6.S951 Modern Mathematical Statistics Fall 2024

Lecture 4 — September 17, 2024

Prof. Stephen Bates

Scribe: Russell Legate-Yang

1 Outline

Agenda:

1. Minimax in Gaussian linear model
2. Least-squares with model misspecification

Last time:

1. Statistical decision theory
2. Sufficiency
3. Bayes optimality, minimax optimality, admissibility
4. Example: Bayes and minimax in Gaussian linear model

Next two weeks:

1. Statistical decision theory for prediction ("statistical learning theory")
2. Optimal hypothesis testing
3. Information-theoretic minimax lower bounds

Remarks on homework:

1. New version on Canvas corrects typos
2. Check Piazza

2 Minimax optimality in Gaussian linear model

Recall Setup for Gaussian linear model:

$$\begin{aligned}
\vec{Y} &\in \mathbb{R}^n \\
Z &\in \mathbb{R}^{n\times d},\ \text{fixed and constant matrix} \\
\vec{\varepsilon} &\in \mathbb{R}^n \\
\vec{\varepsilon} &\sim \mathcal{N}(0_n, I_n\sigma^2) \\
\theta &\in \mathbb{R}^d \\
\vec{Y} &= Z\theta + \vec{\varepsilon} \\
L(a,\theta) &= \ell(a-\theta),\ \text{loss function, } \ell \text{ bowl-shaped} \\
\hat{\theta}^{LS} &= \arg\min_{\theta'} ||\vec{Y} - Z\theta'||^2,\ \text{least squares estimator} \\
\implies \hat{\theta}^{LS} &= (Z^TZ)^{-1}Z^T\vec{Y},\ \hat{\theta}^{LS} \sim \mathcal{N}(\theta, (Z^TZ)^{-1}\sigma^2)
\end{aligned}$$

We analyzed the Bayes optimal procedure with prior $Q = \mathcal{N}(0, \tau^2 I_d)$ for θ:

$$\begin{aligned}
R_B^*(Q) &= \inf_A \ R_B(A,Q) \\
&= \mathbb{E}[\ell(\Sigma_\tau^{1/2}W)],
\end{aligned}$$

where $W \sim \mathcal{N}(0, I_d)$ and $\Sigma_\tau = ((Z^TZ)/\sigma^2 + I_d/\tau^2)^{-1}$.

Today, we aim to show that $\hat{\theta}^{LS}$ is minimax in the Gaussian linear model.

Theorem 1. *Suppose Z has rank d. Then $\hat{\theta}^{LS}$ is minimax optimal in the Gaussian linear model.*

Proof. Given the distribution of $\hat{\theta}^{LS}$, we know that the risk of $\hat{\theta}^{LS}$ does not depend on θ:

$$\mathbb{E}[L(\hat{\theta}^{LS}, \theta)] = \mathbb{E}[\ell((Z^TZ/\sigma^2)^{-1/2}W)],$$

since $\hat{\theta}^{LS} - \theta \sim \mathcal{N}(0, (Z^TZ)^{-1}\sigma^2)$. Hence, the minimax risk of $\hat{\theta}^{LS}$ equals this constant risk:

$$\begin{aligned}
R_M(\hat{\theta}^{LS}) &= \sup_\theta \mathbb{E}[\ell((Z^TZ/\sigma^2)^{-1/2}W)] \\
&= \mathbb{E}[\ell((Z^TZ/\sigma^2)^{-1/2}W)].
\end{aligned}$$

Therefore, it suffices to show that $R_M(\hat{\theta}) \geq \mathbb{E}[\ell((Z^TZ/\sigma^2)^{-1/2}W)]$ for any estimator $\hat{\theta}$.

We split the proof into 3 cases.

Case 1: $d = n$ and $Z = I_n$. Consider the risk of the Bayes estimator A^* for prior $Q = \mathcal{N}(0_d, \tau^2 I_d)$. We have:

$$\begin{aligned}
R_B(A^*, Q) &= \mathbb{E}\left[\ell((Z^TZ/\sigma^2 + I_d\tau^2)^{-1/2}W)\right] && (1) \\
&= \mathbb{E}\left[\ell((1/\sigma^2 + 1/\tau^2)^{-1/2}I_dW)\right] \ \text{since } Z = I_d = I_n. && (2)
\end{aligned}$$

Since the Bayes risk at any prior Q is a lower bound for the minimax risk of any procedure $\hat{\theta}$,

$$\begin{aligned} R_M(\hat{\theta}) &\geq R_B(A^*, Q) \\ &= \mathbb{E}\left[\ell((1/\sigma^2 + 1/\tau^2)^{-1/2} I_d W)\right]. \end{aligned}$$

Taking the limit as $\tau \to \infty$ and using the monotone convergence theorem,

$$R_M(\hat{\theta}) \geq \mathbb{E}[\ell(\sigma I_d W)],$$

as desired.

Case 2: $d = n$ **and** $Z \neq I_n$**.** Let $\hat{\theta}(\vec{Y})$ be a generic estimator. The risk of $\hat{\theta}$ is

$$\begin{aligned} R(\hat{\theta}, \theta) &= \mathbb{E}[\ell(\hat{\theta}(\underbrace{Z\theta + \varepsilon}_{\vec{Y}}) - \theta)] \\ &= \mathbb{E}[\tilde{\ell}(Z\hat{\theta}(Z\theta + \varepsilon) - Z\theta)], \text{ where } \tilde{\ell}(x) = \ell(Z^{-1}x). \end{aligned}$$

Consider the "rotated" Gaussian linear model problem where:

- We view the parameter as $Z\theta$ instead of θ
- We view the constant design matrix as I_n instead of Z
- We use the loss $L(a, \theta) = \tilde{\ell}(a - \theta)$. One can check that $\tilde{\ell}$ is bowl-shaped.

Note that this rotated problem fits into case 1. Then the risk in this rotated problem when the estimator is $Z\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$ is

$$R_{I,\tilde{\ell}}(Z\hat{\theta}, Z\theta) = \mathbb{E}[\tilde{\ell}(Z\hat{\theta} - Z\theta)],$$

where the subscripts in $R_{I,\tilde{\ell}}$ indicate that this is the risk in the rotated problem. Hence,

$$\begin{aligned} R(\hat{\theta}, \theta) &= R_{I,\tilde{\ell}}(Z\hat{\theta}, Z\theta) \\ \implies \sup_{\theta} R(\hat{\theta}, \theta) &= \sup_{\theta} R_{I,\tilde{\ell}}(Z\hat{\theta}, Z\theta) \\ &= \sup_{Z\theta} R_{I,\tilde{\ell}}(Z\hat{\theta}, Z\theta), \text{ } Z \text{ invertible so sup over } \theta \text{ same as sup over } Z\theta \\ &\geq \mathbb{E}[\tilde{\ell}((I_n^T I_n/\sigma^2)^{-1/2} W)] \text{ by case 1} \\ &= \mathbb{E}[\ell(Z^{-1}\sigma I_n W] \text{ by def of } \tilde{\ell} \\ &= \mathbb{E}[\ell((Z^T Z/\sigma^2)^{-1/2} W)] \text{ since } Z^{-1}\sigma I_n W \sim \mathcal{N}(0_n, (Z^T Z/\sigma^2)^{-1}). \end{aligned}$$

So $R_M(\hat{\theta}) \geq \mathbb{E}[\ell((Z^T Z/\sigma^2)^{-1/2} W)]$ as desired.

Case 3: $d < n$**.** This case will reduce to case 2 by sufficiency.

Recall: $\vec{Y} = Z\theta + \vec{\varepsilon}$.

Define $U = Z(Z^T Z)^{-1/2}$. U has d orthonormal columns and these columns span the column space of Z. So UU^T projects onto the column space of Z.

Lemma 2. *$U^T\vec{Y}$ is sufficient.*

Suppose the lemma holds for now. We have $U^T\vec{Y} \sim \mathcal{N}((Z^T Z)^{1/2}\theta, \sigma^2 I_d)$. Since $U^T\vec{Y} \in \mathbb{R}^d$, we have reduced from n dimensions to d dimensions. We can therefore view $U^T\vec{Y}$ as coming from a $d-$dimensional Guassian linear model with parameter θ and constant matrix $(Z^T Z)^{1/2}$. For any estimator $A : \mathbb{R}^d \to \mathbb{R}^d$, case 2 implies

$$\begin{aligned}\sup_\theta R_d(A;\theta) &\geq \mathbb{E}[\ell((((Z^T Z)^{1/2})^T((Z^T Z)^{1/2})/\sigma^2)^{-1/2}W)] \\ &= \mathbb{E}[\ell((Z^T Z/\sigma^2)^{-1/2})W],\end{aligned}$$

where R_d denotes the risk in the reduced, d-dimensional model.

Because $U^T\vec{Y}$ is sufficient in the n-dimensional model, for every estimator $\hat{\theta}(Y) : \mathbb{R}^n \to \mathbb{R}^d$ in the original model, there is an estimator $A : \mathbb{R}^d \to \mathbb{R}^d$ (possibly randomized) in the reduced model with the same risk. Hence,

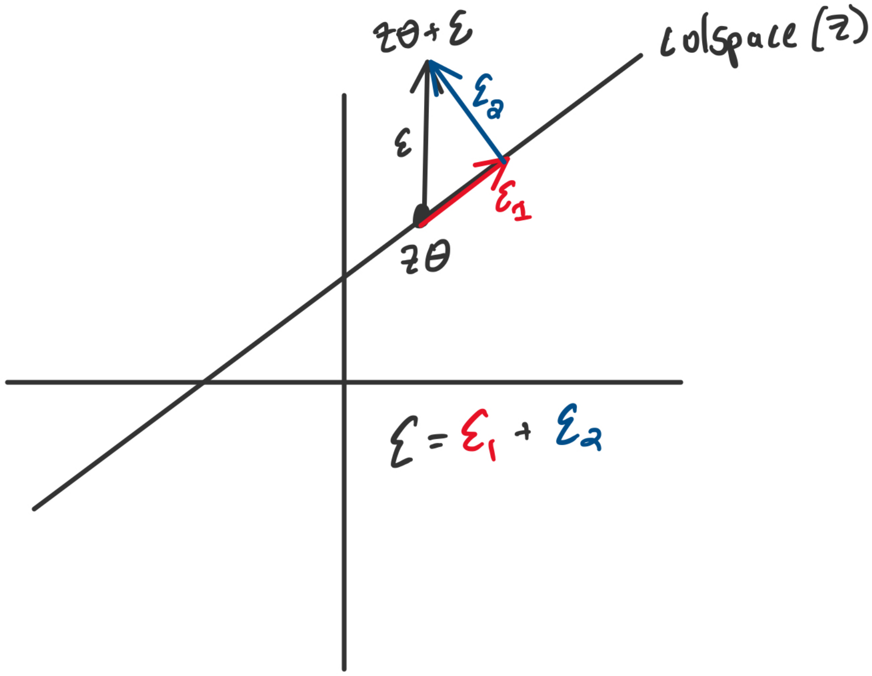
$$\begin{aligned}\sup_\theta R(\hat{\theta};\theta) &= \sup_\theta R_d(A;\theta) \\ &\geq \mathbb{E}[\ell((Z^T Z/\sigma^2)^{-1/2})W],\end{aligned}$$

as desired.

Now, a proof sketch for the lemma.

Proof. Decompose ε by projecting it onto the column space of Z: $\varepsilon = \varepsilon_1 + \varepsilon_2$ where $\varepsilon_1 \in \text{colspace}(Z)$ and $\varepsilon_2 \in \text{colspace}(Z)^\perp$. Because the orthogonal projection operation is linear, ε_1 and ε_2 jointly Gaussian and uncorrelated. Marginally, ε_1 is the Gaussian restricted to the column space of Z, and ε_2 is the Gaussian restricted to the orthogonal complement of the column space of Z. Since ε_1 and ε_2 are jointly Gaussian and uncorrelated, $\varepsilon_1 \perp \varepsilon_2$. Therefore, the distribution of the data conditional on $Z\theta + \varepsilon_1$ does not depend on θ since ε_2 is independent Gaussian noise. So $Z\theta + \varepsilon_1$ is sufficient for θ.

It remains to show that $U^T\vec{Y}$ is sufficient for θ. Note that $Z\theta + \varepsilon_1$ is the orthogonal projection of $Z\theta + \varepsilon$ on the column space of Z: $Z\theta$ already lies in the column space of Z and ε_1 is defined to have this property. This orthogonal projection is exactly $UU^T\vec{Y}$. So $U^T\vec{Y}$ is sufficient (since Z is a known non-random matrix, U is also known and non-random).



□

This concludes the proof of the theorem. □

3 Least-squares with a misspecified model

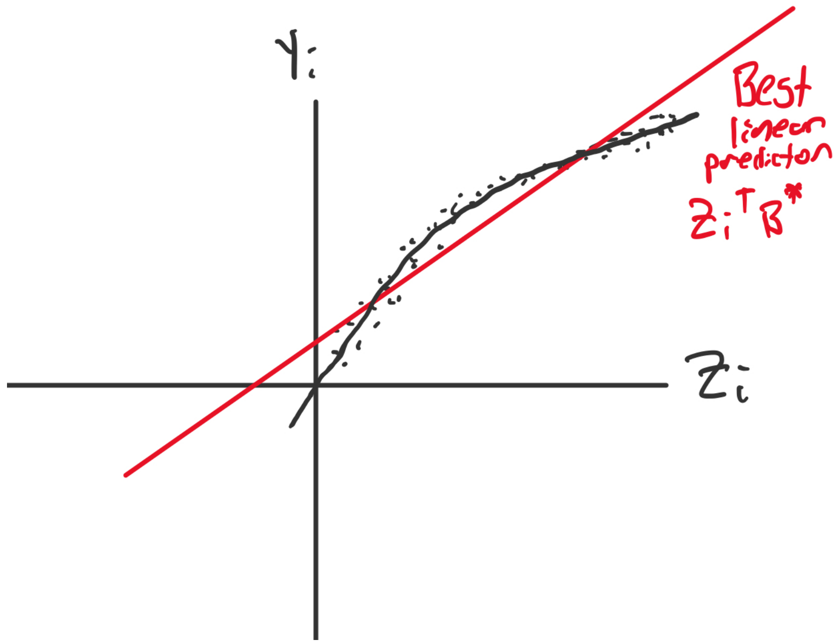
Consider an iid sample of $(Z_i, Y_i) \in \mathbb{R}^{d+1}$ from the statistical model:

$$\mathcal{P} = \{P_\theta^n : P_\theta \text{ is a distribution on } \mathbb{R}^{d+1} \text{ with finite fourth moments}\}.$$

In particular, Z_i is random. What is the behavior of the least-squares estimator in this setting?

Definition 3. *(Least squares target) The best linear prediction of Y_i given Z_i is*

$$\beta^* = \arg\min_\beta \mathbb{E}[(Y_i - Z_i^T \beta)^2]$$



β^* is a sensible target to think about in many situations even when the data are not linear.

Proposition 4. *(Consistency of least-squares for β^*)*

$$\hat{\beta}^{LS} = (Z^T Z)^{-1} Z^T \vec{Y} \xrightarrow{p} \beta^*$$

Proof. This argument is a proof outline.

Divide by n in the numerator and denominator:

$$\hat{\beta}^{LS} = (Z^T Z/n)^{-1} Z^T \vec{Y}/n.$$

By the law of large numbers and a continuous mapping theorem,

$$\begin{aligned}
(Z^T Z/n)^{-1} &\xrightarrow{p} \mathbb{E}[Z_i Z_i']^{-1} \in \mathbb{R}^{d \times d} \\
Z^T \vec{Y}/n &\xrightarrow{p} \mathbb{E}[Y_i Z_i] \in \mathbb{R}^d \\
\implies \hat{\beta}^{LS} &\xrightarrow{p} \mathbb{E}[Z_i Z_i']^{-1} \mathbb{E}[Y_i Z_i].
\end{aligned}$$

One can show that $\beta^* = \mathbb{E}[Z_i Z_i']^{-1} \mathbb{E}[Y_i Z_i]$. □

We will later derive an approximate distribution for $\hat{\beta}^{LS}$.

Claim 5.

$$\hat{\beta}^{LS} \sim_{approx} \mathcal{N}(\beta^*, \Sigma/n).$$

We will derive Σ in part 3 of the course.

Compared to its behavior in the Gaussian linear model, the least-squares estimator in a misspecified model has larger total variance. Variance includes the original noise (ε) and the noise coming from misspecificiation (where the best linear predictor diverges from the conditional expectation of Y_i given Z_i). However, as the $1/n$ factor in the approximate variance suggests, $\hat{\beta}^{LS}$ remains root-n consistent.