# An Information-theoretic Approach to Unsupervised Feature Extraction for High-Dimensional Data

Shao-Lun Huang

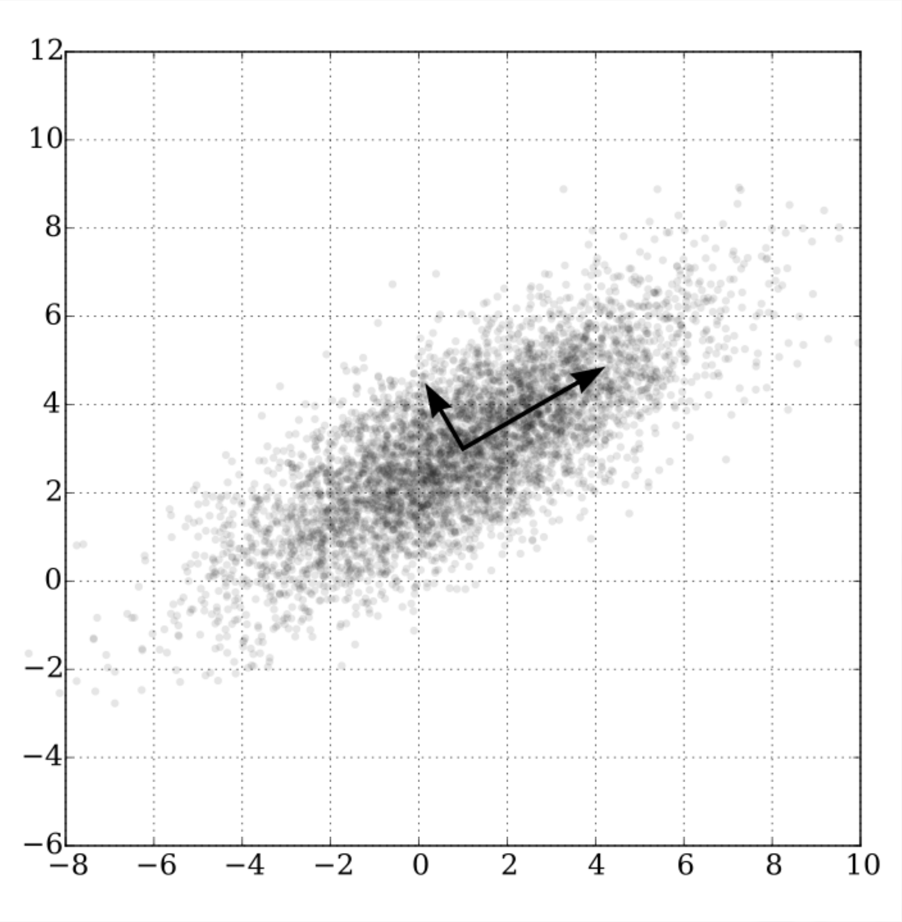
Tsinghua-Berkeley Shenzhen Institute (TBSI)

Joint work with Xiangxiang Xu (Tsinghua) and Lizhong Zheng (MIT)

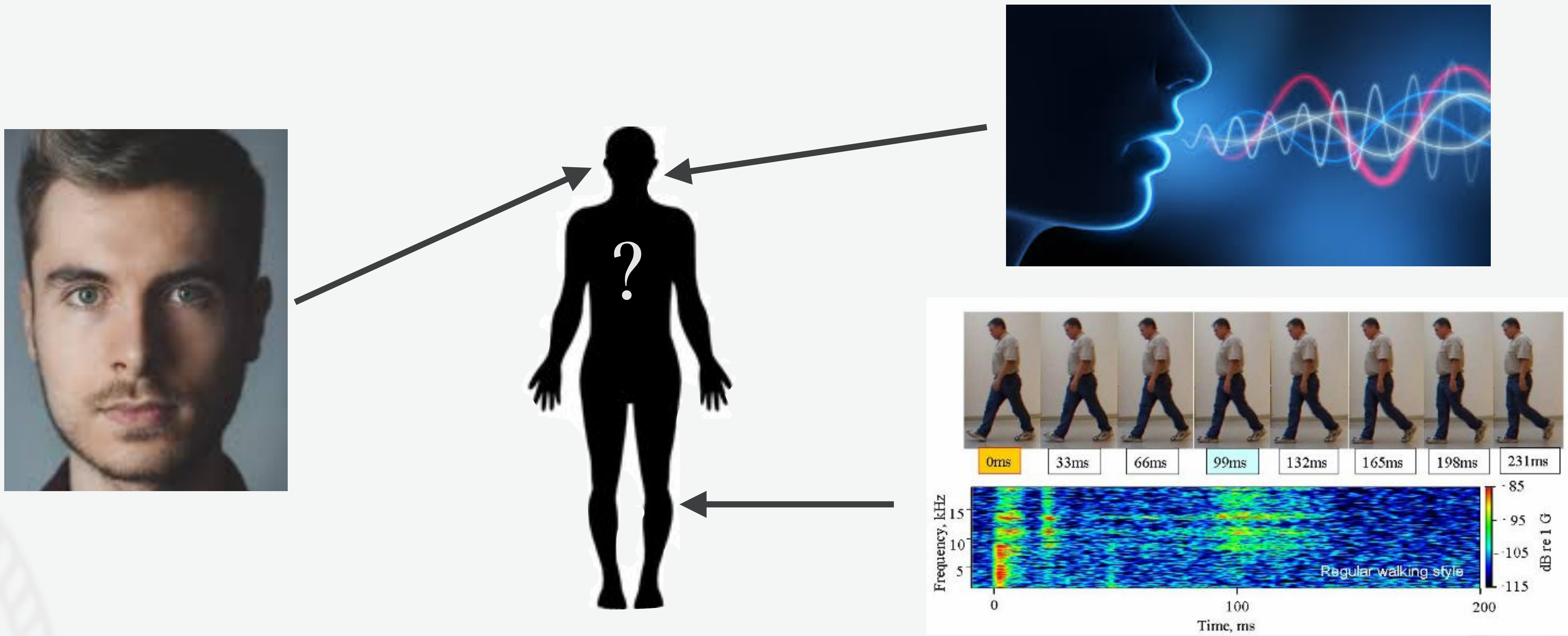
2019 Conference on Entropy and Its Applications

# Principal Component Analysis

- Search the direction that different dimensions of data are aligned.
- Extract the common randomness between different dimensions.
- How to formalize this idea by information theory?

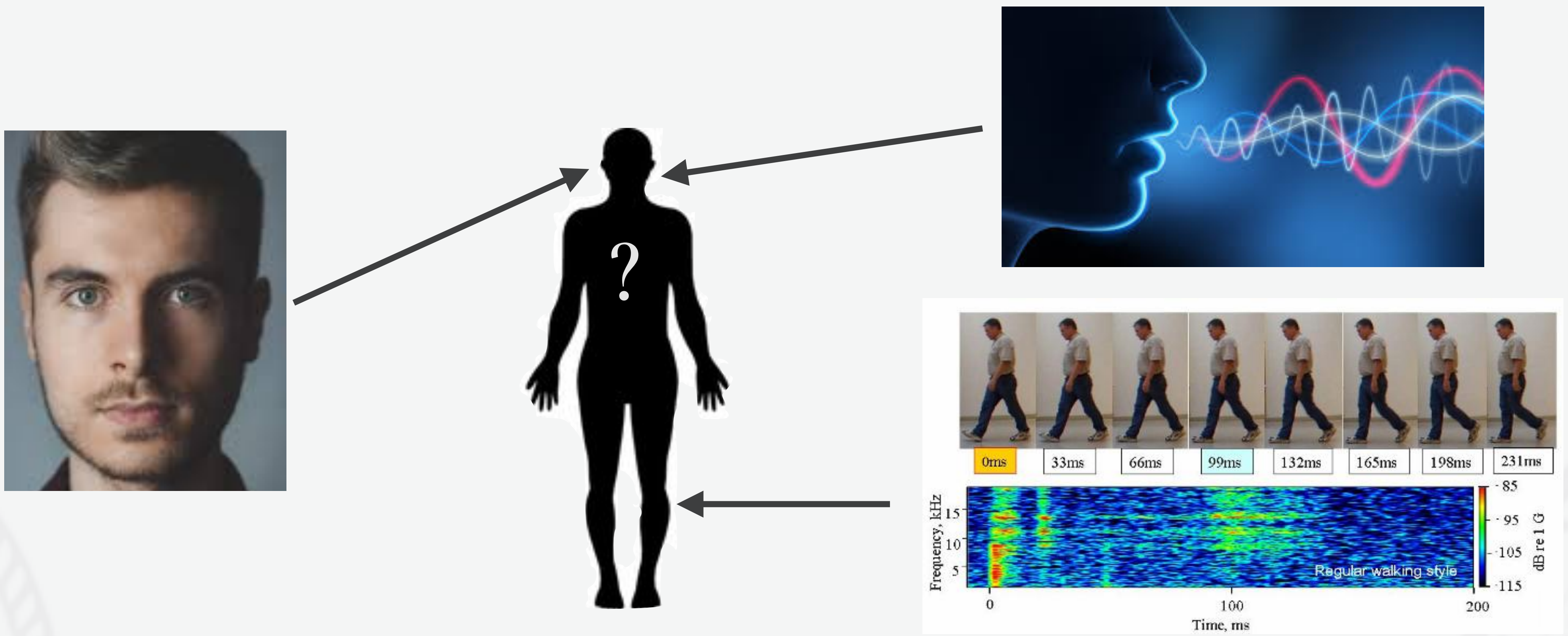


# Multimodal Data Analyses



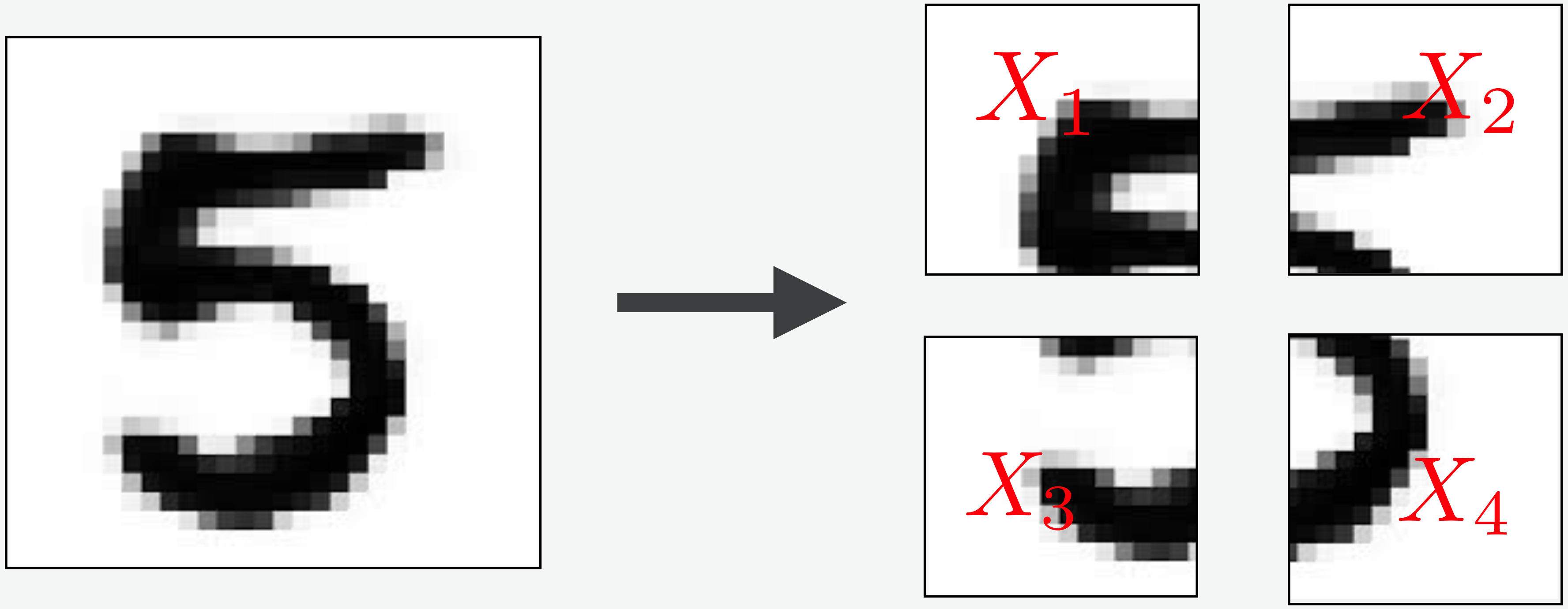

- Learning from multimodal sensory data.
  - Need to find informative representations for data.

# Multimodal Data Analyses



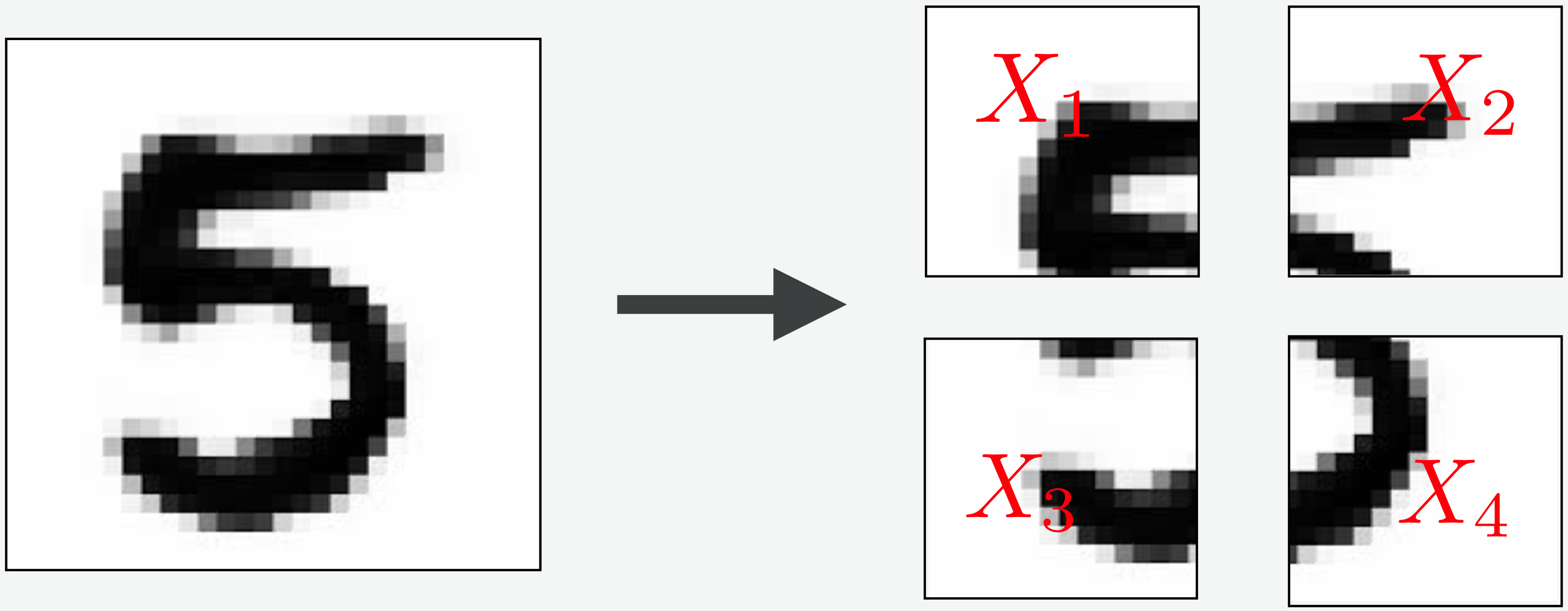

- Learning from multimodal sensory data.
  - Need to find informative representations for data.
  - Extracting the representation to describe the common structure between multimodal data.

# The MNIST Problem



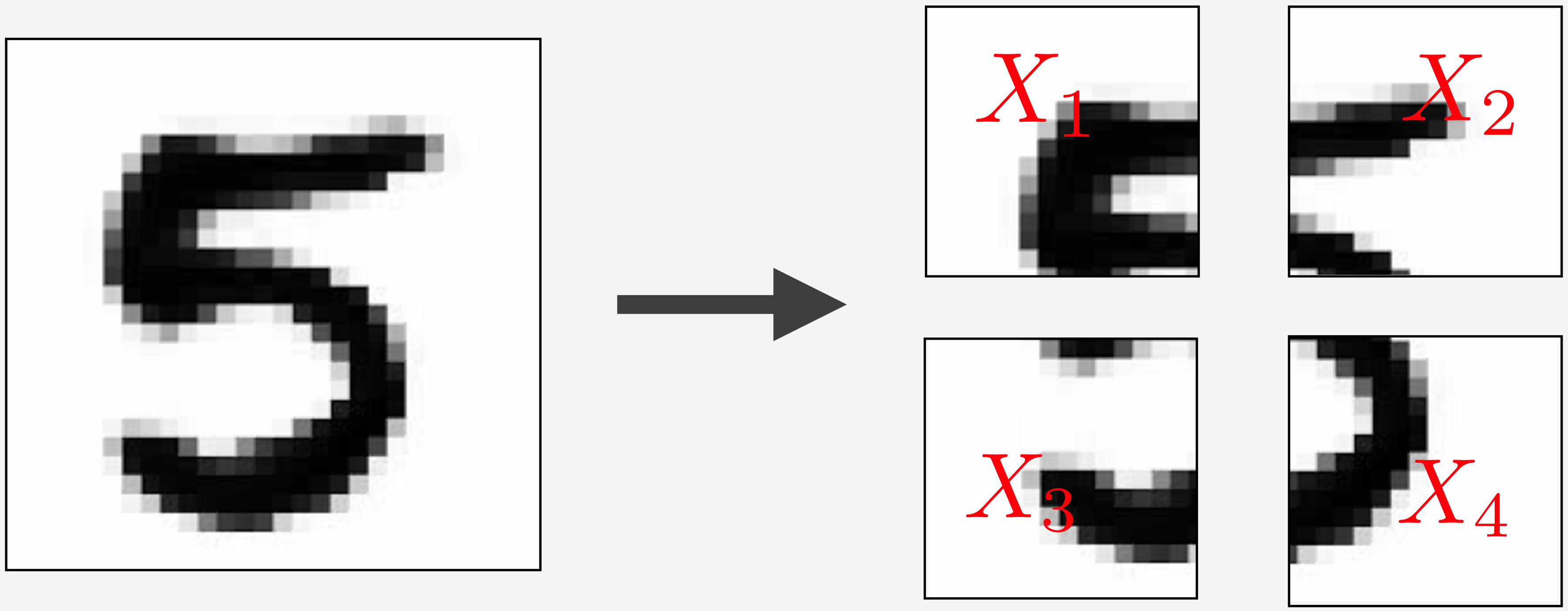

- MNIST hand written digits problem:
    - Divide image into subareas, extract features for each subarea.

# The MNIST Problem




- MNIST hand written digits problem:
  - Divide image into subareas, extract features for each subarea.
- The extracted features should describing the common information (the label) shared by different subareas.

# The MNIST Problem




- MNIST hand written digits problem:
  - Divide image into subareas, extract features for each subarea.
- The extracted features should describing the common information (the label) shared by different subareas.
- How to extract the common structure shared between data variables?

# Mathematical Formulation

- Given pairwise dependent discrete random variables $X_1, \ldots, X_d$, with joint distribution $P_{X_1, \ldots, X_d}$.
  - Observed sampled vectors generated i.i.d. from $P_{X_1, \ldots, X_d}$.

# Mathematical Formulation

- Given pairwise dependent discrete random variables $X_1, \ldots, X_d$, with joint distribution $P_{X_1, \ldots, X_d}$.
  - Observed sampled vectors generated i.i.d. from $P_{X_1, \ldots, X_d}$.
- Want to find the function $f(X_1, X_2, \ldots, X_d)$ that conveys much information about the common structure between $X_1, \ldots, X_d$.

# Mathematical Formulation

- Given pairwise dependent discrete random variables $X_1, \ldots, X_d$, with joint distribution $P_{X_1, \ldots, X_d}$.
    - Observed sampled vectors generated i.i.d. from $P_{X_1, \ldots, X_d}$.
- Want to find the function $f(X_1, X_2, \ldots, X_d)$ that conveys much information about the common structure between $X_1, \ldots, X_d$.
    - Need an information metric to measure the commonness.

# Mathematical Formulation

- Given pairwise dependent discrete random variables $X_1, \ldots, X_d$, with joint distribution $P_{X_1, \ldots, X_d}$.
  - Observed sampled vectors generated i.i.d. from $P_{X_1, \ldots, X_d}$.
- Want to find the function $f(X_1, X_2, \ldots, X_d)$ that conveys much information about the common structure between $X_1, \ldots, X_d$.
  - Need an information metric to measure the commonness.
  - Better to be computable by efficiently algorithms from data.

# The Common Information Measure

- Given discrete random variables $X_1$, ..., $X_d$, the **Watanabe's total correlation** is a measurement for their common information:

$$C(X_1, \ldots, X_d) \triangleq D(P_{X_1 \cdots X_d} \| P_{X_1} \cdots P_{X_d})$$

# The Common Information Measure

- Given discrete random variables $X_1, \ldots, X_d$, the **Watanabe's total correlation** is a measurement for their common information:

$$C(X_1, \ldots, X_d) \triangleq D(P_{X_1 \cdots X_d} \| P_{X_1} \cdots P_{X_d})$$

- For attribute $U$ of $X_1, \ldots, X_d$ with conditional distribution $P_{X_1, \ldots, X_d | U}$, we monitor the loss of total correlation given $U$:

$$C(X_1, \ldots, X_d) - C(X_1, \ldots, X_d | U) = \sum_{i=1}^{d} I(U; X_i) - I(U; X_1, \ldots, X_d)$$

  - The amount of common information contained in $U$.

# The Common Information Measure

- Given discrete random variables $X_1, \ldots, X_d$, the **Watanabe's total correlation** is a measurement for their common information:

$$C(X_1, \ldots, X_d) \triangleq D(P_{X_1 \cdots X_d} \| P_{X_1} \cdots P_{X_d})$$

- For attribute $U$ of $X_1, \ldots, X_d$ with conditional distribution $P_{X_1, \ldots, X_d|U}$, we monitor the loss of total correlation given $U$:

$$C(X_1, \ldots, X_d) - C(X_1, \ldots, X_d|U) = \sum_{i=1}^{d} I(U; X_i) - I(U; X_1, \ldots, X_d)$$

  - The amount of common information contained in $U$.

- Learn the most informative feature about the common information = find $U$ maximize the total correlation loss.

# Extract Informative Features From Data

$$\max_{P_{UX_1\cdots X_d}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Once $P_{X_1,\ldots,X_d|U}$ is solved, the log-likelihood function to detect $U$ leads to the representation of data for the common structure:

$$f_u(x_1,\ldots,x_d) = \log \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|u)}{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}$$

# Extract Informative Features From Data

$$\max_{P_{UX_1\cdots X_d}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Once $P_{X_1,\ldots,X_d|U}$ is solved, the log-likelihood function to detect $U$ leads to the representation of data for the common structure:

$$f_u(x_1,\ldots,x_d) = \log \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|u)}{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}$$

- Information sieve: restrict the cardinality of $U$ [Ver Steeg *et. al*, 14]
  - The optimal solution has no systematic structure

# Extract Informative Features From Data

$$\max_{P_{UX_1\cdots X_d}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Once $P_{X_1,\ldots,X_d|U}$ is solved, the log-likelihood function to detect $U$ leads to the representation of data for the common structure:

$$f_u(x_1,\ldots,x_d) = \log \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|u)}{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}$$

- Information sieve: restrict the cardinality of $U$ [Ver Steeg *et. al*, 14]
  - The optimal solution has no systematic structure
- Want to add an extra constraint $I(U;X_1\cdots X_d) \leq \frac{1}{2}\epsilon^2$ for small $\epsilon$
  - Can focus on the most significant low-dimensional feature.
  - A geometric structure for optimally decomposing common information.

# Extract Informative Features From Data

$$\max_{P_{UX_1\cdots X_d}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Once $P_{X_1,\ldots,X_d|U}$ is solved, the log-likelihood function to detect $U$ leads to the representation of data for the common structure:

$$f_u(x_1,\ldots,x_d) = \log \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|u)}{P_{X_1\cdots X_d}(x_1,\ldots,x_d)} \simeq \frac{P_{X_1\cdots X_d|U=u} - P_{X_1\cdots X_d}}{P_{X_1\cdots X_d}}$$

- Information sieve: restrict the cardinality of $U$ [Ver Steeg *et. al*, 14]
  - The optimal solution has no systematic structure
- Want to add an extra constraint $I(U;X_1\cdots X_d) \leq \frac{1}{2}\epsilon^2$ for small $\epsilon$
  - Can focus on the most significant low-dimensional feature.
  - A geometric structure for optimally decomposing common information.

# How to Find Optimal Features?

$$\max_{\substack{P_{UX_1\cdots X_d}: \\ I(U;X_1,\ldots,X_d)\leq\frac{1}{2}\epsilon^2}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Information vector:

$$\psi_i(x_i) = \frac{P_{X_i|U}(x_i|0) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \qquad \phi(x_1,\ldots,x_d) = \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|0) - P_{X_1\cdots X_d}(x_1,\ldots,x_d)}{\epsilon\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}}$$

# How to Find Optimal Features?

$$\max_{\substack{P_{UX_1\cdots X_d}:\\ I(U;X_1,\ldots,X_d)\le \frac{1}{2}\epsilon^2}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Information vector:

$$\psi_i(x_i) = \frac{P_{X_i|U}(x_i|0) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \qquad \phi(x_1,\ldots,x_d) = \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|0) - P_{X_1\cdots X_d}(x_1,\ldots,x_d)}{\epsilon\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}}$$

- Correspondence to log-likelihood functions:

$$\psi_i(x_i) = \sqrt{P_{X_i}(x_i)}f_i(x_i) \qquad \phi(x_1,\ldots,x_d) = \sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}f(x_1,\ldots,x_d)$$

# How to Find Optimal Features?

$$\max_{\substack{P_{UX_1\cdots X_d}:\\ I(U;X_1,\ldots,X_d)\le \frac{1}{2}\epsilon^2}} \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d)$$

- Information vector:

$$\psi_i(x_i) = \frac{P_{X_i|U}(x_i|0) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \qquad \phi(x_1,\ldots,x_d) = \frac{P_{X_1\cdots X_d|U}(x_1,\ldots,x_d|0) - P_{X_1\cdots X_d}(x_1,\ldots,x_d)}{\epsilon\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}}$$

- Correspondence to log-likelihood functions:

$$\psi_i(x_i) = \sqrt{P_{X_i}(x_i)} f_i(x_i) \qquad \phi(x_1,\ldots,x_d) = \sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)} f(x_1,\ldots,x_d)$$

- Approximate the K-L divergence:

$$I(U;X_i) \simeq \frac{1}{2}\epsilon^2\|\psi_i\|^2 \qquad I(U;X_1,\ldots,X_d) \simeq \frac{1}{2}\epsilon^2\|\phi\|^2$$

# Linear Transform of Information Vectors

$$\max \sum_{i=1}^{d} \|\psi_i\|^2, \quad \text{subject to: } \|\phi\|^2 \leq 1$$

- Linear transform between information vectors:

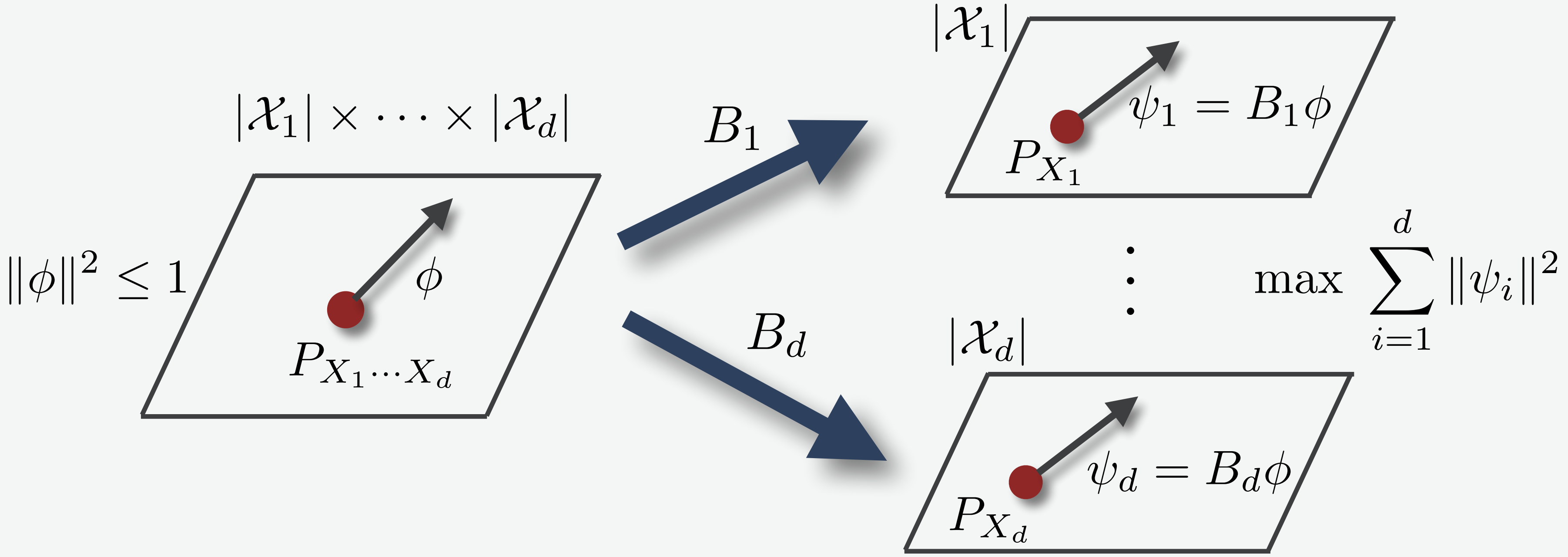
$$\psi_i = B_i \cdot \phi, \quad B_i(\hat{x}_i; (x_1, \ldots, x_d)) = \begin{cases} \frac{\sqrt{P_{X_1 \cdots X_d}(x_1, \ldots, x_d)}}{\sqrt{P_{X_i}(\hat{x}_i)}} & \text{if } \hat{x}_i = x_i, \\ 0 & \text{otherwise.} \end{cases}$$

# Linear Transform of Information Vectors

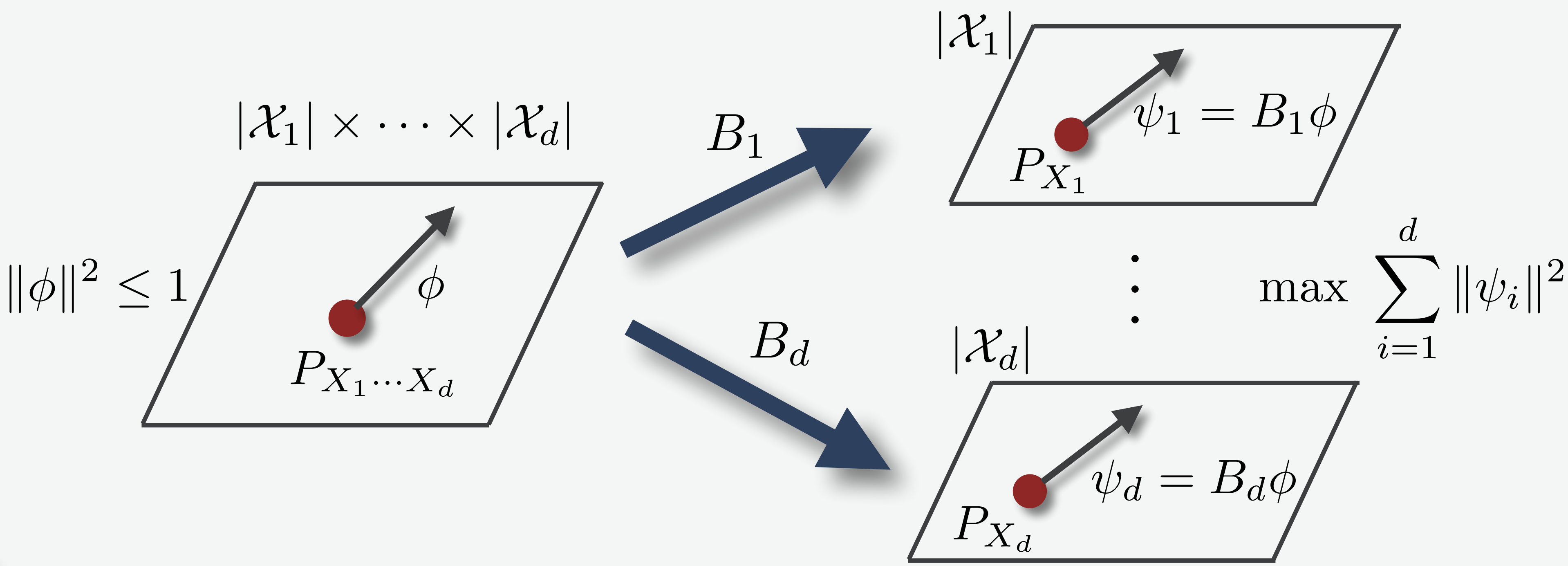
$$\max \sum_{i=1}^{d} \|\psi_i\|^2, \quad \text{subject to: } \|\phi\|^2 \leq 1$$

- Linear transform between information vectors:

$$\psi_i = B_i \cdot \phi, \quad B_i(\hat{x}_i; (x_1, \ldots, x_d)) = \begin{cases} \frac{\sqrt{P_{X_1 \cdots X_d}(x_1, \ldots, x_d)}}{\sqrt{P_{X_i}(\hat{x}_i)}} & \text{if } \hat{x}_i = x_i, \\ 0 & \text{otherwise.} \end{cases}$$



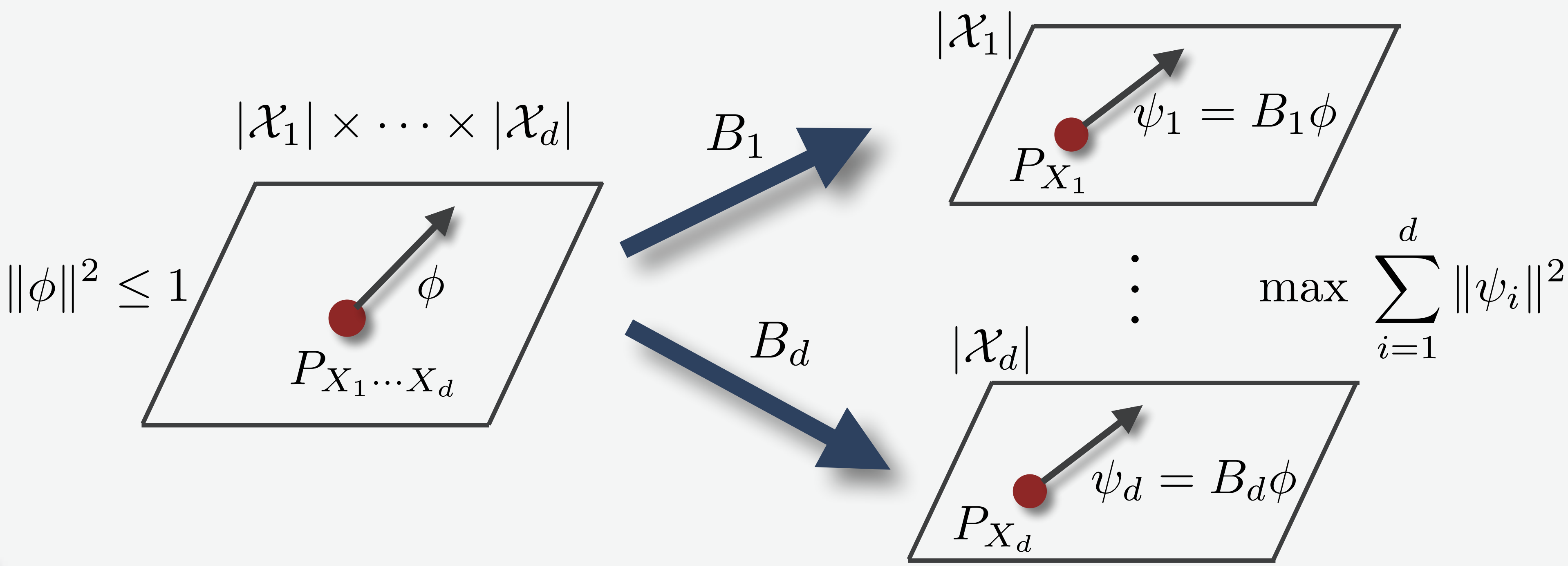
# The Geometric Interpretation




$$B_i^T = \begin{bmatrix} \ddots & & \mathbf{0} \\ & \sqrt{P_{X_1 \cdots X_d}} & \\ \mathbf{0} & & \ddots \end{bmatrix} \cdot \mathbb{I}_i^T \cdot \begin{bmatrix} \ddots & & \mathbf{0} \\ & \sqrt{P_{X_i}^{-1}} & \\ \mathbf{0} & & \ddots \end{bmatrix}$$

Projection matrix: $\mathbb{I}_i(\hat{x}_i; (x_1, \ldots, x_d)) = \begin{cases} 1 & \text{if } \hat{x}_i = x_i, \\ 0 & \text{otherwise.} \end{cases}$

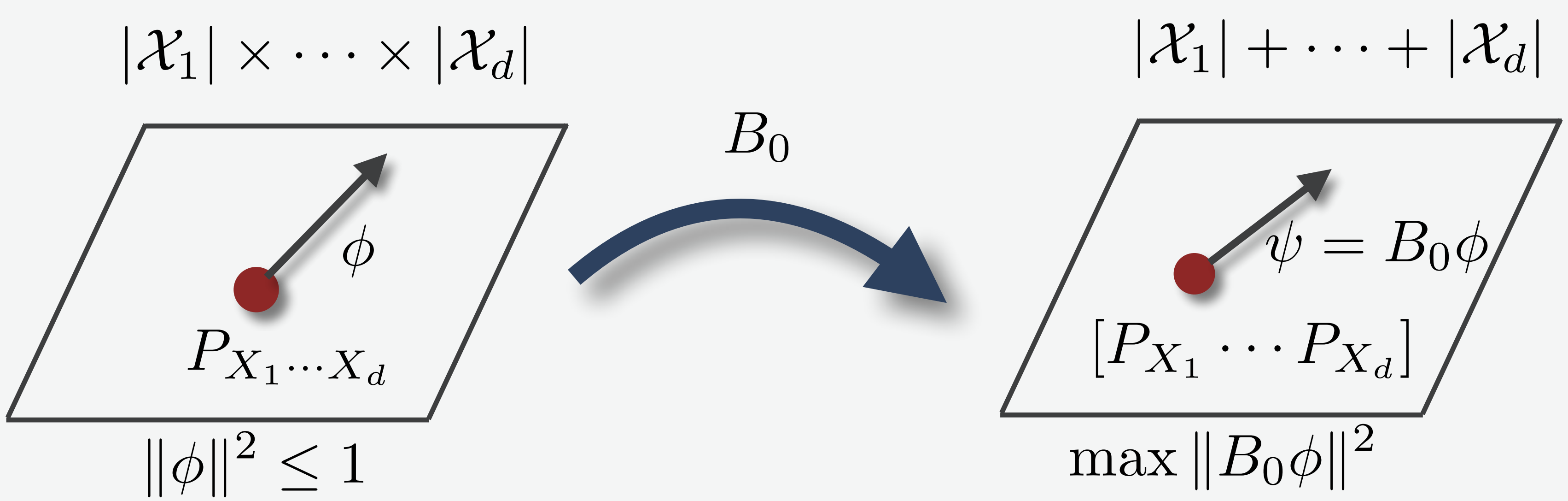
# The Geometric Interpretation




- Merge to a single linear transform:

$$\sum_{i=1}^{d} \|\psi_i\|^2 = \left\| \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_d \end{bmatrix} \phi \right\|^2 \qquad B_0 = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_d \end{bmatrix}$$

# The Geometric Interpretation




$$\psi = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_d \end{bmatrix}$$

$$B_0^T = \begin{bmatrix} \ddots & & \mathbf{0} \\ & \sqrt{P_{X_1 \cdots X_d}} & \\ \mathbf{0} & & \ddots \end{bmatrix} \begin{bmatrix} \mathbb{I}_1^T \cdots \mathbb{I}_d^T \end{bmatrix} \begin{bmatrix} \sqrt{P_{X_1}^{-1}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{P_{X_d}^{-1}} \end{bmatrix}$$

- Solve the singular value decomposition of $B_0$.

# The Correspondence

$$B_0^T = \begin{pmatrix} \ddots & & \mathbf{0} \\ & \sqrt{P_{X_1\cdots X_d}} & \\ \mathbf{0} & & \ddots \end{pmatrix} \begin{pmatrix} \mathbb{I}_1^T \cdots \mathbb{I}_d^T \end{pmatrix} \begin{pmatrix} \sqrt{P_{X_1}^{-1}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{P_{X_d}^{-1}} \end{pmatrix} \Bigg) |\mathcal{X}_1| + \cdots + |\mathcal{X}_d|$$

Information vectors: $\phi$ $\xrightarrow{B_0 \phi = \sigma_1 \psi}$ $\psi = \left[\psi_1^{\mathrm{T}} \psi_2^{\mathrm{T}} \cdots \psi_d^{\mathrm{T}}\right]^{\mathrm{T}}$

$\phi \xleftarrow{\sigma_1 \phi = B_0^T \psi} \psi$

$$f = \frac{\phi}{\sqrt{P_{X_1\cdots X_d}}} \qquad f_i(x_i) = \frac{\psi_i(x_i)}{\sqrt{P_{X_i}(x_i)}}$$

Log-likelihood: $f \longleftarrow \{f_1(X_1), \ldots, f_d(X_d)\}$

$$f(X_1, \ldots X_d) = f_1(X_1) + \cdots f_d(X_d)$$

# The Correspondence

$$B_0^T = \begin{bmatrix} \ddots & & \mathbf{0} \\ & \sqrt{P_{X_1\cdots X_d}} & \\ \mathbf{0} & & \ddots \end{bmatrix} \begin{bmatrix} \mathbb{I}_1^T & \cdots & \mathbb{I}_d^T \end{bmatrix} \left( \begin{bmatrix} \sqrt{P_{X_1}^{-1}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{P_{X_d}^{-1}} \end{bmatrix} \right) |\mathcal{X}_1| + \cdots + |\mathcal{X}_d|$$

Information vectors: $\phi$ $\underset{\sigma_1 \phi = B_0^T \psi}{\overset{B_0 \phi = \sigma_1 \psi}{\rightleftarrows}}$ $\psi = \left[\psi_1^{\mathrm{T}} \psi_2^{\mathrm{T}} \cdots \psi_d^{\mathrm{T}}\right]^{\mathrm{T}}$

$$f = \frac{\phi}{\sqrt{P_{X_1 \cdots X_d}}} \qquad\qquad f_i(x_i) = \frac{\psi_i(x_i)}{\sqrt{P_{X_i}(x_i)}}$$

Log-likelihood: $f$ $\longleftarrow$ $\{f_1(X_1), \ldots, f_d(X_d)\}$

$$f(X_1, \ldots X_d) = f_1(X_1) + \cdots f_d(X_d)$$

The optimal solution: $P^*(x_1, \ldots, x_d | u) = P(x_1, \ldots, x_d) \left(1 + \epsilon h(u) \sum_{i=1}^{d} f_i(x_i)\right)$

# Compute Singular Vectors

- Easier to compute the left singular vectors of $B_0$, lower dimension.

$$B \triangleq B_0 \cdot B_0^T = \underbrace{\begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1d} \\ B_{21} & B_{22} & \cdots & B_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ B_{d1} & B_{d2} & \cdots & B_{dd} \end{bmatrix}}_{|\mathcal{X}_1| + |\mathcal{X}_2| + \cdots + |\mathcal{X}_d|}$$

$$\dim(B_{ij}) = |\mathcal{X}_i| \times |\mathcal{X}_j|$$

$B_{ii}$ : Identity matrix

$$B_{ij}(x_i; x_j) = \frac{P_{X_i X_j}(x_i, x_j)}{\sqrt{P_{X_i}(x_i)}\sqrt{P_{X_j}(x_j)}}$$

# Algorithm to Compute Singular Vectors

$$\begin{bmatrix} \psi_1 \\ \vdots \\ \psi_d \end{bmatrix} \leftarrow \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1d} \\ B_{21} & B_{22} & \cdots & B_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ B_{d1} & B_{d2} & \cdots & B_{dd} \end{bmatrix} \cdot \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_d \end{bmatrix}$$

| Power iteration algorithm | Multivariate alternative conditional expectation (MACE) algorithm |
|---|---|
| **Initialize:** Pick arbitrary vectors $\psi_i$ | **Initialize:** Pick arbitrary nonzero function $f_i$ |
| **Repeat:** $\psi_i \leftarrow \psi_i + \sum_{j \neq i} B_{ij} \psi_j$ | **Repeat:** $f_i(X_i) \leftarrow f_i(X_i) + \mathbb{E}\left[ \sum_{j \neq i} f_j(X_j) \middle| X_i \right]$ |
| **Regulate:** Scale $\psi_i \leftarrow \frac{\psi_i}{\lVert \psi \rVert}$ | **Regulate:** Scale $f_i(X_i) \leftarrow f_i(X_i) / \sqrt{\mathbb{E}\left[ \sum_{i=1}^{d} f_i^2(X_i) \right]}$ |



清华-伯克利深圳学院 TBSI Tsinghua-Berkeley Shenzhen Institute

# Compute Multiple Features

- A singular vector corresponds to a feature function.
  - The top $k$ eigenvectors => top $k$ informative features.
  - Solve $k$ eigenfunctions by MACE with Gram-Schmidt process.

**Algorithm 2** The Computation of $\vec{f}^{(k)}$

**Require** : The data samples of variables $X_1, \ldots, X_n$, and previously computed functions $\vec{f}^{(1)}, \ldots, \vec{f}^{(k-1)}$.

1. Initialization: randomly pick zero-mean functions $\vec{f}^{(k)} = (f_1^{(k)}, \ldots, f_d^{(k)})$.

**repeat** :

2a. $f_i^{(k)}(X_i) \leftarrow f_i^{(k)}(X_i) + \mathbb{E}\left[ \sum_{j \neq i} f_j^{(k)}(X_j) \middle| X_i \right]$.

2b. $f_i^{(k)}(X_i) \leftarrow f_i^{(k)}(X_i) / \sqrt{\mathbb{E}\left[ \sum_{i=1}^{d} (f_i^{(k)}(X_i))^2 \right]}$.

3. $\vec{f}^{(k)} \leftarrow \vec{f}^{(k)} - \sum_{m=1}^{k-1} \langle \vec{f}^{(m)}, \vec{f}^{(k)} \rangle \cdot \vec{f}^{(m)}$

**until** $\vec{f}^{(k)}$ converges.

# Extracting Common Bits Patterns

- Given a sequence of binary independent bits $\{b_1, \ldots, b_m\}$, $b_i=\{1,-1\}$, suppose each random variable $X_i$ is composed of a subset of bits.
  - For example: $X_1=\{b_1,b_2,b_3\}$, $X_2=\{b_1,b_2\}$, $X_3=\{b_1,b_3\}$, $X_4=\{b_1\}$

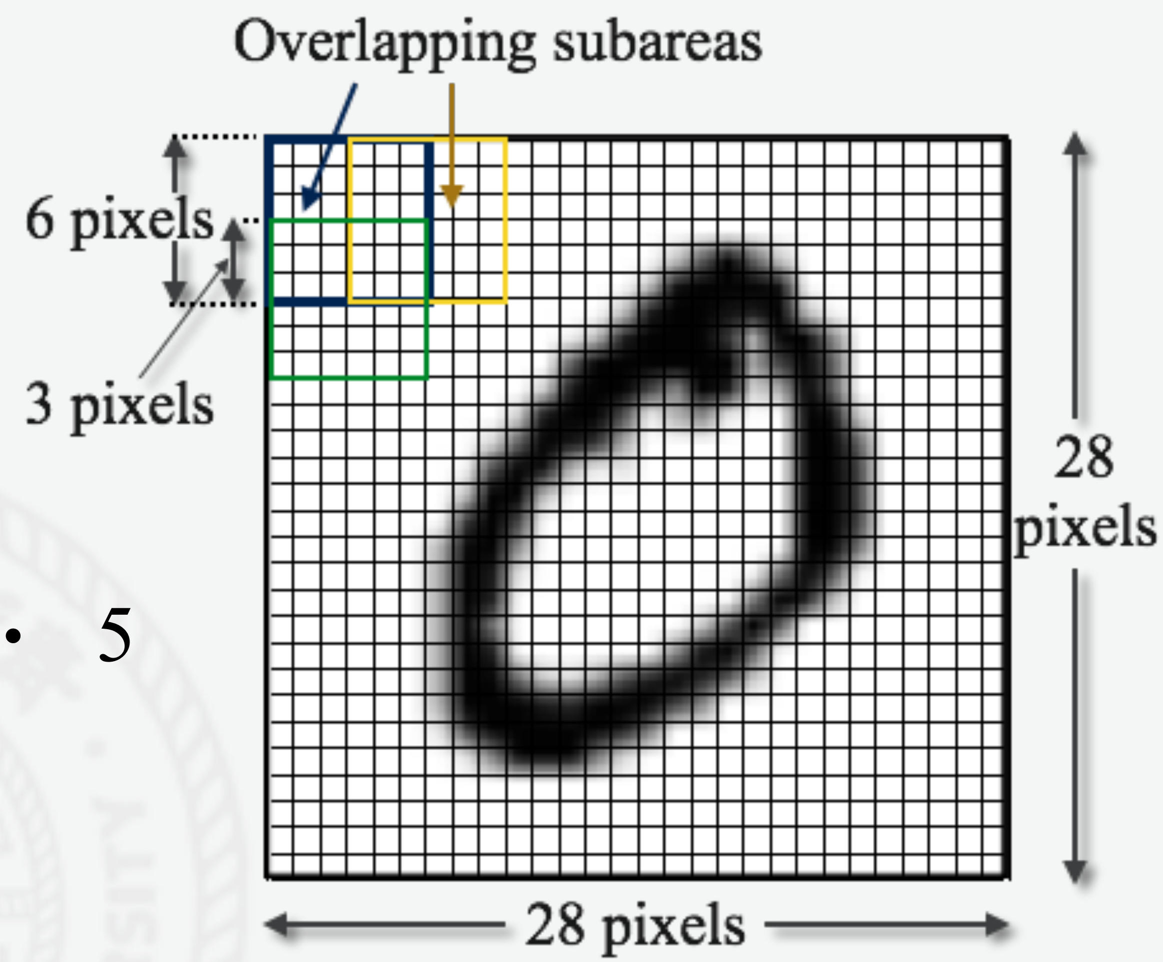
# Extracting Common Bits Patterns

- Given a sequence of binary independent bits $\{b_1, \ldots, b_m\}$, $b_i=\{1,-1\}$, suppose each random variable $X_i$ is composed of a subset of bits.
  - For example: $X_1=\{b_1,b_2,b_3\}$, $X_2=\{b_1,b_2\}$, $X_3=\{b_1,b_3\}$, $X_4=\{b_1\}$
  - The MACE algorithm counts and extracts the bit pattern that appears the most among the random variables.

$$f^{(1)}(X_1, X_2, X_3, X_4) = \sqrt{\lambda_1} b_1, \ \ f_i^{(1)}(X_i) = \frac{1}{\sqrt{\lambda_1}} b_1, \quad \text{eigenvalue } \lambda_1 = 4.$$

| Feature Function | $b_1$ | $b_2$ | $b_3$ | $b_1 \oplus b_2$ | $b_2 \oplus b_3$ | $b_1 \oplus b_3$ | $b_1 \oplus b_2 \oplus b_3$ |
|---|---|---|---|---|---|---|---|
| Eigen-value | 4 | 2 | 2 | 2 | 1 | 2 | 1 |

# MNIST Digits Recognition

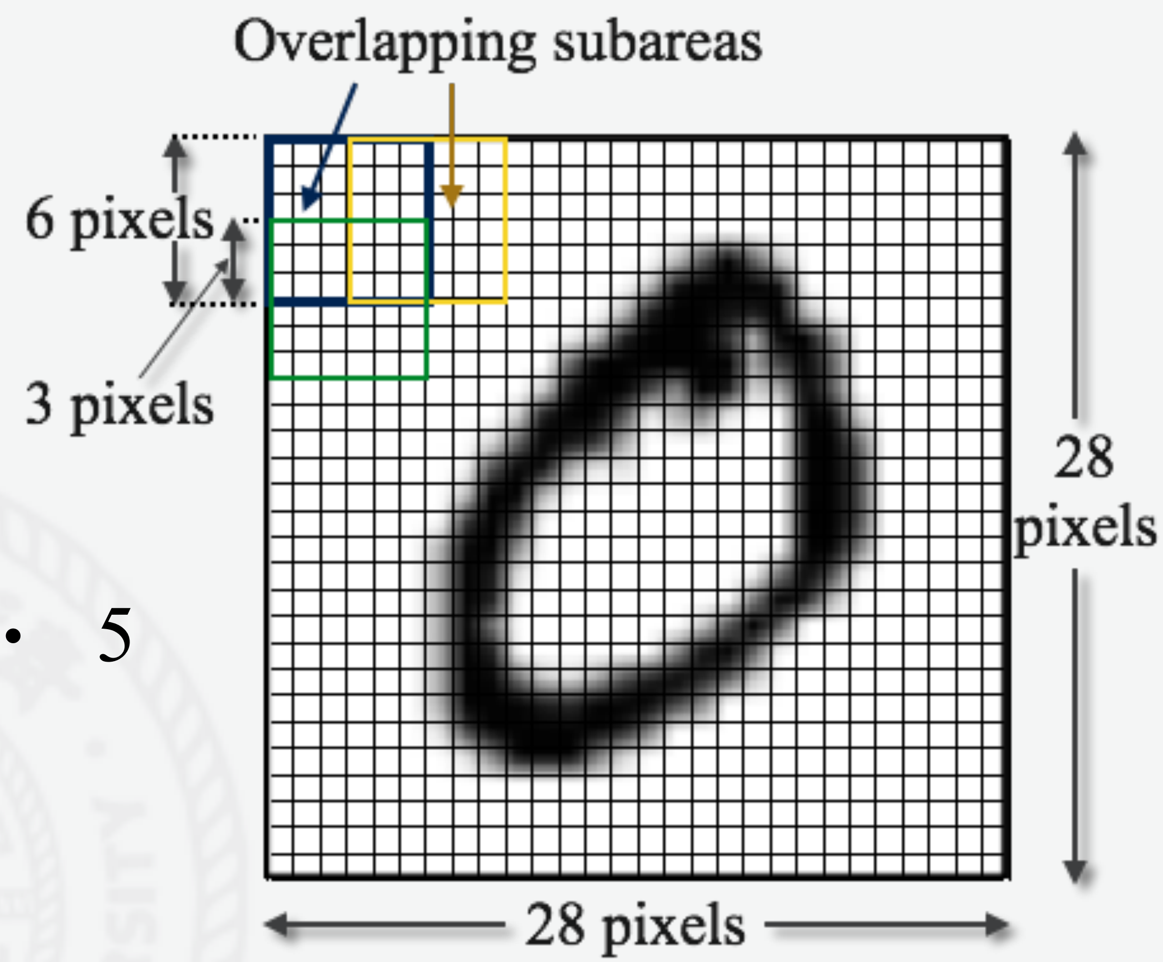



- 5

| $k$ | 4 | 8 | 12 | 16 | 20 | 24 |
|---|---|---|---|---|---|---|
| Error rate (%) | 4.74 | 2.44 | 2.36 | 2.21 | 2.15 | 2.08 |

- Divide into 8*8=64 subareas
  - Each with 6*6 pixels

# MNIST Digits Recognition

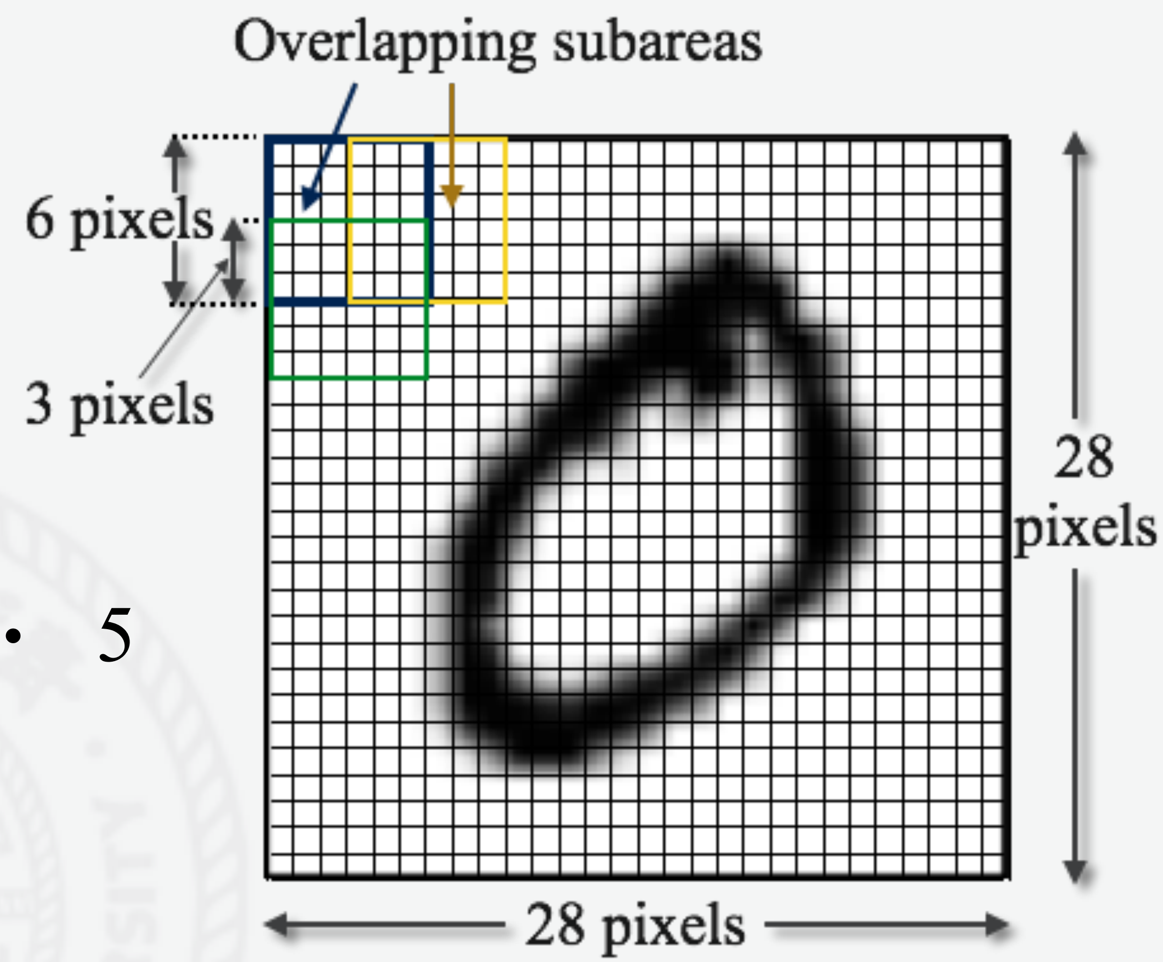



- 5

| $k$ | 4 | 8 | 12 | 16 | 20 | 24 |
|---|---|---|---|---|---|---|
| Error rate (%) | 4.74 | 2.44 | 2.36 | 2.21 | 2.15 | 2.08 |

- Divide into 8*8=64 subareas
  - Each with 6*6 pixels
- Quantize each subarea as a discrete random variable

# MNIST Digits Recognition




- 5

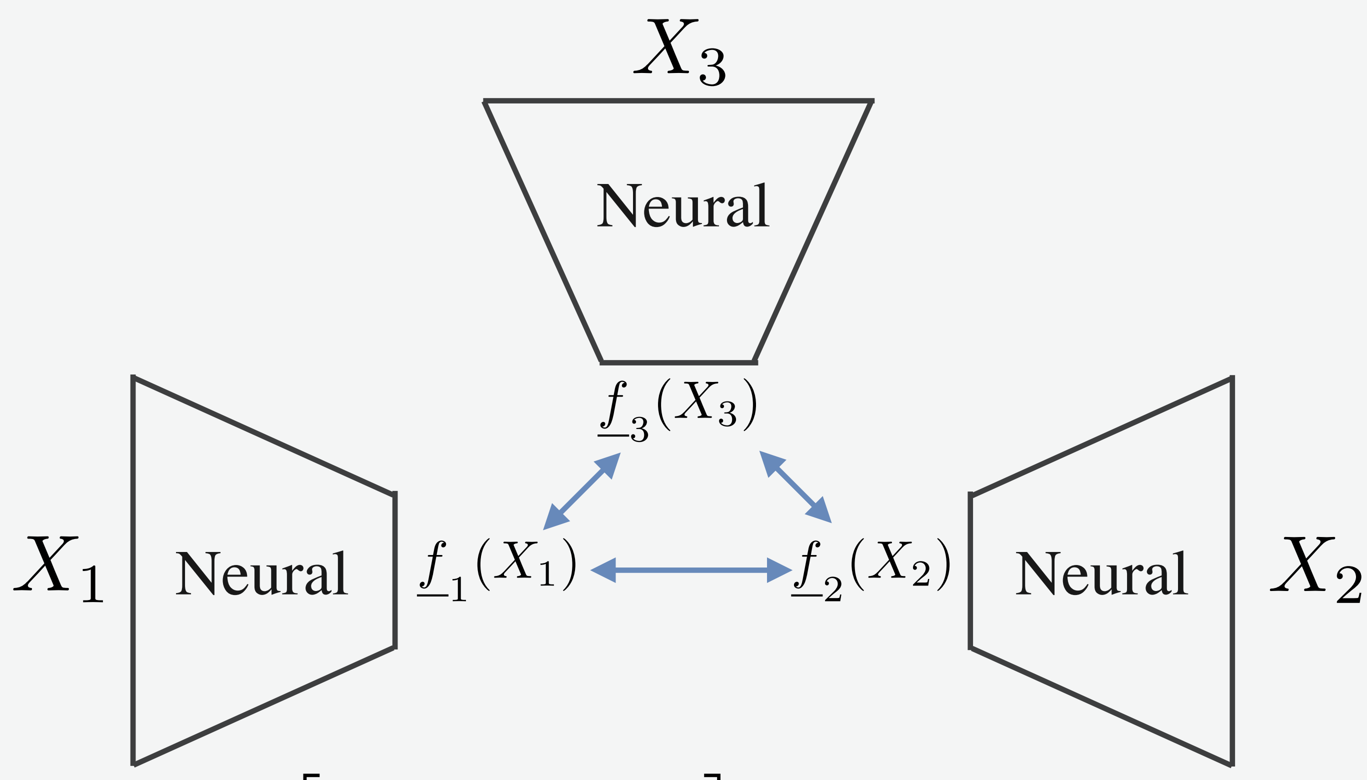
| *k* | 4 | 8 | 12 | 16 | 20 | 24 |
|---|---|---|---|---|---|---|
| Error rate (%) | 4.74 | 2.44 | 2.36 | 2.21 | 2.15 | 2.08 |

- Divide into 8*8=64 subareas
  - Each with 6*6 pixels
- Quantize each subarea as a discrete random variable
- Train the top *k* eigenvectors, extract 64*k*-dimensional features
  - Classify digits by the extracted features by SVM
  - Comparable to 3-layer NN
  - Extract features without labels



清华-伯克利深圳学院
Tsinghua-Berkeley Shenzhen Institute
TBSI

# Deep Common Structure Extraction




$$\max_{\underline{f}_i \in \mathbb{R}^k,\ i=1,\ldots,d} \mathbb{E}\left[\sum_{i\neq j} \underline{f}_i^{\mathrm{T}}(X_i)\,\underline{f}_j(X_j)\right] - \frac{1}{2}\sum_{i\neq j} \operatorname{trace}\left\{\operatorname{cov}\left(\underline{f}_i(X_i)\right)\operatorname{cov}\left(\underline{f}_j(X_i)\right)\right\}$$

- For continuous data, extract the common structure by deep neural networks, maximize the joint correlation.
  - Combining different types of data: images, texts, audios, …