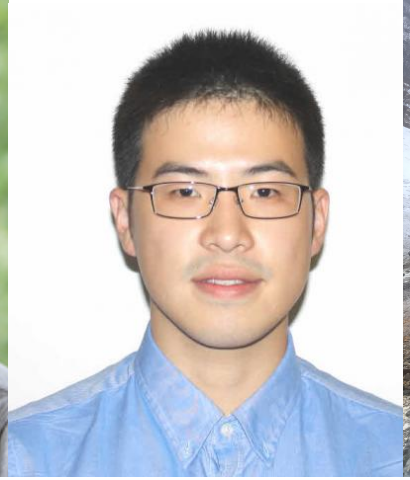
Robust Estimation and Generative Adversarial Nets

Yuan YAO
HKUST

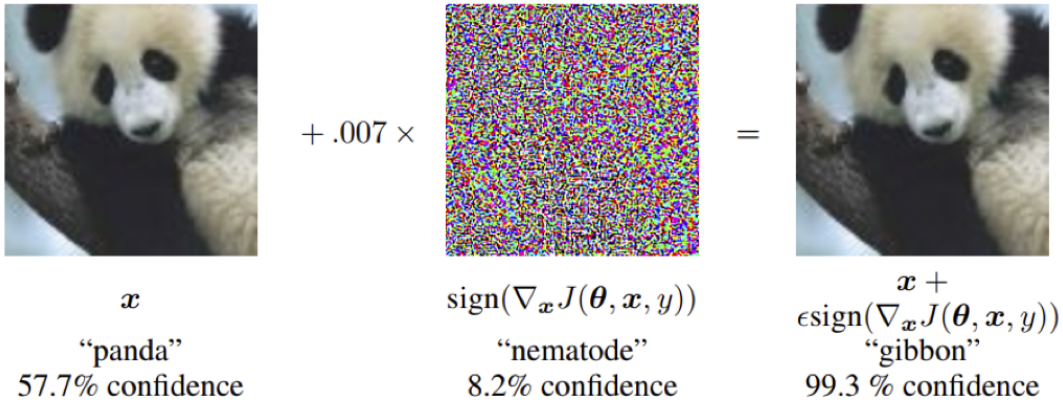
Chao Gao (Chicago)



Jiyu Liu (Yale)

Weizhi Zhu (HKUST)

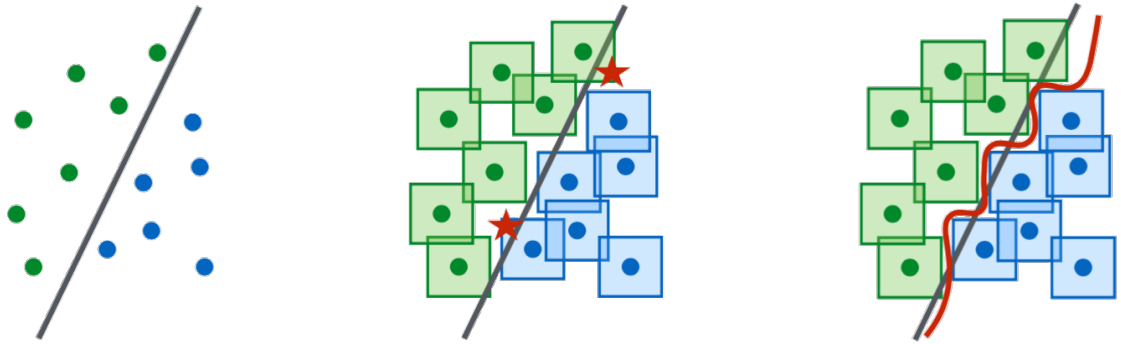
Deep Learning is Notoriously Not Robust!



[Goodfellow et al., 2014]

- Imperceivable adversarial examples are ubiquitous to fail neural networks
- How can one achieve robustness?

Robust Optimization



- Traditional training:

$$\min_{\theta} J_n(\theta, \mathbf{z} = (x_i, y_i)_{i=1}^n)$$

  - e.g. square or cross-entropy loss as negative log-likelihood of logit models

- Robust optimization:

$$\min_{\theta} \max_{\|\epsilon_i\| \le \delta} J_n(\theta, \mathbf{z} = (x_i + \epsilon_i, y_i)_{i=1}^n)$$

  - robust to any distributions, yet perhaps too conservative

Distributional Robust Optimization

- Distributional Robust Optimization:

$$\min_{\theta} \max_{\epsilon} \mathbb{E}_{\mathbf{z} \sim P_\epsilon \in \mathcal{D}}[J_n(\theta, \mathbf{z})]$$

  - $\mathcal{D}$ is a set of ambiguous distributions, e.g. Wasserstein ambiguity set
  - intermediate approach with statistically contaminated distributions
  - *sometimes, contamination might be unstructured...*

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

parameter of interest

arbitrary contamination

[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 ?

Robust Maxmum-Likelihood Does not work!

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

Robust Maxmum-Likelihood Does not work!

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

$$\ell(\theta, Q) = \text{negative log-likelihood} = \sum_{i=1}^{''} (\theta - X_i)^2$$

$$\sim (1-\epsilon)\mathbb{E}_{\mathcal{N}(\theta)}(\theta - X)^2 + \epsilon \mathbb{E}_Q(\theta - X)^2$$

the sample mean

$$\hat{\theta}_{mean} = \frac{1}{n}\sum_{i=1}^{n} X_i = \arg\min_{\theta} \ell(\theta, Q)$$

$$\min_{\theta}\max_{Q} \ell(\theta, Q) \geq \max_{Q}\min_{\theta} \ell(\theta, Q) = \max_{Q} \ell(\hat{\theta}_{mean}, Q) = \infty$$

Medians

1. **Coordinatewise median**

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

Medians

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]

	Coordinatewise Median	Tukey's Median
statistical precision (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$
statistical precision (with contamination)	$\frac{p}{n} + p\epsilon^2$	$\frac{p}{n} + \epsilon^2$: minimax [Chen-Gao-Ren'15]
computational complexity	Polynomial	NP-hard [Amenta et al. '00]

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

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

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

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

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

[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:

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

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

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

Multi-task Regression Depth

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

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

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

Multi-task Regression Depth

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

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

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

$m = 1,$

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

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.

Covariance Matrix

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

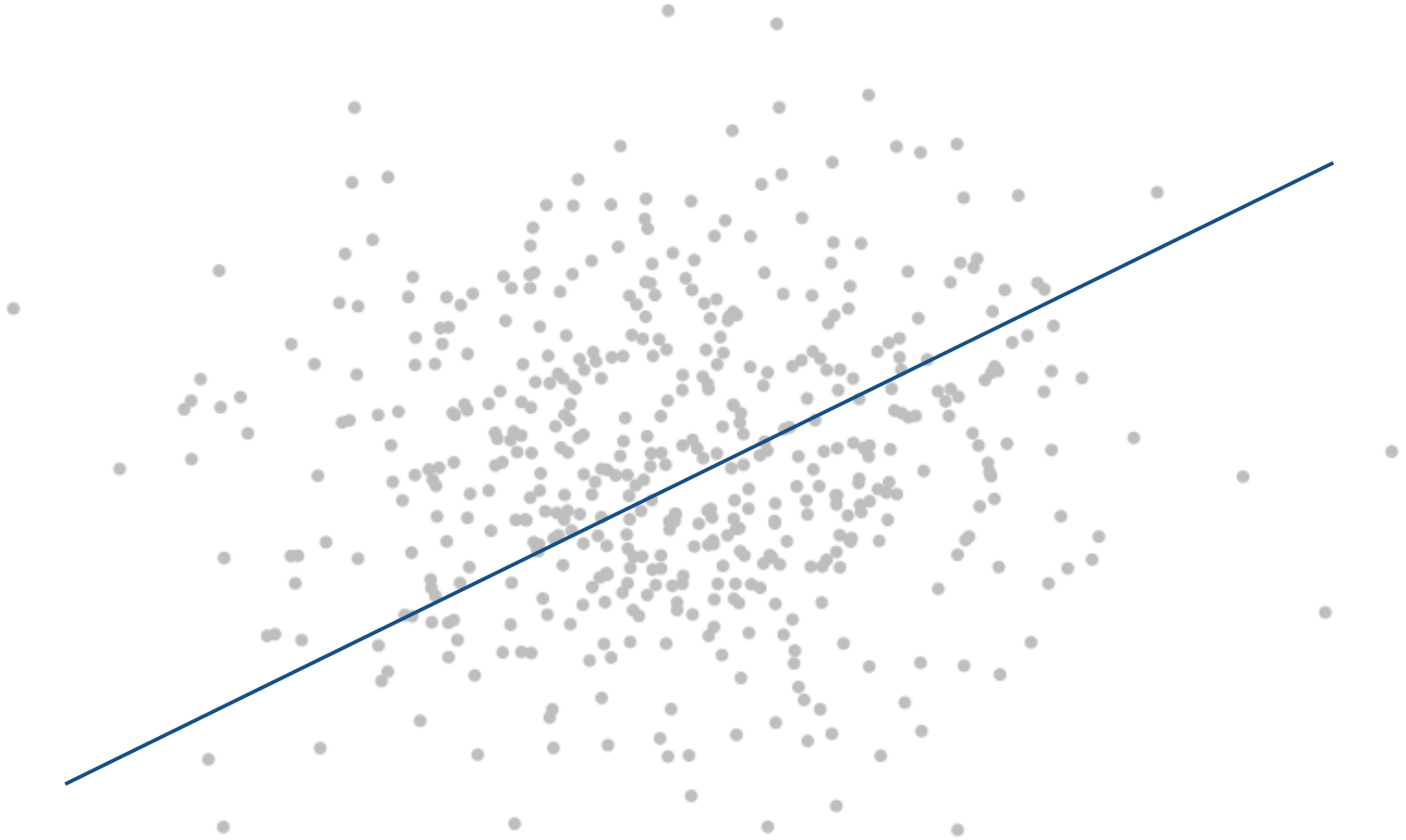
Covariance Matrix

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

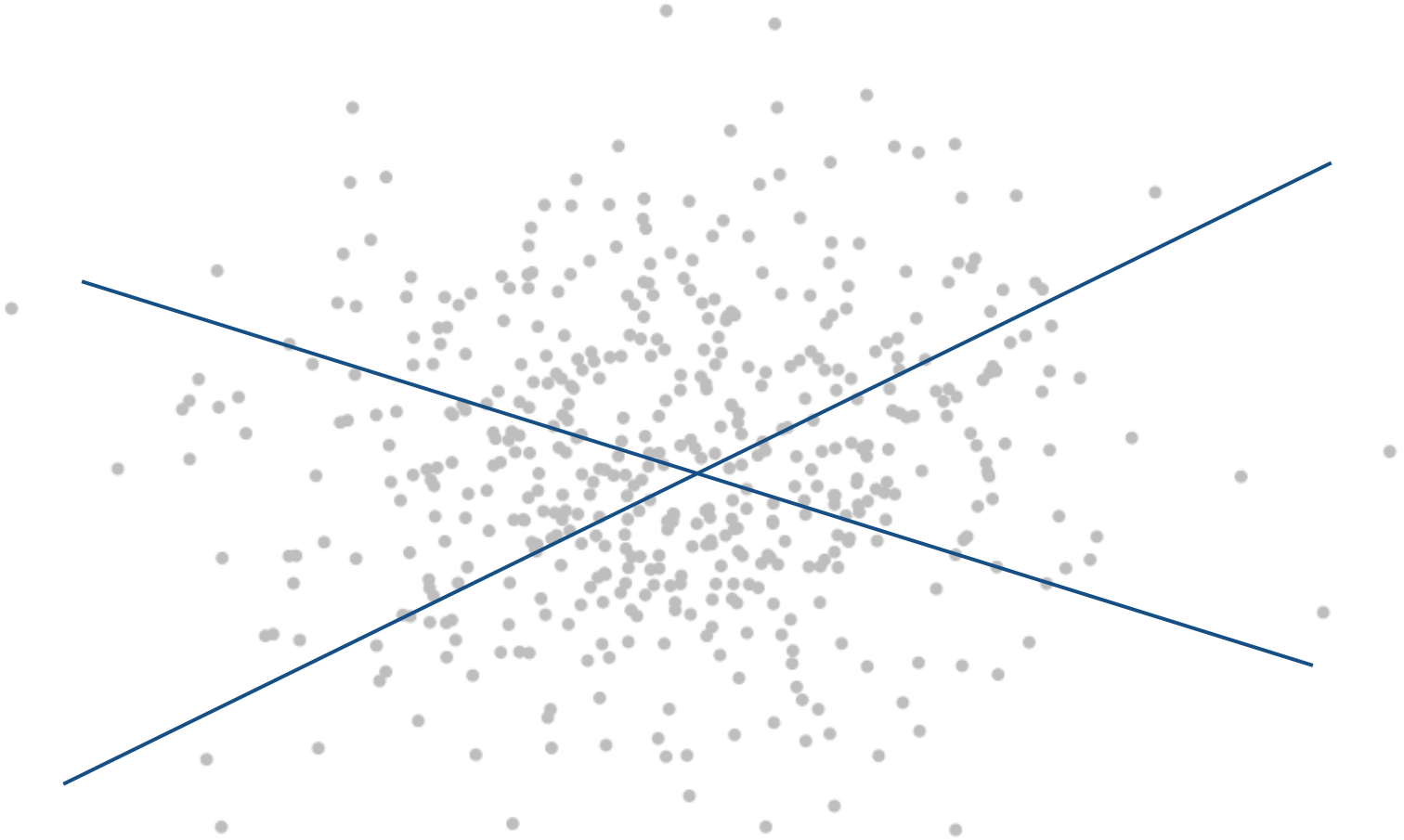
how to estimate ?

Covariance Matrix

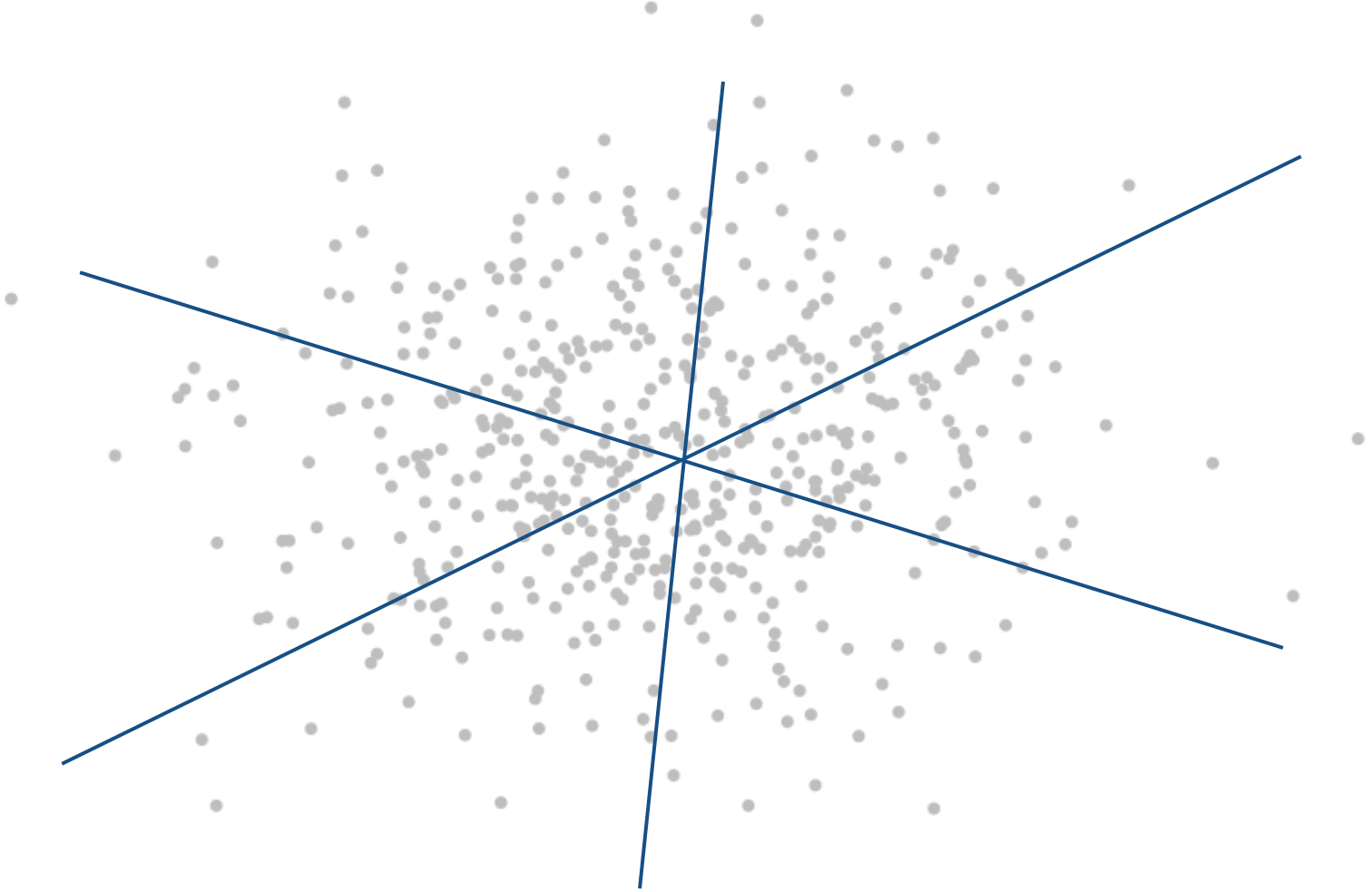
Covariance Matrix



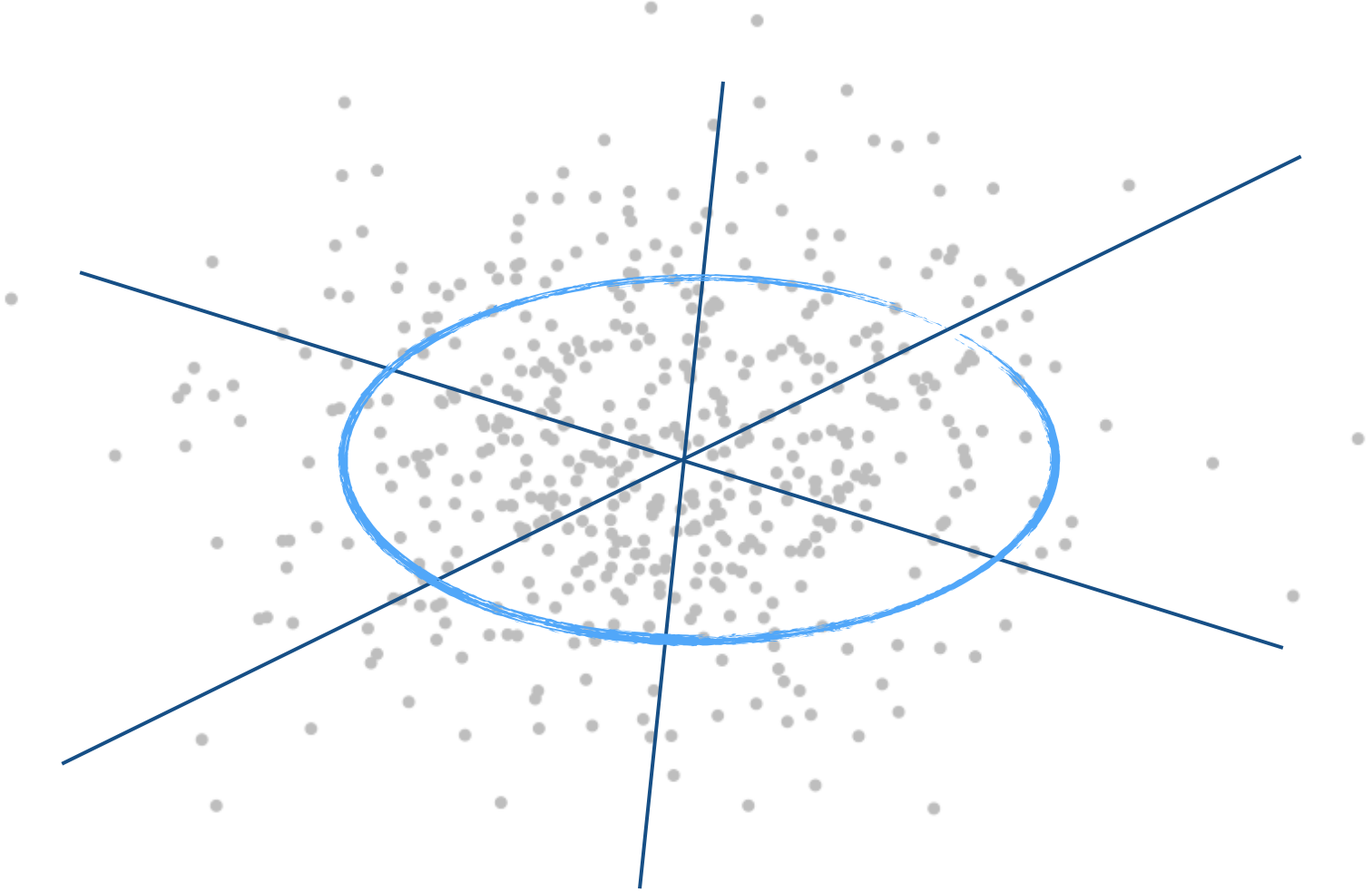
Covariance Matrix



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	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{\sigma^2}{\kappa^2}\frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
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	$\|\cdot\|_{\mathrm{F}}^2$	$\frac{s\log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$