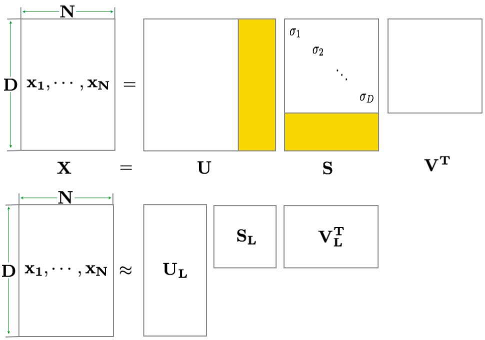
Latent Linear Models

Singular Value Decomposition(SVD)



- $X = USV^T$
- Time: $O(ND \min(N, D))$
- $X^T X = VDV^T$ (row space)
  $XX^T = UDU^T$ (column)
  $D = S^2$ (eigenvalues)
- Truncated SVD:

$$X \approx X_L = U_{:,1:L} S_{1:L,1:L} V^T_{1:L,:}$$

Parameters: $L(N + D + 1)$

Principle Component Analysis(PCA)

1. A analysis (or statistical) view

Suppose a random variable $x \in \mathbb{R}^D$ with zero-mean $\mathbb{E}(x) = 0$, and try to find $L < D$ principle components $y \in \mathbb{R}^d$:

- y_i's are 'uncorrelated' linear combinations of x:

$$y_i = u_i^T x \in R, u_i \in R^D$$

- The variance of y_i is maximized subject to

$$u_i^T u_i = 1, i = 1, \cdots, d \quad \text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_L) > 0$$

[Note: I'll use i to index features and j for samples.]

PCA Solution

Define the <u>covariance matrix</u>:

$$\Sigma_x = \mathbb{E}[xx^T]$$

and suppose $\text{rank}(\Sigma_x) \geq L$. Then the L principle components are given by

$$y_i = u_i^T x$$

where $\{u_i\}_{i=1}^{L}$ are L orthonormal eigenvectors of Σ_x associated with its L largest eigenvalues $\{\lambda_i\}_{i=1}^{d}$. And

$$\lambda_i = \text{Var}(y_i), i = 1, \cdots, L$$

Proof:

For the sake of simplicity, assume Σ_x doesn't have repeated eigenvalues.

$$\mathsf{Var}(y) = \mathsf{Var}(u^T x) = \mathbb{E}(u^T x x^T u) = u^T \Sigma_x u$$

(1) For the first principle component:

$$\max_{u_1 \in \mathbb{R}^D} \; u_1^T \Sigma_x u_1 \quad \text{s.t. } u_1^T u_1 = 1$$

And the Lagrangian is given by:

$$\mathcal{L} = u_1^T \Sigma_x u_1 + \lambda_1 (1 - u_1^T u_1)$$

Necessary condition:

$$\frac{\partial \mathcal{L}}{\partial u_1} = 2\Sigma_x u_1 - 2\lambda_1 u_1 = 0 \Rightarrow \Sigma_x u_1 = \lambda_1 u_1$$

Variance of y_1: $Var(u_1^T x) = u_1^T \Sigma_x u_1 = \lambda_1 u_1^T u_1 = \lambda_1$

So λ_1 is the largest eigenvalue of Σ_x and u_1 is the corresponding eigenvector.

(2) For the second principle component:

$$\max_{u_2 \in R^D} \; u_2^T \Sigma_x u_2 \; s.t. \; u_2^T u_2 = 1, u_2^T u_1 = 0$$

The Lagrangian is:

$$\mathcal{L} = u_2^T \Sigma_x u_2 + \lambda_2 (1 - u_2^T u_2) + \gamma u_2^T u_1$$

Necessary condition:

$$\Sigma_x u_2 - \lambda_2 u_2 + \frac{\gamma}{2} u_1 = 0 \qquad \Rightarrow \qquad \gamma = 0, \Sigma_x u_2 = \lambda_2 u_2, Var(y_2) = \lambda_2$$

(3) For the remaining principle components:

$$\max_{u_i \in R^D} \; u_i^T \Sigma_x u_i \; s.t. \; u_i^T u_i = 1, u_j^T u_i = 0, j = 1, \cdots, i-1$$

- Non-zero mean random variables: $y_i = u_i^T x + a_i$

$$\mu = \mathbb{E}(x), \Sigma_x = \mathbb{E}[(x-\mu)(x-\mu^T)] \text{ with } a_i = -u_i^T \mu$$

- **Sample PCA:** given N i.i.d. samples $\{x_j\}_{j=1}^N$ of the zero-mean r.v. x

  Data matrix: $\mathbf{X} = [x_1, \cdots, x_N]$

$$\Sigma_N = \frac{1}{N} \sum_{j=1}^{N} x_j x_j^T = \mathbf{X}\mathbf{X}^\mathbf{T}$$

  Repationship between PCA and sample PCA:

  if x is Gaussian, the eigenvector of Σ_N is an asymptotically consistent unbiased estimation for the corresponding eigenvector of Σ_x.

PCA via SVD

Data matrix $\mathbf{X} = [x_1, \cdots, x_N] = USV^T$. And

$$\mathbf{X}\mathbf{X}^\mathbf{T} = US^2U^T$$

- **Score:**
  - $y_j = U_d^T x_j$ where the columns of U_d is the first d columns of U.
  - $y_j = S_d^2 V_j^T$ where S_d is the first d singular values.

2. A synthesis (or geometric) view

Given a set of points $x_1, \cdots, x_N \in \mathbb{R}^D$, try to find an (affine) subspace $S \subset \mathbb{R}^D$ of dimension L, $\dim(S) = L$ that best fits these points.

$$x_j = \mu + W_L y_j + \epsilon_j \;\; j = 1, \cdots, N$$

where $W_L \in \mathbb{R}^{D \times L}$ whose columns form a basis for the subspace and y_j is the vector of new coordinates of x_j in the subspace.

$$\min_{\mu, W_d, \{y_j\}} \quad \sum_{j=1}^{N} ||x_j - \mu - \mathbf{W_L} z_j||^2 = ||\mathbf{X} - \mu \mathbf{1^T} - \mathbf{W_L Y}||_F^2$$

$$\text{s.t.} \qquad \mathbf{W_L^T W_L} = \mathbf{I_L}, \qquad \sum_{j=1}^{N} y_j = 0$$

Solution: The Lagrangian:

$$\mathcal{L} = \sum_{j=1}^{N} ||x_j - \mu - W_L y_j||^2 + \gamma^T (\sum_{j=1}^{N} y_j) + \text{trace}[(I_L - W_L^T W_L)\Lambda]$$

- The derivatives of $\mathcal{L}$ with respect to μ, z_j

$$\frac{\partial \mathcal{L}}{\partial \mu} = -2\sum_{j=1}^{N}(x_j - \mu - W_L y_j) = 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$\frac{\partial \mathcal{L}}{\partial z_j} = -2W_L^T(x_j - \mu - W_L y_j) + \gamma = 0 \qquad \Rightarrow \qquad \gamma = 0,\ \hat{z}_j = W_L^T(x_j - \hat{\mu})$$

Now we suppose $\hat{\mu} = 0$

- Optimization over W_L:

$$\min_{W_L} \sum_{j=1}^{N} ||x_j - W_d W_d^T x_j||^2 \qquad \text{s.t.} W_L^T W_L = I_L$$

$$\begin{aligned}
\sum_{j=1}^{N} ||x_j - W_L W_L^T x_j||^2 &= \sum_{j=1}^{N} x_j^T (I_D - W_L W_L^T)^T (I_D - W_L W_L^T) x_j \\
&= \sum_{j=1}^{N} x_j^T (I_D - W_L W_L^T) x_j \\
&= \text{trace}(X^T (I_D - W_d W_d^T) X) \\
&= \text{trace}((I_D - W_L W_L^T) X X^T)
\end{aligned}$$

Optimization problem:

$$\max_{U_d} \text{trace}(W_L W_L^T X X^T) \; s.t. W_L^T W_L = I_L$$

$\Lambda = W_L^T X X^T W_L$.
Then we only need to $\max \text{trace}(\Lambda)$

$\Rightarrow \; \Lambda = diag(\lambda_1, \cdots, \lambda_L)$, W_L is the first L columns of U where $X = U \Sigma V^T$

Model Selection for PCA

Reconstruction error:

$$E(\mathcal{D}, L) = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} ||\mathbf{x_j} - \hat{\mathbf{x}}_\mathbf{j}||^2 = \sum_{i=L+1}^{D} \lambda_i$$

Explained variance:

$$F(\mathcal{D}, L) = \frac{\sum_{i=1}^{L} \lambda_i}{\sum_{k=1}^{D} \lambda_k}$$

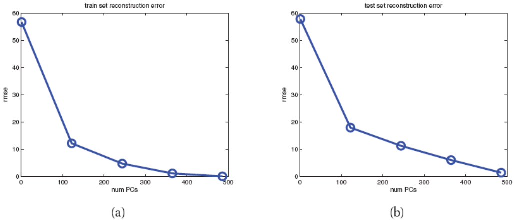


Figure 12.14 Reconstruction error on MNIST vs number of latent dimensions used by PCA. (a) Training set. (b) Test set. Figure generated by `pcaOverfitDemo`.

- **Profile likelihood:** One way to automate the detection of "regime change".

  Suppose some measure of one model:(e.g. eigenvalues here)

  $$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$$

  Partition all these values into two regimes:

  $$\begin{array}{ll} k \leq L & \lambda_k \sim \mathcal{N}(\mu_1, \sigma^2) \\ k > L & \lambda_k \sim \mathcal{N}(\mu_2, \sigma^2) \end{array}$$

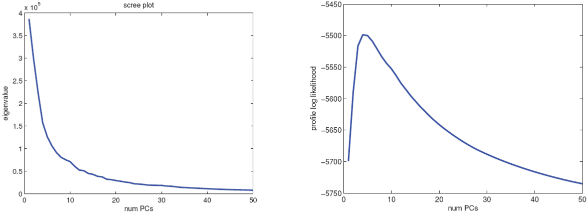
Profile log likelihood:

$$l(L) = \sum_{k=1}^{L} \log \mathcal{N}(\lambda_k | \mu_1(L), \sigma^2(L)) + \sum_{k=L+1}^{N} \log \mathcal{N}(\lambda_k | \mu_2(L), \sigma^2(L))$$

where:

$$\mu_1(L) = \frac{\sum_{k \le L} \lambda_k}{L}, \quad \mu_2(L) = \frac{\sum_{k > L} \lambda_k}{N - L}$$

$$\sigma^2(L) = \frac{\sum_{k \le L} (\lambda_k - \mu_1(L))^2 + \sum_{k > L} (\lambda_k - \mu_2(L))^2}{N}$$



(a) (b)

Figure 12.16 (a) Scree plot for training set, corresponding to Figure 12.14(a). (b) Profile likelihood. generated by pcaOverfitDemo.

Probabilistic PCA(PPCA)

- PCA: find a low-dimensional representation $\{y_j \in \mathbb{R}^L\}$ of a set of sample points $\{x_j \in \mathbb{R}^D\}$
- Generative PCA model:

$$x = \mu + U_d y + \epsilon$$

Treat the low-dimensional representation y and error ϵ as independent random variables. (Here both **Gaussian**)

Suppose the mean and covariance of y and ϵ are denoted respectively as ,

$$(\mu_y, \Sigma_y), \qquad (\mu_\epsilon, \Sigma_\epsilon)$$

Then we can get:

$$\mu_x = \mu + U_d \mu_y + \mu_\epsilon, \qquad \Sigma_x = U_d \Sigma_y U_d^T + \Sigma_\epsilon$$

So in general we cannot rescover the model parameters from μ_x, Σ_x.

PPCA makes the following assumptions:

- For the mean, assume $\mu_y = 0$ and $\mu_\epsilon = 0$.

$$\hat{\mu} = \mu_x$$

- For the covariance, we want Σ_y to capture as much information about Σ_x as possible(full rank) and Σ_ϵ to be as close to zero as possible(same variance, less information).

1. $\Sigma_y = I_d$
2. $\Sigma_\epsilon = \sigma^2 I_D$

$$\Sigma_x = U_d U_d^T + \sigma^2 I_D$$

- Since $U_d U_d^T$ has rank d, $D-d$ eigenvalues of $U_d U_d^T$ must be equal to zero.
- Since σ is as small as possible, the smallest $D-d$ eigenvalues of Σ_x must be equal to each other and equal to σ^2.

$$\sigma^2 = \lambda_{d+1} = \lambda_{d+2} = \cdots = \lambda_D$$

- To find U_d,

$$\Sigma_x = [U_1 \ U_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \sigma^2 I_{D-d} \end{bmatrix} [U_1 \ U_2]^T$$

$$
\begin{aligned}
U_d U_d^T &= \Sigma_x - \sigma^2 I_D \\
&= [U_1 \ U_2] \left[\begin{array}{cc} \Lambda_1 - \sigma^2 I_d & 0 \\ 0 & 0 \end{array} \right] [U_1 \ U_2]^T \\
&= U_1(\Lambda_1 - \sigma^2 I_d) U_1^T
\end{aligned}
$$

So all the solutions for U_d must be of the form

$$U_d = U_1(\Lambda_1 - \sigma^2 I_d)^{1/2} R$$

where R is an arbitrary orthogonal matrix.

PPCA by Maximum Likelihood

Given N i.i.d. samples, $\{x_j\}_{j=1}^N$, estimate the PPCA model parameters μ, U_d, σ.

Model hypothesis:

$$x = \mu + U_d y + \epsilon$$

where:

- $y \sim \mathcal{N}(0, I_d)$
- $\epsilon \sim \mathcal{N}(0, \sigma^2 I_D)$
- $x \sim \mathcal{N}(\mu, U_d U_d^T + \sigma^2 I_D)$

Log-likelihood of $\{x_j\}_{j=1}^N$:

$$\mathcal{L} = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log\det(\Sigma_x) - \frac{1}{2}\sum_{j=1}^{N}(x_j - \mu)^T\Sigma_x^{-1}(x_j - \mu)$$

By the derivative of $\mathcal{L}$ with respect to μ, we could get: $\hat{\mu} = \frac{1}{N}\sum_{j=1}^N x_j$.

So:

$$\mathcal{L} = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log\det(\Sigma_x) - \frac{1}{2}trace(\Sigma_x^{-1}\Sigma_N)$$

where

$$\Sigma_N = \frac{1}{N}\sum_{j=1}^{N}(x_j - \hat{\mu})(x_j - \hat{\mu})^T$$

Furtherly, we can show:

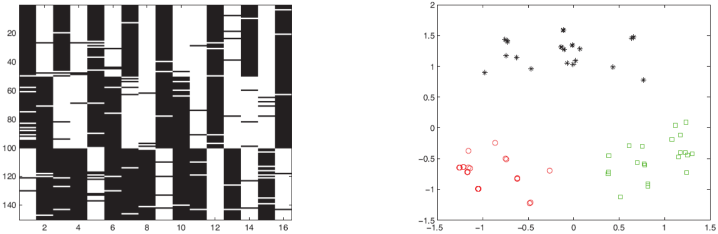
$$\hat{\sigma}^2 = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i \qquad \hat{U}_d = U_1(\Lambda_1 - \hat{\sigma}^2 I)^{1/2} R$$

where $U_1, \Lambda_1, \lambda_i$ is the corresponding eigenvectors and eigenvalues of Σ_N.

PCA for Categorical Data

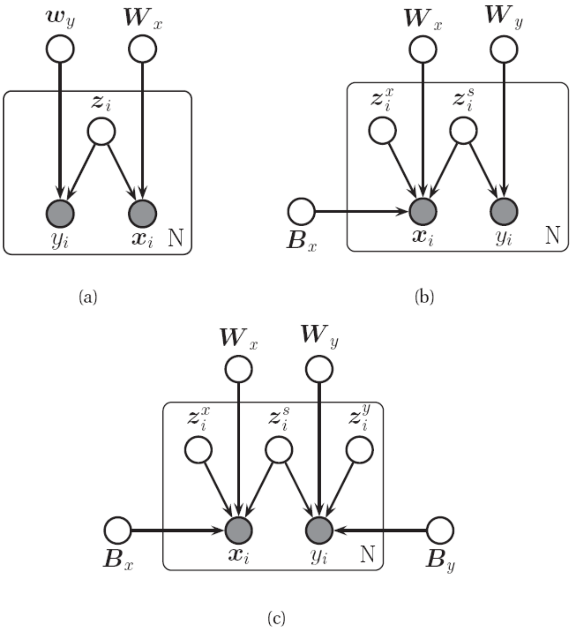
Model hypothesis:

$$
\begin{aligned}
p(\mathbf{z_j}) &= \mathcal{N}(0, \mathbf{I}) && (1)\\
p(\mathbf{y_j}|\mathbf{z_j}, \theta) &= \prod_{i=1}^{D} \mathrm{Cat}(y_{ji}|\mathrm{Softmax}(\mathbf{W^T z_j} + \mathbf{w_0})) && (2)
\end{aligned}
$$

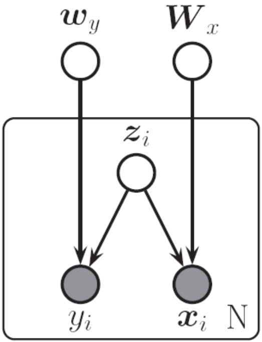


(a) (b)

PCA for paired and multi-view data

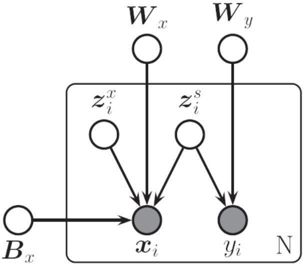


Supervised PCA



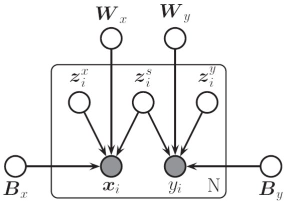
$$
\begin{aligned}
p(\mathbf{z_i}) &= \mathcal{N}(0, \mathbf{I}_L) \\
p(y_i|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_y^T \mathbf{z_i} + \mu_y, \sigma_y^2) \\
p(\mathbf{x_i}|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_x^T \mathbf{z_i} + \mu_x, \sigma_x^2)
\end{aligned}
$$

Partial Least Squares



$$
\begin{aligned}
p(\mathbf{z_i}) &= \mathcal{N}(\mathbf{z}_i^s|0, \mathbf{I}_{L_s})\mathcal{N}(\mathbf{z}_i^x|0, \mathbf{I}_{L_x}) \\
p(y_i|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_y^T \mathbf{z_i^s} + \mu_y, \sigma_y^2) \\
p(\mathbf{x_i}|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_x^T \mathbf{z_i^s} + \mathbf{B_x} \mathbf{z_i^x} + \mu_x, \sigma^2)
\end{aligned}
$$

Canonical Correlation Analysis



$$
\begin{aligned}
p(\mathbf{z_i}) &= \mathcal{N}(\mathbf{z}_i^s|0,\mathbf{I}_{L_s})\mathcal{N}(\mathbf{z}_i^x|0,\mathbf{I}_{L_x})\mathcal{N}(\mathbf{z}_i^y|0,\mathbf{I}_{L_y}) \\
p(y_i|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_y^T\mathbf{z_i^s}+\mathbf{B_y}\mathbf{z_i^y}+\mu_y,\sigma_y^2) \\
p(\mathbf{x_i}|\mathbf{z_i}) &= \mathcal{N}(\mathbf{w}_x^T\mathbf{z_i^s}+\mathbf{B_x}\mathbf{z_i^x}+\mu_x,\sigma^2)
\end{aligned}
$$

Compare between:

- Ridge regression
- Principal components regression
- Partial least squares

- **Ridge regression:**

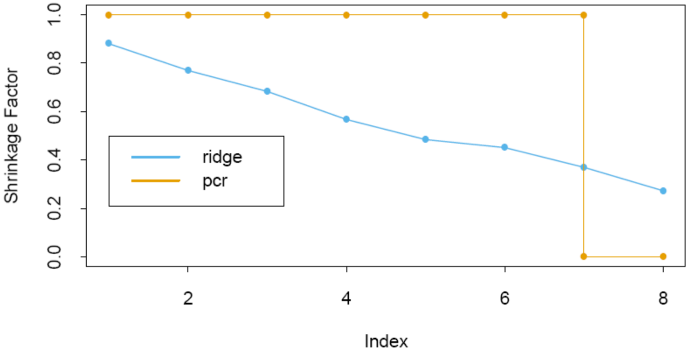
$$\hat{\beta}_{\text{ridge}} = (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T y}$$

Replace $\mathbf{X}$ with its SVD $\mathbf{X} = \mathbf{UDV^T}$:

$$\begin{aligned} \mathbf{X}\hat{\beta}_{\text{ls}} &= \mathbf{X(X^T X)^{-1} X^T y} \\ &= \mathbf{UU^T y} \end{aligned}$$

$$\begin{aligned} \mathbf{X}\hat{\beta}_{\text{ridge}} &= \mathbf{X}(\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X^T y} \\ &= \mathbf{UD}(\mathbf{D^2} + \lambda \mathbf{I})^{-1} \mathbf{DU^T y} \\ &= \sum_{j=1}^{D} \mathbf{u_j} \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u_j^T y} \end{aligned}$$

- **Principal components regression:**



- **Partial least squares:**

  The mth PLS direction $\hat{\psi}_m$ solves:

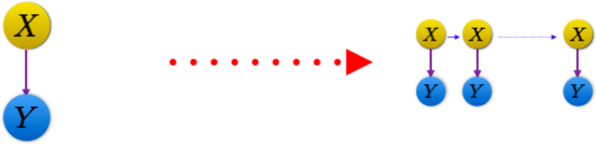
$$
\begin{aligned}
\max_{\alpha} \quad & \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha) \\
\text{s.t.} \quad & ||\alpha|| = 1, \alpha^T \Sigma \hat{\psi}_l = 0, l = 1, \cdots, m-1
\end{aligned}
$$

Factor Analysis(FA)

basic component

Sequential Model

latent variables(hidden states)



X	Y		
discrete	discrete	Mixture of Models	HMM
discrete	continuous		
continuous	continuous	Factor Analysis	State Space Model

Model Hypothesis:

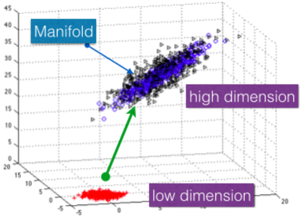
$$\mathbf{Y} = \mu + \mathbf{W}\mathbf{X} + \epsilon$$

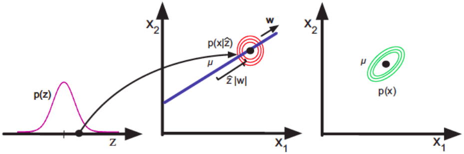
- $\mathbf{X} \perp \epsilon$
- $\mathbf{X} \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $\epsilon \sim \Psi$(diagonal)
- $\mathbf{W}$: factor loading matrix

Inference:

$$Y|X \sim \mathcal{N}(\mu + W\mu_0, \Psi)$$

- $P(Y)$? (Probability of observed variable)
- $P(X|Y)$? (Probability of latent variable)
- $P(X, Y)$





Joint Gaussian

Joint, Marginal, Conditional distributions are all Gaussian.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

- Marginal:

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

- Conditional:

$$X_1 | X_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

$$\mathbf{Y} = \mu + \mathbf{W}\mathbf{X} + \epsilon$$

- $\mathbf{X} \perp \epsilon$
- $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- $\epsilon \sim \Psi$(diagonal)
- $\mathbf{W}$: factor loading matrix

Joint Distribution:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & W^T \\ W & WW^T + \Psi \end{bmatrix} \right)$$

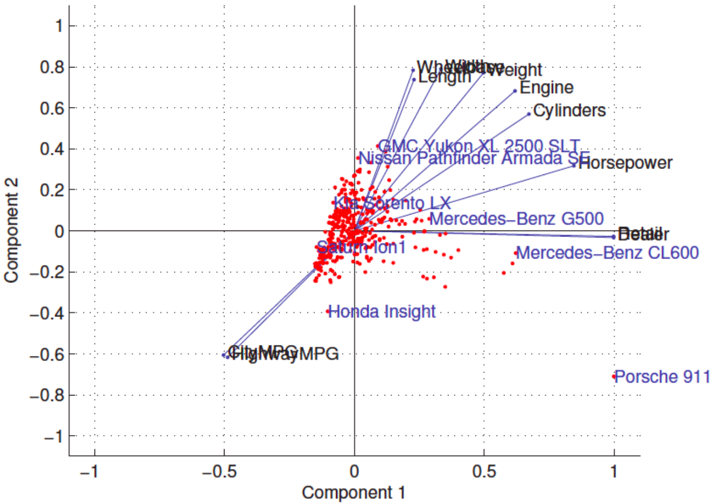
$$X|Y \sim \mathcal{N}(W^T(WW^T + \Psi)^{-1}(Y - \mu), I - W^T(WW^T + \Psi)^{-1}W)$$

Define:

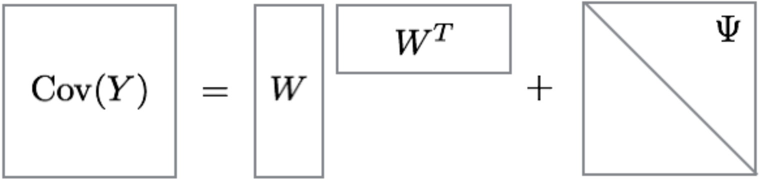
$$\mathsf{Var}_{X|Y} = I - W^T(WW^T + \Psi)^{-1}W = (I + W^T\Psi^{-1}W)^{-1}$$

$$X|Y \sim \mathcal{N}(\mathsf{Var}_{X|Y} W^T\Psi^{-1}(Y - \mu), \mathsf{Var}_{X|Y})$$

rotation=none



Constrainted Covariance Gaussiam



Number of parameters: from $O(D^2)$ to $O(LD + D)$

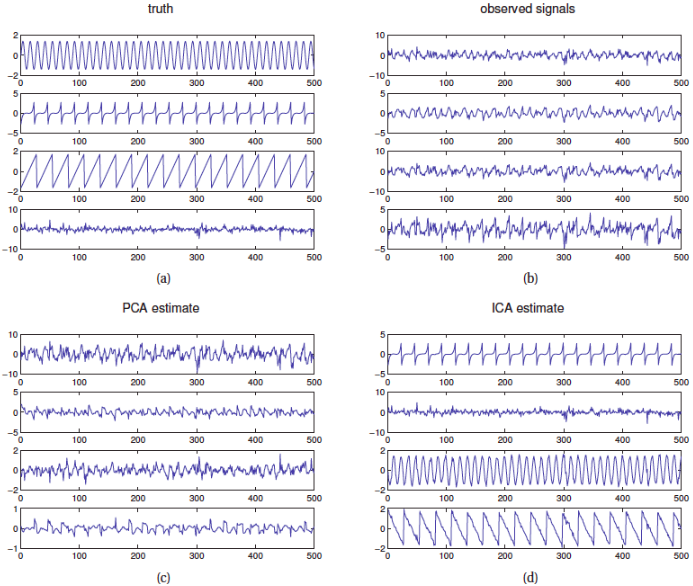
Model Invariance

- Consider any learning algorithms, what they did is to estimate parameters which maximize the liklihood of observed sequences Y.
- Then W always appears in the form of WW^T.
- For any loading matrix W and any orthonormal matrix R,

$$(WR)(WR)^T = WW^T$$

- So which part of model hypothesis makes this happen?

Independent Component Analysis



Suppose zero-mean,

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \epsilon$$

- In this context, $\mathbf{W}$ is called **mixing matrix**.
- Where is the difference?

Suppose zero-mean,

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \epsilon$$

- In this context, $\mathbf{W}$ is called **mixing matrix**.
- Where is the difference?

  Suppose any non-Gaussian distribution for $\mathbf{X}$

$$p(\mathbf{x}_j) = \prod_{i=1}^{L} p(x_{ji})$$

- Why non-Gaussian? Why independent components?

Suppose zero-mean,

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \epsilon$$

- In this context, $\mathbf{W}$ is called **mixing matrix**.
- Where is the difference?
  Suppose any non-Gaussian distribution for $\mathbf{X}$

$$p(\mathbf{x}_j) = \prod_{i=1}^{L} p(x_{ji})$$

- Why non-Gaussian? Why independent components?
  Compare with PCA, we assume x_i to be statistically independent rather than uncorrelated.

Modeling the source densities:

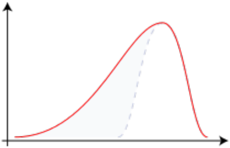
$$\mu_k = \mathbb{E}[(X - \mathbb{E}X)^k]$$

- Skewness:

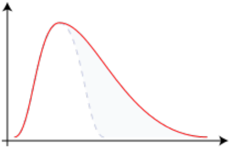
$$\mathsf{skew}(x) = \frac{\mu_3}{\sigma^3}$$

- Kurtosis:

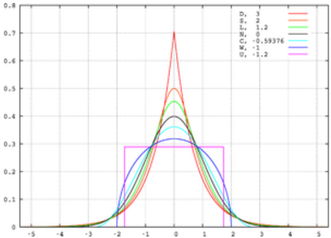
$$\mathsf{kurt}(x) = \frac{\mu_4}{\sigma^4} - 3$$



Negative Skew



Positive Skew



- **Super-Gaussian:** big spike and heavy tails (kurt > 0) e.g. Laplace
  - natural signal (passed through certain linear filters)
- **Sub-Gaussian:** much flatter (kurt < 0)
- **Skewed:** be asymmetric