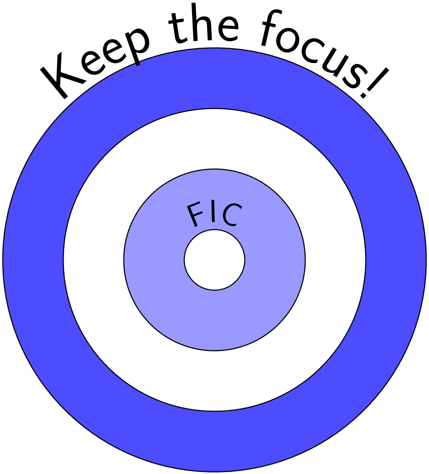
Gerda Claeskens

ORSTAT and Leuven Statistics Research Center

KU LEUVEN



FocuStat Conference May 22-25, 2018 – Oslo

Why is it good to be focused?

A little story: a proud big father takes his little 2 year old daughter to the doctor. Both are having a sore throat and apparently they both need antibiotics to get better. The doctor's focuses:

1. The average dose level needed for the little girl to get cured
2. The average dose level needed for the father to get cured

Relevant information (covariates): age and weight.

Why is it good to be focused?

A little story: a proud big father takes his little 2 year old daughter to the doctor. Both are having a sore throat and apparently they both need antibiotics to get better. The doctor's focuses:

1. The average dose level needed for the little girl to get cured
2. The average dose level needed for the father to get cured

Relevant information (covariates): age and weight.

Focus 1: mean dose level $= \mu$(age girl, weight girl),
Both age and weight are important
Focus 2: mean dose level $= \mu$(age father, weight father).
Perhaps age is not important (knowing that it is an adult),
weight might be important.
Classical model selection criteria (AIC, BIC, etc.) yield one model formula μ(age, weight) that works well on average: perhaps a too high dose for the little girl, not high enough for the father.

Notation

Example: a linear model $Y = \beta_0 + X\beta + \sigma\epsilon$,

- θ_0: length p, parameters included in all considered models.
  A natural choice would be $\theta_0 = (\sigma, \beta_0)$
- γ: length q, parameters on which we perform variable selection. e.g., $\gamma = \beta$.

Likelihood model $f(y|x, \theta_0, \gamma)$
Focus: quantity of interest $\mu_{true} = \mu(\theta_0, \gamma)$
e.g. $\mu(\sigma, \beta_0, \beta) = \beta_0 + x_{\mathrm{new}}\beta$ prediction for a new observation.

Many choices for estimation: $\mu(\widehat{\theta}, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_q)$, $\mu(\widehat{\theta}, \widehat{\gamma}_3, \widehat{\gamma}_5)$, $\mu(\widehat{\theta}, \widehat{\gamma}_1)$, etc.

Many choices for estimation: $\mu(\widehat{\theta}, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_q)$, $\mu(\widehat{\theta}, \widehat{\gamma}_3, \widehat{\gamma}_5)$, $\mu(\widehat{\theta}, \widehat{\gamma}_1)$, etc.

Properties of a good estimator:

$$\left.\begin{array}{l} \bullet \text{ Small or no bias} \\ \bullet \text{ Small variance} \end{array}\right\} \hookrightarrow \text{small MSE} = \text{bias}^2 + \text{var}$$

Select that model $S \subset \{1, \ldots, q\}$ for which the estimated MSE of the estimator $\mu(\widehat{\theta}_0, \widehat{\gamma}_S)$ is the lowest.

$\hookrightarrow$ Need to estimate the MSE of $\mu(\widehat{\theta}_0, \widehat{\gamma}_S)$ in each of the considered models.

The original Focused Information Criterion

Local misspecification framework: $\gamma_{\mathrm{true}} = \gamma_0 + \delta/\sqrt{n}$

This is used to avoid the bias to dominate the MSE expression.

Take $q = \text{length}(\gamma)$ fixed. $S \subseteq \{1, \ldots, q\}$ and let $(\hat{\theta}_S, \hat{\gamma}_S)$ be the MLE estimator.

- In this submodel, the estimator of the focus is $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S)$.
- Taylor expansion:
  $$\sqrt{n}(\hat{\mu}_S - \mu_{true}) \approx \left(\frac{\partial \mu}{\partial \theta}\right)^\top \sqrt{n}(\hat{\theta}_S - \theta_0) + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) - \left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta.$$
- We write $MSE(S)$ the mean squared error of $\sqrt{n}(\hat{\mu}_S - \mu_{true})$.
- The focused information criterion is defined as $FIC(S) = \widehat{MSE(S)}$.

The original Focused Information Criterion (2)

In the classical low-dimensional framework, Claeskens & Hjort (2003) show that for $\widehat{\mu}_S = \mu(\widehat{\theta}, \widehat{\gamma}_S)$, $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$, and with $\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma}$,

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_S \sim N\{E(\Lambda_S),\ \mathrm{Var}(\Lambda_S)\}$$

with mean $E(\Lambda_S) = \omega^\top(I_q - G_S)\delta$ and variance

$$\mathrm{Var}(\Lambda_S) = (\frac{\partial\mu}{\partial\theta})^\top J_{00}^{-1}\frac{\partial\mu}{\partial\theta} + \omega^\top \pi_S^\top J^{11,S}\pi_S\omega$$

where $G_S = \pi_S^\top J^{11,S}\pi_S(J^{11})^{-1}$. Fisher information matrix

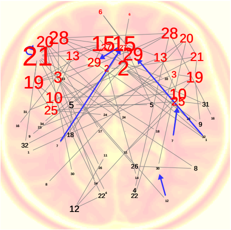
$$J_{\rm full} = \mathrm{Var(Score)} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$$

with inverse $J_{\rm full}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}$ where $J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$.

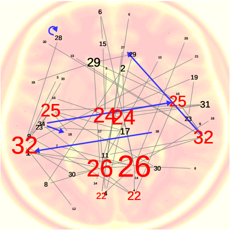
Application to fMRI data

The focus guides the model selection and is more important than the model

Prefrontal cortex



Parietal lobe



68 regions of interest 'ROI';
240 measurements over time
$(X_{1,t}, \ldots, X_{68,t}), t = 1, \ldots, 240.$

Nodewise regression models

Neighborhood selection [Meinshausen–Bühlmann '06]:
X_i 'response node', other X_j $(j \neq i)$ covariates in a linear regression
Lasso to determine the neighborhood of node i

$$\begin{aligned} \hat{ne}_i^\lambda &= \{j \in V : \hat{\theta}_j^i \neq 0\} \\ \hat{E}^{\lambda,\mathrm{AND}} &= \{(i,j) : i \in \hat{ne}_j^\lambda \text{ AND } j \in \hat{ne}_i^\lambda\} \\ \hat{E}^{\lambda,\mathrm{OR}} &= \{(i,j) : i \in \hat{ne}_j^\lambda \text{ OR } j \in \hat{ne}_i^\lambda\} \end{aligned}$$

Instantaneous and temporal effects

Nodewise Gaussian AR1 model, local misspecification, and a penalty

$$
\begin{aligned}
\mathcal{L}(\theta, \gamma) &= \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{k=2}^{n}\frac{y_k - \alpha - \tilde{x}_k^{\mathsf{T}}\beta - \rho y_{k-1}}{2\sigma^2} \\
&\quad -\frac{\lambda_n}{n}\{\sum_{j=1}^{d_\gamma}\psi(|\beta_j - \beta_{j0}|) + \psi(|\rho - \rho_0|)\},
\end{aligned}
$$

where $\theta = (\sigma^2, \alpha)$ and $\gamma = (\rho, \beta)$.
At lth node, estimated focus $\widehat{\mu}_{l;S_l} = \mu(\widehat{\theta}_{S_l}, \widehat{\gamma}_{S_l})$
Graphwise

$$
\mathrm{FIC}(\mathcal{G}(\mathcal{E}_{\mathcal{S}}, \mathcal{V})) = \sum_{l=1}^{p}\widehat{\mathrm{MSE}}(\widehat{\mu}_{l;S_l}).
$$

Penalty functions

Local quadratic approximation to ψ when not differentiable at zero.

- ridge: ℓ_2, $\psi_l(|\gamma_j - \gamma_{j0}|) = (\gamma_j - \gamma_{j0})^2$;
- lasso: ℓ_1, $\psi_l(|\gamma_j - \gamma_{j0}|) = |\gamma_j - \gamma_{j0}|$;
- bridge: $\psi_b(|\gamma_j - \gamma_{j0}|) = |\gamma_j - \gamma_{j0}|^\alpha;\ \alpha > 0$;
- hard thresholding: $\psi_h(|\gamma_j - \gamma_{j0}|) = \lambda^2 - (|\gamma_j - \gamma_{j0}| - \lambda)^2 I(|\gamma_j - \gamma_{j0}| < \lambda)$;
- adaptive lasso: $\psi_{al}(|\gamma_j - \gamma_{j0}|) = w_j|\gamma_j - \gamma_{j0}|$;
- SCAD (first derivative):
  $\psi_s^{'}(|\gamma_j - \gamma_{j0}|) = I(|\gamma_j - \gamma_{j0}| \leq \lambda) + \frac{(a\lambda - |\gamma_j - \gamma_{j0}|)_+}{(a-1)\lambda} I(|\gamma_j - \gamma_{j0}| > \lambda)$; $a > 2$.

Data-driven penalty constant

$$\hat{\lambda}_S = \arg\min_c \mathrm{MSE}(\hat{\mu}_S)\sqrt{n}/\psi^{''}(0)$$

Joint work with E. Pircalabelu, L. Waldorp, S. Jahfari; AoAS

The Focused Information Crit. for high-dimensional data

Joint work with T. Gueuning, SJS

Limitations of original formula

- p or q growing is not supported by the theory.
- Fisher information matrix J is often not invertible, e.g. if $p + q > n$.
- Penalty that performs selection brings additional selection uncertainty with it.

New setting: likelihood model $f(y|x, \theta_0, \gamma)$ with $\dim(\theta) = p$ fixed and $\dim(\gamma) = q_n$ diverging, allowing $p + q_n > n$.

We distinguish two cases:

- The submodel is low-dimensional ($p + |S| < n$)
- The submodel is high-dimensional ($p + |S| \geq n$), requiring a regularized estimator. We construct a desparsified estimator.

FIC in high-dimension: low-dimensional submodel

Likelihood model $f(y|x, \theta_0, \gamma_n)$ with $\gamma_n = \gamma_{0,n} + \delta_n/\sqrt{n}$ of dimension q_n.

We consider a low-dimensional submodel S for which the MLE estimator $(\hat{\theta}_S, \hat{\gamma}_S)$ is available.

Assumptions

- $\left\| \left[\left(\frac{\partial\mu}{\partial\theta}\right)^\top, \left(\frac{\partial\mu}{\partial\gamma_S}\right)^\top \right] \right\|_\infty = K = O(1)$ in a neighborhood of θ_0, γ_0.
- Sparsity condition on δ_n: $s_n = o(n^{1/4})$ with $S_{0,n} = \{j : \delta_{n,j} \neq 0\}$ and $s_n = |S_{0,n}|$

required for the following Taylor series expansion to be valid:

$$\sqrt{n}(\hat{\mu}_S - \mu_{true}) \approx \left(\frac{\partial\mu}{\partial(\theta, \gamma_S)}\right)^\top \begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} - \left(\frac{\partial\mu}{\partial\gamma}\right)^\top \delta.$$

FIC in high-dimension: low-dimensional submodel

We obtain the following limiting mean squared error:

$$\boxed{MSE(S) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^{\top} \left(B_S \delta \delta^{\top} B_S^{\top} + \pi_S^{*t} J_S^{-1} \pi_S^{*} \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}}$$

with $B_S = \pi_S^{*t} J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} - \begin{pmatrix} 0_{p \times q_n} \\ I_{q_n} \end{pmatrix}$ and we define

$$FIC(S) = \widehat{MSE(S)}.$$

Main advantage:

- Only J_S needs to be inverted, not the full information matrix, which might not be invertible.
- This FIC gives exactly the same value as the original FIC when $p + q < n$

FIC in high-dimension: high-dimensional submodel

- If $p + |S| > n$ then the MLE is not available.
  Regularization methods can be used.
  Example, for adaptive lasso (Zou 2006) is proven:
  - Consistent variable selection
    $\lim_{n\to\infty} P\big(\hat{A} = \{j : \hat{\beta}_j \neq 0\} = \{j : \beta_{j,\mathrm{true}} \neq 0\} = A\big) = 1$
  - Asymptotic normality $\sqrt{n}(\hat{\beta}_A - \beta_A) \to N(0, \Sigma_{\mathrm{Adapt.L}})$
- Information not sufficient to directly construct a FIC
  - For the construction of the FIC, the estimator's asymptotic distribution is used to estimate the MSE.
  - We want the full distribution, of all components, not only those of the true active set.

- We use the desparsified estimator introduced by van de Geer, Bühlmann, Ritov & Dezeure (2014, AoS), whose distribution can be tracked.
- We now restrict to a linear model.

$$Y = X_\beta \beta_0 + X_\gamma \gamma_n + \sigma\epsilon$$

- Extensions to GLM and convex loss functions are possible.
- Like in most of the high-dimensional literature, we assume that σ^2 is known. In practice, use $\hat{\sigma}_\epsilon^2 = \mathrm{RSS}/(n - \widehat{\mathrm{df}})$ with $\widehat{\mathrm{df}}$ the number of non-zero coefficients of the penalized estimator of γ_n.
- Protected variables: β_0. Unprotected variables: $\gamma_n = \delta_n/\sqrt{n}$.

A desparsified estimator

Let M_S be a relaxed inverse of $J_S = \frac{1}{n\sigma^2} X_S^{*t} X_S^*$ (by construction not invertible if $p + |S| > n$), obtained by the nodewise regression technique.

$$\begin{aligned}
\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} &= \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} + M_S \frac{1}{n\sigma^2} X_S^{*t} \left(Y - X_S^* \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} \right) \\
&= M_S \frac{1}{n\sigma^2} X_S^{*t} Y + \left(I_{p+|S|} - M_S J_S \right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix}.
\end{aligned}$$

- Interpretation 1: Correction of the Lasso bias, proportional to λ.
- Interpretation 2: Correction of the bias of $M_S \frac{1}{n\sigma^2} X_S^{*t} Y$ (the least squares estimator $J_S^{-1} \frac{1}{n\sigma^2} X_S^{*t} Y$ is not available).

FIC for a high-dimensional subset S

Calculations give

$$\boxed{MSE(S) \approx \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^{\top} \left(B'_S \delta \delta^{\top} B_S'^{\top} + \pi_S^{*\top} M_S J_S M_S^{\top} \pi_S^{*}\right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}}$$

with $B'_S = \left(\pi_S^{*t} J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} - \begin{pmatrix} 0_{p \times q_n} \\ I_{q_n} \end{pmatrix}\right) \left(I_q - \pi_S^{\top} \pi_S\right)$ and we define

$FIC(S) = \widehat{MSE(S)}$.

Particular cases:

- If $S_{0,n} \subseteq S$ then $B'_S \delta = 0_{p+q}$ (no bias).
- If $M_S = J_S^{-1}$ then $B'_S = B_S$, corresponding to the FIC formula for low-dim submodel.

Riboflavin data (R package hdi)

$n = 71$, $p = 4088$ predictors (gene expressions)

y: riboflavin production of the Bacillus subtilis bacteria.
Training set: $n' = 50$, test set: $n'' = 21$.

	Lasso	Best FIC	FIC 1	FIC 2
Avg Squared pred. error (21 focuses)	0.235	0.180	0.177	0.182
Average number of selected variables	27	6.7	4.6	10.7
Number of vars. selected at least once	27	120	77	177
Number of vars. selected at least 3 times	27	5	2	10

- FIC1: stepwise, start with empty set
- FIC2: stepwise, start with lasso selection

- FIC uses much fewer variables than Lasso (6.7 versus 27)
- Number of different variables used by the FIC much larger than for Lasso (120 versus 27).

Minimum mean squared error estimation

Starting assumptions:

- Linear model $Y = X_\theta \theta_0 + X_\gamma \gamma_n + \epsilon$
- Linear focus $\mu_{true} = x_0^\top \begin{pmatrix} \theta_0 \\ \gamma_n \end{pmatrix}$
- Local misspecification $\gamma_n = \delta/\sqrt{n}$

FIC searches among submodels of the big $X = (X_\theta, X_\gamma)$ to produce $\widehat{\mu}_S = \mathcal{X}_S Y$ with $\mathcal{X}_S = \pi_S^{*\top}(X_S^{*\top} X_S^*)^{-1} X_S^{*\top}$.

However,

- Estimation is more important than identifying a submodel
- Relax constraint: search a matrix $\mathcal{X}$ such that MSE of $\widehat{\mu} = \mathcal{X} Y$ is minimized over all matrices $\mathcal{X} \in \mathbb{R}^{(p+q)\times n}$.

Under local misspecification and with a linear focus:

$$\mathrm{MSE}(\mathcal{X}) = x_0^\top \mathcal{X} A \mathcal{X}^\top x_0 - x_0^\top \mathcal{X} B x_0 - x_0^\top B^\top \mathcal{X}^\top x_0 + x_0^\top C x_0$$

with

$$\begin{aligned} A &= X \begin{pmatrix} \sqrt{n}\theta_0 \\ \delta \end{pmatrix}^{\otimes 2} X^\top + n\sigma_\epsilon^2 I_n, \quad B = X \begin{pmatrix} \sqrt{n}\theta_0 \\ \delta \end{pmatrix}^{\otimes 2}, \\ C &= \begin{pmatrix} \sqrt{n}\theta_0 \\ \delta \end{pmatrix}^{\otimes 2}. \end{aligned}$$

This leads to $\mathcal{X}_{opt} = B^\top A^{-1}$ and $\mu_{opt} = x_0^\top B^\top A^{-1} Y$.

With $\beta = (\theta_0^\top, \delta^\top/\sqrt{n})^\top$,

$$\mathcal{X}_{opt} = \beta(X\beta)^\top \left(X\beta\beta^\top X^\top + \sigma_\epsilon^2 I_n\right)^{-1}.$$

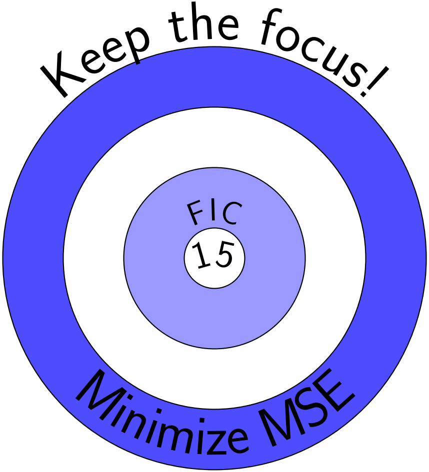
Use initial estimators $\tilde{\beta}$, $\tilde{\sigma}^2$.

$$\widehat{\beta} = \tilde{\beta}(X\tilde{\beta})^\top \left(X\tilde{\beta}\tilde{\beta}^\top X^\top + \tilde{\sigma}_\epsilon^2 I_n\right)^{-1} Y$$

For low dimensions MMSE of Farebrother (1975), Wan and Ohtani (2000). Not known yet for high-dimensional data.

Simulations indicate that FIC and MMSE are competitive, both dominate lasso.

Reiterating



Thank you!

Nodewise regression

Construction of M_S which acts as a relaxed inverse of J_S.
For each $j \in \{1, \ldots, p + |S|\}$ compute

$$\hat{\eta}_j = \operatorname*{arg\,min}_{\eta \in \mathbb{R}^{p+|S|-1}} \frac{1}{2n} \left\| X^*_{S,j} - X^*_{S,-j}\eta \right\|_2^2 + \lambda_j \left\| \eta \right\|_1 ,$$

where $X^*_{S,j}$ is the j th column of X^*_S and $X^*_{S,-j} \in \mathbb{R}^{n\times(p+|S|-1)}$ is X^*_S without its jth column, and we form

$$\hat{A}_S = \begin{bmatrix} 1 & -\hat{\eta}_{1,2} & \cdots & \hat{\eta}_{1,p+|S|} \\ -\hat{\eta}_{2,1} & 1 & \cdots & \hat{\eta}_{2,p+|S|} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\eta}_{p+|S|,1} & -\hat{\eta}_{p+|S|,2} & \cdots & \hat{\eta}_{p+|S|,p+|S|} \end{bmatrix}$$

with components of $\hat{\eta}_j$ indexed by $k \in \{1, \ldots, j-1, j+1, \ldots, p+|S|\}$. We define

$$M_S = \hat{T}_S^{-2} \hat{A}_S$$

with $\hat{T}_S^2 = \operatorname{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_{p+|S|}^2)$ and $\hat{\tau}_j^2 = \frac{1}{n} \left\| X^*_{S,j} - X^*_{S,-j}\hat{\eta}_j \right\|_2^2 + \lambda_j \left\| \hat{\eta}_j \right\|_1$. Back