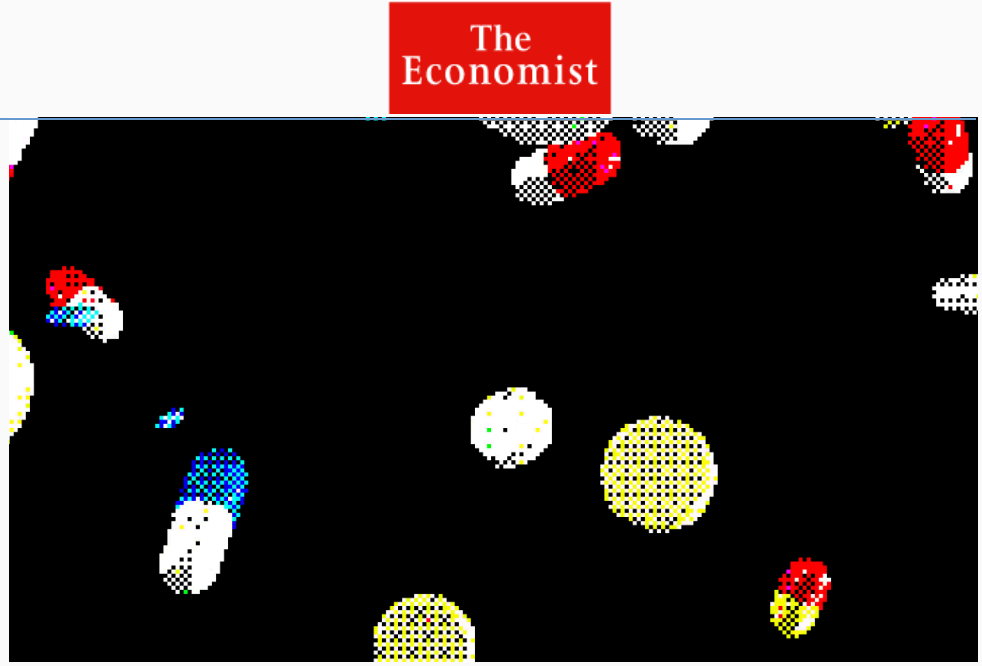
Statistical Theory for Data Science

STA2212H S LEC9101

Week 3

January 20 2026



Today

arxiv paper

1. Recap: delta method, Bayesian inference
2. Misspecified models; point and interval estimation
3. Nonparametric estimation
4. Statistics in the News
5. HW Questions

Upcoming

- Toronto Data Workshop, Wednesday 21 January 2026, noon (EST) on Zoom
  Rachel Porter, University of Notre Dame
  "The CampaignView collection of records about candidates and their policies"
- Thursday 22 January 2026, 11am Hydro 9195/9
  Isaac Gibbs, Job Talk

Recap: Delta method

$X \sim N(0, \sigma^2)$ $\sum X_i^2/n = \hat{\sigma}^2$

- variance-stabilizing transformations

$\uparrow$

$\operatorname{var}(\bar{X})$ depends on $E(\bar{X})$ $\cdot\mu$

then normal approx could be poor

$\operatorname{var} g(\hat{\theta}) \simeq g'(E\hat{\theta})^2 \cdot \operatorname{var}\hat{\theta}$

$g(\cdot)$ 1-1 fn of θ

$\frac{\partial g}{\partial \theta^T}(E\hat{\theta}) \cdot \operatorname{var}\hat{\theta} \frac{\partial g(E\hat{\theta})}{\partial \theta^T}$

Example $X_i \sim Po(\mu)$ $i = 1, \ldots, n$ iid $E(\bar{X}) = \mu$ $\operatorname{var}(\bar{X}) = \mu$

$\operatorname{var}\{g'(\bar{X})\} = c = \frac{1}{n} g'(\mu)^2 \cdot \mu \propto c$

$\operatorname{var}(\bar{X}^{1/2}) \simeq \frac{1}{4n}$ $E\bar{X} \simeq \mu^{1/2}$

$g'(\mu) \propto \frac{1}{\sqrt{\mu}}$ $g(\mu) \propto \sqrt{\mu}$

Recap: Regularity conditions

Choudhury & Das, 2026

HIGH DIMENSIONAL GAUSSIAN AND BOOTSTRAP APPROXIMATIONS IN GENERALIZED LINEAR MODELS

MAYUKH CHOUDHURY AND DEBRAJ DAS

$d = o(n^{1/2})$

$\frac{d}{\sqrt{n}} \to 0$ as $n \to \infty$

Link

ABSTRACT. Generalized Linear Models (GLMs) extend ordinary linear regression by linking the mean of the response variable to covariates through appropriate link functions. This paper investigates the asymptotic behavior of GLM estimators when the parameter dimension d grows with the sample size n. In the first part, we establish Gaussian approximation results for the distribution of a properly centered and scaled GLM estimator uniformly over class of convex sets and Euclidean balls. Using high-dimensional results from Fang and Koike (2024) for the leading Bahadur term, bounding remainder terms as in He and Shao (2000), and applying Nazarov's (2003) Gaussian isoperimetric inequality, we show that Gaussian approximation holds when $d = o(n^{2/5})$ for convex sets and $d = o(n^{1/2})$ for Euclidean balls—the best possible rates matching those for high-dimensional sample means. We further extend these results to the bootstrap approximation when the covariance matrix is unknown. In the second part, when $d >> n$, a natural question is to answer whether all covariates are equally important. To answer that, we employ sparsity in GLM through the Lasso estimator. While Lasso is widely used for variable selection, it cannot

(C.1) $y_i \in \mathbf{R}$ for all i, h is the identity function and $b(u) = u^2/2$ (which is the case for linear regression) or, $y_i \geq 0$ for all i and $-h$ & h_1 are strictly convex (which is the case for logistic, poisson, gamma regression etc).

(C.2) h is thrice continuously differentiable and g^{-1} is twice continuously differentiable.

(C.3) $\boldsymbol{S}_n$ and $\boldsymbol{L}_n$ are positive definite matrices.

(C.4) For some $\alpha_1, \alpha_2, \alpha_3 > 0$, satisfying $\alpha_1 + 2\alpha_2 + 2\alpha_3 < 1/2$ with $0 < \alpha_2 < 1/14$, there exist a constant $0 < c_1 < \infty$, and two other constants $0 < c_2 < c_3 < \infty$, for which following holds: $\lambda_{\min}(\boldsymbol{S}_n) \geq c_1 n^{-\alpha_1}$ and $c_2 n^{-\alpha_2} \leq \lambda_{\min}(\mathbf{E}(\boldsymbol{L}_n)) < \lambda_{\max}(\mathbf{E}(\boldsymbol{L}_n)) \leq c_3 n^{\alpha_3}$.

(C.5) $\max_{i \in \{1,..,n\}} \|\boldsymbol{x}_i\|_\infty = O(1)$.

(C.6) $\|\boldsymbol{\beta}\|_\infty = O(1)$.

(C.7) $n^{-1} \sum_{i=1}^n \mathbf{E}|y_i|^6 = O(1)$.

(C.8) $n^{-1} \sup_{\{||\boldsymbol{\alpha}||, ||\boldsymbol{\kappa}||=1\}} \sum_{i=1}^n |\boldsymbol{\alpha}^\top \boldsymbol{x}_i|^8 |\boldsymbol{\kappa}^\top \boldsymbol{x}_i|^8 = O(1)$.

(C.9.i) $\max_{i \in \{1,..,n\}} |h'(\boldsymbol{x}_i^\top \boldsymbol{\beta})| = O(1)$.

(C.9.ii) $\max_{i \in \{1,..,n\}} |h''(\boldsymbol{x}_i^\top \boldsymbol{\beta})| = O(1)$.

(C.9.iii) $\max_{i \in \{1,..,n\}} |(g^{-1})'(\boldsymbol{x}_i^\top \boldsymbol{\beta})| = O(1)$.

(C.9.iv) $n^{-1} \sum_{i=1}^n \sup_{\{|z_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}| < \delta\}} |(g^{-1})''(z_i)|^8 = O(1)$.

(C.9.v) $n^{-1} \sum_{i=1}^n \sup_{\{|z_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}| < \delta\}} |h'''(z_i)|^{12} = O(1)$.

$E(Y_i|x_i) = \mu_i$

$g(\mu_i) = x_i^\top \beta$

$h(x_i^\top \beta)$

$= g^{-1}(\mu_i)$

Regularity conditions: asymptotic normality

LaE §1.7.2; AoS Thm 9.18; MS Thm 5.4

1. The support of the density for X, i.e., the set $\{x : f(x;\theta) > 0\}$ does not depend on θ.
2. The true value θ_* is contained in an open subset of Θ, and for all θ in this open subset, the density $f(x;\theta)$ is differentiable with respect to θ for all x in the support of the density.
3. $f(x;\theta)$ is three times continuously differentiable with respect to θ for all x in the support of the density.
4. $\mathrm{E}_\theta\{u(\theta;X)\} = 0, \text{ and } \mathrm{E}_\theta\{u(\theta;X)u^\mathsf{T}(\theta;X)\} = \mathrm{E}_\theta\{-\partial u(\theta;X)/\partial\theta^\mathsf{T}\}$, and these expectations exist and are finite for all θ in the open subset defined in 2.
5. The matrix $I_1(\theta) = \mathrm{E}_\theta\{u(\theta;X)u^\mathsf{T}(\theta;X)\}$ is positive definite for all θ in the open subset defined in 2.
6. There exist functions $M_{abc}(\cdot)$ such that

$$\left|\frac{\partial^3}{\partial\theta_a\partial\theta_b\partial\theta_c}\ell(\theta;x)\right| \le M_{abc}(x), \quad ||\theta-\theta_*|| \le \delta \quad \text{and } \mathrm{E}_{\theta_*}\{M_{abc}(X)\} < \infty.$$

$X_1, \ldots, X_n$ *i.i.d.* $f(x;\theta), \theta \in \Theta$

Bayesian inference

Under regularity conditions on the model, and the prior, the posterior is asymptotically normal.

More precisely

$$\int_{a_n}^{b_n} \pi(\theta \mid \boldsymbol{x}) d\theta \xrightarrow{p} \int_a^b \phi(t) dt$$

$$a_n = \widehat{\theta} + a I_n(\widehat{\theta})^{-1/2}, \quad b_n = \widehat{\theta} + b I_n(\widehat{\theta})^{-1/2}$$

$I_n(\theta)$ is the expected Fisher information in the sample, and can be replaced (to this order of approximation) by the observed Fisher information $J(\widehat{\theta})$.

$J(\widehat{\theta})$ in my notation; text defines $J = I^{-1}$ and uses H for $-\ell''$

Informally, $\pi(\theta \mid \boldsymbol{x}) \dot{\sim} N\{\widehat{\theta}, I_n^{-1}(\widehat{\theta})\}$, and an approximate 95% credible interval for θ is $\widehat{\theta} \pm 1.96\widehat{se}$

same as 95% CI

Laplace approximation

Laplace approx =

equivalently, $\pi(\theta|x) \dot{\sim} N\left(\hat{\theta}_\pi, J_{n\pi}^{-1}(\hat{\theta})\right)$ (*)

$$\hat{\theta}_\pi = \arg\sup_\theta \pi(\theta|x)$$

under reg. on $f(x;\theta)$ & $\pi(\theta)$

$$J_{n\pi} = -\frac{\partial^2}{\partial\theta^2} \log \pi(\theta|x)\Big|_{\theta=\hat{\theta}_\pi}$$

$$\pi(\theta|x) = \frac{L(\theta;x)\pi(\theta)}{m(x)} \leftarrow \max_\theta$$

Laplace approximation

$$\pi(\theta \mid x) = L(\theta; x)\pi(\theta) \Big/ \underbrace{\int L(\theta; x)\pi(\theta)\, d\theta}_{\text{den}} \qquad \theta \in \mathbb{R}$$

$$\text{den} = \int e^{\ell(\theta; x)} \pi(\theta)\, d\theta \longrightarrow \int e^{\ell(\theta; x) + \log \pi(\theta)}\, d\theta$$

$$\doteq \int e^{\ell(\hat\theta) + (\theta - \hat\theta)\ell'(\hat\theta) + \frac{1}{2}(\theta - \hat\theta)^2 \ell''(\hat\theta)} \left\{ \pi(\hat\theta) + (\theta - \hat\theta)\pi'(\hat\theta) \right\} d\theta$$

$$= \pi(\hat\theta) e^{\ell(\hat\theta)} \cdot \int e^{-\frac{1}{2}(\theta - \hat\theta)^2 \{-\ell''(\hat\theta)\}} \left\{ 1 + (\theta - \hat\theta)\frac{\pi'(\hat\theta)}{\pi(\hat\theta)} \right\} d\theta$$

$$= \sqrt{2\pi}\, \pi(\hat\theta) e^{\ell(\hat\theta)} j(\hat\theta)^{-1/2} \cdot \int \underbrace{\frac{j(\hat\theta)^{1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2} j(\hat\theta)(\theta - \hat\theta)^2}}_{} \left\{ 1 + \underbrace{(\theta - \hat\theta)\frac{\pi'}{\pi}}_{} \right\} d\theta$$

$$= \sqrt{2\pi}\, \pi(\hat\theta) e^{\ell(\hat\theta)} \mathcal{J}(\hat\theta)^{-\frac{1}{2}}$$

$$\pi(\theta \mid \underline{x}) \doteq \frac{1}{\sqrt{2\pi}} \cdot e^{\ell(\theta) - \ell(\hat\theta)} \cdot \pi(\hat\theta) \mathcal{J}(\hat\theta)^{1/2}$$

$$\theta \in \mathbb{R}^d \doteq \frac{1}{\sqrt{2\pi}^d} e^{\ell(\underline{\theta}) - \ell(\hat{\underline{\theta}})} \cdot \pi(\hat{\underline{\theta}}) \left|\mathcal{J}(\hat{\underline{\theta}})\right|^{1/2}$$

$d_n \uparrow n$

($\hat\theta$ = mle)

$\mathcal{J}(\hat\theta)$ = obs'd

INLA integrated nested Laplace approx^n

(normal version)

Choosing priors

MS p.287 ff

- conjugate priors
- non-informative priors
- convenience priors
- minimally/weakly informative priors
- hierarchical priors
- empirical Bayes

Posterior is in the same class as prior

same form

$f(x;\theta) = e^{\eta(\theta)\cdot t(x) - \kappa(\theta) - d(x)}$

$\pi(\theta;\alpha) =$ flat, "ignorance"

$x \sim N(\mu, \sigma^2)$

$\mu \sim N(a, b)$ & β_j i.i.d $N(b, \tilde{\sigma})$

$\sigma^2 \sim$ Inv. Gamma(,)

Gelman

$\beta \sim N(\underline{0}, \tau^2 \cdot I)$; $\tau^2 \sim$ Inv. G( )

HW 2 : estimate parameters of the prior

$$\pi(\underset{\sim}{\theta} \mid \underset{\sim}{x}) = \frac{L(\theta; \underset{\sim}{x})\pi(\underset{\sim}{\theta};\alpha,\beta)}{\int L(\theta; \underset{\sim}{x})\pi(\theta;\alpha,\beta)\,d\theta} \equiv m(\underset{\sim}{x};\alpha,\beta)$$

could use $m(\underset{\sim}{x};\alpha,\beta)$ + $x_1, \ldots, x_n$ to estimate α & β

non-informative

- $\pi(\theta) \propto I^{1/2}(\theta)$ $I^{1/2}(\theta) = \pi(\theta)$ $I^{1/2}(\tau) = \pi(\tau)$ $\tau = \tau(\theta)$
- Example: $X \sim Bin(n, \theta)$ $I(\theta) = n/\{\theta(1-\theta)\}, \quad 0 < \theta < 1$
- Jeffreys' prior for multiparameter $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$ not recommended even by Jeffreys $N(\mu, \sigma^2)$ (nfbc) $\pi(\mu, \sigma) = \frac{1}{\sigma^2}$
- for regression-scale models Jeffreys recommends $\pi(\beta, \sigma) \propto 1/\sigma$ ← leads to improper dist'n
- For normal theory linear regression, the conjugate prior is Normal for β and inverse gamma for σ, with $\beta \perp \sigma$. These are often the defaults.
- Posterior means are typically weighted averages of MLE and prior mean, with weight on MLE $\to 1$ with n one Bayes estimator $E(\theta|x)$

sample estimate → prior mean

Example: Bivariate normal

Table 3.1 *Scores from two tests taken by 22 students,* `mechanics` *and* `vectors`.

	1	2	3	4	5	6	7	8	9	10	11
`mechanics`	7	44	49	59	34	46	0	32	49	52	44
`vectors`	51	69	41	70	42	40	40	45	57	64	61

	12	13	14	15	16	17	18	19	20	21	22
`mechanics`	36	42	5	22	18	41	48	31	42	46	63
`vectors`	59	60	30	58	51	63	38	42	69	49	63

Table 3.1 shows the scores on two tests, `mechanics` and `vectors`, achieved by $n = 22$ students. The sample correlation coefficient between the two scores is $\hat{\theta} = 0.498$,

$$\hat{\theta} = \sum_{i=1}^{22}(m_i - \bar{m})(v_i - \bar{v}) \Big/ \left[\sum_{i=1}^{22}(m_i - \bar{m})^2 \sum_{i=1}^{22}(v_i - \bar{v})^2\right]^{1/2}, \quad (3.10)$$

with m and v short for `mechanics` and `vectors`, $\bar{m}$ and $\bar{v}$ their averages. We wish to assign a Bayesian measure of posterior accuracy to the true correlation coefficient θ, "true" meaning the correlation for the hypothetical population of all students, of which we observed only 22.

If we assume that the joint (m, v) distribution is bivariate normal (as

$\underline{m} \sim N(\mu_m, \sigma_m^2)$

$\underline{v} \sim N(\mu_v, \sigma_v^2)$

$\text{corr}(\underline{m}, \underline{v})$ is θ

Cov. $\theta \sigma_v \sigma_m$

$L(\mu_m, \mu_v, \sigma_v^2, \sigma_m^2, \theta; \underline{m}, \underline{v})$

could find a f- of $\hat{\theta}$, θ only to use as a Lik.

$$f(\hat{\theta} \mid \theta) = \frac{1}{\pi}(n-2)(1-\theta^2)^{(n-1)/2}(1-\hat{\theta}^2)^{(n-4)/2} \int_0^\infty \frac{1}{(\cosh(w) - \theta\hat{\theta})^{n-1}} dw$$

↑

$\pi(\theta|\hat{\theta}) \propto f(\hat{\theta}|\theta)\pi(\theta)$

Example: Bivariate normal

EH §3.1

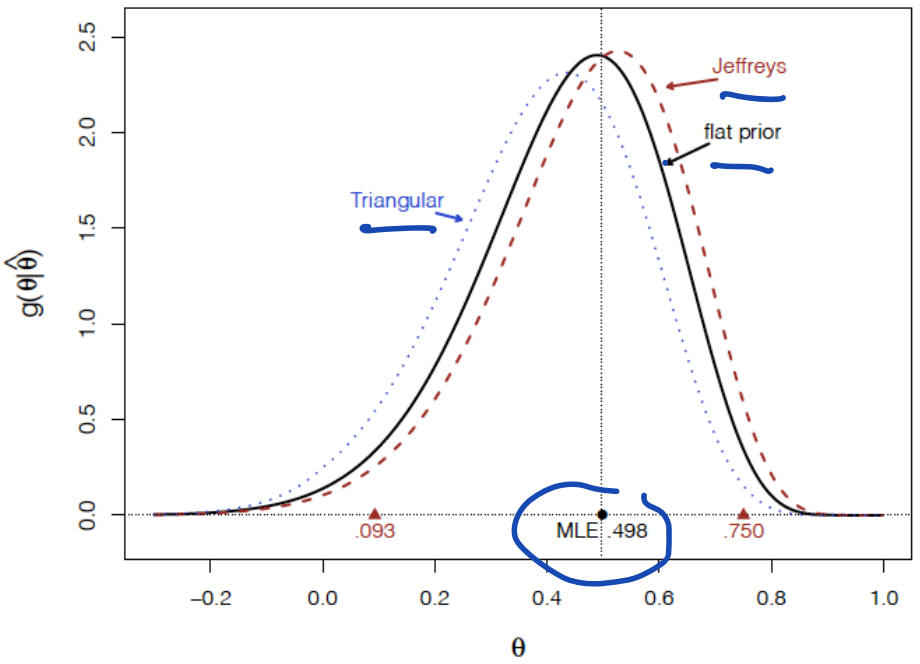


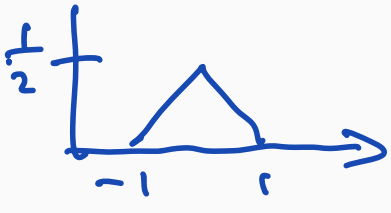
Figure 3.2 Student scores data; posterior density of correlation θ for three possible priors.

$$\pi_J = \frac{1}{1-\theta^2}$$

$$-1 \le \theta \le 1$$

$$\pi_{flat} = 1$$

π_{triang}



Example: Binomial

SM Ex.11.11

Table 11.2 Mortality rates r/m from cardiac surgery in 12 hospitals (Spiegelhalter *et al.*, 1996b, p. 15). Shown are the numbers of deaths r out of m operations.

A	0/47	*B*	18/148	*C*	8/119	*D*	46/810	*E*	8/211	*F*	13/196
G	9/148	*H*	31/215	*I*	14/207	*J*	8/97	*K*	29/256	*L*	24/360

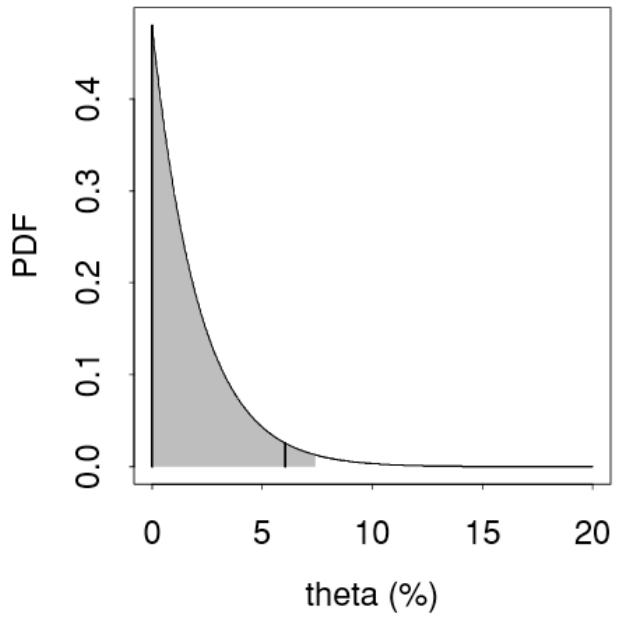
provided the mode lies inside the parameter space. Here $\tilde{J}(\theta)$ is the second deriva-

$\hat{P}(\text{dying in A}) = 0$

prior for hospital A $Beta(1, 1)$ $\frac{1}{49}$ posterior mean $\tilde{\theta}_A$ posterior mean

Example: Binomial

SM Ex.11.11



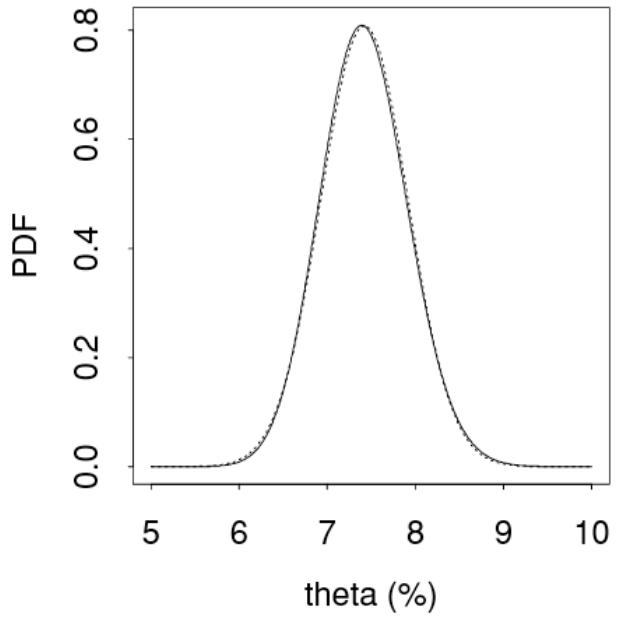


Figure 11.1 Cardiac surgery data. Left panel: posterior density for θ_A, showing boundaries of 0.95 highest posterior credible interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ (shaded). Right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).

↓ same θ

put all hospitals together; 208 failures '

(or HW2)

Marginalization

- Bayes posterior carries all the information about θ, given $\boldsymbol{x}$ — by definition
- probabilities for any set A computed using the posterior distribution
- $\mathrm{pr}(\boldsymbol{\Theta} \in A \mid \boldsymbol{x}) = \int_A \pi(\underline{\theta} \mid \underline{x})\, d\underline{\theta}$
- if $\theta = (\psi, \boldsymbol{\lambda})$, ... $\leftarrow \int \pi(\psi, \underline{\lambda} \mid \underline{x})\, d\underline{\lambda} = \pi_m(\psi \mid \underline{x})$ marginal post.
- or, if $\psi = \psi(\theta)$ $\rightarrow \int_{\{\underline{\theta} : \pi(\underline{\theta}) = \psi\}} \pi(\underline{\theta} \mid \underline{x})\, d\underline{\theta}$
- in this context, 'flat' priors can have a large influence on the marginal posterior

Example: many normal means

Stein, 1959

$X_i \sim N(\mu_i, \frac{1}{n})$ $\quad X_i \sim N(\mu_i, 1/n)$ $\quad \pi(\mu_i) = 1 \quad i = 1, \dots, k$

$\sqrt{n} X_i \sim N(\mu_i \sqrt{n}, 1)$

$n X_i^2 \sim \chi^2_1(\mu_i^2 n)$

$\pi(\mu_i | x_i) \sim N(x_i, 1/n)$

$E(\psi | \underset{\sim}{X}) = n + \Sigma X_i^2$

$$\pi(\underset{\sim}{\mu} | \underset{\sim}{x}) \propto \prod_{i=1}^{k} e^{-\frac{n}{2}(x_i - \mu_i)^2}$$

$E[\quad] = n + n + \Sigma \mu_i^2$

$$\psi = \sum_{i=1}^{n} \mu_i^2 \qquad \int_{\{\underset{\sim}{\mu} : \Sigma \mu_i^2 = \psi\}} \pi(\underset{\sim}{\mu} | \underset{\sim}{x}) \, d\underset{\sim}{\mu} = \pi(\psi | \underset{\sim}{x})$$

$$\text{posterior of } n \sum_{i=1}^{k} \mu_i^2 \sim \chi^2_k\left(n \sum_{i=1}^{k} x_i^2\right) \qquad n \sum_{i=1}^{k} X_i^2 \sim \chi^2_k\left(n \sum_{i=1}^{k} \mu_i^2\right)$$

$n \cdot 2($

$E(\tilde{\psi}_B | \underset{\sim}{X}) = \frac{k}{n} + \Sigma X_i^2$ $\quad E(E(\tilde{\psi}_B | \underset{\sim}{X}) = \frac{2k}{n} + \Sigma \mu_i^2$

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

notation

- what is $\hat{\theta}_n$ estimating ?

$\hat{\theta}_n \xrightarrow{p} ??$ if it converges

e.g. $X_i \sim N(\mu, 1)$

$\bar{X} \to E_F(X_i)$

$= \int x_i \, dF(x_i)$

let $\theta(F)$ be defined as

$$E_F\{\ell'(\theta; X_i)\} = 0$$

$\theta = \theta(F)$

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$
- maximum likelihood estimator based on model:

notation

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?
- define the parameter $\theta(F)$ by (estimand)

$$\int_{-\infty}^{\infty} \ell'\{\theta(F); x\} dF(x) = 0$$

-

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

-

$$\sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}] dF(x))^2}$$

$$0 = \sum \ell'(\hat{\theta}; x_i) = \sum \ell'(\theta(F); x_i) + \{\hat{\theta} - \theta(F)\} \sum \ell''(\theta(F); x_i)$$

$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \doteq \frac{\frac{1}{\sqrt{n}} \sum \ell'(\theta(F); X_i)}{-\frac{1}{n} \sum \ell''(\theta(F); X_i)} + \dots + R_n$$

$$\frac{1}{\sqrt{n}} \sum \ell'(\theta(F); X_i) \xrightarrow{d} N\left(0, \operatorname{var}_F\{\ell'(\theta(F); X_1)\}\right)$$

$$-\frac{1}{n} \sum \ell''(\quad) \xrightarrow{P} E_F\{-\ell''(\theta(F); X_1)\}$$

no Bartlett identity

Misspecified models

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{\theta(F); x\}]^2 dF(x)}{(\int [\ell''\{\theta(F); x\}]^2 dF(x))^2}$$
- more generally, for $\theta \in \mathbb{R}^p$,
$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N_p\{0, G^{-1}(F)\}$$
- $$G(F) = J(F)I^{-1}(F)J(F),$$
- $$J(F) = \int -\ell''\{\theta(F); x_i\} dF(x_i), \quad I(F) = \int \{\ell'(\theta(F); x_i)\}\{\ell'(\theta(F); x_i)\}^T dF(x_i)$$

Godambe information

sandwich variance

M-estimators

- Consider maximizing a function other than the log-likelihood function
- VdV notation $M_n(\theta; \boldsymbol{x}) \equiv \frac{1}{n}\Sigma_{i=1}^n m(\theta; x_i)$ — M est
- Typically found by solving $M_n'(\theta; \boldsymbol{x}) = 0$, or

VdV Z-estimators

$$\Psi_n(\theta; \boldsymbol{x}) = \sum_{i=1}^n \Psi(\theta; x_i) = 0$$

Z - estimators

- SM notation: $g(\boldsymbol{Y}; \theta) = \Sigma_{i=1}^n g(Y_i; \theta)$ is a (set of) estimating equations
- $g(y; \theta)$ is an unbiased estimating function if

$$\mathrm{E}\{g(Y; \theta)\} = 0 = \int g(y; \theta) f(y; \theta) dy$$

model $f(y;\theta)$ not using

- the score function is an unbiased estimating function
- the solution of $g(\boldsymbol{y}; \theta) = 0$ is an M-estimator

Score eq^n to est. θ

- Examples: moment estimators
- Example: median
- Example: Huber estimator

- Examples: moment estimators $g(x;\theta) = (x^k - \mu_k)$ SM Ex.7.15

$\frac{1}{n}\sum X_i^k = \hat{\mu}_k$
$= E(X)^k$

- Example: median $g(x;\theta) = \text{sign}(x - \theta)$ VdV Ex.5.4

$\frac{1}{n}\sum \text{sign}(X_i - \theta) = 0$
$\hat{\theta}_n = \text{median}(X_1, \ldots, X_n)$

- Example: Huber estimator $g(x;\theta) = g(x - \theta)$, where SM Ex 7.19

$$g(x;\theta) = \begin{cases} -k, & x \le \theta - k, \\ x, & -k < x - \theta < k \\ k, & x \ge \theta + k \end{cases}$$

$\overline{X}_n$ "truncated at $\pm k$"

- Example: quantile regression Koenker[1]

$$\widehat{\beta}_\tau = \arg\min_\beta \rho_\tau(y_i - x_i^T\beta), \quad \rho_\tau(u) = u\{\tau - I(u \le 0)\}$$

[1] Koenker, Roger, and Kevin F. Hallock (2001). Quantile Regression. *J. Econ. Persp.* **15**, 143–56.

- $0 = g(\boldsymbol{x}; \widetilde{\theta}) =$

$\hat{\theta}_n$ $\frac{1}{n}\sum g(x;\hat{\theta}) =$ $0 = g(\boldsymbol{x};\widetilde{\theta}) = g(\boldsymbol{x};\theta) + (\widetilde{\theta}-\theta)\dot{g}(\boldsymbol{x};\theta) + \ldots$

$\dot{g}(x;\theta) = \frac{\partial}{\partial\theta} g(x;\theta)$

$$\widetilde{\theta} - \theta \doteq \frac{\frac{1}{n}\sum g(x_i;\theta)}{-\frac{1}{n}\dot{g}(x_i;\theta)} \doteq \frac{\frac{1}{n}\sum g(x_i;\theta)}{\mathrm{E}\{-\dot{g}(X;\theta)\}}$$

$$\mathrm{E}(\widetilde{\theta}-\theta) \doteq 0, \quad \mathrm{var}(\widetilde{\theta}-\theta) \doteq \frac{1}{n}\mathrm{var}\{g(X;\theta)\}/\mathrm{E}^2\{\dot{g}(X;\theta)\} \tag{1}$$

$$\sqrt{n}(\widetilde{\theta}-\theta) \xrightarrow{d} N\{0, G^{-1}(\theta)\}, \tag{2}$$

$$G(\theta) = E\{-\frac{\partial}{\partial\theta^T} g(X;\theta)\}[\mathrm{var}\{g(X;\theta)\}]^{-1} E\{-\frac{\partial}{\partial\theta} g(X;\theta)\} \tag{3}$$

$g(x_i; \theta)$

7 · *Estimation and Hypothesis Testing*

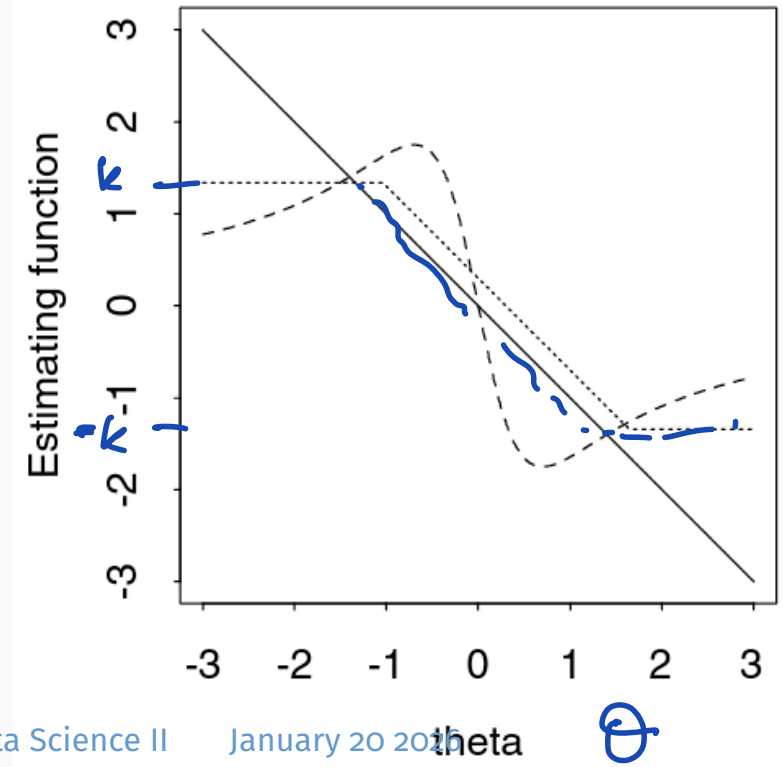


Figure 7.3 Estimating functions. Left: construction of $g(y;\theta)$ (heavy) as the sum of $g(y_j;\theta)$ for a sample of size $n = 3$ shown by the rug. The lines $g = 0$ (dots) and $\theta = \tilde{\theta}$ (dashes) are also shown. Right: estimating functions for the mean (solid), the Huber estimator (dots) and a redescending M-estimator (dashes), slightly offset to avoid overplotting.

Interval estimation

AoS 6.3.2; MS 7.1

- a $(1-\alpha)$-level confidence interval for $\theta \in \mathbb{R}$ is $\{L(\boldsymbol{X}), U(\boldsymbol{X})\}$, with

$$\operatorname{pr}_\theta\{L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\} \geq 1-\alpha$$

for all θ,

"with = for some θ" (MS)

equality?

- similarly upper and lower confidence bounds

$\underset{\sim}{X} \sim f(\underset{\sim}{x}; \theta)$

$$P_{r_\theta}\{L(\underset{\sim}{X}) \leq \theta\} \geq 1-\alpha$$

$$\int_{-\infty}^{\theta} \ell(\underset{\sim}{x}) f(\underset{\sim}{x}; \theta)\, d\underset{\sim}{x} \geq 1-\alpha$$

$$P_{r_\theta}\{u(\underset{\sim}{X}) \geq \theta\} \geq 1-\alpha$$

$$\int_{\theta}^{\infty} u(\underset{\sim}{x}) f(\underset{\sim}{x}; \theta)\, dx \geq 1-\alpha$$

Interval estimation

- a $(1-\alpha)$-level confidence interval for $\theta \in \mathbb{R}$ is $\{L(\boldsymbol{x}), U(\boldsymbol{x})\}$, with

$$\mathrm{pr}\{L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\} \geq 1-\alpha$$

equality?

- similarly upper and lower confidence bounds

- a $(1-\alpha)$-level confidence region for $\theta \in \Theta$, is a set $R(\boldsymbol{X}) \subset \Theta$, with

$$\mathrm{pr}\{\theta \in R(\boldsymbol{X})\} \geq 1-\alpha$$

$\theta \in \mathbb{R}^p$ (say)

Interval estimation

- a $(1-\alpha)$-level confidence interval for $\theta \in \mathbb{R}$ is $\{L(\boldsymbol{X}), U(\boldsymbol{X})\}$, with

$$\mathrm{pr}\{L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\} \geq 1-\alpha$$

equality?

- similarly upper and lower confidence bounds

- a $(1-\alpha)$-level confidence region for $\theta \in \Theta$, is a set $R(\boldsymbol{X}) \subset \Theta$, with

$$\mathrm{pr}\{\theta \in R(\boldsymbol{X})\} \geq 1-\alpha$$

- a pivotal quantity $h(\boldsymbol{x};\theta)$ is a function of $\boldsymbol{X}$ and θ with a known distribution: we can find a, b s.t.

continuity

$$\mathrm{pr}\{a \leq h(\boldsymbol{X};\theta) \leq b\} \geqslant 1-\alpha$$

if $X \sim$ const

- inversion of this gives a $(1-\alpha)$ confidence region for θ

Binomial example

AoS Ex 6.15,7

$X \sim Binom(n, \theta), \quad 0 \leq \theta \leq 1, \qquad \widehat{\theta}(x) = x/n$

$\widehat{\theta} \dot{\sim} N\{\theta, (1-\theta)\theta/n\}$

approximate $1-\alpha$ CI version 1 — Wald

$$\widehat{\theta} \pm 1.96 \sqrt{\widehat{\theta}(1-\widehat{\theta})/n}$$

approximate $1-\alpha$ CI version 2 — Wilson

$$P_\theta\left\{|\widehat{\theta} - \theta| \leq z_{\alpha/2}\left\{\frac{(1-\theta)\theta}{n}\right\}^{1/2}\right\} = 1-\alpha$$

guaranteed $1-\alpha$ CI — nonparametric

$$-z\sqrt{\frac{\theta(1-\theta)}{n}} \leq (\widehat{\theta} - \theta) \leq z\sqrt{(1-\theta)\theta/n}$$

Hoeffding's inequality:

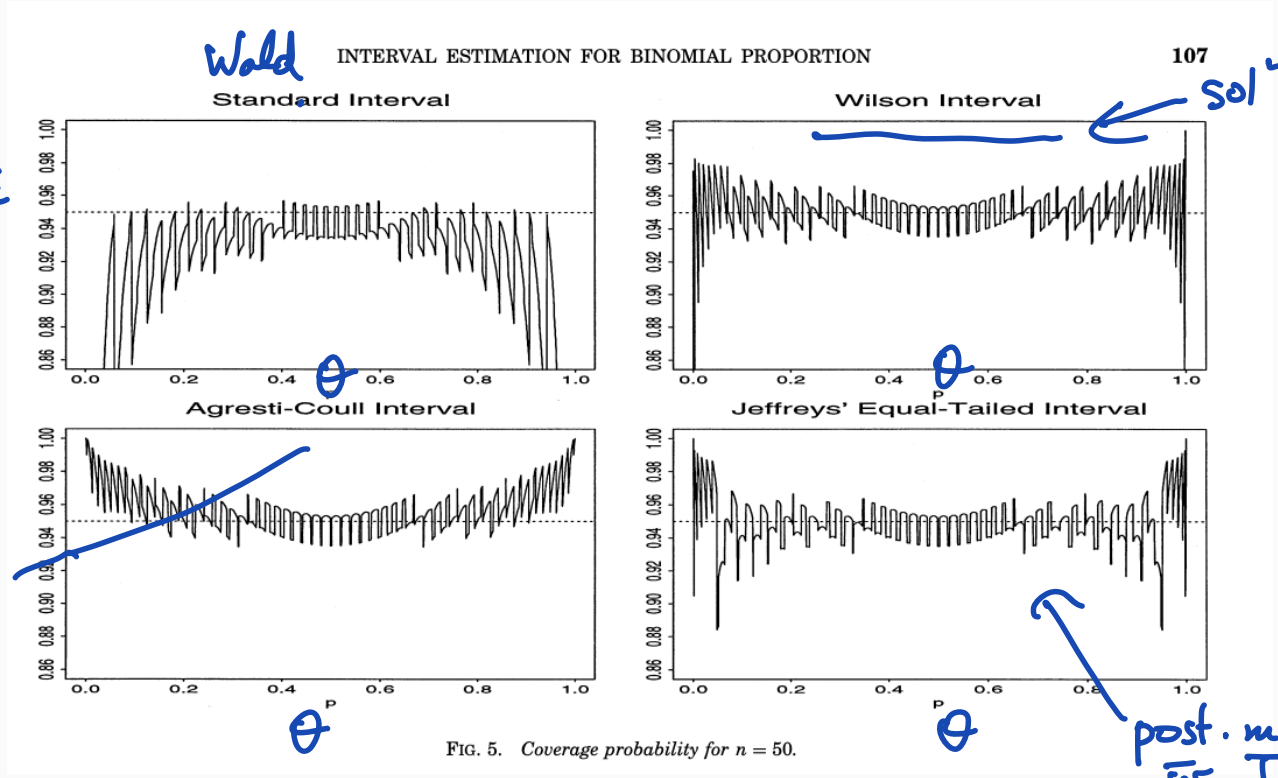
$$(\widehat{\theta} - \epsilon_n, \widehat{\theta} + \epsilon_n), \quad \epsilon_n^2 = \log(2/\alpha)/(2n)$$

$$\mathrm{pr}(\theta \in \widehat{\theta} - \epsilon_n, \widehat{\theta} + \epsilon_n) \geq 1 - \alpha$$

no normal approx

Binomial example

Brown et al.[2]



[2] Brown, L.D., Cai, T., DasGupta, A. (2001). Interval Estimation for a Binomial Proportion, *Statistical Science* **16**, 101–133. link

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

plug-in est'rs

- empirical cdf

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

- properties:

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

any fixed t

- pointwise approximate confidence limits $\hat{F}_n(t) \pm z_{1-\alpha/2}[\hat{F}_n(t)\{1 - \hat{F}_n(t)\}]^{1/2}$

- recall $X_1, \ldots, X_n$, i.i.d. $F(\cdot)$ plug-in est'rs
- empirical cdf

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{(i)} \leq t\}$$

- properties:

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

any fixed t

- pointwise approximate confidence limits $\hat{F}_n(t) \pm z_{1-\alpha/2}[\hat{F}_n(t)\{1 - \hat{F}_n(t)\}]^{1/2}$
- 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

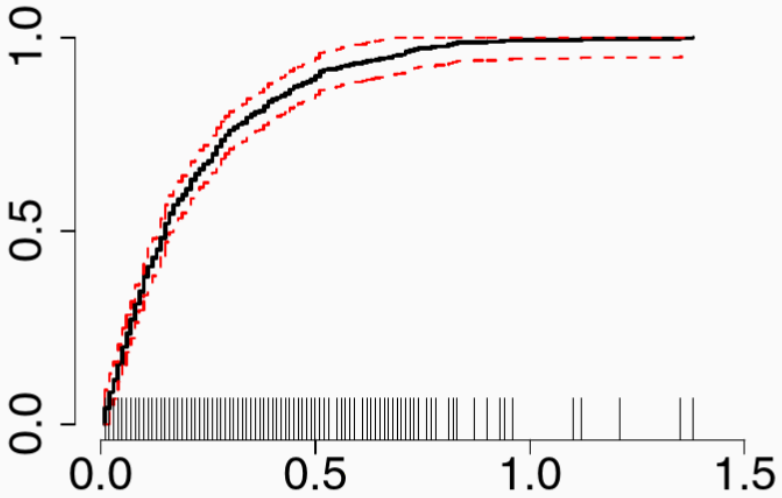


FIGURE 7.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

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 $\widehat{F}_n(.6) - \widehat{F}_n(.4) = .93 - .84 = .09$. ■

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

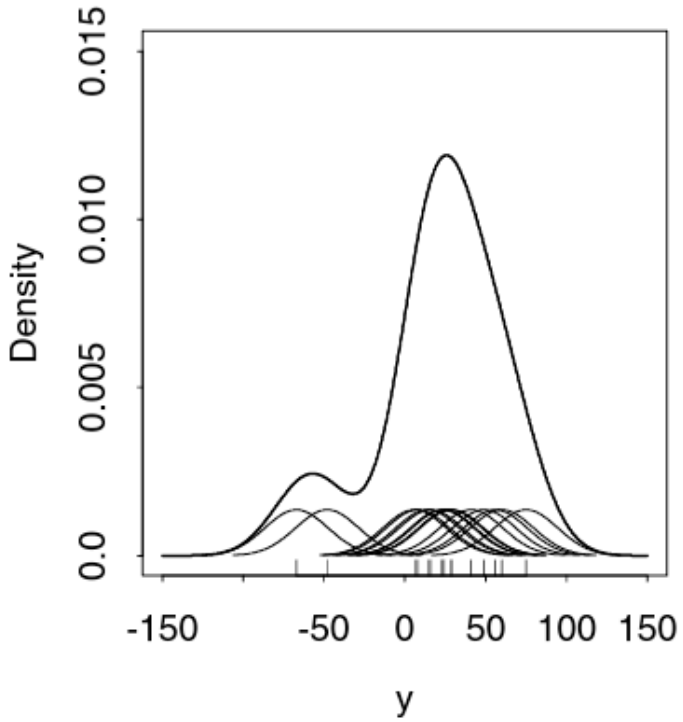
$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- with a symmetric kernel function, for small h,

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

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

- mean-squared error



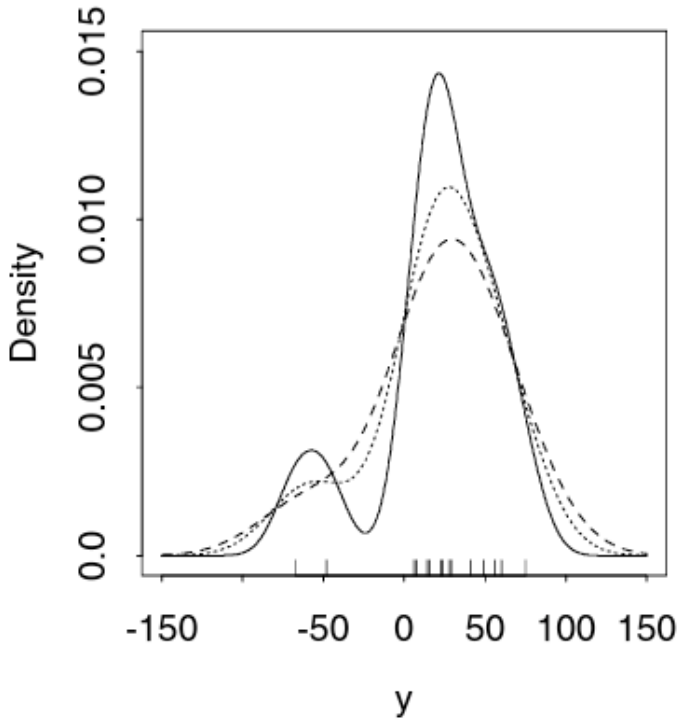


Figure 7.2 Kernel density estimates for maize data. Left: construction of kernel estimate (heavy) as sum of 15 scaled normal densities centred at the y_j, with $h = 19.5$. Right: density estimates with $h = 13.3$ (solid), $h = 23.2$ (dots) and $h = 30$ (dashes).

Density estimation

AoS 20.3, SM 7.1.2

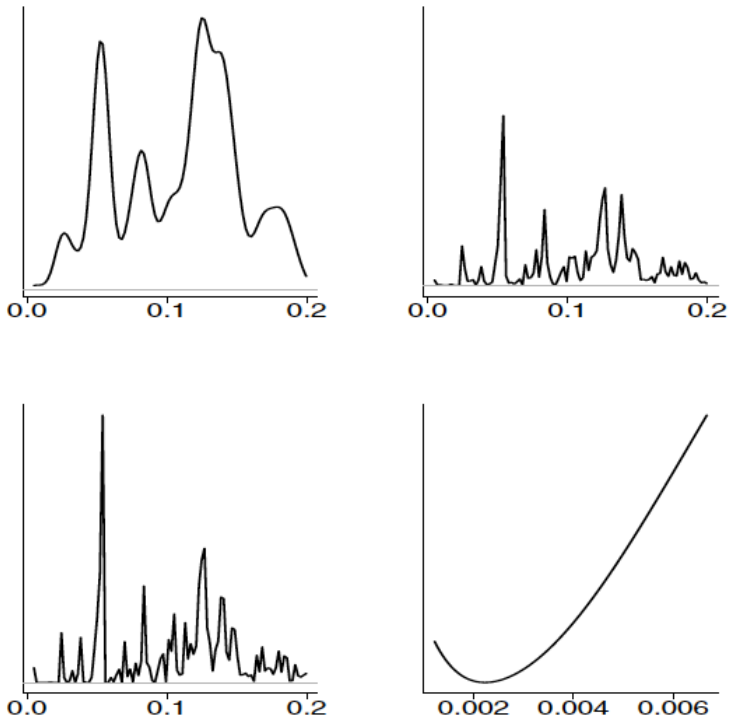


FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value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

318 20. Nonparametric Curve Estimation

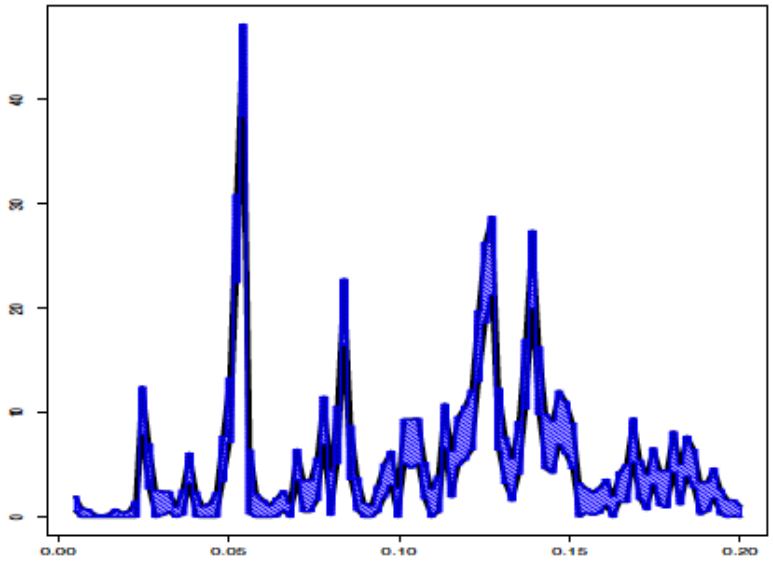
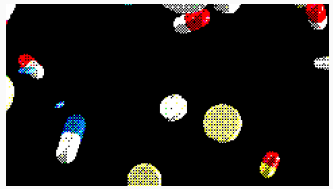


FIGURE 20.7. 95 percent confidence bands for kernel density estimate for the astronomy data.

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

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

Link 

My trail:

“Published in 2025”
`https://www2.unlearn.ai/l/1055293/2025-04-29/4d7f2/1055293/1757437116rfdemBsS/PD_C`

ASA Substack Dec 2025
`https://asabiopreport.substack.com/p/unleashing-ai-generated-digital-twins`

arxiv paper `https://arxiv.org/pdf/2405.01488`