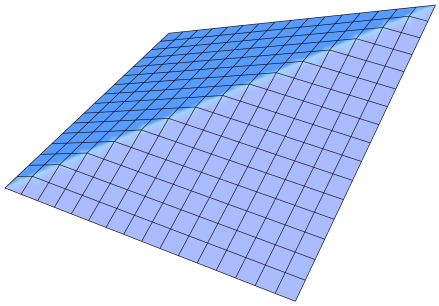

STA 4508H: Topics in Likelihood Inference

Prof. Nancy Reid • Winter 2022 • University of Toronto
W 2:00-5:00PM in SS 1087

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Wednesday, January 12, 2022

1. Likelihood Foundations

The Role and Scope of Likelihood

A National Post graphic once described the probability of depression at different ages as a "percentage likelihood." The quantity in that graphic is a probability rather than a likelihood, although its shape happened to resemble that of a log likelihood function. This small misuse of terminology is a useful reminder that probability and likelihood are closely related but answer inverse questions.

Statistical inference begins with a scientific question. A probability model and a choice of parameter translate that question into mathematical and statistical terms. This is the starting principle in Cox's *Principles of Statistical Inference*. For example, if astronomers regard the number of observed jets from a distant star as Poisson, then the scientific question becomes a question about the parameter of a Poisson model. Choosing a plausible model and parametrization is often the hardest part of the analysis; once they have been chosen, a common inferential theory can be carried from one application to another.

The likelihood function is the probability model viewed as a function of its unknown parameter. It makes probability modelling central, supplies more information than a point estimate or a test of a single hypothesis, and provides a broadly accepted basis for inference. As applications become more complicated, however, the ordinary likelihood may be difficult to evaluate or may not be the appropriate object. This motivates a large family of related constructions: marginal likelihood, conditional likelihood, profile likelihood, adjusted profile likelihood, semiparametric likelihood, partial likelihood, quasi-likelihood, composite likelihood, empirical likelihood, penalized likelihood, simulated likelihood, indirect inference, bootstrap likelihood, hierarchical likelihood, weighted likelihood, pseudolikelihood, local likelihood, and sieve likelihood. Some of these address nuisance parameters or nearly nonparametric models, some deliberately work with misspecified models, and others replace an intractable likelihood by a computable analogue.

The Likelihood Function and Model Examples

Let Y take values in a sample space $\mathcal{Y}$. A **parametric model** is a family

$$\big\{ f(y;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \big\},$$

where $f(\,\cdot\,;\boldsymbol{\theta})$ is a density with respect to a fixed dominating measure. Thus f may be an ordinary density with respect to Lebesgue measure or a probability mass function with respect to counting measure.

Definition 1.1 For observed data y, a **likelihood function** for $\boldsymbol{\theta}$ is any function satisfying

$$L(\boldsymbol{\theta}; y) = c(y) f(y;\boldsymbol{\theta}) \propto f(y;\boldsymbol{\theta}),$$

where $c(y) > 0$ does not depend on $\boldsymbol{\theta}$. The corresponding **log likelihood function** is

$$\ell(\boldsymbol{\theta}; y) = \log L(\boldsymbol{\theta}; y) = \log f(y;\boldsymbol{\theta}) + a(y),$$

where $a(y) = \log c(y)$.

Only relative likelihood values are well defined. This freedom is essential: if $z = g(y)$ is a one-to-one transformation of the data, then the Jacobian changes the density by a factor depending on the observation but not on $\boldsymbol{\theta}$. The likelihoods based on y and z are therefore proportional and contain the same information about $\boldsymbol{\theta}$.

For responses $\boldsymbol{y} = (y_1, \ldots, y_n)$ and covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the joint model may be written as $f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$. If the responses are conditionally independent, then

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \qquad \text{and} \qquad \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) + a(\boldsymbol{y}).$$

The conditioning covariates are commonly suppressed when they play no role in the theoretical calculation.

The parameter dimension need not be small. In a genomic regression, for instance, $n = 30$ patients might be accompanied by measurements on $p = 30{,}000$ genes. The dimension may exceed the sample size or grow with it. The parameter can even be a function, as in

$$Y_i = m(x_i) + \varepsilon_i, \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where the only initial restriction is that the unknown function m be twice differentiable. A **regular**

model has fixed parameter dimension $p < n$ as n increases. Regular models share a convenient general theory; nonregular models can fail to be regular in many different ways.

The following models illustrate the range already covered by Definition 1.1.

Example 1.2 Suppose that $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. With $\boldsymbol{\theta} = (\mu, \sigma^2)$,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

In a Gaussian linear regression with $\mathbb{E}[Y_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, the likelihood becomes

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\right),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

A flexible mean function can be constructed from known basis functions $B_1, \ldots, B_J$:

$$m(x) = \sum_{j=1}^{J} \phi_j B_j(x).$$

The Gaussian likelihood then has the same form with $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ replaced by $\sum_{j=1}^{J} \phi_j B_j(x_i)$, and the parameter is $\boldsymbol{\theta} = (\phi_1, \ldots, \phi_J, \sigma^2)$. Polynomial, trigonometric, and B-spline bases are common choices. Such a construction is often called nonparametric even though it is a finite-dimensional parametric model for fixed J; the description becomes more natural when J grows with the sample size.

Example 1.3 Consider the first-order autoregressive model

$$Y_i = \mu + \rho(Y_{i-1} - \mu) + \varepsilon_i, \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

If μ is treated as fixed and $\boldsymbol{\theta} = (\rho, \sigma^2)$, then dependence changes the product representation to

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f_0(y_0; \boldsymbol{\theta}) \prod_{i=1}^{n} f(y_i \mid y_{i-1}; \boldsymbol{\theta}),$$

where f_0 supplies the initial distribution. More complicated autoregressive and moving average models use the same principle.

Example 1.4 Suppose that $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. Writing $Y_{(n)}$ for the sample maximum,

$$L(\theta; \boldsymbol{y}) = \theta^{-n} \mathbb{1}\{\theta \geq Y_{(n)}\}.$$

The likelihood is zero below the sample maximum, attains its maximum at $Y_{(n)}$, and then decreases as θ^{-n}. Hence $\hat{\theta} = Y_{(n)}$. The support depends on θ, so this is a nonregular model and the usual differentiable likelihood theory does not apply directly.

Example 1.5 Let $0 < Y_1 < \cdots < Y_n < \tau$ be the event times from a nonhomogeneous Poisson process with rate function λ. Its log likelihood is

$$\ell(\lambda; \boldsymbol{y}) = \sum_{i=1}^{n} \log(\lambda(y_i)) - \int_0^{\tau} \lambda(u)\, \mathrm{d}u.$$

The sum records the rate at the observed event times, while the integral records the stretches of continuous time in which no event occurred. Unless λ is parameterized in some way, it is an infinite-dimensional parameter. This expression is developed in Davison's *Statistical Models*.

Example 1.6 For a k-class classification problem, write $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{ik})$, where exactly one component equals 1 and the rest equal 0. If

$$p_{ic} = \mathbb{P}(Y_{ic} = 1 \mid \boldsymbol{x}_i; \boldsymbol{\theta}),$$

then the multinomial log likelihood is

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic} \log p_{ic}.$$

Its negative is the cross-entropy loss used in classification. A useful model links the probabilities across observations through a common finite-dimensional parameter rather than assigning a separate probability vector to each observation.

A small data set from Davison records ten failure times of a spring under stress:

$$225,\ 171,\ 198,\ 189,\ 189,\ 135,\ 162,\ 135,\ 117,\ 162.$$

The four panels in Figure 1 compare likelihood shapes for an exponential model, a two-parameter Weibull model, and a Cauchy location model with fixed unit scale. Every curve is normalized

relative to its maximum, so its absolute vertical scale is irrelevant. The Weibull contours are visibly nonelliptical, while the Cauchy likelihood contains several local modes. A numerical optimizer in a higher-dimensional version of the Cauchy problem could easily stop at a local rather than global maximum.

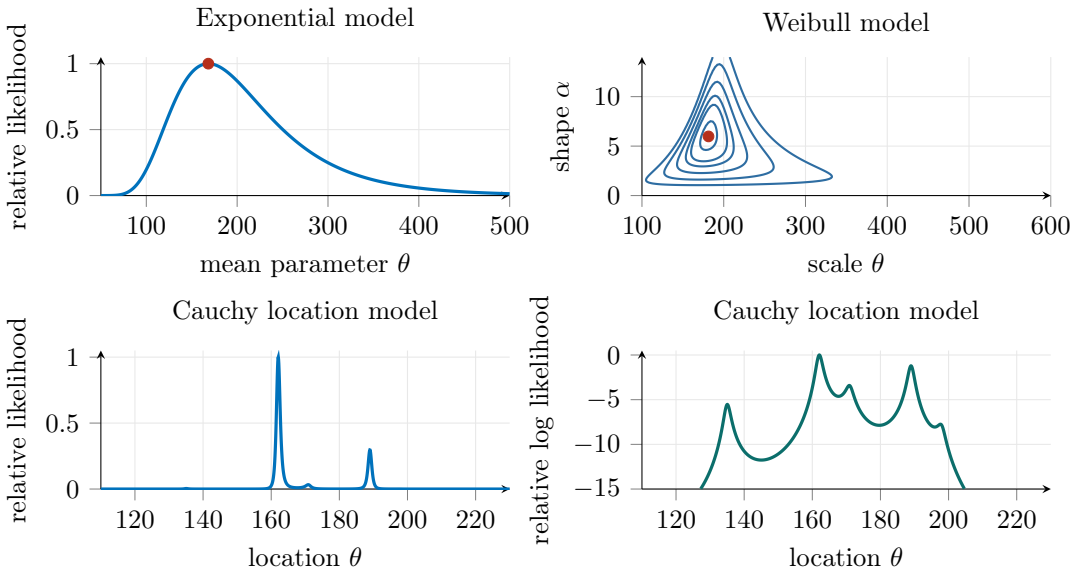

Figure 1

Complicated and Intractable Likelihoods

Models with latent variables or spatial dependence often have likelihoods that are mathematically explicit but difficult to evaluate.

Example 1.7 For cluster i and observation r, consider the latent variable model

$$Z_{ir} = \boldsymbol{x}_{ir}^\top \boldsymbol{\beta} + b_i + \varepsilon_{ir}, \qquad b_i \sim \mathcal{N}(0, \sigma_b^2), \qquad \text{and} \qquad \varepsilon_{ir} \sim \mathcal{N}(0, 1).$$

Only the binary response

$$Y_{ir} = \mathbb{1}\{Z_{ir} > 0\}$$

is observed. The random intercept b_i is shared by all observations in a family, school, hospital, or other cluster, and therefore induces within-cluster dependence. Conditional on b_i,

$$p_{ir}(b_i) := \mathbb{P}(Y_{ir} = 1 \mid b_i) = \Phi(\boldsymbol{x}_{ir}^\top \boldsymbol{\beta} + b_i),$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

The conditional responses are independent, but the random effect is not observed and must be integrated out. The likelihood is therefore

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{r=1}^{n_i} p_{ir}(b_i)^{y_{ir}} \left[1 - p_{ir}(b_i)\right]^{1-y_{ir}} \right\} \phi(b_i; 0, \sigma_b^2) \, \mathrm{d}b_i.$$

Each cluster contributes a one-dimensional integral. The more general specification

$$Z_{ir} = \boldsymbol{x}_{ir}^\top \boldsymbol{\beta} + \boldsymbol{w}_{ir}^\top \boldsymbol{b}_i + \varepsilon_{ir}$$

allows random slopes and requires a multidimensional integral for each cluster. This model motivates the pairwise likelihood calculations of Renard, Molenberghs, and Geys.

Example 1.8 Let $Y(s)$ be observed over a spatial region $\mathcal{S} \subseteq \mathbb{R}^d$, where $d \geq 2$, and suppose that

$$\mathbb{E}[Y(s) \mid U(s)] = g(\boldsymbol{x}(s)^\top \boldsymbol{\beta} + U(s)).$$

Here, U is a stationary Gaussian random field with mean zero and covariance

$$\mathrm{Cov}(U(s), U(s')) = \sigma^2 \rho(s - s'; \alpha).$$

For observations $y_i = Y(s_i)$ at n locations, let $\boldsymbol{u} = (u_1, \ldots, u_n)$ and let $\boldsymbol{\Sigma}$ denote the covariance matrix of the latent field at those locations. The likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^n} \left\{ \prod_{i=1}^{n} f(y_i \mid u_i; \boldsymbol{\theta}) \right\} \phi_n(\boldsymbol{u}; \boldsymbol{0}, \boldsymbol{\Sigma}) \, \mathrm{d}u_1 \cdots \mathrm{d}u_n.$$

Unlike the likelihood in Example 1.7, this integral does not factor into independent low-dimensional pieces. Its dimension is the number of observed spatial locations. See Diggle and Ribeiro's *Model-based Geostatistics* for a systematic treatment.

Some likelihoods contain a normalizing constant whose definition is simple but whose computation

is prohibitive.

Example 1.9 Let $G = (V, E)$ be a graph with n nodes and binary node values $y_i \in \{-1, 1\}$. An **Ising model** has probability mass function

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{(i,j)\in E} \theta_{ij} y_i y_j \right),$$

where θ_{ij} measures the interaction between adjacent nodes. The **partition function** is

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in\{-1,1\}^n} \exp\left(\sum_{(i,j)\in E} \theta_{ij} y_i y_j \right).$$

The sum has 2^n terms, so direct evaluation is generally infeasible for a large graph. This computational obstruction motivates pseudolikelihood and other approximations; one high-dimensional approach is developed by Ravikumar, Wainwright, and Lafferty.

The graph in Figure 2 illustrates the local interaction structure of an **Ising model:** only node pairs joined by an edge contribute to the exponent.

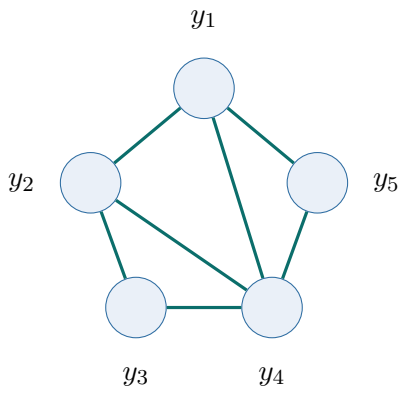

Figure 2

The three examples identify distinct sources of difficulty. Clustered probit data require many low-dimensional integrals, spatial generalized linear models require a single high-dimensional integral, and an Ising model requires a sum over an exponentially large state space. The various likelihood analogues introduced earlier are designed to preserve useful inferential structure while avoiding one of these obstacles.

Likelihood as a Basis for Inference

The term *likelihood* was introduced by R. A. Fisher in his 1922 paper on the mathematical foundations of theoretical statistics. Fisher distinguished the likelihood of one parameter value from its probability: likelihood compares the relative frequencies with which different hypothetical parameter values would have produced the observed sample. It is therefore not a probability distribution over the parameter. This distinction was central to Fisher's attempt to formulate a theory of statistical evidence that did not require assigning prior probabilities to unknown constants.

Definition 1.10 Let $\hat{\boldsymbol{\theta}}$ maximize $L(\boldsymbol{\theta})$. The **relative likelihood** is

$$R_L(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}' \in \Theta} L(\boldsymbol{\theta}')} = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}.$$

It takes values in $[0, 1]$ and is unchanged by multiplication of the likelihood by a positive constant that depends only on the data.

The reciprocal ratio $L(\hat{\boldsymbol{\theta}})/L(\boldsymbol{\theta})$ measures the evidence against a proposed parameter value. The likelihood paradigm in Royall's *Statistical Evidence* uses ratios below 3 as a rough indication of strong plausibility, ratios from 3 to 10 as evidence against the proposed value, and ratios above 10 as stronger evidence against it. A finer descriptive scale for Definition 1.10 regards relative likelihood above $1/3$ as strongly supported, between $1/10$ and $1/3$ as supported, between $1/100$ and $1/10$ as weakly supported, between $1/1000$ and $1/100$ as poorly supported, and below $1/1000$ as very poorly supported. These thresholds are interpretive conventions rather than confidence levels from sampling theory.

Likelihood also supplies the part of Bayesian updating that depends on the data. If $\pi(\boldsymbol{\theta})$ is a prior density, then

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} f(\boldsymbol{y}; \boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})\, \mathrm{d}\boldsymbol{\vartheta}}.$$

This posterior permits probability statements about the parameter, such as $\mathbb{P}(\theta > 0 \mid \boldsymbol{y}) = 0.23$. By contrast, frequentist uses of the likelihood rely on derived quantities and their distributions under repeated sampling from the model.

Definition 1.11 The **likelihood principle** states that two experiments producing proportional likelihood functions for the same parameter should lead to the same inference about that parameter.

Bayesian updating respects Definition 1.11 directly because the data enter through the likelihood. Procedures based on sampling distributions require separate examination. Further reasons for working with likelihood include the following: the likelihood map is sufficient; in transformation models the likelihood can act as an exact pivotal quantity, as developed by Hinkley; its standard summary statistics have known limiting distributions; and likelihood combined with derivatives on the sample space can yield higher-order approximations. The approximate theory begins with the score and information.

Score and Information

For a single observation and a vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$, the **score vector** is

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}}.$$

The **observed information matrix** and **expected information matrix** are, respectively,

$$\mathbf{j}_1(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \qquad \text{and} \qquad \mathbf{i}_1(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})^\top].$$

When the score has mean zero, expected information is also the covariance matrix of the score.

For independent and identically distributed observations,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) \qquad \text{and} \qquad \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}) + a(\boldsymbol{y}).$$

Writing $\boldsymbol{U}_i$ for the score contribution of observation i,

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}) = O_{\mathbb{P}}(\sqrt{n}).$$

This stochastic order applies when the score is evaluated at the data-generating parameter. At a fixed incorrect parameter value, its expectation is generally of order n, so the score itself is generally $O_{\mathbb{P}}(n)$. The **maximum likelihood estimator** is

$$\hat{\boldsymbol{\theta}} = \arg\sup_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \boldsymbol{y}).$$

At the maximum, the observed information has order n, and the expected information for the full

sample is additive:

$$\mathbf{j}_n(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = O_{\mathbb{P}}(n) \qquad \text{and} \qquad \mathbf{i}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{U}_n(\boldsymbol{\theta})\boldsymbol{U}_n(\boldsymbol{\theta})^\top] = n\mathbf{i}_1(\boldsymbol{\theta}).$$

The equality between the two forms of expected information follows from the Bartlett identities.

Proposition 1.12 Suppose that the support of $f(y; \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, differentiation may be passed under the integral sign, and the required moments exist. Then the following identities hold:

1. $$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{U}(\boldsymbol{\theta})] = \mathbf{0}.$$

2. $$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})^\top] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{j}_1(\boldsymbol{\theta})] = \mathbf{i}_1(\boldsymbol{\theta}).$$

Proof. Since f is a density,

$$1 = \int_{\mathcal{Y}} f(y; \boldsymbol{\theta})\, \mathrm{d}y.$$

Differentiating with respect to $\boldsymbol{\theta}$ gives

$$\mathbf{0} = \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\, \mathrm{d}y = \int_{\mathcal{Y}} \frac{\partial \ell(\boldsymbol{\theta}; y)}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\, \mathrm{d}y = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{U}(\boldsymbol{\theta})],$$

which proves (1). Differentiating once more gives

$$\mathbf{0} = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} + \boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})^\top\right].$$

Rearranging proves (2). □

The support condition matters. In the uniform model of Example 1.4, differentiating the expression $f(y; \theta) = 1/\theta$ inside its support gives the apparent score $-1/\theta$, whose expectation is not zero. The missing boundary contribution arises because the upper endpoint depends on θ.

Repeated differentiation yields higher-order Bartlett identities. In the scalar case, define

$$i_{rst} = \mathbb{E}_\theta \left[\left(\frac{\partial \ell}{\partial \theta} \right)^r \left(\frac{\partial^2 \ell}{\partial \theta^2} \right)^s \left(\frac{\partial^3 \ell}{\partial \theta^3} \right)^t \right].$$

The notation is extended by adding positions for fourth and higher derivatives. The first identities are

$$i_{10} = 0, \qquad i_{01} + i_{20} = 0, \qquad i_{001} + 3i_{11} + i_{30} = 0, \qquad \text{and} \qquad i_{0001} + 4i_{101} + 3i_{02} + 6i_{21} + i_{40} = 0.$$

The analogous vector formulas are tensor identities. Their validity requires the same fixed support, differentiability, and moment conditions as Proposition 1.12.

Limiting Distributions and Approximate Pivots

For clarity, first suppose that θ is scalar and that $Y_1, \ldots, Y_n$ are independent and identically distributed. By (1) and (2),

$$U_n(\theta) = \sum_{i=1}^n U_i(\theta), \qquad \mathbb{E}_\theta[U_n(\theta)] = 0, \qquad \text{and} \qquad \mathrm{Var}_\theta(U_n(\theta)) = n i_1(\theta).$$

Provided that $0 < i_1(\theta) < \infty$, the central limit theorem gives

$$\frac{U_n(\theta)}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, i_1(\theta)).$$

Independent and identically distributed observations are sufficient but not necessary. The same conclusion follows whenever a suitable central limit theorem applies to the score sum, including under Lindeberg–Feller conditions or appropriate weak dependence.

Assume that the maximum likelihood estimator is an interior solution of $U_n(\hat{\theta}) = 0$. A Taylor expansion of the score around the true value gives

$$0 = U_n(\hat{\theta}) = U_n(\theta) + (\hat{\theta} - \theta) U_n'(\theta) + R_n.$$

Under the regularity conditions that make the remainder negligible and ensure $-U_n'(\theta)/n \to i_1(\theta)$ in probability,

$$\hat{\theta} - \theta = \frac{U_n(\theta)}{i_n(\theta)} \{1 + o_{\mathbb{P}}(1)\},$$

where $i_n(\theta) = n i_1(\theta)$. Consequently,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, i_1(\theta)^{-1}).$$

Thus maximum likelihood estimators are asymptotically normal. The omitted detail is control of the Taylor remainder: it contains powers of $\hat{\theta} - \theta$ multiplied by higher derivatives of the log likelihood, and the regularity conditions must make it smaller than the leading terms.

A second Taylor expansion, now of the log likelihood around $\hat{\theta}$, gives

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) + R_n.$$

The linear term vanishes because $\ell'(\hat{\theta}) = 0$. Replacing the observed curvature by its probability limit yields

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^2 i_n(\theta)\{1 + o_{\mathbb{P}}(1)\} \xrightarrow{\mathrm{d}} \chi^2_1.$$

For a d-dimensional regular parameter $\boldsymbol{\theta}$, the corresponding results are

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{i}_1(\boldsymbol{\theta})^{-1}) \qquad \text{and} \qquad 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})\} \xrightarrow{\mathrm{d}} \chi^2_d.$$

These limiting distributions lead to two familiar forms of approximate confidence region. The normal approximation gives

$$\hat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}_d(\boldsymbol{\theta}, \mathbf{j}_n(\hat{\boldsymbol{\theta}})^{-1}),$$

while the likelihood ratio region contains values for which the log likelihood has not fallen too far below its maximum.

In a scalar illustration, a Wald interval based on an estimate 21.5 is $(19.5, 23.5)$, whereas the likelihood interval obtained from

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \leq \chi^2_{1,0.95} \approx 3.84$$

is $(18.6, 23.0)$. The comparison in Figure 3 shows the resulting asymmetry: the likelihood interval follows the shape of the likelihood rather than imposing a symmetric normal approximation.

In generalized linear model software, coefficient tables based on estimated standard errors use the Wald approximation, while an analysis of deviance or a comparison of nested models uses the likelihood ratio approximation. Both arise from the central limit theorem applied to the score, but their behaviour in finite samples can differ when the log likelihood is skewed or the model is unusually complex.

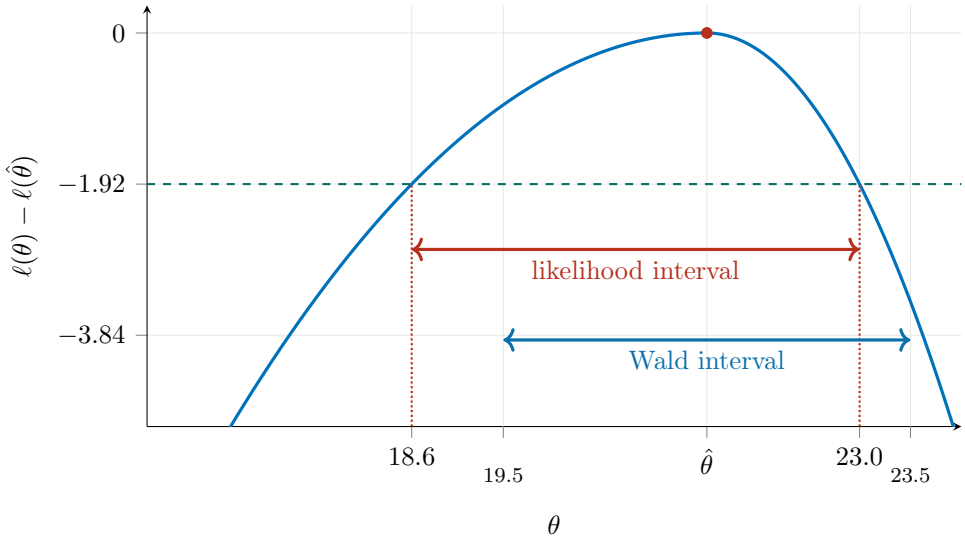

Figure 3

Definition 1.13 A **pivotal quantity** is a function of the data and the parameter whose sampling distribution does not depend on the unknown parameter. An approximate pivotal quantity has a known limiting distribution.

For scalar θ, three likelihood-based roots are

$$r_U(\theta) = U_n(\theta) j_n(\hat{\theta})^{-1/2},$$
$$r_E(\theta) = (\hat{\theta} - \theta) j_n(\hat{\theta})^{1/2} \qquad \text{and} \qquad r(\theta) = \text{sign}(\hat{\theta} - \theta) \left[2\{\ell(\hat{\theta}) - \ell(\theta)\} \right]^{1/2}.$$

Under regularity conditions, each is approximately $\mathcal{N}(0, 1)$ when evaluated at the true parameter. The subscripts distinguish the score root r_U and the estimator or Wald root r_E; the third quantity is the signed likelihood root.

The usual Student t statistic illustrates Definition 1.13: it depends on the sample through the sample mean and sample standard deviation, depends on the unknown mean, and has a known Student t distribution. The likelihood roots play the same role approximately.

For observed data $\boldsymbol{y}^o$, write $r_U^o(\theta)$ for the realized score root. The corresponding **p-value function** is

$$p_U(\theta) = \Phi\{r_U^o(\theta)\},$$

where Φ is the standard normal distribution function. Similarly,

$$p_E(\theta) = \Phi\{r_E^o(\theta)\} \qquad \text{and} \qquad p_R(\theta) = \Phi\{r^o(\theta)\}.$$

For testing $H_0 : \theta = \theta_0$, one evaluates the chosen function at θ_0. An approximate 95% confidence interval consists of parameter values for which the corresponding root lies between -1.96 and 1.96, or equivalently for which the p-value function lies between 0.025 and 0.975. Its intersections with 0.025 and 0.975 give the confidence limits, while an intersection with 0.5 provides a central point estimate.

The two panels in Figure 4 use one observation $y = 1$ from the exponential rate model

$$f(y;\theta) = \theta \exp(-\theta y), \quad y > 0.$$

Here,

$$r_U(\theta) = \frac{1}{\theta} - 1, \qquad r_E(\theta) = 1 - \theta, \qquad \text{and} \qquad r(\theta) = \text{sign}(1-\theta)\{2(\theta - 1 - \log\theta)\}^{1/2}.$$

The three roots agree to first order near the maximum likelihood estimate $\hat{\theta} = 1$, but differ away from it. Their transformed curves in Figure 4 provide p-values for every proposed value of θ from the same observed data.

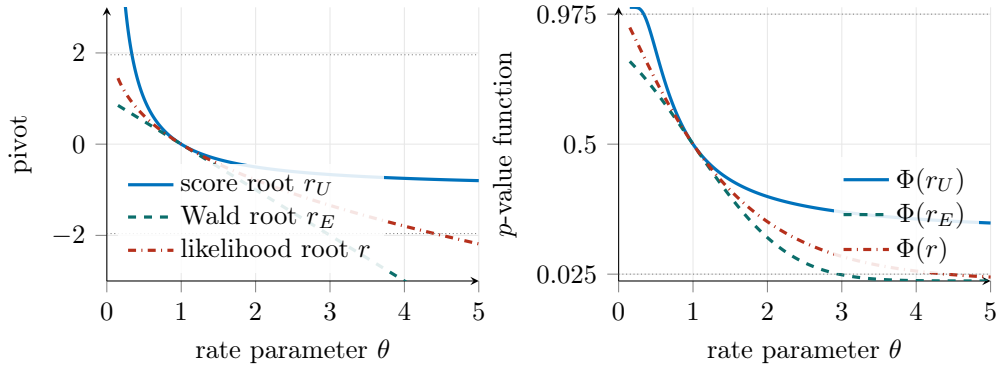

FIGURE 4

When an exact p-value function is available, the quality of an approximation can be checked either as a function of θ for fixed data or as a function of the data for fixed θ_0. If no exact calculation is available, simulation can provide the comparison. For a vector parameter, the usual strategy is to select one component as the parameter of interest and handle the remaining components as

nuisance parameters, leading naturally to profile likelihood.

Lecture 2 — Wednesday, January 19, 2022

2. Likelihood with Nuisance Parameters

Observable Data and Random Effects

A likelihood is proportional to the density of the observed, or at least observable, random variables. This requirement explains why latent random effects are integrated out. Consider first the ordinary Gaussian regression model

$$Y_{ij} = \mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \varepsilon_{ij}, \qquad \text{and} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

with independent errors. The noise is not observed separately, but it directly induces the marginal distribution

$$Y_{ij} \sim \mathcal{N}(\mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}, \sigma^2).$$

There is no visible integral because translating the error distribution immediately gives the density of the response.

Now add a random intercept:

$$Y_{ij} = \mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \qquad \text{and} \qquad b_i \sim \mathcal{N}(0, \sigma_b^2),$$

where the random intercepts and errors are mutually independent. Conditional on b_i, the responses in group i are independent, and hence

$$f(\boldsymbol{y}_i \mid b_i; \boldsymbol{\beta}, \sigma^2) = \prod_{j=1}^{n_i} \phi\big(y_{ij}; \mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i, \sigma^2\big).$$

The random intercept is not observable. The contribution of group i to the likelihood must therefore be its marginal density,

$$f(\boldsymbol{y}_i; \boldsymbol{\beta}, \sigma^2, \sigma_b^2) = \int_{\mathbb{R}} \left\{ \prod_{j=1}^{n_i} \phi\big(y_{ij}; \mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b, \sigma^2\big) \right\} \phi(b; 0, \sigma_b^2)\, \mathrm{d}b.$$

This is the same marginalization principle used in Example 1.7. The Gaussian model is exceptional only because the integral may be evaluated without numerical integration. In particular,

$$\mathbb{E}[Y_{ij}] = \mu + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}, \qquad \text{and} \qquad \operatorname{Var}(Y_{ij}) = \sigma^2 + \sigma_b^2,$$

and, for distinct observations in the same group,

$$\operatorname{Cov}(Y_{ij}, Y_{ij'}) = \sigma_b^2.$$

Thus $\boldsymbol{Y}_i$ is multivariate normal with a known covariance pattern. When the conditional response model is Bernoulli rather than Gaussian, this shortcut disappears and the integral must be retained.

Observed and Expected Information

The asymptotic results of Proposition 1.12 and the subsequent limiting theory admit several asymptotically equivalent standardizations. In the scalar case, for example,

$$\frac{U_n(\theta)}{i_n(\theta)^{1/2}}, \qquad \frac{U_n(\theta)}{i_n(\hat{\theta})^{1/2}}, \qquad \text{and} \qquad \frac{U_n(\theta)}{j_n(\hat{\theta})^{1/2}}$$

all converge in distribution to $\mathcal{N}(0,1)$ under the usual regularity conditions. The equivalence follows because

$$\frac{i_n(\hat{\theta})}{i_n(\theta)} \xrightarrow{\mathbb{P}} 1 \qquad \text{and} \qquad \frac{j_n(\hat{\theta})}{i_n(\theta)} \xrightarrow{\mathbb{P}} 1,$$

so Slutsky's theorem may be applied to the central limit theorem for the score.

First-order asymptotic theory cannot distinguish among these versions, but their performance in finite samples need not be identical. The observed information reflects the curvature of the likelihood for the data actually obtained. The expected information averages that curvature over hypothetical repetitions of the experiment. The analysis in Efron and Hinkley's comparison of observed and expected information gives a frequentist justification for the common preference for $j_n(\hat{\theta})$, particularly when inference is interpreted conditionally on features of the realized sample.

For vector parameters, the same reasoning is matrix-valued. If

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta}}, \qquad \text{and} \qquad \mathbf{j}_n(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

then

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{i}_1(\boldsymbol{\theta})),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{i}_1(\boldsymbol{\theta})^{-1}).$$

The likelihood ratio statistic converges to χ^2_d. These results produce confidence regions for the full vector, but most scientific questions concern only a small part of it.

Parameters of Interest and Nuisance Parameters

Partition the d-dimensional parameter as

$$\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda}), \qquad \boldsymbol{\psi} \in \mathbb{R}^q, \qquad \text{and} \qquad \boldsymbol{\lambda} \in \mathbb{R}^{d-q},$$

where $\boldsymbol{\psi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is the nuisance parameter. The score and information matrices inherit the same partition:

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_\psi(\boldsymbol{\theta}) \\ \boldsymbol{U}_\lambda(\boldsymbol{\theta}) \end{pmatrix}, \qquad \text{and} \qquad \mathbf{i}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{i}_{\psi\psi} & \mathbf{i}_{\psi\lambda} \\ \mathbf{i}_{\lambda\psi} & \mathbf{i}_{\lambda\lambda} \end{pmatrix},$$

and

$$\mathbf{j}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{j}_{\psi\psi} & \mathbf{j}_{\psi\lambda} \\ \mathbf{j}_{\lambda\psi} & \mathbf{j}_{\lambda\lambda} \end{pmatrix}.$$

Definition 2.1 The **efficient information** for $\boldsymbol{\psi}$ in the presence of $\boldsymbol{\lambda}$ is the Schur complement

$$\mathbf{i}_{\psi\psi\cdot\lambda}(\boldsymbol{\theta}) = \mathbf{i}_{\psi\psi} - \mathbf{i}_{\psi\lambda}\mathbf{i}_{\lambda\lambda}^{-1}\mathbf{i}_{\lambda\psi}.$$

If $\mathbf{i}^{\psi\psi}$ denotes the upper-left block of $\mathbf{i}^{-1}$, block matrix inversion gives

$$\mathbf{i}^{\psi\psi} = \mathbf{i}_{\psi\psi\cdot\lambda}^{-1}.$$

Consequently, the marginal limiting distribution of the estimator of the parameter of interest is

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{\text{d}} \mathcal{N}_q\left(\mathbf{0}, \mathbf{i}_{1,\psi\psi\cdot\lambda}(\boldsymbol{\theta})^{-1}\right).$$

The subtraction in Definition 2.1 quantifies the information lost because $\boldsymbol{\lambda}$ is unknown. If the cross information blocks vanish, estimation of the nuisance parameter causes no first-order loss of information about $\boldsymbol{\psi}$.

Profile Likelihood

For each fixed value of $\boldsymbol{\psi}$, maximize the log likelihood over the nuisance parameter.

Definition 2.2 The **constrained maximum likelihood estimator** of $\boldsymbol{\lambda}$ is

$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}} = \arg\sup_{\boldsymbol{\lambda}} \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}).$$

The **profile log likelihood** for $\boldsymbol{\psi}$ is

$$\ell_p(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}).$$

Geometrically, the curve

$$\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$$

passes through the highest point of the likelihood surface above every fixed value of $\boldsymbol{\psi}$. Slicing the surface along this curve produces the profile log likelihood, as illustrated in Figure 5.

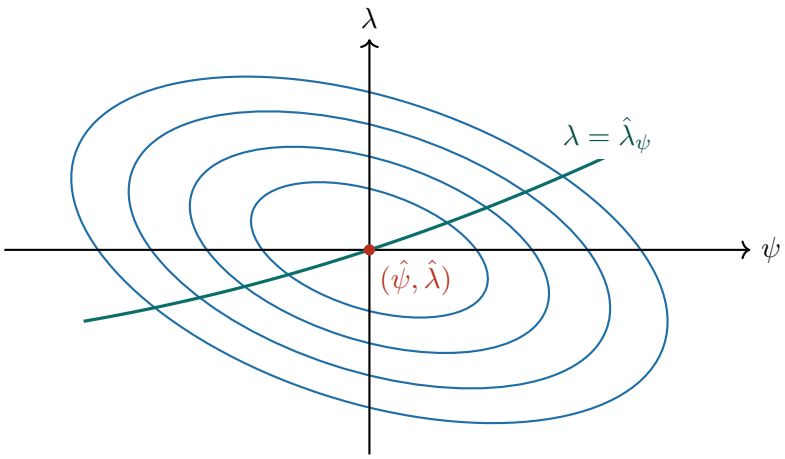

Figure 5

The maximum of the profile log likelihood is the $\boldsymbol{\psi}$ component of the full maximum likelihood estimator. To see this analytically, suppose first that ψ is scalar. By the chain rule,

$$\ell_p'(\psi) = \ell_\psi(\psi, \hat{\boldsymbol{\lambda}}_\psi) + \ell_\lambda(\psi, \hat{\boldsymbol{\lambda}}_\psi)^\top \frac{\mathrm{d}\hat{\boldsymbol{\lambda}}_\psi}{\mathrm{d}\psi} = \ell_\psi(\psi, \hat{\boldsymbol{\lambda}}_\psi),$$

because the constrained estimator satisfies

$$\ell_\lambda(\psi, \hat{\boldsymbol{\lambda}}_\psi) = \mathbf{0}.$$

Thus $\ell_p'(\hat{\psi}) = 0$, together with the constrained score equation, gives the full likelihood equations at

$$\hat{\boldsymbol{\theta}} = (\hat{\psi}, \hat{\boldsymbol{\lambda}}).$$

This argument applies to an interior maximum of a regular model. Boundary examples such as Example 1.4 require direct maximization instead.

Definition 2.3 For scalar ψ, the **observed profile information** is

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

Implicit differentiation of $\ell_\lambda(\psi, \hat{\boldsymbol{\lambda}}_\psi) = \mathbf{0}$ gives

$$\frac{\mathrm{d}\hat{\boldsymbol{\lambda}}_\psi}{\mathrm{d}\psi} = -\ell_{\lambda\lambda}^{-1}\ell_{\lambda\psi},$$

where the derivatives are evaluated at $(\psi, \hat{\boldsymbol{\lambda}}_\psi)$. It follows that

$$j_p(\psi) = j_{\psi\psi} - \mathbf{j}_{\psi\lambda}\mathbf{j}_{\lambda\lambda}^{-1}\mathbf{j}_{\lambda\psi}.$$

The profile information is therefore the observed analogue of the efficient information in Definition 2.1.

The profile likelihood need not be proportional to the density of an observable random variable, so it is not generally a likelihood in the strict sense of Definition 1.1. Nevertheless, it behaves like an ordinary likelihood to first order when the parameter dimension is fixed. For q-dimensional $\boldsymbol{\psi}$, the score, Wald, and profile likelihood ratio statistics may be written as

$$\begin{aligned} w_U(\boldsymbol{\psi}) &= \boldsymbol{U}_\psi(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})^\top \mathbf{i}_{\psi\psi\cdot\lambda}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})^{-1}\boldsymbol{U}_\psi(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}), \\ w_E(\boldsymbol{\psi}) &= (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^\top \left\{\mathbf{i}^{\psi\psi}(\hat{\boldsymbol{\theta}})\right\}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}), \\ w(\boldsymbol{\psi}) &= 2\{\ell_p(\hat{\boldsymbol{\psi}}) - \ell_p(\boldsymbol{\psi})\}. \end{aligned}$$

Under the same regularity conditions used for the full likelihood,

$$w_U(\boldsymbol{\psi}) \xrightarrow{\mathrm{d}} \chi^2_q, \qquad w_E(\boldsymbol{\psi}) \xrightarrow{\mathrm{d}} \chi^2_q, \qquad \text{and} \qquad w(\boldsymbol{\psi}) \xrightarrow{\mathrm{d}} \chi^2_q.$$

When ψ is scalar, the corresponding signed roots have exactly the familiar form

$$\begin{aligned} r_U(\psi) &= \ell_p'(\psi) j_p(\hat{\psi})^{-1/2}, \\ r_E(\psi) &= (\hat{\psi} - \psi) j_p(\hat{\psi})^{1/2}, \\ r_p(\psi) &= \operatorname{sign}(\hat{\psi} - \psi) \left[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \right]^{1/2}. \end{aligned}$$

Each is asymptotically $\mathcal{N}(0,1)$. Hence profile likelihood supports the same first-order p-value functions and confidence intervals as an ordinary scalar likelihood. This is the precise sense in which eliminating a nuisance parameter of fixed dimension creates no first-order inferential difference.

Failure of Ordinary Profiling

First-order validity does not guarantee an accurate finite sample approximation. Ordinary profiling can become too concentrated because maximizing over nuisance parameters fails to account for the uncertainty used to estimate them.

Example 2.4 Consider the Gaussian linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $\mathbf{X}$ has rank p, and take $\psi = \sigma^2$. Apart from a constant that depends only on the data,

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

For every fixed σ^2, the constrained estimator is

$$\hat{\boldsymbol{\beta}}_{\sigma^2} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y},$$

which does not depend on σ^2. If the residual sum of squares is denoted by

$$R = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

then

$$\ell_p(\sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{R}{2\sigma^2},$$

and its maximizer is

$$\hat{\sigma}^2 = \frac{R}{n}.$$

Since $R/\sigma^2 \sim \chi^2_{n-p}$,

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2.$$

The estimate is systematically too small, especially when p is not negligible relative to n.

The usual degrees-of-freedom correction replaces n by $n-p$:

$$\ell_a(\sigma^2) = -\frac{n-p}{2}\log\sigma^2 - \frac{R}{2\sigma^2}, \qquad \text{and} \qquad \hat{\sigma}^2_a = \frac{R}{n-p}.$$

The adjustment both shifts the maximizer and broadens the curve, as Figure 6 illustrates, reflecting the uncertainty consumed in estimating $\boldsymbol{\beta}$.

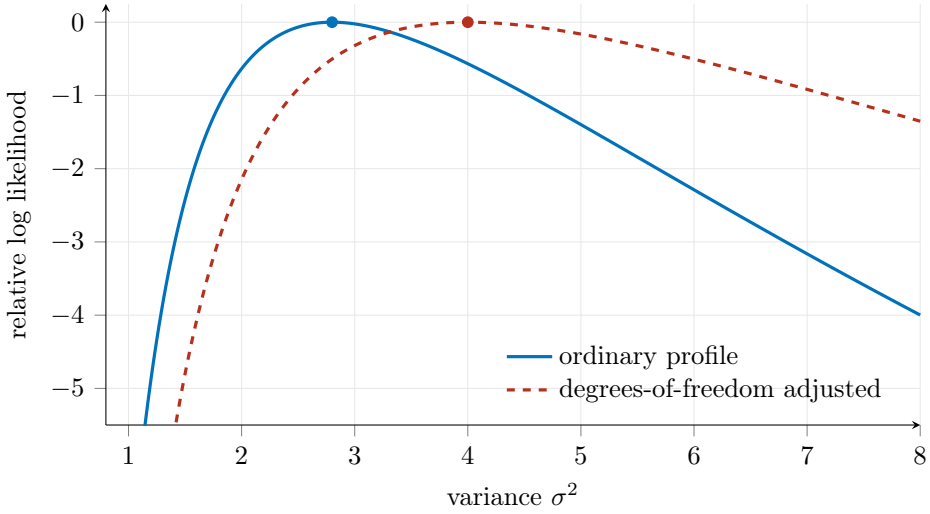

Figure 6

Example 2.5 The **Neyman–Scott model** makes the difficulty persist asymptotically. Suppose that

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \qquad j = 1, \ldots, k, \qquad \text{and} \qquad i = 1, \ldots, m,$$

independently. The variance σ^2 is common to every group, while the m means are nuisance parameters. Profiling gives

$$\hat{\sigma}^2 = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

For fixed k,

$$\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \frac{k-1}{k}\sigma^2 \quad \text{as } m \to \infty.$$

Adding groups supplies more observations, but it also adds one nuisance mean per group. The parameter dimension therefore increases with the sample size, violating a central assumption of likelihood theory for fixed dimensions. When $k = 2$, the probability limit is only $\sigma^2/2$.

A related incidental parameter problem arises from many matched pairs of binary observations. One convenient parametrization is

$$\text{logit}(p_{i1}) = \psi + \lambda_i \qquad \text{and} \qquad \text{logit}(p_{i2}) = \lambda_i,$$

where ψ is a common log odds ratio and each pair has its own nuisance intercept. With only one binary observation under each condition, the number of nuisance parameters grows with the number of pairs. In the parametrization used here, ordinary profiling yields the nonconsistent limit

$$\hat{\psi}_p \xrightarrow{\mathbb{P}} 2\psi$$

rather than ψ. Conditional likelihood will remove the λ_i exactly.

Indeed, write n_{10} and n_{01} for the numbers of discordant pairs of the two possible types. For either type, profiling over its nuisance intercept gives $\hat{\lambda}_i = -\psi/2$. The profiled likelihood is therefore proportional to

$$q(\psi)^{2n_{10}}\{1-q(\psi)\}^{2n_{01}}, \qquad q(\psi) = \frac{\exp(\psi/2)}{1+\exp(\psi/2)},$$

so $\hat{\psi}_p = 2\log(n_{10}/n_{01})$. Since the ratio of the two discordant-pair probabilities is $\exp(\psi)$, the asserted probability limit follows.

These examples distinguish two regimes. If the nuisance dimension is fixed while n increases, profile likelihood has the standard first-order theory. If the nuisance dimension grows with n, the profile estimator can remain biased and the profile likelihood can remain too narrow.

Adjusted Profile Likelihood

Definition 2.6 An **adjusted profile log likelihood** has the form

$$\ell_A(\psi) = \ell_p(\psi) + A(\psi),$$

where $A(\psi) = O_{\mathbb{P}}(1)$. The adjustment leaves the leading $O_{\mathbb{P}}(n)$ term unchanged while correcting lower-order distortion caused by nuisance parameter estimation.

A particularly important construction requires an orthogonal parametrization.

Definition 2.7 The scalar parameter of interest ψ and nuisance parameter $\boldsymbol{\lambda}$ are **orthogonal** if

$$\mathbf{i}_{\psi\lambda}(\psi, \boldsymbol{\lambda}) = \mathbf{0}^\top$$

throughout the parameter space.

Under Definition 2.7, the **Cox–Reid adjustment** to the profile log likelihood is

$$\ell_a(\psi) = \ell_p(\psi) - \frac{1}{2} \log \left| \mathbf{j}_{\lambda\lambda}(\psi, \hat{\boldsymbol{\lambda}}_\psi) \right|.$$

The determinant measures the observed information about the nuisance parameter at the constrained fit. A large change in this information across ψ signals that ordinary profiling has removed different amounts of nuisance parameter uncertainty at different points.

For Example 2.4,

$$\mathbf{j}_{\beta\beta}(\sigma^2) = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}.$$

Therefore,

$$-\frac{1}{2} \log |\mathbf{j}_{\beta\beta}(\sigma^2)| = \frac{p}{2} \log \sigma^2 + \text{constant},$$

and the Cox–Reid adjustment converts the coefficient of $\log \sigma^2$ from $-n/2$ to $-(n-p)/2$. It recovers exactly the degrees-of-freedom-adjusted curve plotted above.

Historical Note (2026) Sir David Cox died on January 18, 2022, one day before this material was presented in class. His contributions ranged from the foundations of likelihood to the proportional hazards model and conditional inference. He had also been Nancy Reid's postdoctoral mentor and a long-standing collaborator. Their joint paper on parameter orthogonality and approximate conditional inference developed the adjustment above and connected it to the modified profile likelihood of Barndorff-Nielsen. The construction is a first systematic repair to ordinary profiling; subsequent conditional and marginal methods explain more directly what information the repair is recovering.

Lecture 3 — Wednesday, January 26, 2022

Efficient Scores as Residuals

The efficient score has a useful geometric interpretation that extends from finite-dimensional nuisance parameters to semiparametric models. Recall that its first-order form is

$$\boldsymbol{U}_{\psi\cdot\lambda} = \boldsymbol{U}_\psi - \mathbf{i}_{\psi\lambda}\mathbf{i}_{\lambda\lambda}^{-1}\boldsymbol{U}_\lambda.$$

The second term is the linear projection of $\boldsymbol{U}_\psi$ onto the span of the nuisance score $\boldsymbol{U}_\lambda$. Thus $\boldsymbol{U}_{\psi\cdot\lambda}$ is the residual after removing the part of the score for $\boldsymbol{\psi}$ that can be explained by changes in $\boldsymbol{\lambda}$. Its covariance is the efficient information from Definition 2.1.

This is analogous to the Frisch–Waugh–Lovell decomposition in linear regression. To estimate one coefficient β_j, first regress the jth covariate on the remaining covariates and retain the residuals. A simple regression of the response on those residuals gives the same coefficient as the full multiple regression. In both settings, the relevant variation is obtained by projecting away the nuisance directions.

Approximate Bayesian Inference

Likelihood theory treats the parameter as fixed and derives sampling distributions for statistics. Bayesian inference instead begins with a prior density $\pi(\boldsymbol{\theta})$ and regards the parameter as random. The posterior density is

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\exp(\ell(\boldsymbol{\theta}; \boldsymbol{y}))\pi(\boldsymbol{\theta})}{\int \exp(\ell(\boldsymbol{t}; \boldsymbol{y}))\pi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}}. \tag{3.1}$$

Once this density is available, posterior means, medians, modes, credible intervals, and probabilities of parameter regions follow by integration.

Suppose first that θ is scalar. A quadratic expansion about the maximum likelihood estimator gives

$$\ell(\theta) = \ell(\hat\theta) - \frac{1}{2}(\theta - \hat\theta)^2 j(\hat\theta) + R_n(\theta),$$

because $\ell'(\hat\theta) = 0$. If the prior is positive and smooth near the true value and the remainder is uniformly negligible over the region where the posterior concentrates, substitution into (3.1) gives

$$\pi(\theta \mid \boldsymbol{y}) \mathrel{\dot\sim} \mathcal{N}(\hat\theta, j(\hat\theta)^{-1}). \tag{3.2}$$

This normal approximation has the same centre and variance as the first-order sampling approximation for the maximum likelihood estimator, but its interpretation is different. In (3.2), the data are fixed and the probability statement concerns θ. In the frequentist approximation, θ is fixed and the probability statement concerns $\hat{\theta}$ under repeated samples.

Theorem 3.1 (Bernstein–von Mises theorem) Under standard regularity and prior positivity conditions, let

$$a_n = \hat{\theta}_n + a\, j(\hat{\theta}_n)^{-1/2} \qquad \text{and} \qquad b_n = \hat{\theta}_n + b\, j(\hat{\theta}_n)^{-1/2},$$

where $a < b$ are fixed. Then

$$\int_{a_n}^{b_n} \pi(\theta \mid \boldsymbol{Y})\, \mathrm{d}\theta \xrightarrow{\mathbb{P}} \Phi(b) - \Phi(a).$$

Equivalently, the posterior distribution of

$$(\theta - \hat{\theta}_n) j(\hat{\theta}_n)^{1/2}$$

converges to the standard normal distribution.

The exact assumptions in Theorem 3.1 matter. Under stronger smoothness, moment, identifiability, and posterior-concentration conditions, Berry–Esseen-type refinements can bound the distance between the standardized posterior distribution function F_n and Φ at order $n^{-1/2}$. Positivity and continuous differentiability of the prior near the true value are not, by themselves, sufficient for that rate. The theorem formalizes the phrase that the data eventually overwhelm a fixed prior.

Let

$$h(\theta) = \ell(\theta) + \log \pi(\theta),$$

and let $\hat{\theta}_\pi$ maximize h. Expanding about $\hat{\theta}_\pi$ gives the equivalent approximation

$$\pi(\theta \mid \boldsymbol{y}) \dot{\sim} \mathcal{N}(\hat{\theta}_\pi, j_\pi(\hat{\theta}_\pi)^{-1}), \qquad \text{and} \qquad j_\pi(\theta) = -h''(\theta).$$

Under the same conditions,

$$\hat{\theta}_\pi - \hat{\theta} = O_{\mathbb{P}}(n^{-1}).$$

The posterior mode and maximum likelihood estimator may therefore differ in a finite sample, but not at the $n^{-1/2}$ scale of first-order inference.

The disappearance of the prior in a theorem for large samples is not an argument that the prior is unimportant in practice. With little data, a genuinely informative prior can substantially improve

inference when it accurately represents previous scientific knowledge. The same influence can be harmful if the prior is poorly justified. A Bayesian analysis is therefore most informative when the origin and sensitivity of the prior are made explicit.

Laplace Approximation

The preceding normal approximation expands both numerator and denominator in (3.1). A more accurate approximation retains the numerator and applies the quadratic expansion only to the denominator. For $\boldsymbol{\theta} \in \mathbb{R}^d$, multivariate Gaussian integration gives

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{|\mathbf{j}(\hat{\boldsymbol{\theta}})|^{1/2}}{(2\pi)^{d/2}} \exp\Big(\ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}})\Big) \frac{\pi(\boldsymbol{\theta})}{\pi(\hat{\boldsymbol{\theta}})} \{1 + O_{\mathbb{P}}(n^{-1})\}. \tag{3.3}$$

This is the **Laplace approximation** to the posterior density. It preserves skewness in the likelihood and retains the prior ratio. Its relative error is typically of order n^{-1}, compared with the order $n^{-1/2}$ error of the elementary normal approximation.

An equivalent expression expands the log posterior h about its mode:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{|\mathbf{j}_\pi(\hat{\boldsymbol{\theta}}_\pi)|^{1/2}}{(2\pi)^{d/2}} \exp\Big(h(\boldsymbol{\theta}) - h(\hat{\boldsymbol{\theta}}_\pi)\Big) \{1 + O_{\mathbb{P}}(n^{-1})\}.$$

Laplace approximation is especially valuable because the quantities in it are usually already available from a likelihood fit: the maximizer, the log likelihood ratio, and the observed information.

Now partition $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, with $\boldsymbol{\psi} \in \mathbb{R}^q$. Integrating the joint posterior over $\boldsymbol{\lambda}$ and applying Laplace approximation at $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ gives

$$\pi_m(\boldsymbol{\psi} \mid \boldsymbol{y}) \doteq \frac{|\mathbf{j}_p(\hat{\boldsymbol{\psi}})|^{1/2}}{(2\pi)^{q/2}} \exp\Big(\ell_p(\boldsymbol{\psi}) - \ell_p(\hat{\boldsymbol{\psi}})\Big) \left\{ \frac{|\mathbf{j}_{\lambda\lambda}(\hat{\boldsymbol{\theta}})|}{|\mathbf{j}_{\lambda\lambda}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})|} \right\}^{1/2} \frac{\pi(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})}{\pi(\hat{\boldsymbol{\theta}})}. \tag{3.4}$$

Consequently, if the prior factors as $\pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\lambda} \mid \boldsymbol{\psi})\pi(\boldsymbol{\psi})$, then, up to a term depending only on the data,

$$\log \pi_m(\boldsymbol{\psi} \mid \boldsymbol{y}) \doteq \ell_p(\boldsymbol{\psi}) - \frac{1}{2} \log |\mathbf{j}_{\lambda\lambda}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})| + \log \pi(\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}} \mid \boldsymbol{\psi}) + \log \pi(\boldsymbol{\psi}). \tag{3.5}$$

The likelihood part of (3.5) is precisely the Cox–Reid adjustment in an orthogonal parametrization. Thus marginal Bayesian calculation and adjusted frequentist profiling arrive at the same correction from different directions.

Marginal and Conditional Likelihood

An alternative to maximizing out nuisance parameters is to find an observable statistic whose distribution does not involve them. Suppose the data can be reduced to $(\boldsymbol{T}_1, \boldsymbol{T}_2)$. A factorization

$$f(\boldsymbol{t}_1, \boldsymbol{t}_2; \boldsymbol{\psi}, \boldsymbol{\lambda}) = f_m(\boldsymbol{t}_1; \boldsymbol{\psi}) f_c(\boldsymbol{t}_2 \mid \boldsymbol{t}_1; \boldsymbol{\psi}, \boldsymbol{\lambda})$$

defines the **marginal likelihood**

$$L_m(\boldsymbol{\psi}) = f_m(\boldsymbol{t}_1; \boldsymbol{\psi}).$$

Conversely, a factorization

$$f(\boldsymbol{t}_1, \boldsymbol{t}_2; \boldsymbol{\psi}, \boldsymbol{\lambda}) = f_c(\boldsymbol{t}_1 \mid \boldsymbol{t}_2; \boldsymbol{\psi}) f_m(\boldsymbol{t}_2; \boldsymbol{\psi}, \boldsymbol{\lambda})$$

defines the **conditional likelihood**

$$L_c(\boldsymbol{\psi}) = f_c(\boldsymbol{t}_1 \mid \boldsymbol{t}_2; \boldsymbol{\psi}).$$

Unlike a profile likelihood, each is the density or mass function of an observable random variable. The Bartlett identities and ordinary likelihood asymptotics therefore apply directly, with the corresponding marginal or conditional information. Information about $\boldsymbol{\psi}$ in the discarded factor may nevertheless be lost.

In the Gaussian linear model of Example 2.4,

$$\frac{R}{\sigma^2} \sim \chi^2_{n-p}$$

and the residual sum of squares R is independent of $\hat{\boldsymbol{\beta}}$. The marginal density of R therefore supplies a genuine likelihood for σ^2, with log likelihood

$$\ell_m(\sigma^2) = -\frac{n-p}{2} \log \sigma^2 - \frac{R}{2\sigma^2} + \text{constant}.$$

It produces the degrees-of-freedom correction directly. Restricted maximum likelihood in linear mixed models generalizes this residual marginal likelihood.

For the matched binary pairs introduced after Example 2.5, condition on the pair total

$$T_i = Y_{i1} + Y_{i2}.$$

Pairs with $T_i = 0$ or $T_i = 2$ carry no conditional information about ψ. For a discordant pair,

$$\mathbb{P}(Y_{i1} = 1 \mid T_i = 1) = \frac{\exp(\psi)}{1 + \exp(\psi)},$$

which is free of λ_i. The product over discordant pairs is an exact conditional likelihood for ψ, and the growing collection of nuisance intercepts disappears without estimation.

Conditional Likelihood in Exponential Families

The preceding cancellation is a general property of exponential families. Suppose

$$f(y_i; \boldsymbol{\psi}, \boldsymbol{\lambda}) = \exp\Big(\boldsymbol{\psi}^\top \boldsymbol{s}(y_i) + \boldsymbol{\lambda}^\top \boldsymbol{t}(y_i) - k(\boldsymbol{\psi}, \boldsymbol{\lambda})\Big) h(y_i).$$

For independent observations, let

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{s}(Y_i) \qquad \text{and} \qquad \boldsymbol{T} = \sum_{i=1}^{n} \boldsymbol{t}(Y_i).$$

The joint density of $(\boldsymbol{S}, \boldsymbol{T})$ has the form

$$f(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\psi}, \boldsymbol{\lambda}) = \exp\Big(\boldsymbol{\psi}^\top \boldsymbol{s} + \boldsymbol{\lambda}^\top \boldsymbol{t} - n k(\boldsymbol{\psi}, \boldsymbol{\lambda})\Big) \tilde{h}(\boldsymbol{s}, \boldsymbol{t}).$$

Dividing by the marginal density of $\boldsymbol{T}$ cancels every factor involving $\boldsymbol{\lambda}$, leaving

$$f(\boldsymbol{s} \mid \boldsymbol{t}; \boldsymbol{\psi}) = \exp\Big(\boldsymbol{\psi}^\top \boldsymbol{s} - n \tilde{k}_{\boldsymbol{t}}(\boldsymbol{\psi})\Big) \tilde{h}_{\boldsymbol{t}}(\boldsymbol{s}). \tag{3.6}$$

Thus the conditional distribution is itself an exponential family with canonical parameter $\boldsymbol{\psi}$.

For example, let

$$Y_i \sim \mathrm{Bin}(m_i, p_i) \qquad \text{and} \qquad \mathrm{logit}(p_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

The joint mass function can be written as

$$f(\boldsymbol{y}; \boldsymbol{\beta}) = \exp\left(\sum_{j=1}^{p} \beta_j S_j - \sum_{i=1}^{n} m_i \log\Big(1 + \exp\Big(\boldsymbol{x}_i^\top \boldsymbol{\beta}\Big)\Big)\right) \prod_{i=1}^{n} \binom{m_i}{y_i},$$

where

$$S_j = \sum_{i=1}^{n} x_{ij} Y_i.$$

If β_5 is the parameter of interest, conditioning on $\boldsymbol{S}_{-(5)}$ removes the remaining regression coeffi-

cients:

$$f_c(S_5 \mid \boldsymbol{S}_{-(5)}; \beta_5) \propto \exp\Big(\beta_5 S_5 - \tilde{k}(\beta_5)\Big) h(S_5).$$

The normalizing sum may be computationally demanding, but the construction is exact. Inference from this conditional likelihood can differ modestly from ordinary logistic regression because it accounts for nuisance estimation through conditioning rather than profiling.

Exact marginal or conditional reductions are uncommon outside special model classes. Laplace approximation then suggests

$$\ell_c(\psi) \doteq \ell_p(\psi) - \frac{1}{2} \log |\mathbf{j}_{\lambda\lambda}(\psi, \hat{\boldsymbol{\lambda}}_\psi)|,$$

under parameter orthogonality. This is again the Cox–Reid adjusted profile log likelihood. More invariant higher-order versions include the Fraser and Barndorff-Nielsen adjustments, which add Jacobian terms describing how the constrained nuisance estimate changes relative to the full estimate. Their common purpose is to approximate the conditional likelihood while respecting changes of nuisance parametrization.

The historical thread behind these ideas continued in Cox's 1972 paper on regression models and life tables. That work introduced a regression method for censored survival times, when comparable regression tools for this data structure were not yet available. Its partial likelihood will provide the central example of eliminating an infinite-dimensional nuisance parameter.

Lecture 4 — Wednesday, February 02, 2022

3. Likelihood beyond Fully Specified Models

Poisson Processes and Function Valued Parameters

A semiparametric model combines a finite-dimensional parameter of interest with an infinite-dimensional nuisance parameter. A Poisson process provides a direct example. Suppose $N = n$ events are observed at

$$0 < Y_1 < \cdots < Y_N < \tau$$

with rate function $\lambda(\cdot)$. The likelihood of the observed event times is

$$L(\lambda(\cdot); \boldsymbol{y}) = \left\{ \prod_{i=1}^{n} \lambda(y_i) \right\} \exp\left(- \int_0^\tau \lambda(u)\, \mathrm{d}u \right). \tag{4.1}$$

The product accounts for events at the observed times; the exponential factor accounts for the intervals in which no event occurred. Hence

$$\ell(\lambda(\cdot); \boldsymbol{y}) = \sum_{i=1}^{n} \log \lambda(y_i) - \int_0^\tau \lambda(u)\, \mathrm{d}u.$$

For a spatial Poisson process, the same expression holds with integration over the spatial observation region. The rate function can be left unrestricted or given a finite-dimensional form such as $\lambda(y_1, y_2) = 100 \exp(-a y_1)$. Increasing a concentrates events toward smaller y_1 without changing their distribution along the second coordinate.

Censoring and Hazard Functions

Let Y_i^0 be a failure time with distribution function $F(\cdot; \boldsymbol{\theta})$, and let C_i be an independent censoring time with distribution function G. The observable variables are

$$Y_i = \min(Y_i^0, C_i) \qquad \text{and} \qquad D_i = \mathbb{1}\{Y_i = Y_i^0\}.$$

Thus $D_i = 1$ marks an observed failure, whereas $D_i = 0$ marks a right-censored observation. The joint density is

$$f(y_i, d_i; \boldsymbol{\theta}) = \{ f(y_i; \boldsymbol{\theta})[1 - G(y_i)] \}^{d_i} \{ [1 - F(y_i; \boldsymbol{\theta})] g(y_i) \}^{1 - d_i}.$$

When the censoring mechanism is independent and contains no parameter of interest, its factors may be omitted from the likelihood for $\boldsymbol{\theta}$.

Define the hazard and cumulative hazard by

$$\lambda(y; \boldsymbol{\theta}) = \frac{f(y; \boldsymbol{\theta})}{1 - F(y; \boldsymbol{\theta})} \qquad \text{and} \qquad \Lambda(y; \boldsymbol{\theta}) = \int_0^y \lambda(u; \boldsymbol{\theta})\, \mathrm{d}u = -\log(1 - F(y; \boldsymbol{\theta})).$$

Apart from terms involving only G, the log likelihood becomes

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{ d_i \log \lambda(y_i; \boldsymbol{\theta}) - \Lambda(y_i; \boldsymbol{\theta}) \}. \tag{4.2}$$

Censored observations contribute through survival to their censoring times even though their exact

failure times are unknown.

Proportional Hazards and Partial Likelihood

The **proportional hazards model** specifies

$$\lambda(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \lambda_0) = \lambda_0(y) \exp\Big(\boldsymbol{x}^\top \boldsymbol{\beta}\Big), \tag{4.3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the parameter of interest and the baseline hazard $\lambda_0(\cdot)$ is unrestricted. Substitution into (4.2) gives

$$\ell(\boldsymbol{\beta}, \lambda_0) = \sum_{i=1}^{n} \Big[d_i \{ \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log \lambda_0(y_i) \} - \Lambda_0(y_i) \exp\Big(\boldsymbol{x}_i^\top \boldsymbol{\beta}\Big) \Big].$$

Suppose for simplicity that the observed failure times are distinct. At a failure time y_i, let

$$R_i = \{ j : y_j \geq y_i \}$$

be the risk set immediately before the failure. Conditional on one member of R_i failing at y_i, the probability that it is individual i is

$$\frac{\exp\big(\boldsymbol{x}_i^\top \boldsymbol{\beta}\big)}{\sum_{j \in R_i} \exp\Big(\boldsymbol{x}_j^\top \boldsymbol{\beta}\Big)}.$$

Multiplying these contributions produces Cox's partial likelihood,

$$L_{\text{part}}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp\big(\boldsymbol{x}_i^\top \boldsymbol{\beta}\big)}{\sum_{j \in R_i} \exp\Big(\boldsymbol{x}_j^\top \boldsymbol{\beta}\Big)} \right\}^{d_i}. \tag{4.4}$$

The unspecified baseline hazard cancels. The construction in Figure 7 compares who failed with everyone who could have failed at that instant, rather than modelling the absolute chance of a failure at that time.

Write

$$A_i(\boldsymbol{\beta}) = \sum_{j \in R_i} \exp\Big(\boldsymbol{x}_j^\top \boldsymbol{\beta}\Big).$$

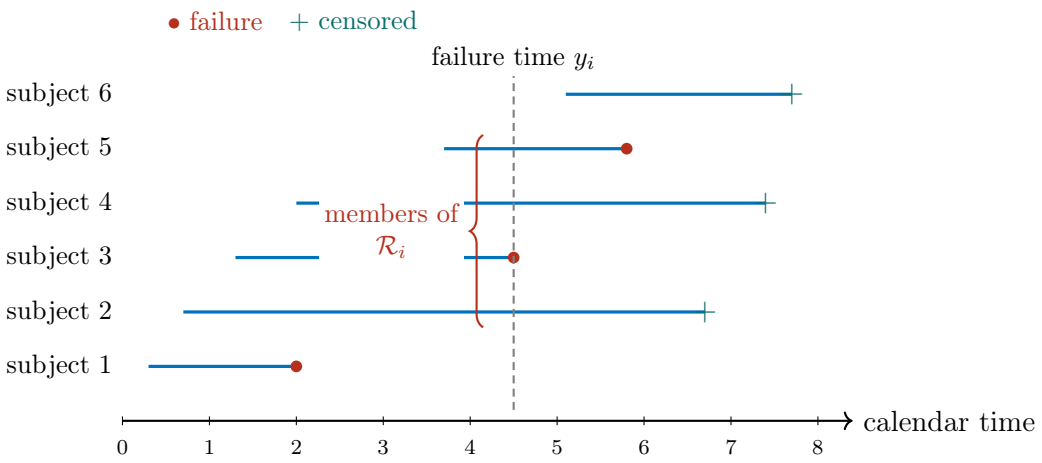

Figure 7

The partial log likelihood, score, and observed information are

$$\ell_{\text{part}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \{\boldsymbol{x}_i^\top \boldsymbol{\beta} - \log A_i(\boldsymbol{\beta})\}, \tag{4.5}$$

$$\boldsymbol{U}_{\text{part}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \left\{ \boldsymbol{x}_i - \frac{\boldsymbol{A}_i'(\boldsymbol{\beta})}{A_i(\boldsymbol{\beta})} \right\}, \tag{4.6}$$

$$\mathbf{j}_{\text{part}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \left\{ \frac{\mathbf{A}_i''(\boldsymbol{\beta})}{A_i(\boldsymbol{\beta})} - \frac{\boldsymbol{A}_i'(\boldsymbol{\beta}) \boldsymbol{A}_i'(\boldsymbol{\beta})^\top}{A_i(\boldsymbol{\beta})^2} \right\}. \tag{4.7}$$

Although the risk set contributions are dependent, martingale limit theory establishes the usual first-order conclusions:

$$\hat{\boldsymbol{\beta}}_{\text{part}} \dot{\sim} \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{j}_{\text{part}}(\hat{\boldsymbol{\beta}}_{\text{part}})^{-1})$$

and

$$2\{\ell_{\text{part}}(\hat{\boldsymbol{\beta}}_{\text{part}}) - \ell_{\text{part}}(\boldsymbol{\beta})\} \xrightarrow{\text{d}} \chi^2_d.$$

Cox originally called (4.4) a conditional likelihood. It can also be motivated as a marginal likelihood for the ranks of the failure times. The original insight was not immediate: Cox later recalled that the essential conditioning argument occurred to him while he was ill with a fever, after years of thinking about the problem. Later retellings sometimes exaggerated the time it took him to recover the argument after returning to work; the underlying episode nevertheless illustrates how the risk set formulation broke a long-standing impasse.

Semiparametric Profile Likelihood

A third justification for (4.4) views it as a profile likelihood. Consider first uncensored failure times and write the cumulative baseline hazard as a step function with jumps only at the observed failures. Any increase between observed failures lowers the survival factor without increasing a density contribution, so a maximizing cumulative hazard need not jump elsewhere. For fixed $\boldsymbol{\beta}$, maximization over the jump at y_i gives

$$\Delta\hat{\Lambda}_{0,\boldsymbol{\beta}}(y_i) = \left\{ \sum_{j \in R_i} \exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right) \right\}^{-1}. \tag{4.8}$$

Substitution of (4.8) into the full likelihood yields (4.4), up to a factor independent of $\boldsymbol{\beta}$. Censoring can be incorporated with the same risk set structure.

This profiling step replaces an arbitrary function by one jump size for every observed failure. The nuisance dimension therefore increases with the sample size, so ordinary likelihood theory for fixed dimensions does not justify the result. The required extension uses nuisance score directions. If $\boldsymbol{U}_\theta$ is the ordinary score for a finite-dimensional parameter $\boldsymbol{\theta}$, its semiparametric efficient score is

$$\tilde{\boldsymbol{U}}_\theta = \boldsymbol{U}_\theta - \mathrm{Proj}_{\mathcal{T}_g}(\boldsymbol{U}_\theta),$$

where $\mathcal{T}_g$ is the closed linear span of all scores generated by smooth perturbations of the nuisance function g. This is the infinite-dimensional analogue of projecting $\boldsymbol{U}_\psi$ away from $\boldsymbol{U}_\lambda$. The efficient information is

$$\tilde{\mathbf{i}}_1(\boldsymbol{\theta}_0) = \mathrm{Var}(\tilde{\boldsymbol{U}}_{\theta,1}(\boldsymbol{\theta}_0)).$$

Under model-specific regularity conditions, the profile log likelihood admits the local expansion

$$\ell_p(\tilde{\boldsymbol{\theta}}_n) - \ell_p(\boldsymbol{\theta}_0) = (\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \sum_{i=1}^n \tilde{\boldsymbol{U}}_{\theta,i}(\boldsymbol{\theta}_0) - \frac{n}{2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \tilde{\mathbf{i}}_1(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_{\mathbb{P}}(1), \tag{4.9}$$

uniformly over suitable local sequences. Consequently,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \tilde{\mathbf{i}}_1(\boldsymbol{\theta}_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\boldsymbol{U}}_{\theta,i}(\boldsymbol{\theta}_0) + o_{\mathbb{P}}(1),$$

and the standard normal and likelihood ratio limits follow.

The proof proceeds through a **least favourable submodel**: a finite-dimensional path through the

true distribution whose score reproduces the efficient score. This path turns the difficult nuisance directions into a locally finite-dimensional likelihood problem. The result is not automatic for every semiparametric model; the required approximation and regularity conditions must be verified model by model. The development by Murphy and van der Vaart includes proportional hazards, current-status, frailty, and partially missing-data examples.

Empirical Likelihood

For a probability measure $\mathbb{P}$ and observations $y_1, \ldots, y_n$, define

$$\mathcal{L}_{\text{emp}}(\mathbb{P}; \boldsymbol{y}) = \prod_{i=1}^{n} \mathbb{P}(\{y_i\}). \tag{4.10}$$

This is a likelihood on the masses assigned to the observed singleton sets; it is not the product of density values. If $\mathbb{P}$ ranges over all probability measures and the observations are distinct, write

$$p_i = \mathbb{P}(\{y_i\}), \qquad p_i \geq 0, \qquad \text{and} \qquad \sum_{i=1}^{n} p_i = 1.$$

The arithmetic–geometric mean inequality shows that (4.10) is maximized at

$$\hat{p}_i = \frac{1}{n}, \quad i = 1, \ldots, n.$$

The empirical distribution

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}$$

is therefore the nonparametric maximum likelihood estimator in this sense. Equivalently, for an arbitrary distribution function F, the likelihood is the product of its jumps

$$L(F) = \prod_{i=1}^{n} \{F(y_i) - F(y_i-)\}.$$

A continuous F assigns zero mass to each observed singleton and hence has zero empirical likelihood.

To make inference about a functional $T(F) = \theta$, profile over all probability masses satisfying the constraint:

$$R(\theta) = \sup_{F: T(F) = \theta} \frac{L(F)}{L(F_n)}. \tag{4.11}$$

For the mean functional, (4.11) becomes

$$R(\theta) = \max\left\{\prod_{i=1}^{n} np_i : p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i y_i = \theta\right\}.$$

Lagrange multipliers give

$$\hat{p}_i(\theta) = \frac{1}{n\{1+\alpha(y_i-\theta)\}}, \qquad \text{and} \qquad \sum_{i=1}^{n} \frac{y_i-\theta}{1+\alpha(y_i-\theta)} = 0. \tag{4.12}$$

These positive weights exist when θ lies in the interior of the convex hull of the observations; the multiplier must also satisfy $1+\alpha(y_i-\theta) > 0$ for every i. If the observations are independent and identically distributed with mean θ_0 and variance in $(0,\infty)$, Owen's nonparametric Wilks theorem states that

$$-2\log R(\theta_0) \xrightarrow{\mathrm{d}} \chi^2_1. \tag{4.13}$$

The result is striking because n masses have been profiled out. Its special constraint geometry restores a likelihood ratio limit that does not hold for arbitrary nuisance problems whose dimension grows. This theory originated in Owen's 1988 empirical likelihood construction.

Maximizers need not exist in every infinite-dimensional model. For example, in the semiparametric logistic model

$$\mathbb{P}(Y=1 \mid V, W) = \frac{\exp(\theta V + \eta(W))}{1+\exp(\theta V+\eta(W))},$$

an unrestricted η can interpolate the binary responses by tending to positive or negative infinity at each observed W_i. A roughness penalty such as

$$\ell(\theta,\eta) - \alpha_n^2 \int \{\eta^{(k)}(w)\}^2\,\mathrm{d}w$$

prevents this degeneration, but the resulting penalized estimator requires a separate analysis.

Composite Likelihood

Suppose $\boldsymbol{Y} \in \mathbb{R}^m$ has joint density $f(\boldsymbol{y};\boldsymbol{\theta})$. Instead of evaluating the full density, select observable events A_k and nonnegative weights w_k.

Definition 4.1 The **composite log likelihood** is

$$c\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{k \in \mathcal{K}} w_k \ell_k(\boldsymbol{\theta}; \boldsymbol{y}),$$

where each ℓ_k is the log likelihood for the event A_k.

Common choices include

$$\begin{aligned}
c\ell_{\text{ind}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{r=1}^{m} w_r \log f_r(y_r; \boldsymbol{\theta}), \\
c\ell_{\text{pair}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{r<s} w_{rs} \log f_{rs}(y_r, y_s; \boldsymbol{\theta}), \\
c\ell_{\text{cond}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{r=1}^{m} w_r \log f(y_r \mid \boldsymbol{y}_{(-r)}; \boldsymbol{\theta}).
\end{aligned}$$

Time series constructions may condition on the preceding observation, while spatial constructions may condition on neighbouring sites. Each component is a genuine likelihood contribution, but their weighted product is generally not the full likelihood and may count dependent information repeatedly.

Let

$$\boldsymbol{U}_c(\boldsymbol{\theta}) = \frac{\partial c\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and define the sensitivity and variability matrices by

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2 c\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right] \qquad \text{and} \qquad \mathbf{J}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(\boldsymbol{U}_c(\boldsymbol{\theta})).$$

Each component score has expectation zero, so the first Bartlett identity survives. The cross-covariances among overlapping components prevent the second Bartlett identity from holding in general; consequently, $\mathbf{H} \neq \mathbf{J}$.

Definition 4.2 The **Godambe information** is

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{H}(\boldsymbol{\theta})^\top.$$

The name recognizes V. P. Godambe, who spent much of his career at the University of Waterloo and developed the theory of optimal estimating functions. If $\mathbf{H} = \mathbf{J}$, then $\mathbf{G} = \mathbf{H}$, recovering Fisher information.

For independent vector observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, let $\hat{\boldsymbol{\theta}}_c$ maximize the summed composite log likelihood. Taylor expansion of the composite score gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}_d(\mathbf{0}, \mathbf{G}_1(\boldsymbol{\theta})^{-1}).$$

The unadjusted composite likelihood ratio does not generally converge to χ^2_d. Instead,

$$2\{c\ell(\hat{\boldsymbol{\theta}}_c) - c\ell(\boldsymbol{\theta})\} \xrightarrow{\mathrm{d}} \sum_{a=1}^{d} \mu_a Z_a^2, \tag{4.14}$$

where the Z_a are independent standard normal random variables and the μ_a are the eigenvalues of $\mathbf{J}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})^{-1}$. For scalar θ, this reduces to

$$J(\theta)H(\theta)^{-1}\chi^2_1.$$

The unadjusted curve is often too concentrated because it counts overlapping information more than once. Figure 8 shows how Godambe scaling can restore curvature comparable to that of the full likelihood.

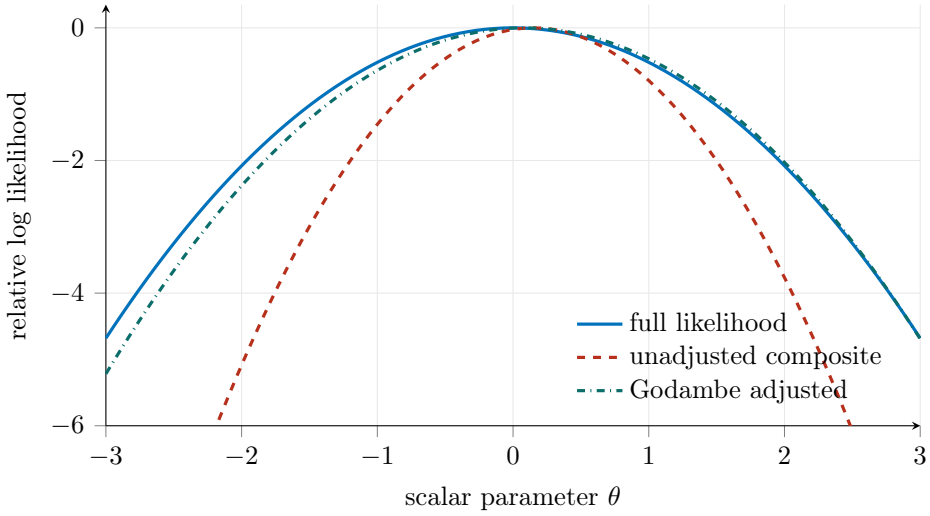

FIGURE 8

Example 4.3 Let

$$\boldsymbol{Y}_i \sim \mathcal{N}_m(\mathbf{0}, \mathbf{R}(\rho)),$$

where every marginal variance equals one and every pairwise correlation equals ρ. A pairwise composite likelihood multiplies the $\binom{m}{2}$ bivariate normal marginal densities for each

observation. If

$$S_{\mathrm{W}} = \sum_{i=1}^{n}\sum_{s=1}^{m}(Y_{is} - \bar{Y}_{i\cdot})^2 \qquad \text{and} \qquad S_{\mathrm{B}} = \sum_{i=1}^{n} Y_{i\cdot}^2,$$

then, up to an additive constant,

$$\begin{aligned} c\ell(\rho) = &- \frac{nm(m-1)}{4}\log(1-\rho^2) - \frac{m-1+\rho}{2(1-\rho^2)}S_{\mathrm{W}} \\ &- \frac{(m-1)(1-\rho)}{2m(1-\rho^2)}S_{\mathrm{B}}, \end{aligned}$$

whereas the full log likelihood is

$$\ell(\rho) = -\frac{n(m-1)}{2}\log(1-\rho) - \frac{n}{2}\log(1+(m-1)\rho) - \frac{S_{\mathrm{W}}}{2(1-\rho)} - \frac{S_{\mathrm{B}}}{2m\{1+(m-1)\rho\}}.$$

The full maximum likelihood estimator is more efficient because the pairwise construction does not recover all of the joint dependence. In this model, the relative efficiency remains fairly high for moderate m, but its exact value depends on both m and ρ. More components of a fixed number of vectors do not force the pairwise estimator's variance to zero; consistency requires the number n of independent vectors to increase.

Composite likelihood was formalized by Lindsay, and the symmetric-normal comparison was developed further by Cox and Reid in their work on pseudolikelihood from marginal densities. The construction trades some efficiency for tractability when a full high-dimensional dependence model is unavailable or computationally prohibitive.

Lecture 5 — Wednesday, February 09, 2022

Likelihood under Model Misspecification

Suppose $Y_1, \ldots, Y_n$ are independent observations from a true density g, but inference uses a parametric family $\{f(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ that need not contain g. The Kullback–Leibler divergence from g to the fitted density is

$$\mathcal{K}(\boldsymbol{\theta}) = \int \log\left(\frac{g(y)}{f(y;\boldsymbol{\theta})}\right) g(y)\,\mathrm{d}y. \tag{5.1}$$

It is nonnegative and vanishes exactly when the two densities agree almost everywhere. Define the **least false parameter**

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{K}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_G[\ell(\boldsymbol{\theta}; Y)]. \tag{5.2}$$

Under suitable compactness, identifiability, and uniform convergence conditions, the maximum likelihood estimator from the working family satisfies

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*.$$

If $g = f(\,\cdot\,; \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0$, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. Otherwise, the estimator converges to the member of the working family closest to the truth in the sense of (5.1).

Example 5.1 Suppose the true distribution is lognormal,

$$\log Y \sim \mathcal{N}(\mu, \sigma^2),$$

but an exponential density with mean θ is fitted:

$$f(y; \theta) = \frac{1}{\theta} \exp(-y/\theta), \quad y > 0.$$

Then

$$\mathbb{E}_G[\ell(\theta; Y)] = -\log\theta - \frac{\mathbb{E}_G[Y]}{\theta},$$

so

$$\theta^* = \mathbb{E}_G[Y] = \exp\bigl(\mu + \sigma^2/2\bigr).$$

The working model maximum likelihood estimator is $\hat{\theta} = \bar{Y}$, which converges to this value. Although the exponential shape is wrong, its mean parameter still has a clear interpretation under the true model.

Let $\boldsymbol{U}(\boldsymbol{\theta}; Y) = \partial \ell(\boldsymbol{\theta}; Y)/\partial \boldsymbol{\theta}$. Differentiating the criterion in (5.2) shows that

$$\mathbb{E}_G[\boldsymbol{U}(\boldsymbol{\theta}^*; Y)] = \mathbf{0}.$$

Define

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}_G\left[-\frac{\partial^2 \ell(\boldsymbol{\theta}; Y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right], \tag{5.3}$$

$$\mathbf{J}(\boldsymbol{\theta}) = \mathrm{Var}_G(\boldsymbol{U}(\boldsymbol{\theta}; Y)). \tag{5.4}$$

Because expectation is taken under g, not generally under $f(\,\cdot\,; \boldsymbol{\theta})$, the second Bartlett identity

need not hold and $\mathbf{H} \neq \mathbf{J}$. Taylor expansion of the score equation and the central limit theorem yield

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathcal{N}_d\left(\mathbf{0}, \mathbf{H}_1^{-1}\mathbf{J}_1\mathbf{H}_1^{-\top}\right). \tag{5.5}$$

where the matrices are evaluated at $\boldsymbol{\theta}^*$. This is the sandwich covariance, or equivalently the inverse Godambe information.

Example 5.2 Suppose an independent sample from G is analysed using a normal working model with parameters (μ, σ^2). The least false parameters are the true mean and variance under G, and the working model estimators are

$$\hat{\mu} = \bar{Y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Let

$$\mu_r = \mathbb{E}_G[(Y - \mu)^r].$$

Then

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathrm{d}} \mathcal{N}_2\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}\right\}.$$

The estimators are asymptotically correlated when G is skewed, and the variance estimator depends on the fourth central moment. Under a normal distribution, $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$, recovering the familiar covariance.

Example 5.3 Treat $\mathbf{X}$ as fixed and suppose the true regression model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \qquad \mathbb{E}_G[\boldsymbol{\varepsilon}] = \mathbf{0}, \qquad \text{and} \qquad \mathrm{Var}_G(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{R},$$

but the working model incorrectly uses $\mathbf{R} = \mathbf{I}$. Ordinary least squares remains unbiased:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{Y} \qquad \text{and} \qquad \mathbb{E}_G[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0.$$

Its true covariance is

$$\mathrm{Var}_G(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{R}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}. \tag{5.6}$$

For a model containing only an intercept and with equicorrelation $R_{ij} = \rho$,

$$\mathrm{Var}_G(\bar{Y}) = \frac{\sigma^2}{n}\{1 + (n-1)\rho\}.$$

Even weak positive dependence can severely reduce the effective sample size. With $n = 100$ and $\rho = 0.1$, the true standard error is approximately 3.3 times the independence model standard error.

The growth of this variance inflation with ρ is shown in Figure 9.

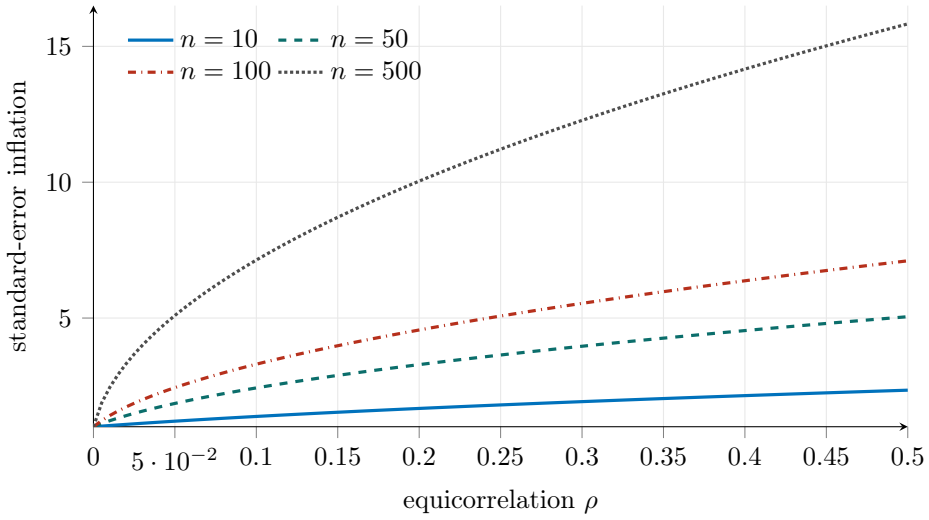

FIGURE 9

These examples distinguish complete misspecification from methods built deliberately around a working model. Composite likelihood replaces the full density by lower-dimensional likelihood components. Generalized estimating equations specify means and a working covariance. Both require sandwich inference because their curvature alone does not measure sampling variability.

Composite Likelihood with Nuisance Parameters

Let $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, and let

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{\psi}} = (\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})$$

maximize the composite log likelihood subject to fixed $\boldsymbol{\psi}$. If $\mathbf{G}^{\psi\psi}$ denotes the upper-left block of $\mathbf{G}^{-1}$, then

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_c - \boldsymbol{\psi}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{G}_1^{\psi\psi}(\boldsymbol{\theta})).$$

The profile composite likelihood ratio

$$w_c(\boldsymbol{\psi}) = 2\{c\ell(\hat{\boldsymbol{\theta}}_c) - c\ell(\tilde{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\}$$

again has a weighted sum of chi-squared limits. The weights are determined by the appropriate parameter-of-interest blocks of the sensitivity and Godambe matrices. Scalar rescaling or more general moment matching can restore a chi-squared reference approximation.

In practice, a direct estimator of the sensitivity is

$$\hat{\mathbf{H}} = -\frac{\partial^2 c\ell(\hat{\boldsymbol{\theta}}_c)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

whereas independent vector observations permit

$$\hat{\mathbf{J}} = \sum_{i=1}^{n} \boldsymbol{U}_{c,i}(\hat{\boldsymbol{\theta}}_c)\boldsymbol{U}_{c,i}(\hat{\boldsymbol{\theta}}_c)^\top.$$

Estimating $\mathbf{J}$ can be unstable because it depends on cross-covariances among all component scores.

The same correction changes model selection penalties. Akaike's criterion uses the parameter dimension because $\mathbf{H} = \mathbf{J}$ under a correctly specified full likelihood. Under misspecification, the effective dimension becomes $\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})$, giving **Takeuchi's criterion**

$$\mathcal{C}_{\text{Takeuchi}} = -2c\ell(\hat{\boldsymbol{\theta}}_c) + 2\,\mathrm{tr}(\hat{\mathbf{H}}\hat{\mathbf{G}}^{-1}). \tag{5.7}$$

A composite analogue of the Bayesian information criterion replaces the factor two by $\log n$, with care required when tuning parameters or nonstandard effective dimensions are involved.

Composite Likelihood Applications

The range of applications is best understood through the obstacle in the full model that each construction avoids:

1. Dichotomized multivariate normal data provide a basic latent variable example. Let

$$Y_{ir} = \mathbb{1}\{Z_{ir} > 0\}, \qquad \text{and} \qquad \boldsymbol{Z}_i \sim \mathcal{N}_m(\mathbf{0}, \mathbf{R}(\rho)).$$

   The full likelihood requires m-dimensional normal orthant probabilities. A pairwise likelihood needs only the four bivariate probabilities $p_{11}, p_{10}, p_{01}, p_{00}$ for each pair. In the equicorrelated model, its asymptotic efficiency relative to full maximum likelihood remains near one for weak dependence and is about 0.85 even at $\rho = 0.98$ in the reported comparison.

2. Multilevel probit regression extends the same construction by adding random effects. Sup-

pose

$$Z_{ir} = \boldsymbol{x}_{ir}^\top \boldsymbol{\beta} + B_i + \varepsilon_{ir}, \qquad B_i \sim \mathcal{N}(0, \sigma_b^2), \qquad \text{and} \qquad \varepsilon_{ir} \sim \mathcal{N}(0,1),$$

and observe only $Y_{ir} = \mathbb{1}\{Z_{ir} > 0\}$. The full likelihood integrates the product of conditional Bernoulli probabilities over every random effect. A pairwise likelihood uses bivariate normal probabilities for (Y_{ir}, Y_{is}). Its computational cost is governed by pairs rather than the dimension of a random effect vector. Simulations showed roughly twenty per cent efficiency loss for regression coefficients and a larger loss for variance components, which are intrinsically more difficult to estimate.

3. Longitudinal count data provide a discrete example. For repeated Poisson counts with correlated gamma random effects,

$$Y_{ir} \mid U_{ir} \sim \text{Poisson}(U_{ir} \exp\Big(\boldsymbol{x}_{ir}^\top \boldsymbol{\beta}\Big)) \qquad \text{and} \qquad \text{Corr}(U_{ir}, U_{is}) = \rho^{|r-s|},$$

the joint density has a combinatorial number of terms. A weighted pairwise likelihood based on $f(Y_{ir}, Y_{is}; \boldsymbol{\beta})$ avoids the full expansion; weights can be chosen so that it agrees with the full likelihood when $\rho = 0$.

4. Ordinal longitudinal responses can also be represented by latent variables. Pain scores recorded on a six-point ordinal scale may be modelled by

$$Y_{ij}^* = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + U_i + \varepsilon_{ij}, \qquad \text{and} \qquad Y_{ij} = k \iff a_{k-1} < Y_{ij}^* < a_k.$$

A random intercept alone gives constant within-subject correlation. Adding a time-decaying error correlation permits nearby observations to be more strongly related. The full likelihood then requires a high-dimensional normal rectangle probability, whereas the pairwise likelihood uses only bivariate rectangles.

5. Spatial extremes create a combinatorial density calculation. If Z_j is the normalized componentwise maximum at location j, a max-stable model has

$$\mathbb{P}(Z_1 \le z_1, \ldots, Z_m \le z_m) = \exp(-V(z_1, \ldots, z_m)).$$

Obtaining the full density requires mixed derivatives indexed by partitions of the locations; the number of terms grows combinatorially. Pairwise composite likelihood needs only bivariate exponent functions. This made likelihood comparison feasible for models of annual maximum rainfall at dozens of Swiss monitoring stations.

6. Ising models create an intractable normalizing constant. For binary or spin-valued nodes on a graph,

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{(j,k) \in E} \theta_{jk} y_j y_k \right),$$

the normalizing constant $Z(\boldsymbol{\theta})$ is often intractable. The conditional distributions of the nodes are logistic and do not involve this constant. Summing their log likelihoods gives a conditional composite likelihood; adding a sparsity penalty to the edge parameters supports graph selection in high dimensions.

These successful examples do not imply that adding components must help. Godambe information can decrease when more component likelihoods are included, and a pairwise likelihood can be less efficient than an independence likelihood. Weights do not always repair the loss. Marginal models must also be compatible with at least one joint distribution, and a parameter identifiable in the full likelihood need not remain identifiable in a composite likelihood. Conditional specifications benefit from the Hammersley–Clifford theory under appropriate positivity conditions; no equally general compatibility theorem exists for arbitrary collections of marginals.

Quasi-Likelihood

For a generalized linear model, suppose

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, \phi) = \exp\left[\frac{y_i\theta_i - c(\theta_i)}{\phi} + h(y_i, \phi)\right].$$

Then

$$\mathbb{E}[Y_i] = \mu_i = c'(\theta_i), \qquad \operatorname{Var}(Y_i) = \phi V(\mu_i) = \phi c''(\theta_i), \qquad \text{and} \qquad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

Standard variance functions include $V(\mu) = 1$ for the normal family, μ^2 for the gamma family, μ for the Poisson family, and $\mu(1-\mu)$ for the Bernoulli family.

The likelihood score for $\boldsymbol{\beta}$ can be written entirely in terms of the mean, variance function, and link:

$$\boldsymbol{g}(\boldsymbol{y}; \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}. \tag{5.8}$$

This motivates an estimator even when no complete response distribution is specified. Assume only

$$\mathbb{E}[Y_i] = \mu_i, \qquad \operatorname{Var}(Y_i) = \phi V(\mu_i), \qquad \text{and} \qquad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

and define $\tilde{\boldsymbol{\beta}}$ by $\boldsymbol{g}(\boldsymbol{y}; \tilde{\boldsymbol{\beta}}) = \mathbf{0}$. The estimating equation is unbiased, and the estimator has the general sandwich covariance

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\top}, \qquad \mathbf{A} = \mathbb{E}\left[-\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}^\top}\right], \qquad \text{and} \qquad \mathbf{B} = \operatorname{Var}(\boldsymbol{g}).$$

For (5.8), $\mathbf{A} = \mathbf{B}$, so the second Bartlett identity is recovered despite the absence of a full density.

Definition 5.4 The **quasi-log-likelihood** is

$$Q(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i(\boldsymbol{\beta})} \frac{y_i - u}{\phi V(u)} \, \mathrm{d}u. \tag{5.9}$$

Its derivative with respect to $\boldsymbol{\beta}$ is the estimating equation (5.8).

Quasi-likelihood permits mean–variance relationships that do not correspond to a familiar exponential family. Its special integrated form is available for independent scalar responses; it does not automatically extend to arbitrary correlated vectors.

Generalized Estimating Equations

Now let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^\top$ collect repeated measurements from subject or cluster i. Specify a marginal mean model

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$$

and a working covariance matrix $\mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$. With

$$\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top},$$

the **generalized estimating equation (GEE)** is

$$\sum_{i=1}^{n} \mathbf{D}_i^\top \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{5.10}$$

The dependence parameter $\boldsymbol{\alpha}$ may be estimated under a convenient working structure such as independence, exchangeable correlation, or first-order autoregression. If the marginal mean is correctly specified and clusters are independent, the estimator of $\boldsymbol{\beta}$ remains consistent even when the working covariance is wrong. Its covariance must then be estimated by the cluster-level sandwich formula.

Unlike scalar quasi-likelihood, there is generally no function whose gradient is exactly (5.10). Liang and Zeger's generalized estimating equations therefore belong most naturally to the broader theory of unbiased estimating functions developed by Godambe. The common goal is to retain reliable inference for scientifically meaningful mean parameters without requiring a complete and correct joint distribution.

Lecture 6 — Wednesday, February 16, 2022

4. High Dimensional Likelihood

Why Dimension Changes the Theory

Classical likelihood asymptotics hold the parameter dimension p fixed while the sample size n tends to infinity. Several earlier constructions violate this description:

1. Profiling the baseline hazard in proportional hazards regression introduces roughly one nuisance jump per observed failure, yet partial likelihood inference for the regression parameter of fixed dimension remains valid.
2. Empirical likelihood assigns one mass to every observation, yet its likelihood ratio for a moment parameter of fixed dimension has a chi-squared limit.
3. Incidental parameter models introduce one or more nuisance parameters per group. If the number of observations per group stays fixed, ordinary profile likelihood can be inconsistent.

Dimension alone therefore does not determine whether classical inference survives. The geometry of the nuisance parameters and the rate at which their dimension grows are decisive.

Why Empirical Likelihood Has a Wilks Limit

For inference about a population mean μ, the constrained empirical likelihood weights in (4.12) satisfy

$$\hat{p}_i(\mu) = \frac{1}{n\{1+\alpha(Y_i-\mu)\}}, \qquad \text{and} \qquad \sum_{i=1}^{n} \frac{Y_i-\mu}{1+\alpha(Y_i-\mu)} = 0.$$

Let

$$S_n(\mu) = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2.$$

At the true mean μ_0, the proof of (4.13) first establishes

$$\alpha = O_{\mathbb{P}}(n^{-1/2})$$

and then expands the constraint to obtain

$$\alpha = \frac{\bar{Y} - \mu_0}{S_n(\mu_0)} + o_{\mathbb{P}}(n^{-1/2}).$$

Because

$$-2\log R(\mu_0) = 2\sum_{i=1}^{n} \log(1 + \alpha(Y_i - \mu_0)),$$

a second Taylor expansion gives

$$-2\log R(\mu_0) = \frac{n(\bar{Y} - \mu_0)^2}{S_n(\mu_0)} + o_{\mathbb{P}}(1) \xrightarrow{\mathrm{d}} \chi^2_1. \tag{6.1}$$

Thus the empirical likelihood ratio is asymptotically equivalent to the square of a studentized mean. The n fitted masses do not behave like an arbitrary collection of nuisance parameters: the normalization and moment constraints force their local effect through a single Lagrange multiplier of order $n^{-1/2}$.

Stratified Incidental Parameters

The general incidental parameter arrangement has observations

$$Y_{ij} \sim f(\,\cdot\,; \psi, \lambda_i), \qquad j = 1, \ldots, m, \qquad \text{and} \qquad i = 1, \ldots, q.$$

The scalar ψ is common to all strata, and each stratum has its own nuisance parameter λ_i. The total sample size is $n = mq$. Three asymptotic regimes are possible:

1. If $m \to \infty$ with fixed q, each nuisance parameter is estimated increasingly well and ordinary theory for fixed dimensions applies.
2. If $q \to \infty$ with fixed m, new nuisance parameters arrive as quickly as new groups. The profile likelihood can fail, as in Example 2.5.
3. If both m and q increase, validity depends on their relative rates.

Other examples of the same structure include a common mean with stratum-specific variances, matched binary pairs with a common log odds ratio, and gamma samples with a common shape but stratum-specific scales.

Example 6.1 Suppose

$$Y_{ij} \sim \text{Gamma}(\psi, \lambda_i), \qquad j = 1, \ldots, m, \qquad \text{and} \qquad i = 1, \ldots, q,$$

where ψ is the common shape and λ_i is a scale parameter. Conditioning on each group total removes the scales exactly. If

$$s = \sum_{i=1}^{q} \sum_{j=1}^{m} \log \left(\frac{Y_{ij}}{\sum_{r=1}^{m} Y_{ir}} \right),$$

then, apart from a data-dependent constant, the exact conditional, ordinary profile, and modified profile log likelihoods are

$$\begin{aligned} \ell_c(\psi) &= \psi s + q \log \Gamma(m\psi) - mq \log \Gamma(\psi), \\ \ell_p(\psi) &= \psi s + qm\psi \log(m\psi) - mq\psi - mq \log \Gamma(\psi), \\ \ell_M(\psi) &= \psi s + q(m\psi - 1/2) \log(m\psi) - mq\psi - mq \log \Gamma(\psi). \end{aligned}$$

The modified profile replaces the troublesome middle term by the leading Stirling approximation to $\log \Gamma(m\psi)$. Although the algebraic change is only of lower order per stratum, it accumulates across q strata and can transform severely distorted profile inference into an approximation very close to the exact conditional result.

For a broad class of stratified models, ordinary profile likelihood may require

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

whereas modified profile likelihood can retain its usual calibration under the weaker condition

$$q = o(n^{3/4}).$$

These rates are equivalent to the conditions on the two indices in Sartori's analysis of stratum nuisance parameters. The improvement is a concrete high-dimensional benefit of the adjustment introduced in Definition 2.6.

Increasing Dimension Asymptotics

Let $p = p_n$ denote the dimension of the fitted parameter. It is useful to distinguish several regimes:

1. In the classical regime, p is fixed as $n \to \infty$.

2. In the increasing but low dimensional regime, $p/n \to 0$, often with a stronger rate such as $p^{3/2}/n \to 0$ for asymptotic normality.
3. In the moderate dimensional regime, $p/n \to \kappa \in (0, 1)$. Classical maximum likelihood estimators may have nonnegligible bias and variance inflation.
4. In the high dimensional regime, p is comparable to or larger than n.
5. In the ultra high dimensional regime, p may grow exponentially with n, so $\log p$ rather than p appears in feasible rates.

The boundary is model dependent. Portnoy's increasing parameter likelihood theory showed that consistency may survive under $p/n \to 0$ while normal and likelihood ratio approximations generally demand stronger restrictions. In moderate dimensional logistic regression, the estimator can require explicit bias and variance corrections even when $p < n$. A finite data set does not reveal which asymptotic regime generated it, so these results are guidance rather than an automatic diagnostic.

For a more explicit result, consider independent observations

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + Z_i$$

and an estimator defined by

$$\sum_{i=1}^{n} \boldsymbol{x}_i \psi(Y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Under monotonicity of ψ, design conditions, and

$$\frac{p \log p}{n} \longrightarrow 0,$$

there is a consistent solution. Stronger conditions such as

$$\frac{p^{3/2} \log n}{n} \longrightarrow 0$$

permit asymptotic normality of suitable contrasts $\boldsymbol{a}_n^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. This contrast formulation is essential: when p changes with n, convergence does not concern an entire parameter vector of fixed dimension.

Penalized Linear Regression when $p > n$

Consider

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \qquad \text{and} \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

with $p > n$. The least squares criterion has many minimizers, so $\boldsymbol{\beta}$ is not identifiable without extra structure. Ridge and lasso estimators impose two standard penalties:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{6.2}$$

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{6.3}$$

The covariate columns must be placed on comparable scales before their coefficients receive a common penalty. An intercept is usually left unpenalized. Ridge shrinks every coefficient continuously, whereas the corners of the ℓ_1 constraint in (6.3) set many coefficients exactly to zero. Cross-validation commonly selects λ.

As the penalty weakens and the total ℓ_1 norm grows, variables enter the lasso solution at different points, as shown in Figure 10. Unlike a ridge path, several coefficient paths remain exactly zero over an initial range.

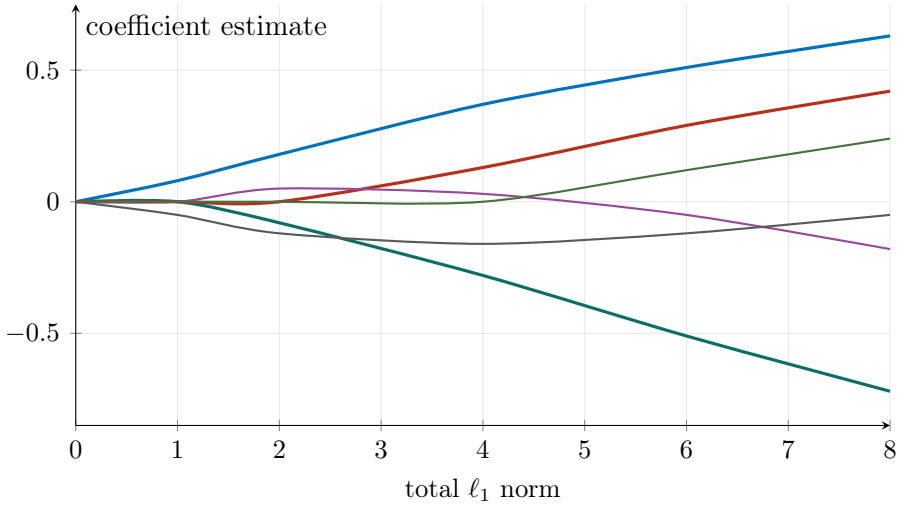

Figure 10

There are three distinct inferential goals:

1. Prediction concerns $\mathbf{X}\boldsymbol{\beta}$ or $\boldsymbol{x}_{\text{new}}^{\top}\boldsymbol{\beta}$. It may be possible even when $\boldsymbol{\beta}$ itself is not identifiable.
2. Coefficient estimation requires additional design conditions, such as restricted eigenvalue or compatibility conditions, to distinguish sparse directions.
3. Support recovery seeks
$$S = \{j : \beta_j \neq 0\}.$$
Exact recovery typically requires a beta-min condition ensuring that every nonzero signal exceeds the noise scale, often of order $\sqrt{\log(p)/n}$, together with design assumptions. The weaker screening goal asks only that $\hat{S}$ contain S with high probability.

Sparsity is the structural assumption that makes these problems estimable. If
$$|S| \ll \frac{n}{\log p},$$
then useful prediction and estimation can be possible even for exponentially large p. Without sparsity or some comparable low complexity structure, high-dimensional coefficient inference is ill posed. The early asymptotic analysis of lasso-type estimators by Knight and Fu established, among other results, that zero coefficients can produce limiting distributions with point mass at zero.

Post Selection Inference

Penalization changes both the centre and distribution of an estimator, so the usual least squares standard errors are invalid after variable selection. Fitting an ordinary regression to variables selected on the same data is also invalid: the reported t-statistics ignore the selection event that made those variables appear promising.

Three broad strategies were developed for this setting.

1. Multisample splitting selects $\hat{S}$ using one random half of the data, fits the selected model on the other half, and assigns
$$P_j = 1$$
to unselected variables. For selected variables, a Bonferroni correction multiplies the p-value from the second half by $|\hat{S}|$. Repeating the split reduces sensitivity to a single random partition, but the resulting corrected p-values are dependent and must be aggregated carefully.

Splitting is conceptually clean but sacrifices information.

2. Debiasing constructs
$$\hat{\beta}_{j,\mathrm{db}} = \hat{\beta}_{j,\mathrm{lasso}} - \widehat{\mathrm{bias}}_j$$
so that, under sparsity and design assumptions,
$$\frac{\hat{\beta}_{j,\mathrm{db}} - \beta_j}{\hat{\sigma}\sqrt{w_j}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$
Debiasing restores approximately normal marginal inference, but it produces a nonzero estimate for every coordinate and therefore requires multiplicity control across p tests.

3. Stability selection uses repeated subsampling to record how often each variable is selected. Variables with high selection probability are retained, with theoretical bounds on expected false selections under additional assumptions.

In a gene expression example with $n = 71$ and $p = 4088$, the methods selected very different numbers of features: ordinary lasso selected about thirty, multisample splitting selected one, debiased ridge selected none, and stability selection selected three. This disagreement is not merely computational; it reflects the difficulty of attaching uncertainty statements to a data-dependent model. The comparison and methods are reviewed in Bühlmann, Kalisch, and Meier's account of high-dimensional inference.

The same principles extend beyond Gaussian linear regression. A generalized linear model may use the penalized likelihood
$$\arg\min_{\beta_0,\boldsymbol{\beta}} \left\{ -\frac{1}{n}\ell(\beta_0, \boldsymbol{\beta}; \boldsymbol{y}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Sample splitting and stability selection apply directly, while debiasing uses a weighted least squares approximation. A marginal alternative fits one feature at a time, but this creates p tests and requires explicit control of the familywise error rate or false discovery rate. High-dimensional inference therefore replaces the single classical likelihood approximation by an interlocking set of structural, computational, and multiplicity assumptions.

Appendix: Lecture and Tutorial Index