Dimension reduction : Bayesian Modeling and Inference

Kaniav Kamary

PSPM-Institut Camille Jordan

INSA, Lyon

June 19, 2026

Outline

Machine Learning?

- Machine learning is fundamentally about learning patterns from data.

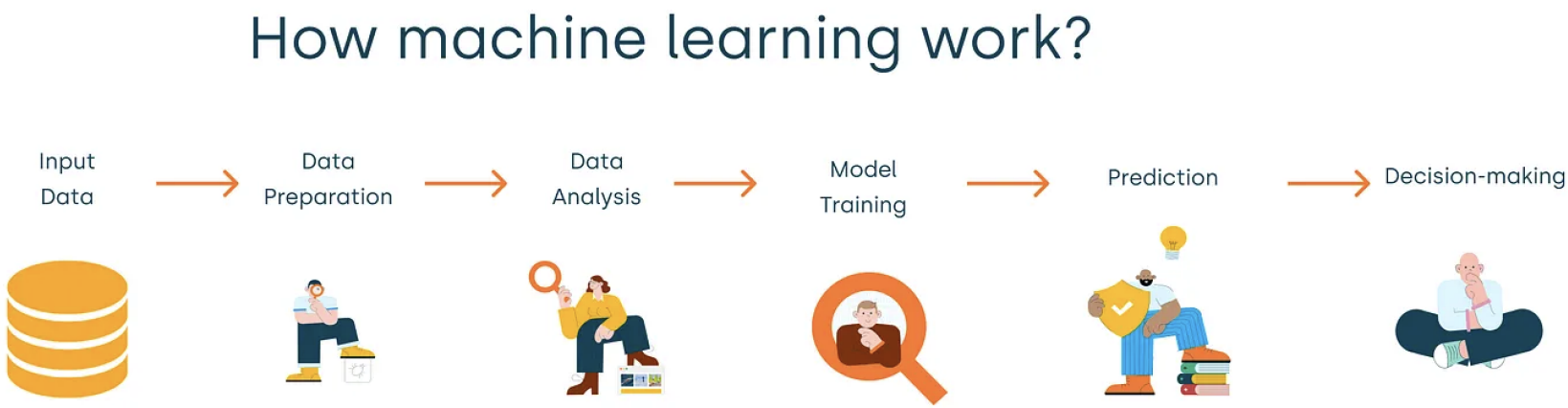


Image source

At its core, machine learning is concerned with inferring latent structures from observed data.

Machine Learning?

- Input / output data : image, text, speech and tabular

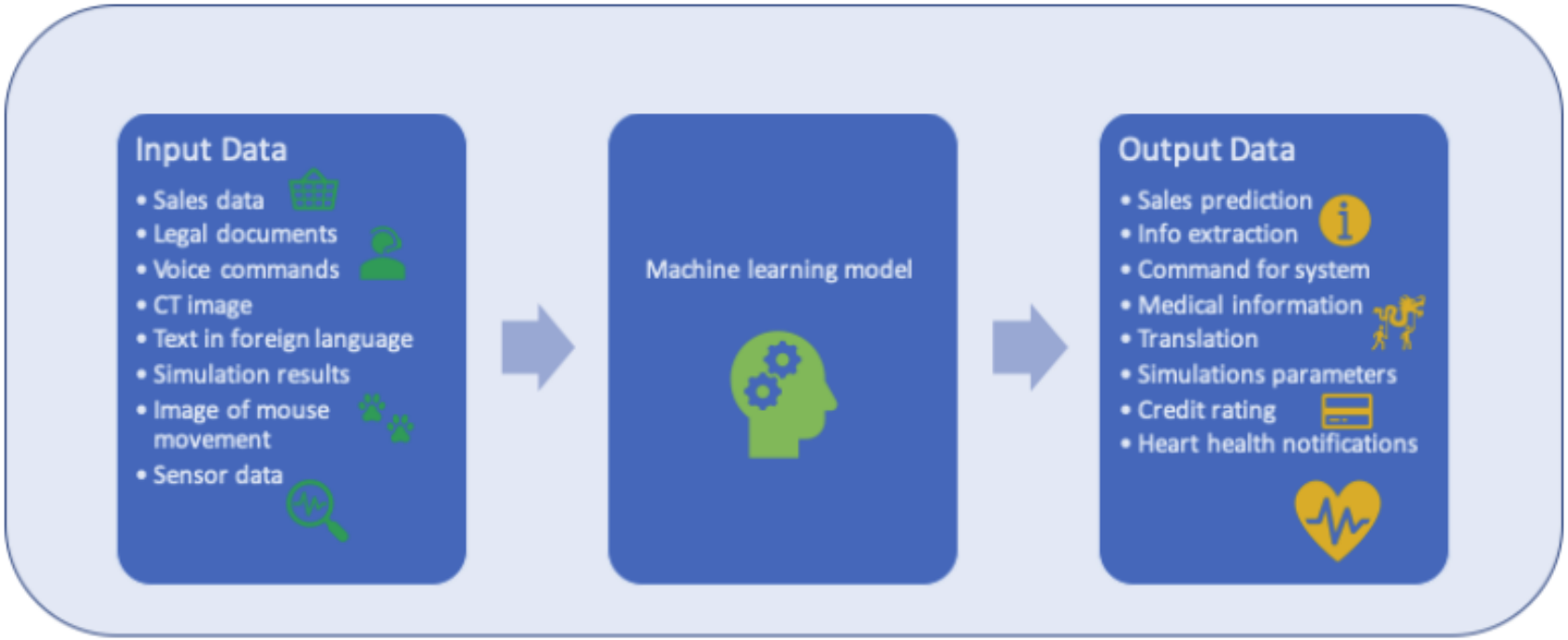


Image source

Fundamental question :

How can we model the mechanisms that generate data, and recover the hidden structures responsible for the observations?

Machine Learning?

Machine learning types by data content

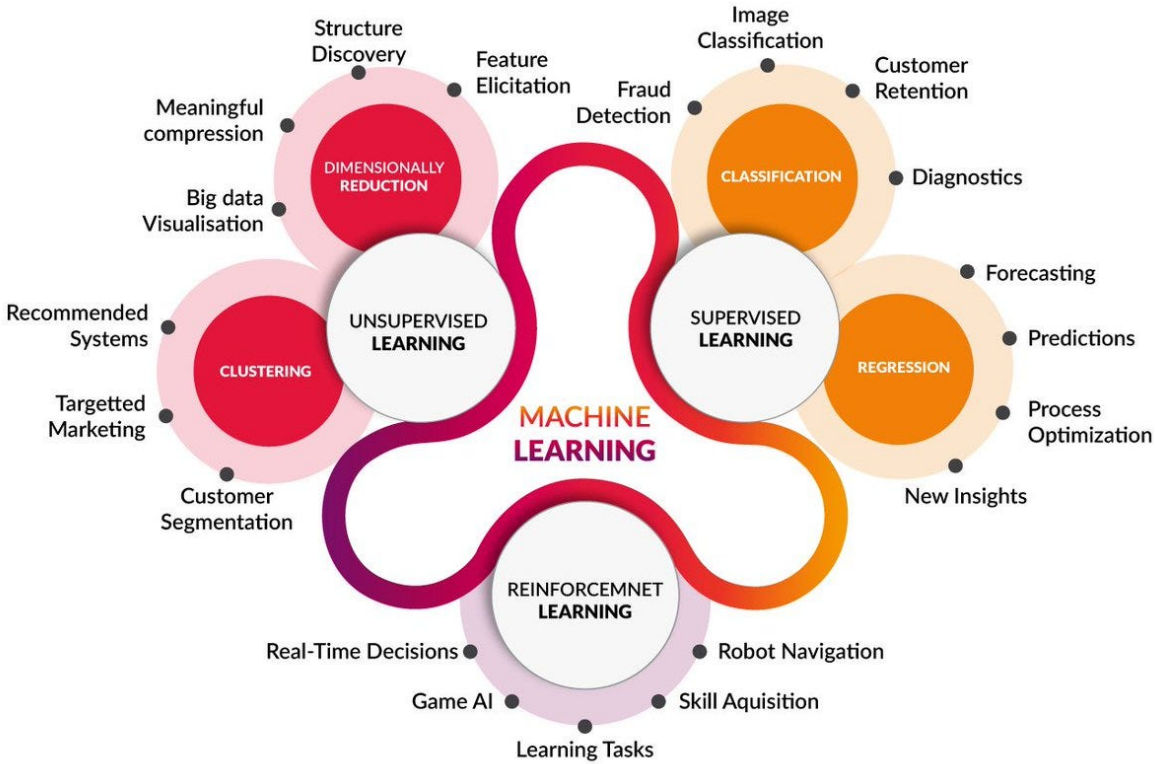


Image source

Supervised learning
- needs labeled data

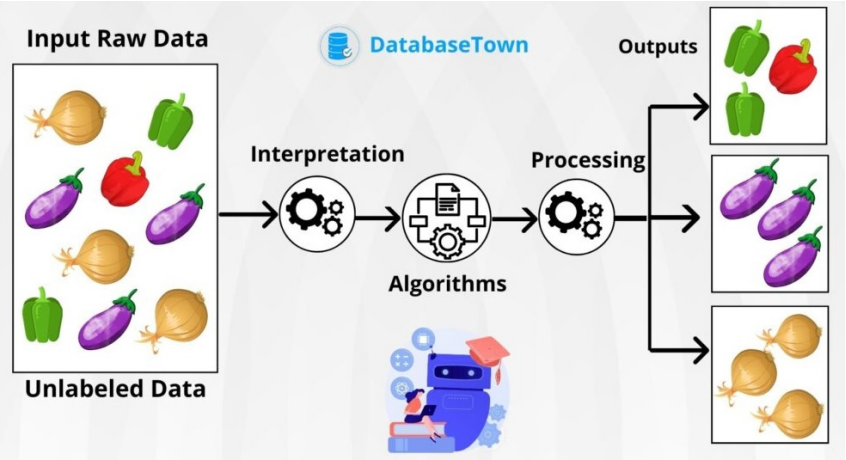
Unsupervised learning
- finds groups with similarities

Reinforcement learning
- finds a policy to achieve a goal

Machine Learning?

Unsupervised Machine learning

Clustering



Dimensionality reduction

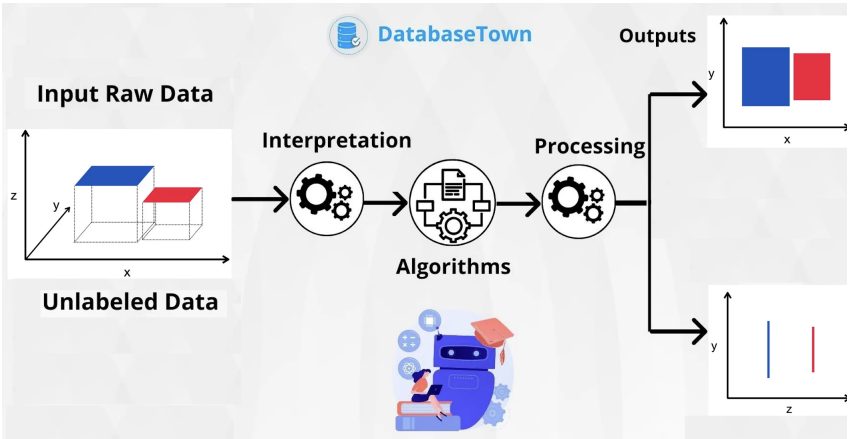


Image source

Why Bayesian Modeling in Machine Learning?

- Classical approaches : finding a single "best" estimate of unknown quantities.
- Bayesian methods : unknown quantities as random variables and quantify uncertainty about them.

Modern Bayesian machine learning builds probabilistic models of data and uses inference algorithms to learn hidden structures while quantifying uncertainty.

Principal Component Analysis

Classical PCA :

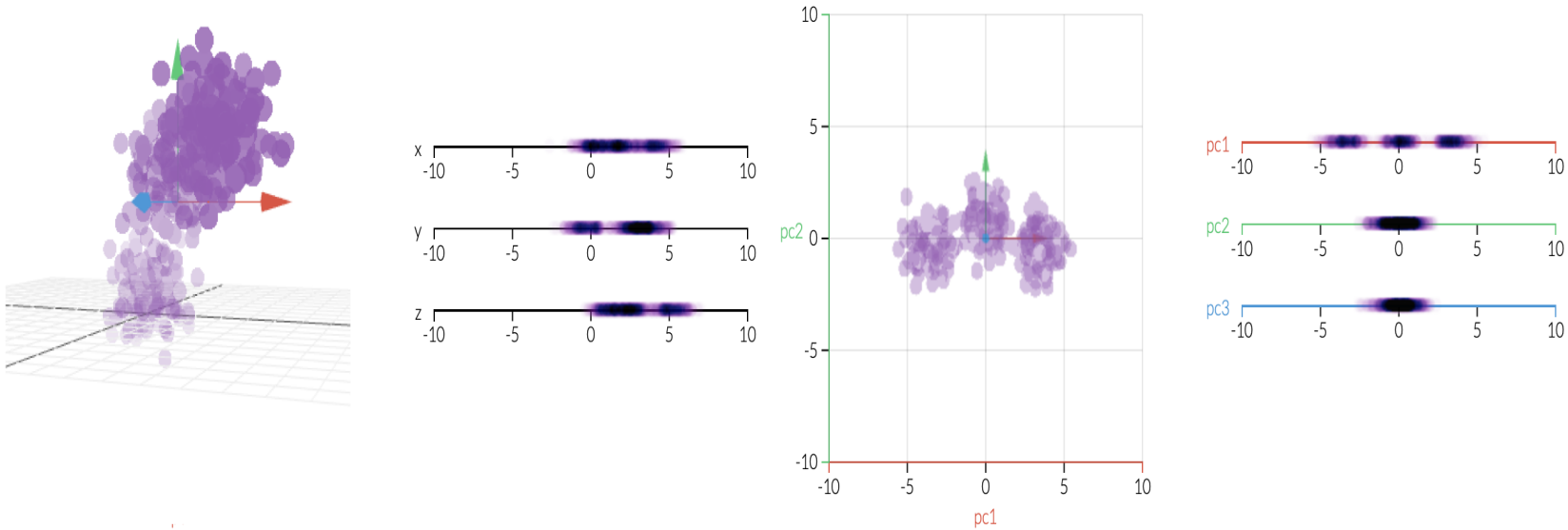
- a widely covered machine learning method frequently used for analyzing unsupervised datasets,
- linear projection of a data set,
- finding directions explaining maximum variance,
- for which the sum-of-squares reconstruction cost is minimized.

[Bishop, C. (1998)]

Used for several purposes :

- Data visualization : visualize multidimensional data in two or three dimensions,
- Uncorrelated variables : generates new uncorrelated variables,
- Dimension reduction : perform data dimension reduction in machine learning.

Principal Component Analysis



Data source

- horizontal axis PC1 has the most variation : 73% of the total variance,
- vertical axis PC2, the second-most : 25% of the total variance,
- third axis PC3, the least : 2% of the total variance.

Principal Component Analysis

For a data set of observed d-dimensional vectors $\{t_n\}$ where $n \in \{1, \ldots, N\}$,

Classical PCA can be broken down into five steps :

I. Standardize the range of continuous initial variables :

$$t_n' = \frac{t_n - \bar{t}}{\sigma(t)},$$

II. Compute the covariance matrix to identify correlations

$$S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \bar{t})(t_n - \bar{t})^T,$$

III. Compute the eigenvectors u_i and eigenvalues λ_i of the covariance matrix to identify the principal components

$$\forall\ i = 1, \ldots, d : \quad Su_i = \lambda_i u_i,$$

IV. Create a feature vector to decide which principal components to keep ($q < d$),

V. Recast the data along the principal components axes : reduced-dimensionality representation of the data set is defined by $x_n = U_q^T (t_n - \bar{t})$ where $U_q = (u_1, \ldots, u_q)$

Principal Component Analysis

Classical PCA :

- But does not define a probability distribution (significant limitation),
- Where does uncertainty enter?
- How do we choose the number of components?
- How can PCA handle missing data?
- Can PCA be embedded into a probabilistic machine learning framework?

PCA can be reformulated as the maximum likelihood solution of a specific **latent variable model**.

[Tipping, M. & Bishop, C. (1999)]

Latent variable models

Suppose we observe $\mathscr{D} = \{\boldsymbol{t}\}$ with $\boldsymbol{t} \in \mathbb{R}^d$. Many datasets contain hidden factors :

- Intelligence behind exam scores
- Customer preferences behind purchases
- Biological processes behind gene expression
- Hidden directions of variation in high-dimensional data

General form of a latent variable model :

$$\boldsymbol{t} = f(\boldsymbol{x}) + \epsilon$$

where

- $\boldsymbol{x}$: latent variable
- ϵ : measurement noise

Probabilistic PCA

For $q < d$, $\{\boldsymbol{x}_i\}$'s are q-dimensional latent variables :

$$\boldsymbol{t}_i = \boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i,$$

$$\boldsymbol{W} : d \times q \text{ matrix}, \quad \boldsymbol{\mu} : d - \text{dimensional vector}, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d).$$

[Tipping, M. & Bishop, C. (1999)]

Thus,

$$\boldsymbol{t}_i | \boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2 \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d).$$

By the formulation above, PCA becomes a probabilistic generative model.

If $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_q)$, then the marginal distribution of the observed variable is :

$$\boldsymbol{t}_i | \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{M}), \quad \boldsymbol{M} = \boldsymbol{W}^T \boldsymbol{W} + \sigma^2 \mathbf{I}_q,$$

and the conditional distribution of $\boldsymbol{x}_i$:

$$\boldsymbol{x}_i | \boldsymbol{t}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2 \sim \mathcal{N}(\boldsymbol{M}^{-1} \boldsymbol{W}^T (\boldsymbol{t}_i - \boldsymbol{\mu}), \sigma^2 \boldsymbol{M}^{-1}).$$

Probabilistic PCA

Maximum likelihood estimation :

The log likelihood under the observed data set $\mathscr{D} = \{\boldsymbol{t}_i\}$:

$$\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2; \mathscr{D}) = -\frac{N}{2}\left[d\ln(2\pi) + \ln|\boldsymbol{M}| + \mathrm{Tr}[\boldsymbol{M}^{-1}S]\right],$$

where S is the sample covariance matrix and

$$\boldsymbol{\mu}_{\mathrm{ML}} = \bar{\boldsymbol{t}}, \quad \boldsymbol{W}_{\mathrm{ML}} = \boldsymbol{U}_q\sqrt{\boldsymbol{\Lambda}_q - \sigma^2_{\mathrm{ML}}\mathbf{I}_q}, \quad \sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i.$$

where the columns of $\boldsymbol{U}_q$ are eigenvectors of S^2, with corresponding eigenvalues in the diagonal matrix $\boldsymbol{\Lambda}_q$.

The maximum of the likelihood is achieved when the q largest eigenvalues are chosen.

Question : But how to choose q? **Solution :** Bayesian PCA.

Bayesian PCA

Bayesian method to determine automatically an appropriate effective dimensionality :

Under the probabilistic reformulation of the PCA,

Step.1 introduce a prior distribution over the parameters of the model : μ, W, σ^2,

Step.2 compute and infer the posterior distribution $p(\mu, W, \sigma^2|\mathscr{D})$,

by focusing on the posterior distribution of W to obtain the effective dimensionality of the latent space, e.i., number of retained principal components.

Instead of estimating a single projection matrix, Bayesian PCA learns a distribution over possible projection matrices.

Bayesian PCA

Bishop, (1998)'s strategy to determine an appropriate effective dimensionality :

- start with many components $q = d - 1$,
- hierarchical prior $p(W|\boldsymbol{\alpha})$ over the matrix $\boldsymbol{W}$,

$$\forall j \in 1, \ldots, q : \boldsymbol{w}_j | \alpha_j \sim \mathcal{N}(0, \alpha_j^{-1} \mathbf{I}_d); \quad \alpha_j \sim \mathcal{G}(a_{\alpha_j}, b_{\alpha_j})$$

- infer the posterior distribution of the parameters.

Interpretation :

- each hyper-parameter α_j controls a column of the matrix W,
- large $\alpha_j \to \infty$, implies $\boldsymbol{w}_j \to 0$ (since $\text{Var}(\boldsymbol{w}_j) = 1/\alpha_j$),
- irrelevant components automatically shrink toward zero.

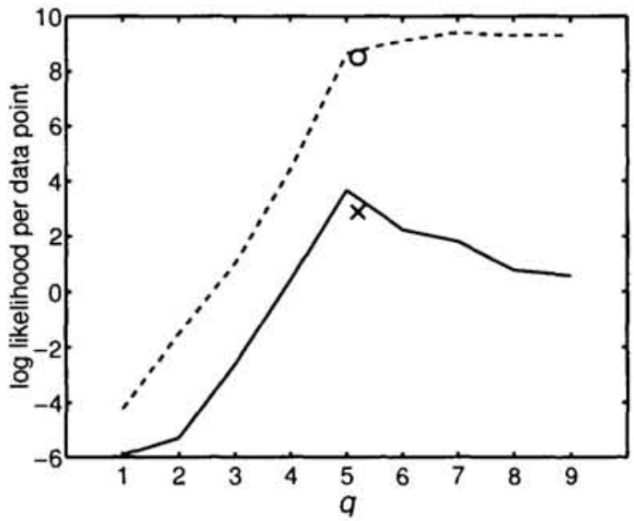
The effective dimensionality of the model is the number of vectors $\boldsymbol{w}_j$ whose values remain non-zero.

[Neal & MacKay, (1995)], [Bishop, (1998)]

Bayesian PCA : Simulation Study

Example

- 10 data sets in 10 dimensions,
- generated from a Gaussian distribution,
- having standard deviations in 5 directions given by $\{1.0, 0.8, 0.6, 0.4, 0.2\}$ and standard deviation 0.04 in the remaining 5 directions,



Training set (dashed curve), Test set (solid curve)

- q_{eff} is averaged over 10 independent experiments,

[Bishop, (1998)]

Bayesian PCA

Alternative Bayesian procedure :

- assign a **mixture prior distribution** to $\boldsymbol{w}_j$:

$$\boldsymbol{w}_j | p, \alpha \sim (1-p)\delta_0(\boldsymbol{w}_j) + p(1-\delta_0(\boldsymbol{w}_j))\mathcal{N}\left(0, \frac{1}{\alpha}\mathbf{I}_d\right), \quad j = 1, \ldots, q,$$

where $p \in]0,1[$, $\delta_0(w) = \begin{cases} 1 & \text{if } w = 0; \\ 0 & \text{otherwise.} \end{cases}$

- $$p \sim \mathcal{B}eta(c_0, c_1), \qquad \alpha \sim \mathcal{G}(a_\alpha, b_\alpha),$$

- $$\boldsymbol{\mu} | \beta \sim \mathcal{N}\left(0, \frac{1}{\beta}\mathbf{I}_d\right), \quad \beta \sim \mathcal{G}(a_\beta, b_\beta), \quad \tau = \frac{1}{\sigma^2} \sim \mathcal{G}(a_\tau, b_\tau).$$

[Sook Oh and Kim, (2010)]

Bayesian PCA

Mixture prior distribution for $\boldsymbol{w}_j$:

$$\boldsymbol{w}_j|p,\alpha \sim (1-p)\delta_0(\boldsymbol{w}_j) + p(1-\delta_0(\boldsymbol{w}_j))\mathcal{N}\left(0, \frac{1}{\alpha}\mathbf{I}_d\right), \quad j = 1,\ldots,q,$$

$$p \sim \mathcal{B}eta(c_0, c_1), \qquad \alpha \sim \mathcal{G}(a_\alpha, b_\alpha),$$

$$\boldsymbol{\mu}|\beta \sim \mathcal{N}\left(0, \frac{1}{\beta}\mathbf{I}_d\right), \quad \beta \sim \mathcal{G}(a_\beta, b_\beta), \quad \tau = \frac{1}{\sigma^2} \sim \mathcal{G}(a_\tau, b_\tau).$$

- **Advantages :**
  - posterior inference on the parameters by applying a simple Gibbs sampler
  - data dimension reduction is performed automatically during the MCMC procedure
- **Disadvantages :**
  - high number of hyper-parameters to be identified
  - how to select the hyper-parameters a_α, b_α?

Bayesian PCA

Mixture prior distribution for $\boldsymbol{w}_j$:

$$\boldsymbol{w}_j|p,\alpha \sim (1-p)\delta_0(\boldsymbol{w}_j) + p(1-\delta_0(\boldsymbol{w}_j))\mathcal{N}\left(0, \frac{1}{\alpha}\mathrm{I}_d\right), \quad j=1,\ldots,q,$$

$$p \sim \mathcal{B}eta(c_0,c_1), \qquad \alpha \sim \mathcal{G}(a_\alpha, b_\alpha),$$

$$\boldsymbol{\mu}|\beta \sim \mathcal{N}\left(0, \frac{1}{\beta}\mathrm{I}_d\right), \quad \beta \sim \mathcal{G}(a_\beta, b_\beta), \quad \tau = \frac{1}{\sigma^2} \sim \mathcal{G}(a_\tau, b_\tau).$$

- **If $a_\alpha = 3$, sensitivity of q to the choice of b_α:**

Posteror estimates for parameters given b_α and b_τ.

b_α $(=b_\tau)$	$\hat{q}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\sigma}^2$
0.1	5	0.54	27.07	0.004
0.5	4	0.45	14.11	0.0142
1	4	0.45	10.05	0.0143
2	3	0.36	5.42	0.0512

[Sook Oh and Kim, (2010)]

Bayesian PCA

Noninformative prior modeling :

Mixture prior distribution for $\boldsymbol{w}_j$:

$$\boldsymbol{w}_j|p,\alpha \sim (1-p)\delta_0(\boldsymbol{w}_j) + p(1-\delta_0(\boldsymbol{w}_j))\mathcal{N}\left(0, \frac{1}{\alpha}\mathbf{I}_d\right), \quad j = 1,\ldots,q,$$

Noninformative prior for p and exponential distribution for α :

$$p \sim \mathscr{B}eta(c_0, c_1),$$

Jeffreys prior for $(\boldsymbol{\mu}, \sigma^2)$:

$$\pi(\boldsymbol{\mu}, \tau) \propto \frac{1}{\sqrt{\tau}}, \quad \tau = \frac{1}{\sigma^2}$$

And for α :

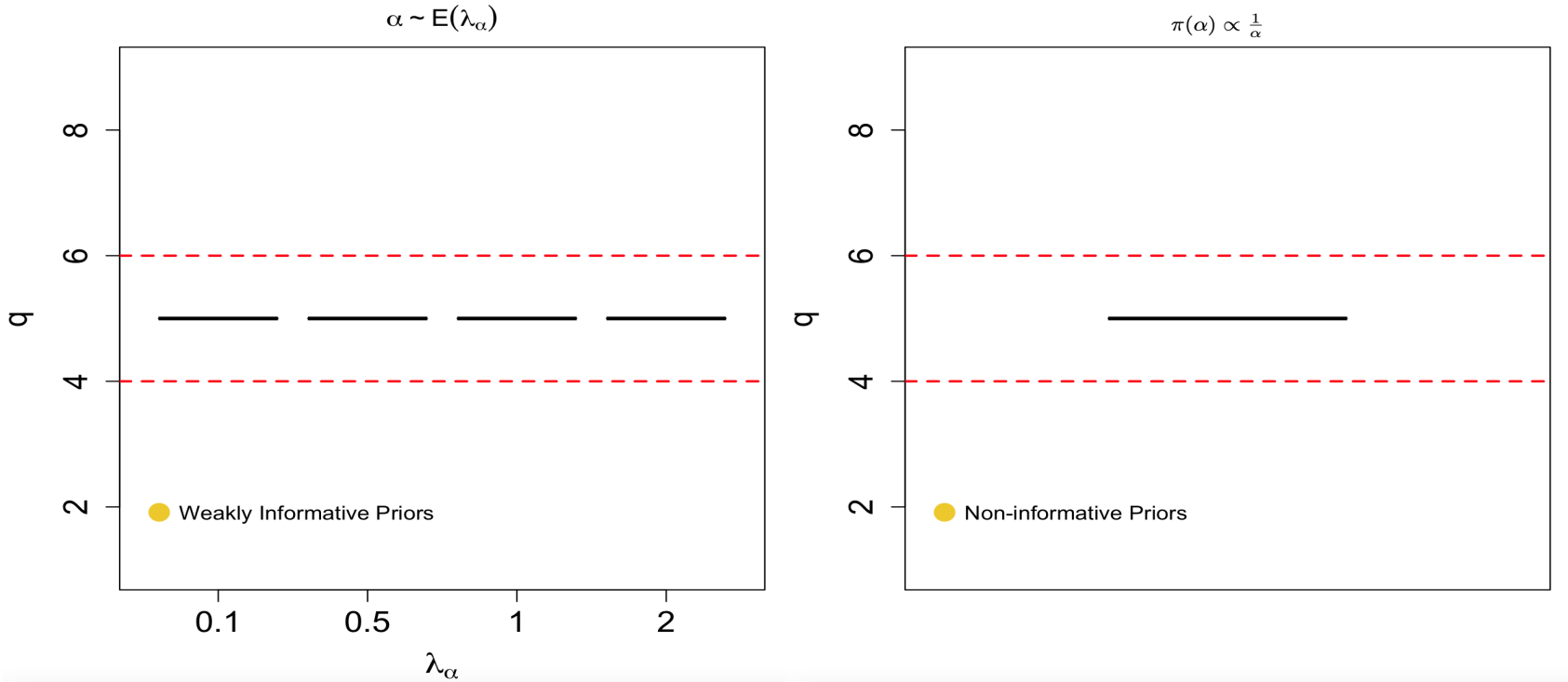
$$\alpha \sim \mathscr{E}xp(\lambda_\alpha), \quad \text{or} \quad \pi(\alpha) \propto 1/\alpha.$$

Resulting conditional posterior distributions with closed-form PDF, let us to use a Gibbs sampler to obtain the posterior estimates of the parameters.

Bayesian PCA : Illustrations

50 independent data simulations, each of size $n = 20$ points with $d = 10$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$:

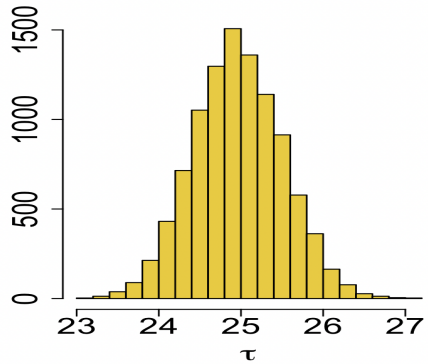
- σ^2: in 5 directions given by $(1.0, 0.8, 0.6, 0.4, 0.2)^2$ and 0.04^2 in the remaining 5 directions

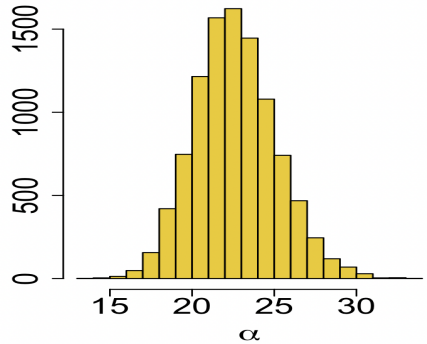


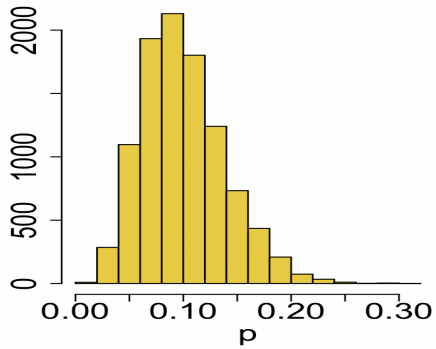
Note that for each λ_α, 50 values of $\hat{q}$ are the sample average of the MCMC draws (with 9000 discarded draws as burn-in).

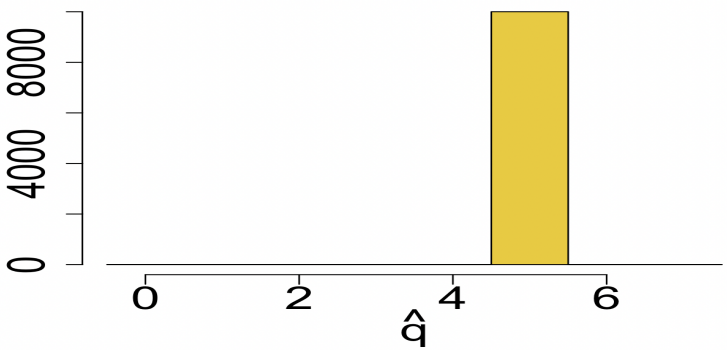
Bayesian PCA : Illustrations

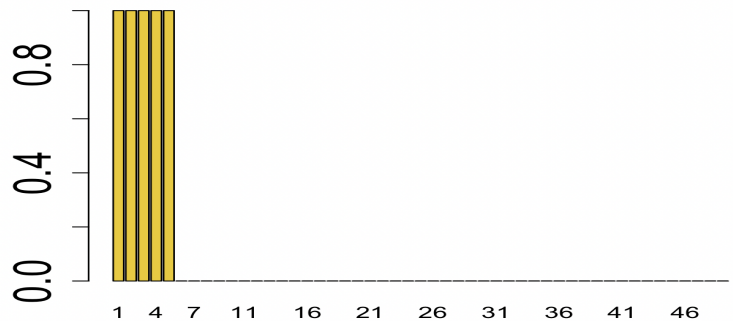
$n = 100$ with $d = 50$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ with $\sigma^2 = 2$: in 5 directions and 0.04 in the remaining 45 directions;

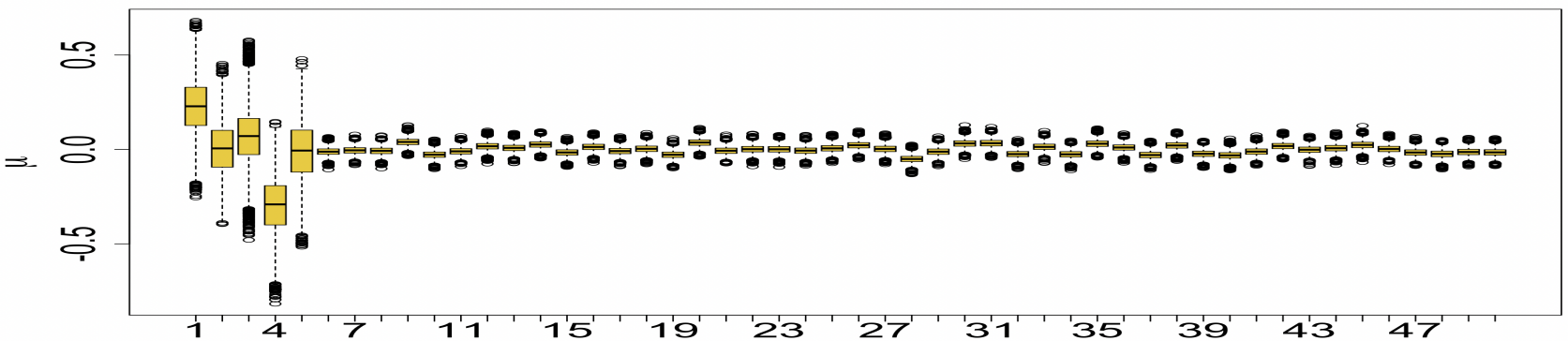






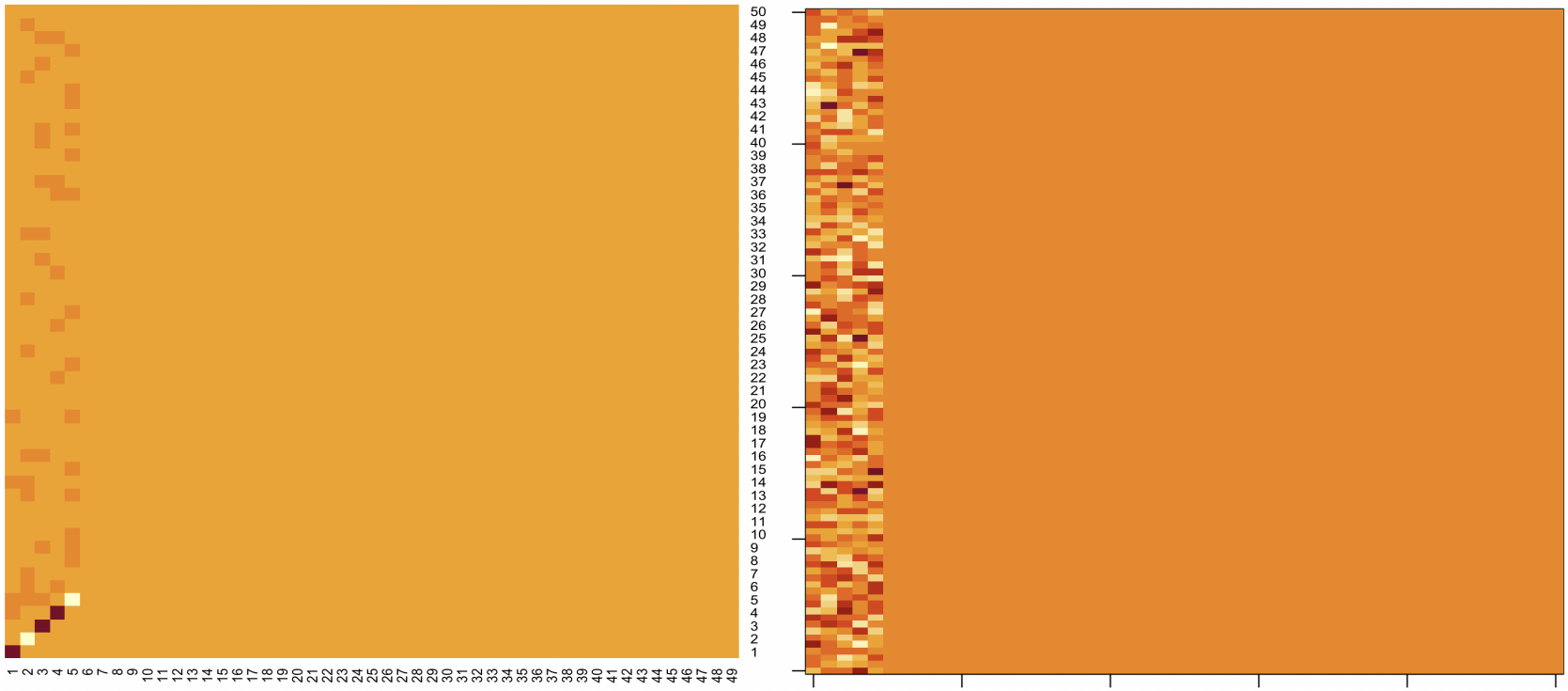






Bayesian PCA : Illustrations

$n = 100$ with $d = 50$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ with $\sigma^2 = 2$: in 5 directions and 0.04 in the remaining 45 directions;



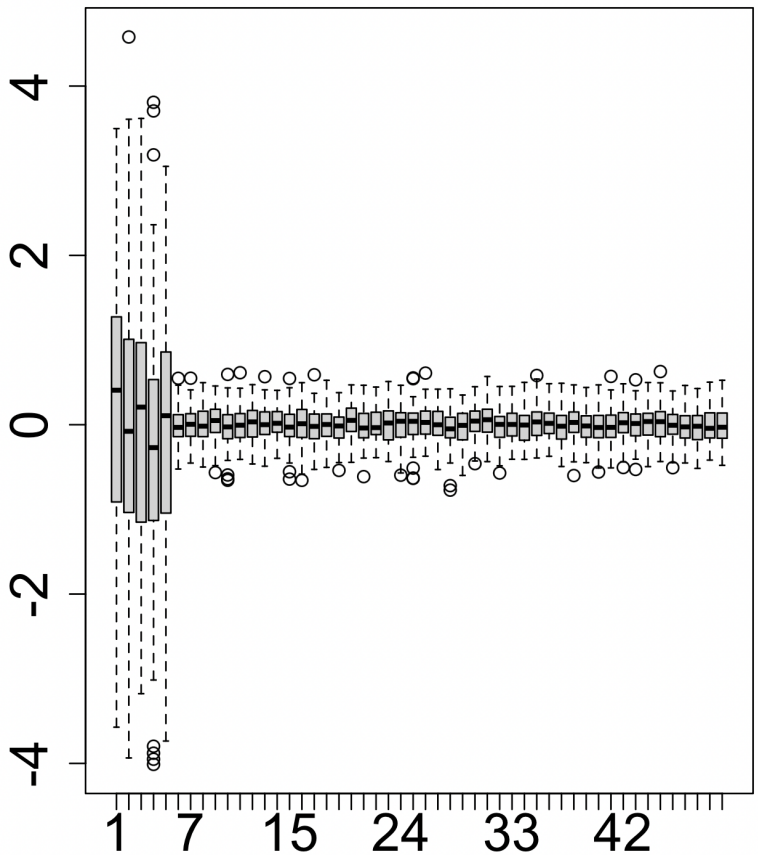
Posterior mean of :

(Left) $\widehat{W}$; (Right) $\widehat{X}$.

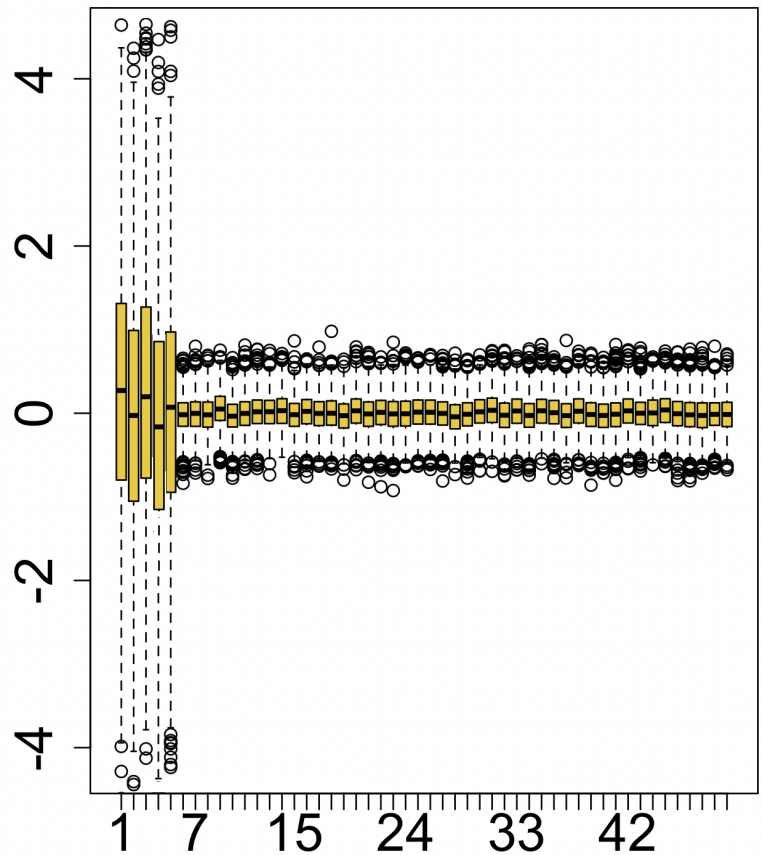
Bayesian PCA : Illustrations

$n = 100$ with $d = 50$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ with $\sigma^2 = 2$: in 5 directions and 0.04 in the remaining 45 directions;

Uncertainty quantification?



(Left) Data distribution



(Right) Posterior predictive distribution

Bayesian PCA : Illustrations

$n = 100$ with $d = 50$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ with $\sigma^2 = 2$: in 5 directions and 0.04 in the remaining 45 directions;

Uncertainty quantification?

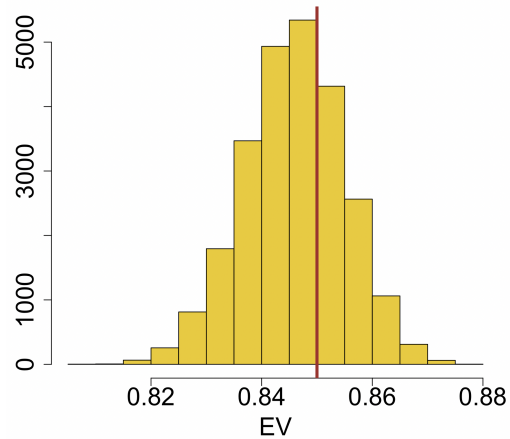
- Posterior distribution of q :

d	1	2	3	4	5	6	7	> 7
$\mathbb{P}(q = d \vert ...)$	$2e{-}5$	$3.6e{-}4$	$6.4e{-}4$	$3.32e{-}3$	0.995	$2.2e{-}4$	$6e{-}5$	0

- Bayesian explained variance :

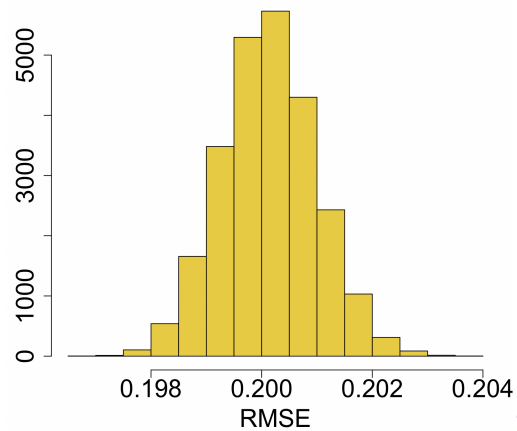
$$EV = \frac{1}{L} \sum_{l=1}^{L} \frac{tr\left(\boldsymbol{W}^{(l)} \boldsymbol{W}^{(l)\,T}\right)}{tr\left(\boldsymbol{W}^{(l)} \boldsymbol{W}^{(l)\,T}\right) + d\sigma^{2\,(l)}}; \quad \widehat{EV} = 0.85$$

L : number of MCMC iterations.



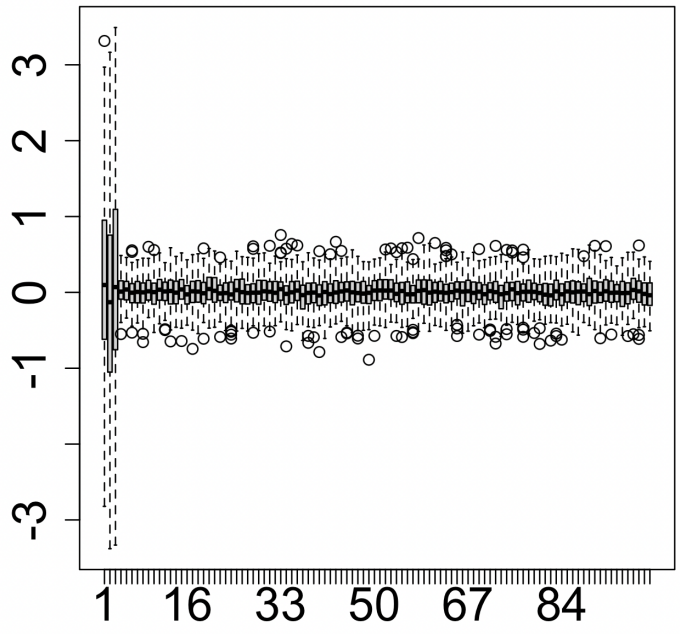
- Posterior distribution of the RMSE :

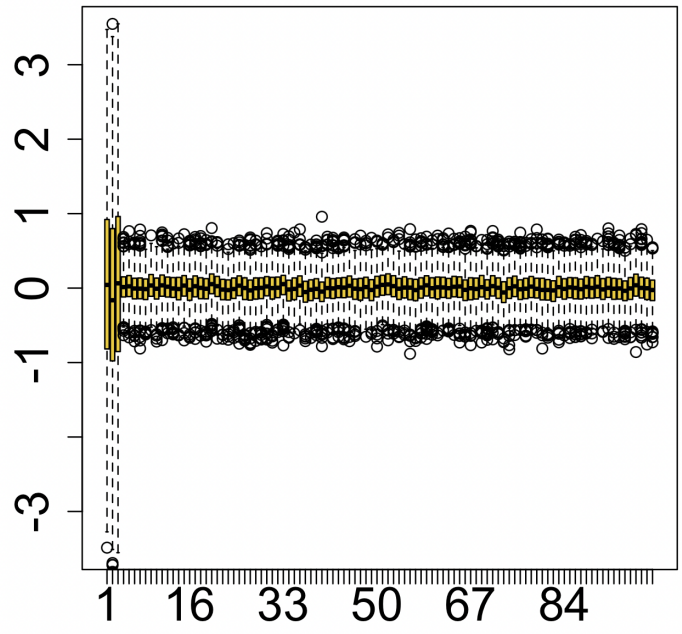
$$\widehat{\boldsymbol{t}}_i = \hat{\mu}^{(l)} + \boldsymbol{W}^{(l)} \boldsymbol{x}_i; \qquad \text{RMSE}^{(l)} = \sqrt{\frac{1}{nd} \sum_{i=1}^{n} \|\boldsymbol{t}_i - \widehat{\boldsymbol{t}}_i\|^2}.$$

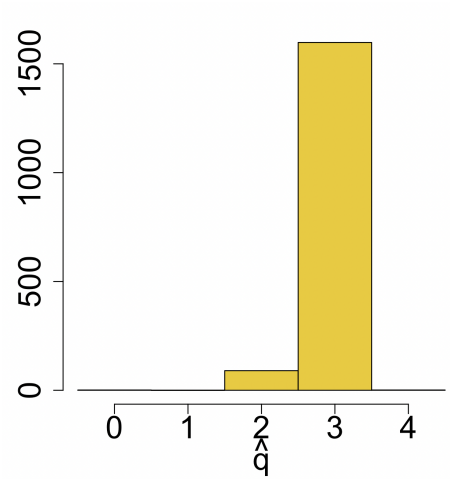


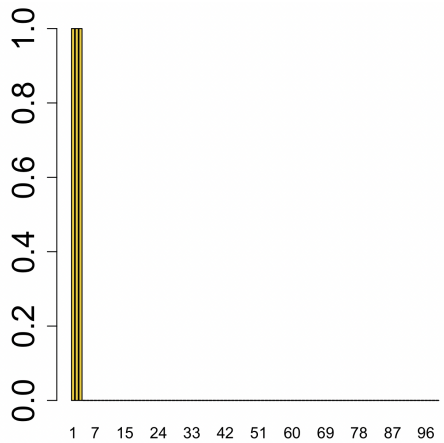
Bayesian PCA : Illustrations

$n = 120$ with $d = 100$ from $\mathcal{N}_d(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$ with $\sigma^2 = 2$: in 3 directions and 0.04 in the remaining 97 directions;









Conclusion

- Classical PCA :
  - choose the latent space dimensionality q manually,
- Bayesian PCA :
  - offers a significant advantage in allowing the model discover hidden structure in the data automatically,
  - quantify uncertainty,
  - using noninformative prior modeling leads to reduce the number of model hyper-parameter,
  - easy implementation by applying a Gibbs sampler algorithm,
  - particularly advantageous for small data sets in high dimensions.

References

- Bishop, C. (1998). *Bayesian pca*. Advances in neural information processing systems, 11.
- Tipping, M., & Bishop, C. (1999). *Probabilistic principal component analysis.* Journal of the Royal Statistical Society, 11(3), 611-622.
- Oh, H. S., & Kim, D. G. (2010). *Bayesian principal component analysis with mixture priors*. Journal of the Korean Statistical Society, 39(3), 387-396.


Thank you