Parametric or Nonparametric: The Focused Information Criterion Approach

Martin Jullum

Nils Lid Hjort

University of Oslo

martinju@math.uio.no

March 4, 2014

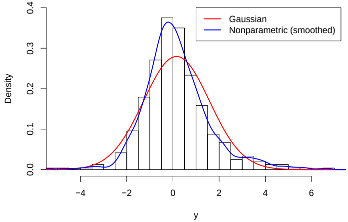
Outline

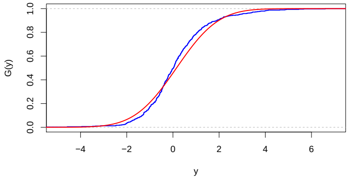
Model selection

- Unknown underlying distribution G for data $Y_1, \ldots, Y_n$
- Parametric approach
  - Restrict G to a specific parametric family F_θ
  - Estimate parameter θ (e.g. by ML) and use $\widehat{G} = F_{\widehat{\theta}}$
- Nonparametric approach
  - Let the data speak for themselves, no structural assumptions
  - Estimate G by ecdf: $\widehat{G}_n(y) = n^{-1}\#\{Y_i \leq y\}$.
  - Smoother versions: Local kernel smoothers
- Several appropriate models – which one should we trust?

Model selection

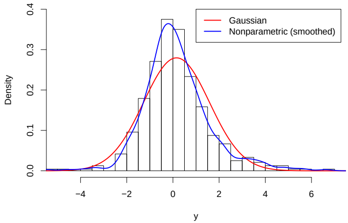
- Unknown underlying distribution G for data $Y_1, \ldots, Y_n$
- Parametric approach
  - Restrict G to a specific parametric family F_θ
  - Estimate parameter θ (e.g. by ML) and use $\widehat{G} = F_{\widehat{\theta}}$
- Nonparametric approach
  - Let the data speak for themselves, no structural assumptions
  - Estimate G by ecdf: $\widehat{G}_n(y) = n^{-1}\#\{Y_i \leq y\}$.
  - Smoother versions: Local kernel smoothers
- Several appropriate models – which one should we trust?

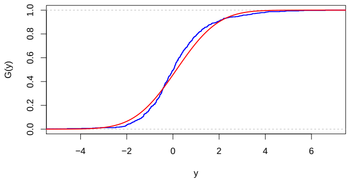




Model selection

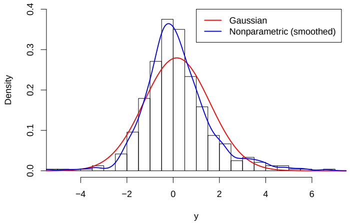
- Unknown underlying distribution G for data $Y_1, \ldots, Y_n$
- Parametric approach
    - Restrict G to a specific parametric family F_θ
    - Estimate parameter θ (e.g. by ML) and use $\widehat{G} = F_{\widehat{\theta}}$
- Nonparametric approach
    - Let the data speak for themselves, no structural assumptions
    - Estimate G by ecdf: $\widehat{G}_n(y) = n^{-1}\#\{Y_i \leq y\}$.
    - Smoother versions: Local kernel smoothers
- Several appropriate models – which one should we trust?

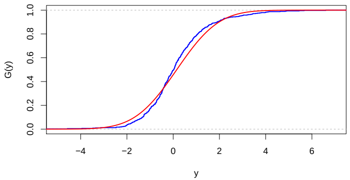




Model selection

- Unknown underlying distribution G for data $Y_1, \ldots, Y_n$
- Parametric approach
  - Restrict G to a specific parametric family F_θ
  - Estimate parameter θ (e.g. by ML) and use $\widehat{G} = F_{\widehat{\theta}}$
- Nonparametric approach
  - Let the data speak for themselves, no structural assumptions
  - Estimate G by ecdf: $\widehat{G}_n(y) = n^{-1}\#\{Y_i \leq y\}$.
  - Smoother versions: Local kernel smoothers
- Several appropriate models – which one should we trust?





Classical model selection

- Information criteria
  - AIC$= 2$ log-likelihood$_{\text{max}} - 2\dim(\theta)$
  - BIC$= 2$ log-likelihood$_{\text{max}} - (\log n)\dim(\theta)$
  - DIC, GIC, TIC, etc...
- Select the model optimizing the information criterion
- Cannot handle nonparametrics

Classical model selection

- Information criteria
  - AIC$= 2\ \text{log-likelihood}_{\text{max}} - 2\dim(\theta)$
  - BIC$= 2\ \text{log-likelihood}_{\text{max}} - (\log n)\dim(\theta)$
  - DIC, GIC, TIC, etc...
- Select the model optimizing the information criterion
- Cannot handle nonparametrics

Classical model selection

- Information criteria
  - AIC$= 2$ log-likelihood$_{\text{max}} - 2\,\text{dim}(\theta)$
  - BIC$= 2$ log-likelihood$_{\text{max}} - (\log n)\,\text{dim}(\theta)$
  - DIC, GIC, TIC, etc...
- Select the model optimizing the information criterion
- Cannot handle nonparametrics

Classical model selection

- Information criteria
  - AIC$= 2$ log-likelihood$_{\text{max}} - 2\,\text{dim}(\theta)$
  - BIC$= 2$ log-likelihood$_{\text{max}} - (\log n)\,\text{dim}(\theta)$
  - DIC, GIC, TIC, etc...
- Select the model optimizing the information criterion
- Cannot handle nonparametrics

Goodness of fit measures

- Cramér–von Mises: $\int(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \mathrm{d}F(y;\widehat{\theta})$
- Kolmogorov–Smirnov: $\sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$
- Categorical data: Pearson's chi-squared: $\sum_{j=1}^{k} \frac{(N_j - nf_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Select F if the goodness of fit measure $< \kappa_\alpha$ with significance level α
- Problem: Need to choose significance level α
  - Why 0.05 or 0.01? Why not 0.03682?

Goodness of fit measures

- Cramér–von Mises: $\int(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \mathrm{d}F(y;\widehat{\theta})$
- Kolmogorov–Smirnov: $\sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$
- Categorical data: Pearson's chi-squared: $\sum_{j=1}^{k} \frac{(N_j - nf_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Select F if the goodness of fit measure $< \kappa_\alpha$ with significance level α
- Problem: Need to choose significance level α
  - Why 0.05 or 0.01? Why not 0.03682?

Goodness of fit measures

- Cramér–von Mises: $\int(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \mathrm{d}F(y;\widehat{\theta})$
- Kolmogorov–Smirnov: $\sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$
- Categorical data: Pearson's chi-squared: $\sum_{j=1}^{k} \frac{(N_j - nf_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Select F if the goodness of fit measure $< \kappa_\alpha$ with significance level α
- Problem: Need to choose significance level α
  - Why 0.05 or 0.01? Why not 0.03682?

Goodness of fit measures

- Cramér–von Mises: $\int (\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \mathrm{d}F(y;\widehat{\theta})$
- Kolmogorov–Smirnov: $\sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$
- Categorical data: Pearson's chi-squared: $\sum_{j=1}^{k} \frac{(N_j - nf_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Select F if the goodness of fit measure $< \kappa_\alpha$ with significance level α
- Problem: Need to choose significance level α
  - Why 0.05 or 0.01? Why not 0.03682?

Goodness of fit measures

- Cramér–von Mises: $\int(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \mathrm{d}F(y;\widehat{\theta})$
- Kolmogorov–Smirnov: $\sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$
- Categorical data: Pearson's chi-squared: $\sum_{j=1}^{k} \frac{(N_j - nf_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Select F if the goodness of fit measure $< \kappa_\alpha$ with significance level α
- Problem: Need to choose significance level α
    - Why 0.05 or 0.01? Why not 0.03682?

Parametrics or nonparametrics

- No good criterion for model selection among parametrics and nonparametrics
- Different models have strengths and weaknesses on different parts of the data space
- Why you are doing the analysis should reflect your choice of model

Goal

Create a focused or interest driven model selection criterion for selection among a set of parametric and nonparametric models

Parametrics or nonparametrics

- No good criterion for model selection among parametrics and nonparametrics
- Different models have strengths and weaknesses on different parts of the data space
- Why you are doing the analysis should reflect your choice of model

Goal

Create a focused or interest driven model selection criterion for selection among a set of parametric and nonparametric models

Parametrics or nonparametrics

- No good criterion for model selection among parametrics and nonparametrics
- Different models have strengths and weaknesses on different parts of the data space
- Why you are doing the analysis should reflect your choice of model

Goal

Create a focused or interest driven model selection criterion for selection among a set of parametric and nonparametric models

Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A functional μ of the distribution G: $\mu(G)$

Examples

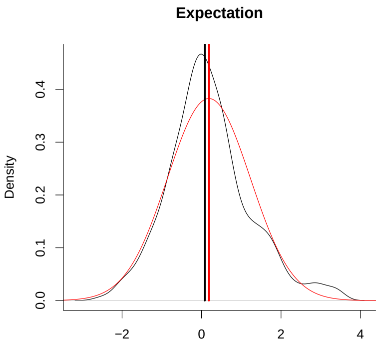
- Expectation: $\mu(G) = \mathrm{E}_G[Y_i]$
- $Pr\{Y_i > 2\}$: $\mu(G) = 1 - G(2)$
- Interquartile range: $\mu(G) = G^{-1}(3/4) - G^{-1}(1/4)$

Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A functional μ of the distribution G: $\mu(G)$

Examples

- Expectation: $\mu(G) = \mathrm{E}_G[Y_i]$
- $Pr\{Y_i > 2\}$: $\mu(G) = 1 - G(2)$
- Interquartile range: $\mu(G) = G^{-1}(3/4) - G^{-1}(1/4)$

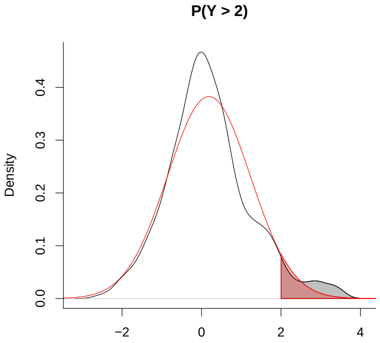


Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A functional μ of the distribution G: $\mu(G)$

Examples

- Expectation: $\mu(G) = \mathrm{E}_G[Y_i]$
- $Pr\{Y_i > 2\}$: $\mu(G) = 1 - G(2)$
- Interquartile range: $\mu(G) = G^{-1}(3/4) - G^{-1}(1/4)$

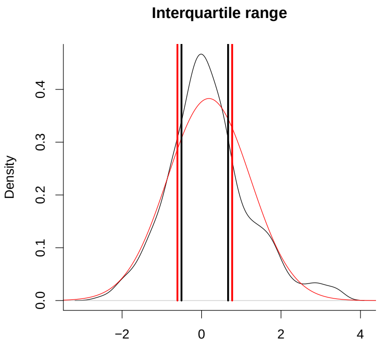


Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A functional μ of the distribution G: $\mu(G)$

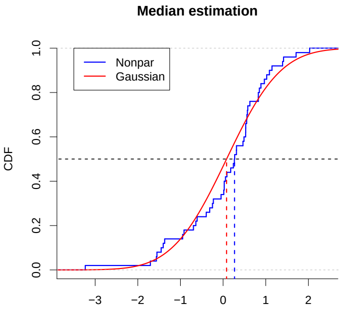
Examples

- Expectation: $\mu(G) = \mathrm{E}_G[Y_i]$
- $Pr\{Y_i > 2\}$: $\mu(G) = 1 - G(2)$
- Interquartile range: $\mu(G) = G^{-1}(3/4) - G^{-1}(1/4)$



Simple illustration

- I.i.d. univariate observations $Y = (Y_1, \ldots, Y_n)^{\mathrm{t}}$
- Focus parameter of interest: $\mu(G) = G^{-1}(1/2)$, the median of the unknown data generating distribution G
- Gaussian or nonparametric?
- Nonparametric sample median $\mathrm{med}(Y)$ or the Gaussian alternative $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$?



Criterion idea

- Performance measure: Mean squared error (mse) of the focus parameter estimator $\widehat{\mu}_M$

$$\mathsf{E}\left[(\widehat{\mu}_M - \mu_{\text{true}})^2\right] = \text{bias}^2(\widehat{\mu}_M) + \text{Var}(\widehat{\mu}_M)$$

Basic idea

- For each candidate model M with estimator $\widehat{\mu}_M = \mu(\widehat{G}_M)$:
  Estimate the mean squared error (mse) as squared bias + variance:

$$\text{FIC}(M) = \widehat{\text{mse}}(\widehat{\mu}_M) = \widehat{\text{bias}^2}(\widehat{\mu}_M) + \widehat{\text{Var}}(\widehat{\mu}_M)$$

- Choose the model and estimator with the smallest estimated mse

Criterion idea

- Performance measure: Mean squared error (mse) of the focus parameter estimator $\widehat{\mu}_M$

$$\mathsf{E}\left[(\widehat{\mu}_M - \mu_{\mathrm{true}})^2\right] = \mathrm{bias}^2(\widehat{\mu}_M) + \mathrm{Var}(\widehat{\mu}_M)$$

Basic idea

- For each candidate model M with estimator $\widehat{\mu}_M = \mu(\widehat{G}_M)$: Estimate the mean squared error (mse) as squared bias + variance:

$$\mathsf{FIC}(M) = \widehat{\mathrm{mse}}(\widehat{\mu}_M) = \widehat{\mathrm{bias}^2}(\widehat{\mu}_M) + \widehat{\mathrm{Var}}(\widehat{\mu}_M)$$

- Choose the model and estimator with the smallest estimated mse

I.i.d derivation – notation

I.i.d. data $Y_1, \ldots, Y_n$ from an unknown distribution G

Focus parameter $\mu = \mu(\cdot)$

- True value: $\mu_{\mathrm{true}} = \mu(G)$
- Nonparametric estimator: $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$
- Parametric estimators: $\widehat{\mu}_{\mathrm{pm}} = \mu(F_{\widehat{\theta}}) = \mu_F(\widehat{\theta})$
- Parametric least false value: $\mu_{0,\mathrm{pm}} = \mu(F_{\theta_0}) = \mu_F(\theta_0)$

I.i.d derivation – notation

I.i.d. data $Y_1, \ldots, Y_n$ from an unknown distribution G

Focus parameter $\mu = \mu(\cdot)$

- True value: $\mu_{\mathrm{true}} = \mu(G)$
- Nonparametric estimator: $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$
- Parametric estimators: $\widehat{\mu}_{\mathrm{pm}} = \mu(F_{\widehat{\theta}}) = \mu_F(\widehat{\theta})$
- Parametric least false value: $\mu_{0,\mathrm{pm}} = \mu(F_{\theta_0}) = \mu_F(\theta_0)$

I.i.d derivation – notation

I.i.d. data $Y_1, \ldots, Y_n$ from an unknown distribution G

Focus parameter $\mu = \mu(\cdot)$

- True value: $\mu_{\rm true} = \mu(G)$
- Nonparametric estimator: $\widehat{\mu}_{\rm np} = \mu(\widehat{G}_n)$
- Parametric estimators: $\widehat{\mu}_{\rm pm} = \mu(F_{\widehat{\theta}}) = \mu_F(\widehat{\theta})$
- Parametric least false value: $\mu_{0,\rm pm} = \mu(F_{\theta_0}) = \mu_F(\theta_0)$

I.i.d derivation – notation

I.i.d. data $Y_1, \ldots, Y_n$ from an unknown distribution G

Focus parameter $\mu = \mu(\cdot)$

- True value: $\mu_{\rm true} = \mu(G)$
- Nonparametric estimator: $\widehat{\mu}_{\rm np} = \mu(\widehat{G}_n)$
- Parametric estimators: $\widehat{\mu}_{\rm pm} = \mu(F_{\widehat{\theta}}) = \mu_F(\widehat{\theta})$
- Parametric least false value: $\mu_{0,\rm pm} = \mu(F_{\theta_0}) = \mu_F(\theta_0)$

I.i.d derivation – notation

I.i.d. data $Y_1, \ldots, Y_n$ from an unknown distribution G

Focus parameter $\mu = \mu(\cdot)$

- True value: $\mu_{\mathrm{true}} = \mu(G)$
- Nonparametric estimator: $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$
- Parametric estimators: $\widehat{\mu}_{\mathrm{pm}} = \mu(F_{\widehat{\theta}}) = \mu_F(\widehat{\theta})$
- Parametric least false value: $\mu_{0,\mathrm{pm}} = \mu(F_{\theta_0}) = \mu_F(\theta_0)$

Basic setup

Under weak regularity conditions (omitted) we have

- Nonparametric: $\widehat{\mu}_{\mathrm{np}} = \mu_{\mathrm{true}} + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_\mu(Y_i; G) + o_p(n^{-1/2})$
  - $\mathrm{IF}_\mu(y; G)$ the influence function of the functional $\mu(G)$: $\frac{\partial}{\partial\lambda}\mu(H + \lambda(\delta_y - H))\big|_{\lambda=0}$
  - δ_y is point mass at y
- Parametric: $\widehat{\theta} = \theta_0 + J^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i; \theta_0) + o_p(n^{-1/2})$
  - $U(y; \theta) = \partial \log f(y; \theta)/\partial\theta$ is the score function
  - $J = -\mathrm{E}\left[\partial U(Y_i; \theta_0)/\partial\theta\right]$ is the information matrix
  - $K = \mathrm{Var}(U(Y_i; \theta_0))$
- $U(Y_i; \theta_0)$,$\mathrm{IF}_\mu(Y_i; G)$ have zero means and finite variances

Basic setup

Under weak regularity conditions (omitted) we have

- Nonparametric: $\widehat{\mu}_{\mathrm{np}} = \mu_{\mathrm{true}} + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_\mu(Y_i; G) + o_p(n^{-1/2})$
  - $\mathrm{IF}_\mu(y; G)$ the influence function of the functional $\mu(G)$: $\frac{\partial}{\partial\lambda}\mu(H + \lambda(\delta_y - H))\big|_{\lambda=0}$
  - δ_y is point mass at y
- Parametric: $\widehat{\theta} = \theta_0 + J^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i; \theta_0) + o_p(n^{-1/2})$
  - $U(y; \theta) = \partial \log f(y; \theta)/\partial\theta$ is the score function
  - $J = -\mathrm{E}\left[\partial U(Y_i; \theta_0)/\partial\theta\right]$ is the information matrix
  - $K = \mathrm{Var}(U(Y_i; \theta_0))$
- $U(Y_i; \theta_0)$,$\mathrm{IF}_\mu(Y_i; G)$ have zero means and finite variances

Basic setup

Under weak regularity conditions (omitted) we have

- Nonparametric: $\widehat{\mu}_{\mathrm{np}} = \mu_{\mathrm{true}} + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_{\mu}(Y_i; G) + o_p(n^{-1/2})$
  - $\mathrm{IF}_{\mu}(y; G)$ the influence function of the functional $\mu(G)$: $\frac{\partial}{\partial\lambda}\mu(H + \lambda(\delta_y - H))\big|_{\lambda=0}$
  - δ_y is point mass at y
- Parametric: $\widehat{\theta} = \theta_0 + J^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i; \theta_0) + o_p(n^{-1/2})$
  - $U(y; \theta) = \partial \log f(y; \theta)/\partial\theta$ is the score function
  - $J = -\mathrm{E}\left[\partial U(Y_i; \theta_0)/\partial\theta\right]$ is the information matrix
  - $K = \mathrm{Var}(U(Y_i; \theta_0))$
- $U(Y_i; \theta_0)$,$\mathrm{IF}_{\mu}(Y_i; G)$ have zero means and finite variances

Basic setup

Under weak regularity conditions (omitted) we have

- Nonparamertic: $\widehat{\mu}_{\mathrm{np}} = \mu_{\mathrm{true}} + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_{\mu}(Y_i; G) + o_p(n^{-1/2})$
  - $\mathrm{IF}_{\mu}(y; G)$ the influence function of the functional $\mu(G)$: $\frac{\partial}{\partial\lambda}\mu(H + \lambda(\delta_y - H))\big|_{\lambda=0}$
  - δ_y is point mass at y
- Parametric: $\widehat{\theta} = \theta_0 + J^{-1}\frac{1}{n}\sum_{i=1}^{n} U(Y_i; \theta_0) + o_p(n^{-1/2})$
  - $U(y;\theta) = \partial \log f(y;\theta)/\partial\theta$ is the score function
  - $J = -\mathrm{E}\left[\partial U(Y_i;\theta_0)/\partial\theta\right]$ is the information matrix
  - $K = \mathrm{Var}(U(Y_i;\theta_0))$
- $U(Y_i;\theta_0)$,$\mathrm{IF}_{\mu}(Y_i; G)$ have zero means and finite variances

Key limiting distribution

CLT+Slutsky+delta method gives

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\rm np}-\mu_{\rm true}\\ \widehat{\mu}_{\rm pm}-\mu_{0,\rm pm}\end{pmatrix}\xrightarrow{L} N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}V_{\rm np}(G) & V_{\rm c}(G,\theta_0)\\ V_{\rm c}(G,\theta_0) & V_{\rm pm}(\theta_0)\end{pmatrix}\right)$$

where

- $V_{\rm np}(G)=\mathrm{E}\left[\mathrm{IF}_\mu(Y_i;G)\right]=\int \mathrm{IF}_\mu(y;G)^2\mathrm{d}G(y)$
- $V_{\rm pm}(\theta_0)=c^{\rm t}J^{-1}KJ^{-1}c$
- $V_{\rm c}(G,\theta_0)=c^{\rm t}J^{-1}d$
- $c=\partial\mu(F_{\theta_0})/\partial\theta$
- $d=\mathrm{Cov}(U(Y_i;\theta_0),\mathrm{IF}_\mu(Y_i;G))=\int U(y;\theta_0)\mathrm{IF}_\mu(y;G)\,\mathrm{d}G(y)$

Mse approximation

- $\mathrm{mse}(\widehat{\mu}_M) = \mathsf{E}\,[(\widehat{\mu}_M - \mu_{\mathrm{true}})^2] = \mathrm{bias}^2(\widehat{\mu}_M) + \mathrm{Var}(\widehat{\mu}_M)$
- From the limiting distribution:

$$\text{Nonparametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{np}}) \approx 0 + \frac{1}{n} V_{\mathrm{np}}(G)$$

$$\text{Parametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{pm}}) \approx b(\theta_0, G)^2 + \frac{1}{n} V_{\mathrm{pm}}(\theta_0)$$

where $b(\theta_0, G) = \mu_{0,\mathrm{pm}} - \mu_{\mathrm{true}} = \mu(F_{\theta_0}) - \mu(G)$

Mse approximation

- $\mathrm{mse}(\widehat{\mu}_M) = \mathsf{E}\,[(\widehat{\mu}_M - \mu_{\mathrm{true}})^2] = \mathrm{bias}^2(\widehat{\mu}_M) + \mathrm{Var}(\widehat{\mu}_M)$
- From the limiting distribution:

$$\text{Nonparametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{np}}) \approx 0 + \frac{1}{n} V_{\mathrm{np}}(G)$$

$$\text{Parametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{pm}}) \approx b(\theta_0, G)^2 + \frac{1}{n} V_{\mathrm{pm}}(\theta_0)$$

where $b(\theta_0, G) = \mu_{0,\mathrm{pm}} - \mu_{\mathrm{true}} = \mu(F_{\theta_0}) - \mu(G)$

Mse estimation

Insert empirical analogues for unknown quantities:
$\widehat{\theta}$ for θ_0 and $\widehat{G}_n$ for G

FIC scheme

$$\text{Nonparametric: } \mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n} V_{\mathrm{np}}(\widehat{G}_n)$$

$$\text{Parametric: } \mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left\{0, b(\widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(\widehat{\theta}, \widehat{G}_n)\right\} + \frac{1}{n} V_{\mathrm{pm}}(\widehat{\theta})$$

The criterion selects the model with the smallest FIC score

- $b(\widehat{\theta}, \widehat{G}_n) = \widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}}; \quad V_{\mathrm{b}} = V_{\mathrm{pm}} + V_{\mathrm{np}} - 2V_{\mathrm{c}}$
- $b(\widehat{\theta}, \widehat{G}_n)^2$ overestimates $b(\theta_0, G)^2$:

$$\begin{aligned} E[b(\widehat{\theta}, \widehat{G}_n)^2] &= (E[b(\widehat{\theta}, \widehat{G}_n)])^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n)) \\ &\approx b(\theta_0, G)^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n)) \end{aligned}$$

Mse estimation

Insert empirical analogues for unknown quantities:
$\widehat{\theta}$ for θ_0 and $\widehat{G}_n$ for G

FIC scheme

$$\text{Nonparametric: } \mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n} V_{\mathrm{np}}(\widehat{G}_n)$$

$$\text{Parametric: } \mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left\{0, b(\widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(\widehat{\theta}, \widehat{G}_n)\right\} + \frac{1}{n} V_{\mathrm{pm}}(\widehat{\theta})$$

The criterion selects the model with the smallest FIC score

- $b(\widehat{\theta}, \widehat{G}_n) = \widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}}; \quad V_{\mathrm{b}} = V_{\mathrm{pm}} + V_{\mathrm{np}} - 2V_{\mathrm{c}}$
- $b(\widehat{\theta}, \widehat{G}_n)^2$ overestimates $b(\theta_0, G)^2$:

$$\begin{aligned} E[b(\widehat{\theta}, \widehat{G}_n)^2] &= (E[b(\widehat{\theta}, \widehat{G}_n)])^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n)) \\ &\approx b(\theta_0, G)^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n)) \end{aligned}$$

Mse estimation

Insert empirical analogues for unknown quantities:
$\widehat{\theta}$ for θ_0 and $\widehat{G}_n$ for G

FIC scheme

Nonparametric: $\mathrm{FIC}(\widehat{\mu}_{\mathrm{np}}) = \frac{1}{n} V_{\mathrm{np}}(\widehat{G}_n)$

Parametric: $\mathrm{FIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left\{0, b(\widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(\widehat{\theta}, \widehat{G}_n)\right\} + \frac{1}{n} V_{\mathrm{pm}}(\widehat{\theta})$

The criterion selects the model with the smallest FIC score

- $b(\widehat{\theta}, \widehat{G}_n) = \widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}}$; $\quad V_{\mathrm{b}} = V_{\mathrm{pm}} + V_{\mathrm{np}} - 2V_{\mathrm{c}}$
- $b(\widehat{\theta}, \widehat{G}_n)^2$ overestimates $b(\theta_0, G)^2$:

$$
\begin{aligned}
E[b(\widehat{\theta}, \widehat{G}_n)^2] &= (E[b(\widehat{\theta}, \widehat{G}_n)])^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n)) \\
&\approx b(\theta_0, G)^2 + \mathrm{Var}(b(\widehat{\theta}, \widehat{G}_n))
\end{aligned}
$$

Illustration: Running through the cdf

- Sample ($n = 100$) of school averaged grade data from the math part of SAT in Pennsylvania 2009
- Sequentially focus on $\mu(y) = G(y) = Pr\{Y_i \leq y\}$

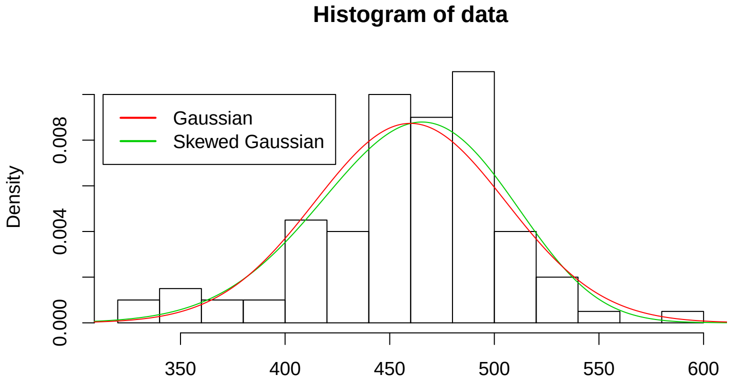


Illustration: Running through the cdf

- Sample ($n = 100$) of school averaged grade data from the math part of SAT in Pennsylvania 2009
- Sequentially focus on $\mu(y) = G(y) = Pr\{Y_i \leq y\}$

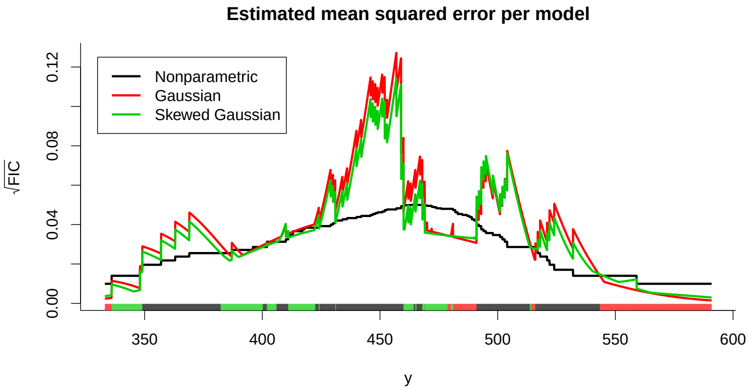
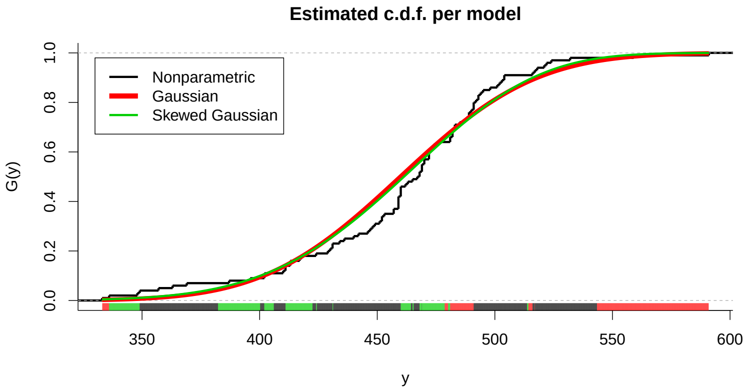


Illustration: Running through the cdf

- Sample ($n = 100$) of school averaged grade data from the math part of SAT in Pennsylvania 2009
- Sequentially focus on $\mu(y) = G(y) = Pr\{Y_i \leq y\}$



Averaged Focused Information Criterion (AFIC)

- Focus on a (weighted) set of focus parameters simultaneously
- Performance measure: risk $= \int \mathrm{E}\,[(\widehat{\mu}(t) - \mu_{\mathrm{true}}(t))^2]\,\mathrm{d}W(t)$, for some cumulative weight function W

AFIC scheme

$$\text{Nonparametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n} V_{\mathrm{np}}(t; \widehat{G}_n)\mathrm{d}W(t)$$

$$\text{Parametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left[0, \int \{b(t; \widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(t; \widehat{\theta}, \widehat{G}_n)\}\,\mathrm{d}W(t)\right] + \int \frac{1}{n} V_{\mathrm{pm}}(t; \widehat{\theta})\,\mathrm{d}W(t)$$

The criterion selects the model with the smallest AFIC score

- Estimate W empirically if it depends on unknown quantities

Averaged Focused Information Criterion (AFIC)

- Focus on a (weighted) set of focus parameters simultaneously
- Performance measure: risk $= \int \mathrm{E}\,[(\widehat{\mu}(t) - \mu_{\mathsf{true}}(t))^2]\,\mathrm{d}W(t)$, for some cumulative weight function W

AFIC scheme

$$\text{Nonparametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n} V_{\mathrm{np}}(t; \widehat{G}_n)\mathrm{d}W(t)$$

$$\text{Parametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left[0, \int \{b(t;\widehat{\theta},\widehat{G}_n)^2 - \frac{1}{n}V_{\mathrm{b}}(t;\widehat{\theta},\widehat{G}_n)\}\,\mathrm{d}W(t)\right] + \int \frac{1}{n} V_{\mathrm{pm}}(t;\widehat{\theta})\,\mathrm{d}W(t)$$

The criterion selects the model with the smallest AFIC score

- Estimate W empirically if it depends on unknown quantities

Averaged Focused Information Criterion (AFIC)

- Focus on a (weighted) set of focus parameters simultaneously
- Performance measure: risk $= \int \mathrm{E}\left[(\widehat{\mu}(t) - \mu_{\text{true}}(t))^2\right] \mathrm{d}W(t)$, for some cumulative weight function W

AFIC scheme

$$\text{Nonparametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n} V_{\mathrm{np}}(t; \widehat{G}_n) \mathrm{d}W(t)$$

$$\text{Parametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left[0, \int \left\{b(t; \widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(t; \widehat{\theta}, \widehat{G}_n)\right\} \mathrm{d}W(t)\right] + \int \frac{1}{n} V_{\mathrm{pm}}(t; \widehat{\theta}) \, \mathrm{d}W(t)$$

The criterion selects the model with the smallest AFIC score

- Estimate W empirically if it depends on unknown quantities

Averaged Focused Information Criterion (AFIC)

- Focus on a (weighted) set of focus parameters simultaneously
- Performance measure: risk $= \int \mathrm{E}\,[(\widehat{\mu}(t) - \mu_{\text{true}}(t))^2]\,\mathrm{d}W(t)$, for some cumulative weight function W

AFIC scheme

$$\text{Nonparametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n} V_{\mathrm{np}}(t; \widehat{G}_n)\mathrm{d}W(t)$$

$$\text{Parametric: } \mathrm{AFIC}(\widehat{\mu}_{\mathrm{pm}}) = \max\left[0, \int \{b(t; \widehat{\theta}, \widehat{G}_n)^2 - \frac{1}{n} V_{\mathrm{b}}(t; \widehat{\theta}, \widehat{G}_n)\}\,\mathrm{d}W(t)\right] + \int \frac{1}{n} V_{\mathrm{pm}}(t; \widehat{\theta})\,\mathrm{d}W(t)$$

The criterion selects the model with the smallest AFIC score

- Estimate W empirically if it depends on unknown quantities

Criteria properties

Robust mse estimation

- Consistent variance estimators
- Consistent and asymptotically unbiased* squared bias estimators
- Estimation outside model conditions

FIC asymptotics

- Parametrics biased ($\mu_{0,\mathrm{pm}} \neq \mu_{\mathrm{true}}$): $Pr\left\{\text{Select pm}\right\} \to 0$
- Parametrics correct: $Pr\left\{\text{Select pm}\right\} \to \chi_1^2(2) \approx 1 - 0.157$
- Implicit focused hypothesis test of parametric model

AFIC asymptotics depend on μ and W

Criteria properties

Robust mse estimation

- Consistent variance estimators
- Consistent and asymptotically unbiased* squared bias estimators
- Estimation outside model conditions

FIC asymptotics

- Parametrics biased ($\mu_{0,\mathrm{pm}} \neq \mu_{\mathrm{true}}$): $Pr\,\{\text{Select pm}\} \to 0$
- Parametrics correct: $Pr\,\{\text{Select pm}\} \to \chi^2_1(2) \approx 1 - 0.157$
- Implicit focused hypothesis test of parametric model

AFIC asymptotics depend on μ and W

Criteria properties

Robust mse estimation

- Consistent variance estimators
- Consistent and asymptotically unbiased* squared bias estimators
- Estimation outside model conditions

FIC asymptotics

- Parametrics biased ($\mu_{0,\mathrm{pm}} \neq \mu_{\mathrm{true}}$): $Pr\left\{\text{Select pm}\right\} \to 0$
- Parametrics correct: $Pr\left\{\text{Select pm}\right\} \to \chi_1^2(2) \approx 1 - 0.157$
- Implicit focused hypothesis test of parametric model

AFIC asymptotics depend on μ and W

AFIC curiosity 1

- Consider count data $(N_1, \ldots, N_k), (p_1, \ldots, p_k), \sum p_j = 1, \sum N_j = n$
- Focus on all p_j, with weight $1/p_j$
- Direct comparison between parametrics and nonparametrics reduces to

$$X_n = n \sum \frac{(\widehat{p}_{\mathrm{np},j} - \widehat{p}_{\mathrm{pm},j})^2}{\widehat{p}_{\mathrm{np},j}} \text{ vs. 2df}$$

- Implicit test level for test of pm true with df$=1, \ldots, 10$: $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035.$

AFIC curiosity 1

- Consider count data $(N_1, \ldots, N_k), (p_1, \ldots, p_k), \sum p_j = 1, \sum N_j = n$
- Focus on all p_j, with weight $1/p_j$
- Direct comparison between parametrics and nonparametrics reduces to

$$X_n = n \sum \frac{(\widehat{p}_{\mathrm{np},j} - \widehat{p}_{\mathrm{pm},j})^2}{\widehat{p}_{\mathrm{np},j}} \text{ vs. } 2\mathrm{df}$$

- Implicit test level for test of pm true with df$=1, \ldots, 10$: $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035.$

AFIC curiosity 1

- Consider count data $(N_1, \ldots, N_k), (p_1, \ldots, p_k), \sum p_j = 1, \sum N_j = n$
- Focus on all p_j, with weight $1/p_j$
- Direct comparison between parametrics and nonparametrics reduces to

$$X_n = n \sum \frac{(\widehat{p}_{\mathrm{np},j} - \widehat{p}_{\mathrm{pm},j})^2}{\widehat{p}_{\mathrm{np},j}} \text{ vs. } 2\mathrm{df}$$

- Implicit test level for test of pm true with df$=1, \ldots, 10$: $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035.$

AFIC curiosity 2

- Focus on the complete $G(y)$, weighted by $W(y) = F(y;\theta_0)$
- Direct comparison between parametrics and nonparametrics reduces to

$$\mathrm{CvM}_n = \int n(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \, \mathrm{d}F(y;\widehat{\theta}) \text{ vs. } \kappa$$

- If $G = F \sim N(\xi, \sigma^2)$: $Pr\{\text{Select pm}\} \to 1 - 0.062$
- Replacing $\mathrm{d}W(y)$ by $1\,\mathrm{d}y$ gives $Pr\{\text{Select pm}\} \to 1 - 0.049$

AFIC curiosity 2

- Focus on the complete $G(y)$, weighted by $W(y) = F(y;\theta_0)$
- Direct comparison between parametrics and nonparametrics reduces to

$$\mathrm{CvM}_n = \int n(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \,\mathrm{d}F(y;\widehat{\theta}) \text{ vs. } \kappa$$

- If $G = F \sim N(\xi, \sigma^2)$: $Pr\,\{\text{Select pm}\} \to 1 - 0.062$
- Replacing $\mathrm{d}W(y)$ by $1\,\mathrm{d}y$ gives $Pr\,\{\text{Select pm}\} \to 1 - 0.049$

AFIC curiosity 2

- Focus on the complete $G(y)$, weighted by $W(y) = F(y;\theta_0)$
- Direct comparison between parametrics and nonparametrics reduces to

$$\mathrm{CvM}_n = \int n(\widehat{G}_n(y) - F(y;\widehat{\theta}))^2 \,\mathrm{d}F(y;\widehat{\theta}) \text{ vs. } \kappa$$

- If $G = F \sim N(\xi, \sigma^2)$: $Pr\,\{\text{Select pm}\} \to 1 - 0.062$
- Replacing $\mathrm{d}W(y)$ by $1\,\mathrm{d}y$ gives $Pr\,\{\text{Select pm}\} \to 1 - 0.049$

AFIC curiosity 2

- Focus on the complete $G(y)$, weighted by $W(y) = F(y; \theta_0)$
- Direct comparison between parametrics and nonparametrics reduces to
$$\mathrm{CvM}_n = \int n(\widehat{G}_n(y) - F(y; \widehat{\theta}))^2 \, \mathrm{d}F(y; \widehat{\theta}) \text{ vs. } \kappa$$
- If $G = F \sim N(\xi, \sigma^2)$: $Pr\,\{\text{Select pm}\} \to 1 - 0.062$
- Replacing $\mathrm{d}W(y)$ by $1\,\mathrm{d}y$ gives $Pr\,\{\text{Select pm}\} \to 1 - 0.049$

Extensions to other data types

Same strategy, but more mathematics:

- Multivariate data, categorical data, time series, comparison across several populations
- Hazard rate models: Kaplan–Meier vs. parametrics or Nelson–Aalen vs. parametrics
- Cox regression vs. parametric regression

Other strategies:

- Standard regression setting
- Density estimation

Extensions to other data types

Same strategy, but more mathematics:

- Multivariate data, categorical data, time series, comparison across several populations
- Hazard rate models: Kaplan–Meier vs. parametrics or Nelson–Aalen vs. parametrics
- Cox regression vs. parametric regression

Other strategies:

- Standard regression setting
- Density estimation

Summary

- Focus driven model selection with a nonparametric alternative
- Rank models according to $\widehat{\mu}_M$'s estimated risk
- Robustifies parametric model selection by including the nonparametric candidate model
- AFIC allows several focus parameters to be handled simultaneously
- FIC and AFIC **Are** model selection schemes, but may also justify the use of different significance levels in hypothesis testing of models
- R-function automatically performing FIC for the handled situations
    - `folk.uio.no/martinju/FIC`