Parametric or nonparametric, that's the question

Martin Jullum

jullum@nr.no

Norwegian Computing Center

Joint work with Nils Lid Hjort

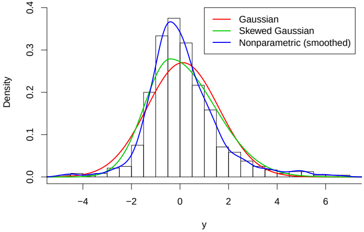
May 23, 2018

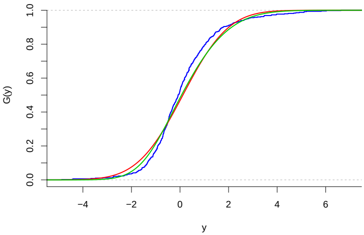
The question!

- Consider i.i.d. data $Y_1, \ldots, Y_n$ from some unknown underlying distribution G
- 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
- $K+1$ different appropriate models – which one should we trust?

The question!

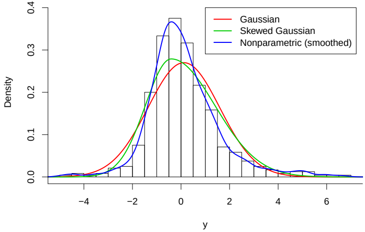
- Consider i.i.d. data $Y_1, \ldots, Y_n$ from some unknown underlying distribution G
- Parametric approach
  - Restrict G to a specific parametric family F_θ
  - Estimate parameter θ (e.g. by ML) and use $\widehat{G} = F_{\widehat{\theta}}$
- Nonparametric approach
  - Let the data speak for themselves,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

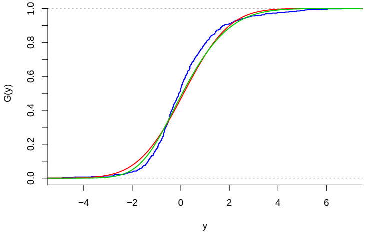




The question!

- Consider i.i.d. data $Y_1, \ldots, Y_n$ from some unknown underlying distribution G
- 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
- $K + 1$ different appropriate models – which one should we trust?





Classical parametric 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 parametric 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 parametric 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 parametric 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

Goodness of fit measures

- Measures the quality of the parametric fit compared to nonparametric
  - 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: $\sum_{j=1}^{k} \frac{(N_j - n f_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$
- Problematic to use for selecting between nonparametric and k parametric models
  - How to set thresholds for rejecting the parametric models?
  - How to penalize parametric model complexity?

Goodness of fit measures

- Measures the quality of the parametric fit compared to nonparametric
  - 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: $\sum_{j=1}^{k} \frac{(N_j - n f_j(\widehat{\theta}))^2}{f_j(\widehat{\theta})}$

- Problematic to use for selecting between nonparametric and k parametric models
  - How to set thresholds for rejecting the parametric models?
  - How to penalize parametric model complexity?

Parametrics or nonparametrics

- No appropriate criterion for model selection among parametrics and nonparametrics in the literature
- 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

Our proposed solution

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

Parametrics or nonparametrics

- No appropriate criterion for model selection among parametrics and nonparametrics in the literature
- 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

Our proposed solution

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

Parametrics or nonparametrics

- No appropriate criterion for model selection among parametrics and nonparametrics in the literature
- 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

Our proposed solution

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 quantity $\mu = T(G)$ mapping the distribution G to a scalar (multivariate situation later)

Examples

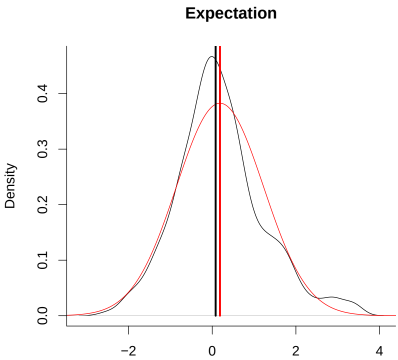
- Expectation: $\mu = \mathrm{E}_G\,(Y_i) = \int y\,\mathrm{d}G(y)$
- $\Pr\,(Y_i > 2)$: $\mu = 1 - G(2)$
- Interquartile range: $\mu = G^{-1}(3/4) - G^{-1}(1/4)$

Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A quantity $\mu = T(G)$ mapping the distribution G to a scalar (multivariate situation later)

Examples

- Expectation: $\mu = \mathsf{E}_G\,(Y_i) = \int y\,\mathrm{d}G(y)$
- Pr $(Y_i > 2)$: $\mu = 1 - G(2)$
- Interquartile range: $\mu = G^{-1}(3/4) - G^{-1}(1/4)$

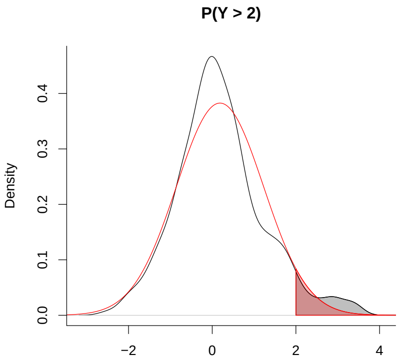


Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A quantity $\mu = T(G)$ mapping the distribution G to a scalar (multivariate situation later)

Examples

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

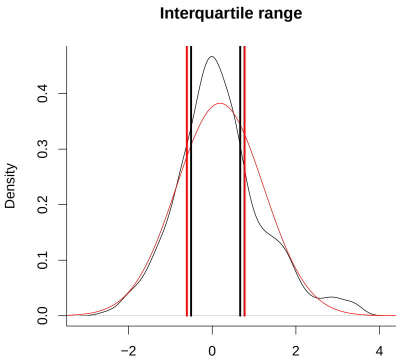


Focus parameter

- A general population quantity of interest – not(!) a model specific parameter
- A quantity $\mu = T(G)$ mapping the distribution G to a scalar (multivariate situation later)

Examples

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



Criterion idea

- Model selection problem $\rightarrow$ Best estimator for μ
  - Compare estimated performance of $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$ and all the different $\widehat{\mu}_{\mathrm{pm}} = \mu(F_{\widehat{\theta}})$
  - Select model with best performing estimator $\widehat{\mu}_M$
- Performance measure:

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

Basic idea: Focused information criterion (FIC)

- Estimate the mean squared error (mse) as squared bias + variance:

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

- Choose the model and estimator with the smallest FIC score

Criterion idea

- Model selection problem $\rightarrow$ Best estimator for μ
  - Compare estimated performance of $\widehat{\mu}_{\mathsf{np}} = \mu(\widehat{G}_n)$ and all the different $\widehat{\mu}_{\mathsf{pm}} = \mu(F_{\widehat{\theta}})$
  - Select model with best performing estimator $\widehat{\mu}_M$
- Performance measure:

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

Basic idea: Focused information criterion (FIC)

- 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 FIC score

FIC construction: Derivation details (I)

- Let (generically) $\mu_{0,\mathrm{pm}} = T(F_{\theta_0})$ be the least false focus parameter value

Under weak regularity conditions:

- Nonparametric: $\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}}) \to_d \mathrm{N}(0, v_{\mathrm{np}})$
- Parametrics: $\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_{0,\mathrm{pm}}) \to_d \mathrm{N}(0, v_{\mathrm{pm}})$

Large sample approximate risks

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

$$\text{Parametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{pm}}) \approx \widetilde{\mathrm{mse}}(\widehat{\mu}_{\mathrm{pm}}) = b^2 + \frac{1}{n} v_{\mathrm{pm}}$$

where $b = \mu_{0,\mathrm{pm}} - \mu_{\mathrm{true}}$

FIC construction: Derivation details (I)

- Let (generically) $\mu_{0,\mathrm{pm}} = T(F_{\theta_0})$ be the least false focus parameter value

Under weak regularity conditions:

- Nonparametric: $\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}}) \to_d \mathrm{N}(0, v_{\mathrm{np}})$
- Parametrics: $\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_{0,\mathrm{pm}}) \to_d \mathrm{N}(0, v_{\mathrm{pm}})$

Large sample approximate risks

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

$$\text{Parametric: } \mathrm{mse}(\widehat{\mu}_{\mathrm{pm}}) \approx \widetilde{\mathrm{mse}}(\widehat{\mu}_{\mathrm{pm}}) = b^2 + \frac{1}{n} v_{\mathrm{pm}}$$

where $b = \mu_{0,\mathrm{pm}} - \mu_{\mathrm{true}}$

FIC construction: Derivation details (II)

- The convergence in distribution also holds jointly:

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\rm np}-\mu_{\rm true}\\ \widehat{\mu}_{\rm pm}-\mu_{0,\rm pm}\end{pmatrix}\to_d \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}v_{\rm np} & v_{\rm c}\\ v_{\rm c} & v_{\rm pm}\end{pmatrix}\right)$$

- $\Rightarrow \sqrt{n}(\widehat{b}-b)\to_d \mathrm{N}(0,v_b)$, with $\widehat{b}=\widehat{\mu}_{\rm pm}-\widehat{\mu}_{\rm np}$ and $v_b=v_{\rm pm}+v_{\rm np}-2v_c$
- $\Rightarrow \mathsf{E}\{(\widehat{b})^2\}=(\mathsf{E}\{\widehat{b}\})^2+\mathsf{Var}(\widehat{b})\approx b^2+\frac{1}{n}v_b$

General estimates of risk components

- Squared parametric bias: b^2 estimated by $\max\{0,\widehat{b}^2-\frac{1}{n}\widehat{v}_b\}$
- Variances: $v_{\rm np}$ estimated by $\widehat{v}_{\rm np}$, and $v_{\rm pm}$ estimated by $\widehat{v}_{\rm pm}$

FIC scores

Nonparametric: $\mathrm{FIC}(\widehat{\mu}_{\rm np})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm np})=0+\dfrac{1}{n}\widehat{v}_{\rm np}$

Parametric: $\mathrm{FIC}(\widehat{\mu}_{\rm pm})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm pm})=\max\{0,(\widehat{b})^2-\dfrac{1}{n}\widehat{v}_{\rm b}\}+\dfrac{1}{n}\widehat{v}_{\rm pm}$

FIC construction: Derivation details (II)

- The convergence in distribution also holds jointly:

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\rm np}-\mu_{\rm true}\\ \widehat{\mu}_{\rm pm}-\mu_{0,\rm pm}\end{pmatrix}\to_d {\rm N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}v_{\rm np} & v_{\rm c}\\ v_{\rm c} & v_{\rm pm}\end{pmatrix}\right)$$

- $\Rightarrow \sqrt{n}(\widehat{b}-b)\to_d {\rm N}(0,v_b)$, with $\widehat{b}=\widehat{\mu}_{\rm pm}-\widehat{\mu}_{\rm np}$ and $v_b=v_{\rm pm}+v_{\rm np}-2v_c$
- $\Rightarrow \mathsf{E}\{(\widehat{b})^2\}=(\mathsf{E}\{\widehat{b}\})^2+\mathsf{Var}(\widehat{b})\approx b^2+\frac{1}{n}v_b$

General estimates of risk components

- Squared parametric bias: b^2 estimated by $\max\{0,\widehat{b}^2-\frac{1}{n}\widehat{v}_b\}$
- Variances: $v_{\rm np}$ estimated by $\widehat{v}_{\rm np}$, and $v_{\rm pm}$ estimated by $\widehat{v}_{\rm pm}$

FIC scores

Nonparametric: $\mathrm{FIC}(\widehat{\mu}_{\rm np})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm np})=0+\dfrac{1}{n}\widehat{v}_{\rm np}$

Parametric: $\mathrm{FIC}(\widehat{\mu}_{\rm pm})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm pm})=\max\{0,(\widehat{b})^2-\dfrac{1}{n}\widehat{v}_{\rm b}\}+\dfrac{1}{n}\widehat{v}_{\rm pm}$

FIC construction: Derivation details (II)

- The convergence in distribution also holds jointly:

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{\rm np}-\mu_{\rm true}\\ \widehat{\mu}_{\rm pm}-\mu_{0,\rm pm}\end{pmatrix}\to_d \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}v_{\rm np} & v_{\rm c}\\ v_{\rm c} & v_{\rm pm}\end{pmatrix}\right)$$

- $\Rightarrow \sqrt{n}(\widehat{b}-b)\to_d \mathrm{N}(0,v_b)$, with $\widehat{b}=\widehat{\mu}_{\rm pm}-\widehat{\mu}_{\rm np}$ and $v_b = v_{\rm pm}+v_{\rm np}-2v_c$
- $\Rightarrow \mathsf{E}\{(\widehat{b})^2\} = (\mathsf{E}\{\widehat{b}\})^2+\mathsf{Var}(\widehat{b})\approx b^2+\frac{1}{n}v_b$

General estimates of risk components

- Squared parametric bias: b^2 estimated by $\max\{0,\widehat{b}^2-\frac{1}{n}\widehat{v}_b\}$
- Variances: $v_{\rm np}$ estimated by $\widehat{v}_{\rm np}$, and $v_{\rm pm}$ estimated by $\widehat{v}_{\rm pm}$

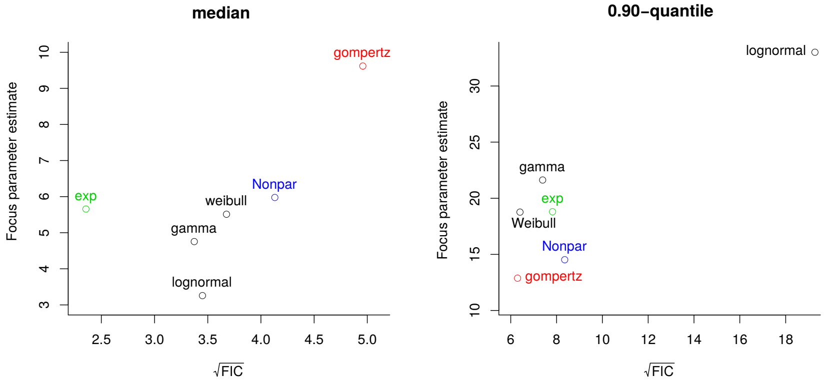
FIC scores

Nonparametric: $\mathrm{FIC}(\widehat{\mu}_{\rm np})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm np})=0+\dfrac{1}{n}\widehat{v}_{\rm np}$

Parametric: $\mathrm{FIC}(\widehat{\mu}_{\rm pm})=\widehat{\mathrm{mse}}(\widehat{\mu}_{\rm pm})=\max\{0,(\widehat{b})^2-\dfrac{1}{n}\widehat{v}_{\rm b}\}+\dfrac{1}{n}\widehat{v}_{\rm pm}$

Applying the FIC

- 1 data set, 5+1 competing candidate models
- 2 different focus parameters ⇒ 2 different model ranking lists



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

Nonparametric: $\mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n}\widehat{v}_{\mathrm{np}}(t)\mathrm{d}W(t)$

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

The criterion selects the model with the smallest AFIC score

Averaged Focused Information Criterion (AFIC)

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

AFIC scheme

Nonparametric: $\mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n}\widehat{v}_{\mathrm{np}}(t)\mathrm{d}W(t)$

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

The criterion selects the model with the smallest AFIC score

Averaged Focused Information Criterion (AFIC)

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

AFIC scheme

Nonparametric: $\mathrm{AFIC}(\widehat{\mu}_{\mathrm{np}}) = \int \frac{1}{n}\widehat{v}_{\mathrm{np}}(t)\mathrm{d}W(t)$

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

The criterion selects the model with the smallest AFIC score

Data types and situations

General setup requirements

- Some data generating mechanism G
- Focus parameters which can be written as a functional T of the data generating mechanism G: $\mu = T(G)$
- Focus parameter μ estimated by plug-in for each model M: $\widehat{\mu}_M = T(\widehat{G}_M)$

As seen: Standard i.i.d. situation

- G is cumulative distribution function
- Nonparametric estimation: Cumulative distribution function $\widehat{G}_n = n^{-1}\#\{Y_i \leq y\}$ or smoothed versions
- Parametric models: Estimated by ML or the like
- Typical focus parameters: Smooth functions of means and quantiles

Data types and situations

General setup requirements

- Some data generating mechanism G
- Focus parameters which can be written as a functional T of the data generating mechanism G: $\mu = T(G)$
- Focus parameter μ estimated by plug-in for each model M: $\widehat{\mu}_M = T(\widehat{G}_M)$

As seen: Standard i.i.d. situation

- G is cumulative distribution function
- Nonparametric estimation: Cumulative distribution function $\widehat{G}_n = n^{-1}\#\{Y_i \leq y\}$ or smoothed versions
- Parametric models: Estimated by ML or the like
- Typical focus parameters: Smooth functions of means and quantiles

Data types and situations (II)

Dependency modelling in stationary Gaussian time series models

- $G =$ Unknown spectral measure/distribution
- Nonparametric estimation: Integrated periodogram $\widehat{G}_n(\omega) = \int_{-\pi}^{\omega} I_n(u)\,\mathrm{d}u$ for I_n the periodogram
- Parametric estimation: Typically ARMA-models, but also more general parametric models for the spectral distribution
- Typical focus parameters: Differences in spectral distribution, covariance lags and correlation lags

Data types and situations (III)

Censoring

- Semiparametric Cox regression vs. fully parametric proportional hazard regression models
  - $G = (A(\cdot), \beta)$ corresponding to the hazard rate function: $\alpha(s) \exp(x^{\mathrm{t}}\beta)$, with α unspecified baseline hazard and A its cumulative
  - Semiparametric Cox regression: $\widehat{\beta}_{\mathsf{cox}}$ via Cox's partial likelihood, $\widehat{A}_{\mathsf{cox}}(\cdot) =$ Breslow estimator
  - Parametric estimation: Joint ML estimation of θ, β, with a parametric hazard rate function α_θ (exponential, Weibull, Gompertz, ...)
  - Typical focus parameters: Survival probabilities, quantiles and cumulative hazards, conditional on covariate values
- Without covariates: Nelson–Aalen or Kaplan–Meier estimators vs. parametric survival models

FIC/AFIC asymptotics

Consider one of the parametric models M: with least false model specification F_{θ_0} and least false focus parameter value $\mu_{0,\mathrm{pm}}$:

- If M is fully correct ($G = F_{\theta_0}$)
  - Pr (FIC selects M over nonparametric) $\to \chi_1^2(2) \approx 0.843$
  - No general results for AFIC (depends on μ and weight W)
- If M is biased ($\mu_{\mathrm{true}} \neq \mu_{0,\mathrm{pm}}$)
  - Pr (FIC/AFIC selects M) $\to 0$

$\Rightarrow$ If all parametric candidate models are biased:
Pr (FIC/AFIC selects nonparametric) $\to 1$

- Insurance mechanism against parametric misspecification

FIC/AFIC asymptotics

Consider one of the parametric models M: with least false model specification F_{θ_0} and least false focus parameter value $\mu_{0,\mathrm{pm}}$:

- If M is fully correct $(G = F_{\theta_0})$
  - Pr (FIC selects M over nonparametric) $\to \chi_1^2(2) \approx 0.843$
  - No general results for AFIC (depends on μ and weight W)
- If M is biased $(\mu_{\mathrm{true}} \neq \mu_{0,\mathrm{pm}})$
  - Pr (FIC/AFIC selects M) $\to 0$

$\Rightarrow$ If all parametric candidate models are biased:
Pr (FIC/AFIC selects nonparametric) $\to 1$

- Insurance mechanism against parametric misspecification

FIC/AFIC asymptotics

Consider one of the parametric models M: with least false model specification F_{θ_0} and least false focus parameter value $\mu_{0,\mathrm{pm}}$:

- If M is fully correct ($G = F_{\theta_0}$)
  - Pr (FIC selects M over nonparametric) $\to \chi_1^2(2) \approx 0.843$
  - No general results for AFIC (depends on μ and weight W)
- If M is biased ($\mu_{\mathrm{true}} \neq \mu_{0,\mathrm{pm}}$)
  - Pr (FIC/AFIC selects M) $\to 0$

$\Rightarrow$ If all parametric candidate models are biased:
Pr (FIC/AFIC selects nonparametric) $\to 1$

- Insurance mechanism against parametric misspecification

Special case of AFIC for categorical data

- 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 $1/p_j$
- Comparing AFIC_{pm} vs. AFIC_{np} is equivalent to comparing

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

- X_n is Pearson's chi squared test statistic*
- $\Rightarrow$ New motivation for Pearson chi-squared test, testing correctness of parametric model "pm". Test level chosen by the theory itself
  - df$=1, \ldots, 10$ gives asymptotic test level of $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035$

Special case of AFIC for categorical data

- 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 $1/p_j$
- Comparing AFIC_{pm} vs. AFIC_{np} is equivalent to comparing

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

- X_n is Pearson's chi squared test statistic*
- $\Rightarrow$ New motivation for Pearson chi-squared test, testing correctness of parametric model "pm". Test level chosen by the theory itself
  - df$=1, \ldots, 10$ gives asymptotic test level of $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035$

Special case of AFIC for categorical data

- 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 $1/p_j$
- Comparing AFIC_{pm} vs. AFIC_{np} is equivalent to comparing

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

- X_n is Pearson's chi squared test statistic*
- $\Rightarrow$ New motivation for Pearson chi-squared test, testing correctness of parametric model "pm". Test level chosen by the theory itself
  - df$=1, \ldots, 10$ gives asymptotic test level of $0.157, 0.135, 0.112, 0.092, 0.075, 0.062, 0.051, 0.042, 0.035$

Summary

- Answer our model selection question by relying on a focus driven model selection with a nonparametric alternative
- Rank models according to $\widehat{\mu}_M$'s estimated risk
- 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 parametric models