A Brief Tutorial on OLS and Geometry.

OLS → Ordinary Least Squares

$\theta^* \in \mathbb{R}^d$

Observations: $x_i = \langle A_i, \theta^* \rangle + \eta_i \quad , \quad i = 1, 2, \dots, n$

$A_i \to \in \mathbb{R}^d$

$\eta_i \to$ 1-subG noise.

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$A = \overset{\longleftarrow d \longrightarrow}{\begin{pmatrix} A_1^T \\ \vdots \\ A_n^T \end{pmatrix}} \; \Big\updownarrow n$$

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{n} \left(x_i - \langle A_i, \theta \rangle \right)^2 + \lambda \|\theta\|^2$$

x_i → observations (r.vs).

A_i → known (deterministic)

$\lambda\|\theta\|^2$ → if $\lambda > 0$, regularized least squares.

$$\boxed{\hat{\theta} = \text{estimate}}$$

$(\hat{\theta} - \theta^*)$ is a random vector in $\mathbb{R}^d$.

Facts:

① $$\hat{\theta} = V^{-1} \sum_{i=1}^{n} A_i x_i = V^{-1} A^T \cdot X$$

where $$V = \sum_{i=1}^{n} A_i A_i^T = (A^T \cdot A).$$

$\therefore \quad \hat{\theta} = (A^T A)^{-1} A^T \cdot x$, OLS estimator.

---

② $$E\left[(\hat{\theta} - \theta^*) \cdot (\hat{\theta} - \theta^*)^T \right] = V^{-1}$$

↳ covariance matrix

with $E[\hat{\theta}] = \theta^*$

---

③ V^{-1} is a real, symmetric p.d. matrix,

i.e., $V^{-1} = U^T \Lambda U$ $\quad (U^T = U^{-1})$

↪ unitary matrix

↳ diagonal matrix with $\lambda_i > 0 \quad \forall\, i = 1, .., d$.

(Assumption: $\{A_1, .., A_n\}$ spans $\mathbb{R}^d$)

$\{\lambda_i\}$ are the variance terms along the basis vectors $\{u_i\}$.

---

④ Confidence Ellipsoid = C.

$$C = \{ x \in \mathbb{R}^d : \; x^T V x \leq 1 \}.$$

$$= \{ x \in \mathbb{R}^d : \; x^T (U^T \Lambda U)^{-1} x \leq 1 \}$$

$$U^{-1} \Lambda^{-1} (U^T)^{-1}$$

$= U^{T}$ $= U$

$$= \{x \in \mathbb{R}^d : x^T U^T \Lambda^{-1} U x \leq 1\}$$

$$= \{x \in \mathbb{R}^d : (Ux)^T \Lambda^{-1} (Ux) \leq 1\}$$

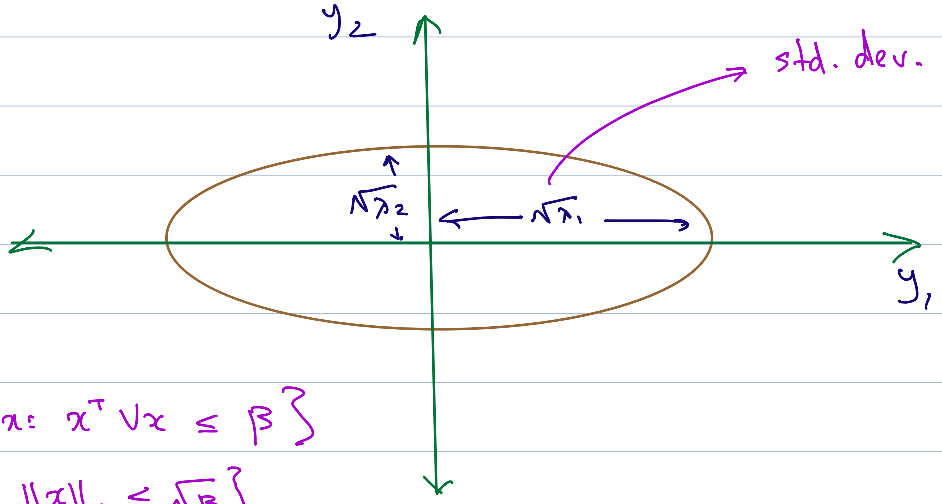
$$= \{y \in \mathbb{R}^d : y^T \Lambda^{-1} y \leq 1\}$$

rotate coordinate axes to the $\{u_i\}$ basis.

$$= \left\{y \in \mathbb{R}^d : \sum_{i=1}^{d} \left(\frac{y_i}{\sqrt{\lambda_i}}\right)^2 \leq 1\right\}$$

i.e., $|y_i| \leq \sqrt{\lambda_i}$, $i = 1, \ldots, d$

and



Note: $\{x : x^T V x \leq \beta\}$
$= \{x : \|x\|_V \leq \sqrt{\beta}\}$.

i.e., axes of ellipsoid are the standard deviations in the new basis $\{u_i\}_{i=1}^{d}$.

Thus, $\{x \in \mathbb{R}^d : x^T V x \leq \beta\}$ is

a confidence ellipsoid, where the confidence level can be chosen by scaling $\beta > 0$ appropriately.

⑤ Volume of this confidence ellipsoid:

$$C = \{x \in \mathbb{R}^d : x^T V x \leq 1\}.$$

$$\text{Vol}(C) = \text{Volume}(\text{Unit-Sphere}) \sqrt{\lambda_1 \cdots \lambda_d}$$

(i.e., product of axes, scaled by the volume of unit sphere).

$$\text{Vol}(\text{Unit Sphere}) = \left(\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} \right).$$

Gamma function

Now, recall V^{-1} is p.d. $\quad$ det(·) = determinant(·)

$$\therefore \quad \det(V^{-1}) = \left(\prod_{i=1}^{d} \lambda_i \right).$$

and $$\det(V) = \frac{1}{\left(\prod_{i=1}^{d} \lambda_i\right)},$$

$\therefore$

$$\text{minimize } \mathrm{Vol}\,(\overset{\substack{\text{confidence ellipsoid}\\ \downarrow}}{C}) \equiv \text{maximize } \det(V)$$

$$\equiv \text{maximize } \log\det(V).$$

---

e.g.: When we study Kiefer-Wolfowitz Thm, we will maximize

$$f(\pi) = \log\det(V(\pi)),$$

where $V(\pi)$ is a real, symmetric p.d. matrix, and π is a finite-dim. vector. The intuition for this problem comes from minimizing the volume of a confidence ellipsoid associated with an unbiased estimator.

---

⑥ $Z = \langle \hat{\theta} - \theta^*, x \rangle, \qquad x \in \mathbb{R}^d$

i.e., Z is the projection of the estimation error along x.

Notice:

ⓐ $Z = \text{Linear-Transformation}(\eta_1, \eta_2, \cdots, \eta_n)$

measurement noise; $\eta_i \sim$ 1-subGaussian

i.e., $E\left[e^{\alpha \eta_i}\right] \leq e^{\alpha^2/2}$, $\alpha \in \mathbb{R}$.

ⓑ $$E[Z] = E\left[\langle \hat{\theta} - \theta^*, x \rangle\right]$$

$$= \langle \underbrace{E[\hat{\theta} - \theta^*]}_{=0}, x \rangle = 0.$$

$= 0$, unbiased estimator.

$$\text{Var}(Z) = E\left[Z^2\right]$$

$$= E\left[x^T (\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T x\right]$$

$$= x^T V^{-1} x = \|x\|^2_{V^{-1}}.$$

Thus, $Z \sim \|x\|_{V^{-1}}$ - subGaussian, i.e.,

for any $\delta \in (0,1)$, we have

$$P\left(Z \geq \sqrt{2 \|x\|_{V^{-1}}^{2} \ln(1/\delta)}\right) \leq \delta.$$
