Statistical Theory for Data Science

STA2212H S LEC9101

Week 4

January 27 2026

Asymptotics for estimating a diverging number of parameters — with and without sparsity

Jana Gauss and Thomas Nagler

Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)

April 8, 2025

Abstract

We consider high-dimensional estimation problems where the number of parameters diverges with the sample size. General conditions are established for consistency, uniqueness, and asymptotic normality in both unpenalized and penalized estimation settings. The conditions are weak and accommodate a broad class of estimation problems, including ones with non-convex and group structured penalties. The wide applicability of the results is illustrated through diverse examples, including generalized linear models, multi-sample inference, and stepwise estimation procedures.

1 Introduction

In modern applications, statisticians are facing increasingly complex and high-dimensional problems. Many data sets have a huge number of variables, calling for similarly many parameters p. In other scenarios, the number of variables is moderate, but adequately modeling the data requires highly complex, non-linear models with many parameters. The traditional fixed-p-large-n paradigm is inadequate in such situations.

This article adopts an asymptotic perspective, allowing both the sample size n and the number of parameters p_n to diverge. We consider general parametric problems where the estimator $\hat{\boldsymbol{\theta}}$ solves an estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}) = \mathbf{0} \in \mathbb{R}^{p_n}, \tag{1}$$

with some function $\phi\colon \mathbb{R}^{p_n} \to \mathbb{R}^{p_n}$. A classical example are risk minimization problems, where ϕ is the gradient of a loss function. The estimating equation framework is also suited for more complex

Asymptotics for estimating a diverging number of parameters — with and without sparsity

Jana Gauss and Thomas Nagler

Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)

April 8, 2025

Abstract

We consider high-dimensional estimation problems where the number of parameters diverges with the sample size. General conditions are established for consistency, uniqueness, and asymptotic normality in both unpenalized and penalized estimation settings. The conditions are weak and accommodate a broad class of estimation problems, including ones with non-convex and group structured penalties. The wide applicability of the results is illustrated through diverse examples, including generalized linear models, multi-sample inference, and stepwise estimation procedures.

1 Introduction

In modern applications, statisticians are facing increasingly complex and high-dimensional

SEMIPARAMETRIC EFFICIENT EMPIRICAL HIGHER ORDER INFLUENCE FUNCTION ESTIMATORS

By Lin Liu*, Rajarshi Mukherjee†, Whitney K. Newey‡, James M. Robins§

Robins et al. (2008) applied the theory of higher order influence functions (HOIFs) to derive an estimator of the mean ψ of an outcome Y in a missing data model with Y missing at random conditional on a vector X of continuous covariates; their estimator, in contrast to other existing estimators but ours, is semiparametric efficient under the minimal Hölder smoothness conditions derived in Robins et al. (2009b), together with an additional (non-minimal) Hölder smoothness condition on the density g of X, because that particular estimator depends on a non-parametric estimate of g. In this paper, we introduce a new HOIF estimator that has the same asymptotic properties as the previous one, but imposes no smoothness requirement on g. This improvement is significant for two reasons. First, one rarely has the knowledge about the smoothness properties of g. Second, even when g is smooth, and even if X is just multivariate with fixed dimensions, accurate nonparametric estimation of its density is generally not feasible at the sample sizes often encountered in practice. In fact, to our knowledge, this new HOIF estimator to be studied here remains the *only* semiparametric efficient estimator of ψ under minimal Hölder smoothness conditions, despite the rapidly growing literature on causal effect estimation. We also show that our estimator can be generalized to the entire class of functionals considered by Robins et al. (2008) which includes the average effect of a treatment on a response Y when a vector X suffices to control for confounding and the expected conditional variance of Y given X. Simulation experiments are also conducted, which demonstrate that our new estimator outperforms previous ones proposed in earlier works on HOIFs in finite samples, when g is not very smooth.

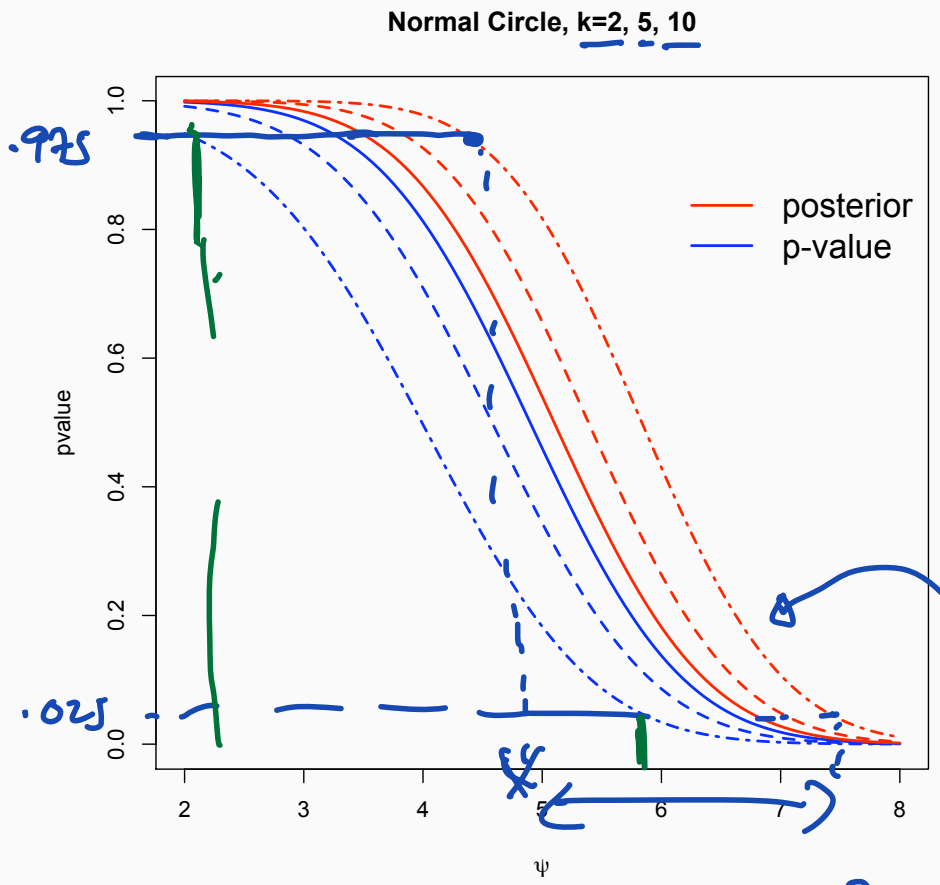
1. Introduction. Robins et al. (2008), together with a companion technical report Robins et al. (2016) containing more details, introduced novel U-statistic based estimators of a class of nonlinear functionals in semi- and non-parametric models. Construction of these estimators was based on the theory of Higher Order Influence Functions (henceforth referred to as HOIFs) (Robins et al., 2008). HOIFs are U-statistics that represent higher order derivatives of a functional. The

Today

arxiv paper

1. Midterm 2 shift from March 3 to March 10? ✓
2. Recap: Stein's problem (notes corrected), sandwich variance, *M*-estimators (recent arxiv), interval estimation
3. Nonparametric estimation
4. Hypothesis and signficance testing
5. Questions/Review re Midterm Feb 3 7.10 – 8.00 pm
6. Papers re project

Recap: Stein's example



$X_i \sim N(\mu_i, 1)$ ind't $i = 1, \dots, k$

$\pi(\mu) = 1$ $\quad \psi = \sum_{i=1}^{k} \mu_i^2$

$E(\psi \mid \underset{\sim}{x})$ post. mean $= k + \Sigma x_i^2$

$E_{model} E(\psi \mid \underset{\sim}{x}) = k + k + \psi$

$\psi \mid \underset{\sim}{x} \sim \chi^{2\prime}_k(\Sigma x_i^2)$ exact

$P_\pi(\Psi \geq \psi \mid \underset{\sim}{x})$ Bayes surv. f.

p-value function $p(\psi) = P_{\underset{\sim}{x} \mid \mu}\{\Sigma x_i^2 \leq \Sigma x_i^2\}$ ntbc.

Recap: Sandwich variance

MS 5.5, SM 8.4

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x; \theta), \theta \in \Theta$
- maximum likelihood estimator based on model: $\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$
- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2), \quad \sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}]^2 dF(x))^2} = \frac{\mathrm{var}_F\{\ell'(\theta; X)\}}{\mathrm{E}_F\{-\ell''(\theta; X)\}^2}$$

Recap: Sandwich variance

MS 5.5, SM 8.4

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$ $\quad X_i \sim F(\cdot)$
- maximum likelihood estimator based on model: $\sum_{i=1}^n \ell'(\hat{\theta}_n; X_i) = 0$
-

$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2), \quad \sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}]^2 dF(x))^2} = \frac{\mathrm{var}_{\hat{F}_n}\{\ell'(\theta; X)\}}{\mathrm{E}_{\hat{F}_n}\{-\ell''(\theta; X)\}^2}$$

- model is correct; estimator is root of $\sum_{i=1}^n \psi(\theta; x_i) = 0$ $\quad \tilde{\theta}_n$
- assume $\psi(\theta; X_i)$ has expected value 0, finite variance, and is differentiable

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\{0, G^{-1}(\theta)\},$$

$$G(\theta) = E\{-\frac{\partial}{\partial \theta^T}\psi(X; \theta)\}[\mathrm{var}\{\psi(X; \theta)\}]^{-1} E\{-\frac{\partial}{\partial \theta}\psi(X; \theta)\}$$

Recap: Confidence interval $\theta \in \mathbb{R}$

AoS 6.3.2; MS 7.1

- upper, lower, central $C_n(x) = (-\infty, a_n]$ or $[b_n, \infty)$ or $[a_n(x), b_n(x)]$
- length and coverage $b_n(x) - a_n(x)$ (↑ $1-\alpha$) $\Rightarrow \Pr_\theta\{a_n(x) \le \theta \le b_n(x)\} \ge 1-\alpha \ \forall \theta$ or $E\{b_n(x) - a_n(x)\}$
- exact or approximate
- inversion of pivotal quantity $S(x;\theta)$ with known dist: $G_S(s)$ "$= 1-\alpha$ some θ"
- posterior credible intervals

$$\int_a^b \pi(\theta|x)\,d\theta = 1-\alpha$$

e.g. $\int_{-\infty}^{a} \pi(\theta|x)\,d\theta = \alpha/2$

$P_\theta\{S(x;\theta) \le s^\alpha\} = \alpha$ (↑ find s^α)

invert that $= G_S^{-1}(\alpha)$ to get a bd for θ

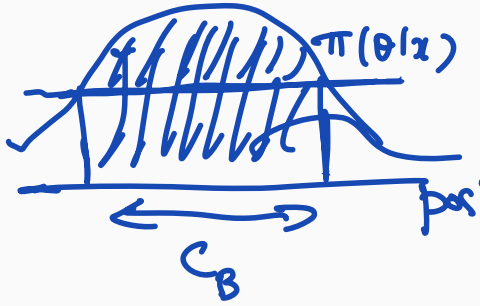
Confidence regions; HPD regions

$\theta \in \mathbb{R}^p \quad \mathbb{R}^k \quad \mathbb{R}^d$

$R(\underline{x})$ $C_n(\underline{x}) = \{\underline{\theta} : P_{n,\underline{\theta}}\{R(\underline{X}) \ni \underline{\theta}\} \geq 1-\alpha$ gen'n Conf. int.

$\longrightarrow C_B(\underline{x}) = \{\underline{\theta} : \pi(\underline{\theta} \mid \underline{x}) \geq k_\alpha\}$

credible region for $\underline{\theta}$



k_α chosen s.t.

$P_n(\underline{\theta}_n \in C_B \mid \underline{x}) = 1-\alpha$

approx. $C_n(\underline{x})$ $(\hat{\underline{\theta}} - \underline{\theta})^T I_n(\underline{\theta})(\hat{\theta} - \theta) \leq \chi^2_{p(1-\alpha)}$

C_B $\{(\theta - \hat{\theta}_\pi)^T J_\pi(\theta)(\theta - \hat{\theta}_\pi)\} \leq \chi^2_p(1-\alpha)$ e.g.

Confidence intervals for sub-parameters

- model $Y \sim f(y; \psi, \lambda), \quad \psi \in \mathbb{R}^d, \lambda \in \mathbb{R}^{p-d}, \quad \theta = (\psi, \lambda)$ $\quad y = (y_1, \ldots, y_n)$
- profile log-likelihood function $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ maximize over λ

$j_p(\psi) = -\ell_p''(\psi)$

- approximate pivotal quantities MS Thm 7.4,5

c.R. for ψ

$\{\psi : (\hat{\psi} - \psi)^T j_p(\hat{\psi})(\quad) \le \chi^{2(1-\alpha)}_d\}$

$(\hat{\psi} - \psi)^T j_p(\hat{\psi})(\hat{\psi} - \psi) \sim \chi^2_p$ under $f(y; \psi, \lambda)$

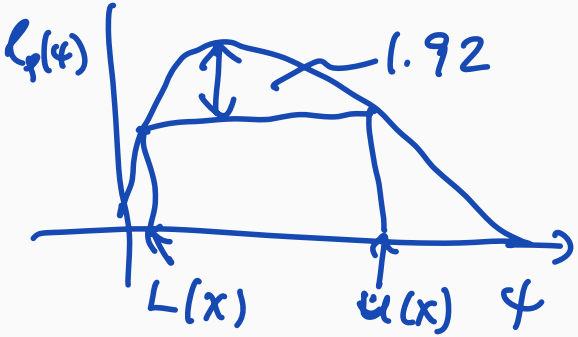
$2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \sim \chi^2_p$

$\hat{\psi} \pm 1.96\, j_p^{-1/2}(\hat{\psi})$

- approximate confidence regions

$\{\psi : (\hat{\psi} - \psi)^T j_p(\hat{\psi})(\hat{\psi} - \psi) \le \chi^2_{d,1-\alpha}\}$

$\{\psi : 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \le \chi^2_{d,1-\alpha}\}$



- recall $X_1, \ldots, X_n$, i.i.d. $F(\cdot)$

plug-in est'rs

$X_{(1)} \le \cdots \le X_{(n)}$

- empirical cdf

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{(i)} \le x\} = \frac{1}{n} \sum 1\{X_i \le x\}$$

- properties:

$$\mathrm{E}\{\hat{F}_n(x)\} = F(x), \quad \mathrm{var}\{\hat{F}_n(x)\} = \frac{1}{n} F(x)\{1 - F(x)\}$$

Binomial

any fixed x

- pointwise approximate confidence limits

$$\hat{F}_n(x) \pm z_{1-\alpha/2} [\hat{F}_n(x)\{1 - \hat{F}_n(x)\}]^{1/2}$$

for any fixed x has coverage $\approx 1-\alpha$

Empirical cdf

$X_1, \ldots, X_n$ i.i.d. $F(\cdot)$

$\{F(x);\ x \in \mathcal{X}\}$ stoch. proc.

Glivenko-Cantelli Theorem:

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{p} 0$$

$\{\widehat{F}_n(x);\ x \in \mathcal{X}\}$ empirical process

Dvoretzky-Kiefer-Wolfowitz Inequality

$$\mathrm{pr}\{sup_x|\widehat{F}_n(x) - F(x)| > \epsilon\} \leq 2\exp(-2n\epsilon^2)$$

simultaneous confidence band : $\mathrm{pr}\{L(t) \leq F(t) \leq U(t) \text{ for all } t\} \geq 1 - \alpha$:

$$L(t) = \max\{\hat{F}_n(t) - \epsilon_n, 0\}, \quad U(t) = \min\{\hat{F}_n(t) - \epsilon_n, 1\}, \quad \epsilon_n = \left\{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)\right\}^{1/2}$$

98 7. Estimating the CDF and Statistical Functionals

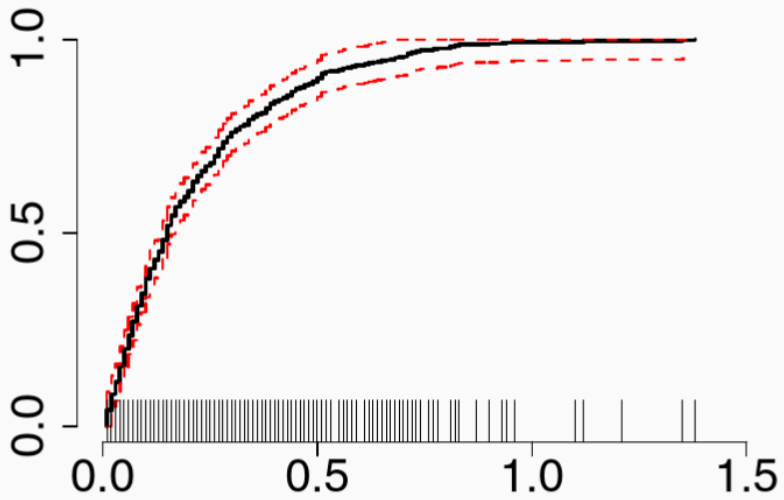


FIGURE 7.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

7.2 Example (Nerve Data). Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. Figure 7.1 shows the empirical CDF $\widehat{F}_n$. The data points are shown as small vertical lines at the bottom of the plot. Suppose we want to estimate the fraction of waiting times between .4 and .6 seconds. The estimate is $\widehat{F}_n(.6) - \widehat{F}_n(.4) = .93 - .84 = .09$. ■

$\hat{F}_n(x) = \text{e.c.d.f.}$

$\hat{F}_n'(x) = \frac{1}{n}$ at each x_i $i = 1, \dots, n$

- $X_1, \dots, X_n$ i.i.d., $X_i \sim f(\cdot)$
- kernel density estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

$K(\cdot)$ kernel f-n

h: scale

- with a symmetric kernel function, for small h,

e.g. $N(0, 1)$

$$\mathrm{E}\{\hat{f}(x)\} = f(x) + \frac{1}{2} h^2 f''(x) + O(h^4),$$

biased est.

$$\mathrm{var}\{\hat{f}(x)\} = \frac{1}{nh} f(x) \int K^2(u) du$$

fixed x

- mean-squared error

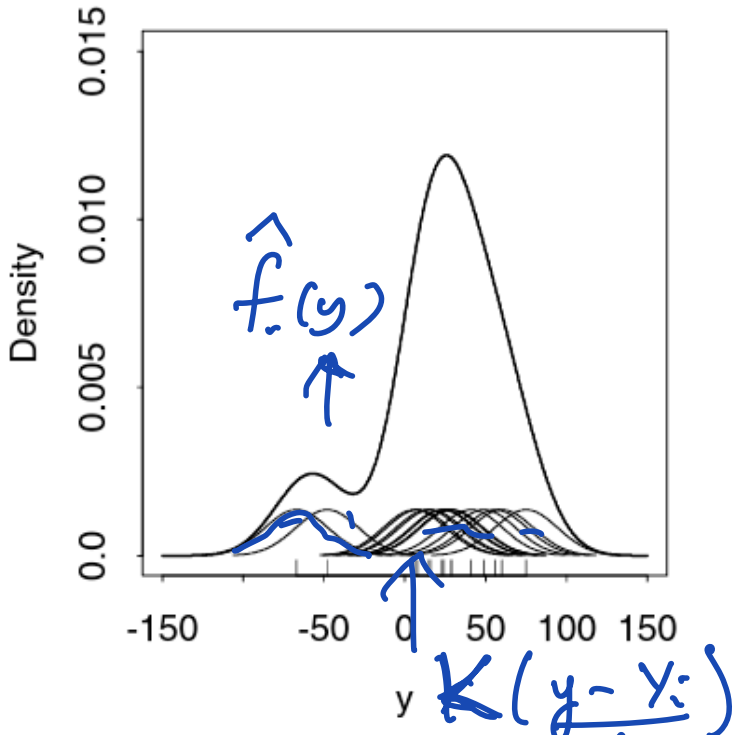
$$\mathrm{MSE} \doteq \frac{h^4}{4}\{f''(x)\}^2 + \frac{1}{nh} f(x) \int K^2(u) du \quad \left\{ + ? O(h) ? \right.$$

$\lambda \uparrow$

$\frac{\partial}{\partial h}\ldots = 0 \Rightarrow h^* = h^*(x)$

$h \propto n^{-1/5}$ gives MSE $\sim O(n^{-4/5})$

306 7 · Estimation and Hypothesis Testing



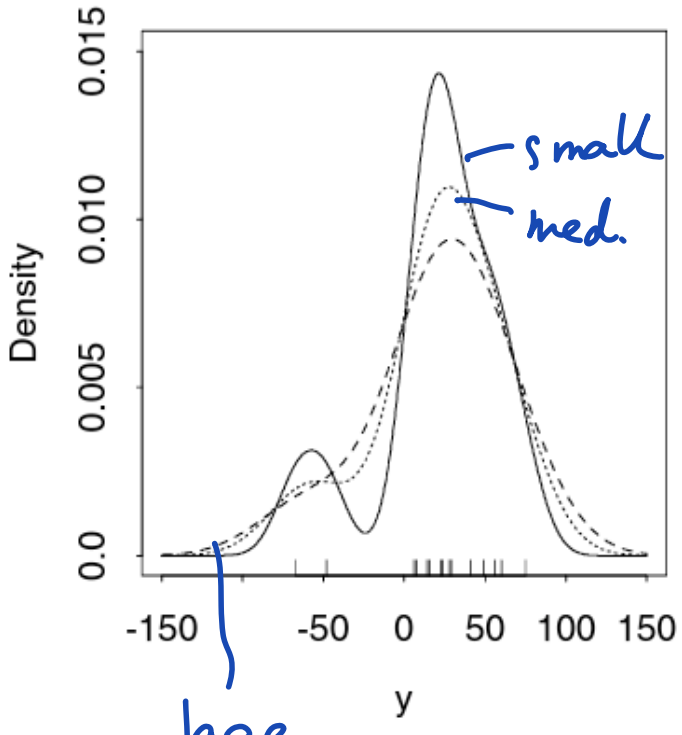


Figure 7.2 Kernel density estimates for maize data. Left: construction of kernel estimate (heavy) as sum of 15 scaled normal densities centred at the y_j, with $h = 19.5$. Right: density estimates with $h = 13.3$ (solid), $h = 23.2$ (dots) and $h = 30$ (dashes).

Density estimation

AoS 20.3, SM 7.1.2

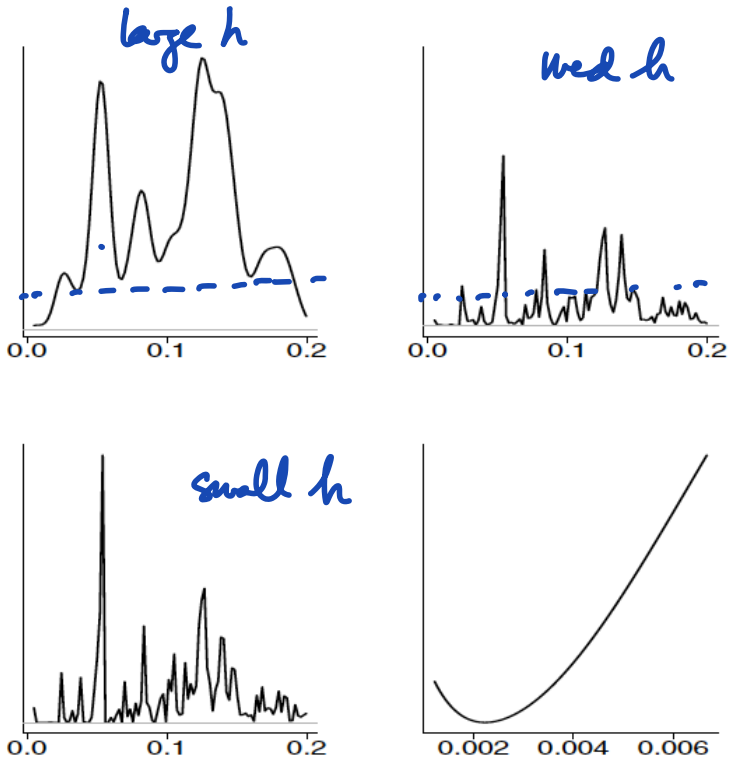


FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value of h where the curve is a minimum.

318 20. Nonparametric Curve Estimation

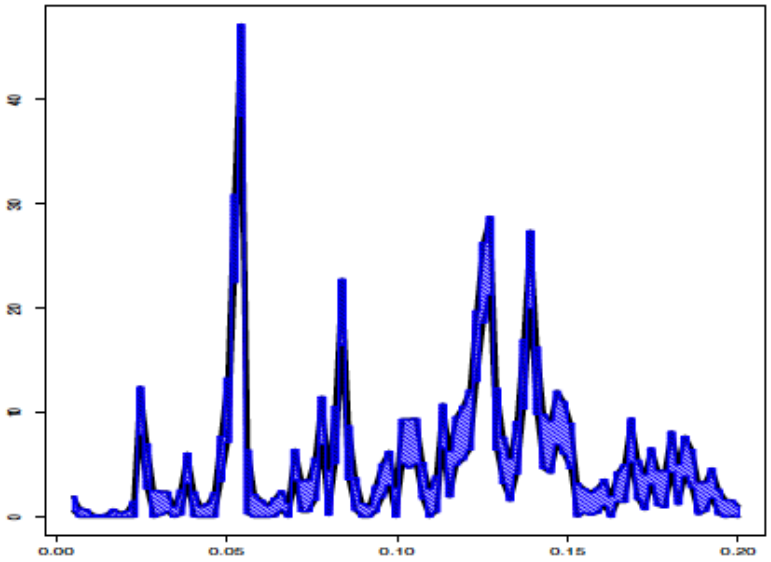


FIGURE 20.7. 95 percent confidence bands for kernel density estimate for the astronomy data.

$\hat{\theta} \pm 1.96\,\hat{se}$

for θ

20.9 Definition. *A pair of functions* $(\ell_n(x), u_n(x))$ *is a* $1-\alpha$ **confidence band (or confidence envelope)** *if*

$$\mathbb{P}\Big(\ell(x) \leq \overline{f}_n(x) \leq u(x) \quad \text{for all } x\Big) \geq 1-\alpha. \tag{20.16}$$

$\bar{f}_h(x)$ is a smoothed version of the density estimator and

re-defines what we're est'g

$$f(x) \longrightarrow \int_{x-\varepsilon}^{x+\varepsilon} f(z)\,dz = \bar{f}$$

$K(\,)$

To construct confidence bands, we use something similar to histograms. Again, the confidence band is for the smoothed version,

$$\overline{f}_n = \mathbb{E}(\widehat{f}_n(x)) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\, du,$$

of the true density f. [4] Assume the density is on an interval (a, b). The band is

$$\ell_n(x) = \widehat{f}_n(x) - q\,\mathsf{se}(x), \quad u_n(x) = \widehat{f}_n(x) + q\,\mathsf{se}(x) \tag{20.27}$$

where

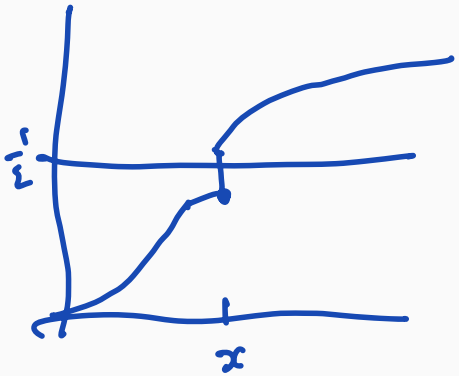
$$\begin{aligned}
\mathsf{se}(x) &= \frac{s(x)}{\sqrt{n}}, \\
s^2(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(x) - \overline{Y}_n(x))^2, \\
Y_i(x) &= \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \\
q &= \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right), \\
m &= \frac{b-a}{\omega}
\end{aligned}$$

where ω is the width of the kernel. In case the kernel does not have finite width then we take ω to be the effective width, that is, the range over which the kernel is non-negligible. In particular, we take $\omega = 3h$ for the Normal

Statistical functional

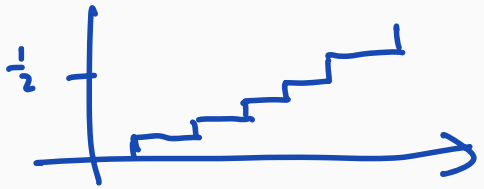
- Define an estimand $\theta = T(F)$ or $\theta(F)$ $\quad T : \mathcal{F} \to \Theta$
- Substitution (plug-in) estimator $\widehat{\theta}_n = T(\widehat{F}_n)$ $\quad \widehat{F}_n$ e.c.d.f.
- Examples of functionals: moments, quantiles, solution of $\int \psi\{x; T(F)\} dF(x) = 0$

$$E(X^3) = \int x^3 dF(x) = T_3(F) \qquad \frac{1}{n}\sum X_i^3 = T_3(\hat{F}_n)$$



$$T_m = \text{median}(F) = F^{-1}(\tfrac{1}{2}) = \inf\{x : F(x) \geq \tfrac{1}{2}\}$$

$$\Rightarrow \quad T_m(\hat{F}_n) = \hat{F}_n^{-1}(\tfrac{1}{2})$$



$$\int \psi(x, T(\hat{F}_n)) d\hat{F}_n(x) = 0$$

$$0 = \frac{1}{n}\sum \psi(x_i, T(\hat{F}_n))$$

Statistical functional

- Define an estimand $\theta = T(F)$ or $\theta(F)$ $T : \mathcal{F} \to \Theta$
- Substitution (plug-in) estimator $\widehat{\theta}_n = T(F_n)$
- Examples of functionals: moments, quantiles, solution of $\int \psi\{x; T(F)\} dF(x) = 0$

- Lorenz curve; Gini index MS Ex 4.21; F on $[0, \infty)$

$$q_F(t) = \frac{\int_0^t F^{-1}(s) ds}{\int_0^1 F^{-1}(s) ds}, \qquad \theta(F) = 2 \int_0^1 \{t - q_F(t)\} dt$$

Statistical functional

AoS 7.2, MS 4.5,6

- Define an estimand $\theta = T(F)$ or $\theta(F)$ $T : \mathcal{F} \to \Theta$
- Substitution (plug-in) estimator $\widehat{\theta}_n = T(F_n)$
- Examples of functionals: moments, quantiles, solution of $\int \psi\{x; T(F)\}dF(x) = 0$

- Lorenz curve; Gini index MS Ex 4.21; F on $[0, \infty)$

$$q_F(t) = \frac{\int_0^t F^{-1}(s)ds}{\int_0^1 F^{-1}(s)ds}, \qquad \theta(F) = 2\int_0^1 \{t - q_F(t)\}dt$$

- Estimator: $\widehat{\theta}_n$ solves $\frac{1}{n}\sum_{i=1}^n \psi(X_i; \theta) = 0$ $= \int \psi\{x; T(\widehat{F}_n)\}d\widehat{F}_n(x)$
- $F^{-1}(t) = \inf\{x : F(x) \geq t\}$, $\widehat{\theta}_n = 2\int_0^1 \{t - q_{\widehat{F}_n}(t)\}dt$

Influence function

AoS 7.2, MS 4.5,6

- Define an estimand $\theta = T(F)$, estimator $\widehat{\theta}_n = T(\widehat{F}_n)$ $\quad T : \mathcal{F} \to \Theta$
- Influence function $\quad$ if limit exists

$$\phi(x; F) = \frac{d}{dt}\theta\{(1-t)F + t\delta(x)\}\Big|_{t=0}$$

- If we have $\widehat{\theta}_n - \theta(F) = \frac{1}{n}\sum_{i=1}^n \phi(X_i; F) + R_n$, $\quad \sqrt{n}R_n \xrightarrow{p} 0$, then $\quad \mathrm{E}\{\phi(X_i; F)\} = 0$

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N\{0, \sigma^2(F)\}$$

$$\sigma^2(F) = \int \phi^2(x; F) dF(x)$$

MS Ex.4.35

Classic reference:

Van der Vaart & Wellner (1996). *Weak Convergence and Empirical Processes*, Springer.

Hypothesis and significance testing

- model $Y \sim f(y; \psi, \lambda), \quad \psi \in \mathbb{R}^d, \lambda \in \mathbb{R}^{p-d}, \quad \theta = (\psi, \lambda)$ $\quad y = (y_1, \ldots, y_n)$
- approximate pivotal quantities MS Thm 7.4,5

likelihood-based

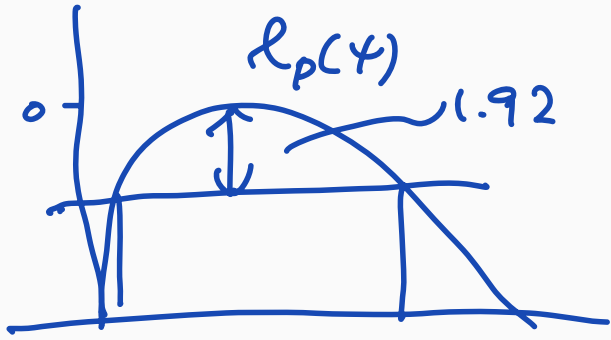
$$(\widehat{\psi} - \psi)^{\mathrm{T}} j_{\mathrm{p}}(\widehat{\psi})(\widehat{\psi} - \psi)$$
$$2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}$$

- approximate confidence regions

$$\{\psi : (\widehat{\psi} - \psi)^{\mathrm{T}} j_{\mathrm{p}}(\widehat{\psi})(\widehat{\psi} - \psi) \leq \chi^2_{d,1-\alpha}\}$$
$$\{\psi : 2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} \leq \chi^2_{d,1-\alpha}\}$$

? check ?

- generalized likelihood ratio ~~test~~ pivot



Aside: Generalized likelihood ratio test

AoS Thm 10.2, MS 7.4

- Textbook version

$$\Lambda = \frac{\sup_{\theta \in \Theta} f(\boldsymbol{X}; \theta)}{\sup_{\theta \in \Theta_0} f(\boldsymbol{X}; \theta)} = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}$$

- Theorem 7.5 (MS)

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_d$$

- constraint $\theta \in \Theta_0$ has d restrictions on $\theta = (\theta_1, \ldots, \theta_p)$
- version using profile constrains $\psi = (\theta_1, \ldots, \theta_d)$, $d < p$
- motivated as a test of $H_0 : \theta \in \Theta_0$ vs $H_A : \theta \notin \Theta_0$

Formal Theory of hypothesis testing

$X_1, \ldots, X_n \sim f(\boldsymbol{x}; \theta)$ $\theta \in \Theta$

AoS range $\boldsymbol{x} \in \mathcal{X}$

- Null and alternative hypothesis $H_0 : \theta \in \Theta_0$ $H_1 : \theta \in \Theta_0$
- Rejection region $R \subset \mathcal{X}$ sample space if $\underset{\sim}{x} \in R$ "reject H_0" $\underset{\sim}{x} \notin R$ "do not reject H_0"
- Test statistic and critical value $\{\underset{\sim}{x} : t(\underset{\sim}{x}) \geq k\}$ some $t(\underset{\sim}{X})$ & some k
- Test function $\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } \underset{\sim}{x} \in R \\ 0 & \text{o.w.} \end{cases} = \begin{cases} 1 & t(\underset{\sim}{x}) > k \\ 0 & \text{o.w.} \end{cases}$
- Type I and Type II error $P_{H_0}(\text{reject } H_0)$ $P_{H_1}(\text{do not reject } H_0)$

$X_1, \ldots, X_n \sim f(\boldsymbol{x}; \theta)$ AoS range $\boldsymbol{x} \in \mathcal{X}$

- Null and alternative hypothesis: $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$
- Rejection region: $R \subset \mathcal{X}$; if $\boldsymbol{x} \in R$ "reject" H_0 AoS 10.0
  $R = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 1\}$
- Test statistic and critical value: $R = \{\boldsymbol{x} \in \mathcal{X} : t(\boldsymbol{x}) > k\}$ k to be chosen
- Test (decision) function: $\phi : \mathcal{X} \to \{0, 1\}$ MS 7.3
  $\phi(\boldsymbol{X}) = 1$ decide $\theta \in \Theta_1$, else decide $\theta \in \Theta_0$
- Type I and Type II error: $\text{pr}\{\boldsymbol{X} \in R\}$, $\theta \in \Theta_0$, $\text{pr}\{\boldsymbol{X} \notin R\}$, $\theta \in \Theta_1$
  $\text{pr}\{t(\boldsymbol{X}) \geq c\}$, $\text{pr}\{t(\boldsymbol{X}) < c\}$

- goal is to identify R, $\phi(\cdot)$, or $T = t(\boldsymbol{X})$ with small Type I and Type II errors

- goal is to identify R, $\phi(\cdot)$, or $T = t(\boldsymbol{X})$ with small Type I and Type II errors
- can't reduce both errors at once — see text following Ex. 7.10
- classical solution: require Type 1 error $\leq \alpha$ — size α
- subject to this constraint, minimize type 2 error

- goal is to identify R, $\phi(\cdot)$, or $T = t(\boldsymbol{X})$ with small Type I and Type II errors
- can't reduce both errors at once — see text following Ex. 7.10
- classical solution: require Type 1 error $\leq \alpha$ — size α
- subject to this constraint, minimize type 2 error
- find a test statistic, $T = t(\boldsymbol{X})$, and define $R = \{\boldsymbol{x} : t(x) \geq c\}$ s.t.

$$\mathrm{pr}_{\theta \in \Theta_0}\{t(\boldsymbol{X}) \geq c\} \leq \alpha, \quad \text{then } \max \mathrm{pr}_{\theta \in \Theta_1}\{t(\boldsymbol{X}) \geq c\}$$

- power function

$$\beta(\theta) = \mathrm{pr}_\theta\{t(X) \geq c\}$$

(implicit assn: large $t(\underset{\sim}{X})$ is evidence against H_0)

AoS assumes $\Theta = \mathbb{R}$; , pr ↑ in $\theta > \theta_0$

Neyman-Pearson lemma

AoS Thm 10.30, MS Thm 7.2

- for testing simple H_0 against simple H_1
- test statistic

$$T = \frac{L(\theta_1; \boldsymbol{x})}{L(\theta_0; \boldsymbol{x})} = \frac{f(\boldsymbol{x}; \theta_1)}{f(\boldsymbol{x}; \theta_0)}$$

- critical region

$$\{\boldsymbol{x} : t(\boldsymbol{x}) \geq k\}$$

- Choose $k = k_\alpha$ to satisfy

$$\mathrm{pr}_{H_0}(T \geq k_\alpha) = \alpha$$

- This test is a most powerful test of H_0 against H_1 at level α

smallest T2 error

largest 1 − T2 error

A neatly-typed proof (from SM 7.3)

Let R be the rejection region for the test based on $T = f_1(\boldsymbol{x})/f_0(\boldsymbol{s})$

$$R = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq k_\alpha\}$$

Let R' be some other rejection region also of size α $\leq \alpha$

$$\alpha = \int_R f_0(\boldsymbol{x})d\boldsymbol{x} = \int_{R'} f_0(\boldsymbol{x})d\boldsymbol{x}$$

$$\int_{R-R'} f_0(\boldsymbol{x})d\boldsymbol{x} = \int_{R'-R} f_0(\boldsymbol{x})d\boldsymbol{x}$$

$A - B \triangleq A \cap B^c$

$R - R' \subset R$

$R' - R \subset R^c$

On LHS $f_1(\boldsymbol{x}) \geq k_\alpha f_0(\boldsymbol{x})$.

On RHS $f_1(\boldsymbol{x}) < k_\alpha f_0(\boldsymbol{x})$.

$$\int_{R-R'} f_1(\boldsymbol{x})d\boldsymbol{x} \geq \int_{R'-R} f_1(\boldsymbol{x})d\boldsymbol{x}$$

Add integral over intersection $R \cap R'$

A neatly-typed proof (from MS)

Let $\phi(\boldsymbol{x})$ be the test function for the test based on T.
Let $\psi(\boldsymbol{x})$ be any other function that maps $\boldsymbol{x}$ to $[0, 1]$.
If

$$\mathrm{E}_{H_0}\{\psi(\boldsymbol{X})\} \leq \mathrm{E}_{H_0}\{\phi(\boldsymbol{X})\} = \alpha$$

then it must follow that

$$\mathrm{E}_{H_1}\{\psi(\boldsymbol{X})\} \leq \mathrm{E}_{H_1}\{\phi(\boldsymbol{X})\}$$

Proof: $\forall\, \boldsymbol{x}$,

$$\psi(\boldsymbol{x})\{f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x})\} \leq \phi(\boldsymbol{x})\{f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x})\}$$

Integrate and re-arrange terms to get the result

p-values

- The formal theory of testing imagines a decision to "reject H_0" or not, according as $X \in R$ or $X \notin R$, for some defined region $R \subset \mathcal{X}$ e.g. $|Z| > 1.96$
- This is useful for deriving the form of optimal tests, but not useful in practice.
- Doesn't distinguish between $Z = 1.97$ and $Z = 19.7$, for example.
- P-values give more precise information about the null hypothesis

$H_0: \theta = \theta_0$ vs $\theta = \theta_1$

NP-lemma $\longrightarrow$ $H_0: \theta = \theta_0$ vs $\theta > \theta_0$
exp'l family ω scalar θ UMP

p-values

- The formal theory of testing imagines a decision to "reject H_0" or not, according as $X \in R$ or $X \notin R$, for some defined region $R \subset \mathcal{X}$ e.g. $|Z| > 1.96$
- This is useful for deriving the form of optimal tests, but not useful in practice.
- Doesn't distinguish between $Z = 1.97$ and $Z = 19.7$, for example.
- P-values give more precise information about the null hypothesis
- AoS definition: p-value $= \inf\{\alpha : T(X_n) \in R_\alpha\}$ Def 10.11
- MS definition: $p(\boldsymbol{x}) = \inf\{\alpha : \phi_\alpha(\boldsymbol{x}) = 1\}$ 7.5
- SM definition $p_{obs} = \Pr_{H_0}\{T(X_n) \geq t_{obs}\}$

$t_{obs} = T(\underset{\sim}{x})$

$|Z|$ - some 1.97

$p < 0.05$

$p = 0.052$

$|Z| = 4.5$ $p = 10^{-9}$

Hypothesis tests and significance tests

- Hypothesis tests typically means:
  - H_0, H_1
  - critical/rejection region $R \subset \mathcal{X}$,
  - level α, power $1 - \beta$
  - conclusion: "reject H_0 at level α" or "do not reject H_0 at level α"
  - planning: maximize power for some relevant alternative

minimize type II error

Hypothesis tests and significance tests

- Hypothesis tests typically means:
  - H_0, H_1
  - critical/rejection region $R \subset \mathcal{X}$,
  - level α, power $1 - \beta$
  - conclusion: "reject H_0 at level α" or "do not reject H_0 at level α"
  - planning: maximize power for some relevant alternative — minimize type II error

- Significance tests typically means:
  - H_0,
  - test statistic T
  - observed value t^{obs},
  - p-value $p^{obs} = \Pr(T \geq t^{obs}; H_0)$
  - alternative hypothesis often only implicit — large T points to alternative

Example: Sign test

SM Ex.7.26

- $X_1, \ldots, X_n$ i.i.d. $F(\cdot)$
- $H_0 : \mu = \mu_0$, $\mu = F^{-1}(1/2)$ median of distribution
- $H_1 : \mu > \mu_0$

both H composite

- test statistic

$$T = \sum_{i=1}^{n} 1\{X_i > \mu_0\}$$

- under H_0,

$$T \sim Binom(n, 1/2)$$

- p-value

$$p_{obs} = \mathrm{pr}_{H_0}(T \geq t_{obs}) = \sum_{r=t_{obs}}^{n} \binom{n}{r} \frac{1}{2^n} \doteq 1 - \Phi\left\{ \frac{2(t_{obs} - n/2)}{n^{1/2}} \right\}.$$

Power of the sign test

SM Ex.7.30

- $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0$ $\mu = F^{-1}(1/2)$
- Test statistic $T = \sum_{i=1}^{n} 1\{X_i > \mu_0\}$
- Rejection region $R = \{T \geq c_\alpha\}$
- $c_\alpha \approx n/2 - n^{1/2} z_\alpha / 2$ Normal approx

- Power = $\mathrm{pr}_{H_1}(\text{reject } H_0) = \mathrm{pr}_{H_1}(T \geq c_\alpha)$ Need distribution of T under H_1
- to calculate power we need values for μ and for F

Power of the sign test

SM Ex.7.30

- $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0$
- Test statistic $T = \sum_{i=1}^{n} 1\{X_i > \mu_0\}$
- Rejection region $R = \{T \geq c_\alpha\}$
- $c_\alpha \approx n/2 - n^{1/2} z_\alpha / 2$

no Normal ass^n needed

$\mu = F^{-1}(1/2)$

Normal approx

- Power = $\mathrm{pr}_{H_1}(\text{reject } H_0) = \mathrm{pr}_{H_1}(T \geq c_\alpha)$
- to calculate power we need values for μ and for F

Need distribution of T under H_1

- SM assumes F is $N(\mu, \sigma^2)$, so

$\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$

$$\mathrm{pr}_{\mu_1}(T \geq c_\alpha) = \mathrm{pr}_{\mu_1}(T \geq n/2 - n^{1/2} z_\alpha / 2) \doteq \Phi\left\{ \frac{n\Phi(n^{-1/2}\delta) - n/2 + n^{1/2} z_\alpha}{[n\Phi(n^{-1/2}\delta)\{1 - \Phi(n^{-1/2}\}]} \right\}$$

$$\doteq \Phi\{z_\alpha + \delta(2/\pi)^{1/2}\}$$

- test based on $\bar{X}$ has power $\Phi(z_\alpha + \delta)$

… power of sign test

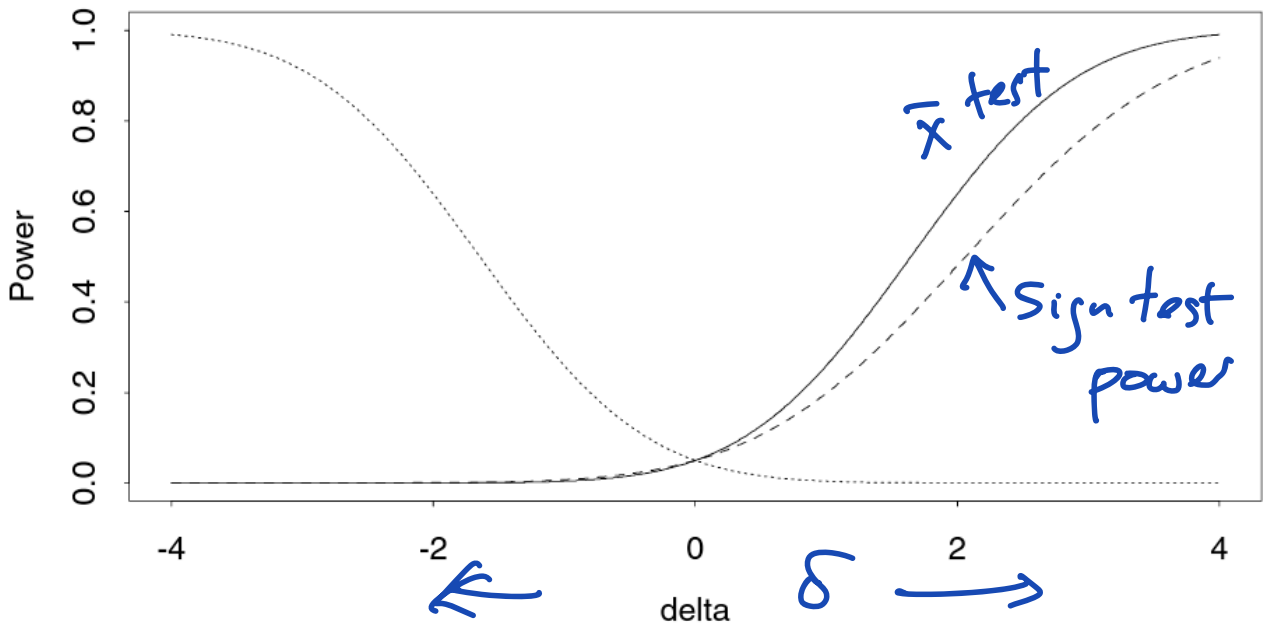


Figure 7.6 Power functions for a test of whether the mean of a $N(\mu, \sigma^2)$ random sample of size n equals μ_0 against the alternative $\mu = \mu_1$, as a function of $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$. The test size is $\alpha = 0.05$. The solid curve is the power function for a test of $\mu_1 > \mu_0$ based on $\overline{y}$, and the dashed line is the power function for the sign test. Both critical regions are of form $\overline{y} > t_\alpha$. The dotted curve is the power function for $\overline{y}$ when the critical region is $\overline{y} < t_\alpha$.

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

$\hat{\sigma}^2 \xrightarrow{p} ?$

$E(\hat{\sigma}^2) = \dots$

$?\ P(|\bar{x}-\mu| > \varepsilon) < \delta\, \mathrm{var}(\bar{X})\ ?$

WLLN $\hat{\sigma}^2 \xrightarrow{p} \dots$

$\mathrm{var}(\hat{\sigma}^2) \to 0$