Robust Estimation and Generative Adversarial Nets

Chao Gao
University of Chicago

@IPAM, May 2019

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]

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

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

Multivariate Location Depth

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

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

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

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

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

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

[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:

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

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:

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

empirical version:

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

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:

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

empirical version:

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

[Mizera, 2002]

Multi-task Regression Depth

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

Multi-task Regression Depth

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

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

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

Multi-task Regression Depth

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

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

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

$m = 1,$

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

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.

Covariance Matrix

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

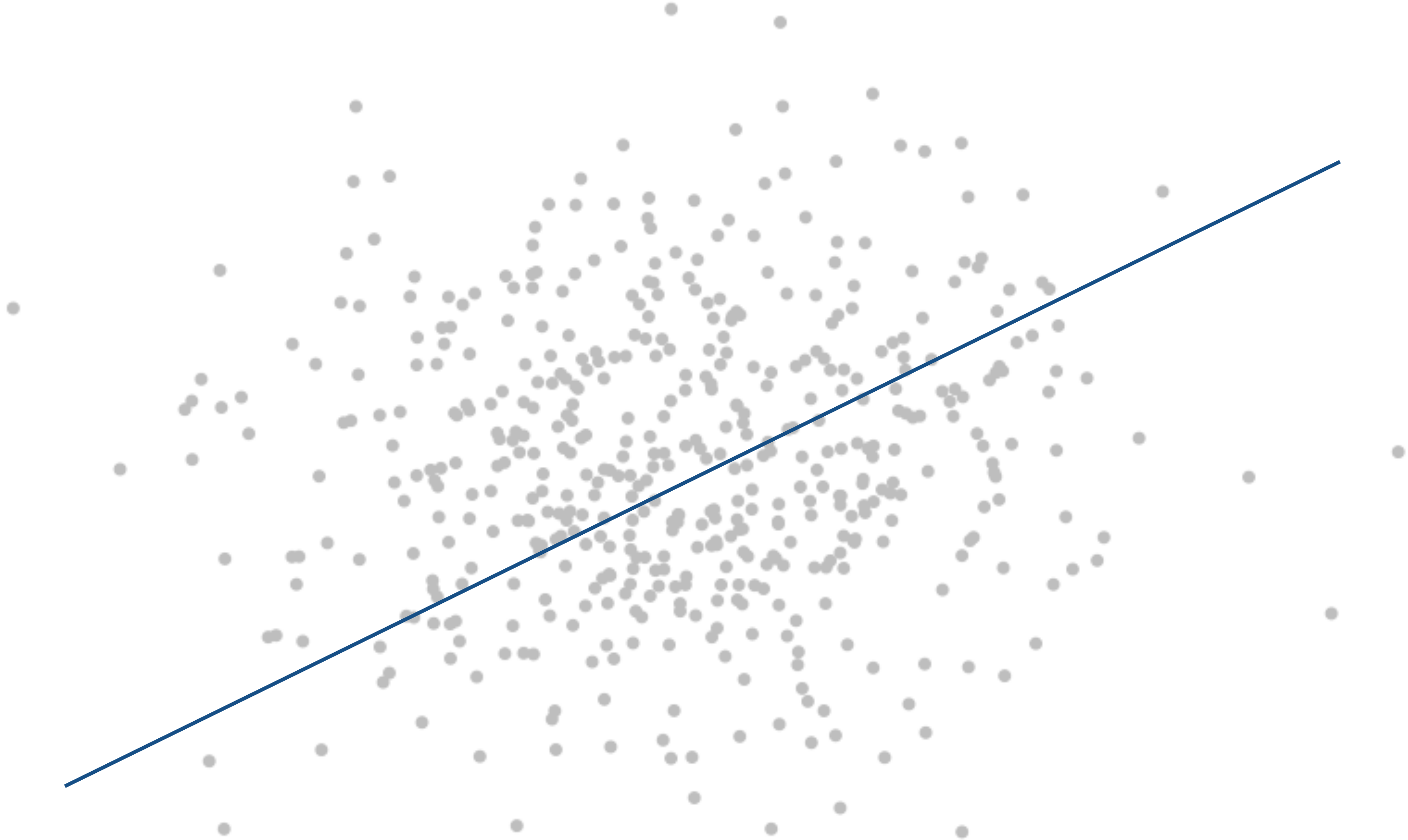
Covariance Matrix

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

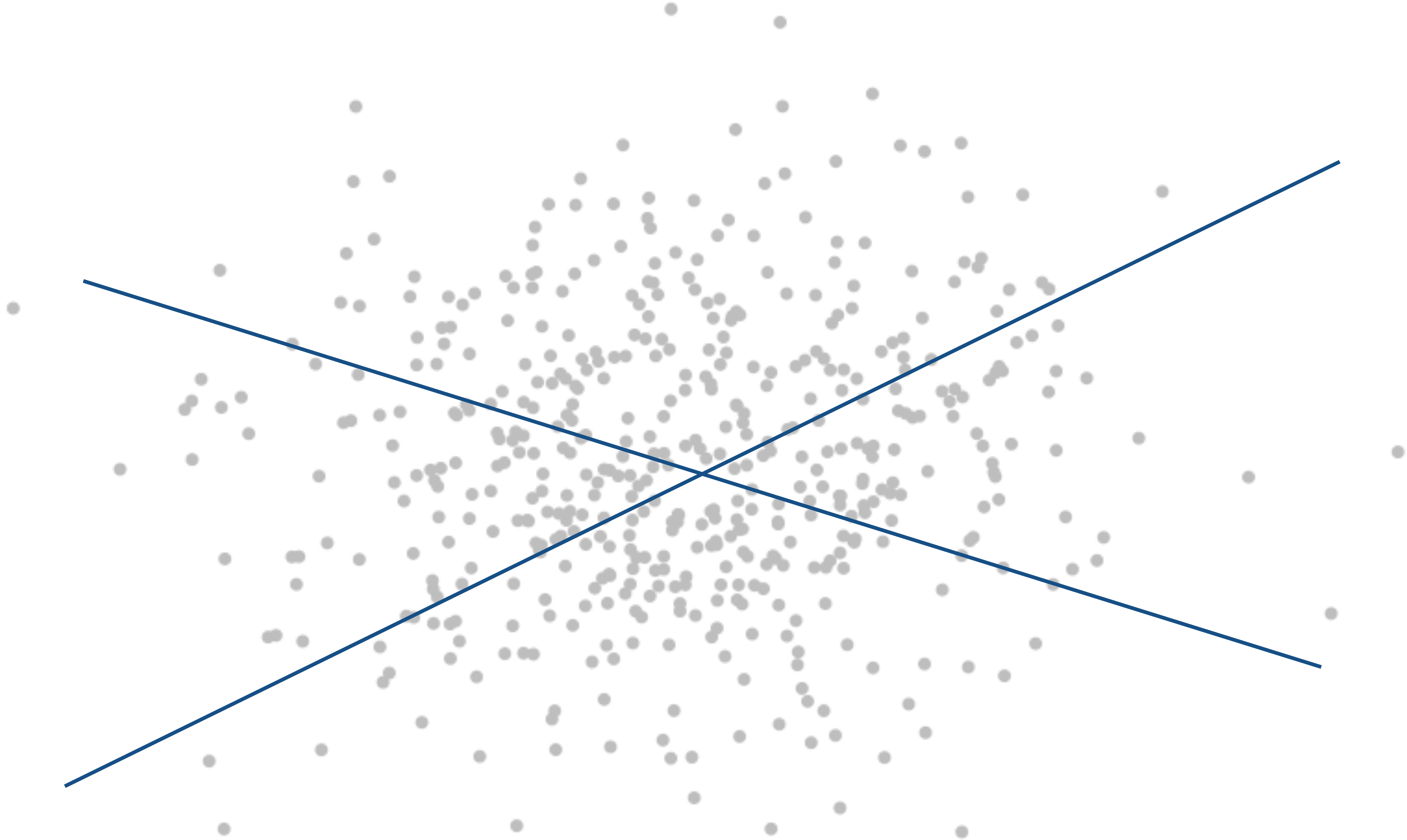
how to estimate ?

Covariance Matrix

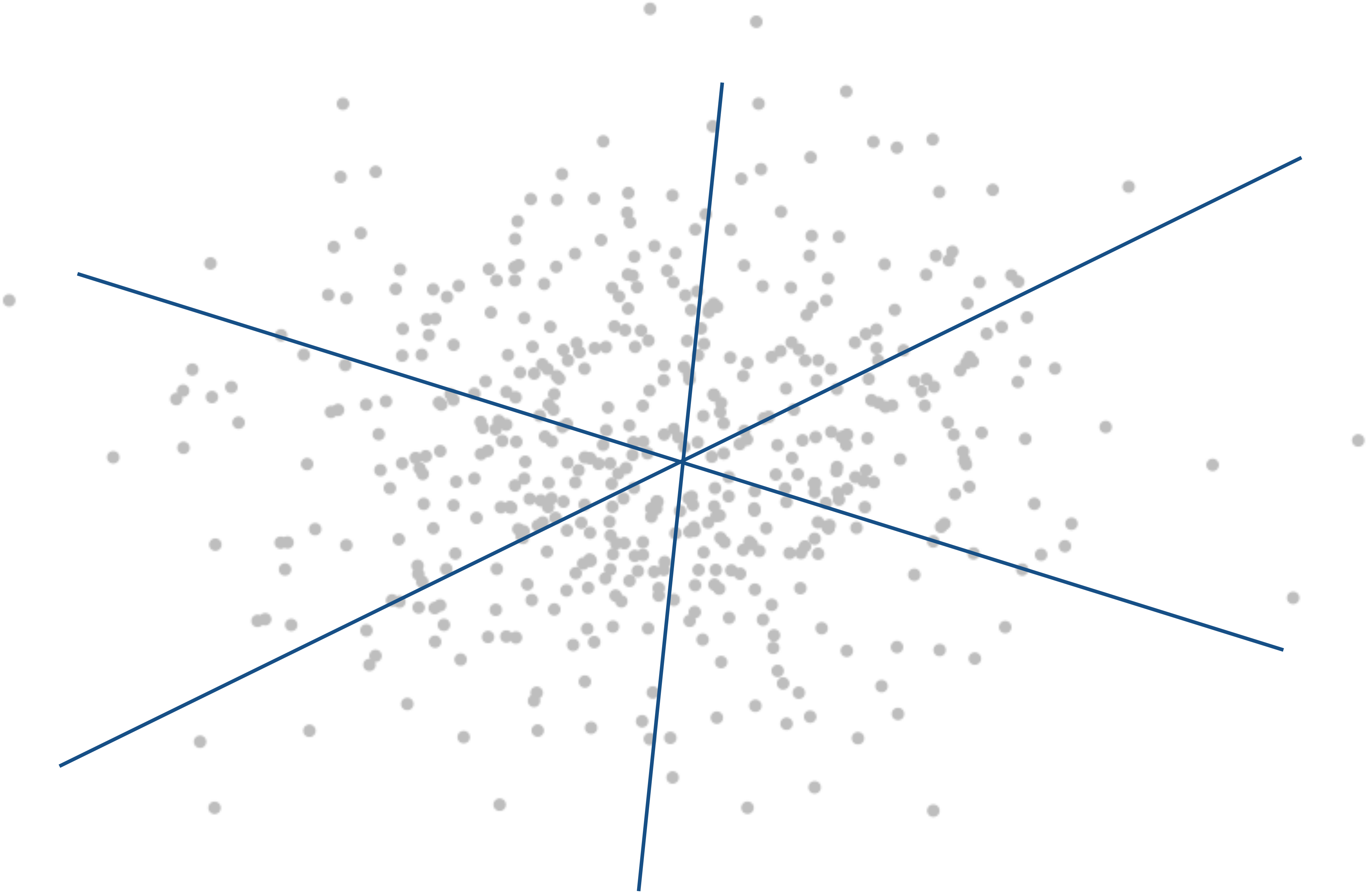
Covariance Matrix



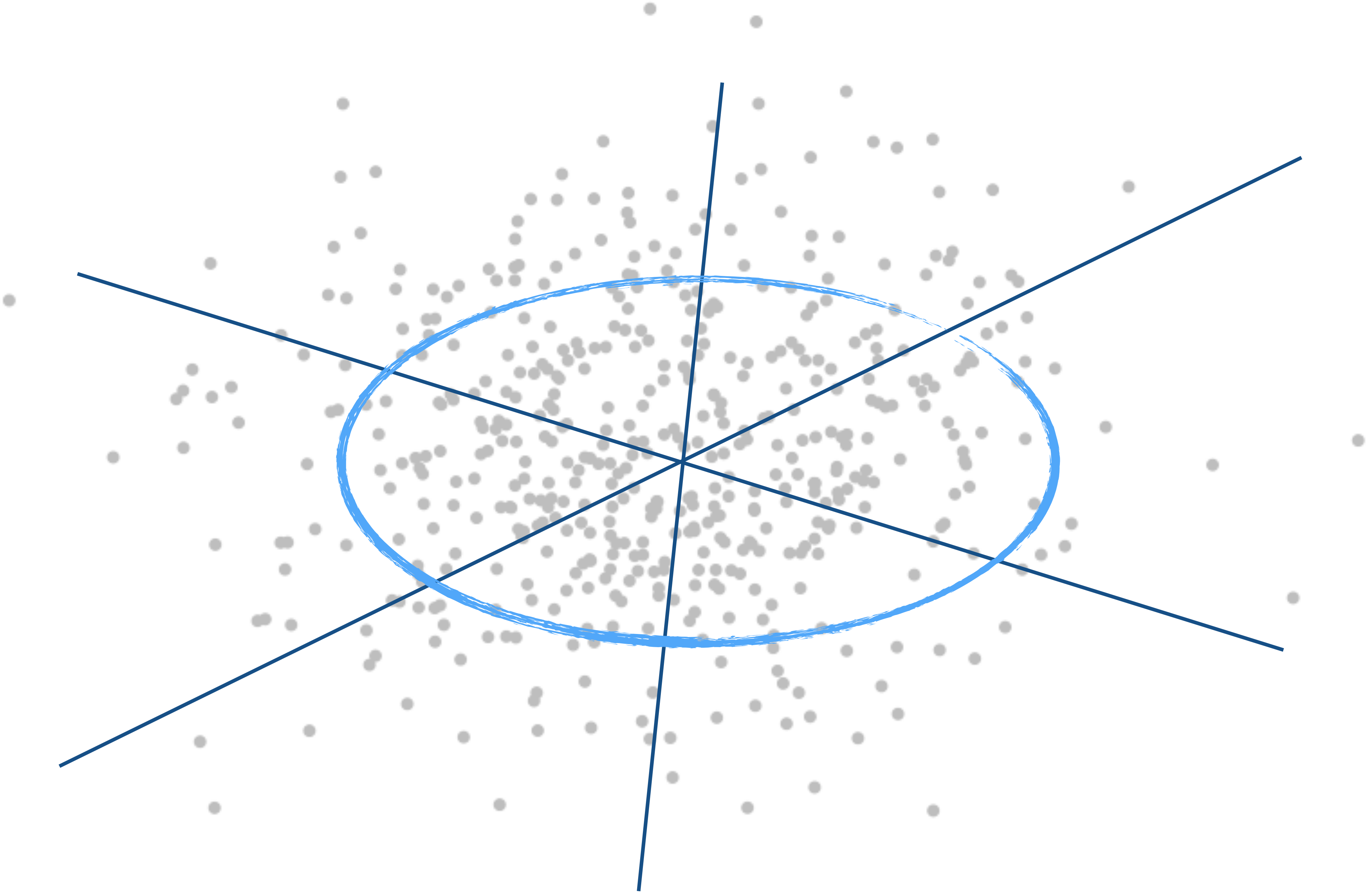
Covariance Matrix



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 $C > 0$,

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

with high probability uniformly over Σ, 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	$\|\cdot\|^2$	$\frac{p}{n} \vee \epsilon^2$
reduced rank regression	$\|\cdot\|_F^2$	$\frac{\sigma^2}{\kappa^2}\frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
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	$\|\cdot\|_F^2$	$\frac{s \log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$

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
Collier and Dalalyan

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

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

[Nguyen, Wainwright, Jordan]

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

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

[Goodfellow et al.

f-Learning

Jensen-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

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

[Goodfellow et al., Baraud and Birge]

f-Learning

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

[Goodfellow et al., Baraud and Birge]

TV-Learning

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

[Yatracos, 1985]

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

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

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

TV-Learning

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

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

$r \to 0$

TV-Learning

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

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

$r \to 0$

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

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

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

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

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

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

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

$r \to 0$

(related to) matrix depth

$$\max_{\Sigma} \min_{\|u\|=1} \left[\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{|u^T X_i|^2 \leq u^T\Sigma u\} - \mathbb{P}(\chi_1^2 \leq 1) \right) \wedge \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{|u^T X_i|^2 > u^T\Sigma u\} - \mathbb{P}(\chi_1^2 > 1) \right) \right]$$

robust statistics community

deep learning community

robust statistics community

f-Learning

f-GAN

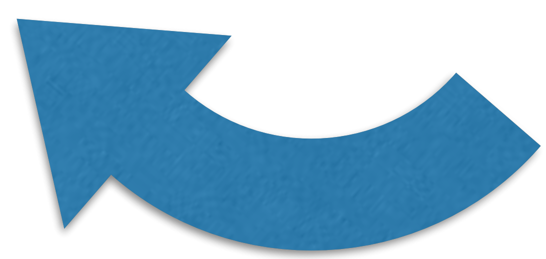
deep learning community

robust statistics community

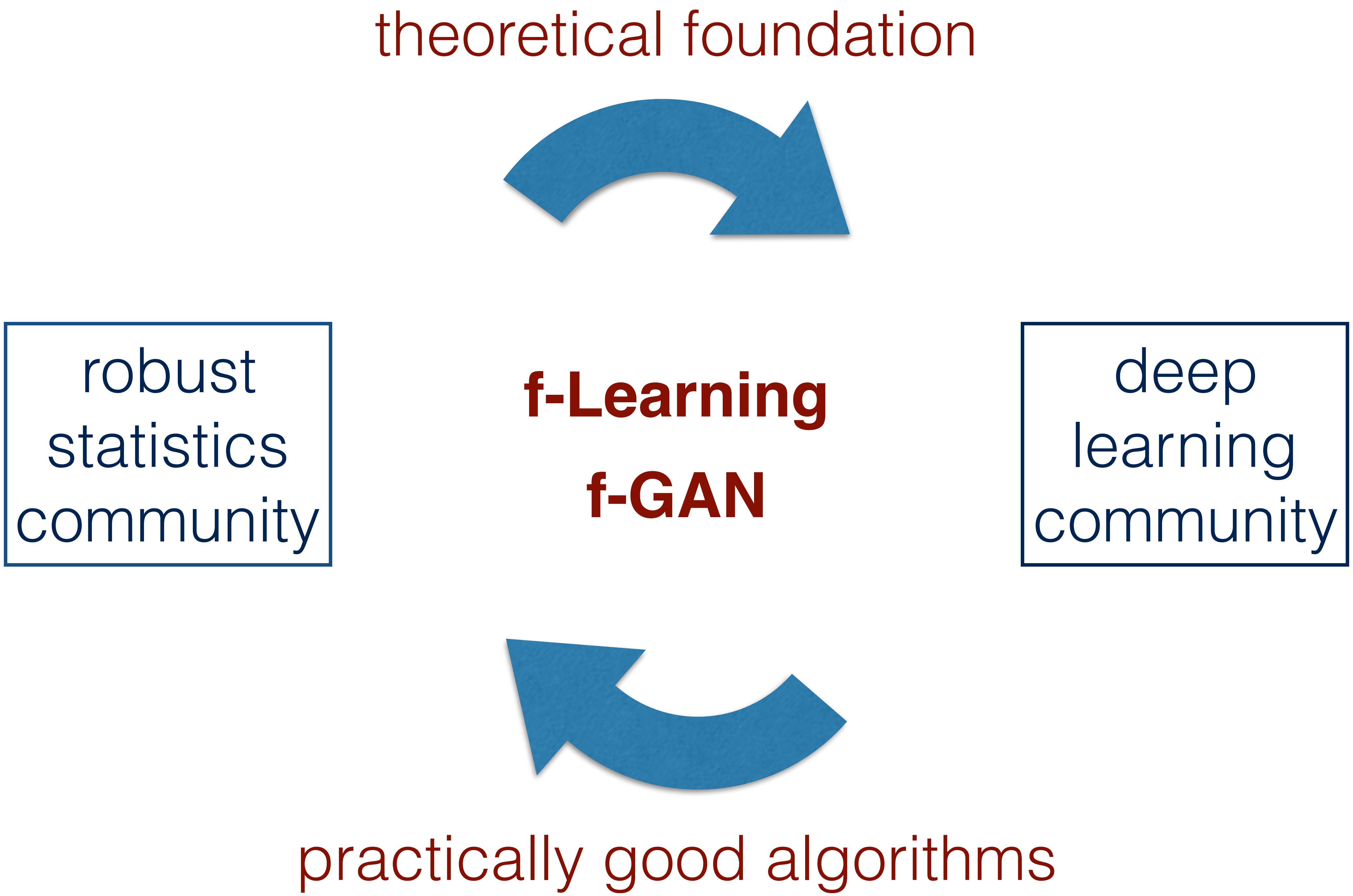
f-Learning

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

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

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

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

$N(\eta, I_p)$

logistic regression classifier

TV-GAN

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

logistic regression classifier

$N(\eta, I_p)$

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

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

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

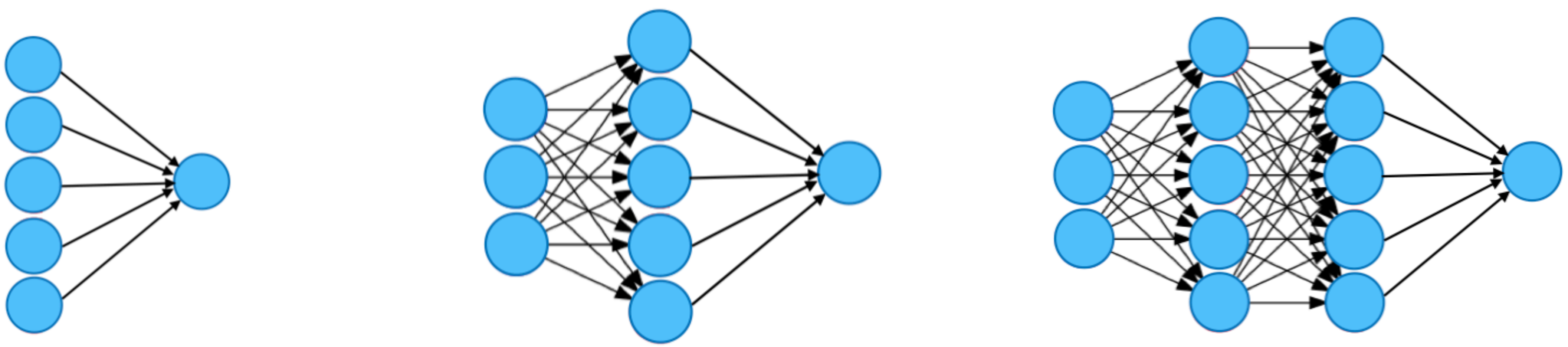
$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$

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

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

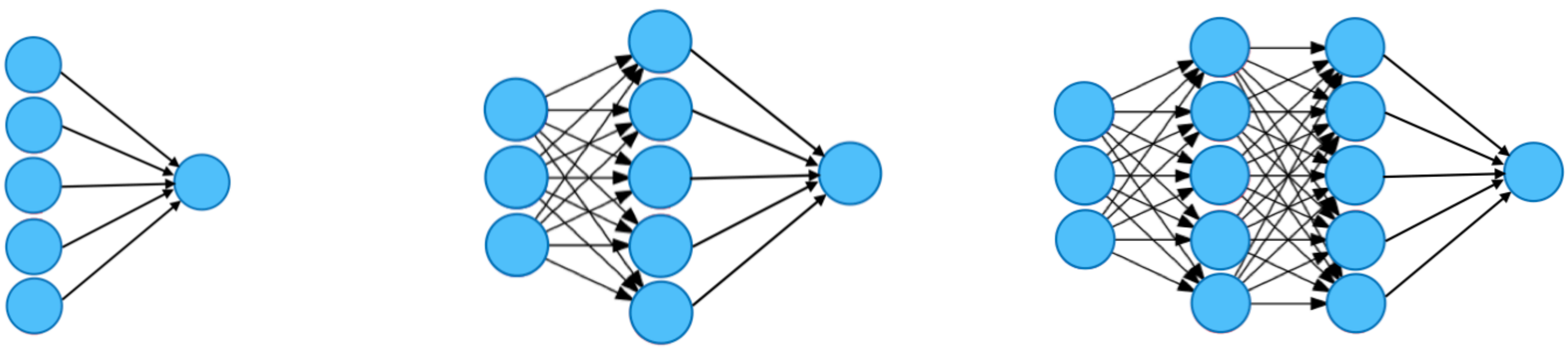


JS-GAN

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

numerical experiment

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



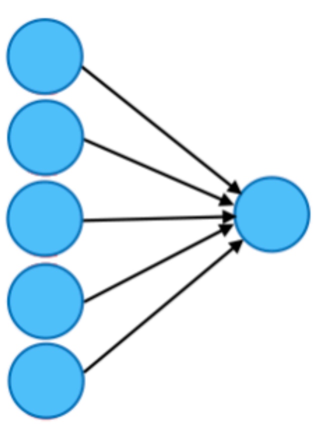
$$\widehat{\theta} \approx (1-\epsilon)\theta + \epsilon \widetilde{\theta}$$

JS-GAN

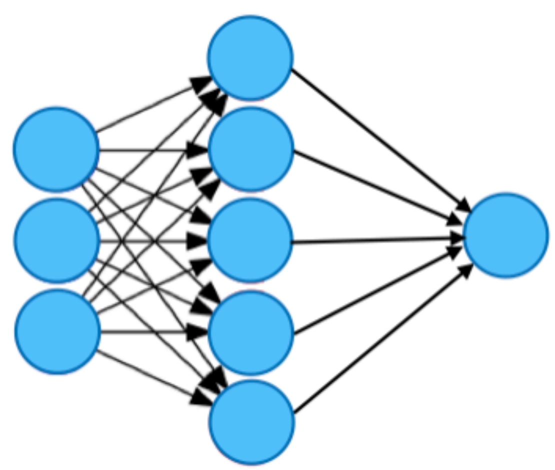
$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\text{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

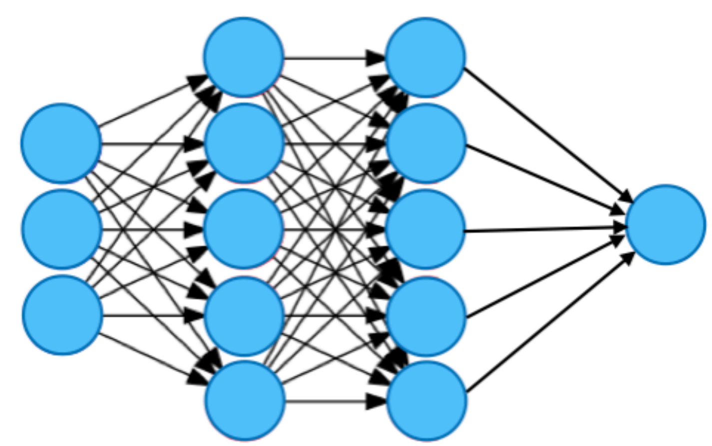
$$X_1, ..., X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$



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



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



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

JS-GAN

A classifier with hidden layers leads to robustness. Why?

JS-GAN

A classifier with hidden layers leads to robustness. Why?

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

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

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

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

JS-GAN

unknown covariance?

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

JS-GAN

unknown covariance?

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

$$(\widehat{\theta}, \widehat{\Sigma}) = \underset{\eta, \Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(\eta, \Gamma)} \log(1 - T(X)) \right]$$

JS-GAN

unknown covariance?

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

$$(\widehat{\theta}, \widehat{\Sigma}) = \underset{\eta,\Gamma}{\text{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(\eta,\Gamma)} \log(1 - T(X)) \right]$$

no need to change the discriminator class

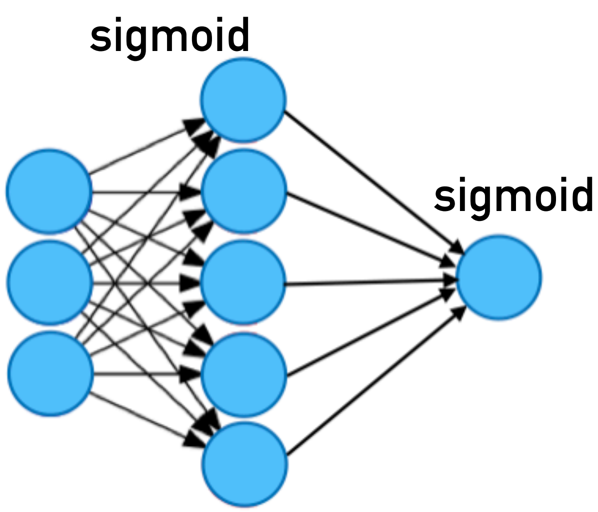
Covariance Matrix

JS-GAN

$$\widehat{\Sigma} = \operatorname*{argmin}_{\Gamma} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

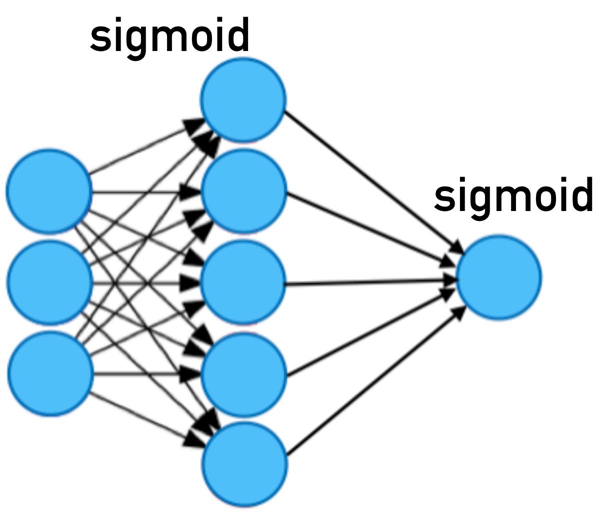
JS-GAN

$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



JS-GAN

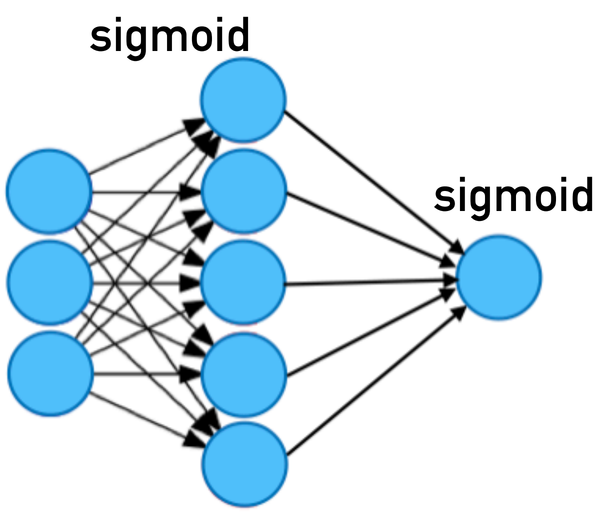
$$\widehat{\Sigma} = \underset{\Gamma}{\mathrm{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



optimal for mean estimation

JS-GAN

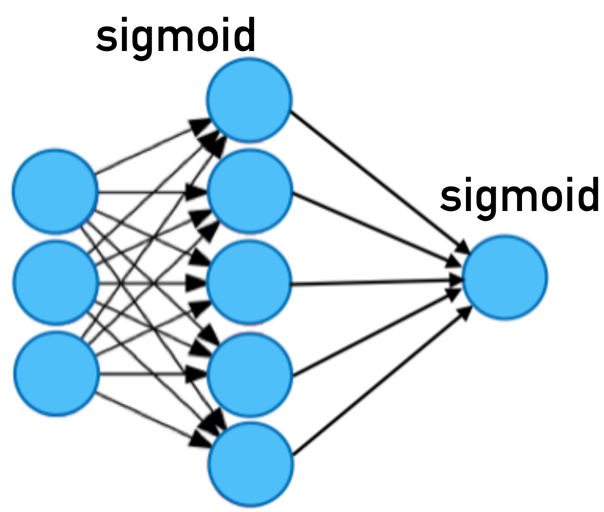
$$\widehat{\Sigma} = \underset{\Gamma}{\mathrm{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



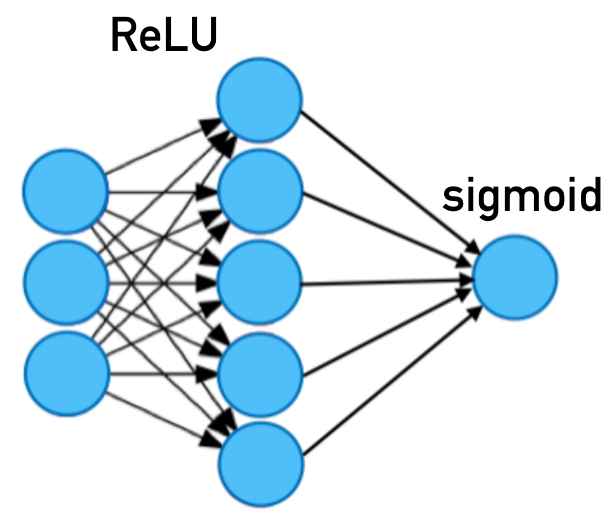
optimal for mean estimation but **inconsistent** for covariance estimation

JS-GAN

$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

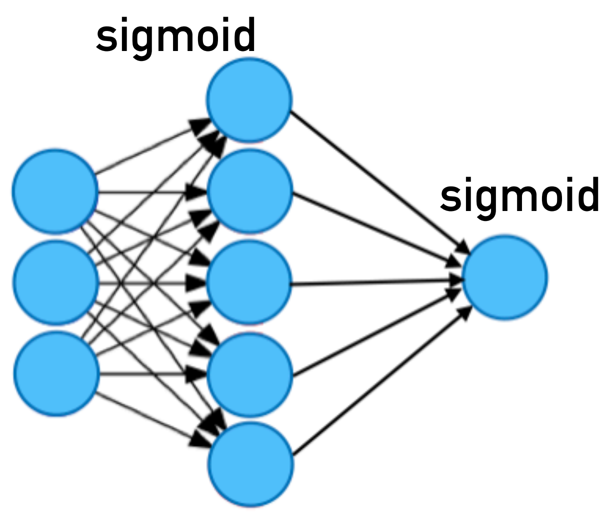


optimal for mean estimation but **inconsistent** for covariance estimation

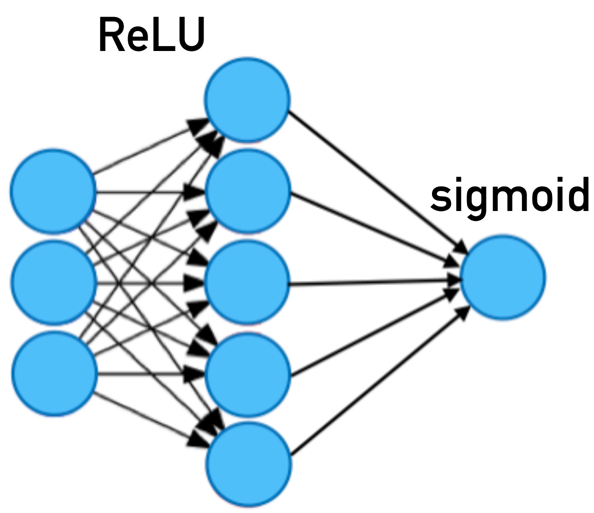


JS-GAN

$$\widehat{\Sigma} = \operatorname*{argmin}_{\Gamma} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



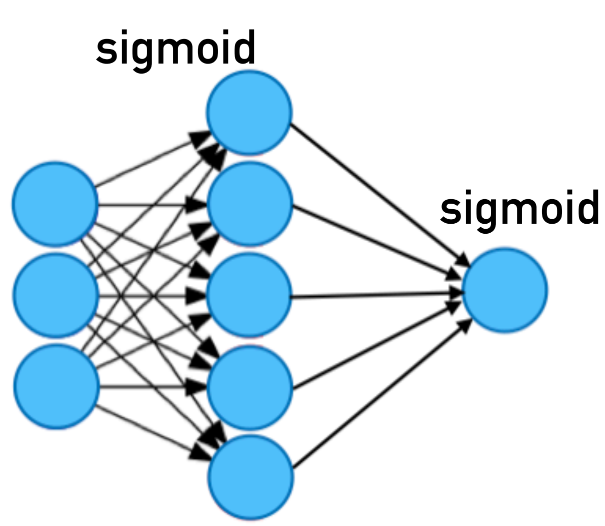
optimal for mean estimation but **inconsistent** for covariance estimation



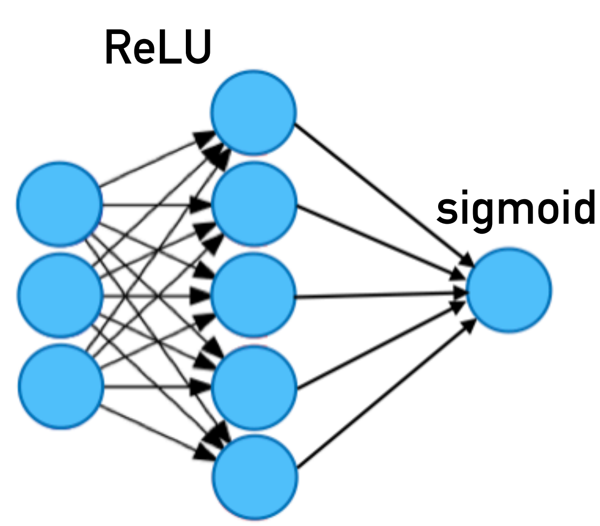
optimal without contamination

JS-GAN

$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



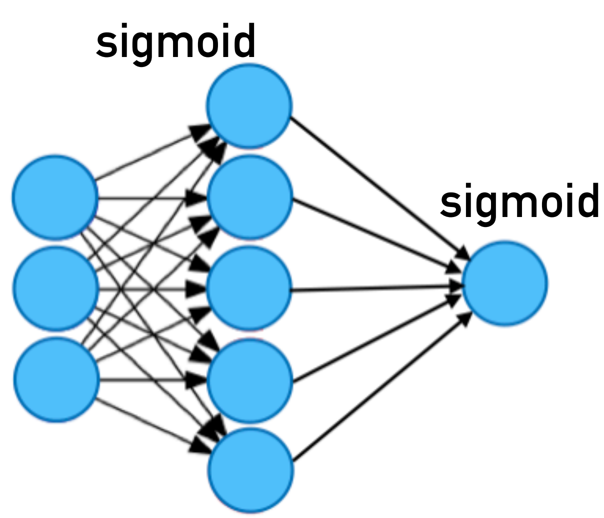
optimal for mean estimation but **inconsistent** for covariance estimation

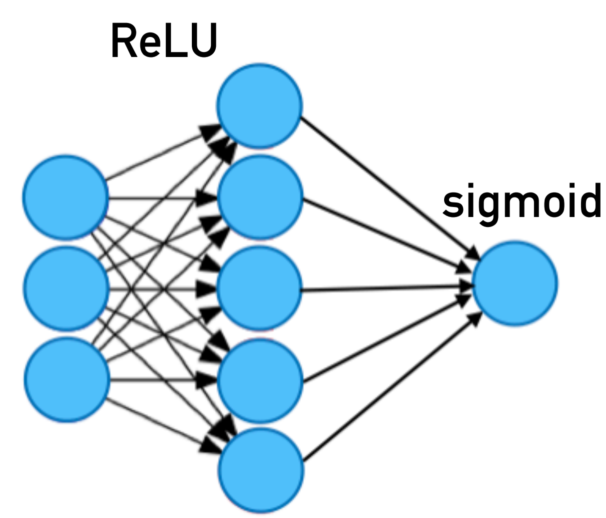


optimal without contamination but **not robust**

JS-GAN

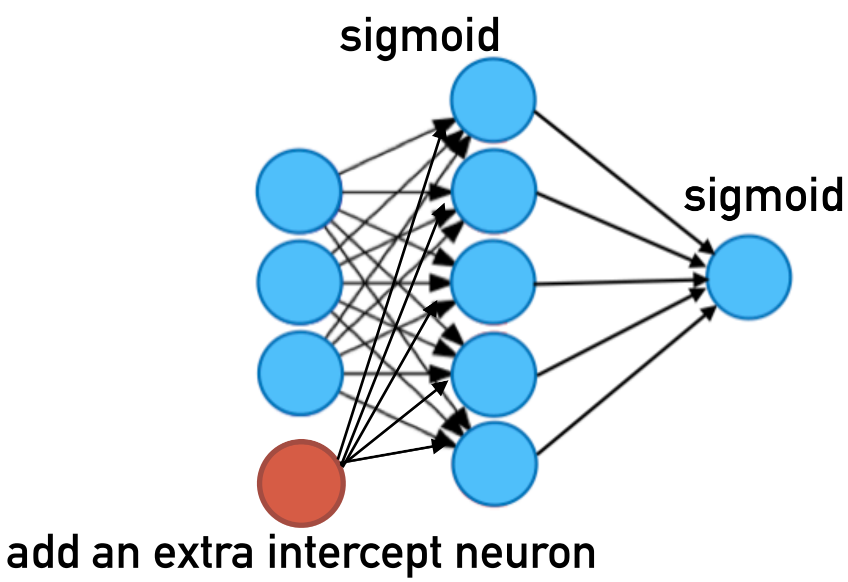
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

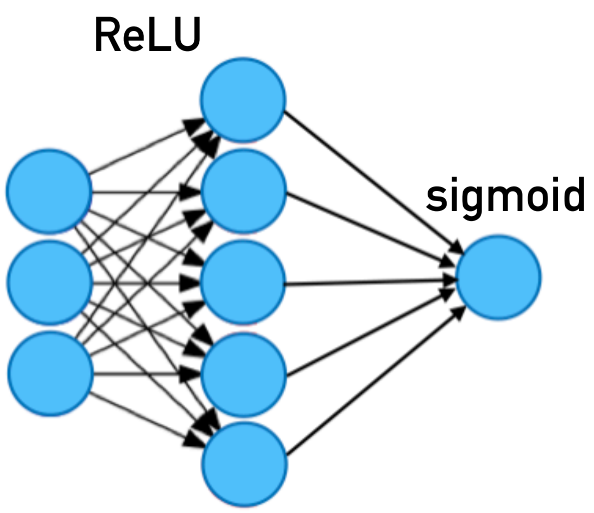




JS-GAN

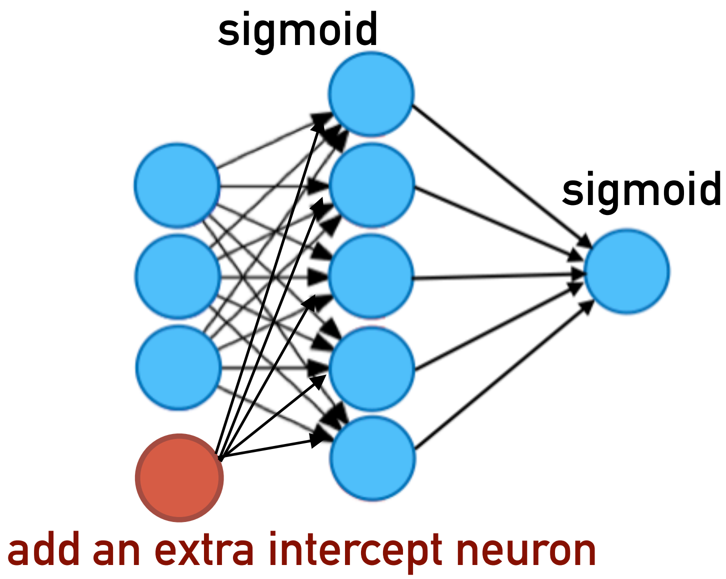
$$\widehat{\Sigma} = \underset{\Gamma}{\mathrm{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

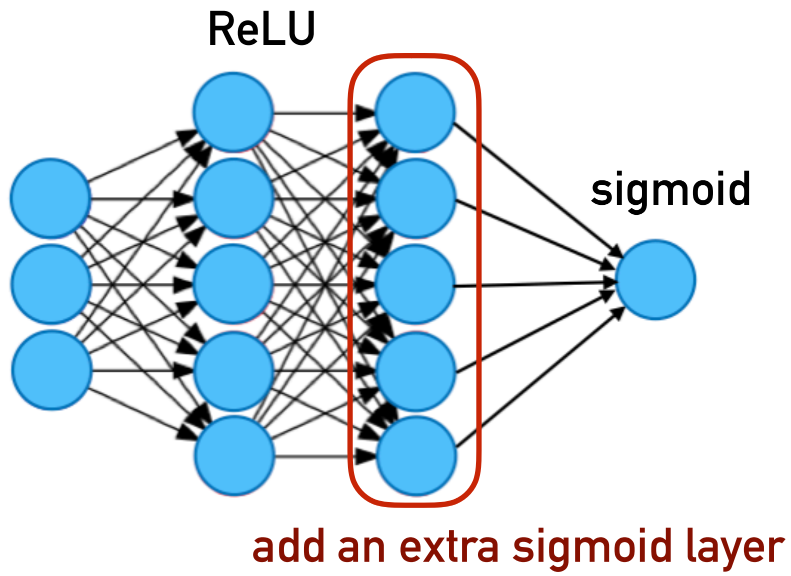




JS-GAN

$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$





JS-GAN

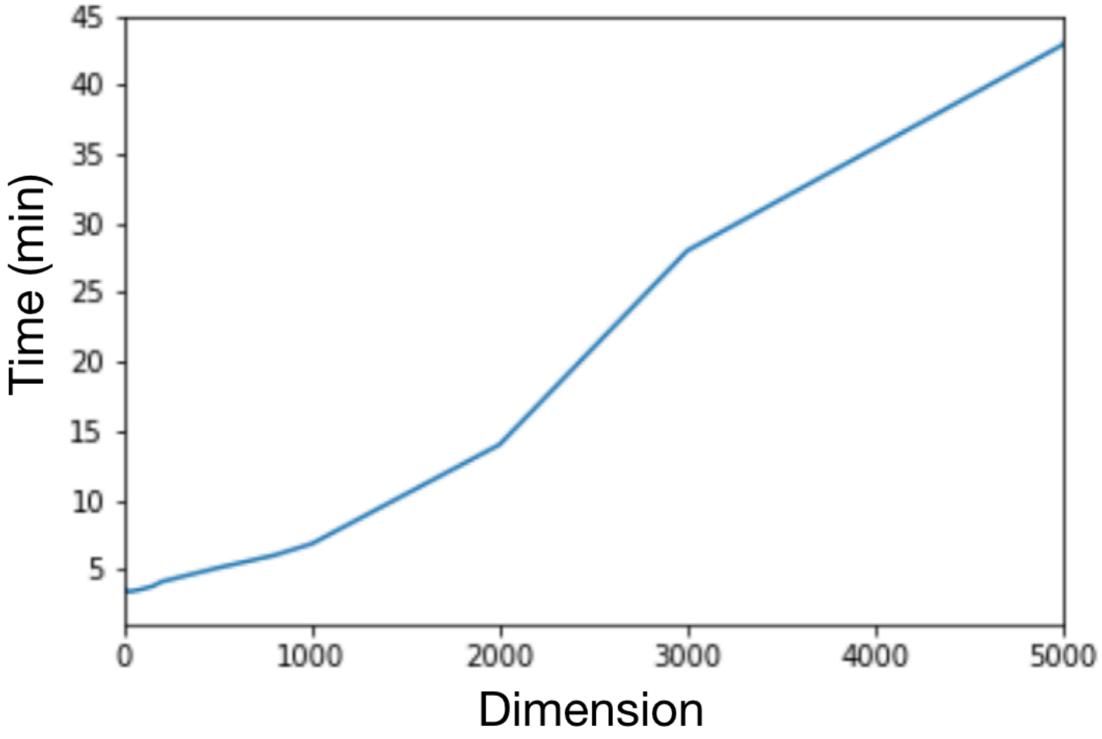
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

Theorem [GYZ19]. For the above two neural network classes, we have

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}}^2 \lesssim \begin{cases} \frac{p}{n} + \epsilon^2 & \text{(2-layer sigmoid with intercept)} \\ \frac{p \log p}{n} + \epsilon^2 & \text{(3-layer ReLU)} \end{cases}$$

with high probability uniformly over Σ, Q .

JS-GAN



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

Thank You