

Model selection in semiparametric models

Gerda Claeskens

gerda.claeskens@econ.kuleuven.be



Co-authors: R.J. Carroll, N.L. Hjort

Addressing the 'quiet scandal of statistics':
Don't hide the model selection step

Semiparametric likelihood

Variable selection in parametric part of the model, without assuming nonparametric part to be known

$$\sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \alpha_{\text{true}}, \theta_{\text{true}}(X_i)\}$$

Examples:

- Partially linear model: $Y_i = Z_i^t \alpha + \theta(X_i) + \varepsilon_i$
- $Y_i \sim N(Z_i^t \alpha; \sigma^2(X_i))$, $\theta(X_i) = \log \sigma(X_i)$
- $\text{Logit}\{P(\text{Success})\} = Z_i^t \alpha + \theta(X_i)$

Main references

- Hjort & Claeskens (2003, JASA) FMA
Frequentist model averaging estimators.
Distributions of estimators after model selection.
- Claeskens & Hjort (2003, JASA) FIC
The focussed information criterion.
Adapt selection criterion to depend on the quantity you wish to estimate with the selected model.

For full likelihood parametric models only

We extend FMA to general semiparametric models

Defs & model assumptions

Log likelihood function:

$$\sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \alpha_{\text{true}}, \theta_{\text{true}}(X_i)\}$$

(1) full model: $\alpha_{\text{full}}^t = (\beta^t, \gamma^t)$

(2) Reduced model: $\alpha_{\text{red}}^t = (\beta^t, 0_q^t)$

Local misspecification assumption: $\gamma_{\text{true}} = \delta / \sqrt{n}$

Denote the partial derivatives $\mathcal{L}_\theta(\cdot)$, $\mathcal{L}_\alpha(\cdot)$, $\mathcal{L}_{\alpha,\alpha}(\cdot)$, etc.

! All expectations are wrt true distribution of Y_i .

Estimation of nonparametric part

Define $\theta(x, \alpha)$ as the solution to

$$E[\mathcal{L}_\theta\{Y, Z, \alpha, \theta(X, \alpha)\} | X = x] = 0.$$

Of course, $\theta(\cdot, \alpha_{\text{true}}) = \theta_{\text{true}}(\cdot)$.

Local linear estimator $\{\hat{\theta}(x; \alpha_S), \hat{\theta}_1(x; \alpha_S)\}$ is the maximizer, with respect to (ψ_0, ψ_1) , of

$$n^{-1} \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \alpha_S, \psi_0 + \psi_1(X_i - x)\} K_h(X_i - x),$$

where $K_h(\cdot) = K(\cdot/h)/h$.

If the first partial derivatives of the likelihood exist,
set of estimating equations in the semiparametric model:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_\theta\{Y_i, Z_i, \alpha_S, \psi_0 + \psi_1(X_i - x)\} K_h(X_i - x) \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} = 0.$$

○ Claeskens & Van Keilegom (2003, Ann.Stat)

$$\sup_x |\hat{\theta}(x, \alpha) - \theta(x, \alpha)| = O(\{nh/\log(n)\}^{-1/2} + h^2) \text{ a.s.}$$

$$\sup_x \left| \hat{\theta}_1(x, \alpha) - \frac{\partial}{\partial x} \theta(x, \alpha) \right| = O(\{nh^3/\log(n)\}^{-1/2} + h^2) \text{ a.s.}$$

○ Claeskens & Carroll (2005)

$$\sup_x \left| \frac{\partial}{\partial \alpha} \hat{\theta}(x, \alpha) - \frac{\partial}{\partial \alpha} \theta(x, \alpha) \right| = O[\{nh/\log(n)\}^{-1/2} + h^2] \text{ a.s.}$$

Profile likelihood estimation of α

Given $\hat{\theta}(x, \alpha_S)$, we define $\hat{\alpha}_S$ as the solution to

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \alpha_S, \hat{\theta}(X_i, \alpha_S)\} \\ &= n^{-1} \sum_{i=1}^n \left[\mathcal{L}_\alpha\{Y_i, Z_i, \alpha_S, \hat{\theta}(X_i, \alpha_S)\} \right. \\ &\quad \left. + \mathcal{L}_\theta\{Y_i, Z_i, \alpha_S, \hat{\theta}(X_i, \alpha_S)\} \frac{\partial}{\partial \alpha_S} \hat{\theta}(X_i, \alpha_S) \right]. \end{aligned}$$

Least favorable curve

The local linear estimator $\hat{\theta}(\cdot, \alpha)$ is a consistent estimator of the least favorable curve $\theta^*(\cdot, \alpha)$ which minimizes

$$E \left[\frac{d}{d\alpha} \mathcal{L}\{Y, Z, \alpha_{\text{true}}, \theta^*(X, \alpha_{\text{tr}})\} \frac{d}{d\alpha} \mathcal{L}\{Y, Z, \alpha_{\text{tr}}, \theta^*(X, \alpha_{\text{tr}})\}^t \mid X \right]$$

Asymptotic theory for param. part

Define the semiparametric information bound

$$\mathcal{S}(\alpha_{\text{true}}) = \text{cov}\left\{\frac{d}{d\alpha}\mathcal{L}\{Y, Z, \alpha_{\text{true}}, \theta(X, \alpha_{\text{true}})\}\right\}$$

$$\mathcal{S}(\alpha) = \begin{pmatrix} S_{\beta\beta}(\alpha) & S_{\beta\gamma}(\alpha) \\ S_{\gamma\beta}(\alpha) & S_{\gamma\gamma}(\alpha) \end{pmatrix}, \quad \mathcal{S}^{-1}(\alpha) = \begin{pmatrix} S^{\beta\beta}(\alpha) & S^{\beta\gamma}(\alpha) \\ S^{\gamma\beta}(\alpha) & S^{\gamma\gamma}(\alpha) \end{pmatrix}.$$

Under the local misspecification assumption, that is,

$\gamma_{\text{true}} = n^{-1/2}\delta$, and when working in the full model,

$$n^{1/2}(\hat{\alpha}_{\text{full}} - \alpha_{\text{true}}) \xrightarrow{d} \mathbf{N}\{0, \mathcal{S}^{-1}(\alpha_{\text{true}})\}$$

Under the local misspecification assumption, that is,

$\gamma_{\text{true}} = n^{-1/2}\delta$, and when working in the reduced model,

$$n^{1/2}(\hat{\beta}_{\text{red}} - \beta_{\text{true}}) \xrightarrow{d} \mathbf{N}\{S_{\beta\beta}^{-1}(\alpha_{\text{true}})S_{\beta\gamma}(\alpha_{\text{true}})\underline{\delta}, S_{\beta\beta}^{-1}(\alpha_{\text{true}})\}$$

Under the local misspecification assumption,

$$n^{1/2}\{\mu(\hat{\alpha}_{\text{full}}) - \mu(\alpha_{\text{true}})\} \xrightarrow{d} \Lambda_{\text{full}} = \frac{\partial\mu}{\partial\alpha} \cdot \mathbf{N}\{0, S^{-1}(\alpha_{\text{true}})\}$$

$$n^{1/2}\{\mu(\hat{\alpha}_{\text{red}}) - \mu(\alpha_{\text{true}})\} \xrightarrow{d} \Lambda_{\text{red}} = \frac{\partial\mu}{\partial\beta} \cdot \mathbf{N}\{S_{\beta\beta}^{-1}(\alpha_{\text{true}})S_{\beta\gamma}(\alpha_{\text{true}})\underline{\delta}, S_{\beta\beta}^{-1}(\beta_{\text{true}})\} - \frac{\partial\mu}{\partial\gamma}\underline{\delta}.$$

Model selection

- First: distribution of estimators in each of the models.
- Next: define weights or apply model selection method.

For example AIC or BIC based on the semiparametric profile loglikelihood

$$n^{-1} \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \alpha_S, \hat{\theta}(X_i, \alpha_S)\},$$

assigns a data-driven AIC/BIC weight to each model.

e.g. weight = 1 if model selected, 0 otherwise.

- Combine weights with estimators.

Main result

We consider cases where the weights depend on

$$\hat{\delta}_{\text{full}} = n^{1/2} \hat{\gamma}_{\text{full}} \xrightarrow{d} D = \mathbf{N}(\delta, S^{\gamma}).$$

Estimators after model selection take the form

$$\hat{\mu} = c(\hat{\delta}_{\text{full}}) \mu(\hat{\alpha}_{\text{full}}) + \{1 - c(\hat{\delta}_{\text{full}})\} \mu(\hat{\alpha}_{\text{red}}),$$

with $0 \leq c(\hat{\delta}_{\text{full}}) \leq 1$.

Under the local misspecification assumption,

$$n^{1/2} \{\hat{\mu} - \mu(\alpha_{\text{true}})\} \xrightarrow{d} c(D) \Lambda_{\text{full}} + \{1 - c(D)\} \Lambda_{\text{red}}.$$

Danish malignant melanoma study

Anderson, Borgan, Gill & Keiding (1993)

- t_i , survival time after operation (in days);
- c_i , 1: dead from illness, 2: censored, 4: other cause;
- x_1 , woman = 1/man = 2;
- z_1 , thickness of the tumour;
- z_2 , infection infiltration level (resistance high 1, low 4);
- z_3 , presence or not of epithelioid cells;
- z_4 , presence or not of ulceration;
- z_5 , invasion depth (at levels 1, 2, 3);
- z_6 , age at the operation (in years).

Which of z_1, \dots, z_6 to include, for which purposes?

Framework for model selection

Hazard rate in the Cox model:

$$h_i(u) = h_0(u) \exp(x_i^t \beta + z_i^t \gamma) \quad \text{for } i = 1, \dots, n,$$

with $x_i = (x_{i,1}, \dots, x_{i,p})^t$ protected,
 $z_i = (z_{i,1}, \dots, z_{i,q})^t$ to choose from.

We **focus** on selecting models to estimate $\mu = \mu(\beta, \gamma, H_0)$, like the survival probability, median survival time, relative risk, average relative risk man vs. woman, etc.

For each $S \subset \{1, \dots, q\}$: $\hat{\mu}_S = \mu(\hat{\beta}_S, \hat{\gamma}_S, 0_{S^c}, \hat{H}_{0,S})$,
where $(\hat{\alpha}_S, \hat{\gamma}_S)$ are Cox estimators in the model that only
uses $z_{i,j}$ for $j \in S$, and

$$\hat{H}_{0,S}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \exp(x_i^t \hat{\beta}_S + z_{i,S}^t \hat{\gamma}_S)}.$$

Too big S : too much variance, big confidence intervals,
low test power.

Too small S : bias, not capturing all effects.

Which S ? Should we **average** across models?
Balancing **model bias** \longleftrightarrow **estimation variability**.

Limit distribution theory

Local misspecification framework:

$$h_{i,\text{true}}(u) = h_0(u) \exp(x_i^t \beta + z_i^t \delta / \sqrt{n}) \quad \text{for } i = 1, \dots, n,$$

The partial log-likelihood:

$$\begin{aligned} \log L_n(\beta, \gamma) &= \sum_{i=1}^n \int_0^\tau \left[x_i^t \beta + z_i^t \gamma \right. \\ &\quad \left. - \log \left\{ \sum_{i=1}^n Y_i(u) \exp(x_i^t \beta + z_i^t \gamma) \right\} \right] dN_i(u) \end{aligned}$$

$$Y_i(u) = I\{t_i \geq u\} \quad \text{and} \quad dN_i(u) = I\{t_i \in [u, u + du], \delta_i = 1\}.$$

Information matrix

Standard conditions imply that the $(p + q) \times (p + q)$ matrix of - 2nd derivatives/ n :

$$-I_n(\beta, \delta/\sqrt{n})/n \xrightarrow{p} J_{\text{full}} = \begin{pmatrix} J_{\beta\beta} & J_{\beta\gamma} \\ J_{\gamma\beta} & J_{\gamma\gamma} \end{pmatrix}.$$

All large-sample theorems will involve J_{full} and its inverse

$$J_{\text{full}}^{-1} = \begin{pmatrix} J^{\beta\beta} & J^{\beta\gamma} \\ J^{\gamma\beta} & J^{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} J^{\beta\beta} & J^{\beta\gamma} \\ J^{\gamma\beta} & K \end{pmatrix}.$$

Estimated as $\hat{J}_{\text{full}} = -I_n(\hat{\beta}_{\text{full}}, \hat{\gamma}_{\text{full}})/n$.

Limit distribution

$$\Omega_S = \pi_S^t K_S \pi_S K^{-1},$$

$$\omega = J_{\gamma\beta} J_{\beta\beta}^{-1} \frac{\partial \mu}{\partial \beta} - \frac{\partial \mu}{\partial \gamma},$$

$$\kappa = \{J_{\gamma\beta} J_{\beta\beta}^{-1} F_0(t) - F_1(t)\} \frac{\partial \mu}{\partial H_0},$$

$\Lambda_{n,S} = \sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ tends in distribution to Λ_S , which is normal with mean and variance

$$(\omega - \kappa)^t (I - \Omega_S) \delta \quad \text{and} \quad \tau_0^2 + (\omega - \kappa)^t \Omega_S K \Omega_S^t (\omega - \kappa).$$

A list of focus parameters

(i) Relative hazard:

$$\mu_1 = \frac{h(s|x, z)}{h(s|x_0, z)} = \exp\{(x - x_0)^t \beta\}.$$

It has $\omega = \exp(x^t \beta)(J_{\gamma\beta} J_{\beta\beta}^{-1} x - z)$ while $\kappa = 0$.

(ii) Baseline hazard: $\mu_2 = H_0(t)$ separately. Here $\omega = 0$ and $\kappa = J_{\gamma\beta} J_{\beta\beta}^{-1} F_0(t) - F_1(t)$.

(iii) Survival probability for a given individual:

$$\begin{aligned}\mu_3 &= S(t|x, z) = \exp\{-\exp(x^t \beta + z^t \gamma) H_0(t)\}, \\ \omega &= -S(t|x, z) H_0(t) (J_{10} J_{00}^{-1} x - z), \\ \kappa &= -S(t|x, z) \exp(x^t \beta) \{J_{10} J_{00}^{-1} F_0(t) - F_1(t)\}.\end{aligned}$$

The Focussed Information Criterion

Idea: Find limiting risk, then estimate, and select model with smallest risk estimate.

We have $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S$. Limiting risk for $\hat{\mu}_S$:

$$\begin{aligned} \text{risk}(S) &= \text{variance} + \text{bias}^2 \\ &= \tau_0^2 + (\omega - \kappa)^t \{ (I - \Omega_S) \delta \delta^t (I - \Omega_S)^t + \Omega_S K \Omega_S^t \} (\omega - \kappa). \end{aligned}$$

$$\begin{aligned} \widehat{\text{risk}}(S) &= \widehat{\tau}_0^2 + (\widehat{\omega} - \widehat{\kappa})^t \{ (I - \widehat{\Omega}_S) (\widehat{\delta} \widehat{\delta}^t - \widehat{K}) (I - \widehat{\Omega}_S)^t \\ &\quad + \widehat{\Omega}_S \widehat{K} \widehat{\Omega}_S^t \} (\widehat{\omega} - \widehat{\kappa}). \end{aligned}$$

FIC for Cox regression models

Simplify and remove terms not depending on S . Consider

$$\tilde{\psi}_{\text{full}} = (\omega - \kappa)^t D \quad \text{and} \quad \tilde{\psi}_S = (\omega - \kappa)^t \Omega_S D$$

for estimating $\psi = \omega^t \delta$ based on $D \sim \mathbf{N}_q(\delta, K)$. We find

$$\mathbf{FIC}_S = (\tilde{\psi}_{\text{full}} - \tilde{\psi}_S)^2 + 2(\omega - \kappa)^t \pi_S^t K_S \pi_S (\omega - \kappa).$$

The **focussed information criterion**, or **FIC**, takes the model with smallest estimated risk.

Model averaging estimators

Model selection schemes, like the AIC, BIC and the FIC, take the form

$$\hat{\mu} = \sum_S w_n(S|\hat{\delta}_{\text{full}}) \hat{\mu}_S,$$

where $w_n(S|\hat{\delta})$ is indicator for the chosen set.

More generally: any data-dependent $w_n(S|\hat{\delta})$ with sum 1.

May smooth across all models, or only over some of them, like the nested sequence $\emptyset, \{1\}, \{1, 2\}, \dots, \text{full}$.

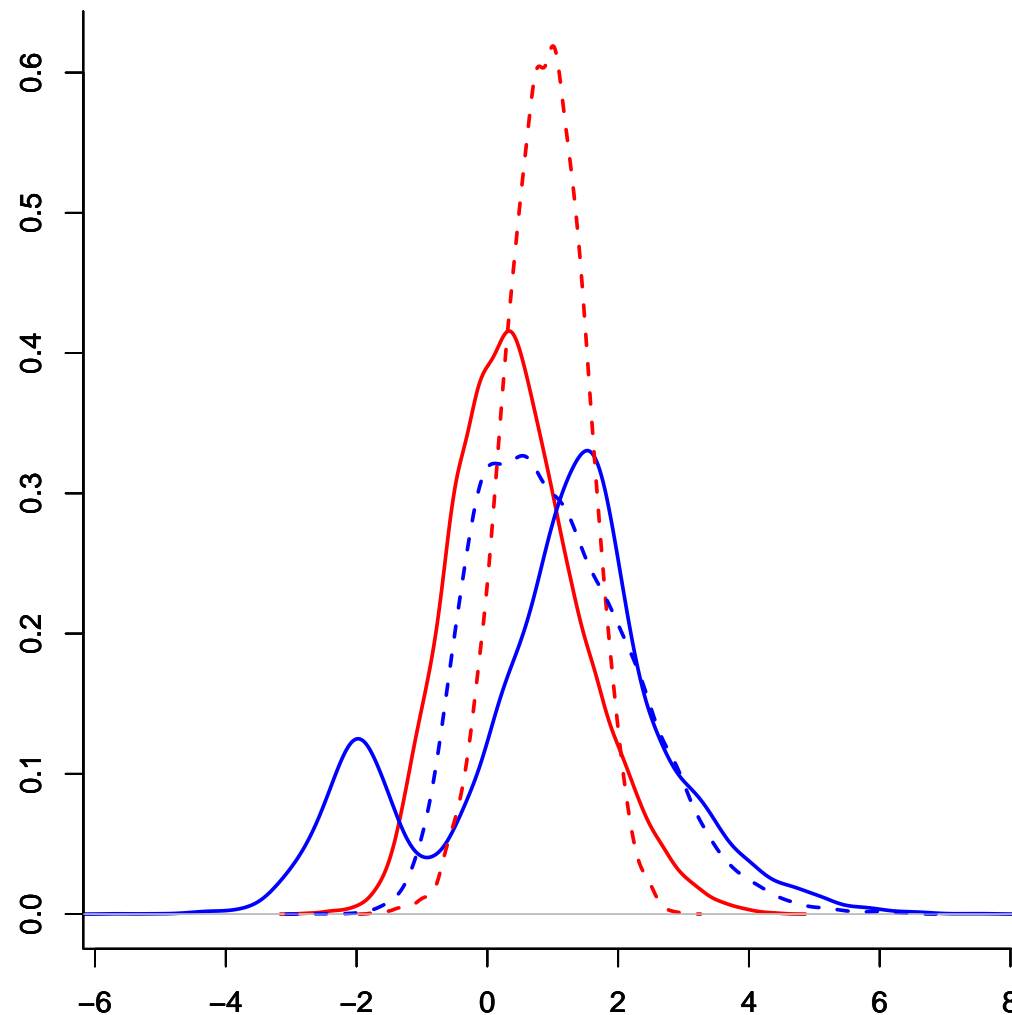
Main result

Hjort & Claeskens (2005)

For a general FMA (frequentist model average) estimator, if $w_n(S|\hat{\delta}) \xrightarrow{d} w(S|D)$ for each S ,

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \rightarrow \Lambda = \Lambda_0 + (\omega - \kappa)^t \{\delta - \hat{\delta}(D)\}.$$

Here $\Lambda_0 \sim \mathbf{N}(0, \tau_0^2)$, independent of $D \sim \mathbf{N}_q(\eta, K)$, and $\hat{\delta}(D) = \sum_S w(S|D)\Omega_S D$.



Focus $\mu_2 = H_0(t)$. Densities Λ for $\sqrt{n}(\hat{\mu}_2 - \mu_2)$, for **FIC**, **AIC** (solid lines), **smoothed FIC**, **smoothed AIC** (dashed lines), based on 10,000 simulations from Λ at $\hat{\delta} = \sqrt{n}\hat{\gamma}_{\text{full}}$.

$$\text{AIC}_{n,S} = 2 \log L_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) - 2(p + |S|)$$

$$\text{BIC}_{n,S} = 2 \log L_{n,S}(\hat{\beta}_S, \hat{\gamma}_S) - (p + |S|) \log n,$$

Highest value gives the best model.

$$\text{AIC}_{n,S} : z_2, z_3, z_4, z_5, z_6$$

$$\text{BIC}_{n,S} : z_4, z_5$$

Use this model **for ALL purposes.**

$$\text{FIC1: } \mu_1 = \exp\{(x - x_0)^t \beta + (z - z_0)^t \gamma\}$$

$$\text{FIC2: } \mu_2 = H_0(t) \text{ at time } t = 3 \text{ years.}$$

$$\text{FIC3: } \mu_3 = S(t|x, z).$$

vars	AIC	vars	BIC	vars	FIC1	vars	FIC2	vars	FIC3
23456	-527.2	45	-542.7	\emptyset	2.84	5	3.33	5	0.15
3456	-528.3	14	-543.0	2	4.11	25	4.09	6	0.18
12346	-528.6	345	-544.0	26	4.17	235	4.11	46	0.18
123456	-528.7	24	-544.7	25	4.30	56	4.15	16	0.18
2346	-529.6	4	-544.9	256	4.35	356	4.15	56	0.19
13456	-529.7	3456	-544.9	6	4.63	125	4.40	146	0.19

6 Best values of the information criteria AIC, BIC and FIC for focus parameters: (1) relative risk, (2) cumulative hazard, (3) survival probability. 'vars' indicate the selected variables among z_1, \dots, z_6 .

Extensions and concluding remarks

- * The limit distribution $\Lambda = \Lambda(\delta)$ is a non-linear mixture of biased normals, and is often quite non-normal.
- * Can simulate from $\Lambda(\hat{\delta}_{\text{full}})$ to derive approximate confidence intervals etc.
- * Pretesting, backward and forward regression: special cases of the averaged estimators.

Papers obtainable from website:

<http://www.econ.kuleuven.be/gerda.claeskens/public>