22. Additional Topics Related to Likelihood

Information Criteria

Akaike's Information criterion is given by

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2k,$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the maximized log likelihood and k is the dimension of the model parameter space.

- AIC $= -2\ell(\hat{\boldsymbol{\theta}}) + 2k$ can be used to determine which of multiple models is "best" for a given data set.
- Small values of AIC are preferred.
- The $+2k$ portion of AIC can be viewed as a penalty for model complexity.

Schwarz's Bayesian Information Criterion is given by

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + k\ln(n)$$

BIC is the same as AIC except the penalty for model complexity is greater for BIC (when $n \geq 8$) and grows with n.

- AIC and BIC can each be used to compare models even if they are not nested (i.e., even if one is not a special case of the other as in our reduced vs. full model comparison discussed previously).
- However, if REML likelihoods are used, compared models must have the same model for the response mean.
- Different models for the mean would yield different error contrasts and different datasets for computation of maximized REML likelihoods.

Large n Theory for MLEs

- Suppose $\boldsymbol{\theta}$ is a $k \times 1$ parameter vector.
- Let $\ell(\boldsymbol{\theta})$ denote the log likelihood function.
- Under regularity conditions discussed in, e.g., Shao, J.(2003) *Mathematical Statistics*, 2^{nd} Ed. Springer, New York; we have the following.

There is an estimator $\hat{\boldsymbol{\theta}}$ that solves the score equations $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ and has the following properties.

1. Consistency of $\hat{\boldsymbol{\theta}}$:

   $\hat{\boldsymbol{\theta}}$ is a (weakly) consistent estimator of $\boldsymbol{\theta}$.

   This means that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$, i.e.,

   $$\lim_{n \to \infty} Pr[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon] = 0 \text{ for any } \varepsilon > 0.$$

2. Asymptotic normality of $\hat{\boldsymbol{\theta}}$:

   For sufficiently large n, $\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$, where

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right] \\
&= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \\
&= \left[-E\left\{\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\}\right]_{i,j \in \{1,\ldots,k\}}
\end{aligned}$$

- $\boldsymbol{I}(\boldsymbol{\theta})$ is known as the *Fisher Information* matrix.
- $\boldsymbol{I}(\boldsymbol{\theta})$ can be approximated by the replacing the unknown $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$ to obtain $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$.
- An alternative approximation is given by the *observed Fisher Information* matrix:

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) \equiv \left. \frac{-\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

In practice, when n is sufficiently large, we use the approximation

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N(\boldsymbol{\theta}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})),$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ can be either $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})$ or $\hat{\boldsymbol{I}}^{-1}(\hat{\boldsymbol{\theta}})$.

Although such statements do a reasonable job of conveying the idea of approximations we use, they are not mathematically rigorous.

When we say something like

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N(\boldsymbol{\theta}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})) \text{ for sufficiently large } n,$$

we mean that as n grows to infinity,

$$[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

converges in distribution to a standard multivariate normal random vector $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$:

$$[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{z}$$

Note that

$$[\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{z}$$

implies

$$\{[\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}'\{[\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})]^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\} \xrightarrow{d} \boldsymbol{z}'\boldsymbol{z}$$

which implies

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'[\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})]^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{\bullet}{\sim} \chi^2_k$$

for sufficiently large n.

A Simple Example

- Suppose $y_1, \ldots, y_n \overset{i.i.d.}{\sim} \text{Poisson}(\theta)$.
- For $y_i \in \{0, 1, 2, \ldots\}\ \forall\ i = 1, \ldots, n$,

$$
\begin{aligned}
L(\theta|\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\
\ell(\theta|\boldsymbol{y}) &= \sum_{i=1}^{n} [y_i \ln(\theta) - \theta - \ln(y_i!)] \\
&= \ln(\theta) \sum_{i=1}^{n} y_i - n\theta - \sum_{i=1}^{n} \ln(y_i!) \\
\frac{\partial \ell(\theta|\boldsymbol{y})}{\partial \theta} &= \frac{1}{\theta} \sum_{i=1}^{n} y_i - n
\end{aligned}
$$

Thus, the score equation is

$$\frac{1}{\theta}\sum_{i=1}^{n} y_i - n = 0.$$

The only solution to the score equation is $\hat{\theta} = \bar{y}_.$

Result (1) on slide 7 implies $\hat{\theta} = \bar{y}_.$ converges in probability to θ.

In this case, we also know that $\bar{y}_.$ converges in probability to θ by the (Weak) Law of Large Numbers (WLLN).

Result (2) on slide 8 implies $\hat{\theta} = \bar{y}_{\cdot} \overset{\bullet}{\sim} N(\theta, I^{-1}(\theta))$.

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2 \ell(\theta|\boldsymbol{y})}{\partial\theta\partial\theta}\right] = -E\left[\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\sum_{i=1}^{n} y_i - n\right)\right] \\ &= -E\left[-\frac{1}{\theta^2}\sum_{i=1}^{n} y_i\right] = \frac{1}{\theta^2}\sum_{i=1}^{n} E(y_i) = \frac{n}{\theta} \end{aligned}$$

Therefore, $I^{-1}(\theta) = \theta/n$ in this case.

Thus, result (2) on slide 8 implies $\hat{\theta} = \bar{y}_{\cdot} \overset{\bullet}{\sim} N(\theta, \theta/n)$, which is also implied by the Central Limit Theorem (CLT).

To get an estimate of the variance of $\hat{\theta} = \bar{y}_.$, we can use

$$I^{-1}(\hat{\theta}) = \hat{\theta}/n = \bar{y}_./n$$

Alternatively, the inverse of the observed Fisher information in this case is

$$\hat{I}^{-1}(\hat{\theta}) = \left[\frac{-\partial^2 \ell(\theta)}{\partial\theta\partial\theta}\bigg|_{\theta=\hat{\theta}} \right]^{-1} = \left[\frac{1}{\hat{\theta}^2} \sum_{i=1}^{n} y_i \right]^{-1} = \left[\frac{n\bar{y}_.}{\bar{y}_.^2} \right]^{-1} = \bar{y}_./n$$

Thus, $I^{-1}(\hat{\theta}) = \hat{I}^{-1}(\hat{\theta})$ in this case.

Substituting in this consistent estimator for $I^{-1}(\theta)$, we have $\hat{\theta} = \bar{y}_. \overset{\bullet}{\sim} N(\theta, \bar{y}_./n)$

Wald Tests and Confidence Intervals

Suppose for large n that

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N(\boldsymbol{\theta}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})).$$

Then a confidence interval for $\boldsymbol{c}'\boldsymbol{\theta}$ that has confidence level approximately equal to $1 - \alpha$ is

$$\boldsymbol{c}'\hat{\boldsymbol{\theta}} \pm z_{1-\alpha/2}\sqrt{\boldsymbol{c}'\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\boldsymbol{c}},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the $N(0, 1)$ distribution.

Likewise, a test of $H_0 : \boldsymbol{c}'\boldsymbol{\theta} = d$ can be based on the test statistic

$$\frac{\boldsymbol{c}'\hat{\boldsymbol{\theta}} - d}{\sqrt{\boldsymbol{c}'\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\boldsymbol{c}}},$$

which has a distribution that is approximately $N(0, 1)$ under H_0.

Likewise, if $\boldsymbol{C}$ is a $q \times k$ matrix of rank q, a test of $H_0 : \boldsymbol{C\theta} = \boldsymbol{d}$ can be based on the test statistic

$$(\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d})'[\boldsymbol{C}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d}),$$

which has a distribution that is approximately χ^2_q under H_0.

Multivariate Delta Method

- Suppose $\boldsymbol{g}$ is a function from $\mathbb{R}^k$ to $\mathbb{R}^m$, i.e.,

$$\text{for } \boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{g}(\boldsymbol{\theta}) = \begin{bmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \\ \vdots \\ g_m(\boldsymbol{\theta}) \end{bmatrix}$$

  for some functions $g_1, \ldots, g_m$.
- Suppose $\boldsymbol{g}$ is differentiable with derivative matrix

$$\boldsymbol{D} \equiv \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_k} & \cdots & \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix}.$$

Now suppose $\hat{\theta}$ has mean θ and variance $\mathrm{Var}(\hat{\theta})$. Then Taylor's Theorem implies

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

which implies

$$E[\boldsymbol{g}(\hat{\boldsymbol{\theta}})] \approx \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = g(\boldsymbol{\theta})$$

and

$$\mathrm{Var}[\boldsymbol{g}(\hat{\boldsymbol{\theta}})] \approx \mathrm{Var}[\boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{D}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \boldsymbol{D}'\mathrm{Var}(\hat{\boldsymbol{\theta}})\boldsymbol{D}.$$

- If $\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N(\boldsymbol{\theta}, \mathrm{Var}(\hat{\boldsymbol{\theta}}))$, it follows that

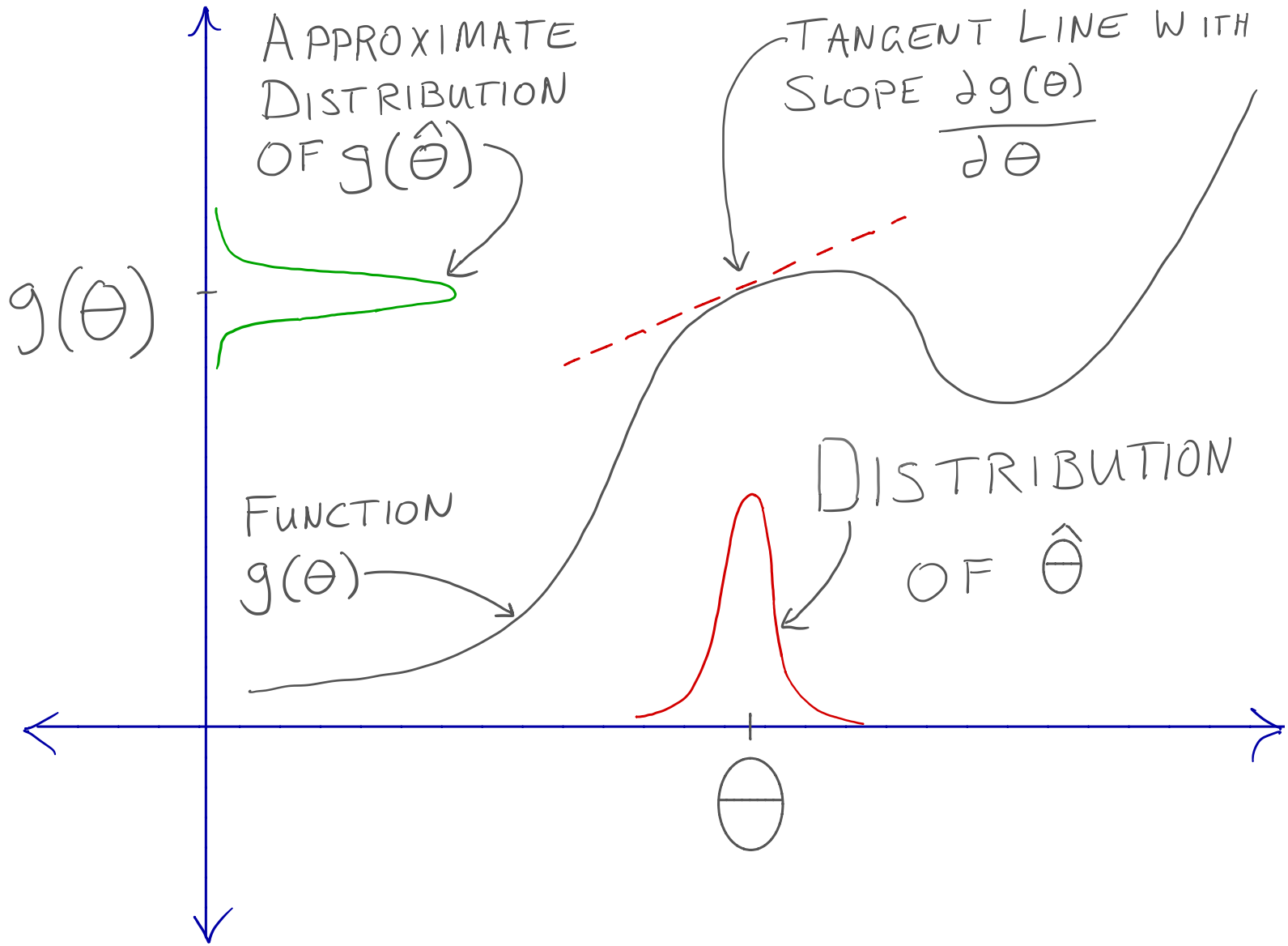
$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \overset{\bullet}{\sim} N(\boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{D}'\mathrm{Var}(\hat{\boldsymbol{\theta}})\boldsymbol{D}).$$

- In practice, we often need to estimate $\boldsymbol{D}$ by replacing $\boldsymbol{\theta}$ in $\boldsymbol{D}$ with $\hat{\boldsymbol{\theta}}$ to obtain $\hat{\boldsymbol{D}}$.

- Similarly, we often need to replace $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ with an estimate $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$.

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \overset{\bullet}{\sim} N(\boldsymbol{g}(\boldsymbol{\theta}), \hat{\boldsymbol{D}}'\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{D}})$$

Delta Method Example with $k = 1$



Likelihood Ratio Based Inference

Suppose we wish to test the null hypothesis that a reduced model provides an adequate fit to a dataset relative to a more general full model that includes the reduced model as a special case.

- Define Λ as

$$\frac{\text{Reduced Model Maximized Likelihood}}{\text{Full Model Maximized Likelihood}}.$$

- Λ is known as the *likelihood ratio*.
- $-2\ln(\Lambda)$ is known as the *likelihood ratio test statistic*.
- Tests based on $-2\ln(\Lambda)$ are called *likelihood ratio tests*.

- Under the regularity conditions in Shao (2003) mentioned previously, the likelihood ratio test statistic $-2\ln(\Lambda)$ is approximately distributed as central $\chi^2_{k_f-k_r}$ under the null hypothesis, where k_f and k_r are the dimensions of the parameter space under the full and reduced models, respectively.

- This approximation can be reasonable if n is "sufficiently large."

Likelihood Ratio Tests and Confidence Regions for a Subvector of the Full Model Parameter Vector θ

- Suppose $\boldsymbol{\theta}$ is $k \times 1$ vector and is partitioned into vectors $\boldsymbol{\theta}_1$ $k_1 \times 1$ and $\boldsymbol{\theta}_2$ $k_2 \times 1$, where $k = k_1 + k_2$ and $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}$.
- Consider a test of $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{d}_1$.

- Suppose $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ maximizes $\ell\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}\right)$ over $\boldsymbol{\theta}_2$ for any fixed value of $\boldsymbol{\theta}_1$.

- Then $2\left[\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\begin{bmatrix} \boldsymbol{d}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{d}_1) \end{bmatrix}\right)\right]$ is approximately $\chi^2_{k_1}$ under the null hypothesis by our previous result when n is "sufficiently large."

Also,

$$Pr\left\{2\left[\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\begin{bmatrix}\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\end{bmatrix}\right)\right] \leq \chi^2_{k_1,1-\alpha}\right\} \approx 1-\alpha$$

which implies

$$Pr\left\{\ell\left(\begin{bmatrix}\boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\end{bmatrix}\right) \geq \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\chi^2_{k_1,1-\alpha}\right\} \approx 1-\alpha.$$

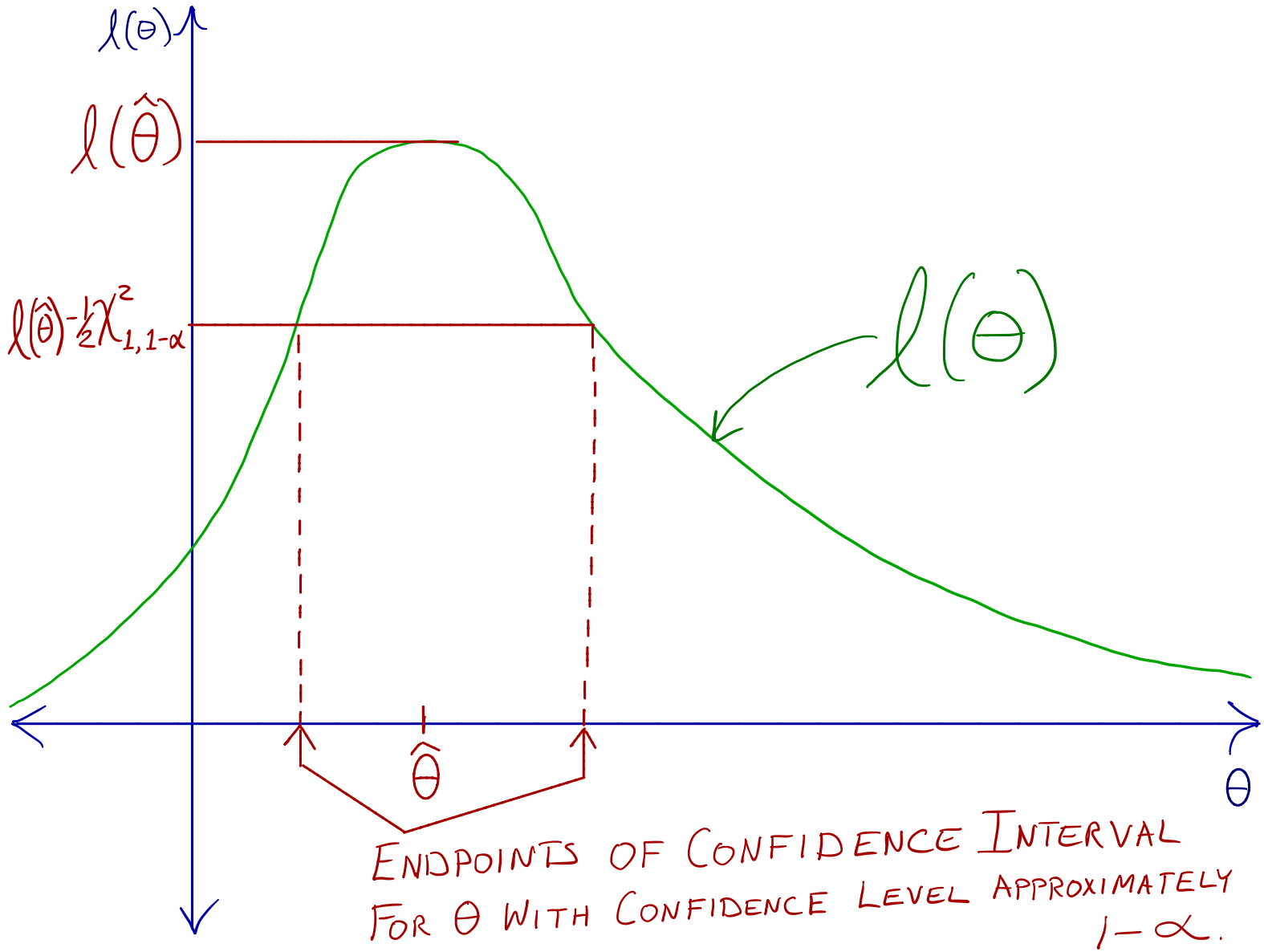
- Thus, the set of values of $\boldsymbol{\theta}_1$ that, when maximizing over $\boldsymbol{\theta}_2$, yield a maximized likelihood within $\frac{1}{2}\chi^2_{k_1,1-\alpha}$ of the likelihood maximized over all $\boldsymbol{\theta}$, form a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}_1$.

- Such a confidence region is known as a *profile likelihood confidence region* because

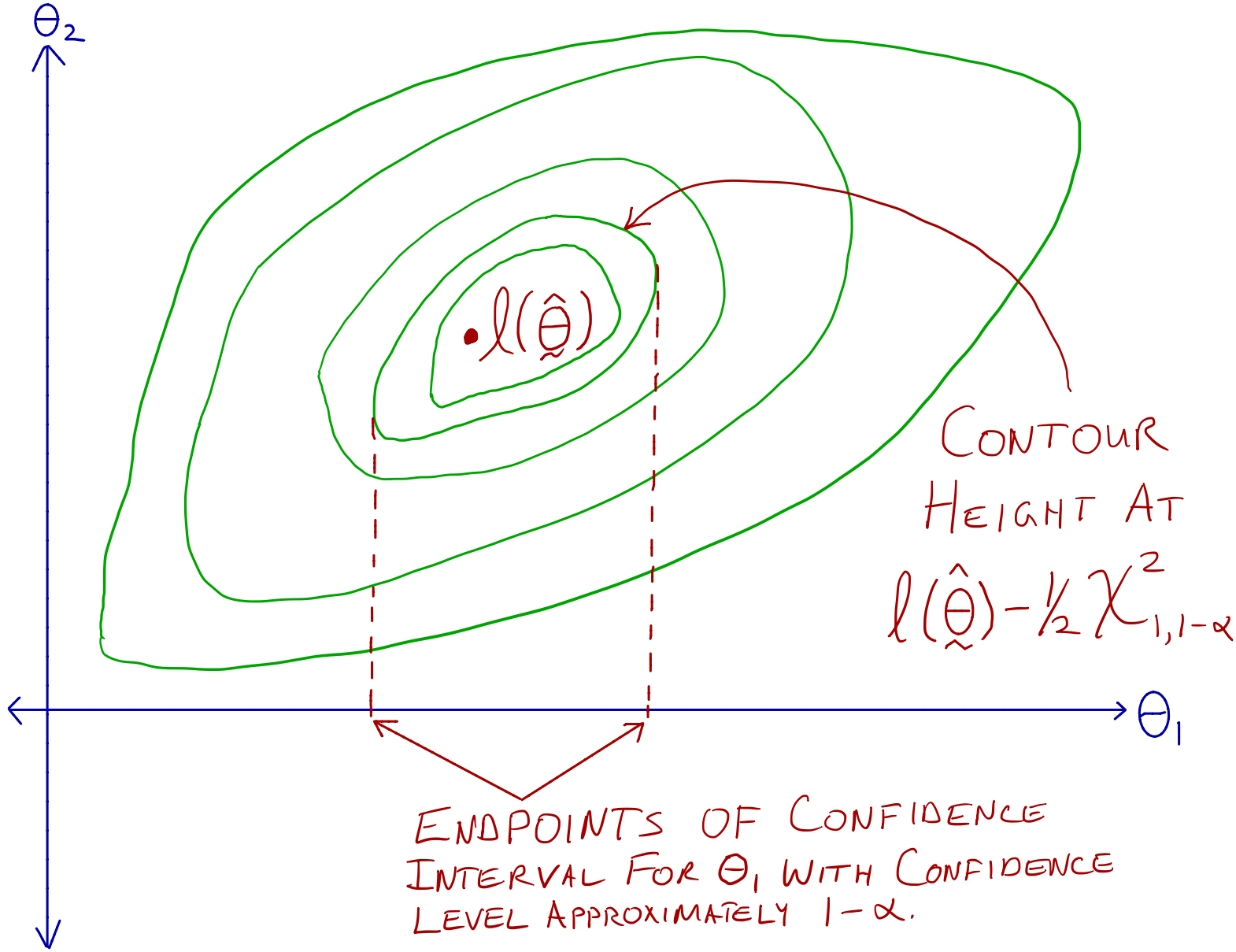
$$\ell\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) \end{bmatrix}\right)$$

  is the *profile log likelihood* for $\boldsymbol{\theta}_1$.

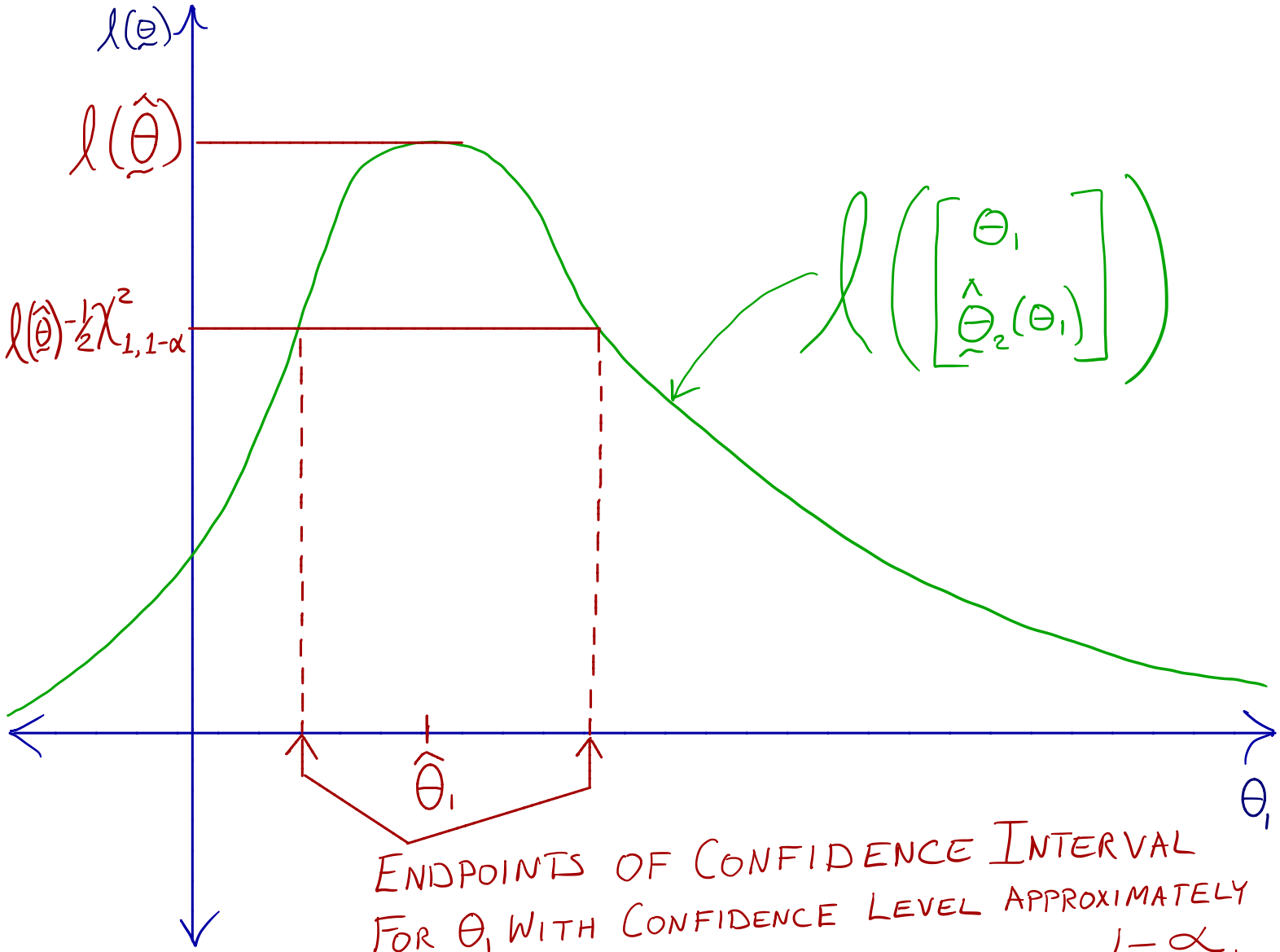
Sketch for the Case of $k = 1$



Sketch for the Case of $k = 2$



Sketch for the Case of $k_1 = 1$ and k_2 Arbitrary



Warnings

- The normal and χ^2 approximations mentioned in these notes may be crude if sample sizes are not sufficiently large.
- The regularity conditions mentioned in these notes do not hold if the true parameter falls on the boundary of the parameter space. Thus, as an example, testing $H_0 : \sigma_u^2 = 0$ is not covered by the methods presented here.