

MISCELLANEOUS 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

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

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 therefore different datasets for computation of maximized REML likelihoods.

LARGE n THEORY FOR MLEs

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

- There is an estimator $\hat{\theta}$ that solves the likelihood equations $\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{0}$ and is a (weakly) consistent estimator of θ , i.e.,

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

- For sufficiently large n ,

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta)),$$

where

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

- $I(\boldsymbol{\theta})$ is known as the *Fisher Information* matrix.
- $I(\boldsymbol{\theta})$ can be approximated by the *observed Fisher Information* matrix, which is given by

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

- $I(\boldsymbol{\theta})$ and $\hat{I}(\hat{\boldsymbol{\theta}})$ may depend on unknown nuisance parameters. In such cases, nuisance parameters are replaced by consistent estimators.

A Simple Example

- Suppose $y_1, \dots, y_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$.
- If we are interested in inference for μ , we can take $\theta = \mu$ and treat σ^2 as a nuisance parameter.
- It is straightforward to show that \bar{y} is the unique solution to the likelihood equation.
- Furthermore, it is straightforward to show that

$$I(\theta) = \hat{I}(\hat{\theta}) = n/\sigma^2.$$

A Simple Example (continued)

Thus, we have

$$\hat{\theta} = \bar{y} \cdot \sim N(\theta = \mu, I^{-1}(\theta) = \sigma^2/n)$$

and

$$\bar{y} \cdot \dot{\sim} N(\mu, s^2/n)$$

for sufficiently large n , where

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y} \cdot)^2}{n - 1}.$$

WALD TESTS AND CONFIDENCE INTERVALS

Suppose for large n that

$$\hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} N(\mathbf{0}, \mathbf{I})$$

and

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \hat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi_k^2.$$

For example, suppose $\hat{\theta}$ is the MLE of θ and $\hat{\mathbf{I}}(\hat{\theta})$ is the observed information matrix. Then under regularity conditions, we have

$$[\hat{\mathbf{I}}(\hat{\theta})]^{1/2}(\hat{\theta} - \theta) \overset{\sim}{\sim} N(\mathbf{0}, \mathbf{I})$$

and

$$(\hat{\theta} - \theta)' \hat{\mathbf{I}}(\hat{\theta}) (\hat{\theta} - \theta) \overset{\sim}{\sim} \chi_k^2$$

for sufficiently large n .

An approximate $100(1 - \alpha)\%$ confidence interval for θ_i is

$$\hat{\theta}_i \pm z_{1-\alpha/2} \sqrt{\hat{\Sigma}_{ii}},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the $N(0, 1)$ distribution and $\hat{\Sigma}_{ii}$ is element (i, i) of $\hat{\Sigma}$.

An approximate p -value for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$Pr[\chi_k^2 \geq (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)],$$

where χ_k^2 is a central χ^2 random variable with k degrees of freedom.

MULTIVARIATE DELTA METHOD

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

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

for some functions g_1, \dots, g_m .

- Suppose g is differentiable with derivative matrix

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

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

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

which implies

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

and

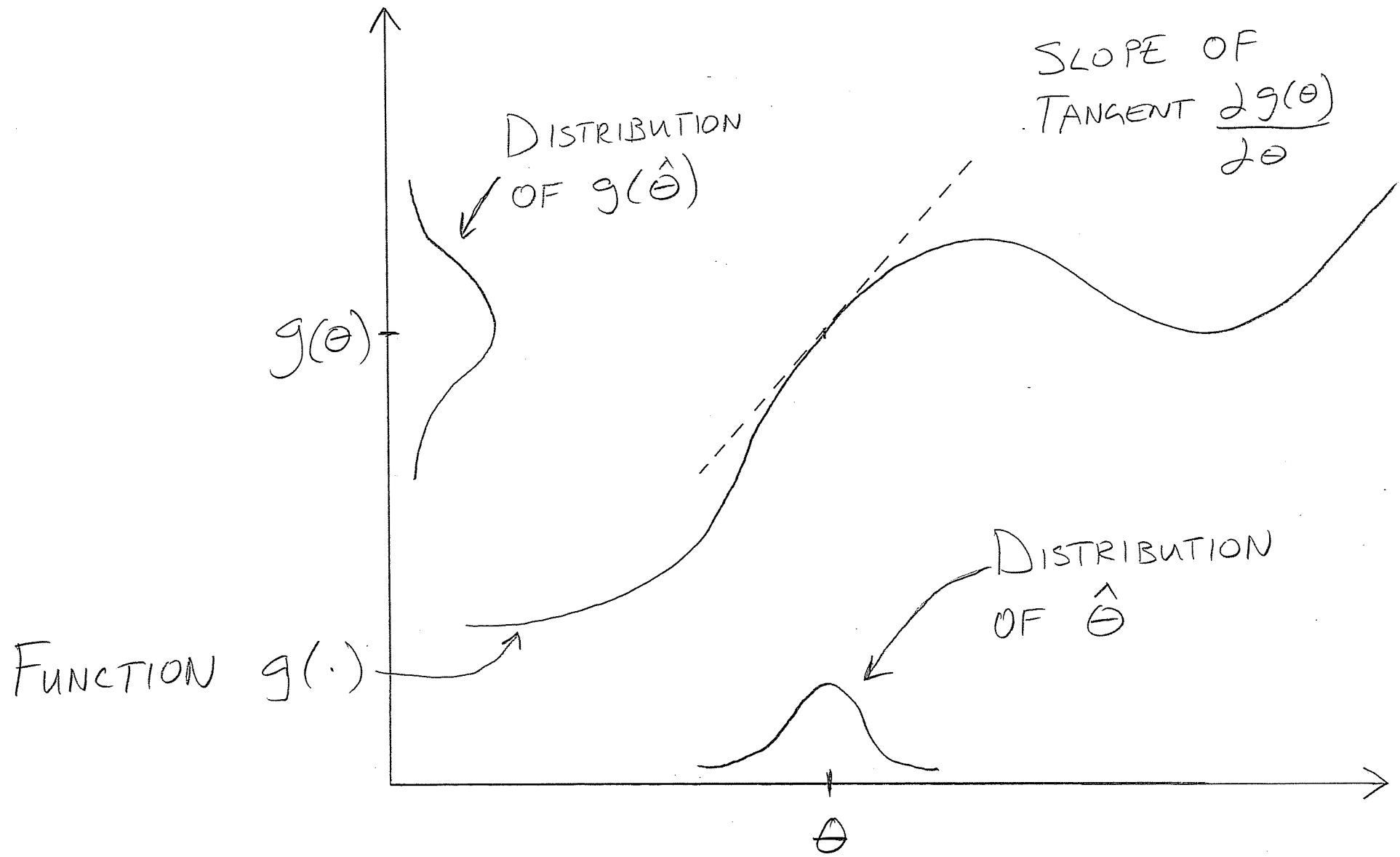
$$\text{Var}[\mathbf{g}(\hat{\theta})] \approx \text{Var}[\mathbf{g}(\theta) + \mathbf{D}'(\hat{\theta} - \theta)] = \mathbf{D}'\Sigma\mathbf{D}.$$

- If $\hat{\theta} \sim N(\theta, \Sigma)$, it follows that

$$\mathbf{g}(\hat{\theta}) \sim N(\mathbf{g}(\theta), \mathbf{D}'\Sigma\mathbf{D}).$$

- In practice, we often need to estimate \mathbf{D} by replacing θ in \mathbf{D} with $\hat{\theta}$ to obtain $\hat{\mathbf{D}}$.
- Similarly, we often need to replace Σ with an estimate $\hat{\Sigma}$.

THE DELTA METHOD



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

- Under the regularity conditions in Shao (2003) mentioned previously, the likelihood ratio test statistic $-2 \log \Lambda$ is approximately distributed as central $\chi_{k_f - k_r}^2$ 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 θ is $k \times 1$ vector and is partitioned into vectors θ_1 $k_1 \times 1$ and θ_2 $k_2 \times 1$, where $k = k_1 + k_2$ and $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$.
- Consider a test of $H_0 : \theta_1 = \theta_{10}$.

- Suppose $\hat{\theta}$ is the MLE of θ and $\hat{\theta}_2(\theta_1)$ maximizes $\ell\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}\right)$ over θ_2 for any fixed value of θ_1 .
- Then $2\left[\ell(\hat{\theta}) - \ell\left(\begin{bmatrix} \theta_{10} \\ \hat{\theta}_2(\theta_{10}) \end{bmatrix}\right)\right]$ is approximately $\chi_{k_1}^2$ under the null hypothesis by our previous result when n is “sufficiently large.”

Also,

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

which implies

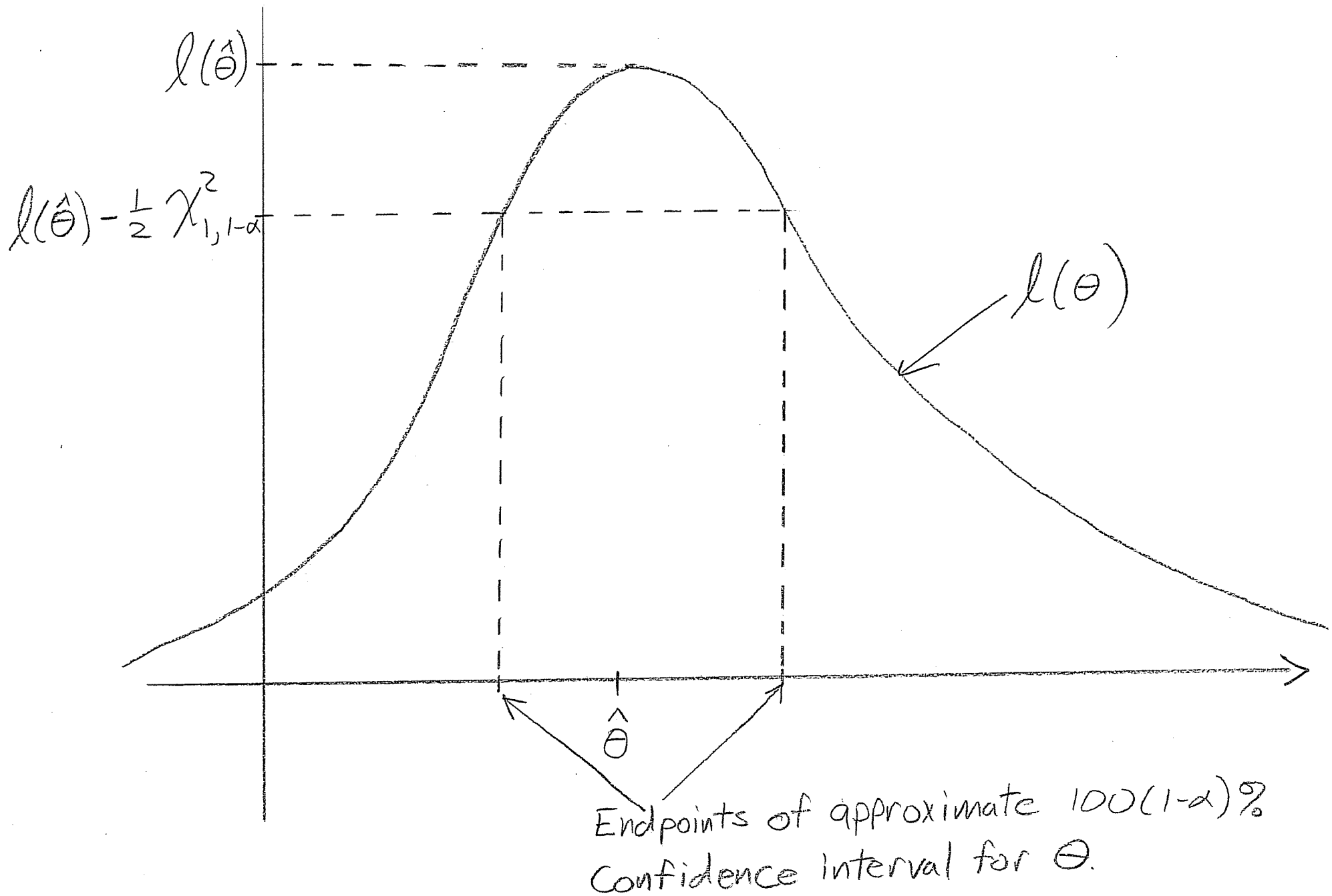
$$Pr \left\{ \ell \left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) \end{bmatrix} \right) \geq \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{k_1, 1-\alpha}^2 \right\} \approx 1 - \alpha.$$

- Thus, the set of values of θ_1 that, when maximizing over θ_2 , yield a maximized likelihood within $\frac{1}{2}\chi_{k_1, 1-\alpha}^2$ of the likelihood maximized over all θ , form a $100(1 - \alpha)\%$ confidence region for θ_1 .
- Such a confidence region is known as a *profile likelihood confidence region* because

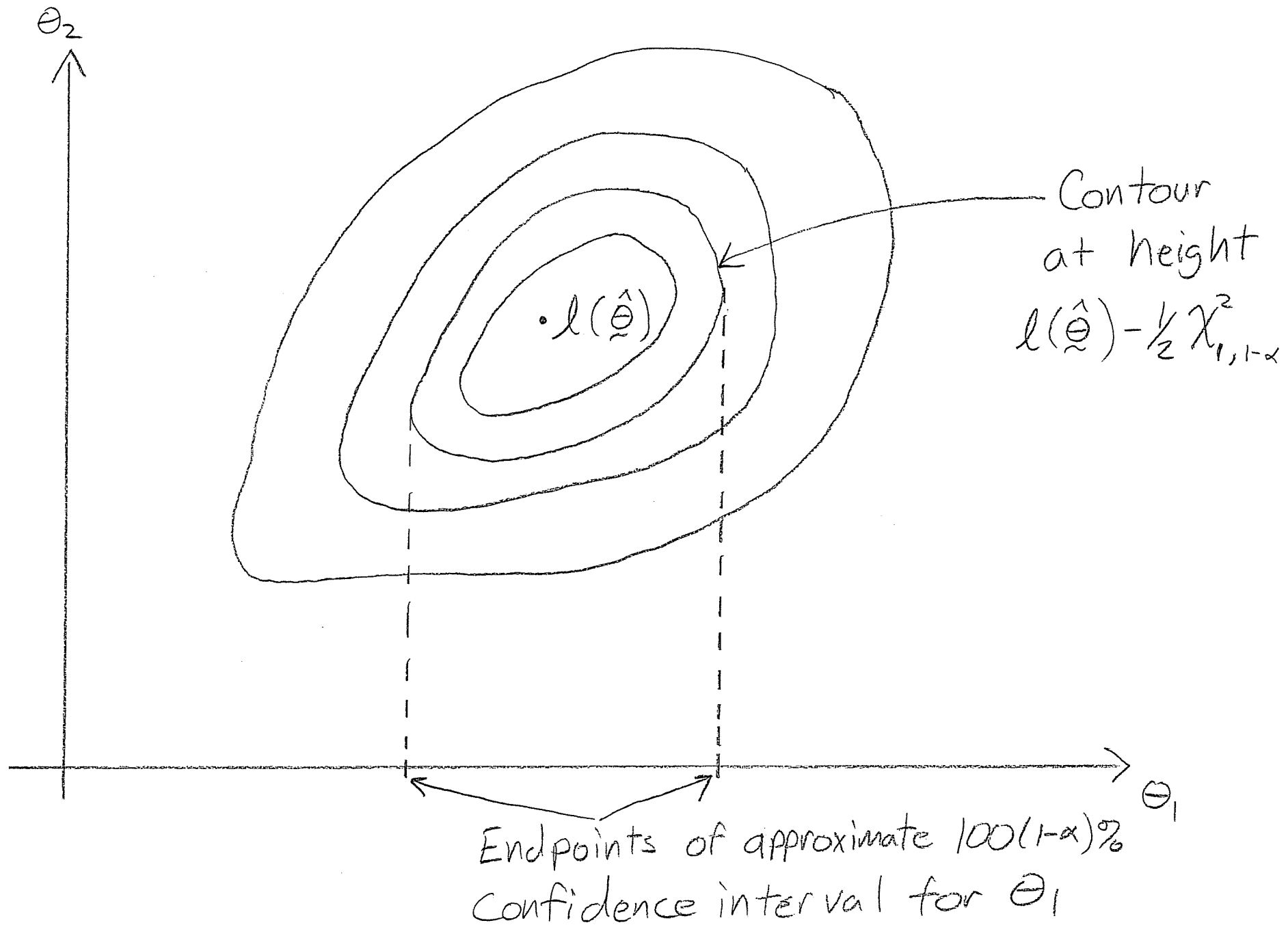
$$\ell \left(\begin{bmatrix} \theta_1 \\ \hat{\theta}_2(\theta_1) \end{bmatrix} \right)$$

is the *profile log likelihood* for θ_1 .

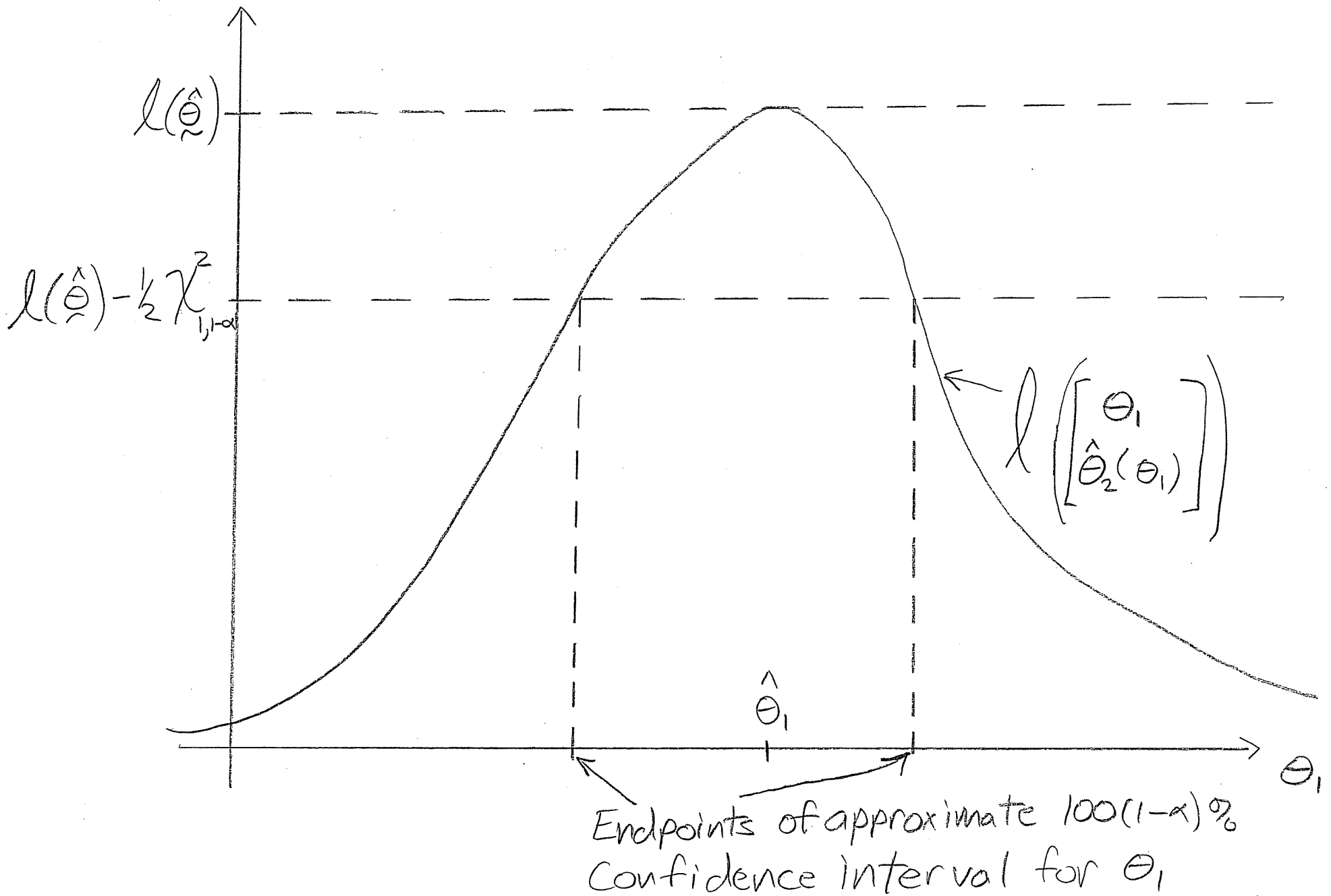
Sketch for the case $k=1$:



Sketch for the case $k=2$:



Sketch for the case $K_1 = 1$, K_2 arbitrary



A Warning

- 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.