Model selection for FA/PCA

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

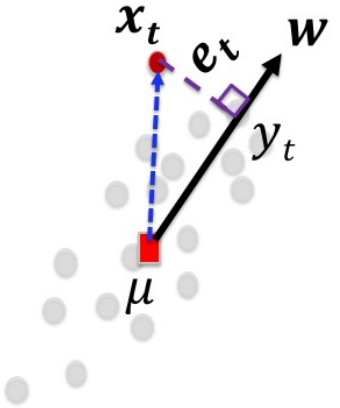
2021-04-25

Outline

- **Recall**
  - **Probabilistic PCA, Factor Analysis (FA)**
- Model selection for PCA/FA
- Practice on experimental comparisons

PCA by minimizing MSE

$$J(w) = \frac{1}{N}\sum_{t=1}^{N} ||\boldsymbol{x_t} - (\boldsymbol{x_t^T w})\boldsymbol{w}||^2$$



$$\boldsymbol{x_t^T x_t} - (\boldsymbol{x_t^T w})\boldsymbol{w^T x_t} - \boldsymbol{x_t^T}(\boldsymbol{x_t^T w})\boldsymbol{w} + (\boldsymbol{x_t^T w})\boldsymbol{w^T}(\boldsymbol{x_t^T w})\boldsymbol{w}$$
$$= \boldsymbol{x_t^T x_t} - \boldsymbol{w^T}(\boldsymbol{x_t x_t^T})\boldsymbol{w}$$

$$||w|| = 1$$

Introduce a Lagrange multiplier λ

$$L(\{\boldsymbol{x_t}\}, \boldsymbol{w}) = J(\{\boldsymbol{x_t}\}, \boldsymbol{w}) - \lambda \cdot (\boldsymbol{w^T w} - 1)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} - \lambda \cdot \frac{\partial(\boldsymbol{w^T w} - 1)}{\partial \boldsymbol{w}} = -2(\Sigma_x \boldsymbol{w}) - \lambda \cdot 2\boldsymbol{w} = \boldsymbol{0}$$

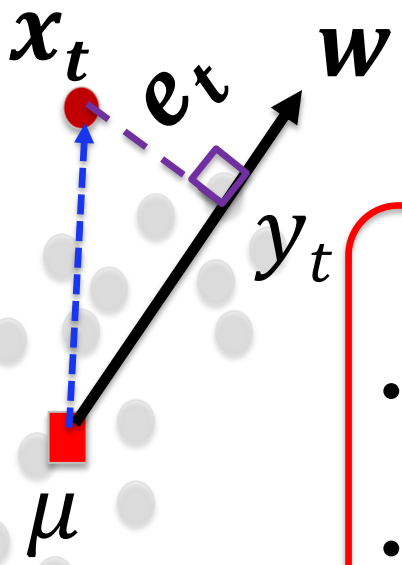
$$\Sigma_x = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x_t x_t^T}$$

$$\boxed{\Sigma_x \boldsymbol{w} = (-\lambda) \cdot \boldsymbol{w}}$$

Eigenvalues and Eigenvectors

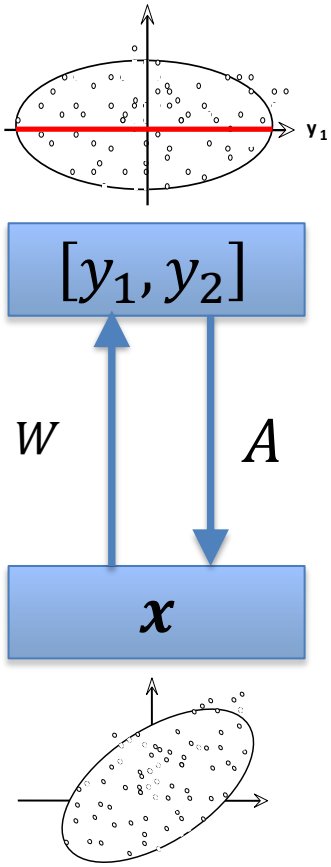
Factor Analysis (FA) model

Continuous latent variable $\boldsymbol{y}$



For the *t*-th data point:

- Randomly sample a y_t:
  $\boldsymbol{y}_t \sim G(\boldsymbol{y}|\boldsymbol{0}, \Sigma_y)$;
- Randomly generate a noise e_t
  $\boldsymbol{e}_t \sim G(\boldsymbol{e}|0, \sigma^2 I)$
- Generate $\boldsymbol{x}_t$ by:
  $\boldsymbol{x}_t = A\boldsymbol{y}_t + \boldsymbol{\mu} + \boldsymbol{e}_t$



$$||\boldsymbol{w}|| = 1$$

$$y_t = \boldsymbol{x}_t^T \boldsymbol{w}$$

$$\boldsymbol{e}_t = ||\boldsymbol{x}_t - y_t \boldsymbol{w}||^2$$

EM algorithm for FA

E-Step: $p^{old}(\boldsymbol{y}|\boldsymbol{x}) = \dfrac{G(\boldsymbol{y}|0,I)G(\boldsymbol{x}|A\boldsymbol{y}+\boldsymbol{\mu},\sigma^2 I)}{G(\boldsymbol{x}|\boldsymbol{\mu},AA^T+\sigma^2 I)}$

$$E[\boldsymbol{y}|\boldsymbol{x}] = W\boldsymbol{x} \qquad W = A^T(AA^T+\sigma^2 I)^{-1}$$

$$E[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{x}] = I - WA + W\boldsymbol{x}\boldsymbol{x}^T W^T$$

M-Step: max $Q\left(p^{old}(\boldsymbol{y}|\boldsymbol{x}),\Theta\right)$

$$Q = \int p^{old}(\boldsymbol{y}|\boldsymbol{x}) \cdot \ln[G(\boldsymbol{y}|0,I)G(\boldsymbol{x}|A\boldsymbol{y}+\boldsymbol{\mu},\sigma^2 I)]\, d\boldsymbol{y}$$

$$A^{new} = \left(\sum_{t=1}^{N} \boldsymbol{x}_t (E[\boldsymbol{y}|\boldsymbol{x}_t])^T\right)\left(\sum_{t=1}^{N} E[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{x}_t]\right)^{-1}$$

$$\sigma^{2^{new}} = \frac{1}{Nd} Tr\left\{\sum_{t=1}^{N}\{\boldsymbol{x}_t\boldsymbol{x}_t^T - A^{new}E[\boldsymbol{y}|\boldsymbol{x}_t]\boldsymbol{x}_t^T\}\right\}$$

Maximum likelihood FA implements PCA

$$p(\boldsymbol{y}) = G(\boldsymbol{y}|\boldsymbol{0}, I), \qquad p(\boldsymbol{x}|\boldsymbol{y}) = G(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\mu}, \Sigma_e),$$

$$p(\boldsymbol{x}|\Theta) = \int p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y} = G\left(\boldsymbol{x}|\boldsymbol{\mu}, \mathrm{AA}^{\mathrm{T}} + \Sigma_e\right),$$

$$\max_{\Theta} \log\left\{\prod_{t=1}^{N=1} p(\boldsymbol{x_t}|\Theta)\right\}$$

Maximum Likelihood

$$\boldsymbol{\Sigma}_e = \sigma_e^2 \mathbf{I}_n$$

assume $\boldsymbol{\mu} = \mathbf{0}$

PCA

$$\begin{cases} \hat{\mathbf{A}}_{n\times m}^{ML} = \mathbf{U}_{n\times m}(\mathbf{D}_m - \hat{\sigma}_e^2)^{\frac{1}{2}}\mathbf{R}^T, \quad \mathbf{D}_m = diag[s_1, \ldots, s_m] \\ \hat{\sigma}_e^{2,ML} = \frac{1}{n-m}\sum_{i=m+1}^{n} s_i, \end{cases}$$

U is eigenvectors of sample cov.

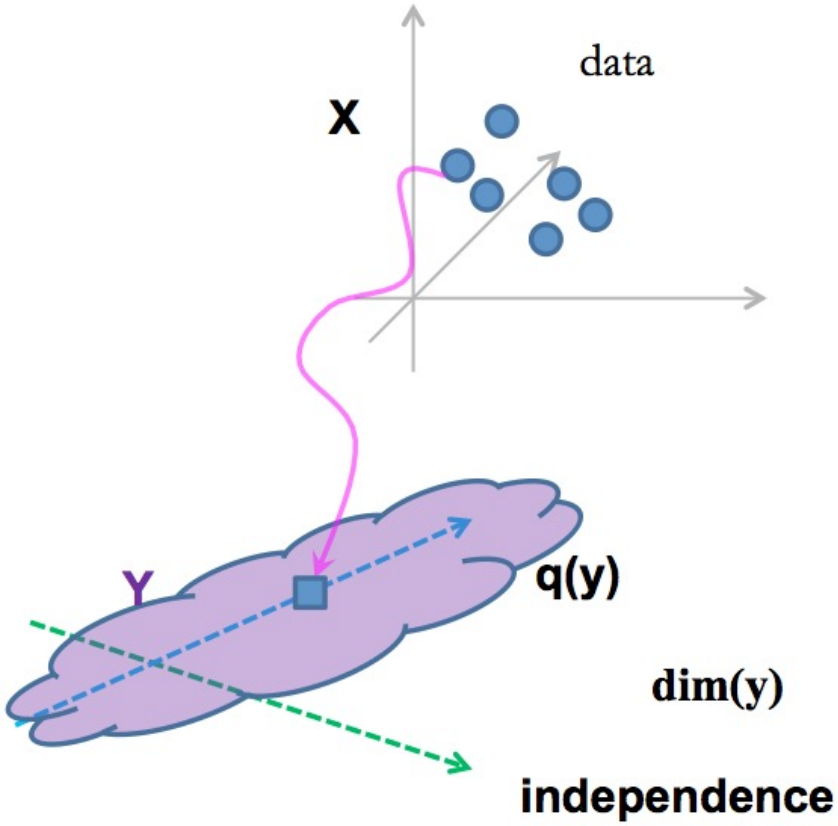
Outline

- Recall
  - Probabilistic PCA, Factor Analysis (FA)

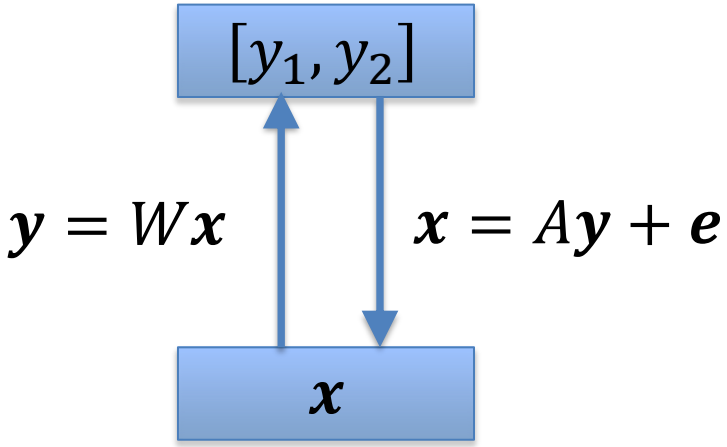
- **Model selection for PCA/FA**

- Practice on experimental comparisons

Dimensionality reduction



$m = \dim(\boldsymbol{y}) = ?$



$n = \dim(\boldsymbol{x}) > m$

Determining *m* is a model selection problem.

Choosing m by keeping 99% variance

$m = \dim(\boldsymbol{y}) = ?$

- Pick m to be the smallest value so that 99% of variance is retained, i.e.,

Average squared projection error

$$\frac{\frac{1}{N}\sum_{t=1}^{N} ||x_t - \hat{x}_t||^2}{\frac{1}{N}\sum_{t=1}^{N} ||x_t||^2} \leq 0.01 \qquad (1\%)$$

Total variation in the data

Where $\hat{x}_t$ is a reconstruction of x_t by

$$\hat{x}_t = W y_t = W(W^T x_t) \qquad \hat{x}_t = y_t w = (x_t^T w) w$$

Choosing m by keeping 99% variance

$m = \dim(\boldsymbol{y}) = ?$

- Eigen-decomposition

$$\Sigma_x \approx U \Lambda U^T$$

$$U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{\mathrm{m}}]^{\mathrm{T}}$$

$$\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_m]$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$$

Covariance matrix (assume $\boldsymbol{x}_t$ zero mean):

$$\Sigma_x = \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^T = \frac{1}{N} X X^T$$

$$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]_{n \times N}$$

- Equivalently, we pick

Sum of selected eigenvalues

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \leq 0.01 \qquad (1\%)$$

Sum of all eigenvalues

Question

- Why are they equivalent?

$$\frac{\frac{1}{N}\sum_{t=1}^{N}||x_t - \hat{x}_t||^2}{\frac{1}{N}\sum_{t=1}^{N}||x_t||^2} \leq 0.01 \quad \Longleftrightarrow \quad \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{j=1}^{n}\lambda_j} \leq 0.01$$

The effects of dim(y)

$$\boldsymbol{x} = \boldsymbol{a}y_1 + \boldsymbol{a}_2 y_2 + \cdots + \boldsymbol{a}_{m^*} y_{m^*} + \boldsymbol{e}$$

Assume $\boldsymbol{\mu} = \mathbf{0}$

$$\boldsymbol{x} = A\boldsymbol{y} + \boldsymbol{e}$$

$$\boldsymbol{y} = [y_1, \cdots, y_{m^*}]^T_{m^* \times 1}$$

$$p(\boldsymbol{y}) = G(\boldsymbol{y}|\mathbf{0}, I),$$

Under-fitting (big bias): $m < m^*$

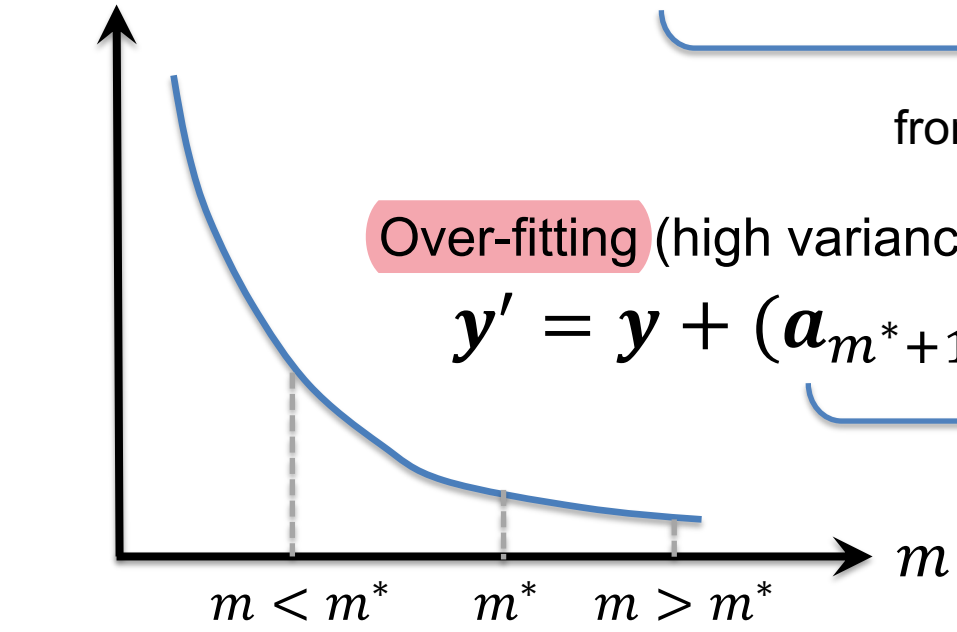
$$\boldsymbol{e}' = \boldsymbol{a}_{m+1} y_{m+1} + \cdots + \boldsymbol{a}_{m^*} y_{m^*} + \boldsymbol{e}$$

from signal $\boldsymbol{y}$

Over-fitting (high variance): $m > m^*$

$$\boldsymbol{y}' = \boldsymbol{y} + (\boldsymbol{a}_{m^*+1} y_{m^*+1} + \cdots + \boldsymbol{a}_m y_m)$$

from noise $\boldsymbol{e}$



Bias-variance decomposition

We want to find a function $\hat{f}(y)$, that approximates the true function $f(y)$ as well as possible:

$$x = f(y) + \epsilon$$

$$E[\epsilon] = 0$$
$$E[x] = f$$
$$Var[x] = Var[\epsilon]$$

The expected error

$$E\left[\left(x - \hat{f}(y)\right)^2\right] = E\left[x^2 + \hat{f}^2 - 2x\hat{f}\right]$$

$$= Var[x] + (E[x])^2 + Var\left[\hat{f}\right] + \left(E\left[\hat{f}\right]\right)^2 - 2fE[\hat{f}]$$

$$= Var[x] + Var\left[\hat{f}\right] + \left(f - E\left[\hat{f}\right]\right)^2$$

$$= Var[\epsilon] + Var\left[\hat{f}\right] + Bias\left[\hat{f}\right]^2$$

Model selection for FA

Probabilistic model

$$p(X_N \mid \Theta_K)$$

Candidate models:

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K \subseteq \cdots$$

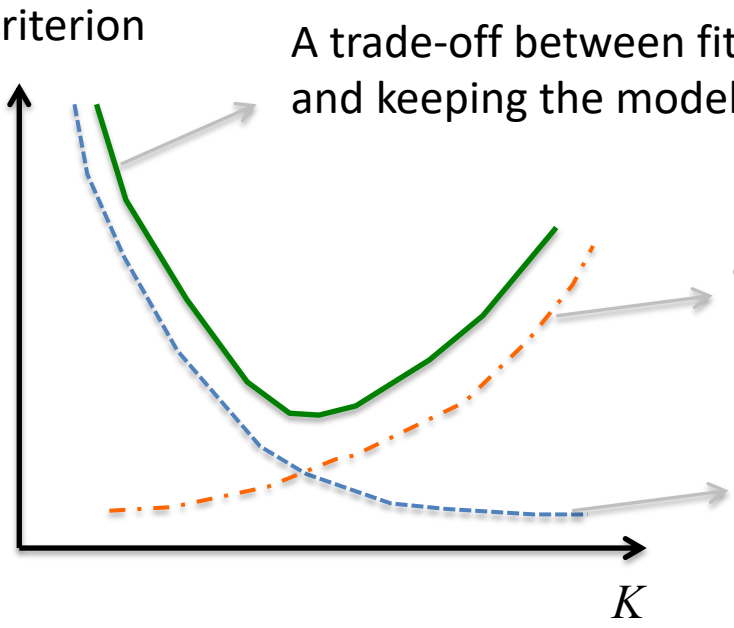
Factor Analysis

$$p(\boldsymbol{y}) = G(\boldsymbol{y}|\boldsymbol{0}, I),$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = G(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\mu}, \Sigma_e),$$

$$p(\boldsymbol{x}|\Theta) = \int p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y} = G\left(\boldsymbol{x}\middle|\boldsymbol{\mu}, \mathrm{A}\mathrm{A}^{\mathrm{T}} + \Sigma_e\right),$$

$$K = m = \dim(\boldsymbol{y})$$



A trade-off between fitting the data well and keeping the model simple

Models become more and more complex as K increases.

Negative log-likelihood (or fitting error) to make sure the model fit the data well.

Akaike's Information Criterion (AIC) $\ln p(X_N \mid \hat{\Theta}_K) - d_k$

Bayesian Information Criterion (BIC) $\ln p(X_N \mid \hat{\Theta}_K) - \frac{1}{2} d_k \ln N$

d_k: number of free parameters

N: sample size

Two-stage model selection

Given a sample set $\mathcal{X}_N = \{\mathbf{x}_t\}_{t=1}^N$, the task of FA modeling consists of parameter learning and model selection, i.e., selecting the number of factors m. Given m, the parameter set $\mathbf{\Theta}_m$ is usually estimated by Maximum Likelihood (ML) learning. Including it as a subtask, model selection problem is usually tackled by the following classical two-stage procedure:

- **Stage I:** Compute $\hat{\mathbf{\Theta}}_m = \hat{\mathbf{\Theta}}(\mathcal{X}_N, m)$ for each $m \in [m_{low}, m_{up}]$ with m_{low} and m_{up} given. Normally, $\hat{\mathbf{\Theta}}_m$ is the ML estimator,

$$\begin{aligned}\hat{\mathbf{\Theta}}_m^{ML} &= \arg\max_{\Theta_m} \ln p(\mathcal{X}_N|\mathbf{\Theta}_m) \\ &= \arg\min_{\Theta_m} \mathcal{E}_L(\mathcal{X}_N|\mathbf{\Theta}_m), \end{aligned} \tag{2}$$

where $\mathcal{E}_L(\mathcal{X}_N|\mathbf{\Theta}_m) = -\frac{2}{N}\ln p(\mathcal{X}_N|\mathbf{\Theta}_m)$ is denoted as **NLL** (negative log-likelihood).

- **Stage II:** Estimate $\hat{m} = \arg\min_m \mathcal{E}_{Cri}$, where $\mathcal{E}_{Cri}$ is formulated according to an information criterion (Cri), e.g.,

$$\mathcal{E}_{Cri}(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m) = \mathcal{E}_L(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m) + \frac{\rho_N d_m}{N} \tag{3}$$

$$d_m = nm + 1 - \frac{m(m-1)}{2} \tag{4}$$

$$\rho_N = \begin{cases} 0; & \text{for } \textbf{NLL} \\ 2; & \text{for } \textbf{AIC} \\ \ln N; & \text{for } \textbf{BIC/MDL} \\ \ln N + 1; & \text{for } \textbf{CAIC} \\ 2\ln(\ln N); & \text{for } \textbf{HQC} \end{cases} \tag{5}$$

where d_m by eq.(4) is the number of free parameters of FA model over the real number field, and $d_m = m(2n - m) + 1$ when considering complex number field. In the

Variational FA

$$\mathbf{\Theta}_m = \{\mathbf{A}, \sigma_e^2\}$$

$$a_k^\alpha = b_k^\alpha = a^\phi = b^\phi = 10^{-3}$$

$$\mu = \mathbf{0}, \mathbf{\Sigma}_e = \sigma_e^2 \mathbf{I}_n$$

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu} + \mathbf{e},$$

$$p(\mathbf{x}|\mathbf{y}) = G(\mathbf{x}|\mathbf{A}\mathbf{y} + \boldsymbol{\mu}, \mathbf{\Sigma}_e),$$

$$p(\mathbf{y}) = G(\mathbf{y}|\mathbf{0}, \mathbf{\Sigma}_y)$$

$$p(\mathbf{A}|\boldsymbol{\alpha}) = \prod_{k=1}^m G(\mathbf{a}_k|\mathbf{0}, \tfrac{1}{\alpha_k^\alpha}\mathbf{I}_n)$$

$$p(\boldsymbol{\alpha}|\boldsymbol{a}^\alpha, \boldsymbol{b}^\alpha) = \prod_{k=1}^m \Gamma(\alpha_k|a_k^\alpha, b_k^\alpha)$$

$$p(\varphi) = \Gamma(\varphi|a^\varphi, b^\varphi), \varphi = (\sigma_e^2)^{-1}$$

$$\mathcal{F}(q_Y, q_\mathbf{A}, q_{\boldsymbol{\alpha}}, q_\varphi) = \int q_Y q_\mathbf{A} q_{\boldsymbol{\alpha}} q_\varphi \ln \frac{p(\mathcal{X}_N, Y, \mathbf{\Theta}|m)}{q_Y q_\mathbf{A} q_{\boldsymbol{\alpha}} q_\varphi} dY d\mathbf{A} d\boldsymbol{\alpha} d\varphi dY$$

Stage I: get $\mathcal{F}(m)$, $m \in [m_{low}, m_{up}]$;

Stage II: estimate $\hat{m} = \arg\max_m \mathcal{F}(m)$.

VBEM for FA

Initialization: get $\{q_{\mathbf{A}}, q_{\boldsymbol{\alpha}}, q_{\varphi}\}$ by random initialization.

E-Step: get q_Y based on $q_\theta = q_{\mathbf{A}} q_{\boldsymbol{\alpha}} q_{\varphi}$:

$$q_Y = \prod_t G(\mathbf{y}_t | \Sigma_y \langle \mathbf{A}^T \varphi \rangle_{q_\theta} \mathbf{x}_t, \Sigma_y),$$

where $\Sigma_y = (\mathbf{I} + \langle \mathbf{A}^T \varphi \mathbf{A} \rangle_{q_\theta})^{-1}$, $\langle \bullet \rangle_{q_\theta} = E_{q_\theta}[\bullet]$.

M-Step: get q_{θ_i} based q_{θ_ℓ} and q_Y, where

$$q_{\theta_i}, q_{\theta_\ell} \in \{q_{\mathbf{A}}, q_{\boldsymbol{\alpha}}, q_{\varphi}\}, \ell \neq i :$$

$$q_{\mathbf{A}} = \prod_{j=1}^{n} G(\mathbf{a}_j | \mu_{a,j}, \Sigma_{a,j}),$$

$$\Sigma_{a,j} = \left(\langle \varphi \rangle_{q_\varphi} \sum_t \langle \mathbf{y}_t \mathbf{y}_t^T \rangle_{q_Y} + diag(\langle \boldsymbol{\alpha} \rangle_{q_\alpha}) \right)^{-1},$$

$$\mu_{a,j} = \Sigma_{a,j} \langle \varphi_j \rangle_{q_\varphi} \left(\sum_t \mathbf{x}_{jt} \langle \mathbf{y}_t \rangle_{q_Y} \right),$$

$$q_{\boldsymbol{\alpha}} = \prod_{k=1}^{m} \Gamma \left(\alpha_k | a_k^\alpha + \frac{n}{2}, b_k^\alpha + \frac{\langle \mathbf{a}_k \mathbf{a}_k^T \rangle_{q_{\mathbf{A}}}}{2} \right),$$

$$q_{\varphi} = \Gamma \left(\varphi | a^\varphi + \frac{N}{2}, b^\varphi + \frac{\sum_t \langle (\mathbf{x}_{jt} - \mathbf{a}_j^T \mathbf{y}_t)^2 \rangle_{q_A q_Y}}{2} \right).$$

Outline

- Recall
  - Probabilistic PCA, Factor Analysis (FA)
- Model selection for PCA/FA
- **Practice on experimental comparisons**

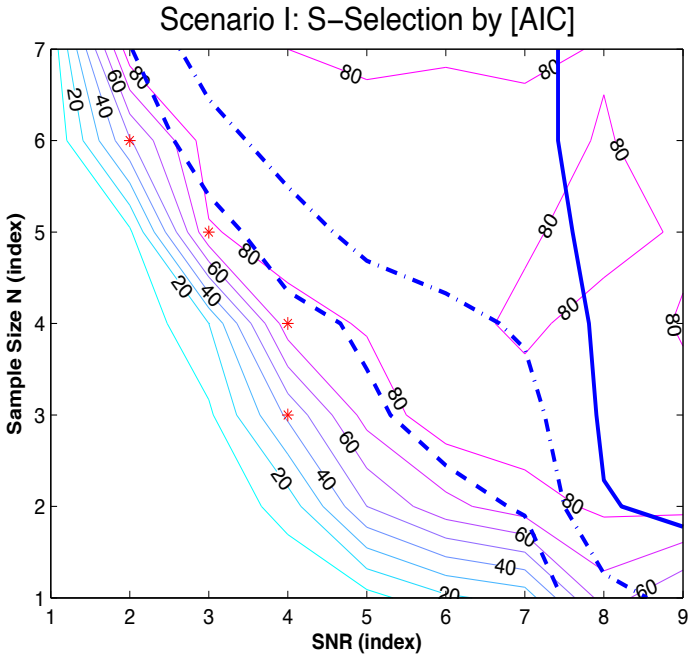
An experimental study

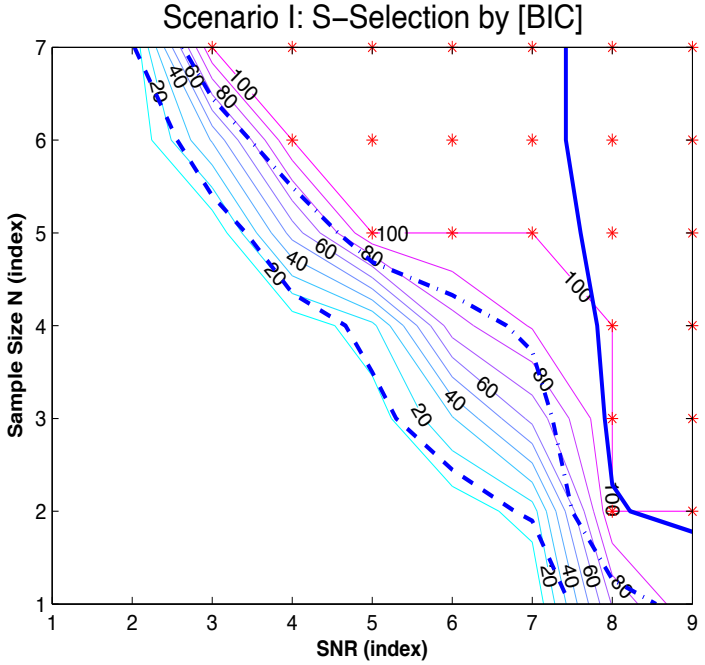
Experimental settings

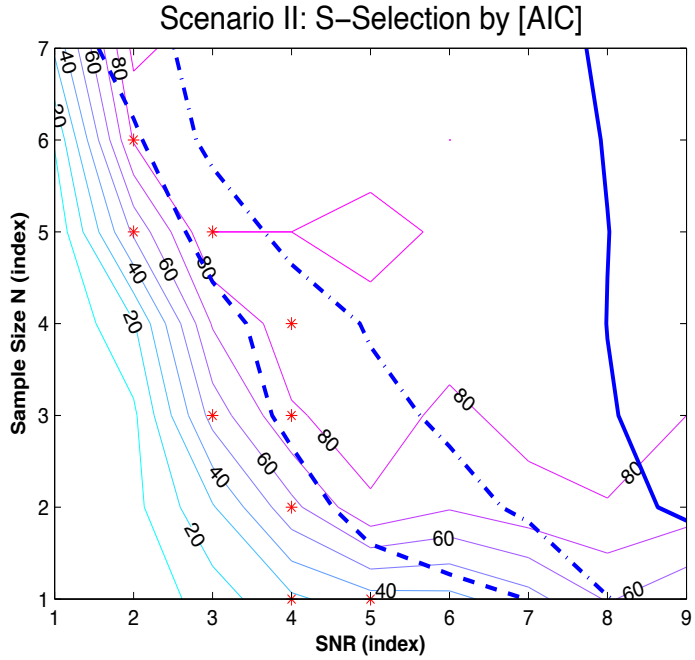
id	features f	values $V(f)$
f_1	sample size N	$\{25, 50, 75, 100, 200, 400, 800\}$
f_2	SNR $\gamma_o = \lambda_{m^*}/\sigma_e^2 + 1$	$\{1.2, 1.5, 2, 2.5, 3,$ $3.5, 4, 8, 16\}$
f_3	$\lambda_i \sim \mathcal{U}([\lambda_{min}, \lambda_{max}])$	$[1, 1]$, $[1, 10]$, $[1, 50]$
f_4	$\{n, m\} = \{\dim(x), \dim(y)\}$	$\{15, 5\}$

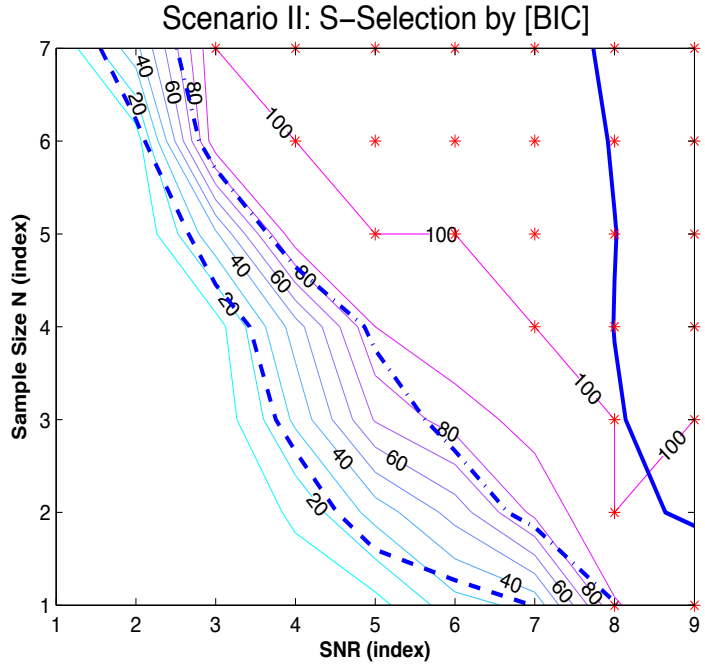
Each setting: $\Omega(f_1, f_2, f_3, f_4), \quad f_4 : \{n = 15, m = 5\}.$

scenarios (τ)	*criteria:* {AIC,BIC, . . .}	*categories*(c)
I $(f_3 : \mathcal{U}([1, 1]))$	$\forall$*criterion*,	**U:** $\hat{m} < m^*$
II $(f_3 : \mathcal{U}([1, 10]))$	$\forall (f_1, f_2) \in V(f_1) \times V(f_2)$	**S:** $\hat{m} = m^*$
III $(f_3 : \mathcal{U}([1, 50]))$	$R_{c,\tau}^{(cri)}(f_1, f_2)$	**O:** $\hat{m} > m^*$


Scenario I: S−Selection by [AIC]
Sample Size N (index)
SNR (index)


Scenario I: S−Selection by [BIC]
Sample Size N (index)
SNR (index)


Scenario II: S−Selection by [AIC]
Sample Size N (index)
SNR (index)


Scenario II: S−Selection by [BIC]
Sample Size N (index)
SNR (index)

Thank you!