Lecture 2

Interlude: Joint Gaussian variables

Random variables $\underline{A}_1, \cdots, \underline{A}_m$ on same prob. space are jointly Gaussian if $\exists \mu, \Sigma$,
$(A_1, \cdots, A_m) \sim N(\mu, \Sigma)$.

Fact 1: If vars are jointly Gaussian, then so is any linear transformation of them

$$\underline{A}_1,\ \underline{A}_2,\ \underline{A}_3,\quad \underline{A}_1 + \underline{A}_2,\quad \underline{A}_1 - \underline{A}_2,\quad 0.1\underline{A}_1 + 10\underline{A}_2 - 5\underline{A}_3$$

Fact 2: Jointly Gaussian $\underline{A}, \underline{B}$ indep. $\Leftrightarrow$ uncorrelated

$$\mathbb{E}[\underline{A}\cdot\underline{B}] = \mathbb{E}[\underline{A}]\cdot\mathbb{E}[\underline{B}] \Leftrightarrow \underline{A} \perp \underline{B}$$

Minimax optimality of LS

design $\quad X \in \mathbb{R}^{n\times d}, \quad \frac{1}{n}X^T X = I_d$

hidden $\quad \underline{\beta} \sim N(0, t\cdot I_d)$ (prior)

$\qquad\qquad \underline{w} \sim N(0, I_n)$ (noise)

$\qquad\qquad \underline{\beta} \perp \underline{w}$ indep.

observation $\quad \underline{y} = X\underline{\beta} + \underline{w}$ $\qquad$ Bayes-optimal estimator, $\hat{\underline{\beta}}_{mean} = \mathbb{E}[\underline{\beta} \mid \underline{y}]$

remains to show

$$\mathbb{E}\left[\|\hat{\underline{\beta}}_{mean} - \underline{\beta}\|^2\right] = \frac{d}{1+1/t}$$

$$\boxed{\textbf{Lemma} \quad \{\underline{\beta} \mid \underline{y}\} = N\Big(\underbrace{\tfrac{1}{n+1/t}X^T y}, \tfrac{1}{n+1/t} I_d\Big) \qquad \text{Connection from LS: } \hat{\underline{\beta}}_{LS} = \tfrac{1}{n}X^T y}$$

Proof: (lemma) Consider "diagonal" case $X = \begin{pmatrix}\sqrt{n}\cdot I_d \\ 0\end{pmatrix}$ w.l.o.g.

so $\underline{y}_i = \sqrt{n}\,\underline{\beta}_i + \underline{w}_i$ for $i \in [d]$

Claim: $\underline{y}_i \perp \underbrace{\underline{\beta}_i - \alpha\cdot\underline{y}_i}$, for $\alpha \ldots$

$$\underline{\beta} = \alpha\cdot\underbrace{\begin{pmatrix}y_1 \\ \vdots \\ y_d\end{pmatrix}}_{\mu(\underline{y})} + \underline{\underline{z}}, \quad \underline{\underline{z}} \perp \underline{y}$$

$\underline{\beta} = \alpha\begin{pmatrix}y_1 \\ y_d\end{pmatrix} + \underline{\beta} - \cancel{\alpha y}$

$*$ β_i, w are Gaussian
so y_i is also Gaussian because of jointly Gaussian

$$\{\underline{\beta} - \mu(\underline{y}) \mid \underline{y}\} = \{\underline{z}\}, \quad \underline{z} \overset{\text{Fact 1}}{\sim} N(\mathbb{E}[\underline{z}], \mathbb{E}[\underline{z}\underline{z}^T])$$

$\underline{z}_i = (1 - \alpha\sqrt{n})\beta_i - \alpha\cdot w_i$ $\quad * \; z = \beta_i - \alpha\cdot y_i = \beta_i - \alpha(\sqrt{n}\beta_i + w_i)$
$= \beta_i - \alpha\sqrt{n}\,\beta_i - \alpha w_i = \beta_i(1 - \alpha\sqrt{n}) - \alpha\cdot w_i$

- $\mathbb{E}[\underline{z}] = (1 - \alpha\sqrt{n})\underbrace{\cancel{\mathbb{E}[\beta_i]}}_{0} - \alpha\cdot\underbrace{\cancel{\mathbb{E}[w_i]}}_{0} = 0$
- $\mathbb{E}[zz^T] = (1 - \alpha\sqrt{n})^2\,\mathbb{E}[\underline{\beta}\underline{\beta}^T]$ $\quad * \; zz^T = (1 - \alpha\sqrt{n})^2 \beta\beta^T$
$\quad - 2(1-\alpha\sqrt{n})\cdot\alpha\cdot\underbrace{\cancel{\mathbb{E}[\beta]}}_{0}\cdot\underbrace{\cancel{\mathbb{E}[w]}}_{0}$ $\quad * \beta \perp w \quad -2(1-\alpha\sqrt{n})\beta\cdot\alpha w$

$+ \alpha^2 \cdot \mathbb{E}[WW^T]$ $\quad + \alpha^2 WW^T$

$= (1-\alpha\sqrt{n})^2 t \cdot I_d + \alpha^2 \cdot I_d = \alpha' \cdot I_d \quad * \ \underline{\beta} \sim N(0, t \cdot I_d),\ W \sim N(0, I_n)$

$\longrightarrow \{\underline{\beta} - \mu(\underline{y}) \mid \underline{y}\} = \{\underline{z}\} = N(0, \alpha' \cdot I_d)$

$\longrightarrow \{\underbrace{\underline{z} + \mu(\underline{y})}_{} \mid \underline{y}\} = N(\mu(\underline{y}), \alpha' \cdot I_d) \longrightarrow$ Let $\alpha = \frac{\sqrt{n}}{n + 1/t}$ then

$z = \beta - \mu(y)$

$\beta - \mu(y) + \mu(y) = \beta$

$\alpha' = t(1-\alpha\sqrt{n})^2 + \alpha^2$

$= t\left(1 - \frac{n}{n+1/t}\right)^2 + \frac{n}{(n+1/t)^2}$

$= t\left(1 + \frac{n^2}{(n+1/t)^2} - \frac{2n}{n+1/t}\right) + \frac{n}{(n+1/t)^2}$

$= t + \frac{tn^2}{(n+1/t)^2} - \frac{2tn}{n+1/t} + \frac{n}{(n+1/t)^2}$

$= \frac{1}{n+1/t}$???

Proof (claim)

$\underline{z}_i = (1-\alpha\sqrt{n})\underline{\beta}_i - \alpha \cdot \underline{W}_i \qquad *$ Remember $\underline{y}_i = \sqrt{n}\,\underline{\beta}_i + \underline{W}_i$

How to choose α such that $\underline{y}_i \perp z_i$?

Correlation: $\mathbb{E}[z_i \cdot y_i] = \sqrt{n} \cdot (1-\alpha\sqrt{n}) \cdot \mathbb{E}[\beta_i^2] - \alpha \cdot \mathbb{E}[w_i^2]$

$= t \cdot \sqrt{n}(1-\alpha\sqrt{n}) - \alpha$

$* \ z_i \cdot y_i = \sqrt{n}(1-\alpha\sqrt{n})\beta_i^2 \ \cancel{- \sqrt{n}\,\alpha \cdot W_i \beta_i} \ \cancel{+ (1-\alpha\sqrt{n})\beta_i W_i} - \alpha \cdot W_i^2$ 0

y_i, z_i uncorrelated $\iff \alpha = t \cdot \sqrt{n}(1-\alpha\sqrt{n}) = t\sqrt{n} - \alpha t n$

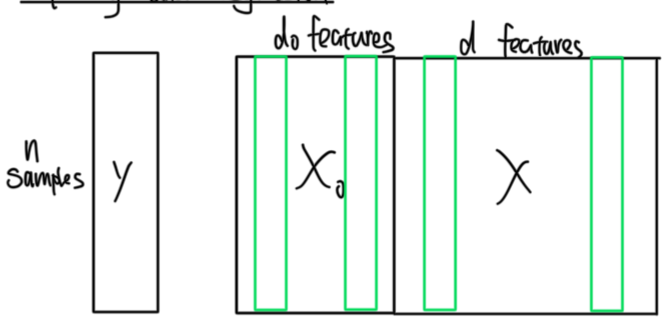
$\iff \alpha + \alpha t n = t\sqrt{n} \iff \alpha(1+tn) = t\sqrt{n}$

$\iff \alpha = \frac{\sqrt{n}}{n+1/t} \iff \underline{y}_i \perp z_i$

$* \ \mathbb{E}[\underline{z}_i] = \mathbb{E}[\underline{y}_i] = 0 \quad \underline{z} \perp \underline{y}$

y_i, z_i are jointly Gaussian (linear transformations of jointly Gaussian $\underline{\beta}, \underline{W}$)

Sparsity and regression



linear model: minimax optimal error is $\frac{d}{n}$

more features also needs more data to reduce the error.

Only small number of features are relevant/useful. $k \ll d$.

Can achieve error $\frac{k}{n} \ll \frac{d}{n}$

Sparse linear model

known $X \in \mathbb{R}^{n \times d}$, $k \in [d]$ $\quad (k \ll n \ll d)$

hidden $\beta^0 \in \mathbb{R}^d$ $\quad$ k-sparse $\quad (\leq k$ non-zeros$)$

observe $\underline{y} = X\beta^0 + \underline{W}$ $\quad$ for $\underline{W} \sim N(0, I_n)$

Brute force (best-subset-selection BSS)

$$\hat{\beta} = \underset{\beta}{\arg\min} \left\{ \|X\beta - y\|^2 \mid \beta \in \mathbb{R}^d \ \text{ k-sparse} \right\}$$

Theorem $\forall \beta^\circ \in \mathbb{R}^d$ k-sparse, w.h.p.

$$\underbrace{\frac{1}{n}\|X(\hat{\beta} - \beta^\circ)\|^2 \le \underbrace{\frac{k}{n}}_{\text{error if relevant features known}} \cdot \underbrace{O\left(\log \frac{2d}{k}\right)}_{\text{cost of not knowing relevant features}}}_{\text{turns out to be statistically necessary}}$$

Proof (Theorem)

Objective function: $f(\beta) = \frac{1}{2n}\|X\beta - y\|^2 = \frac{1}{2n}(X\beta - y)^T(X\beta - y) = \frac{1}{2n}\big(\underbrace{\beta^T X^T X \beta}_{2X^TX\beta} - \underbrace{2y\beta^T X^T}_{2X^Ty} + \underbrace{y^T y}_{0}\big)$

gradient: $\nabla f(\beta^\circ) = \frac{1}{n}X^T(X\beta^\circ - y) = \frac{1}{n}X^TX\beta^\circ - \frac{1}{n}X^Ty = \cancel{\frac{1}{n}X^TX\beta^\circ} - \cancel{\frac{1}{n}X^TX\beta^\circ} - \frac{1}{n}X^T w$

$$= -\frac{1}{n}X^T w$$

behavior around β°: * $f(\beta^\circ + u) = \frac{1}{2n}\|y - X(\beta^\circ + u)\|^2 = \frac{1}{2n}\|(X\beta^\circ + w) - X\beta^\circ - Xu\|^2 = \frac{1}{2n}\|w - Xu\|^2$

$$f(\beta^\circ + u) = f(\beta^\circ) + \underbrace{\langle \nabla f(\beta^\circ), u \rangle}_{-\frac{1}{n}\langle X^T w, u\rangle} + \underbrace{\frac{1}{2n}\|Xu\|^2}_{\text{curvature term}}$$

* Taylor expansion $\|$ $\frac{1}{2n}\left(\|w\|^2 - 2\langle w, Xu\rangle + \|Xu\|^2\right)$

Optimality of $\hat{\beta}$: $f(\beta^\circ) \ge f(\hat{\beta})$

together: let $u = \hat{\beta} - \beta^\circ$

$$\begin{aligned} \cancel{f(\beta^\circ)} &\ge f(\beta^\circ + u) \\ &= f(\beta^\circ) - \frac{1}{n}\langle X^T w, u\rangle + \frac{1}{2n}\|Xu\|^2 \\ &= \cancel{f(\beta^\circ)} - \frac{1}{n}\langle w, Xu\rangle + \underbrace{\frac{1}{2n}\|Xu\|^2}_{\ge 0 \text{ dominate}} \end{aligned}$$

* Move the matrix to the other side and transpose it inner product

$$\Rightarrow \frac{1}{2n}\|Xu\|^2 \le \frac{1}{n}\langle w, Xu\rangle$$

$$\Rightarrow \|Xu\|^2 \le 2\langle w, Xu\rangle$$

$$\Rightarrow \|Xu\|^4 \le 4\langle w, Xu\rangle^2$$

$$\Rightarrow \underbrace{\|Xu\|^2}_{\ge 0\text{, okay to take square at both sides.}} \le 4\frac{\langle w, Xu\rangle^2}{\|Xu\|^2}$$

* u 2k-sparse $u = \underset{\text{k-sparce}}{\hat{\beta}} - \underset{\text{k-sparse}}{\beta^\circ}$

$(u_i \ne 0 \Longrightarrow \beta_i^\circ \ne 0 \vee \hat{\beta}_i \ne 0)$

Suffices to show claim:

$$\max_{\substack{u \in \mathbb{R}^d \\ 2k\text{-sparse}}} \frac{\langle w, Xu\rangle^2}{\|Xu\|^2} \le k \cdot O\left(\log\frac{2d}{k}\right), \quad \text{w.h.p.}$$

Proof (claim)

Reparametrize for every $S \subseteq [d]$, $|S| = 2k$, let $\Phi_S \in \mathbb{R}^{n \times 2k}$ orthogonal
$\text{colspan}(\Phi_S) \supseteq \{X_i \mid i \in S\}$

$$\max_{\substack{u \in \mathbb{R}^d \\ 2k\text{-sparse}}} \frac{\langle \underline{w}, X_{\underline{u}} \rangle^2}{\|X_{\underline{u}}\|^2} \le \max_{\substack{S \subseteq [d] \\ |S| = 2k}} \max_{\substack{v \in \mathbb{R}^{2k} \\ \|v\| = 1}} \underbrace{\langle \underline{w}, \Phi_S v \rangle^2}_{\langle \Phi_{(S)}^T \underline{w}, v \rangle^2 \le \|\Phi_{(S)}^T \underline{w}\|^2 \cdot \cancel{\|v\|^2}^{1}}$$

* Cauchy-Schwarz $\langle a, b \rangle \le \|a\| \cdot \|b\|$

choose S such that $S \supseteq \{i \mid u_i \neq 0\}$, then
$X_u \in \text{colspan}(\Phi_{(S)})$
$\exists v,\ \|v\| = 1,\ \frac{X_u}{\|X_u\|} = \Phi_{(S)} v.$

$$\le \max_{\substack{S \subseteq [d] \\ |S| = 2k}} \|\Phi_S^T \underline{w}\|^2$$

$$\Longrightarrow \max_{\substack{S \subseteq [d] \\ |S| = 2k}} \|\Phi_S^T \underline{w}\|^2 \le k \cdot O\left(\log \frac{2d}{k}\right)$$

$$\underbrace{\Phi_S^T \underline{w}}_{2k \times 2k} \sim N\big(0, \underbrace{\mathbb{E}[\Phi_S^T w w^T \Phi_S]}\big) \Rightarrow N(0, 2k \cdot I_n)$$

$$\Rightarrow \Phi_S^T \underbrace{\cancel{\mathbb{E}[ww^T]}}_{I_n} \Phi_S \Rightarrow 2k \cdot I_n$$ * Φ_S is orthogonal

$\|\Phi_S^T w\|^2 \sim \chi^2_{2k}$ * chi-square

union bound

$$P\left(\max_{\substack{S \subseteq [d] \\ |S| = 2k}} \|\Phi_S^T \underline{w}\|^2 > t \cdot 2k\right) \le \sum_S P\left(\|\Phi_S^T \underline{w}\|^2 > t \cdot 2k\right)$$

$$\le \binom{d}{2k} \underbrace{P\left(\chi^2_{2k} > t \cdot 2k\right)}_{\le e^{-tk/10}}$$

* $t \ge 2$

* $\binom{n}{k} \le \left(\frac{en}{k}\right)^k$, $e \approx 2.718$

* let $t = 30 \ln\left(\frac{2d}{k}\right)$

$$\le \underbrace{\binom{d}{2k}}_{\le \left(\frac{4d}{2k}\right)^{2k} = \exp\left(2k \cdot \ln\left(\frac{4d}{2k}\right)\right)} \cdot e^{-tk/10}$$

?? * choose 4 here for convenience

BSS is NP-hard in worst case!!!
NP-hard does not rule out poly-time estimators with same statistical guarantees as BSS.

$$= e^{2k \cdot \ln(2d/k) - tk/10}$$

$$= e^{2k \cdot \ln(2d/k) - 3k \cdot \ln(2d/k)}$$

$$= e^{-k \cdot \ln(2d/k)} = \left(\frac{2d}{k}\right)^{-k} \le 2^{-k}$$

LASSO

Idea. Replace sparsity constraints by tractable constraints.

$\hat{\beta} = \text{argmin} \{ \|X\beta - y\|^2 \mid \|\beta\| \le \rho \}$ how to choose ρ?

$\hat{\beta} = \arg\min \{ \|X\beta - y\|^2 \mid \underbrace{\|\beta\|_1}_{|\beta_1| + \cdots + |\beta_d|} \le R \}$ how to choose R?

Canonical choice: $R = \|\beta^0\|_1$ (β^0 feasible solution)

normalization: columns $X_1, \dots, X_d$ of X satisfy $\|X_i\|^2 = n$

$$\boxed{\begin{array}{l}\text{Theorem (slow rate)} \\ \text{If } R = \|\beta^0\|_1 \text{, then} \\ \mathbb{E}\left[\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|^2\right] \le \dfrac{\|\beta^0\|_1}{\sqrt{n}} \cdot O(\log 2d)^{1/2}\end{array}}$$

interpretation: consider $\beta^0 \in \{0, \pm 1\}^d$, exactly k non-zero entries $\rightarrow \|\beta^0\|_1 = k$

error bounds: BSS: $\frac{k}{n}$ slow rate: $\frac{k}{\sqrt{n}} = \sqrt{\frac{k^2}{n}}$ * up to log factors.

- error vanishes $k \rightarrow n$ - error vanishes $k^2 \rightarrow n$ ← quadratic gap

Proof (theorem slow rate)

$\underline{u} = \hat{\beta} - \beta_0$ (error vector)

Claim 1: $\frac{1}{2n}\|X\underline{u}\|^2 \le \frac{1}{n}\langle \underline{u}, X^T \underline{w} \rangle$

Hölder's inequality: $\langle \underline{u}, X^T\underline{w} \rangle \le \underbrace{\|\underline{u}\|_1}_{2\|\beta^0\|_1} \cdot \underbrace{\|X^T \underline{w}\|_\infty}_{\max_i |\langle X_i, \underline{w}\rangle|}$

Claim 2: $\|\underline{u}\|_1 \le 2\|\beta^0\|_1$

* Triangle inequality, $\|\hat{\underline{\beta}} - \beta^0\|_1 \le \underbrace{\|\hat{\underline{\beta}}\|_1}_{\le R = \|\beta^0\|_1} + \|\beta^0\|_1 \le 2\|\beta^0\|_1$

Claim 3: $\mathbb{E}[\|X^T\underline{w}\|_\infty] \le O(\log d)^{1/2} \cdot \sqrt{n}$

* Use $\langle X_i, \underline{w} \rangle \sim \mathcal{N}(0, \|X_i\|^2)$ and tail bounds. ??? exercise.

Together:
$$\mathbb{E}\left[\tfrac{1}{2n}\|X\underline{u}\|^2\right] \overset{\text{claim 1}}{\le} \mathbb{E}\left[\tfrac{1}{n}\langle \underline{u}, X^T\underline{w} \rangle\right]$$
$$\overset{\text{Hölder}}{\le} \tfrac{1}{n}\mathbb{E}\left[\|\underline{u}\|_1\right] \cdot \mathbb{E}\left[\|X^T\underline{w}\|_\infty\right]$$
$$\overset{\text{claim 2}}{\le} \tfrac{2}{n}\|\beta^0\|_1 \cdot \mathbb{E}\left[\|X^T\underline{w}\|_\infty\right]$$
$$\overset{\text{claim 3}}{\le} \tfrac{1}{\sqrt{n}}\|\beta^0\|_1 \cdot O(\log d)^{1/2}$$