Principle of Minimum Renyi Correlation: from Marginals to Joint Distribution

Presenter: Farzan Farnia

Stanford University

Information Theory, Learning, Big Data Workshop
Simons Institute
March 17th, 2015

- Based on a joint work with

Meisam Razaviyayn
Stanford

Sreeram Kannan
U of Washington

David Tse
Stanford

Introduction

- **Example:** Genome-Wide Association Studies (GWAS)
  - SNPs $X_i \in \{0, 1, 2\}$

	X_1	X_2	....	$X_{3\times10^6}$
Indiv. 1	1	2		2
Indiv. 2	2	0		1
			$\vdots$	
Indiv. 1000	0	1		2

Introduction

- **Example:** Genome-Wide Association Studies (GWAS)
  - SNPs $X_i \in \{0, 1, 2\}$

	X_1	X_2	....	$X_{3\times10^6}$
Indiv. 1	1	2		2
Indiv. 2	2	0		1
			⋮	
Indiv. 1000	0	1		2

- **Question:** How to model this data by a joint distribution?

Introduction

- Not enough data to estimate ground-truth distribution

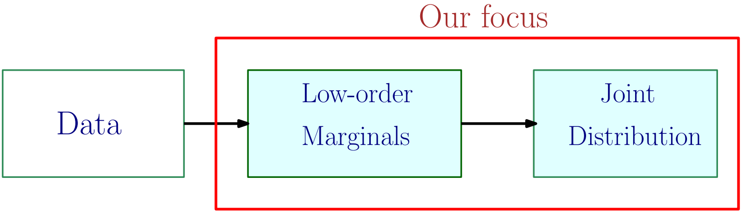
$$\# \text{ SNP sequences: } 3^{3,000,000} \approx 10^{1,400,000}$$
$$\# \text{ atoms in the universe } \approx 4 \times 10^{81}$$

Introduction

- Not enough data to estimate ground-truth distribution
- Low order marginals characterize a class of joint distributions $\mathcal{C}$

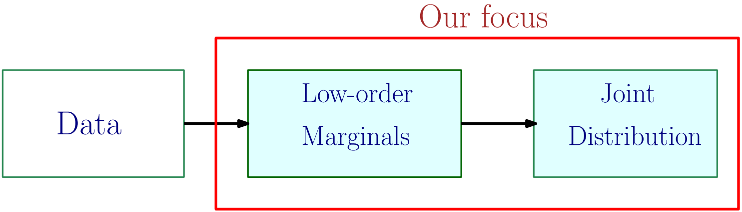
Introduction

- Not enough data to estimate ground-truth distribution
- Low order marginals characterize a class of joint distributions $\mathcal{C}$



Introduction

- Not enough data to estimate ground-truth distribution
- Low order marginals characterize a class of joint distributions $\mathcal{C}$



- **Principle of maximum entropy**: pick the distribution maximizing *Shannon entropy* as a measure of *uncertainty*,

$$\operatorname{argmax}_{\mathbb{P} \in \mathcal{C}} \; \mathrm{H}(\mathbb{P})$$

Introduction

- **Example:** Genome-Wide Association Studies for a particular trait
  - SNPs X_i's, trait Y

	X_1	X_2	....	$X_{3\times 10^6}$	Y
Indiv. 1	1	2		2	1
Indiv. 2	2	0		1	0
			⋮		
Indiv. 1000	0	1		2	1

Introduction

- **Example:** Genome-Wide Association Studies for a particular trait
  - SNPs X_i's, trait Y

	X_1	X_2	⋯⋯⋯	$X_{3\times10^6}$	Y
Indiv. 1	1	2		2	1
Indiv. 2	2	0		1	0
			⋮		
Indiv. 1000	0	1		2	1

- More sensible to minimize $\mathbf{X}$, Y dependence

$$\underset{\mathbb{P}_{\mathbf{X},Y}\in\mathcal{C}}{\operatorname{argmin}} \; \mathrm{D}_{\mathbb{P}}(\mathbf{X};Y)$$

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

- *Mutual information* $I(X;Y)$ [Globerson & Tishby, 2004]:

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

- *Mutual information* $I(X;Y)$ [Globerson & Tishby, 2004]:
  - No efficient algorithms available for its computation

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

- *Mutual information* $I(X;Y)$ [Globerson & Tishby, 2004]:
  - No efficient algorithms available for its computation
- *Pearson correlation coefficient* $\rho(X,Y)$:

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

- *Mutual information* $I(X;Y)$ [Globerson & Tishby, 2004]:
  - No efficient algorithms available for its computation
- *Pearson correlation coefficient* $\rho(X,Y)$:
  - not label-invariant

Principle of minimum Renyi correlation

Question: Which measure of dependence to minimize?

- *Mutual information* $I(X; Y)$ [Globerson & Tishby, 2004]:
  - No efficient algorithms available for its computation
- *Pearson correlation coefficient* $\rho(X, Y)$:
  - not label-invariant
- Renyi maximal correlation

$$\mathcal{R}(X; Y) = \max_{f,g} \rho(f(X), g(Y))$$

Principle of minimum Renyi correlation

- Renyi maximal correlation $\mathcal{R}(X;Y) = \max_{f,g} \rho(f(X), g(Y))$:
  - $0 \le \mathcal{R}(X;Y) \le 1$

Principle of minimum Renyi correlation

- Renyi maximal correlation $\mathcal{R}(X;Y) = \max_{f,g} \rho(f(X), g(Y))$:
  - $0 \leq \mathcal{R}(X;Y) \leq 1$
  - $\mathcal{R}(X;Y) = 0$ if and only if X and Y are independent
  - $\mathcal{R}(X;Y) = 1$ if $\exists h: \; Y = h(X)$

Principle of minimum Renyi correlation

- Renyi maximal correlation $\mathcal{R}(X; Y) = \max_{f,g} \rho(f(X), g(Y))$:
  - $0 \leq \mathcal{R}(X; Y) \leq 1$
  - $\mathcal{R}(X; Y) = 0$ if and only if X and Y are independent
  - $\mathcal{R}(X; Y) = 1$ if $\exists h : \; Y = h(X)$
  - $\mathcal{R}(X; Y) = |\rho(X; Y)|$ if (X, Y) are jointly Gaussian

Principle of minimum Renyi correlation

- Renyi maximal correlation $\mathcal{R}(X;Y) = \max_{f,g}\ \rho(f(X), g(Y)):$
  - $0 \leq \mathcal{R}(X;Y) \leq 1$
  - $\mathcal{R}(X;Y) = 0$ if and only if X and Y are independent
  - $\mathcal{R}(X;Y) = 1$ if $\exists h:\ Y = h(X)$
  - $\mathcal{R}(X;Y) = |\rho(X;Y)|$ if (X, Y) are jointly Gaussian

Principle of minimum Renyi correlation

What about *principle of minimum Renyi correlation*?

$$\underset{\mathbb{P}_{\mathbf{X},Y}\in\mathcal{C}}{\text{argmin}} \;\; \mathcal{R}_{\mathbb{P}}(\mathbf{X}; Y)$$

Principle of minimum Renyi correlation

What about *principle of minimum Renyi correlation*?

$$\underset{\mathbb{P}_{\mathbf{X},Y} \in \mathcal{C}}{\text{argmin}} \ \mathcal{R}_{\mathbb{P}}(\mathbf{X}; Y)$$

- Analytic structure of the minimizer

Principle of minimum Renyi correlation

What about *principle of minimum Renyi correlation*?

$$\underset{\mathbb{P}_{\mathbf{X},Y}\in\mathcal{C}}{\operatorname{argmin}} \ \mathcal{R}_{\mathbb{P}}(\mathbf{X};Y)$$

- Analytic structure of the minimizer
- Computation of the minimizing distribution

Minimum Renyi correlation distribution: continuous setting

- Real-valued $X_1, X_2, \ldots, X_p,\ Y$

Minimum Renyi correlation distribution: continuous setting

- Real-valued $X_1, X_2, \ldots, X_p,\ Y$
- Given
  - First order moment, $\mathbb{E}[(\mathbf{X}\ Y)] = \boldsymbol{\mu} \in \mathbb{R}^{p+1}$
  - Second order moment, $\mathbb{E}[(\mathbf{X}\ Y)^T(\mathbf{X}\ Y)] = \boldsymbol{\Lambda} \in \mathbb{R}^{(p+1)\times(p+1)}$

Minimum Renyi correlation distribution: continuous setting

- Real-valued $X_1, X_2, \ldots, X_p,\ Y$
- Given
  - First order moment, $\mathbb{E}[(\mathbf{X}\ Y)] = \boldsymbol{\mu} \in \mathbb{R}^{p+1}$
  - Second order moment, $\mathbb{E}[(\mathbf{X}\ Y)^T(\mathbf{X}\ Y)] = \boldsymbol{\Lambda} \in \mathbb{R}^{(p+1)\times(p+1)}$

Theorem 1

Jointly Gaussian minimizes Renyi correlation.

Minimum Renyi correlation distribution: continuous setting

- Real-valued $X_1, X_2, \ldots, X_p,\ Y$
- Given
  - First order moment, $\mathbb{E}[(\mathbf{X}\ Y)] = \boldsymbol{\mu} \in \mathbb{R}^{p+1}$
  - Second order moment, $\mathbb{E}[(\mathbf{X}\ Y)^T(\mathbf{X}\ Y)] = \boldsymbol{\Lambda} \in \mathbb{R}^{(p+1)\times(p+1)}$

Theorem 1

Jointly Gaussian minimizes Renyi correlation.

Key reason: linearity of conditional expectation

$$\mathbb{E}[Y|X_1, \ldots, X_p] = \sum_{i=0}^{p} c_i X_i$$

Minimum Renyi correlation distribution: discrete setting

- Discrete $X_i \in \{1, 2, \ldots, m\}$ and $Y \in \{-1, +1\}$

Minimum Renyi correlation distribution: discrete setting

- Discrete $X_i \in \{1, 2, \ldots, m\}$ and $Y \in \{-1, +1\}$
- Given second order marginals characterizing class $\mathcal{C}$:
  - $\Pr(X_i = x, X_j = u)$, $\Pr(X_i = x, Y = y)$

Minimum Renyi correlation distribution: discrete setting

- Discrete $X_i \in \{1, 2, \ldots, m\}$ and $Y \in \{-1, +1\}$
- Given second order marginals characterizing class $\mathcal{C}$:
  - $\Pr(X_i = x, X_j = u)$, $\Pr(X_i = x, Y = y)$

Theorem 2

If there exists $\mathbb{P} \in \mathcal{C}$ with a separable conditional expectation,

$$\mathbb{E}_{\mathbb{P}}\left[Y \,\middle|\, X_1, \ldots X_p\right] = \sum_i \gamma_i(X_i)$$

$\mathbb{P}$ will minimize Renyi correlation.

Renyi minimizer distribution computation: Recipe

- Define $\mathbf{X}_{\mathrm{I}}$ as vector of indicator variables w.r.t. $\mathbf{X}$:

$$\mathbf{X}_{\mathrm{I}\, mi+j} = \begin{cases} 1 & \text{if } \mathbf{X}_i = j \\ 0 & \text{otherwise} \end{cases}$$

Renyi minimizer distribution computation: Recipe

- Define $\mathbf{X}_{\mathrm{I}}$ as vector of indicator variables w.r.t. $\mathbf{X}$:

$$\mathbf{X}_{\mathrm{I}\,mi+j} = \begin{cases} 1 & \text{if } \mathbf{X}_i = j \\ 0 & \text{otherwise} \end{cases}$$

	X_1	....	Y
Indiv. 1	1		1
Indiv. 2	2		0
		⋮	
Indiv. n	0		1

	X_{I_1}	X_{I_2}	X_{I_3}	....	Y
Indiv. 1	0	1	0		1
Indiv. 2	0	0	1		0
				⋮	
Indiv. n	1	0	0		1

Renyi minimizer distribution computation: Recipe

- Find minimizer $\mathbf{z}^*$ (*linear regression* on indicator variables):

$$\mathbf{z}^* \in \underset{\mathbf{z}}{\operatorname{argmin}} \;\; \mathbb{E}\left[(\mathbf{X}_{\mathrm{I}}^T \mathbf{z} - Y)^2\right] \; = \; \underset{\mathbf{z}}{\operatorname{argmin}} \;\; \mathbf{z}^T \mathbf{Q} \mathbf{z} + \mathbf{f}^T \mathbf{z} + 1$$

Renyi minimizer distribution computation: Recipe

- Find minimizer $\mathbf{z}^*$ (*linear regression* on indicator variables):

$$\mathbf{z}^* \in \underset{\mathbf{z}}{\operatorname{argmin}} \ \mathbb{E}\left[(\mathbf{X}_{\mathrm{I}}^T \mathbf{z} - Y)^2\right] = \underset{\mathbf{z}}{\operatorname{argmin}} \ \mathbf{z}^T \mathbf{Q} \mathbf{z} + \mathbf{f}^T \mathbf{z} + 1$$

- Define h as

$$h(X_1, \ldots X_p) = \frac{1}{2}(1 + \mathbf{z}^{*T} \mathbf{X}_{\mathrm{I}})$$

- If $\mathbb{P}$ with separable conditional expectation exists

$$\mathbb{P}(Y = 1 \mid X_1, \ldots X_p) = h(X_1, \ldots X_p)$$

Renyi minimizer distribution computation: Recipe

- Find minimizer $\mathbf{z}^*$ (*linear regression* on indicator variables):

$$\mathbf{z}^* \in \underset{\mathbf{z}}{\text{argmin}} \ \mathbb{E}\left[(\mathbf{X}_{\mathrm{I}}^T \mathbf{z} - Y)^2\right] = \underset{\mathbf{z}}{\text{argmin}} \ \mathbf{z}^T \mathbf{Q} \mathbf{z} + \mathbf{f}^T \mathbf{z} + 1$$

- Define h as

$$h(X_1, \ldots X_p) = \frac{1}{2}(1 + {\mathbf{z}^*}^T \mathbf{X}_{\mathrm{I}})$$

- If $\mathbb{P}$ with separable conditional expectation exists

$$\mathbb{P}(Y = 1 \,|\, X_1, \ldots X_p) = h(X_1, \ldots X_p)$$

Observation: necessary condition for existence of separable $\mathbb{P}$:

$$\forall x_1, \ldots, x_p : \quad 0 \leq h(x_1, \ldots x_p) \leq 1$$

Minimum Renyi correlation distribution: discrete setting

Question: How to check whether this condition holds?

Minimum Renyi correlation distribution: discrete setting

Question: How to check whether this condition holds?

Theorem 3

Under the marginals consistency assumption, separable $\mathbb{P}$ exists if and only if for a minimizer $\mathbf{z}^*$ of

$$\min_{\mathbf{z}} \ \mathbb{E}\left[(\mathbf{X}_I^T\mathbf{z} - Y)^2\right]$$

the separable function

$$h(X_1, \ldots X_p) = \frac{1}{2}(1 + \mathbf{z}^{*T}\mathbf{X}_{\mathrm{I}})$$

satisfies

$$\forall x_1, \ldots, x_p : \quad 0 \le h(x_1, \ldots x_p) \le 1$$

Minimum Renyi correlation distribution: discrete setting

Question: How to check whether this condition holds?

Theorem 3

Under the marginals consistency assumption, separable $\mathbb{P}$ exists if and only if for a minimizer $\mathbf{z}^*$ of

$$\min_{\mathbf{z}} \; \mathbb{E}\left[(\mathbf{X}_I^T\mathbf{z} - Y)^2\right]$$

the separable function

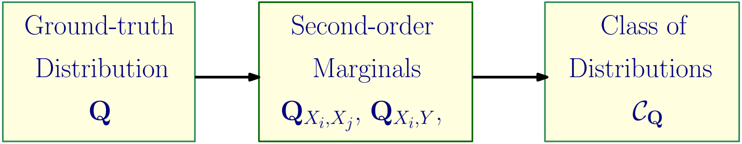
$$h(X_1, \ldots X_p) = \frac{1}{2}(1 + \mathbf{z}^{*T}\mathbf{X}_{\mathrm{I}})$$

satisfies

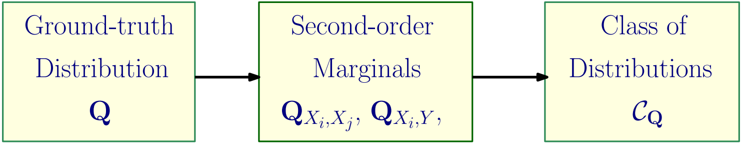
$$\forall x_1, \ldots, x_p : \quad 0 \le h(x_1, \ldots x_p) \le 1$$

Since h is separable, this condition can be checked in $O(mp)$.

Minimum Renyi correlation distribution: discrete setting

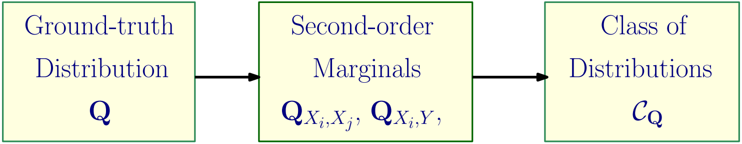


Minimum Renyi correlation distribution: discrete setting



Question: How large is the subset of $\mathbb{Q}$'s for which $\mathcal{C}_{\mathbb{Q}}$ satisfies the condition?

Minimum Renyi correlation distribution: discrete setting



Question: How large is the subset of $\mathbb{Q}$'s for which $\mathcal{C}_{\mathbb{Q}}$ satisfies the condition?

Theorem 4

For $\mathbb{P}$ uniform, there is an $\epsilon > 0$ such that for any $\mathbb{Q}$ in the ϵ-distance from $\mathbb{P}$, $\mathcal{C}_{\mathbb{Q}}$ contains a distribution with separable conditional expectation.

Variable Selection

Variable Selection

- Select a subset of features with highest correlation with the target

$$\max_{|\mathcal{S}| \leq k} \quad \mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y)$$

Variable Selection

- Select a subset of features with highest correlation with the target

$$\max_{|\mathcal{S}|\leq k} \quad \mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y)$$

- Lower bound for Renyi correlation

$$\mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y) \geq \sqrt{1 - \min_{\mathbf{z}} \mathbb{E}\left[(\mathbf{X}_{I,\mathcal{S}}^T \mathbf{z} - Y)^2\right]}$$

Tight under the additive structure assumption.

Variable Selection

- Select a subset of features with highest correlation with the target

$$\max_{|\mathcal{S}|\leq k} \quad \mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y)$$

- Lower bound for Renyi correlation

$$\mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y) \geq \sqrt{1 - \min_{\mathbf{z}} \mathbb{E}\left[(\mathbf{X}_{I,\mathcal{S}}^T \mathbf{z} - Y)^2\right]}$$

Tight under the additive structure assumption.

$$\max_{|\mathcal{S}|\leq k} \quad \sqrt{1 - \min_{\mathbf{z}} \mathbb{E}\left[(\mathbf{X}_{I,\mathcal{S}}^T \mathbf{z} - Y)^2\right]}$$

Variable Selection

- Select a subset of features with highest correlation with the target

$$\max_{|\mathcal{S}|\leq k} \quad \mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y)$$

- Lower bound for Renyi correlation

$$\mathcal{R}(\mathbf{X}_{\mathcal{S}}, Y) \geq \sqrt{1 - \min_{\mathbf{z}} \mathbb{E}\left[(\mathbf{X}_{I,\mathcal{S}}^T \mathbf{z} - Y)^2\right]}$$

Tight under the additive structure assumption.

$$\max_{|\mathcal{S}|\leq k} \quad \sqrt{1 - \min_{\mathbf{z}} \mathbb{E}\left[(\mathbf{X}_{I,\mathcal{S}}^T \mathbf{z} - Y)^2\right]}$$

Variable Selection: LASSO

- Using empirical average, equivalent to

$$\min_{\mathbf{z}} \ \|\mathbf{A}\mathbf{z} - \mathbf{b}\|_2^2$$
$$\text{s.t.} \quad card(\mathbf{z}) \leq k$$

where **A**, **b** sample indicator variables matrix and response

Variable Selection: LASSO

- Using empirical average, equivalent to

$$\min_{\mathbf{z}} \ \|\mathbf{A}\mathbf{z} - \mathbf{b}\|_2^2$$
$$\text{s.t.} \quad card(\mathbf{z}) \leq k$$

where $\mathbf{A}$, $\mathbf{b}$ sample indicator variables matrix and response

- Justification of *group lasso* for feature selection

Summary

- We introduced principle of minimum Renyi correlation as a counterpart for principles of MaxEnt and MinMI.

Summary

- We introduced principle of minimum Renyi correlation as a counterpart for principles of MaxEnt and MinMI.
- In continuous case, jointly Gaussian is a minimizer under fixed first and second order moments

Summary

- We introduced principle of minimum Renyi correlation as a counterpart for principles of MaxEnt and MinMI.
- In continuous case, jointly Gaussian is a minimizer under fixed first and second order moments
- There exists a certain separable structure in discrete minimizing distributions for given first and second order marginals

Summary

- We introduced principle of minimum Renyi correlation as a counterpart for principles of MaxEnt and MinMI.
- In continuous case, jointly Gaussian is a minimizer under fixed first and second order moments
- There exists a certain separable structure in discrete minimizing distributions for given first and second order marginals
- We can compute conditional minimizing distribution by solving a linear regression problem.

Summary

- We introduced principle of minimum Renyi correlation as a counterpart for principles of MaxEnt and MinMI.
- In continuous case, jointly Gaussian is a minimizer under fixed first and second order moments
- There exists a certain separable structure in discrete minimizing distributions for given first and second order marginals
- We can compute conditional minimizing distribution by solving a linear regression problem.
- Principle of minimum Renyi correlation provides an interpretation for group LASSO as a variable selection method

Any Questions?