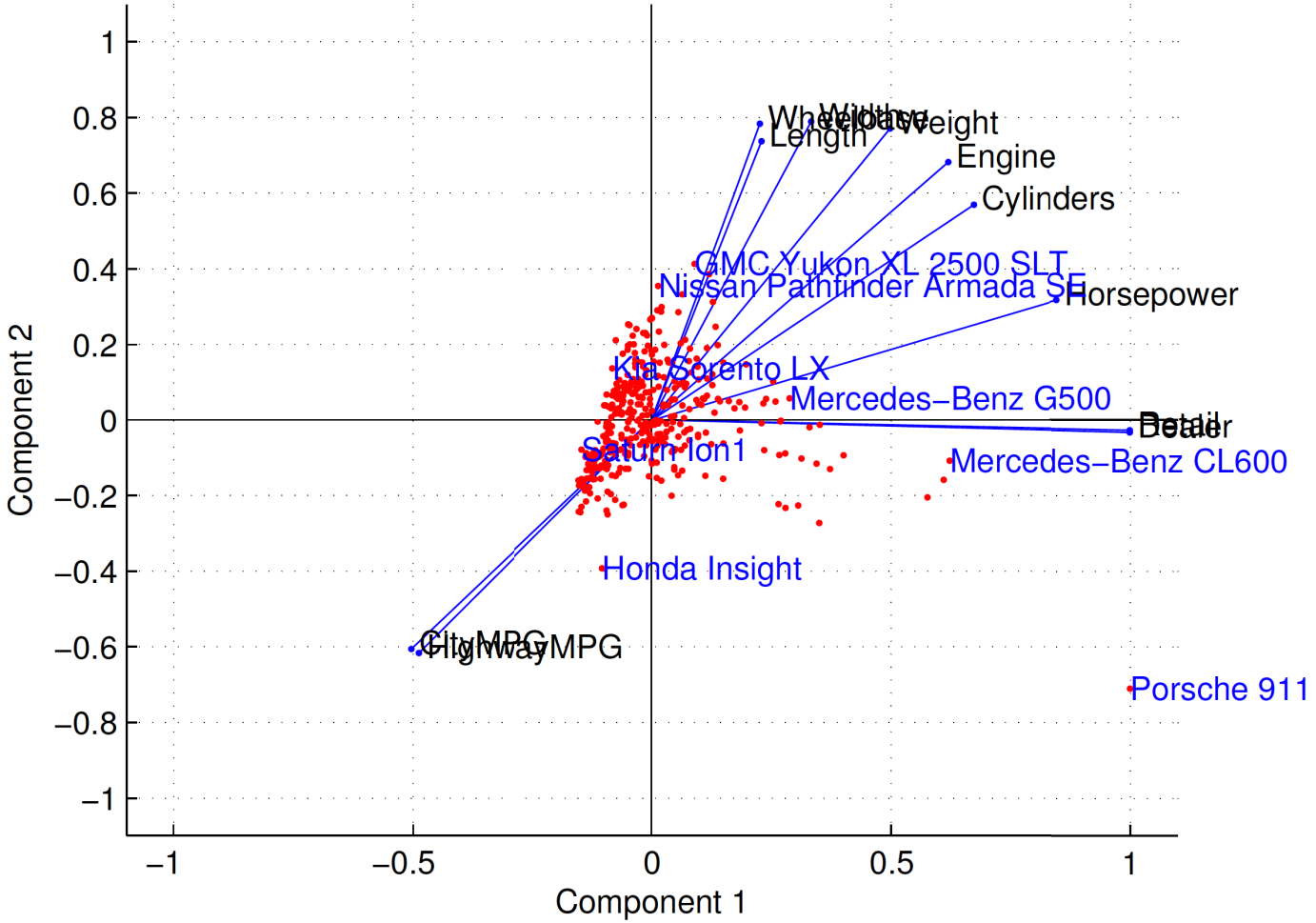
Highway MPG

Weight in pounds

Wheelbase in inches

Length in inches

Width in inches



- Factor analysis models all translations of data equally well
- Factor analysis models all rescalings of data equally well
- Appropriate when modeling vectors measured in varying units