On Estimation of Modal Decompositions

Anuran Makur, Gregory W. Wornell, and Lizhong Zheng

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

IEEE International Symposium on Information Theory 2020

Outline

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- Can we extend these techniques to categorical data?

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]
- **Maximal correlation:** [Geb41], [Rén59], [Wit75]

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]
- **Maximal correlation:** [Geb41], [Rén59], [Wit75]
- **Strong data processing inequalities and related directions:** χ^2-divergence [Sar58], KL divergence [AG76], and recent work on hypercontractivity [AGKN13], contraction coefficients [MZ15], [PW17], [MZ20], functional inequalities [Rag16], estimation theory, security, and privacy [CMM$^+$17], ...

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]
- **Maximal correlation:** [Geb41], [Rén59], [Wit75]
- **Strong data processing inequalities and related directions:** χ^2-divergence [Sar58], KL divergence [AG76], and recent work on hypercontractivity [AGKN13], contraction coefficients [MZ15], [PW17], [MZ20], functional inequalities [Rag16], estimation theory, security, and privacy [CMM$^+$17], ...
- **Lancaster distributions:** Mehler's decomposition [Meh66], Lancaster decompositions [Lan58], [Lan69], orthogonal polynomials [Eag64], [Gri69], [Kou96], [Kou98], and recent work [AZ12], [MZ17], ...

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]
- **Maximal correlation:** [Geb41], [Rén59], [Wit75]
- **Strong data processing inequalities and related directions:** χ^2-divergence [Sar58], KL divergence [AG76], and recent work on hypercontractivity [AGKN13], contraction coefficients [MZ15], [PW17], [MZ20], functional inequalities [Rag16], estimation theory, security, and privacy [CMM$^+$17], ...
- **Lancaster distributions:** Mehler's decomposition [Meh66], Lancaster decompositions [Lan58], [Lan69], orthogonal polynomials [Eag64], [Gri69], [Kou96], [Kou98], and recent work [AZ12], [MZ17], ...
- **Correspondence analysis:** Data visualization [Ben73], [Gre84], [GH87], and recent work on neural networks [HMWZ19], [HSC19], ...

A Brief History of Modal Decompositions

- **Dimensionality reduction:** Principal component analysis (PCA) [Pea01], [Hot33], canonical correlation analysis (CCA) [Hot36]
- **Modal decompositions:** [Hir35]
- **Maximal correlation:** [Geb41], [Rén59], [Wit75]
- **Strong data processing inequalities and related directions:** χ^2-divergence [Sar58], KL divergence [AG76], and recent work on hypercontractivity [AGKN13], contraction coefficients [MZ15], [PW17], [MZ20], functional inequalities [Rag16], estimation theory, security, and privacy [CMM$^+$17], ...
- **Lancaster distributions:** Mehler's decomposition [Meh66], Lancaster decompositions [Lan58], [Lan69], orthogonal polynomials [Eag64], [Gri69], [Kou96], [Kou98], and recent work [AZ12], [MZ17], ...
- **Correspondence analysis:** Data visualization [Ben73], [Gre84], [GH87], and recent work on neural networks [HMWZ19], [HSC19], ...
- **Non-parametric regression:** Alternating conditional expectations (ACE) algorithm [BF85], [Buj85], feature extraction [MKHZ15], [HMZW17], [HMWZ19]

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- Bivariate distribution $P_{X,Y}$ with marginals $P_X, P_Y > 0$

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- Bivariate distribution $P_{X,Y}$ with marginals $P_X, P_Y > 0$
- **Hilbert spaces:**
  Input space: $\mathcal{L}^2(\mathcal{X}, P_X) \triangleq \left\{f : \mathcal{X} \to \mathbb{R} \,\middle|\, \mathbb{E}\left[f(X)^2\right] < +\infty\right\}$ with inner product:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, P_X), \;\; \langle f_1, f_2 \rangle_{P_X} \triangleq \mathbb{E}[f_1(X) f_2(X)] = \sum_{x \in \mathcal{X}} P_X(x) f_1(x) f_2(x) \, ,$$

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- Bivariate distribution $P_{X,Y}$ with marginals $P_X, P_Y > 0$
- **Hilbert spaces:**
  Input space: $\mathcal{L}^2(\mathcal{X}, P_X) \triangleq \left\{ f : \mathcal{X} \rightarrow \mathbb{R} \,\middle|\, \mathbb{E}\left[f(X)^2\right] < +\infty \right\}$ with inner product:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, P_X), \ \ \langle f_1, f_2 \rangle_{P_X} \triangleq \mathbb{E}[f_1(X) f_2(X)] = \sum_{x \in \mathcal{X}} P_X(x) f_1(x) f_2(x) \,,$$

  and induced $\mathcal{L}^2$-norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, P_X), \ \ \|f\|_{P_X}^2 = \mathbb{E}\left[f(X)^2\right].$$

Formal Definitions

- Finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- Bivariate distribution $P_{X,Y}$ with marginals $P_X, P_Y > 0$
- **Hilbert spaces:**
  Input space: $\mathcal{L}^2(\mathcal{X}, P_X) \triangleq \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, \mathbb{E}\left[f(X)^2\right] < +\infty \right\}$ with inner product:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, P_X), \ \ \langle f_1, f_2 \rangle_{P_X} \triangleq \mathbb{E}[f_1(X) f_2(X)] = \sum_{x \in \mathcal{X}} P_X(x) f_1(x) f_2(x) \, ,$$

  and induced $\mathcal{L}^2$-norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, P_X), \ \ \|f\|_{P_X}^2 = \mathbb{E}\left[f(X)^2\right] .$$

  Output space: $\mathcal{L}^2(\mathcal{Y}, P_Y) \triangleq \left\{ g : \mathcal{Y} \to \mathbb{R} \,\middle|\, \mathbb{E}\left[g(Y)^2\right] < +\infty \right\}$

Formal Definitions: Two Equivalent Representations of $P_{X,Y}$

Definition (Conditional Expectation Operator)

$\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, P_X) \rightarrow \mathcal{L}^2(\mathcal{Y}, P_Y)$ maps any $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to $\mathbf{P}_{X|Y} f \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$\forall y \in \mathcal{Y}, \quad (\mathbf{P}_{X|Y} f)(y) \triangleq \mathbb{E}[f(X)|Y = y] \, .$$

Formal Definitions: Two Equivalent Representations of $P_{X,Y}$

Definition (Conditional Expectation Operator)

$\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, P_X) \rightarrow \mathcal{L}^2(\mathcal{Y}, P_Y)$ maps any $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to $\mathbf{P}_{X|Y} f \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$\forall y \in \mathcal{Y}, \;\; \left(\mathbf{P}_{X|Y} f\right)(y) \triangleq \mathbb{E}[f(X)|Y = y] \, .$$

Definition (Divergence Transition Matrix)

The divergence transition matrix (DTM), denoted $\mathbf{B} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$, has (y, x)th entry given by:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; B(x, y) \triangleq \frac{P_{X,Y}(x, y)}{\sqrt{P_X(x) P_Y(y)}} \, .$$

Formal Definitions: Two Equivalent Representations of $P_{X,Y}$

Definition (Conditional Expectation Operator)

$\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ maps any $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to $\mathbf{P}_{X|Y} f \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$\forall y \in \mathcal{Y}, \quad \left(\mathbf{P}_{X|Y} f\right)(y) \triangleq \mathbb{E}[f(X)|Y = y] .$$

Definition (Divergence Transition Matrix)

The divergence transition matrix (DTM), denoted $\mathbf{B} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$, has (y, x)th entry given by:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \quad B(x, y) \triangleq \frac{P_{X,Y}(x, y)}{\sqrt{P_X(x) P_Y(y)}} .$$

Remark: DTMs parallel symmetric normalized *Laplacian matrices*.

Formal Definitions: SVDs and Modal Decompositions

- $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$
- **SVD of Conditional Expectation Operator:**

$$\forall i \in \{0, \ldots, K-1\}, \;\; \mathbf{P}_{X|Y} f_i^* = \sigma_i \, g_i^*$$

  - $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1} \geq 0$ are singular values
  - $f_0^*, \ldots, f_{K-1}^* \in \mathcal{L}^2(\mathcal{X}, P_X)$ are orthonormal right singular vectors
  - $g_0^*, \ldots, g_{K-1}^* \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal left singular vectors

Formal Definitions: SVDs and Modal Decompositions

- $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$
- **SVD of Conditional Expectation Operator:**

$$\forall i \in \{0, \ldots, K-1\}, \quad \mathbf{P}_{X|Y} f_i^* = \sigma_i \, g_i^*$$

  - $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1} \geq 0$ are singular values
  - $f_0^*, \ldots, f_{K-1}^* \in \mathcal{L}^2(\mathcal{X}, P_X)$ are orthonormal right singular vectors
  - $g_0^*, \ldots, g_{K-1}^* \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal left singular vectors

- **SVD of DTM:**

$$\mathbf{B} = \sum_{i=0}^{K-1} \sigma_i \, \psi_i^Y \left(\psi_i^X\right)^{\mathrm{T}}$$

  - $\psi_0^X, \ldots, \psi_{K-1}^X \in \mathbb{R}^{|\mathcal{X}|}$ are orthonormal right singular vectors
  - $\psi_0^Y, \ldots, \psi_{K-1}^Y \in \mathbb{R}^{|\mathcal{Y}|}$ are orthonormal left singular vectors

Formal Definitions: SVDs and Modal Decompositions

- $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$
- **SVD of Conditional Expectation Operator:**

$$\forall i \in \{0, \dots, K-1\}, \;\; \mathbf{P}_{X|Y} f_i^* = \sigma_i \, g_i^*$$

  - $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1} \geq 0$ are singular values
  - $f_0^*, \dots, f_{K-1}^* \in \mathcal{L}^2(\mathcal{X}, P_X)$ are orthonormal right singular vectors
  - $g_0^*, \dots, g_{K-1}^* \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal left singular vectors

- **SVD of DTM:**

$$\mathbf{B} = \sum_{i=0}^{K-1} \sigma_i \, \psi_i^Y \left(\psi_i^X\right)^{\mathrm{T}}$$

  - $\psi_0^X, \dots, \psi_{K-1}^X \in \mathbb{R}^{|\mathcal{X}|}$ are orthonormal right singular vectors
  - $\psi_0^Y, \dots, \psi_{K-1}^Y \in \mathbb{R}^{|\mathcal{Y}|}$ are orthonormal left singular vectors

- **Relation:** $\psi_i^X(x) = f_i^*(x)\sqrt{P_X(x)}$ for $x \in \mathcal{X}$, and $\psi_i^Y(y) = g_i^*(y)\sqrt{P_Y(y)}$ for $y \in \mathcal{Y}$

Formal Definitions: SVDs and Modal Decompositions

Proposition (SVD Structure)

- **Operator Norm:** $\sigma_0 = 1$, $f_0^*(x) = 1$ for all $x \in \mathcal{X}$, and $g_0^*(y) = 1$ for all $y \in \mathcal{Y}$.

Formal Definitions: SVDs and Modal Decompositions

Proposition (SVD Structure)

- **Operator Norm:** $\sigma_0 = 1$, $f_0^*(x) = 1$ for all $x \in \mathcal{X}$, and $g_0^*(y) = 1$ for all $y \in \mathcal{Y}$.
- **Maximal Correlations** [Hir35, Geb41, Sar58, Rén59]**:** Using Courant-Fischer-Weyl,

$$\forall i \in \{1, \ldots, K-1\}, \ \ \sigma_i = \max_{f,g} \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f_i^*(X)g_i^*(Y)] \,,$$

where the maximization is over all $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ such that $\mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[g(Y)^2\right] = 1$ and $\mathbb{E}[f(X)f_j^*(X)] = \mathbb{E}[g(Y)g_j^*(Y)] = 0$ for all $j < i$.

Formal Definitions: SVDs and Modal Decompositions

Proposition (SVD Structure)

- **Operator Norm:** $\sigma_0 = 1$, $f_0^*(x) = 1$ for all $x \in \mathcal{X}$, and $g_0^*(y) = 1$ for all $y \in \mathcal{Y}$.
- **Maximal Correlations** [Hir35, Geb41, Sar58, Rén59]**:** Using Courant-Fischer-Weyl,

$$\forall i \in \{1, \dots, K-1\}, \;\; \sigma_i = \max_{f, g} \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f_i^*(X)g_i^*(Y)] \, ,$$

where the maximization is over all $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ such that $\mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[g(Y)^2\right] = 1$ and $\mathbb{E}[f(X)f_j^*(X)] = \mathbb{E}[g(Y)g_j^*(Y)] = 0$ for all $j < i$.

Proposition (Modal Decomposition of Bivariate Distribution [Hir35, Lan58, Ben73])

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_{X,Y}(x, y) = P_X(x)\, P_Y(y) \left(1 + \sum_{i=1}^{K-1} \sigma_i\, f_i^*(x)\, g_i^*(y) \right)$$

Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$\mathcal{X} = \Big\{$  ,  ,  , $\ldots \Big\}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$\mathcal{X} = \Big\{$  ,  ,  , $\ldots \Big\}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Goal: Embed $\mathcal{X}$ into $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g., clustering.

Motivation: Embedding of Categorical Data into Euclidean Space

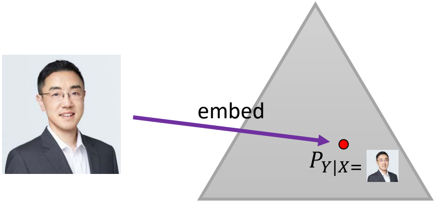
Suppose we have:

$\mathcal{X} = \{$  ,  ,  $, \ldots \}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Goal: Embed $\mathcal{X}$ into $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g., clustering.

"Natural" Embedding: Represent each $x \in \mathcal{X}$ using conditional distribution $P_{Y|X=x} \in \mathbb{R}^{|\mathcal{Y}|}$.



probability simplex

Motivation: Embedding of Categorical Data into Euclidean Space

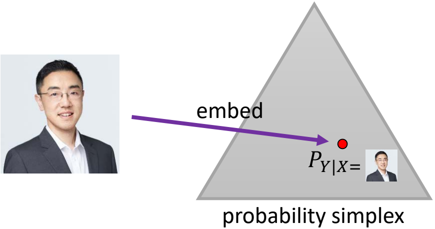
Suppose we have:

$\mathcal{X} = \{$  ,  ,  $, \ldots \}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Goal: Embed $\mathcal{X}$ into $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g., clustering.

"Natural" Embedding: Represent each $x \in \mathcal{X}$ using conditional distribution $P_{Y|X=x} \in \mathbb{R}^{|\mathcal{Y}|}$.



Dimensionality Reduction:

$|\mathcal{Y}|$ is large!

Reduce dimension of embedding.

Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$\mathcal{X} = \Big\{$  ,  ,  $, \ldots \Big\}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Want: Low-dimensional embedding of $\mathcal{X}$ into Euclidean space $\mathbb{R}^k$.

Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$\mathcal{X} = \{$  ,  ,  $, \ldots \}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Modal Decomposition Embedding:

$$P_{Y|X=x} = P_Y + \sum_{i=1}^{K-1} \sigma_i f_i^*(x) \, (g_i^* \cdot P_Y)$$

Motivation: Embedding of Categorical Data into Euclidean Space

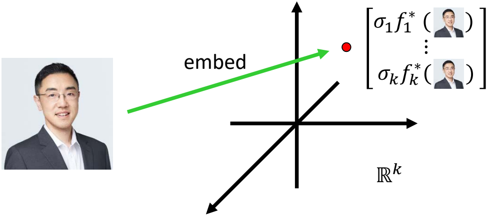
Suppose we have:

$\mathcal{X} = \{$  ,  ,  $, \ldots \}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Modal Decomposition Embedding: When σ_{k+1} is small,

$$\zeta_k : \mathcal{X} \to \mathbb{R}^k, \;\; \zeta_k(x) = \left[\sigma_1 f_1^*(x) \; \cdots \; \sigma_k f_k^*(x)\right]^T$$



Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$\mathcal{X} = \Big\{$  ,  ,  $, \ldots \Big\}$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Modal Decomposition Embedding: When σ_{k+1} is small,

$$\zeta_k : \mathcal{X} \to \mathbb{R}^k, \;\; \zeta_k(x) = \left[\sigma_1 f_1^*(x) \cdots \sigma_k f_k^*(x)\right]^T$$

Diffusion Distance Preservation (cf. Laplacian eigenmaps [BN01], diffusion maps [CL06]):

$$D_{\text{diff}}^2(P_{Y|X=x}, P_{Y|X=x'}) \triangleq \sum_{y \in \mathcal{Y}} \frac{\left(P_{Y|X}(y|x) - P_{Y|X}(y|x')\right)^2}{P_Y(y)} = \left\|\zeta_{K-1}(x) - \zeta_{K-1}(x')\right\|_2^2$$

Motivation: Embedding of Categorical Data into Euclidean Space

Suppose we have:

$$\mathcal{X} = \Big\{$$  ,  ,  $$, \ldots \Big\}$$

$$\mathcal{Y} = \{\text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots\}$$

Modal Decomposition Embedding: When σ_{k+1} is small,

$$\zeta_k : \mathcal{X} \to \mathbb{R}^k, \;\; \zeta_k(x) = \left[\sigma_1 f_1^*(x) \; \cdots \; \sigma_k f_k^*(x)\right]^T$$

Diffusion Distance Preservation (cf. Laplacian eigenmaps [BN01], diffusion maps [CL06]):

$$D_{\text{diff}}^2(P_{Y|X=x}, P_{Y|X=x'}) \triangleq \sum_{y \in \mathcal{Y}} \frac{\left(P_{Y|X}(y|x) - P_{Y|X}(y|x')\right)^2}{P_Y(y)} = \left\|\zeta_{K-1}(x) - \zeta_{K-1}(x')\right\|_2^2$$

$$\approx \left\|\zeta_k(x) - \zeta_k(x')\right\|_2^2 \quad \text{(dimensionality reduction when } k \ll K\text{)}$$

Main Questions

- How do we characterize or identify DTMs?

Main Questions

- How do we characterize or identify DTMs?
- Why do we use DTMs or conditional expectation operators to represent $P_{X,Y}$ (instead of, e.g., information density [HV93])?

Main Questions

- How do we characterize or identify DTMs?
- Why do we use DTMs or conditional expectation operators to represent $P_{X,Y}$ (instead of, e.g., information density [HV93])?
  Known relation to *mutual* χ^2*-information*, ...

Main Questions

- How do we characterize or identify DTMs?
- Why do we use DTMs or conditional expectation operators to represent $P_{X,Y}$ (instead of, e.g., information density [HV93])?
  Known relation to *mutual* χ^2*-information*, ...
- If true distribution $P_{X,Y}$ is *unknown* but we have *training data*, how well can we *learn* $\sigma_1, \ldots, \sigma_k$ and $(f_1^*, g_1^*), \ldots, (f_k^*, g_k^*)$?

Outline

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{$bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals$\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} =$ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{$bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals$\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} =$ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \rightarrow \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{$bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals$\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} =$ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Theorem (Characterization of DTMs)

- $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a distribution in $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} > \mathbf{0}$ (entry-wise) and the spectral norm $\|\mathbf{M}\|_{\mathrm{s}} = 1$:

$$\mathbf{B}(\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\right\}.$$

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{\text{bivariate distributions over } \mathcal{X}\times\mathcal{Y} \text{ with strictly positive marginals}\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} = $ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \rightarrow \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Theorem (Characterization of DTMs)

- $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a distribution in $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} > \mathbf{0}$ (entry-wise) and the spectral norm $\|\mathbf{M}\|_{\mathrm{s}} = 1$:

$$\mathbf{B}(\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\right\}.$$

- More generally, we have:

$$\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \exists\, \psi^X > \mathbf{0},\ \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\psi^X = \psi^X, \text{ and } \exists\, \psi^Y > \mathbf{0},\ \mathbf{M}\mathbf{M}^{\mathrm{T}}\psi^Y = \psi^Y\right\}.$$

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{$bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals$\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} =$ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Theorem (Characterization of DTMs)

- $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a distribution in $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} > \mathbf{0}$ (entry-wise) and the spectral norm $\|\mathbf{M}\|_{\mathrm{s}} = 1$:

$$\mathbf{B}(\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\right\}.$$

- More generally, we have:

$$\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \exists\, \psi^{X} > \mathbf{0},\ \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\psi^{X} = \psi^{X}, \text{ and } \exists\, \psi^{Y} > \mathbf{0},\ \mathbf{M}\mathbf{M}^{\mathrm{T}}\psi^{Y} = \psi^{Y}\right\}.$$

- DTM function is bijective and continuous.

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}} = \{$bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals$\}$
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}} =$ relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Theorem (Characterization of DTMs)

- $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a distribution in $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} > \mathbf{0}$ (entry-wise) and the spectral norm $\|\mathbf{M}\|_{\mathrm{s}} = 1$:

$$\mathbf{B}(\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\right\}.$$

- More generally, we have:

$$\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}) = \Big\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \exists\, \psi^X > \mathbf{0},\ \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\psi^X = \psi^X, \text{ and } \exists\, \psi^Y > \mathbf{0},\ \mathbf{M}\mathbf{M}^{\mathrm{T}}\psi^Y = \psi^Y\Big\}.$$

- DTM function is bijective and continuous. (So, $\mathbf{B}$ is equivalent representation of $P_{X,Y}$.)

Characterization of DTMs

- $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ = {bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with strictly positive marginals}
- $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ = relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$
- DTM function: $\mathbf{B} : \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ so that $\mathbf{B} = \mathbf{B}(P_{X,Y})$

Theorem (Characterization of DTMs)

- $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a distribution in $\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} > \mathbf{0}$ (entry-wise) and the spectral norm $\|\mathbf{M}\|_{\mathrm{s}} = 1$:

$$\mathbf{B}(\mathcal{P}_{\circ}^{\mathcal{X}\times\mathcal{Y}}) = \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\right\}.$$

- More generally, we have:

$$\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}) = \Big\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \exists\, \psi^X > \mathbf{0},\ \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\psi^X = \psi^X, \text{ and } \exists\, \psi^Y > \mathbf{0},\ \mathbf{M}\mathbf{M}^{\mathrm{T}}\psi^Y = \psi^Y\Big\}.$$

- DTM function is bijective and continuous. (So, $\mathbf{B}$ is equivalent representation of $P_{X,Y}$.)

- Proofs utilize *Perron-Frobenius theorem*.

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$ (This defines $P_{X,Y}$!)

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$ (This defines $P_{X,Y}$!)
- Instead of P_X, choose input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for any distribution $Q_X > \mathbf{0}$

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$ (This defines $P_{X,Y}$!)
- Instead of P_X, choose input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for any distribution $Q_X > \mathbf{0}$
- *Operator norm* of $\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is

$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X) \setminus \{\mathbf{0}\}} \frac{\left\|\mathbf{P}_{X|Y} f\right\|_{P_Y}}{\left\|f\right\|_{Q_X}}$$

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$
- Instead of P_X, choose input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for any distribution $Q_X > \mathbf{0}$
- *Operator norm* of $\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is

$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X) \setminus \{\mathbf{0}\}} \frac{\left\|\mathbf{P}_{X|Y} f\right\|_{P_Y}}{\left\|f\right\|_{Q_X}}$$

Theorem (Weak Contraction)

- $\min_{Q_X > \mathbf{0}} \left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} = \left\|\mathbf{P}_{X|Y}\right\|_{P_X \to P_Y} = 1.$
- For any $Q_X > \mathbf{0}$, $\left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y}^2 \geq 1 + \chi^2(P_X || Q_X) \triangleq \sum_{x \in \mathcal{X}} \frac{P_X(x)^2}{Q_X(x)}$.

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$
- Instead of P_X, choose input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for any distribution $Q_X > \mathbf{0}$
- *Operator norm* of $\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is

$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X) \backslash \{\mathbf{0}\}} \frac{\left\|\mathbf{P}_{X|Y} f\right\|_{P_Y}}{\left\|f\right\|_{Q_X}}$$

Theorem (Weak Contraction)

- $\min_{Q_X > \mathbf{0}} \left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} = \left\|\mathbf{P}_{X|Y}\right\|_{P_X \to P_Y} = 1.$
- For any $Q_X > \mathbf{0}$, $\left\|\mathbf{P}_{X|Y}\right\|^2_{Q_X \to P_Y} \geq 1 + \chi^2(P_X || Q_X) \triangleq \sum_{x \in \mathcal{X}} \frac{P_X(x)^2}{Q_X(x)}.$

- Proof uses explicit calculation of adjoint operator $\mathbf{P}^*_{X|Y}$.

Representation of Conditional Expectation Operators

Question: Why use conditional expectation operators with specific choices of Hilbert spaces?

- $\mathbf{P}_{X|Y}$ is characterized by $P_{X|Y}$
- To get SVD of $\mathbf{P}_{X|Y}$, choose output Hilbert space $\mathcal{L}^2(\mathcal{Y}, P_Y)$
- Instead of P_X, choose input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for any distribution $Q_X > \mathbf{0}$
- *Operator norm* of $\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is

$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X) \setminus \{\mathbf{0}\}} \frac{\left\|\mathbf{P}_{X|Y} f\right\|_{P_Y}}{\left\|f\right\|_{Q_X}}$$

Theorem (Weak Contraction)

- $\min_{Q_X > \mathbf{0}} \left\|\mathbf{P}_{X|Y}\right\|_{Q_X \to P_Y} = \left\|\mathbf{P}_{X|Y}\right\|_{P_X \to P_Y} = 1$. (*data processing inequality* for χ^2-divergence)
- For any $Q_X > \mathbf{0}$, $\left\|\mathbf{P}_{X|Y}\right\|^2_{Q_X \to P_Y} \geq 1 + \chi^2(P_X || Q_X) \triangleq \sum_{x \in \mathcal{X}} \frac{P_X(x)^2}{Q_X(x)}$.

- **Answer:** $Q_X^* = P_X$ is unique input Hilbert space that makes $\mathbf{P}_{X|Y}$ a *weak contraction*.

Outline

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \mathbf{1}_{Y_i = y} \, .$$

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \ \ \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \mathbf{1}_{Y_i = y} \, .$$

- Assume P_X and P_Y are known
  (e.g., high-dimensional regime $\max\{|\mathcal{X}|, |\mathcal{Y}|\} \ll n \ll |\mathcal{X}||\mathcal{Y}|$, or extra "unlabeled" data)

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \mathbf{1}_{Y_i = y} \, .$$

- Assume P_X and P_Y are known, and satisfy $P_X, P_Y \geq p_0$ for some constant $p_0 > 0$.

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \, \mathbf{1}_{Y_i = y} \, .$$

- Assume P_X and P_Y are known, and satisfy $P_X, P_Y \geq p_0$ for some constant $p_0 > 0$.
- **Sample Modal Decomposition:**

$$\forall x \in \mathcal{X}, \; \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) = P_X(x) \, P_Y(y) \left(1 + \sum_{i=1}^{K} \hat{\sigma}_i \, \hat{f}_i^*(x) \, \hat{g}_i^*(y) \right)$$

  - Singular value estimates: $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0$
  - $\{\hat{f}_1^*, \ldots, \hat{f}_K^*\} \subset \mathcal{L}^2(\mathcal{X}, P_X)$ and $\{\hat{g}_1^*, \ldots, \hat{g}_K^*\} \subset \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal sets

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \, \mathbf{1}_{Y_i = y} \, .$$

- Assume P_X and P_Y are known, and satisfy $P_X, P_Y \geq p_0$ for some constant $p_0 > 0$.
- **Sample Modal Decomposition:**

$$\forall x \in \mathcal{X}, \, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) = P_X(x) \, P_Y(y) \left(1 + \sum_{i=1}^{K} \hat{\sigma}_i \, \hat{f}_i^*(x) \, \hat{g}_i^*(y) \right)$$

  - Singular value estimates: $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0$
  - $\{\hat{f}_1^*, \ldots, \hat{f}_K^*\} \subset \mathcal{L}^2(\mathcal{X}, P_X)$ and $\{\hat{g}_1^*, \ldots, \hat{g}_K^*\} \subset \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal sets
- Singular vector estimates for all i: $\check{f}_i^*(x) \triangleq \hat{f}_i^*(x) - \mathbb{E}\left[\hat{f}_i^*(X)\right]$, $\check{g}_i^*(y) \triangleq \hat{g}_i^*(y) - \mathbb{E}\left[\hat{g}_i^*(Y)\right]$.

Preliminaries

- Suppose true $P_{X,Y}$ is unknown.
- Observe n training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ with empirical distribution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i = x} \mathbf{1}_{Y_i = y} \,.$$

- Assume P_X and P_Y are known, and satisfy $P_X, P_Y \geq p_0$ for some constant $p_0 > 0$.
- **Sample Modal Decomposition:**

$$\forall x \in \mathcal{X}, \, \forall y \in \mathcal{Y}, \;\; \hat{P}^n_{X,Y}(x,y) = P_X(x)\, P_Y(y) \left(1 + \sum_{i=1}^{K} \hat{\sigma}_i \, \hat{f}_i^*(x)\, \hat{g}_i^*(y)\right)$$

  - Singular value estimates: $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0$
  - $\{\hat{f}_1^*, \ldots, \hat{f}_K^*\} \subset \mathcal{L}^2(\mathcal{X}, P_X)$ and $\{\hat{g}_1^*, \ldots, \hat{g}_K^*\} \subset \mathcal{L}^2(\mathcal{Y}, P_Y)$ are orthonormal sets
- Singular vector estimates for all i: $\check{f}_i^*(x) \triangleq \hat{f}_i^*(x) - \mathbb{E}\big[\hat{f}_i^*(X)\big]$, $\check{g}_i^*(y) \triangleq \hat{g}_i^*(y) - \mathbb{E}\big[\hat{g}_i^*(Y)\big]$.
- **Algorithm:** Compute SVD of *empirical quasi-DTM* using, e.g., *orthogonal iteration* method, *QR iteration* algorithm (or ACE algorithm), *Krylov subspace* methods, etc.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Maximal Correlations

- Estimate $\sigma_1, \ldots, \sigma_k$ using $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ under (squared) ℓ^1-norm loss.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Maximal Correlations

- Estimate $\sigma_1, \ldots, \sigma_k$ using $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ under (squared) ℓ^1-norm loss.

Theorem (Sample Complexity Tail Bound I)

$$\forall\, 0 \leq \delta \leq \frac{1}{p_0}\sqrt{\frac{k}{2}}\,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{n\, p_0^2\, \delta^2}{8k}\right)$$

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Maximal Correlations

- Estimate $\sigma_1, \ldots, \sigma_k$ using $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ under (squared) ℓ^1-norm loss.

Theorem (Sample Complexity Tail Bound I)

$$\forall\, 0 \leq \delta \leq \frac{1}{p_0}\sqrt{\frac{k}{2}}, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{n\, p_0^2\, \delta^2}{8k}\right)$$

- To estimate $\sigma_1, \ldots, \sigma_k$ within fixed error and confidence, n must grow *linearly* with k.

Estimation of $k \in \{1, \dots, K-1\}$ Dominant Maximal Correlations

- Estimate $\sigma_1, \dots, \sigma_k$ using $\hat{\sigma}_1, \dots, \hat{\sigma}_k$ under (squared) ℓ^1-norm loss.

Theorem (Sample Complexity Tail Bound I)

$$\forall\, 0 \leq \delta \leq \frac{1}{p_0}\sqrt{\frac{k}{2}}, \quad \mathbb{P}\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{n\, p_0^2\, \delta^2}{8k}\right)$$

- To estimate $\sigma_1, \dots, \sigma_k$ within fixed error and confidence, n must grow *linearly* with k.
- Proof exploits: 1) vector generalization of *Bernstein's inequality*, and 2) weak *majorization* result for perturbation of singular values known as *Lidskii inequality*.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Maximal Correlations

- Estimate $\sigma_1, \ldots, \sigma_k$ using $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ under (squared) ℓ^1-norm loss.

Theorem (Sample Complexity Tail Bound I)

$$\forall\, 0 \leq \delta \leq \frac{1}{p_0}\sqrt{\frac{k}{2}}\,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{n\, p_0^2\, \delta^2}{8k}\right)$$

- To estimate $\sigma_1, \ldots, \sigma_k$ within fixed error and confidence, n must grow *linearly* with k.
- Proof exploits: 1) vector generalization of *Bernstein's inequality*, and 2) weak *majorization* result for perturbation of singular values known as *Lidskii inequality*.

Corollary (Squared ℓ^1-Risk Bound)

$$\forall\, n \geq 16\log(4kn)\,, \quad \mathbb{E}\left[\left(\sum_{i=1}^{k}\left|\hat{\sigma}_i - \sigma_i\right|\right)^2\right] \leq \frac{6k + 8k\log(nk)}{p_0^2\, n}$$

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq 0 \, .$$

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq 0 \, .$$

- First term equals $\sigma_1^2 + \cdots + \sigma_k^2$ ("rank k approximation" of mutual χ^2-information).

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq 0 \, .$$

- First term equals $\sigma_1^2 + \cdots + \sigma_k^2$.

Theorem (Sample Complexity Tail Bound II)

$$\forall\, 0 \leq \delta \leq 4k \, , \quad \mathbb{P}\left(\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq \delta \right) \leq \left(|\mathcal{X}| + |\mathcal{Y}| \right) \exp\left(-\frac{n \, p_0 \, \delta^2}{64 \, k^2} \right)$$

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq 0 \,.$$

- First term equals $\sigma_1^2 + \cdots + \sigma_k^2$.

Theorem (Sample Complexity Tail Bound II)

$$\forall\, 0 \leq \delta \leq 4k\,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq \delta \right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(- \frac{n\, p_0\, \delta^2}{64\, k^2} \right)$$

- To estimate $f_1^*, \ldots, f_k^*$ within fixed error and confidence, n must be *quadratic* in k.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \geq 0\,.$$

- First term equals $\sigma_1^2 + \cdots + \sigma_k^2$.

Theorem (Sample Complexity Tail Bound II)

$$\forall\, 0 \leq \delta \leq 4k\,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \geq \delta\right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{n\, p_0\, \delta^2}{64\, k^2}\right)$$

- To estimate $f_1^*, \ldots, f_k^*$ within fixed error and confidence, n must be *quadratic* in k.
- Proof exploits: 1) *matrix* generalization of Bernstein's inequality, and 2) singular value stability result known as *Weyl inequality*.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

- Estimate $f_1^*, \ldots, f_k^*$ using $\check{f}_1^*, \ldots, \check{f}_k^*$ under loss function:

$$\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq 0 \,.$$

- First term equals $\sigma_1^2 + \cdots + \sigma_k^2$.

Theorem (Sample Complexity Tail Bound II)

$$\forall\, 0 \leq \delta \leq 4k \,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left\| \mathbf{P}_{X|Y} f_i^* \right\|_{P_Y}^2 - \left\| \mathbf{P}_{X|Y} \check{f}_i^* \right\|_{P_Y}^2 \geq \delta \right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{n \, p_0 \, \delta^2}{64 \, k^2} \right)$$

- To estimate $f_1^*, \ldots, f_k^*$ within fixed error and confidence, n must be *quadratic* in k.
- Proof exploits: 1) *matrix* generalization of Bernstein's inequality, and 2) singular value stability result known as *Weyl inequality*.
- Analogous results hold for estimation of $g_1^*, \ldots, g_k^*$ using $\check{g}_1^*, \ldots, \check{g}_k^*$.

Estimation of $k \in \{1, \ldots, K-1\}$ Dominant Feature Functions

Theorem (Sample Complexity Tail Bound II)

$$\forall\, 0 \leq \delta \leq 4k\,, \quad \mathbb{P}\left(\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \geq \delta\right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{n\, p_0\, \delta^2}{64\, k^2}\right)$$

Corollary (Mean Squared Error Risk Bound)

For every sufficiently large n such that $\frac{p_0 n}{64} \geq \frac{1}{|\mathcal{X}|+|\mathcal{Y}|}$ and $\frac{p_0 n}{4} \geq \log\left(\frac{p_0 n}{64}(|\mathcal{X}| + |\mathcal{Y}|)\right)$,

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2\right)^2\right] \leq \frac{64k^2\left(\log\left(p_0 n(|\mathcal{X}| + |\mathcal{Y}|)\right) - 3\right)}{p_0\, n}$$

Outline

Conclusion

Main Contributions:

- DTMs are entry-wise strictly positive matrices with spectral norm 1.
- Unique Hilbert spaces yield conditional expectation operators that are weak contractions.
- Sample complexity bounds for learning modal decompositions from training data.

Conclusion

Main Contributions:

- DTMs are entry-wise strictly positive matrices with spectral norm 1.
- Unique Hilbert spaces yield conditional expectation operators that are weak contractions.
- Sample complexity bounds for learning modal decompositions from training data.

Main Future Direction:

- Sharpen and generalize sample complexity results using *matrix estimation* ideas.

Thank You!