# Nonparanormal Information Estimation

## Realistic High-Dimensional Dependence Estimation

**Shashank Singh**[1,2] and Barnabás Póczos[2]

[1] Department of Statistics
[2] Machine Learning Department
Carnegie Mellon University

8 August 2017,
ICML, Sydney

# Motivation

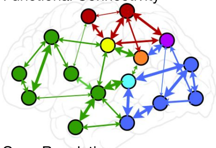
Estimating **dependence strength between variables** is a fundamental ML problem.

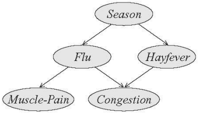
Estimating **dependence strength between variables** is a fundamental ML problem.

Applications to...

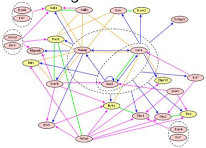
- feature selection [PLD05, SBD+16]
- clustering [ASZAA07]
- learning graphical models [CL68]
- causal discovery [ZPJS11]
- ICA and ISA [LMF03, SPL07]
- EDA and unsupervised learning [VSG16, VSGRG16, Ste17]
- fMRI data analysis [CWBFF09]
- protein structure prediction [Ada04]
- boosting [SGM05]
- fitting deep nonlinear models [HH16]
- ...

Functional Connectivity

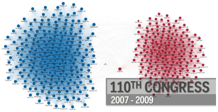





Gene Regulation



Political Influence

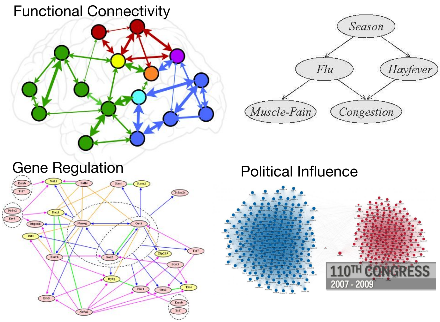


# Motivation

Estimating **dependence strength between variables** is a fundamental ML problem.

Applications to. . .

- feature selection [PLD05, SBD$^+$16]
- clustering [ASZAA07]
- learning graphical models [CL68]
- causal discovery [ZPJS11]
- ICA and ISA [LMF03, SPL07]
- EDA and unsupervised learning [VSG16, VSGRG16, Ste17]
- fMRI data analysis [CWBFF09]
- protein structure prediction [Ada04]
- boosting [SGM05]
- fitting deep nonlinear models [HH16]
- . . .



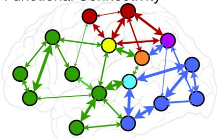
# Motivation

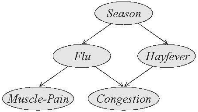
Estimating **dependence strength between variables** is a fundamental ML problem.

Applications to...

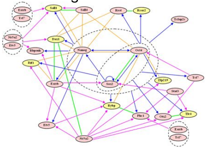
- feature selection [PLD05, SBD$^+$16]
- clustering [ASZAA07]
- learning graphical models [CL68]
- causal discovery [ZPJS11]
- ICA and ISA [LMF03, SPL07]
- EDA and unsupervised learning [VSG16, VSGRG16, Ste17]
- fMRI data analysis [CWBFF09]
- protein structure prediction [Ada04]
- boosting [SGM05]
- fitting deep nonlinear models [HH16]
- ...

Functional Connectivity

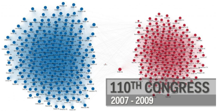





Gene Regulation



Political Influence




**Note: We focus on continuous variables...**
**discrete case is quite different — see next talk.**

# Outline

# Multivariate Mutual Information

## Mutual Information (a.k.a., total correlation [Wat60])

The **mutual information** of a $D$-dimensional random variable $X = (X_1, \cdots, X_D)$ with density $p = p_1 \times \cdots \times p_D$ is

$$I(X) := \mathop{\mathbb{E}}_{X \sim p} \left[ \log \left( \frac{p(x)}{\prod_{j=1}^{D} p_j(x_j)} \right) \right] = D_{KL} \left( p, \prod_{j=1}^{D} p_j \right),$$

where $D_{KL}$ denotes KL divergence.

# Multivariate Mutual Information

## Mutual Information (a.k.a., total correlation [Wat60])

The **mutual information** of a $D$-dimensional random variable $X = (X_1, \cdots, X_D)$ with density $p = p_1 \times \cdots \times p_D$ is

$$I(X) := \mathop{\mathbb{E}}_{X \sim p}\left[\log\left(\frac{p(x)}{\prod_{j=1}^{D} p_j(x_j)}\right)\right] = D_{KL}\left(p, \prod_{j=1}^{D} p_j\right),$$

where $D_{KL}$ denotes KL divergence.

## MI subsumes other information theoretic dependence measures

- Pairwise mutual information: $I(X, Y) = I((X, Y)) - I(X) - I(Y)$
- Conditional mutual information: $I(X|Z) = I((X, Z)) - \sum_{j=1}^{D} I((X_j, Z))$,
- Transfer entropy (a.k.a. “directed information”) $T_{X \to Y}$ between time series

# Multivariate Mutual Information

## Mutual Information (a.k.a., total correlation [Wat60])

The **mutual information** of a $D$-dimensional random variable $X = (X_1, \cdots, X_D)$ with density $p = p_1 \times \cdots \times p_D$ is

$$I(X) := \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(x)}{\prod_{j=1}^D p_j(x_j)}\right)\right] = D_{KL}\left(p, \prod_{j=1}^D p_j\right),$$

where $D_{KL}$ denotes KL divergence.

## MI subsumes other information theoretic dependence measures

- Pairwise mutual information: $I(X, Y) = I((X, Y)) - I(X) - I(Y)$
- Conditional mutual information: $I(X|Z) = I((X, Z)) - \sum_{j=1}^D I((X_j, Z))$,
- Transfer entropy (a.k.a. "directed information") $T_{X \to Y}$ between time series

Paper also discusses **entropy estimation**.

# The Information Estimation Problem

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied:**

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied :**

Gaussian case :

- $X$ jointly Gaussian
- [AG89, CLZ15]
- Minimax MSE : $2D/n$

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied :**

Gaussian case :

- $X$ jointly Gaussian
- [AG89, CLZ15]
- Minimax MSE : $2D/n$

Nonparametric case :

- $X$ has $s$-times differentiable density
- [BM95, L$^+$96, SRH11, SWH13, SP14, KKP$^+$15, SP16, MSHI17]
- Minimax MSE : $\asymp n^{-\frac{8s}{4s+D}}$

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied :**

Gaussian case :

- $X$ jointly Gaussian
- [AG89, CLZ15]
- Minimax MSE : $2D/n$
- Brittle – fails when data are
  - multi-modal
  - heavy-tailed
  - skewed
  - nonlinearly dependent
  - ...

Nonparametric case :

- $X$ has $s$-times differentiable density
- [BM95, L$^+$96, SRH11, SWH13, SP14, KKP$^+$15, SP16, MSHI17]
- Minimax MSE : $\asymp n^{-\frac{8s}{4s+D}}$

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied :**

Gaussian case :

- $X$ jointly Gaussian
- [AG89, CLZ15]
- Minimax MSE : $2D/n$
- Brittle – fails when data are
  - multi-modal
  - heavy-tailed
  - skewed
  - nonlinearly dependent
  - ...

Nonparametric case :

- $X$ has $s$-times differentiable density
- [BM95, L$^+$96, SRH11, SWH13, SP14, KKP$^+$15, SP16, MSHI17]
- Minimax MSE : $\asymp n^{-\frac{8s}{4s+D}}$
- fails when $D$ is bigger than 4-6.

# What do we know about information estimation ?

**Given $n$ IID observations $X_1, \cdots, X_n$ of $X \in \mathbb{R}^D$, estimate $I(X)$.**

**Two cases have been studied :**

Gaussian case :

- $X$ jointly Gaussian
- [AG89, CLZ15]
- Minimax MSE : $2D/n$
- Brittle – fails when data are
  - multi-modal
  - heavy-tailed
  - skewed
  - nonlinearly dependent
  - ...

Nonparametric case :

- $X$ has $s$-times differentiable density
- [BM95, L$^+$96, SRH11, SWH13, SP14, KKP$^+$15, SP16, MSHI17]
- Minimax MSE : $\asymp n^{-\frac{8s}{4s+D}}$
- fails when $D$ is bigger than 4-6.

*"All models are wrong but some are useful."* *[Box79]*

**Often, neither model is useful!**

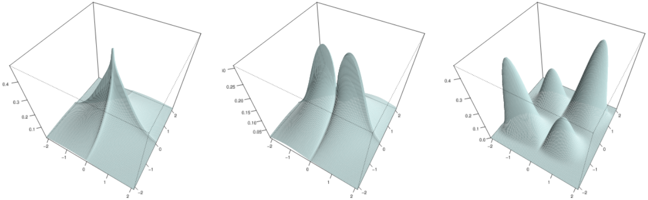
# The Nonparanormal Distribution

**The Nonparanormal (a.k.a. Gaussian copula) Model [LLW09]**

An $\mathbb{R}^D$-valued random variable $X$ has a **nonparanormal distribution** $X \sim \mathcal{NPN}(\Sigma; f)$ if there exist $f_1, ..., f_D : \mathbb{R} \to \mathbb{R}$ such that

$$f(X) = (f_1(X_1), ..., f_D(X_D)) \sim \mathcal{N}(0, \Sigma).$$

$f$ is the **marginal transformation** and $\Sigma$ is the **latent covariance**.



# The Nonparanormal Distribution : two perspectives

- Generalized Gaussian with arbitrary continuous marginals
- Allows, e.g.,. . .
  - multi-modality
  - heavy-tails
  - skew
  - nonlinear dependence
  - . . .

"Additive" model of density estimation

# The Nonparanormal Distribution : two perspectives

- Generalized Gaussian with arbitrary continuous marginals
- Allows, e.g.,...
  - multi-modality
  - heavy-tails
  - skew
  - nonlinear dependence
  - ...

"Additive" model of density estimation

| Gaussian | | Nonparanormal | | Nonparametric |
|---|---|---|---|---|
| $\exp\left(x^T \Sigma x\right)$ | $\overset{\text{generalize}}{\Rightarrow}$ | $\exp\left(f^T(x)\Sigma f(x)\right)$ | $\overset{\text{constrain}}{\Leftarrow}$ | $\exp\left(g(x)\right)$ |

# Our Information Estimators

**Basic Lemma**

If $X \sim \mathcal{NPN}(\Sigma; f)$, then

$$I(X) = -\frac{1}{2} \log |\Sigma|. \tag{1}$$

# Our Information Estimators

**Basic Lemma**

If $X \sim \mathcal{NPN}(\Sigma; f)$, then

$$I(X) = -\frac{1}{2} \log |\Sigma|. \tag{1}$$

- Three latent correlation estimators:
  - $\widehat{\Sigma}_G$ : "Gaussianize" data and calculate empirical correlation
  - $\widehat{\Sigma}_\rho$ : Transform Spearman rank correlation matrix $\rho$
  - $\widehat{\Sigma}_\tau$ : Transform Kendall rank correlation matrix $\tau$

# Our Information Estimators

**Basic Lemma**

If $X \sim \mathcal{NPN}(\Sigma; f)$, then

$$I(X) = -\frac{1}{2} \log |\Sigma|. \tag{1}$$

- Three latent correlation estimators :
  - $\widehat{\Sigma}_G$ : “Gaussianize” data and calculate empirical correlation
  - $\widehat{\Sigma}_\rho$ : Transform Spearman rank correlation matrix $\rho$
  - $\widehat{\Sigma}_\tau$ : Transform Kendall rank correlation matrix $\tau$
- Want to plug $\widehat{\Sigma}_T$ ($T \in \{G, \rho, \tau\}$) into (1) — but not necessarily positive definite!

# Our Information Estimators

**Basic Lemma**

If $X \sim \mathcal{NPN}(\Sigma; f)$, then

$$I(X) = -\frac{1}{2} \log |\Sigma|. \tag{1}$$

- Three latent correlation estimators:
  - $\widehat{\Sigma}_G$ : "Gaussianize" data and calculate empirical correlation
  - $\widehat{\Sigma}_\rho$ : Transform Spearman rank correlation matrix $\rho$
  - $\widehat{\Sigma}_\tau$ : Transform Kendall rank correlation matrix $\tau$
- Want to plug $\widehat{\Sigma}_T$ ($T \in \{G, \rho, \tau\}$) into (1) — but not necessarily positive definite!
- Regularize $\widehat{\Sigma}_T$ to have minimum eigenvalue $z > 0$ (via projection)

# Our Information Estimators

**Basic Lemma**

If $X \sim \mathcal{NPN}(\Sigma; f)$, then

$$I(X) = -\frac{1}{2} \log |\Sigma|. \tag{1}$$

- Three latent correlation estimators:
  - $\widehat{\Sigma}_G$ : "Gaussianize" data and calculate empirical correlation
  - $\widehat{\Sigma}_\rho$ : Transform Spearman rank correlation matrix $\rho$
  - $\widehat{\Sigma}_\tau$ : Transform Kendall rank correlation matrix $\tau$
- Want to plug $\widehat{\Sigma}_T$ ($T \in \{G, \rho, \tau\}$) into (1) — but not necessarily positive definite!
- Regularize $\widehat{\Sigma}_T$ to have minimum eigenvalue $z > 0$ (via projection)
- Plug $\widehat{\Sigma}_{T,z}$ into (1) :

$$\widehat{I}_{T,z} := -\frac{1}{2} \log \left| \widehat{\Sigma}_{T,z} \right|$$

# Theoretical Results

## (Simplified) Upper Bound

Assuming $z \leq \lambda_D(\Sigma)$,

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq \frac{C}{z^2} \frac{D^2}{n}.$$

# Theoretical Results

## (Simplified) Upper Bound

Assuming $z \leq \lambda_D(\Sigma)$,

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq \frac{C}{z^2}\frac{D^2}{n}.$$

## Lower Bound

There exists $C_{n,D} > 0$ such that

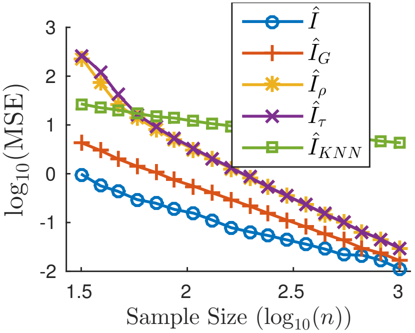
$$\inf_{\widehat{I}} \sup_{\lambda_D(\Sigma)\geq\lambda} \mathbb{E}\left[\left(\widehat{I} - I\right)^2\right] \geq -C\log^2(\lambda_D(\Sigma)).$$

**Constrast Gaussian case**, where distribution of $\widehat{I} - I$ is independent of $\Sigma$.

# Experimental Results

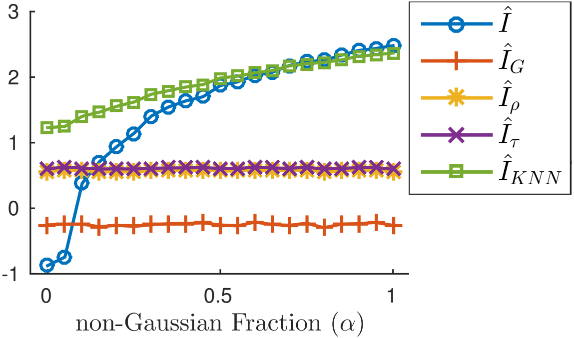
- Synthetic data, with known ground truth
  - [IGK$^+$17] studies applications to neural data analysis
- We compare:
  - Optimal Gaussian estimator $\widehat{I}$ [CLZ15]
  - Our nonparanormal estimators $\widehat{I}_G, \widehat{I}_\rho, \widehat{I}_\tau$
  - Classic nonparametric $k$NN estimator $\widehat{I}_{k\mathrm{NN}}$ [KL87]

# Experimental Results



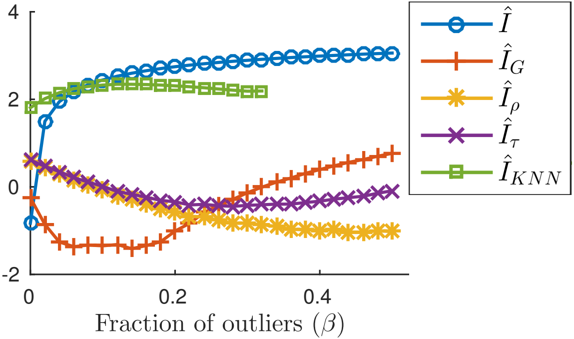

Truly Gaussian data, $D = 25$.

# Experimental Results




Gaussian data partially transformed by $x \mapsto e^x$, $n = 100$, $D = 25$.

# Experimental Results




Gaussian data with random outliers $\pm 5$, $n = 100$, $D = 25$.

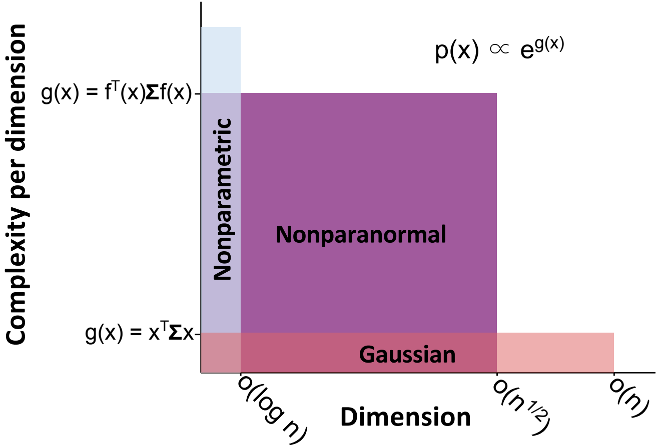



Figure: When can we estimate $I(X)$ consistently?

**See Poster #120**

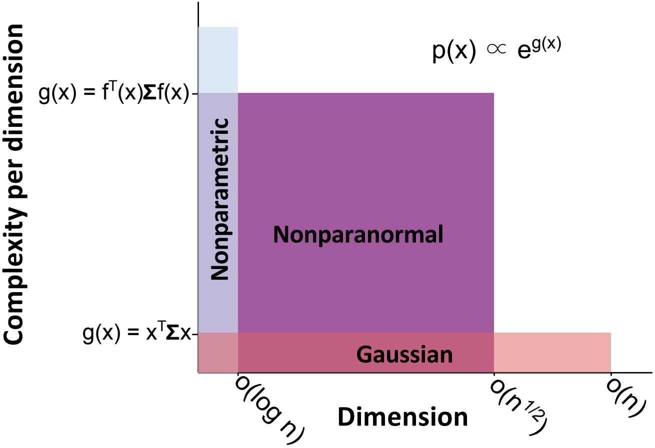



Figure: When can we estimate $I(X)$ consistently?

# Entropy Estimation

$$H(X) = \sum_{j=1}^{D} H(X_j) - I(X)$$

- Depends on marginals through $H(X_1), \cdots, H(X_D)$.
- Can estimate at $O(D^2/n)$ rate under mild smoothness assumptions on marginals.

## Theoretical Results : Lower Bounds

[CLZ15] recently that, in the Gaussian case, the distribution of $\widehat{I} - I$ is independent of $\Sigma$

- Quite surprising, since $I \to \infty$ as $\lambda_D(\Sigma) \to 0$ !

We show this is not possible in the nonparanormal case. Specifically, there exists a constant $C_{n,D}$ such that

$$\inf_{\widehat{I}} \sup_{\lambda_D(\Sigma) \geq \lambda} \mathbb{E}\left[\left(\widehat{I} - I\right)\right] \geq -C \log^2(\lambda_D(\Sigma)).$$

# The "Additive" Model of Density Estimation

$$g(x) = \sum_{j=1}^{D} f_j(x_j)$$

$$p(x) \propto \exp\left(\sum_{j=1}^{D} f_j(x_j)\right) = \prod_{j=1}^{D} \exp\left(f_j(x_j)\right)$$

$$\begin{aligned} p(x) &\propto \exp\left(\sum_{j,k=1}^{D} \sigma_{k,j} f_j(x_j) f_k(x_k)\right) \\ &= \exp\left(f^T(x)\Sigma f(x)\right). \end{aligned}$$

# References I


- Christoph Adami, **Information theory in molecular biology**, Physics of Life Reviews **1** (2004), no. 1, 3–22.
- Nabil Ali Ahmed and DV Gokhale, **Entropy expressions and their estimators for multivariate distributions**, IEEE Trans. on Information Theory **35** (1989), no. 3, 688–692.
- Mehdi Aghagolzadeh, Hamid Soltanian-Zadeh, B Araabi, and Ali Aghagolzadeh, **A hierarchical clustering based on mutual information maximization**, Image Processing, 2007. ICIP 2007. IEEE International Conference on, vol. 1, IEEE, 2007, pp. I–277.
- Lucien Birgé and Pascal Massart, **Estimation of integral functionals of a density**, The Annals of Statistics (1995), 11–29.
- George EP Box, **All models are wrong, but some are useful**, Launer, RL (1979).
- C Chow and Cong Liu, **Approximating discrete probability distributions with dependence trees**, IEEE transactions on Information Theory **14** (1968), no. 3, 462–467.

# References II


- T Tony Cai, Tengyuan Liang, and Harrison H Zhou, **Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions**, J. of Multivariate Analysis **137** (2015), 161–172.
- Barry Chai, Dirk Walther, Diane Beck, and Li Fei-Fei, **Exploring functional connectivities of the human brain using multivariate information analysis**, Advances in neural information processing systems, 2009, pp. 270–278.
- Jacob S Hunter and Nathan O Hodas, **Mutual information for fitting deep nonlinear models**, arXiv preprint arXiv :1612.05708 (2016).
- Robin AA Ince, Bruno L Giordano, Christoph Kayser, Guillaume A Rousselet, Joachim Gross, and Philippe G Schyns, **A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula**, Human brain mapping **38** (2017), no. 3, 1541–1573.
- Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M. Robins, **Nonparametric von mises estimators for entropies, divergences and mutual informations**, Advances in Neural Information Processing Systems, 2015, pp. 397–405.

# References III


- LF Kozachenko and Nikolai N Leonenko, **Sample estimate of the entropy of a random vector**, Problemy Peredachi Informatsii **23** (1987), no. 2, 9–16.
- Béatrice Laurent et al., **Efficient estimation of integral functionals of a density**, The Annals of Statistics **24** (1996), no. 2, 659–681.
- Han Liu, John Lafferty, and Larry Wasserman, **The nonparanormal : Semiparametric estimation of high dimensional undirected graphs**, Journal of Machine Learning Research **10** (2009), no. Oct, 2295–2328.
- E. G. Learned-Miller and J. W. Fisher, **ICA using spacings estimates of entropy**, Journal of Machine Learning Research **4** (2003), 1271–1295.
- Kevin R Moon, Kumar Sricharan, and Alfred O Hero III, **Ensemble estimation of mutual information**, arXiv preprint arXiv :1701.08083 (2017).
- Hanchuan Peng, Fuhui Long, and Chris Ding, **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy**, IEEE Trans. on Pattern Analysis and Machine Intelligence **27** (2005), no. 8, 1226–1238.

# References IV


- Alexander Shishkin, Anastasia Bezzubtseva, Alexey Drutsa, Ilia Shishkov, Ekaterina Gladkikh, Gleb Gusev, and Pavel Serdyukov, **Efficient high-order interaction-aware feature selection based on conditional mutual information**, Advances in Neural Information Processing Systems, 2016, pp. 4637–4645.
- Caifeng Shan, Shaogang Gong, and Peter W McOwan, **Conditional mutual infomation based boosting for facial expression recognition.**, BMVC, 2005.
- Shashank Singh and Barnabás Póczos, **Exponential concentration of a density functional estimator**, Advances in Neural Information Processing Systems, 2014, pp. 3032–3040.
- ______, **Analysis of k-nearest neighbor distances with application to entropy estimation**, arXiv preprint arXiv :1603.08578 (2016).
- Zoltán Szabó, Barnabás Póczos, and András Lőrincz, **Undercomplete blind subspace deconvolution**, Journal of Machine Learning Research **8** (2007), no. May, 1063–1095.
- Kumar Sricharan, Raviv Raich, and Alfred O Hero, **k-nearest neighbor estimation of entropies with confidence**, IEEE International Symposium on Information Theory (ISIT), IEEE, 2011, pp. 1205–1209.

# References V


- Greg Ver Steeg, **Unsupervised learning via total correlation explanation**, arXiv preprint arXiv :1706.08984 (2017).
- Kumar Sricharan, Dennis Wei, and Alfred O Hero, **Ensemble estimators for multivariate entropy estimation**, IEEE Transactions on Information Theory **59** (2013), no. 7, 4374–4388.
- Greg Ver Steeg and Aram Galstyan, **The information sieve**, International Conference on Machine Learning (ICML), 2016.
- Greg Ver Steeg, Shuyang Gao, Kyle Reing, and Aram Galstyan, **Sifting common information from many variables**, arXiv preprint arXiv :1606.02307 (2016).
- Satosi Watanabe, **Information theoretical analysis of multivariate correlation**, IBM J. of research and development **4** (1960), no. 1, 66–82.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf, **Kernel-based conditional independence test and application in causal discovery**, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2011, pp. 804–813.