Robust Estimation: Optimal Rates, Adaptation and Computation

Chao Gao
University of Chicago

@TTIC, August 2018

Huber's Model

$$X_1, ..., X_n \sim (1-\epsilon)P_\theta + \epsilon Q$$

[Huber 1964]

Huber's Model

$$X_1, ..., X_n \sim (1-\epsilon)P_\theta + \epsilon Q$$

parameter of interest

[Huber 1964]

Huber's Model

$$X_1, ..., X_n \sim (1 - \epsilon)P_\theta + \epsilon Q$$

contamination proportion

parameter of interest

[Huber 1964]

Huber's Model

$$X_1, ..., X_n \sim (1-\epsilon)P_\theta + \epsilon Q$$

contamination proportion

contamination

parameter of interest

[Huber 1964]

An Example

$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon Q.$$

An Example

$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon Q.$$

how to estimate ?

An Example

1. **Coordinatewise median**

$$\hat{\theta} = (\hat{\theta}_j), \text{ where } \hat{\theta}_j = \text{Median}(\{X_{ij}\}_{i=1}^{n});$$

An Example

1. **Coordinatewise median**

$$\hat{\theta} = (\hat{\theta}_j), \text{ where } \hat{\theta}_j = \text{Median}(\{X_{ij}\}_{i=1}^n);$$

2. **Tukey's median**

$$\hat{\theta} = \arg\max_{\eta \in \mathbb{R}^p} \min_{||u||=1} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\}.$$

An Example

	coordinatewise median	Tukey's median
breakdown point		

An Example

	coordinatewise median	Tukey's median
breakdown point	$\frac{1}{2}$	$\frac{1}{3}$

An Example

	coordinatewise median	Tukey's median
breakdown point	$\frac{1}{2}$	$\frac{1}{3}$
convergence rate (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$

An Example

	coordinatewise median	Tukey's median
breakdown point	$\frac{1}{2}$	$\frac{1}{3}$
convergence rate (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$
convergence rate (with contamination)	$\frac{p}{n} + p\epsilon^2$	$\frac{p}{n} + \epsilon^2$

An Example

	coordinatewise median	Tukey's median
breakdown point	$\frac{1}{2}$	$\frac{1}{3}$
convergence rate (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$
convergence rate (with contamination)	$\frac{p}{n} + p\epsilon^2$	$\frac{p}{n} + \epsilon^2$ minimax

[CGR15]

Robustness: Where are we now? [1]

Peter J. Huber

University of Bayreuth, Germany

...

3 Breakdown and outlier detection

For a long time, the breakdown point had been a step-child of the robustness literature. The paper by Donoho and Huber (1983) was specifically written to give it more visibility. Recently, I have begun to wonder whether it has given it too much, the suddenly fashionable emphasis on high breakdown point procedures has become counter-productive. One of the most striking

Robustness: Where are we now? [1]

Peter J. Huber

University of Bayreuth, Germany

...

3 Breakdown and outlier detection

For a long time, the breakdown point had been a step-child of the robustness literature. The paper by Donoho and Huber (1983) was specifically written to give it more visibility. Recently, I have begun to wonder whether it has given it too much, the suddenly fashionable emphasis on high breakdown point procedures has become counter-productive. One of the most striking

Why Huber's Model?

Why Huber's Model?

- **unified framework for robustness and accuracy**

Why Huber’s Model?

- **unified framework for robustness and accuracy**
- **implies breakdown point**

Why Huber's Model?

- **unified framework for robustness and accuracy**
- **implies breakdown point**
- **more than breakdown point**

Why Huber's Model?

- **unified framework for robustness and accuracy**
- **implies breakdown point**
- **more than breakdown point**
- **relation to influence function and maxbias**

Multivariate Location Depth

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

Multivariate Location Depth

$$\min_{\|u\|=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i \le u^T \eta\} \right\}$$

Multivariate Location Depth

$$\hat{\theta} = \arg\max_{\eta \in \mathbb{R}^p} \min_{\|u\|=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

Multivariate Location Depth

$$\hat{\theta} = \arg\max_{\eta \in \mathbb{R}^p} \min_{||u||=1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

$$= \arg\max_{\eta \in \mathbb{R}^p} \min_{||u||=1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i > u^T \eta\}.$$

[Tukey, 1975]

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

embedding $$Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$$

Regression Depth

model $y|X \sim N(X^T\beta, \sigma^2)$

embedding $Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$

projection $u^T Xy|X \sim N(u^T XX^T\beta, \sigma^2 u^T XX^T u)$

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

embedding $$Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$$

projection $$u^T Xy|X \sim N(u^T XX^T\beta, \sigma^2 u^T XX^T u)$$

$$\left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) > 0\} \wedge \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) \leq 0\} \right\}$$

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

embedding $$Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$$

projection $$u^T Xy|X \sim N(u^T XX^T\beta, \sigma^2 u^T XX^T u)$$

$$\min_{u \in \mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) > 0\} \wedge \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) \leq 0\} \right\}$$

Regression Depth

model $y|X \sim N(X^T\beta, \sigma^2)$

embedding $Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$

projection $u^TXy|X \sim N(u^TXX^T\beta, \sigma^2 u^TXX^Tu)$

$$\hat{\beta} = \underset{\eta\in\mathbb{R}^p}{\operatorname{argmax}} \min_{u\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^TX_i(y_i - X_i^T\eta) > 0\} \wedge \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^TX_i(y_i - X_i^T\eta) \le 0\} \right\}$$

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

embedding $$Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$$

projection $$u^TXy|X \sim N(u^TXX^T\beta, \sigma^2 u^TXX^Tu)$$

$$\hat{\beta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmax}} \min_{u \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i (y_i - X_i^T \eta) > 0\} \wedge \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{u^T X_i (y_i - X_i^T \eta) \le 0\} \right\}$$

[Rousseeuw & Hubert, 1999]

Tukey’s depth is not a special case of regression depth.

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

$$B \in \mathbb{R}^{p \times m}$$

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

$$B \in \mathbb{R}^{p \times m}$$

population version:

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}$$

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

$$B \in \mathbb{R}^{p \times m}$$

population version:

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}$$

empirical version:

$$\mathcal{D}_{\mathcal{U}}(B, \{(X_i, Y_i)\}_{i=1}^n) = \inf_{U \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\left\langle U^T X_i, Y_i - B^T X_i \right\rangle \geq 0\right\}$$

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

$$B \in \mathbb{R}^{p \times m}$$

population version:

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}$$

empirical version:

$$\mathcal{D}_{\mathcal{U}}(B, \{(X_i, Y_i)\}_{i=1}^n) = \inf_{U \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\left\langle U^T X_i, Y_i - B^T X_i \right\rangle \geq 0\right\}$$

[Mizera, 2002]

Multi-task Regression Depth

$$\boxed{\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}}$$

Multi-task Regression Depth

$$\boxed{\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}}$$

$p = 1, X = 1 \in \mathbb{R},$

$$\mathcal{D}_{\mathcal{U}}(b, \mathbb{P}) = \inf_{u \in \mathcal{U}} \mathbb{P}\left\{u^T (Y - b) \geq 0\right\}$$

Multi-task Regression Depth

$$\boxed{\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}}$$

$p = 1, X = 1 \in \mathbb{R},$

$$\mathcal{D}_{\mathcal{U}}(b, \mathbb{P}) = \inf_{u \in \mathcal{U}} \mathbb{P}\left\{u^T (Y - b) \geq 0\right\}$$

$m = 1,$

$$\mathcal{D}_{\mathcal{U}}(\beta, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{u^T X (y - \beta^T X) \geq 0\right\}$$

Multi-task Regression Depth

Proposition. For any $\delta > 0$,

$$\sup_{B \in \mathbb{R}^{p \times m}} |\mathcal{D}(B, \mathbb{P}_n) - \mathcal{D}(B, \mathbb{P})| \leq C\sqrt{\frac{pm}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$.

Multi-task Regression Depth

Proposition. For any $\delta > 0$,

$$\sup_{B \in \mathbb{R}^{p \times m}} |\mathcal{D}(B, \mathbb{P}_n) - \mathcal{D}(B, \mathbb{P})| \leq C\sqrt{\frac{pm}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$.

Proposition.

$$\sup_{B,Q} |\mathcal{D}(B, (1 - \epsilon P_{B^*}) + \epsilon Q) - \mathcal{D}(B, P_{B^*})| \leq \epsilon$$

Multi-task Regression Depth

$$(X, Y) \sim P_B$$

Multi-task Regression Depth

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

Multi-task Regression Depth

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

$$(X_1, Y_1), ..., (X_n, Y_n) \sim (1 - \epsilon)P_B + \epsilon Q$$

Multi-task Regression Depth

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

$$(X_1, Y_1), ..., (X_n, Y_n) \sim (1-\epsilon)P_B + \epsilon Q$$

Theorem [G17]. For some $C > 0$,

$$\mathsf{Tr}((\widehat{B} - B)^T \Sigma (\widehat{B} - B)) \leq C\sigma^2 \left(\frac{pm}{n} \vee \epsilon^2 \right),$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \leq C \frac{\sigma^2}{\kappa^2} \left(\frac{pm}{n} \vee \epsilon^2 \right),$$

with high probability uniformly over B, Q.

Multi-task Regression Depth

minimax rate achieved by multi-task regression depth	$\frac{pm}{n} \vee \epsilon^2$
separate multiple regression using regression depth	$\frac{pm}{n} \vee m\epsilon^2$

Applications

Sparse Linear Regression

$$\Theta_s = \left\{ \beta \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \leq s \right\}$$

Sparse Linear Regression

$$\Theta_s = \left\{ \beta \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \leq s \right\} \qquad \hat{\beta} = \operatorname*{argmax}_{\beta \in \Theta_s} \mathcal{D}_{\Theta_{2s}}(\beta, \{(X_i, y_i)\}_{i=1}^n)$$

Sparse Linear Regression

$$\Theta_s = \left\{ \beta \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \leq s \right\} \qquad \hat{\beta} = \underset{\beta \in \Theta_s}{\operatorname{argmax}}\, \mathcal{D}_{\Theta_{2s}}(\beta, \{(X_i, y_i)\}_{i=1}^n)$$

Theorem [G17]. For some $C > 0$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta)\|^2 \leq C\sigma^2 \left(\frac{s \log\left(\frac{ep}{s}\right)}{n} \vee \epsilon^2 \right),$$

$$\|\hat{\beta} - \beta\|^2 \leq C \frac{\sigma^2}{\kappa^2} \left(\frac{s \log\left(\frac{ep}{s}\right)}{n} \vee \epsilon^2 \right),$$

$$\|\hat{\beta} - \beta\|_1^2 \leq C \frac{\sigma^2}{\kappa^2} \left(\frac{s^2 \log\left(\frac{ep}{s}\right)}{n} \vee s\epsilon^2 \right),$$

with high probability uniformly over $\beta \in \Theta_s, Q$.

Gaussian Graphical Model

$X \sim P_\Omega : X \sim N(0, \Omega^{-1})$

$X_1, ..., X_n \sim (1-\epsilon)P_\Omega + \epsilon Q$

Gaussian Graphical Model

$$X \sim P_\Omega : X \sim N(0, \Omega^{-1}) \qquad X_1, ..., X_n \sim (1-\epsilon)P_\Omega + \epsilon Q$$

$$\mathcal{F}_s(M) = \left\{ \Omega = \Omega^T \in \mathbb{R}^{p\times p} : M^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M, \max_{1\leq i\leq p} \sum_{j=1}^{p} \mathbb{I}\{\Omega_{ij} \neq 0\} \leq s \right\}$$

Gaussian Graphical Model

$$X \sim P_\Omega : X \sim N(0, \Omega^{-1}) \qquad X_1, ..., X_n \sim (1-\epsilon)P_\Omega + \epsilon Q$$

$$\mathcal{F}_s(M) = \left\{ \Omega = \Omega^T \in \mathbb{R}^{p \times p} : M^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M, \max_{1 \leq i \leq p} \sum_{j=1}^{p} \mathbb{I}\{\Omega_{ij} \neq 0\} \leq s \right\}$$

Theorem [G17]. For some $C > 0$,

$$\|\widehat{\Omega} - \Omega\|_{\ell_1}^2 \leq C \left(\frac{s^2 \log\left(\frac{ep}{s}\right)}{n} \vee s\epsilon^2 \right),$$

with high probability.

Reduced Rank Regression

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

Reduced Rank Regression

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

$$(X_1, Y_1), ..., (X_n, Y_n) \sim (1-\epsilon)P_B + \epsilon Q$$

Reduced Rank Regression

$$(X, Y) \sim P_B : X \sim N(0, \Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

$$(X_1, Y_1), ..., (X_n, Y_n) \sim (1-\epsilon)P_B + \epsilon Q$$

$$\mathcal{A}_r = \{B \in \mathbb{R}^{p \times m} : \text{rank}(B) \leq r\}$$

Reduced Rank Regression

$$(X,Y) \sim P_B : X \sim N(0,\Sigma), \quad Y|X \sim N(B^T X, \sigma^2 I_m)$$

$$(X_1,Y_1),...,(X_n,Y_n) \sim (1-\epsilon)P_B + \epsilon Q$$

$$\mathcal{A}_r = \{B \in \mathbb{R}^{p\times m} : \mathrm{rank}(B) \leq r\}$$

$$\widehat{B} = \operatorname*{argmax}_{B\in\mathcal{A}_r} \mathcal{D}_{\mathcal{A}_{2r}}(B, \{X_i, y_i\}_{i=1}^n)$$

Reduced Rank Regression

Theorem [G17]. For some $C > 0$,

$$\mathsf{Tr}((\widehat{B} - B)^T \Sigma (\widehat{B} - B)) \leq C\sigma^2 \left(\frac{r(p+m)}{n} \vee \epsilon^2 \right),$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \leq C \frac{\sigma^2}{\kappa^2} \left(\frac{r(p+m)}{n} \vee \epsilon^2 \right),$$

$$\|\widehat{B} - B\|_{\mathrm{N}}^2 \leq C \frac{\sigma^2}{\kappa^2} \left(\frac{r^2(p+m)}{n} \vee r\epsilon^2 \right),$$

with high probability uniformly over $B \in \mathcal{A}_r, Q$.

Extension

$X \in \mathcal{H}_x \quad Y \in \mathcal{H}_y$

Extension

$X \in \mathcal{H}_x \quad Y \in \mathcal{H}_y \qquad\qquad \mathcal{F} \subset \{f : \mathcal{H}_x \to \mathcal{H}_y\}$

Extension

$X \in \mathcal{H}_x \quad Y \in \mathcal{H}_y \qquad\qquad \mathcal{F} \subset \{f : \mathcal{H}_x \to \mathcal{H}_y\}$

$$\mathcal{D}_{\mathcal{F}}(f, \mathbb{P}) = \inf_{g \in \mathcal{F}} \mathbb{P}\left\{\langle g(X), Y - f(X)\rangle \geq 0\right\}$$

Extension

$$X \in \mathcal{H}_x \quad Y \in \mathcal{H}_y \qquad\qquad \mathcal{F} \subset \{f : \mathcal{H}_x \to \mathcal{H}_y\}$$

$$\mathcal{D}_{\mathcal{F}}(f, \mathbb{P}) = \inf_{g \in \mathcal{F}} \mathbb{P}\left\{\langle g(X), Y - f(X) \rangle \geq 0\right\}$$

multivariate location	$\mathcal{H}_x = \{1\}$	$\mathcal{H}_y = \mathbb{R}^m$
linear regression	$\mathcal{H}_x = \mathbb{R}^p$	$\mathcal{H}_y = \mathbb{R}$
multiple linear regression	$\mathcal{H}_x = \mathbb{R}^p$	$\mathcal{H}_y = \mathbb{R}^m$
functional linear regression	$\mathcal{H}_x = \mathcal{H}$	$\mathcal{H}_y = \mathbb{R}$
multiple functional linear regression	$\mathcal{H}_x = \mathcal{H}$	$\mathcal{H}_y = \mathbb{R}^m$

Covariance Matrix

$$X_1, ..., X_n \sim (1 - \epsilon)N(0, \Sigma) + \epsilon Q.$$

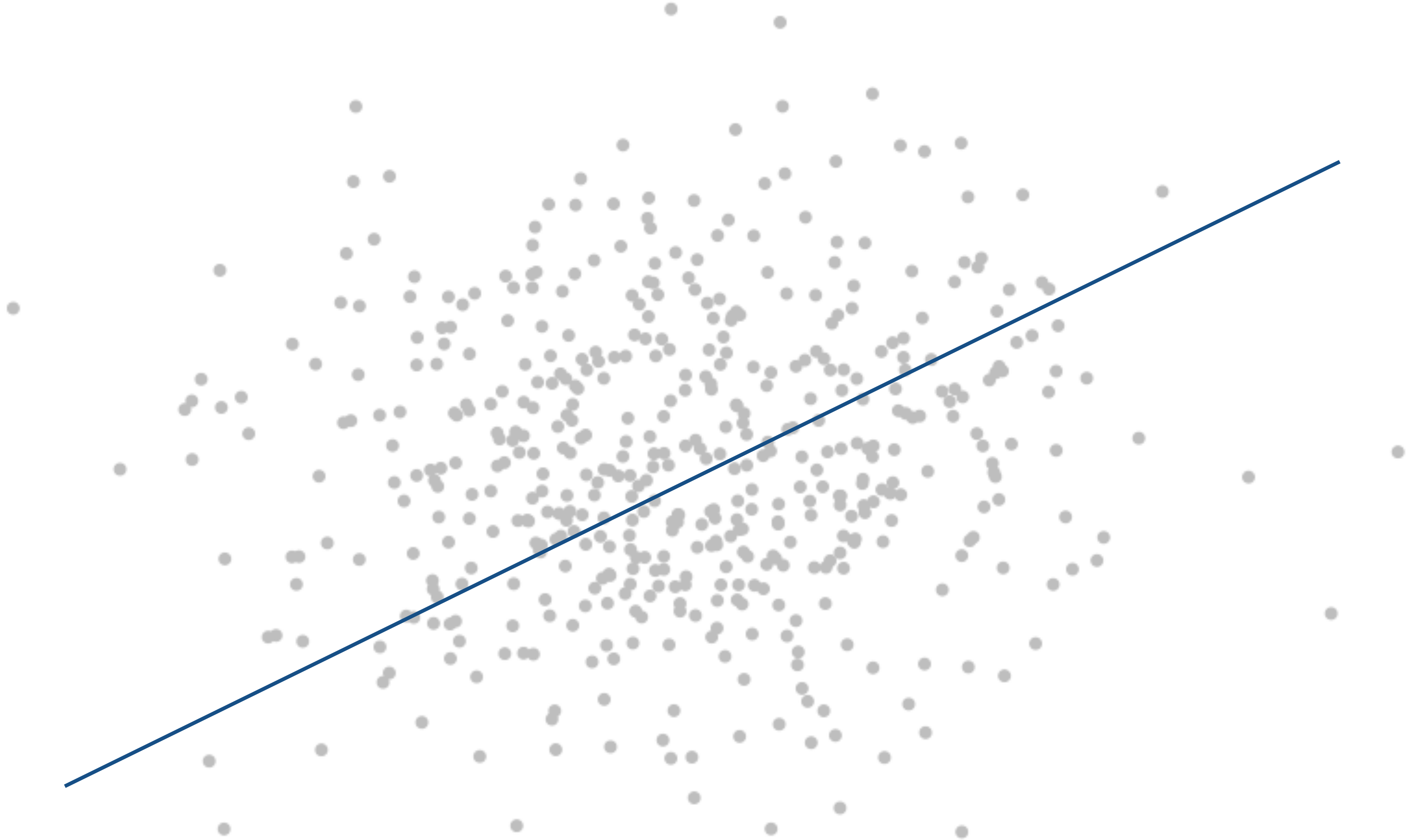
Covariance Matrix

$$X_1, ..., X_n \sim (1-\epsilon)N(0, \Sigma) + \epsilon Q.$$

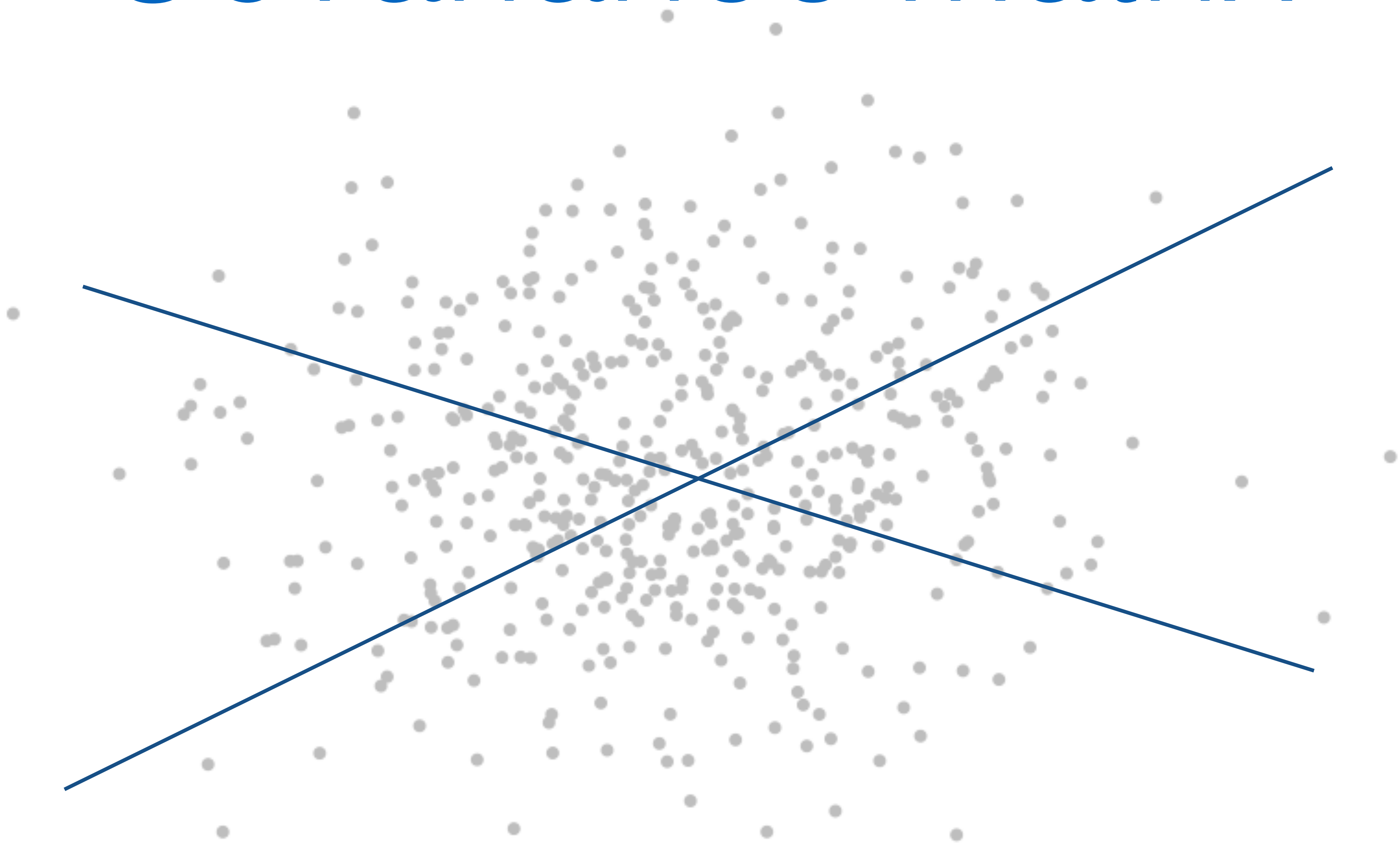
how to estimate ?

Covariance Matrix

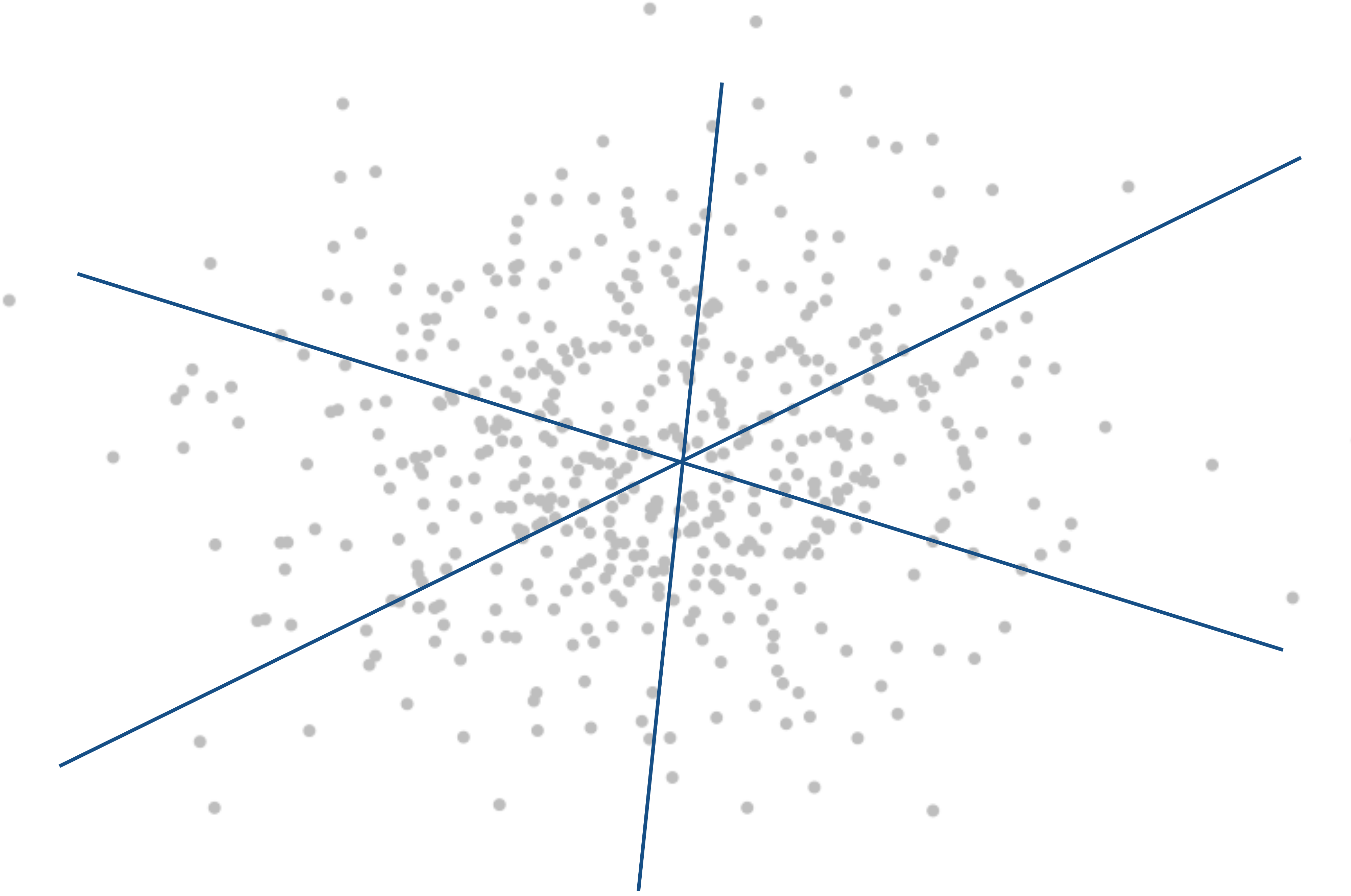
Covariance Matrix



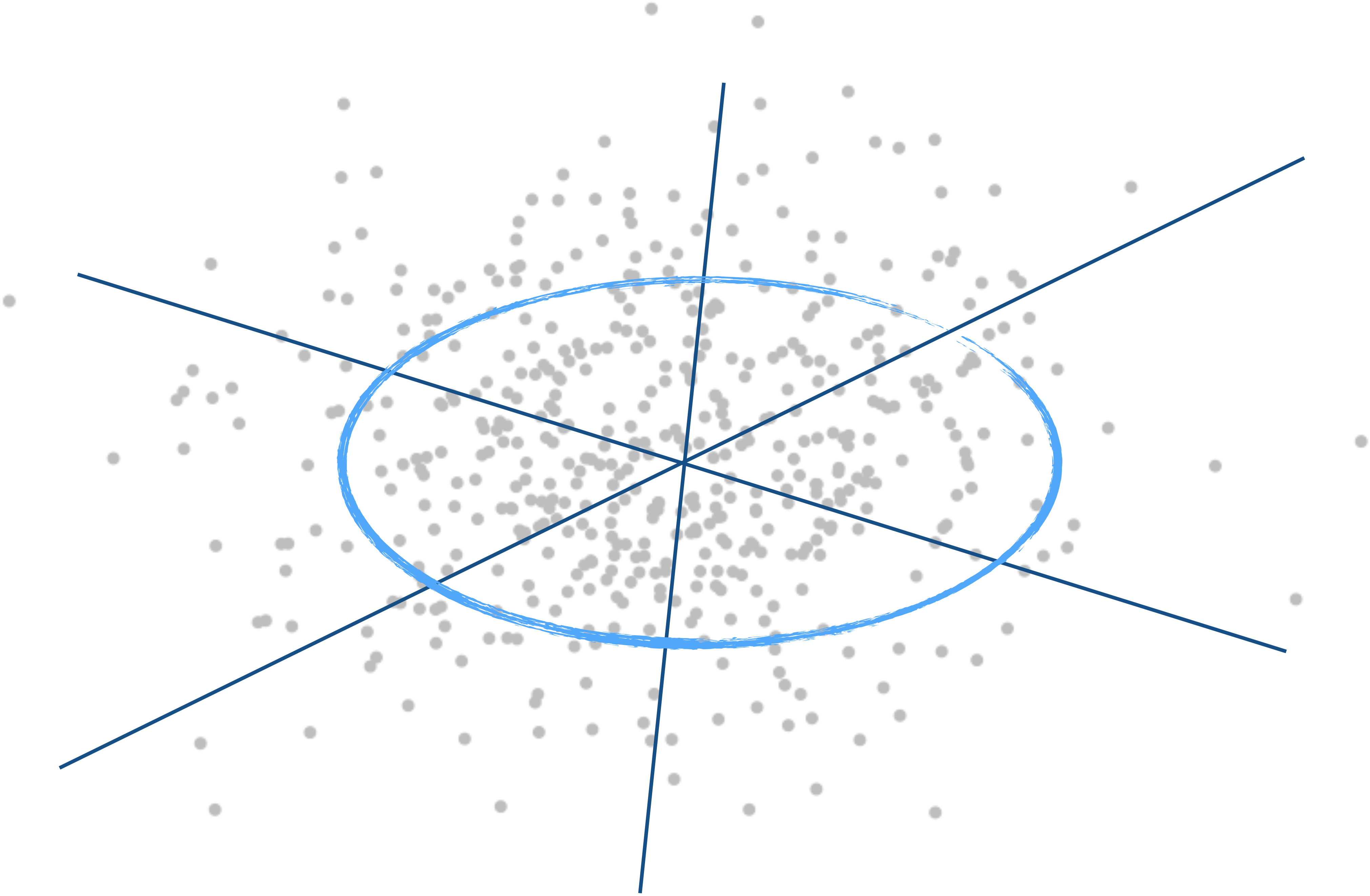
Covariance Matrix



Covariance Matrix



Covariance Matrix



Covariance Matrix

$$\mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) = \min_{\|u\|=1} \min \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \geq u^T \Gamma u\}, \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 < u^T \Gamma u\} \right\}$$

Covariance Matrix

$$\mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) = \min_{\|u\|=1} \min \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \geq u^T \Gamma u\}, \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 < u^T \Gamma u\} \right\}$$

$$\hat{\Gamma} = \arg\max_{\Gamma \succeq 0} \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) \qquad \hat{\Sigma} = \hat{\Gamma}/\beta$$

Covariance Matrix

$$\mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) = \min_{\|u\|=1} \min \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \geq u^T \Gamma u\}, \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 < u^T \Gamma u\} \right\}$$

$$\hat{\Gamma} = \arg\max_{\Gamma \succeq 0} \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) \qquad \hat{\Sigma} = \hat{\Gamma}/\beta$$

Theorem [CGR15]. For some $C > 0$,

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \leq C \left(\frac{p}{n} \vee \epsilon^2 \right)$$

with high probability uniformly over Σ, Q.

Covariance Matrix

$$\mathcal{F}_k = \{\Sigma = (\sigma_{ij}) \succeq 0 : \sigma_{ij} = 0 \text{ if } |i - j| > k\}.$$

Covariance Matrix

$$\mathcal{F}_k = \{\Sigma = (\sigma_{ij}) \succeq 0 : \sigma_{ij} = 0 \text{ if } |i-j| > k\}.$$

$\mathcal{U}_k = \Big\{ u \in S^{p-1} :$

$\Big\}$

Covariance Matrix

$$\mathcal{F}_k = \{\Sigma = (\sigma_{ij}) \succeq 0 : \sigma_{ij} = 0 \text{ if } |i-j| > k\}.$$

$$\mathcal{U}_k = \left\{ u \in S^{p-1} : \right.$$

$$\left. \right\}$$

$$\hat{\Gamma} = \arg\max_{\Gamma \in \mathcal{F}_k} \mathcal{D}_{\mathcal{U}_k}(\Gamma, \{X_i\}_{i=1}^n) \qquad \hat{\Sigma} = \hat{\Gamma}/\beta$$

Covariance Matrix

$$\mathcal{F}_k = \{\Sigma = (\sigma_{ij}) \succeq 0 : \sigma_{ij} = 0 \text{ if } |i-j| > k\}.$$

$$\mathcal{U}_k = \left\{ u \in S^{p-1} : \right.$$

$$\left. \right\}$$

$$\hat{\Gamma} = \arg\max_{\Gamma \in \mathcal{F}_k} \mathcal{D}_{\mathcal{U}_k}(\Gamma, \{X_i\}_{i=1}^n) \qquad \hat{\Sigma} = \hat{\Gamma}/\beta$$

Theorem [CGR15]. For some $C > 0$,

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \leq C\left(\frac{k + \log p}{n} \vee \epsilon^2\right)$$

with high probability uniformly over $\Sigma \in \mathcal{F}_k, Q$.

Covariance Matrix

$$\mathcal{F}_\alpha(M, M_0) = \left\{ \Sigma = (\sigma_{ij}) \in \mathcal{F}(M) : \max_j \sum_{\{i : |i-j| > k\}} |\sigma_{ij}| \leq M_0 k^{-\alpha} \right\}.$$

Covariance Matrix

$$\mathcal{F}_\alpha(M, M_0) = \left\{ \Sigma = (\sigma_{ij}) \in \mathcal{F}(M) : \max_j \sum_{\{i:|i-j|>k\}} |\sigma_{ij}| \le M_0 k^{-\alpha} \right\}.$$

Theorem [CGR15]. Consider the banded estimator with $k = n^{\frac{1}{2\alpha+1}} \wedge p$. For some $C > 0$,

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}}^2 \le C \left[\min\left\{ n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n} \right\} \vee \epsilon^2 \right]$$

with high probability uniformly over $\Sigma \in \mathcal{F}_\alpha, Q$.

Summary

mean	$\|\cdot\|^2$	$\frac{p}{n} \vee \epsilon^2$
reduced rank regression	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{\sigma^2}{\kappa^2}\frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
Gaussian graphical model	$\|\cdot\|_{\ell_1}^2$	$\frac{s^2\log(ep/s)}{n} \vee s\epsilon^2$
covariance matrix	$\|\cdot\|_{\mathrm{op}}^2$	$\frac{p}{n} \vee \epsilon^2$
sparse PCA	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{s\log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$

Summary

mean	$\Vert \cdot \Vert^2$	$\frac{p}{n} \vee \epsilon^2$
reduced rank regression	$\Vert \cdot \Vert_{\mathrm{F}}^2$	$\frac{\sigma^2}{\kappa^2} \frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
Gaussian graphical model	$\Vert \cdot \Vert_{\ell_1}^2$	$\frac{s^2 \log(ep/s)}{n} \vee s\epsilon^2$
covariance matrix	$\Vert \cdot \Vert_{\mathrm{op}}^2$	$\frac{p}{n} \vee \epsilon^2$
sparse PCA	$\Vert \cdot \Vert_{\mathrm{F}}^2$	$\frac{s \log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$

A General Lower Bound

$$\mathbb{P}_{(\epsilon,\theta,Q)} = (1-\epsilon)P_\theta + \epsilon Q \qquad \{\mathbb{P}_{(\epsilon,\theta,Q)} : \theta \in \Theta, Q\}$$

A General Lower Bound

$$\mathbb{P}_{(\epsilon,\theta,Q)} = (1-\epsilon)P_\theta + \epsilon Q \qquad \{\mathbb{P}_{(\epsilon,\theta,Q)} : \theta \in \Theta, Q\}$$

$$\omega(\epsilon, \Theta) = \sup\{L(\theta_1, \theta_2) : \mathsf{TV}(P_{\theta_1}, P_{\theta_2}) \leq \epsilon/(1-\epsilon); \theta_1, \theta_2 \in \Theta\}$$

A General Lower Bound

$\mathbb{P}_{(\epsilon,\theta,Q)} = (1-\epsilon)P_\theta + \epsilon Q \qquad \{\mathbb{P}_{(\epsilon,\theta,Q)} : \theta \in \Theta, Q\}$

$$\omega(\epsilon, \Theta) = \sup\left\{L(\theta_1, \theta_2) : \mathsf{TV}(P_{\theta_1}, P_{\theta_2}) \le \epsilon/(1-\epsilon); \theta_1, \theta_2 \in \Theta\right\}$$

Theorem [CGR15,16]. Suppose $\mathcal{M}(\epsilon)$ is the minimax rate. Then,

$$\mathcal{M}(\epsilon) \gtrsim \mathcal{M}(0) \vee \omega(\epsilon, \Theta)$$

A General Lower Bound

$$\mathbb{P}_{(\epsilon,\theta,Q)} = (1-\epsilon)P_\theta + \epsilon Q \qquad \{\mathbb{P}_{(\epsilon,\theta,Q)} : \theta \in \Theta, Q\}$$

$$\omega(\epsilon, \Theta) = \sup\{L(\theta_1, \theta_2) : \mathsf{TV}(P_{\theta_1}, P_{\theta_2}) \le \epsilon/(1-\epsilon); \theta_1, \theta_2 \in \Theta\}$$

Theorem [CGR15,16]. Suppose $\mathcal{M}(\epsilon)$ is the minimax rate. Then,

$$\mathcal{M}(\epsilon) \gtrsim \mathcal{M}(0) \vee \omega(\epsilon, \Theta)$$

For squared total variation loss, we have

$$\mathcal{M}(\epsilon) \asymp \min_{\delta>0}\left\{\frac{\log \mathcal{N}(\delta, \Theta, \mathsf{TV}(\cdot,\cdot))}{n} + \delta^2\right\} \vee \epsilon^2.$$

Summary

mean	$\|\cdot\|^2$	$\frac{p}{n} \vee \epsilon^2$
reduced rank regression	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{\sigma^2}{\kappa^2}\frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
Gaussian graphical model	$\|\cdot\|_{\ell_1}^2$	$\frac{s^2\log(ep/s)}{n} \vee s\epsilon^2$
covariance matrix	$\|\cdot\|_{\mathrm{op}}^2$	$\frac{p}{n} \vee \epsilon^2$
sparse PCA	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{s\log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$

Computation

Computational Challenges

$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon Q.$$

Computational Challenges

$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon Q.$$

Lai, Rao, Vempala
Diakonikolas, Kamath, Kane, Li, Moitra, Stewart
Balakrishnan, Du, Singh

Advantages of Tukey Median

Advantages of Tukey Median

- **A well-defined objective function**

Advantages of Tukey Median

- **A well-defined objective function**
- **Does not need to know** ϵ

Advantages of Tukey Median

- **A well-defined objective function**
- **Does not need to know** ϵ
- **Does not need to know** Σ

Advantages of Tukey Median

- **A well-defined objective function**
- **Does not need to know** ϵ
- **Does not need to know** Σ
- **Optimal for any elliptical distribution**

A practically good algorithm?

f-Learning

f-Learning

f-divergence $$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ$$

f-Learning

f-divergence $$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ$$

$$f(u) = \sup_t (tu - f^*(t))$$

f-Learning

f-divergence $$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ$$

variational representation $$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

f-Learning

f-divergence

$$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ$$

variational representation

$$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

optimal T

$$T(x) = f'\left(\frac{p(x)}{q(x)}\right)$$

f-Learning

f-divergence $$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ$$

variational representation $$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

$$= \sup_{\tilde{Q}} \left\{ \mathbb{E}_{X\sim P} f'\left(\frac{d\tilde{Q}(X)}{dQ(X)}\right) - \mathbb{E}_{X\sim Q} f^*\left(f'\left(\frac{d\tilde{Q}(X)}{dQ(X)}\right)\right)\right\}$$

f-Learning

$$\max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

$$\max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

f-GAN

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

f-Learning

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

f-GAN

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

f-Learning

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

[Nowozin, Cseke, Tomioka]

f-Learning

f-Learning

Janson-Shannon	$f(x) = x \log x - (x+1) \log(x+1)$	**GAN**

[Goodfellow et al.

f-Learning

Janson-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x \log x$	**MLE**

[Goodfellow et al.

f-Learning

Janson-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x \log x$	**MLE**
Hellinger Squared	$f(x) = 2 - 2\sqrt{x}$	**rho**

[Goodfellow et al., Baraud and Birge]

f-Learning

Janson-Shannon	$f(x) = x\log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x\log x$	**MLE**
Hellinger Squared	$f(x) = 2 - 2\sqrt{x}$	**rho**
Total Variation	$f(x) = (x-1)_+$	**depth**

[Goodfellow et al., Baraud and Birge]

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

$$\boxed{\mathcal{Q}=\left\{N(\theta,I_p):\theta\in\mathbb{R}^p\right\}\qquad\tilde{\mathcal{Q}}=\left\{N(\tilde{\theta},I_p):\tilde{\theta}\in\mathcal{N}_r(\theta)\right\}}$$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

$$\boxed{\mathcal{Q}=\left\{N(\theta,I_p):\theta\in\mathbb{R}^p\right\}\qquad \tilde{\mathcal{Q}}=\left\{N(\tilde{\theta},I_p):\tilde{\theta}\in\mathcal{N}_r(\theta)\right\}}$$

$r\to 0$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

$$\boxed{\mathcal{Q}=\left\{N(\theta,I_p):\theta\in\mathbb{R}^p\right\}\qquad \tilde{\mathcal{Q}}=\left\{N(\tilde{\theta},I_p):\tilde{\theta}\in\mathcal{N}_r(\theta)\right\}}$$

$$\Big\downarrow \; r\to 0$$

Tukey depth $$\max_{\theta\in\mathbb{R}}\min_{\|u\|=1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{u^TX_i\geq u^T\theta\right\}$$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

TV-Learning

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left\{ \frac{\tilde{q}(X_i)}{q(X_i)} \geq 1 \right\} - Q\left(\frac{\tilde{q}}{q} \geq 1 \right) \right\}$$

$$\boxed{\mathcal{Q} = \left\{ N(0, \Sigma) : \Sigma \in \mathbb{R}^{p \times p} \right\} \quad \tilde{\mathcal{Q}} = \left\{ N(0, \tilde{\Sigma}) : \tilde{\Sigma} = \Sigma + ruu^T, \|u\| = 1 \right\}}$$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

$$\mathcal{Q}=\left\{N(0,\Sigma):\Sigma\in\mathbb{R}^{p\times p}\right\}\quad \tilde{\mathcal{Q}}=\left\{N(0,\tilde{\Sigma}):\tilde{\Sigma}=\Sigma+ruu^T,\|u\|=1\right\}$$

$r \to 0$

TV-Learning

$$\min_{Q\in\mathcal{Q}}\max_{\tilde{Q}\in\tilde{\mathcal{Q}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\frac{\tilde{q}(X_i)}{q(X_i)}\geq 1\right\}-Q\left(\frac{\tilde{q}}{q}\geq 1\right)\right\}$$

$$\boxed{\mathcal{Q}=\left\{N(0,\Sigma):\Sigma\in\mathbb{R}^{p\times p}\right\}\quad \tilde{\mathcal{Q}}=\left\{N(0,\tilde{\Sigma}):\tilde{\Sigma}=\Sigma+ruu^T,\|u\|=1\right\}}$$

$r \to 0$

matrix depth $\max_{\Sigma}\min_{\|u\|=1}\min\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{|u^TX_i|^2\geq u^T\Sigma u\},\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{|u^TX_i|^2< u^T\Sigma u\}\right\}$

robust
statistics
community

deep
learning
community

robust
statistics
community

f-Learning

f-GAN

deep
learning
community

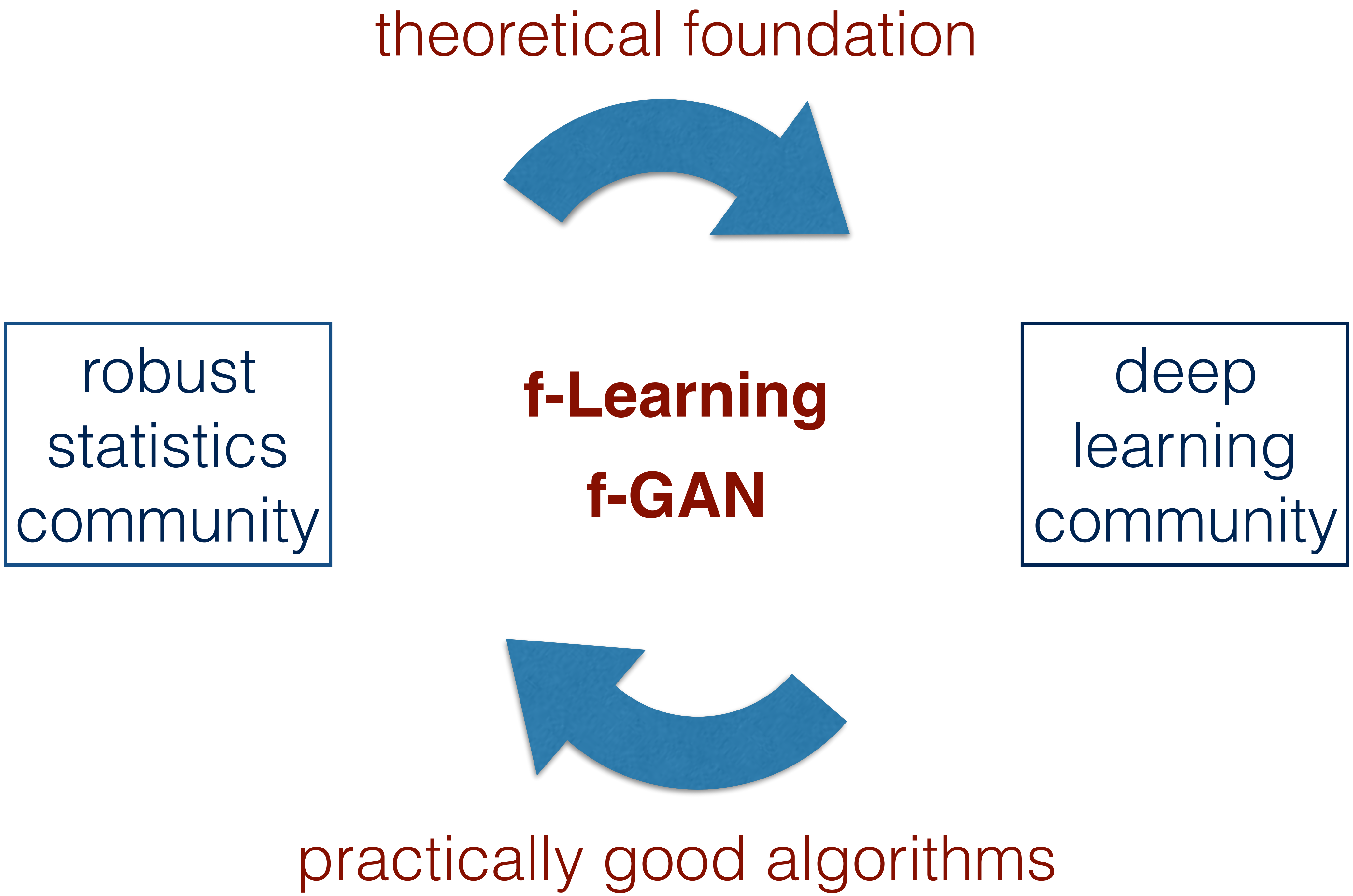
robust statistics community

f-Learning

f-GAN

deep learning community

practically good algorithms


theoretical foundation
robust statistics community
f-Learning
f-GAN
deep learning community
practically good algorithms

TV-GAN

$$\widehat{\theta} = \operatorname*{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1+e^{-w^T X - b}} \right]$$

TV-GAN

$$\widehat{\theta} = \operatorname*{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1+e^{-w^T X - b}} \right]$$

$$N(\eta, I_p)$$

TV-GAN

$$\widehat{\theta} = \underset{\eta}{\operatorname{argmin}} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+e^{-w^T X_i - b}} - E_\eta \frac{1}{1+e^{-w^T X - b}} \right]$$

$N(\eta, I_p)$

logistic regression classifier

TV-GAN

$$\widehat{\theta} = \underset{\eta}{\operatorname{argmin}} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+e^{-w^T X_i - b}} - E_\eta \frac{1}{1+e^{-w^T X - b}} \right]$$

$N(\eta, I_p)$

logistic regression classifier

Theorem [GLYZ18+]. For some $C > 0$,

$$\|\widehat{\theta} - \theta\|^2 \leq C \left(\frac{p}{n} \vee \epsilon^2 \right)$$

with high probability uniformly over $\theta \in \mathbb{R}^p, Q$.

TV-GAN

very hard to optimize!

JS-GAN

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

JS-GAN

$$\widehat{\theta} = \operatorname*{argmin}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

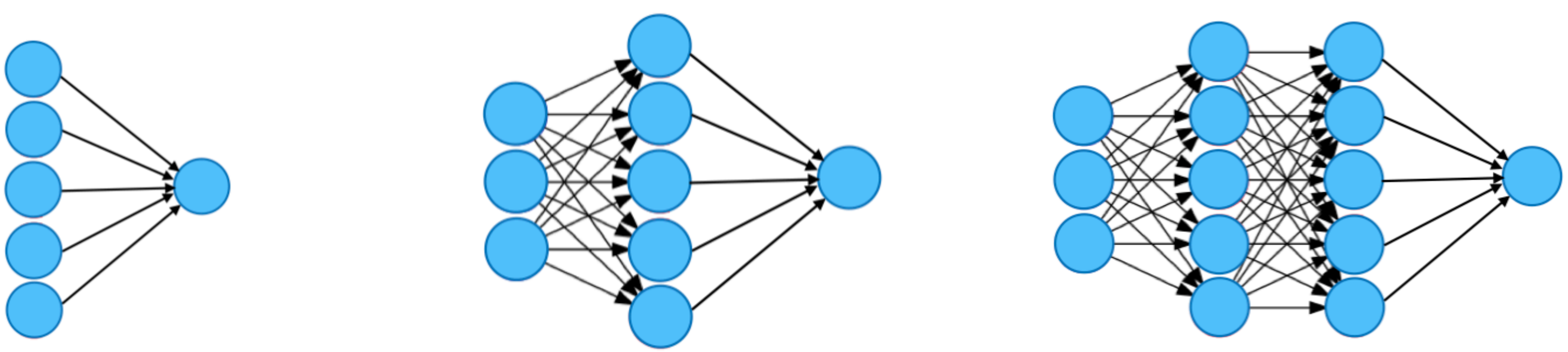
$$X_1, ..., X_n \sim (1 - \epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$

JS-GAN

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

$$X_1, \ldots, X_n \sim (1-\epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$

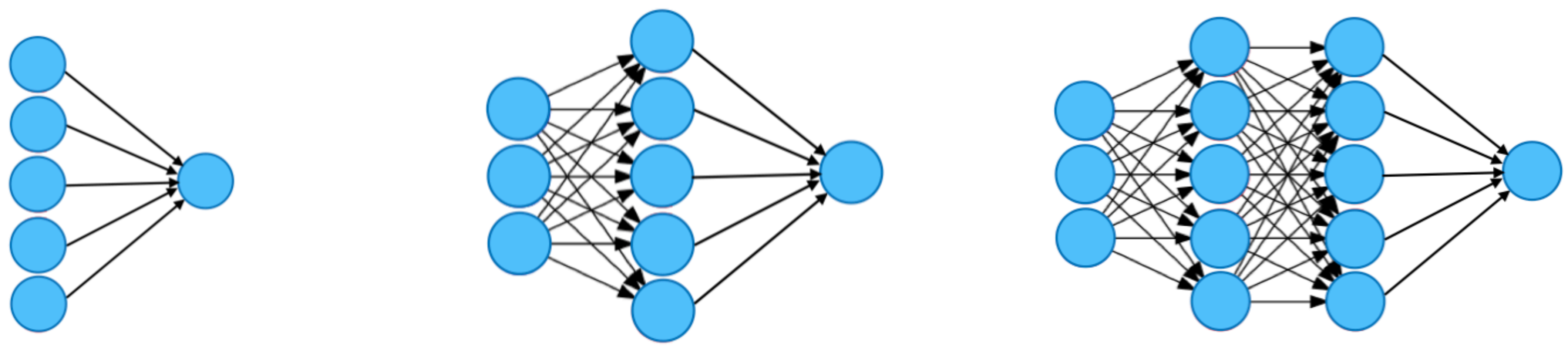


JS-GAN

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

$$X_1, ..., X_n \sim (1-\epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$



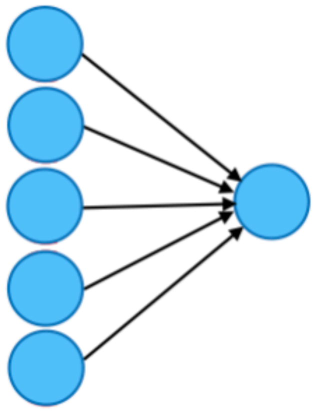
$$\widehat{\theta} \approx (1-\epsilon)\theta + \epsilon \widetilde{\theta}$$

JS-GAN

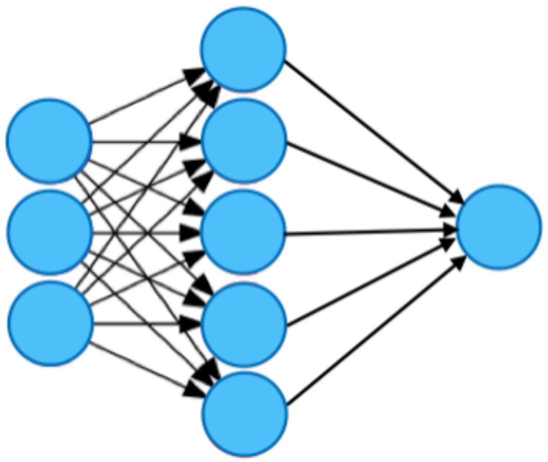
$$\widehat{\theta} = \operatorname*{argmin}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

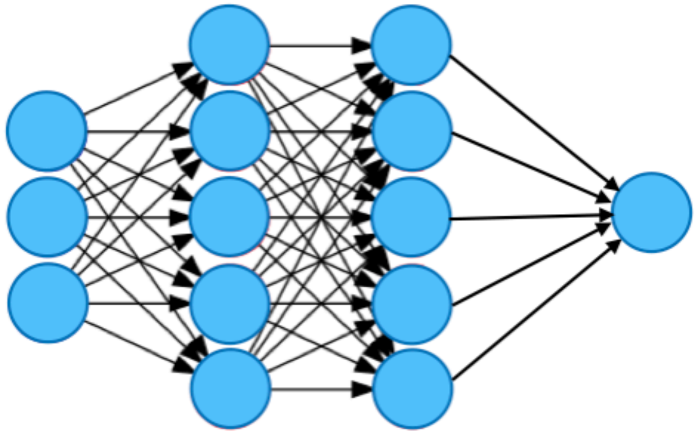
$$X_1, ..., X_n \sim (1-\epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$



$\widehat{\theta} \approx (1-\epsilon)\theta + \epsilon\widetilde{\theta}$



$\widehat{\theta} \approx \theta$



$\widehat{\theta} \approx \theta$

JS-GAN

A classifier with hidden layers leads to robustness. Why?

JS-GAN

A classifier with hidden layers leads to robustness. Why?

$$\mathsf{JS}_g(\mathbb{P},\mathbb{Q}) = \max_{w\in\mathbb{R}^d}\left[\mathbb{P}\log\frac{1}{1+e^{-w^Tg(X)}} + \mathbb{Q}\log\frac{1}{1+e^{w^Tg(X)}}\right] + \log 4.$$

JS-GAN

A classifier with hidden layers leads to robustness. Why?

$$\mathsf{JS}_g(\mathbb{P}, \mathbb{Q}) = \max_{w \in \mathbb{R}^d} \left[\mathbb{P} \log \frac{1}{1 + e^{-w^T g(X)}} + \mathbb{Q} \log \frac{1}{1 + e^{w^T g(X)}} \right] + \log 4.$$

Proposition.

$$\mathsf{JS}_g(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} g(X) = \mathbb{Q} g(X)$$

JS-GAN

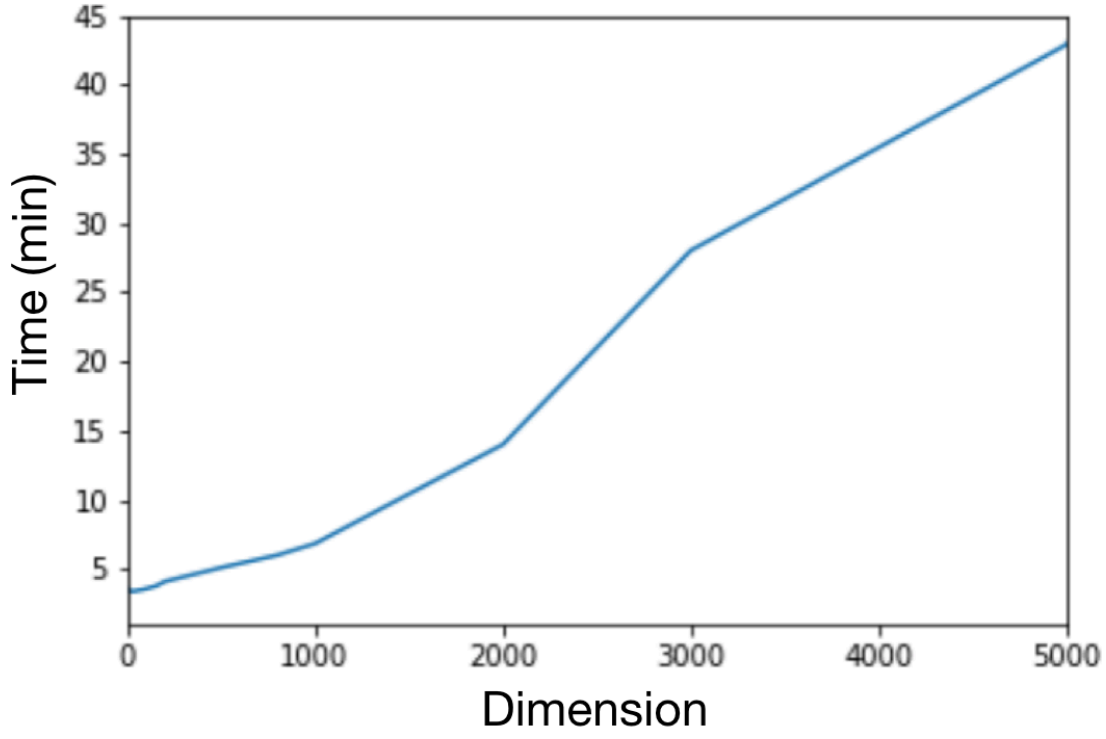
$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

Theorem [GLYZ18+]. For a neural network class $\mathcal{T}$ with at least one hidden layer and appropriate regularization, we have

$$\|\widehat{\theta} - \theta\|^2 \lesssim \begin{cases} \frac{p}{n} + \epsilon^2 & \text{(indicator/sigmoid/ramp)} \\ \frac{p}{n} + \epsilon & \text{(ReLU)} \end{cases}$$

with high probability uniformly over $\theta \in \mathbb{R}^p, Q$.

JS-GAN



Time (min)
Dimension
45
40
35
30
25
20
15
10
5
0
1000
2000
3000
4000
5000

Adaptive Estimation

Robust Density Estimation

$$X_1, ..., X_n \sim (1 - \epsilon) f + \epsilon g$$

Robust Density Estimation

$$X_1, ..., X_n \sim (1-\epsilon) f + \epsilon g$$

$\text{Hölder}(\beta)$

Robust Density Estimation

$$X_1, ..., X_n \sim (1-\epsilon) f + \epsilon g$$

Hölder(β)

arbitrary

Robust Density Estimation

$$X_1, ..., X_n \sim (1-\epsilon)f + \epsilon g$$

$\text{Hölder}(\beta)$

arbitrary

loss function: $|\hat{f}(0) - f(0)|^2$

Robust Density Estimation

$$X_1, ..., X_n \sim (1-\epsilon)f + \epsilon g$$

$\text{Hölder}(\beta)$

arbitrary

Theorem [LG17]. The minimax rate of the problem is given by

$$n^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{2\beta}{\beta+1}}$$

Adaptive Estimation

Theorem [LG17]. When one of (ϵ, β) is known and the other is unknown, the optimal rate is

Adaptive Estimation

Theorem [LG17]. When one of (ϵ, β) is known and the other is unknown, the optimal rate is

$$\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{2\beta}{\beta+1}}$$

Adaptive Estimation

Theorem [LG17]. When one of (ϵ, β) is known and the other is unknown, the optimal rate is

$$\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} \vee \epsilon^{\frac{2\beta}{\beta+1}}$$

adaptation cost: $n \longrightarrow \frac{n}{\log n}$

Adaptive Estimation

Theorem [LG17]. When both (ϵ, β) are unknown,

Adaptive Estimation

Theorem [LG17]. When both (ϵ, β) are unknown,

Adaptation is impossible with any rate!

Adaptive Estimation: Arbitrary Contamination

Definition. *An estimator* $\widehat{f}(0)$ *is called* $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ *rate adaptive if the following holds: for any* $n \geq 1$*, any* $\epsilon \leq 1/2$*, any* $\beta \leq c_1$ *and any* $L \leq c_2$*, we have*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\epsilon,\beta,L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \leq c_3 \left(n^{-r_1(\beta)} \vee \epsilon^{r_2(\beta)}\right).$$

Adaptive Estimation: Arbitrary Contamination

Definition. *An estimator* $\widehat{f}(0)$ *is called* $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ *rate adaptive if the following holds: for any* $n \geq 1$*, any* $\epsilon \leq 1/2$*, any* $\beta \leq c_1$ *and any* $L \leq c_2$*, we have*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\boxed{\epsilon,\beta},L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \leq c_3 \left(n^{-r_1(\beta)} \vee \epsilon^{r_2(\beta)}\right).$$

Adaptive Estimation: Arbitrary Contamination

Definition. *An estimator $\widehat{f}(0)$ is called $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ rate adaptive if the following holds: for any $n \geq 1$, any $\epsilon \leq 1/2$, any $\beta \leq c_1$ and any $L \leq c_2$, we have*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\boxed{\epsilon,\beta},L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \leq c_3\left(n^{-r_1(\beta)} \vee \epsilon^{r_2(\beta)}\right).$$

Lemma. *For any constants $c_1, c_2 > 0$, there exists a constant c_0, such that for any $\beta, \widetilde{\beta} \leq c_1$, and any $L, \widetilde{L} \geq c_2$, and any estimator $\widehat{f}(0)$, one of the following lower bounds must be true,*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\epsilon,\beta,L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \geq c_0 \epsilon^{\frac{2\widetilde{\beta}}{\widetilde{\beta}+1}},$$

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(0,\widetilde{\beta},\widetilde{L})} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \geq c_0 \epsilon^{\frac{2\widetilde{\beta}}{\widetilde{\beta}+1}}.$$

Adaptive Estimation: Arbitrary Contamination

Definition. *An estimator* $\widehat{f}(0)$ *is called* $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ *rate adaptive if the following holds: for any* $n \geq 1$*, any* $\epsilon \leq 1/2$*, any* $\beta \leq c_1$ *and any* $L \leq c_2$*, we have*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\boxed{\epsilon,\beta},L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \leq c_3 \left(n^{-r_1(\beta)} \vee \epsilon^{r_2(\beta)}\right).$$

Lemma. *For any constants* $c_1, c_2 > 0$*, there exists a constant* c_0*, such that for any* $\beta, \widetilde{\beta} \leq c_1$*, and any* $L, \widetilde{L} \geq c_2$*, and any estimator* $\widehat{f}(0)$*, one of the following lower bounds must be true,*

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(\boxed{\epsilon,}\beta,L)} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \geq c_0 \epsilon^{\frac{2\widetilde{\beta}}{\widetilde{\beta}+1}},$$

$$\sup_{(1-\epsilon)f+\epsilon g \in \mathcal{M}(0\boxed{,\widetilde{\beta},}\widetilde{L})} \mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \geq c_0 \epsilon^{\frac{2\widetilde{\beta}}{\widetilde{\beta}+1}}.$$

Summary

Thank You