Week 1
Spring 2009

Lecture 1. Possible topics in this course
Course website: http://www.stat.yale.edu/~hz68/619/
References:
Lecture notes of Lawrence D. Brown, *Shrinkage: Fall 2006*,
(http://www-stat.wharton.upenn.edu/%7Elbrown/teaching/Shrinkage/index.html).
David B. Pollard, *Asymptopia*.
Iain Johnstone, *Function estimation and Gaussian sequence model*.
Review:

Wald (1939), Contributions to the Theory of Statistical Estimation and Testing Hypotheses, *Ann. Math. Stat.*. This paper introduced much of the landscape of modern decision theory, including loss functions, risk functions, admissible decision rules, prior, Bayes decision rules, and minimax decision rules. Wolfowitz described this paper as: " ... probably Wald's most important single paper". The phase "decision theory" was first used by Lehmann.

Example: Observe a normally distributed n-dimensional random variable X,

$$X \sim N(\theta, \Sigma)$$

where θ and Σ are parameters. In the Stat 610, we often assume that the parameter space is of the following form

$$\times = \left\{\theta, \Sigma_{n\times n} : \theta \in S \subset \mathbb{R}^n, \Sigma = \sigma^2 \Psi\right\}$$

where S and Ψ (e.g.,$\Psi = I_{n\times n}$) are known. In this lecture we assume that σ is known, $S = \mathbb{R}^n$ and $\Psi = I_{n\times n}$ for simplicity (in the standard Gaussian linear model, σ is unknown). Our goal is to estimate the mean vector θ. Let $\delta(X)$ be an estimator of θ. A loss function $L(\theta, \delta)$ will be used to measure the resulting error.

Loss function. A commonly used loss function is

$$L(\theta, \delta) = \|\theta - \delta\|^2 = \Sigma_{i=1}^n (\delta_i - \theta_i)^2 .$$

Risk function. The risk function is used to measure how well the estimator does on average

$$R(\theta, \delta) = EL(\theta, \delta) .$$

Admissible. An estimator δ is called to be inadmissible if there is an estimator δ' such that

$$R\left(\theta, \delta'\right) \leq R(\theta, \delta) \text{ for all } \theta \in \mathbb{R}^n, \text{ and } R\left(\theta, \delta'\right) < R(\theta, \delta) \text{ for some } \theta.$$

An estimator is admissible if it is not inadmissible.

Wald (1939) operated with Bayes solution. A brave man?

Minimaxity. An estimator δ^* is minimax if

$$R(\theta, \delta^*) = \inf_{\delta} \sup_{\theta} R(\theta, \delta).$$

Why minimax? In Wald's book he states on page 27:"Nevertheless, since Nature's choice is unknown to the experimenter, it is perhaps not unreasonable to for experimenter to behave as if Nature wanted to maximize the risk. But, even if one is not willing to take this attitude, the theory of games remains of fundamental importance ...".

There may be few statisticians who actively supports the minimax principle as a prescription for action. However the minimax idea has been an essential foundation for advances in many areas of statistical research: asymptotic theory and methodology, hierarchical models, robust estimation, optimal design, and nonparametric function estimation. Why?

Remark 1 *In 1939, Wald "proved" the admissibility of the estimator X. Stein received his Ph.D. in 1947 from Columbia under Wald on sequential analysis. Inspiried by Savage, Stein Started to realize that the inadmissibility was perhaps true.*

Stein (1956), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Norma Distribution. *Proc. 3^{rd} Berk Symp Math. Stat. Prob.*

James and Stein (1961), Estimation with quadratic loss. *Proc. 4^{th} Berkeley Symp. Math. Statist. Prob., 1.*

James-Stein estimator

$$\delta(X) = \left(1 - \frac{c}{\|X\|^2}\right) X,\ c > 0.$$

This estimator is minimax if and only if $0 \leq c \leq 2(n-2)$.

Tentative schedule:

Topic 1 Shrinkage estimation in parametric models (4-6 weeks)

i The Canonical normal means estimation problem. Stein's unbiased estimator of risk.

ii Bayes estimation, minimaxity and Admissibility.

iii Empirical Bayes, hierarchical Bayes and random effects.

Topic 2 Shrinkage estimation in nonparametric models (1-2 weeks, Pinsker bound theory)

i Best linear estimation.

ii Blockwise Stein's estimation and Adaptive minimaxity.

Topic 3 Testing hypothesis and its connection to estimation. (5-6 weeks).

i Neyman-Pearson Lemma and minimax lower bound.

ii Minimax estimation for functional data analysis.

iii Minimax Estimation for covariance matrices estimation.

iv Multiple comparisons and sharp adaptive minimaxity.

Topic 4 Le Cam theory. (0-2 weeks)

i Asymptotic equivalence theory

Lecture 2. The Canonical normal means estimation problem.

Question:

Let $X \sim N\left(\theta, \sigma^2 I_n\right)$ with σ known. Under ordinary squared error loss, is X admissible? *Wald (1939, Ann. Math. Stat.) said "Yes" for all n, but Stein (1956, Proc. 3^{rd} Berk Symp Math. Stat. Prob. 1) said "No" for $n \geq 3$.*

Stein's heuristic arguments from Stein (1956): Let $\sigma = 1$. Write $X = \theta + Z$ with $Z \sim N(0, I_n)$, then

$$\begin{aligned} \|X\|^2 &= \|\theta\|^2 + \|Z\|^2 + 2\theta^T Z \\ &= \|\theta\|^2 + \|Z\|^2 + 2\|\theta\| Y \end{aligned}$$

where $Y = \frac{\theta^T Z}{\|\theta\|} \sim N(0,1)$. For large n we have

$$\|X\|^2 = \|\theta\|^2 + n + O_p\left(\sqrt{\|\theta\|^2 + n}\right)$$

The usual estimator lies outside the set

$$\left\{\theta : \|\theta\|^2 \leq \|X\|^2 - cn\right\}$$

(may assume that $\|\theta\|^2 \leq Mn$). It certainly reasonable to to cut X by a factor

$$\left(\frac{\|X\|^2 - n}{\|X\|^2}\right)^{1/2}$$

to bring the estimate within that sphere. Actually, because of the curvature of the sphere combined with the uncertainty of our knowledge of ξ, the best constant, to within the approximation considered here, turns out to be $\frac{\|X\|^2 - n}{\|X\|^2} = 1 - \frac{n}{\|X\|^2}$. For, consider the class of estimators $\left[1 - h\left(\frac{\|X\|^2}{n}\right)\right] X$, then

$$\begin{aligned} & \left\| \left[1 - h\left(\frac{\|X\|^2}{n}\right)\right] X - \theta \right\|^2 \\ = & \left[1 - h\left(\frac{\|X\|^2}{n}\right)\right]^2 \|X - \theta\|^2 + h^2\left(\frac{\|X\|^2}{n}\right)\|\theta\|^2 + 2\|\theta\| h\left(\frac{\|X\|^2}{n}\right)\left(1 - h\left(\frac{\|X\|^2}{n}\right)\right) Y \\ \approx & \; n\left[\left(1 - h(1+\rho)\right)^2 + h^2(1+\rho)\cdot\rho\right] \end{aligned}$$

where $\rho = \|\theta\|^2 / n$.

An alternative argument (I learned this from Andrew Barron. See also in Brown's lecture notes.): Stein said "It certainly reasonable to cut X

by a factor $\left(\frac{\|X\|^2-n}{\|X\|^2}\right)^{1/2}$ to bring the estimate within that sphere." Why is that factor reasonable? It would be more reasonable to project θ to X and obtain

$$\frac{\langle X,\theta\rangle}{\|X\|^2}X \approx \frac{\|\theta\|^2}{\|X\|^2}X \approx \frac{\|X\|^2-n}{\|X\|^2}X = \left(1-\frac{n}{\|X\|^2}\right)X.$$

Stein's lemma. See Stein (1981, Ann. Stat.) or Stein (1973).

Let Y be a $N(0,1)$ real random variable and $g:\mathbb{R}\to\mathbb{R}$ be an indefinite integral of the Lebesgue measurable function g', essentially the derivative of g. Suppose that $E|g'(Y)|<\infty$. Then

$$E(Yg(Y)) = Eg'(Y).$$

Proof of the lemma: Write

$$\begin{aligned} Eg'(Y) &= \int g'(y)\,\phi(y)\,dy \\ &= \int_0^\infty g'(y)\int_y^\infty x\phi(x)\,dxdy - \int_{-\infty}^0 g'(y)\int_{-\infty}^y x\phi(x)\,dxdy \end{aligned}$$

then apply Fubini's theorem (why?).

Stein's unbiased estimate of the risk (SURE).

Let $X\sim N(\theta,\Sigma_{p\times p})$ with Σ positive definite and $g(X)$ be absolutely continuous, then

$$\begin{aligned} E[(X+g(X)-\theta)^T\Sigma^{-1}(X+g(X)-\theta)] &= E[n+g(X)^T\Sigma^{-1}g(X)+2\nabla\cdot g(X)] \\ E\|(X+g(X)-\theta)\|^2 &= E[tr(\Sigma)+\|g(X)\|^2+2tr(\Sigma\cdot Dg(X))] \end{aligned}$$

where $\nabla g(X)=\sum_{i=1}^p \partial g_i/\partial X_i$ and $Dg(X)$ is a $p\times p$ matrix with $(Dg)_{ij}=\partial g_i/\partial X_j$.

Homework problem 1: prove the identities above.

Example. Let $X\sim N(\theta,1)$ and $\hat\theta = X+h(X)$. Then

$$E\left(\hat\theta-\theta\right)^2 = E\left[1+2h'(x)+h^2(X)\right].$$

Example. Let $X\sim N(\theta,I_{p\times p})$ and $\hat\theta_i=\left(1-\frac{\lambda}{|X_i|}\right)_+X_i$ we have

$$E\left(\hat\theta_i-\theta_i\right)^2 = E\left[1-2I(|X_i|\le\lambda)+X_i^2\wedge\lambda^2\right]$$

and

$$E\left\|\hat\theta-\theta\right\|^2 = ESURE(\lambda,X),\ .$$

where $SURE(\lambda,X)=p-2\cdot\#\{i,|X_i|\le\lambda\}+\sum_{i=1}^p X_i^2\wedge\lambda^2$ is an increasing function on $\left[|X|_{(i)},|X|_{(i+1)}\right)$. Question: if we replace $I_{p\times p}$ by a general covariance matrix $\Sigma_{p\times p}=(\sigma_{ij})$ with $\sigma_{ii}=1$. Do we get the same SURE formula?

James-Stein estimator.

$$\delta_{J-S}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right) X,\ C > 0.$$

Theorem. Let $X \sim N\left(\theta, \sigma^2 I_n\right)$. Let $0 < C \leq 2(n-2)$ (hence $n \geq 3$). Then

$$R(\theta, \delta_{J-S}) = E\left\|(\delta_{J-S}(X) - \theta)\right\|^2 \leq n\sigma^2.$$

Proof of the theorem:

$$E\left\|(\delta_{J-S}(X) - \theta)\right\|^2 = n\sigma^2 - E[\frac{\sigma^4}{\|X\|^2} C(2(n-2) - C)]$$

Question: a similar result for other losses, e.g. $L(\theta, \delta) = \Sigma_{i=1}^n |\delta_i - \theta_i|$.

Lemma: $X \sim N\left(\theta, \sigma^2 I_n\right)$

$$E\left\|(\delta_{J-S}(X) - \theta)\right\|^2 \leq 2\sigma^2 + \frac{(n-2)\sigma^2 \|\theta\|^2}{(n-2)\sigma^2 + \|\theta\|^2} \leq 2\sigma^2 + \frac{n\sigma^2 \|\theta\|^2}{n\sigma^2 + \|\theta\|^2}$$

Proof: Without loss of generality, we assume that $\sigma = 1$. Since $\|X\|^2$ can be seen as a mixture of χ^2_{d+2N} and $N \sim$Poisson$\left(\|\theta\|^2/2\right)$, and

$$E\left\|(\delta_{J-S}(X) - \theta)\right\|^2 = d - (d-2)^2 E_\theta \|X\|^{-2},$$

then

$$E\left\|(\delta_{J-S}(X) - \theta)\right\|^2 = d - (d-2)^2 E\frac{1}{d-2+2N} \underset{\text{Jensen}}{\leq} d - (d-2)^2 \frac{1}{d-2+\|\theta\|^2}.$$

Homework problem 2

Domination of positive-part estimator.

Let $X \sim N\left(\theta, \sigma^2 I_n\right)$. Define

$$\delta_{J-S}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right) X,\ \delta^+_{J-S}(X) = \left(1 - \frac{C\sigma^2}{\|X\|^2}\right)_+ X$$

where $C > 0$. Show that

$$E\left\|\left(\delta^+_{J-S}(X) - \theta\right)\right\|^2 < E\left\|(\delta_{J-S}(X) - \theta)\right\|^2$$

for all $\theta \in \mathbb{R}^n$.

Question: find a Bayes estimator to dominate the positive part of JS estimator?