Adjusting mis-specified likelihood functions

Richard Chandler

richard@stats.ucl.ac.uk

Department of Statistical Science, University College London

Overview of talk

1. **Problem statement** (setup and notation; log-likelihoods; potential applications)
2. **Standard asymptotics for mis-specified likelihoods** (definition of estimator; large-sample properties of estimator and log likelihood ratio)
3. **Adjusting the working log-likelihood** (motivation; options for adjustment; geometry of adjustment in 1-D; multiparameter example; comparing nested models; other applications)
4. **Open questions**

1. Problem statement
2. Standard asymptotics for mis-specified likelihoods
3. Adjusting the working log-likelihood
4. Open questions

Setup and notation

- Data arise as k 'clusters' $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)

Setup and notation

- Data arise as k ‘clusters’ $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)
- $\mathbf{y}_j = (y_{1j} \ \ldots \ y_{n_j j})'$ is vector of n_j observations in cluster j

Setup and notation

- Data arise as k ‘clusters’ $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)
- $\mathbf{y}_j = (y_{1j} \ \ldots \ y_{n_j j})'$ is vector of n_j observations in cluster j
- Let $\mathcal{C}_j$ be conditioning set for $\mathbf{y}_j$ (may include covariates and “history” — allow $\mathbf{y}_i \in \mathcal{C}_j$ for $i < j$, but not for $i \geq j$, hence clusters may be interdependent)

Setup and notation

- Data arise as k ‘clusters’ $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)
- $\mathbf{y}_j = (y_{1j} \ \ldots \ y_{n_j j})'$ is vector of n_j observations in cluster j
- Let $\mathcal{C}_j$ be conditioning set for $\mathbf{y}_j$ (may include covariates and “history” — allow $\mathbf{y}_i \in \mathcal{C}_j$ for $i < j$, but not for $i \geq j$, hence clusters may be interdependent)
- Observations from family of distributions with joint density

$$\prod_{j=1}^{k} f_j \left(\mathbf{y}_j \mid \mathcal{C}_j ; \, \theta, \alpha \right) \; .$$

for parameter vectors θ, α

Setup and notation

- Data arise as k ‘clusters’ $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)
- $\mathbf{y}_j = (y_{1j} \ \ldots \ y_{n_j j})'$ is vector of n_j observations in cluster j
- Let $\mathcal{C}_j$ be conditioning set for $\mathbf{y}_j$ (may include covariates and “history” — allow $\mathbf{y}_i \in \mathcal{C}_j$ for $i < j$, but not for $i \geq j$, hence clusters may be interdependent)
- Observations from family of distributions with joint density

$$\prod_{j=1}^{k} f_j \left(\mathbf{y}_j \mid \mathcal{C}_j ; \, \theta, \alpha \right) \; .$$

for parameter vectors θ, α

- Interested in / tractable model available for low-dimensional margins of joint distributions.

Setup and notation

- Data arise as k 'clusters' $\{\mathbf{y}_j : j = 1, \ldots, k\}$ (k could be 1)
- $\mathbf{y}_j = (y_{1j} \ \ldots \ y_{n_j j})'$ is vector of n_j observations in cluster j
- Let $\mathcal{C}_j$ be conditioning set for $\mathbf{y}_j$ (may include covariates and "history" — allow $\mathbf{y}_i \in \mathcal{C}_j$ for $i < j$, but not for $i \geq j$, hence clusters may be interdependent)
- Observations from family of distributions with joint density

$$\prod_{j=1}^{k} f_j\left(\mathbf{y}_j \mid \mathcal{C}_j;\, \theta, \alpha\right) \ .$$

  for parameter vectors θ, α
- Interested in / tractable model available for low-dimensional margins of joint distributions.
- Conditionally upon $\mathcal{C}_j$, low-dimensional margins are fully determined by $\theta \rightarrow \alpha$ is nuisance parameter for high-dimensional joint structure.

Log-likelihoods

- Full log-likelihood function is $\ell_{\text{FULL}}(\theta, \alpha) = \sum_{j=1}^{k} \log f_j\left(\mathbf{y}_j | \mathcal{C}_j; \theta, \alpha\right)$, but joint distributions usually difficult to model.

Log-likelihoods

- Full log-likelihood function is $\ell_{\text{FULL}}(\theta,\alpha) = \sum_{j=1}^{k} \log f_j\left(\mathbf{y}_j | \mathcal{C}_j; \theta, \alpha\right)$, but joint distributions usually difficult to model.
- **Alternative**: use ‘working’ log-likelihood based on low-dimensional margins:

$$\ell_{\text{WORK}}(\theta) = \sum_{j=1}^{k} \log \tilde{f}_j\left(y_j | \mathcal{C}_j; \theta\right)$$

where $\log \tilde{f}_j\left(y_j | \mathcal{C}_j; \theta\right)$ is contribution from cluster j (**NB** α missing here)

Log-likelihoods

- Full log-likelihood function is $\ell_{\text{FULL}}(\theta,\alpha) = \sum_{j=1}^{k} \log f_j\left(\mathbf{y}_j | \mathcal{C}_j; \theta, \alpha\right)$, but joint distributions usually difficult to model.
- **Alternative**: use ‘working’ log-likelihood based on low-dimensional margins:

$$\ell_{\text{WORK}}(\theta) = \sum_{j=1}^{k} \log \tilde{f}_j\left(y_j | \mathcal{C}_j; \theta\right)$$

where $\log \tilde{f}_j\left(y_j | \mathcal{C}_j; \theta\right)$ is contribution from cluster j (**NB** α missing here)

- **Examples:**
  - **Independence log-likelihood**: $\ell_{\text{IND}}(\theta) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \log f_{ij}\left(y_{ij} | \mathcal{C}_j; \theta\right)$ so that $\log \tilde{f}_j\left(y_j | \mathcal{C}_j; \theta\right) = \sum_{i=1}^{n_j} \log f_{ij}\left(y_{ij} | \mathcal{C}_j; \theta\right)$.

Log-likelihoods

- Full log-likelihood function is $\ell_{\text{FULL}}(\theta,\alpha) = \sum_{j=1}^{k} \log f_j(\mathbf{y}_j|\mathcal{C}_j;\theta,\alpha)$, but joint distributions usually difficult to model.
- **Alternative**: use 'working' log-likelihood based on low-dimensional margins:

$$\ell_{\text{WORK}}(\theta) = \sum_{j=1}^{k} \log \tilde{f}_j(y_j|\mathcal{C}_j;\theta)$$

  where $\log \tilde{f}_j(y_j|\mathcal{C}_j;\theta)$ is contribution from cluster j (**NB** α missing here)
- **Examples:**
  - **Independence log-likelihood**: $\ell_{\text{IND}}(\theta) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \log f_{ij}(y_{ij}|\mathcal{C}_j;\theta)$ so that $\log \tilde{f}_j(y_j|\mathcal{C}_j;\theta) = \sum_{i=1}^{n_j} \log f_{ij}(y_{ij}|\mathcal{C}_j;\theta)$.
  - **(Weighted) log pairwise likelihood**: $\ell_{\text{PAIR}}(\theta) = \sum_{j=1}^{k} w_j \sum_{i_1=1}^{n_j-1} \sum_{i_2=i_1+1}^{n_j} \log f_{i_1,i_2,j}(y_{i_1 j}, y_{i_2 j}|\mathcal{C}_j;\theta)$ so that $\log \tilde{f}_j(y_j|\mathcal{C}_j;\theta) = w_j \sum_{i_1=1}^{n_j-1} \sum_{i_2=i_1+1}^{n_j} \log f_{i_1,i_2,j}(y_{i_1 j}, y_{i_2 j}|\mathcal{C}_j;\theta)$.

Some potential applications

- **Longitudinal studies:**
  - 'Clusters' are patients
  - Can be assumed independent

Some potential applications

- **Longitudinal studies:**
  - 'Clusters' are patients
  - Can be assumed independent
- **Space-time data (multiple time series):**
  - 'Clusters' are observations made at same time instant
  - Temporal autocorrelation may be present — can be handled by including previous observations into conditioning sets $\{\mathcal{C}_j\}$

1. **Problem statement**
2. **Standard asymptotics for mis-specified likelihoods**
3. **Adjusting the working log-likelihood**
4. **Open questions**

Definition of estimator

- ‘Working score’ function is

$$\mathbf{U}(\theta) = \frac{\partial \ell_{\text{WORK}}}{\partial \theta} = \sum_{j=1}^{k} \mathbf{U}_j(\theta) \text{ say.}$$

Definition of estimator

- ‘Working score’ function is

$$\mathbf{U}(\theta) = \frac{\partial \ell_{\text{WORK}}}{\partial \theta} = \sum_{j=1}^{k} \mathbf{U}_j(\theta) \text{ say.}$$

- If data are generated from distribution with $\theta = \theta_0$ then, under general conditions, working score contributions $\{\mathbf{U}_j(\theta_0)\}$ are uncorrelated with zero mean (may need to include ‘history’ into $\mathcal{C}_j$ to ensure this when clusters are interdependent — see Chapter 5 of *Statistical Methods for Spatial-temporal Systems*, eds. Finkenstadt, Held & Isham, CRC Press, 2007).

Definition of estimator

- 'Working score' function is

$$\mathbf{U}(\theta) = \frac{\partial \ell_{\mathrm{WORK}}}{\partial \theta} = \sum_{j=1}^{k} \mathbf{U}_j(\theta) \text{ say.}$$

- If data are generated from distribution with $\theta = \theta_0$ then, under general conditions, working score contributions $\left\{\mathbf{U}_j(\theta_0)\right\}$ are uncorrelated with zero mean (may need to include 'history' into $\mathcal{C}_j$ to ensure this when clusters are interdependent — see Chapter 5 of *Statistical Methods for Spatial-temporal Systems*, eds. Finkenstadt, Held & Isham, CRC Press, 2007).
- Estimator $\hat{\theta}_{\mathrm{WORK}}$ satisfies $\mathbf{U}\left(\hat{\theta}_{\mathrm{WORK}}\right) = \sum_{j=1}^{k} \mathbf{U}_j\left(\hat{\theta}_{\mathrm{WORK}}\right) = \mathbf{0}$.

Large-sample properties of $\hat{\theta}_{\mathrm{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\mathrm{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\left.\frac{\partial^2 \ell_{\mathrm{WORK}}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right), \quad \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j(\theta_0)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j(\theta_0)\,\mathbf{U}_j(\theta_0)'\right]$$

Large-sample properties of $\hat{\theta}_{\mathrm{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\mathrm{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\left.\frac{\partial^2 \ell_{\mathrm{WORK}}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right), \quad \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j(\theta_0)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j(\theta_0)\,\mathbf{U}_j(\theta_0)'\right]$$

- Estimate $\mathbf{H}$ using either expected or observed Hessian at $\hat{\theta}_{\mathrm{WORK}}$, say $\hat{\mathbf{H}}$.

Large-sample properties of $\hat{\theta}_{\mathrm{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\mathrm{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\left.\frac{\partial^2 \ell_{\mathrm{WORK}}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right), \quad \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j(\theta_0)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j(\theta_0)\,\mathbf{U}_j(\theta_0)'\right]$$

- Estimate $\mathbf{H}$ using either expected or observed Hessian at $\hat{\theta}_{\mathrm{WORK}}$, say $\hat{\mathbf{H}}$.
- Estimate $\mathbf{V}$ using empirical counterpart: $\hat{\mathbf{V}} = \sum_{j=1}^{k} \mathbf{U}_j\left(\hat{\theta}_{\mathrm{WORK}}\right)\mathbf{U}_j\left(\hat{\theta}_{\mathrm{WORK}}\right)'$

Large-sample properties of $\hat{\theta}_{\text{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\text{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\left.\frac{\partial^2 \ell_{\text{WORK}}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right), \ \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j\left(\theta_0\right)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j\left(\theta_0\right)\mathbf{U}_j\left(\theta_0\right)'\right]$$

- Estimate $\mathbf{H}$ using either expected or observed Hessian at $\hat{\theta}_{\text{WORK}}$, say $\hat{\mathbf{H}}$.
- Estimate $\mathbf{V}$ using empirical counterpart: $\hat{\mathbf{V}} = \sum_{j=1}^{k} \mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right)\mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right)'$
- Covariance matrix of $\hat{\theta}_{\text{WORK}}$ estimated consistently by robust estimator $\mathcal{R} = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{H}}^{-1}$ — gives Wald tests & confidence regions for components of θ

Large-sample properties of $\hat{\theta}_{\text{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\text{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\left.\frac{\partial^2 \ell_{\text{WORK}}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right), \ \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j\left(\theta_0\right)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j\left(\theta_0\right)\mathbf{U}_j\left(\theta_0\right)'\right]$$

- Estimate $\mathbf{H}$ using either expected or observed Hessian at $\hat{\theta}_{\text{WORK}}$, say $\hat{\mathbf{H}}$.
- Estimate $\mathbf{V}$ using empirical counterpart: $\hat{\mathbf{V}} = \sum_{j=1}^{k} \mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right)\mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right)'$
- Covariance matrix of $\hat{\theta}_{\text{WORK}}$ estimated consistently by robust estimator $\mathcal{R} = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{H}}^{-1}$ — gives Wald tests & confidence regions for components of θ
- Contrast with naïve estimator $\mathcal{N} = -\hat{\mathbf{H}}^{-1}$ (ignores mis-specification of working log- likelihood)

Large-sample properties of $\hat{\theta}_{\text{WORK}}$

- Usual asymptotics hold e.g. for large k, $\hat{\theta}_{\text{WORK}} \sim N\left(\theta_0, \mathbf{H}\mathbf{V}^{-1}\mathbf{H}\right)$ where

$$\mathbf{H} = \mathrm{E}\left(\frac{\partial^2 \ell_{\text{WORK}}}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta_0}\right), \ \mathbf{V} = \mathrm{Var}\left[\sum_{j=1}^{k} \mathbf{U}_j(\theta_0)\right] = \sum_{j=1}^{k} \mathrm{E}\left[\mathbf{U}_j(\theta_0)\,\mathbf{U}_j(\theta_0)'\right]$$

- Estimate $\mathbf{H}$ using either expected or observed Hessian at $\hat{\theta}_{\text{WORK}}$, say $\hat{\mathbf{H}}$.
- Estimate $\mathbf{V}$ using empirical counterpart: $\hat{\mathbf{V}} = \sum_{j=1}^{k} \mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right) \mathbf{U}_j\left(\hat{\theta}_{\text{WORK}}\right)'$
- Covariance matrix of $\hat{\theta}_{\text{WORK}}$ estimated consistently by robust estimator $\mathcal{R} = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{H}}^{-1}$ — gives Wald tests & confidence regions for components of θ
- Contrast with naïve estimator $\mathcal{N} = -\hat{\mathbf{H}}^{-1}$ (ignores mis-specification of working log- likelihood)
- **NB** other techniques required for small k — application-dependent

Large-sample properties of working log likelihood

- Partition θ as $(\phi' \ \ \psi')'$ and let $\tilde{\theta}_{\text{WORK}} = \arg\sup_{\psi=\psi_0} \ell_{\text{WORK}}(\theta)$.

Large-sample properties of working log likelihood

- Partition θ as $(\phi' \;\; \psi')'$ and let $\tilde{\theta}_{\mathrm{WORK}} = \arg\sup_{\psi=\psi_0} \ell_{\mathrm{WORK}}(\theta)$.
- For large k:
  - Distribution of $\Lambda_{WORK} = 2\left[\ell_{WORK}\left(\hat{\theta}_{WORK}\right) - \ell_{WORK}\left(\tilde{\theta}_{WORK}\right)\right]$ is approximately that of weighted sum of χ^2_1 random variables
  - Weights are solution to partition-dependent eigenproblem

Large-sample properties of working log likelihood

- Partition θ as $(\phi' \;\; \psi')'$ and let $\tilde{\theta}_{\mathrm{WORK}} = \arg\sup_{\psi=\psi_0} \ell_{\mathrm{WORK}}(\theta)$.
- For large k:
  - Distribution of $\Lambda_{WORK} = 2\left[\ell_{WORK}\left(\hat{\theta}_{WORK}\right) - \ell_{WORK}\left(\tilde{\theta}_{WORK}\right)\right]$ is approximately that of weighted sum of χ^2_1 random variables
  - Weights are solution to partition-dependent eigenproblem
- Can be used for profile-based inference on components of θ.

1. Problem statement
2. Standard asymptotics for mis-specified likelihoods
3. Adjusting the working log-likelihood
4. Open questions

Motivation

- Profile log-likelihoods useful for tests & confidence regions.

Motivation

- Profile log-likelihoods useful for tests & confidence regions.
- ‘Classical’ approach adjusts ‘naïve’ critical values for tests based on $\ell_{\text{WORK}}(\theta)$, **BUT:**
  - Adjustments difficult to calculate (weighted sum of independent chi-squareds)
  - Adjusted critical values are direction-dependent

Motivation

- Profile log-likelihoods useful for tests & confidence regions.
- ‘Classical’ approach adjusts ‘naïve’ critical values for tests based on $\ell_{\text{WORK}}(\theta)$, **BUT:**
  - Adjustments difficult to calculate (weighted sum of independent chi-squareds)
  - Adjusted critical values are direction-dependent
- **Alternative:** adjust $\ell_{\text{WORK}}(\theta)$ to maintain usual asymptotics (Chandler & Bate, *Biometrika*, 2007):

Motivation

- Profile log-likelihoods useful for tests & confidence regions.
- ‘Classical’ approach adjusts ‘naïve’ critical values for tests based on $\ell_{\text{WORK}}(\theta)$, **BUT:**
  - Adjustments difficult to calculate (weighted sum of independent chi-squareds)
  - Adjusted critical values are direction-dependent
- **Alternative:** adjust $\ell_{\text{WORK}}(\theta)$ to maintain usual asymptotics (Chandler & Bate, *Biometrika*, 2007):
  - Naïve covariance matrix is $\mathcal{N} = -\hat{\mathbf{H}}^{-1} \Rightarrow \ell_{\text{WORK}}(\theta)$ has Hessian $\hat{\mathbf{H}} = -\mathcal{N}^{-1}$.

Motivation

- Profile log-likelihoods useful for tests & confidence regions.
- ‘Classical’ approach adjusts ‘naïve’ critical values for tests based on $\ell_{\text{WORK}}(\theta)$, **BUT:**
  - Adjustments difficult to calculate (weighted sum of independent chi-squareds)
  - Adjusted critical values are direction-dependent
- **Alternative:** adjust $\ell_{\text{WORK}}(\theta)$ to maintain usual asymptotics (Chandler & Bate, *Biometrika*, 2007):
  - Naïve covariance matrix is $\mathcal{N} = -\hat{\mathbf{H}}^{-1} \Rightarrow \ell_{\text{WORK}}(\theta)$ has Hessian $\hat{\mathbf{H}} = -\mathcal{N}^{-1}$.
  - Robust covariance matrix is $\mathcal{R} \Rightarrow$ define adjusted inference function with Hessian $\hat{\mathbf{H}}_{\text{ADJ}} = -\mathcal{R}^{-1}$.

Motivation

- Profile log-likelihoods useful for tests & confidence regions.
- ‘Classical’ approach adjusts ‘naïve’ critical values for tests based on $\ell_{\mathrm{WORK}}(\theta)$, **BUT:**
  - Adjustments difficult to calculate (weighted sum of independent chi-squareds)
  - Adjusted critical values are direction-dependent
- **Alternative:** adjust $\ell_{\mathrm{WORK}}(\theta)$ to maintain usual asymptotics (Chandler & Bate, *Biometrika*, 2007):
  - Naïve covariance matrix is $\mathcal{N} = -\hat{\mathbf{H}}^{-1} \Rightarrow \ell_{\mathrm{WORK}}(\theta)$ has Hessian $\hat{\mathbf{H}} = -\mathcal{N}^{-1}$.
  - Robust covariance matrix is $\mathcal{R} \Rightarrow$ define adjusted inference function with Hessian $\hat{\mathbf{H}}_{\mathrm{ADJ}} = -\mathcal{R}^{-1}$.
  - Borrow profile from $\ell_{\mathrm{WORK}}(\theta)$ — hopefully informative.

Options for adjustment

Horizontal scaling: define $\ell_{\mathrm{ADJ}}(\theta) = \ell_{\mathrm{WORK}}(\theta^*)$, where

$$\theta^* = \hat{\theta}_{\mathrm{WORK}} + \mathbf{M}^{-1}\mathbf{M}_{\mathrm{ADJ}}\left(\theta - \hat{\theta}\right)$$

with $\mathbf{M}'\mathbf{M} = \hat{\mathbf{H}}$, $\mathbf{M}'_{\mathrm{ADJ}}\mathbf{M}_{\mathrm{ADJ}} = \hat{\mathbf{H}}_{\mathrm{ADJ}}$. Possible choices for $\mathbf{M}$, $\mathbf{M}_{\mathrm{ADJ}}$:

- Choleski square roots.
- ‘Minimal rotation’ square roots e.g. $\mathbf{M} = \mathbf{L}\mathbf{D}^{1/2}\mathbf{L}$, where $\mathbf{LDL}$ is spectral decomposition of $\hat{\mathbf{H}}$.

Options for adjustment

Horizontal scaling: define $\ell_{\mathrm{ADJ}}(\theta) = \ell_{\mathrm{WORK}}(\theta^*)$, where

$$\theta^* = \hat{\theta}_{\mathrm{WORK}} + \mathbf{M}^{-1}\mathbf{M}_{\mathrm{ADJ}}(\theta - \hat{\theta})$$

with $\mathbf{M}'\mathbf{M} = \hat{\mathbf{H}}$, $\mathbf{M}'_{\mathrm{ADJ}}\mathbf{M}_{\mathrm{ADJ}} = \hat{\mathbf{H}}_{\mathrm{ADJ}}$. Possible choices for $\mathbf{M}$, $\mathbf{M}_{\mathrm{ADJ}}$:

- Choleski square roots.
- ‘Minimal rotation’ square roots e.g. $\mathbf{M} = \mathbf{LD}^{1/2}\mathbf{L}$, where $\mathbf{LDL}$ is spectral decomposition of $\hat{\mathbf{H}}$.

Vertical scaling: define $\ell_{\mathrm{ADJ}}(\theta)$ as

$$\ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}) + \left\{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}_{\mathrm{ADJ}}(\theta - \hat{\theta}_{\mathrm{WORK}})\right\} \frac{\ell_{\mathrm{WORK}}(\theta) - \ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}}{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}(\theta - \hat{\theta}_{\mathrm{WORK}}}$$

Options for adjustment

Horizontal scaling: define $\ell_{\mathrm{ADJ}}(\theta) = \ell_{\mathrm{WORK}}(\theta^*)$, where

$$\theta^* = \hat{\theta}_{\mathrm{WORK}} + \mathbf{M}^{-1}\mathbf{M}_{\mathrm{ADJ}}(\theta - \hat{\theta})$$

with $\mathbf{M}'\mathbf{M} = \hat{\mathbf{H}}$, $\mathbf{M}'_{\mathrm{ADJ}}\mathbf{M}_{\mathrm{ADJ}} = \hat{\mathbf{H}}_{\mathrm{ADJ}}$. Possible choices for $\mathbf{M}$, $\mathbf{M}_{\mathrm{ADJ}}$:

- Choleski square roots.
- ‘Minimal rotation’ square roots e.g. $\mathbf{M} = \mathbf{LD}^{1/2}\mathbf{L}$, where $\mathbf{LDL}$ is spectral decomposition of $\hat{\mathbf{H}}$.

Vertical scaling: define $\ell_{\mathrm{ADJ}}(\theta)$ as

$$\ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}) + \left\{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}_{\mathrm{ADJ}}(\theta - \hat{\theta}_{\mathrm{WORK}})\right\} \frac{\ell_{\mathrm{WORK}}(\theta) - \ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}}{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}(\theta - \hat{\theta}_{\mathrm{WORK}}}$$

- Options asymptotically equivalent (and identical in quadratic case)

Options for adjustment

Horizontal scaling: define $\ell_{\mathrm{ADJ}}(\theta) = \ell_{\mathrm{WORK}}(\theta^*)$, where

$$\theta^* = \hat{\theta}_{\mathrm{WORK}} + \mathbf{M}^{-1}\mathbf{M}_{\mathrm{ADJ}}(\theta - \hat{\theta})$$

with $\mathbf{M}'\mathbf{M} = \hat{\mathbf{H}}$, $\mathbf{M}'_{\mathrm{ADJ}}\mathbf{M}_{\mathrm{ADJ}} = \hat{\mathbf{H}}_{\mathrm{ADJ}}$. Possible choices for $\mathbf{M}$, $\mathbf{M}_{\mathrm{ADJ}}$:

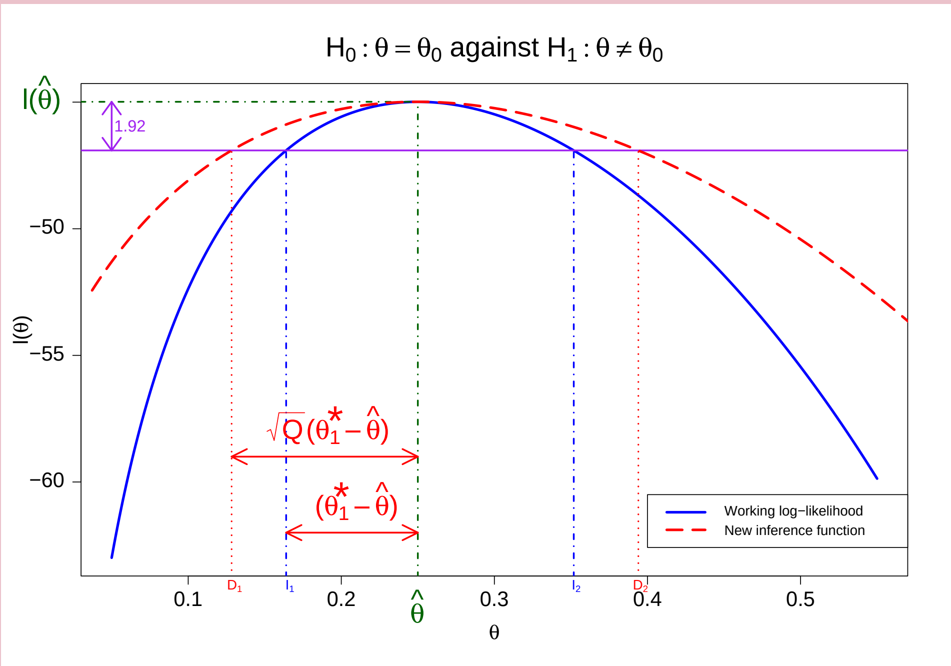
- Choleski square roots.
- 'Minimal rotation' square roots e.g. $\mathbf{M} = \mathbf{L}\mathbf{D}^{1/2}\mathbf{L}$, where $\mathbf{LDL}$ is spectral decomposition of $\hat{\mathbf{H}}$.

Vertical scaling: define $\ell_{\mathrm{ADJ}}(\theta)$ as

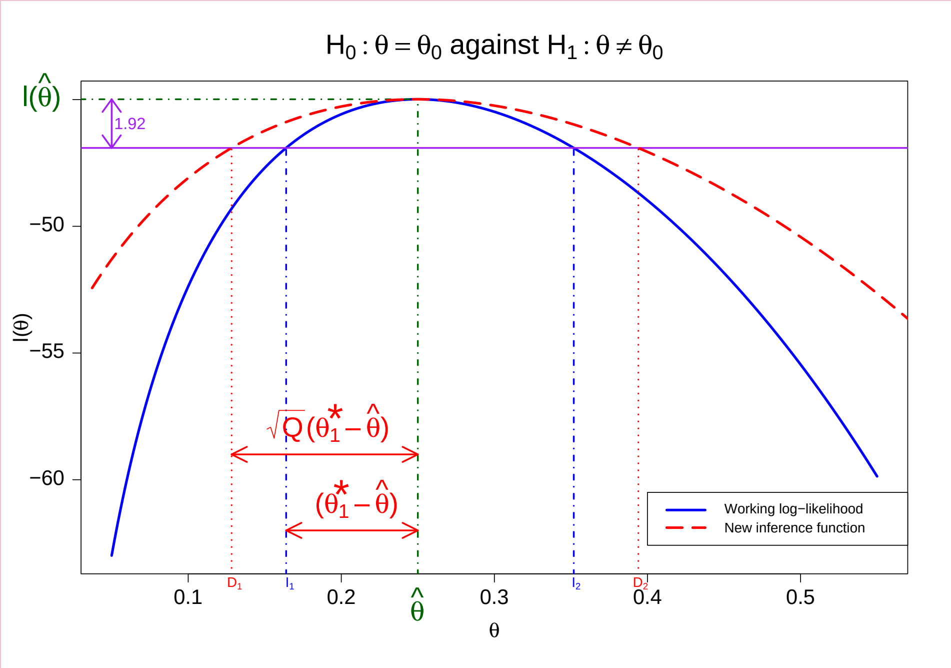
$$\ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}) + \left\{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}_{\mathrm{ADJ}}(\theta - \hat{\theta}_{\mathrm{WORK}})\right\}\frac{\ell_{\mathrm{WORK}}(\theta) - \ell_{\mathrm{WORK}}(\hat{\theta}_{\mathrm{WORK}}}{(\theta - \hat{\theta}_{\mathrm{WORK}})'\hat{\mathbf{H}}(\theta - \hat{\theta}_{\mathrm{WORK}}}$$

- Options asymptotically equivalent (and identical in quadratic case)
- Vertical scaling has practical (and theoretical) advantages

Geometry of adjustment in 1-D

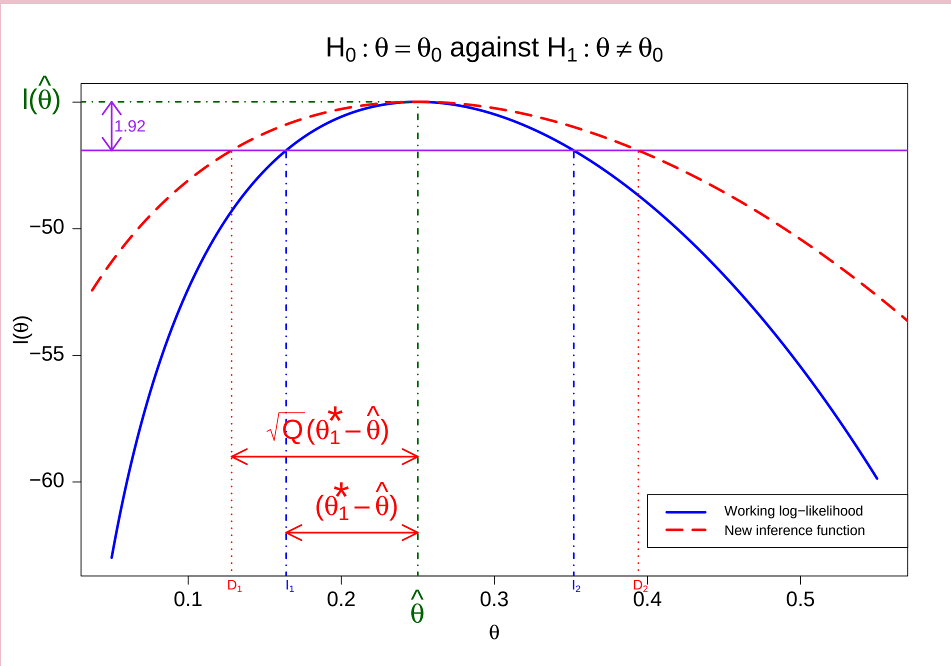


Geometry of adjustment in 1-D



- Horizontal scaling is by ratio of robust to naïve standard errors.

Geometry of adjustment in 1-D



- Horizontal scaling is by ratio of robust to naïve standard errors.
- Vertical scaling is by ratio of robust to naïve variances (same as adjusting critical value)

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.
- Independence log-likelihood for $\theta = \left(\mu_1 \;\; \mu_2 \;\; \sigma_1^2 \;\; \sigma_2^2\right)'$ is
$\ell_{\text{IND}}(\theta) = -\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{2}\left[\log\sigma_i^2 + \sigma_i^{-2}(Y_{ij} - \mu_i)^2\right] + \text{constant}$.

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.
- Independence log-likelihood for $\theta = (\mu_1 \;\; \mu_2 \;\; \sigma_1^2 \;\; \sigma_2^2)'$ is
  $\ell_{\text{IND}}(\theta) = -\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{2}\left[\log\sigma_i^2 + \sigma_i^{-2}(Y_{ij} - \mu_i)^2\right] + \text{constant}$.
- μ- and σ- components of θ are orthogonal in $\ell_{\text{IND}}(\theta)$

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.
- Independence log-likelihood for $\theta = (\mu_1 \;\; \mu_2 \;\; \sigma_1^2 \;\; \sigma_2^2)'$ is
  $\ell_{\text{IND}}(\theta) = -\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{2}\left[\log\sigma_i^2 + \sigma_i^{-2}(Y_{ij} - \mu_i)^2\right] + \text{constant}$.
- μ- and σ- components of θ are orthogonal in $\ell_{\text{IND}}(\theta)$
- Naïve and robust covariance matrices of $\hat{\mu} = \overline{\mathbf{Y}}$ are $\mathcal{N} = k^{-1}\text{diag}\,(\hat{\sigma}_1^2 \;\; \hat{\sigma}_2^2)$; $\mathcal{R} = k^{-1}\hat{\Sigma}$.

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.
- Independence log-likelihood for $\theta = (\mu_1 \;\; \mu_2 \;\; \sigma_1^2 \;\; \sigma_2^2)'$ is
  $\ell_{\text{IND}}(\theta) = -\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{2}\left[\log\sigma_i^2 + \sigma_i^{-2}(Y_{ij} - \mu_i)^2\right] + \text{constant}$.
- μ- and σ- components of θ are orthogonal in $\ell_{\text{IND}}(\theta)$
- Naïve and robust covariance matrices of $\hat{\mu} = \overline{\mathbf{Y}}$ are $\mathcal{N} = k^{-1}\text{diag}\,(\hat{\sigma}_1^2 \;\; \hat{\sigma}_2^2)$; $\mathcal{R} = k^{-1}\hat{\Sigma}$.
- Adjusted profile log-likelihood for μ (horizontal or vertical scaling) is
  $\ell_{\text{ADJ}}(\mu) = -\frac{k}{2}(\overline{\mathbf{Y}} - \mu)'\hat{\Sigma}^{-1}(\overline{\mathbf{Y}} - \mu) + \text{constant}$ — i.e. correct bivariate log-likelihood.

Multiparameter case: a 2-dimensional example

- k bivariate normal pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \ldots, k\}$ with unknown mean μ and covariance matrix Σ.
- Independence log-likelihood for $\theta = (\mu_1 \;\; \mu_2 \;\; \sigma_1^2 \;\; \sigma_2^2)'$ is
  $\ell_{\text{IND}}(\theta) = -\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{2}\left[\log\sigma_i^2 + \sigma_i^{-2}(Y_{ij} - \mu_i)^2\right] + \text{constant}$.
- μ- and σ- components of θ are orthogonal in $\ell_{\text{IND}}(\theta)$
- Naïve and robust covariance matrices of $\hat{\mu} = \overline{\mathbf{Y}}$ are $\mathcal{N} = k^{-1}\text{diag}\,(\hat{\sigma}_1^2 \;\; \hat{\sigma}_2^2)$; $\mathcal{R} = k^{-1}\hat{\Sigma}$.
- Adjusted profile log-likelihood for μ (horizontal or vertical scaling) is
  $\ell_{\text{ADJ}}(\mu) = -\frac{k}{2}(\overline{\mathbf{Y}} - \mu)'\hat{\Sigma}^{-1}(\overline{\mathbf{Y}} - \mu) + \text{constant}$ — i.e. correct bivariate log-likelihood.
- **NB** contours of ℓ_{IND} are always circular — hence classical approach of adjusting critical value is sub-optimal.

Comparing nested models

- Adjustment preserves χ^2 asymptotics by construction $\Rightarrow$ to test $H_0 : \Delta\theta = \delta_0$, use statistic $\Lambda_{\mathrm{ADJ}} = 2\left\{\ell_{\mathrm{ADJ}}\left(\hat{\theta}_{\mathrm{WORK}}\right) - \ell_{\mathrm{ADJ}}\left(\widetilde{\theta}_{\mathrm{ADJ}}\right)\right\}$, where $\widetilde{\theta}_{\mathrm{ADJ}}$ maximises ℓ_{ADJ} under H_0.

Comparing nested models

- Adjustment preserves χ^2 asymptotics by construction $\Rightarrow$ to test $H_0 : \Delta\theta = \delta_0$, use statistic $\Lambda_{\mathrm{ADJ}} = 2\left\{\ell_{\mathrm{ADJ}}\left(\hat{\theta}_{\mathrm{WORK}}\right) - \ell_{\mathrm{ADJ}}\left(\widetilde{\theta}_{\mathrm{ADJ}}\right)\right\}$, where $\widetilde{\theta}_{\mathrm{ADJ}}$ maximises ℓ_{ADJ} under H_0.
- **Problem:** $\widetilde{\theta}_{\mathrm{ADJ}}$ could be difficult / expensive to compute.

Comparing nested models

- Adjustment preserves χ^2 asymptotics by construction $\Rightarrow$ to test $H_0 : \Delta\theta = \delta_0$, use statistic $\Lambda_{\text{ADJ}} = 2\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)\right\}$, where $\widetilde{\theta}_{\text{ADJ}}$ maximises ℓ_{ADJ} under H_0.
- **Problem:** $\widetilde{\theta}_{\text{ADJ}}$ could be difficult / expensive to compute.
- **Alternative:** use asymptotically equivalent statistic based on one-step approximation to $\ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)$:

$$\Lambda^*_{\text{ADJ}} = 2c\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{WORK}}\right)\right\}$$

$$\text{where} \quad c = \frac{\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)'\left[\Delta \mathbf{H}^{-1}_{ADJ}\Delta'\right]^{-1}\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)}{\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)'\hat{\mathbf{H}}_{ADJ}\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)}$$

Comparing nested models

- Adjustment preserves χ^2 asymptotics by construction $\Rightarrow$ to test $H_0 : \Delta\theta = \delta_0$, use statistic $\Lambda_{\text{ADJ}} = 2\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)\right\}$, where $\widetilde{\theta}_{\text{ADJ}}$ maximises ℓ_{ADJ} under H_0.
- **Problem:** $\widetilde{\theta}_{\text{ADJ}}$ could be difficult / expensive to compute.
- **Alternative:** use asymptotically equivalent statistic based on one-step approximation to $\ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)$:

$$\Lambda^*_{\text{ADJ}} = 2c\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{WORK}}\right)\right\}$$

$$\text{where} \quad c = \frac{\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)'\left[\Delta \mathbf{H}^{-1}_{ADJ}\Delta'\right]^{-1}\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)}{\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)'\hat{\mathbf{H}}_{ADJ}\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)}$$

- Λ^*_{ADJ} needs only estimates from working likelihood.

Comparing nested models

- Adjustment preserves χ^2 asymptotics by construction $\Rightarrow$ to test $H_0 : \Delta\theta = \delta_0$, use statistic $\Lambda_{\text{ADJ}} = 2\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)\right\}$, where $\widetilde{\theta}_{\text{ADJ}}$ maximises ℓ_{ADJ} under H_0.
- **Problem:** $\widetilde{\theta}_{\text{ADJ}}$ could be difficult / expensive to compute.
- **Alternative:** use asymptotically equivalent statistic based on one-step approximation to $\ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{ADJ}}\right)$:

$$\Lambda^*_{\text{ADJ}} = 2c\left\{\ell_{\text{ADJ}}\left(\hat{\theta}_{\text{WORK}}\right) - \ell_{\text{ADJ}}\left(\widetilde{\theta}_{\text{WORK}}\right)\right\}$$

$$\text{where} \quad c = \frac{\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)' \left[\Delta\mathbf{H}^{-1}_{ADJ}\Delta'\right]^{-1}\left(\Delta\hat{\theta}_{WORK} - \delta_0\right)}{\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)' \hat{\mathbf{H}}_{ADJ}\left(\hat{\theta}_{WORK} - \tilde{\theta}_{WORK}\right)}$$

- Λ^*_{ADJ} needs only estimates from working likelihood.
- **Details:** Chandler & Bate, *Biometrika*, 2007.

Other applications

- Not restricted to clustered data — applicable in principle whenever 'working' likelihood is used e.g. inference in 'wrong but useful' models (**NB** mis-specification of model *or* likelihood)

Other applications

- Not restricted to clustered data — applicable in principle whenever 'working' likelihood is used e.g. inference in 'wrong but useful' models (**NB** mis-specification of model *or* likelihood)
- Approach not restricted to likelihood-based inference — applicable whenever:
  - Estimation is done by optimising some objective function
  - Resulting estimating equations are (asymptotically) unbiased
  - Robust (and reliable) covariance matrix estimator is available

Other applications

- Not restricted to clustered data — applicable in principle whenever 'working' likelihood is used e.g. inference in 'wrong but useful' models (**NB** mis-specification of model *or* likelihood)
- Approach not restricted to likelihood-based inference — applicable whenever:
  - Estimation is done by optimising some objective function
  - Resulting estimating equations are (asymptotically) unbiased
  - Robust (and reliable) covariance matrix estimator is available
- **Example:** generalised method of moments — $\hat{\theta} = \arg\min_\theta S(\theta; \mathbf{y})$, where:
  - $S(\theta; \mathbf{y}) = \sum_{r=1}^{p} w_r \left[T_r(\mathbf{y}) - \tau_r(\theta)\right]^2$
  - $\{T_r(\mathbf{y}) : r = 1, \ldots, p\}$ are statistics (e.g. sample moments)
  - $\tau_r(\theta) = \mathrm{E}_\theta\left[T_r(\mathbf{y})\right]$ $(r = 1, \ldots, p)$.
  - $\{w_r : r = 1, \ldots, p\}$ are weights (independent of θ).

1. Problem statement
2. Standard asymptotics for mis-specified likelihoods
3. Adjusting the working log-likelihood
4. Open questions

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?**
  Requirements (cf bivariate normal example):

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?**
  Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?**
  Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest
  - $\left|\hat{\theta}_{\mathrm{WORK}} - \hat{\theta}_{\mathrm{FULL}}\right|$ is ‘small enough’ i.e. $\hat{\theta}_{\mathrm{WORK}}$ is efficient

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?** Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest
  - $\left|\hat{\theta}_{\mathrm{WORK}} - \hat{\theta}_{\mathrm{FULL}}\right|$ is ‘small enough’ i.e. $\hat{\theta}_{\mathrm{WORK}}$ is efficient

  NB conditions known for ‘independence’ working log-likelihood in Gaussian linear models — result given by Watson (*Biometrika*, 1972).

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?** Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest
  - $\left|\hat{\theta}_{\mathrm{WORK}} - \hat{\theta}_{\mathrm{FULL}}\right|$ is ‘small enough’ i.e. $\hat{\theta}_{\mathrm{WORK}}$ is efficient

  NB conditions known for ‘independence’ working log-likelihood in Gaussian linear models — result given by Watson (*Biometrika*, 1972).

- **When is ℓ_{ADJ} a bona fide *useful* profile log-likelihood for θ?** Could then argue that adjustment gives full likelihood-based inference under ‘convenient’ model for higher-order structure.

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?** Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest
  - $\left|\hat{\theta}_{\text{WORK}} - \hat{\theta}_{\text{FULL}}\right|$ is ‘small enough’ i.e. $\hat{\theta}_{\text{WORK}}$ is efficient

  NB conditions known for ‘independence’ working log-likelihood in Gaussian linear models — result given by Watson (*Biometrika*, 1972).
- **When is ℓ_{ADJ} a bona fide *useful* profile log-likelihood for θ?** Could then argue that adjustment gives full likelihood-based inference under ‘convenient’ model for higher-order structure.
  - To be useful, need to maintain interpretation of θ

Open questions (1)

- **When does adjustment recover profile log-likelihood for θ asymptotically?** Requirements (cf bivariate normal example):
  - ℓ_{WORK} (approximately) quadratic in region of interest
  - $\left|\hat{\theta}_{\text{WORK}} - \hat{\theta}_{\text{FULL}}\right|$ is ‘small enough’ i.e. $\hat{\theta}_{\text{WORK}}$ is efficient

  NB conditions known for ‘independence’ working log-likelihood in Gaussian linear models — result given by Watson (*Biometrika*, 1972).
- **When is ℓ_{ADJ} a bona fide *useful* profile log-likelihood for θ?** Could then argue that adjustment gives full likelihood-based inference under ‘convenient’ model for higher-order structure.
  - To be useful, need to maintain interpretation of θ
  - Requirement seems to be existence of joint densities $\left\{f_j\left(\mathbf{y}_j | \mathcal{C}_j; \theta, \alpha\right)\right\}$ for which adjustment recovers profile log-likelihood for θ (asymptotically?)

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**
  - In sequence of nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_M$, comparison of (e.g.) $\mathcal{M}_1$ and $\mathcal{M}_2$ could be based on adjusted profiles from $\mathcal{M}_2, \mathcal{M}_3, \ldots$ or $\mathcal{M}_M$ — each model will give different $\hat{\mathbf{H}}_{\mathrm{ADJ}}$, hence adjustment is model-dependent.

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**
  - In sequence of nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_M$, comparison of (e.g.) $\mathcal{M}_1$ and $\mathcal{M}_2$ could be based on adjusted profiles from $\mathcal{M}_2, \mathcal{M}_3, \ldots$ or $\mathcal{M}_M$ — each model will give different $\hat{\mathbf{H}}_{\mathrm{ADJ}}$, hence adjustment is model-dependent.
  - Can base all inference on profiles derived from 'maximal' model $\mathcal{M}_M$ if specified in advance — but not always feasible.

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**
  - In sequence of nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_M$, comparison of (e.g.) $\mathcal{M}_1$ and $\mathcal{M}_2$ could be based on adjusted profiles from $\mathcal{M}_2, \mathcal{M}_3, \ldots$ or $\mathcal{M}_M$ — each model will give different $\hat{\mathbf{H}}_{\mathrm{ADJ}}$, hence adjustment is model-dependent.
  - Can base all inference on profiles derived from 'maximal' model $\mathcal{M}_M$ if specified in advance — but not always feasible.
  - Possible alternative: derive $\hat{\mathbf{H}}_{\mathrm{ADJ}}$ for 'saturated' model (cf deviance for GLMs) — but asymptotic arguments then fail except in special situations e.g. iid clusters.

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**
  - In sequence of nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_M$, comparison of (e.g.) $\mathcal{M}_1$ and $\mathcal{M}_2$ could be based on adjusted profiles from $\mathcal{M}_2, \mathcal{M}_3, \ldots$ or $\mathcal{M}_M$ — each model will give different $\hat{\mathbf{H}}_{\mathrm{ADJ}}$, hence adjustment is model-dependent.
  - Can base all inference on profiles derived from 'maximal' model $\mathcal{M}_M$ if specified in advance — but not always feasible.
  - Possible alternative: derive $\hat{\mathbf{H}}_{\mathrm{ADJ}}$ for 'saturated' model (cf deviance for GLMs) — but asymptotic arguments then fail except in special situations e.g. iid clusters.
  - Other alternatives?

Open questions (2)

- **Adjustment is model-dependent: can this be overcome?**
  - In sequence of nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_M$, comparison of (e.g.) $\mathcal{M}_1$ and $\mathcal{M}_2$ could be based on adjusted profiles from $\mathcal{M}_2, \mathcal{M}_3, \ldots$ or $\mathcal{M}_M$ — each model will give different $\hat{\mathbf{H}}_{\mathrm{ADJ}}$, hence adjustment is model-dependent.
  - Can base all inference on profiles derived from 'maximal' model $\mathcal{M}_M$ if specified in advance — but not always feasible.
  - Possible alternative: derive $\hat{\mathbf{H}}_{\mathrm{ADJ}}$ for 'saturated' model (cf deviance for GLMs) — but asymptotic arguments then fail except in special situations e.g. iid clusters.
  - Other alternatives?

ANY QUESTIONS / SUGGESTIONS?