

Selective inference using randomized group lasso estimators

Gerda Claeskens

Sarah Pirene

ORSTAT and Leuven Statistics Research Center, KU Leuven, Belgium

**Snigdha Panigrahi
and Yiling Huang**

Department of Statistics, University of Michigan, US

Godt Hjort: 4 & 5 December 2023



Down under in the year 2000

- ▶ Both visiting Peter Hall at ANU in Canberra
- ▶ Peter left for a few weeks to visit Europe
- ▶ Start of FMA work
- ▶ One weekend: hike to Telstra tower in Black Mountain

Model choice and compromise estimators in a local environment

Nils & Gerda

A good framework; need more results; version 0.37 or so, as of 4.4.2000

ABSTRACT. Narrow model: $f(y, \theta) = f(y, \theta, \gamma_0)$, real model: $f(y, \theta, \gamma)$ with an extra q -dimensional parameter γ which is not too far away from γ_0 . We consider narrow inference versus various forms of wider inference, including that suggested by model choice criteria like the AIC.

KEY WORDS: Akaike's information criterion, bias and variance balance, compromise estimators, growing models, moderate misspecification, smooth envelopes

1. Introduction and summary

[xx Models are seldom fully correct. (Find some quotes.) Need to study departures from them. Robustness. Model selection. AIC, BIC, others. What happens to estimators-after-selection? Smooth enveloping models. Variable selection in regression models; growing regression structures, with interactions, etc.; should be easy to find motivating examples. Stress the fact that we build a unified apparatus inside which 'everything' can be precisely analysed and compared. xx]

choice framework

Frequentist model averaging

Nils Lid Hjort and Gerda Claeskens

A good framework; need more results; version 0.52 or so, as of 2/6/1

ABSTRACT. Narrow model: $f(y, \theta) = f(y, \theta, \gamma_0)$, real model: $f(y, \theta, \gamma)$ with an extra q -dimensional parameter γ which is not too far away from γ_0 . We consider narrow inference versus various forms of wider inference, including that suggested by model choice criteria like the AIC. Many papers on Bayesian model averaging; this is a contribution in the frequentist direction. Confidence intervals are too short; we investigate what happens. Model selection/estimation problems.

KEY WORDS: Akaike's information criterion, bias and variance balance, compromise estimators, growing models, moderate misspecification, smooth envelopes

1. Introduction and summary

[xx Models are seldom fully correct. (Find some quotes.) Need to study departures from them. Robustness. Model selection. AIC, BIC, others. What happens to estimators-after-selection? Smooth enveloping models. Variable selection in regression models; growing regression structures, with interactions, etc.; should be easy to find motivating examples. Stress the fact that we build a unified apparatus inside which 'everything' can be precisely analysed and compared. Mention Raftery et al., Pötscher, Sen, George... xx]

... framework

Model selection, and next?

- ▶ Back in 2000: take the selected model and pretend as if you knew this model all the way long
- ▶ **FMA (2003, JASA discussion paper)**: under local misspecification framework, provide precise limit variable via model averaging
- ▶ Suggestion for simulation approach (see CUP book 2008: chapter 7)
- ▶ Now in 2023: still need to 'educate' some users of statistics to not pretend to know too much

Highlights of today's talk

- ▶ What?
 - ▶ Inference after selection by group lasso
 - ▶ Applicable to other than Gaussian models
 - ▶ Inference conditional on selection: **selective inference**
- ▶ Results:
 - ▶ Asymptotic post-selection likelihood
 - ▶ Correct inference after selection of variables
 - ▶ Wald-type confidence intervals with finite expected length

Data example: NHANES

National health and nutrition examination survey (US)

- ▶ sample size $n = 5985$
- ▶ Response binary: depressed or not
- ▶ 8 grouped categorical covariates (race, education, marital status, income, alcohol consumption, diet health, milk consumption, diabetes)
- ▶ 9 binary covariates (daily 4/5 or more drinks, ever any kind of alcohol, high blood pressure, high cholesterol level, smoked at least 100 cigarettes in life, asthma, diabetes, treatment for anemia, ever blood transfusion)
- ▶ 6 continuous covariates (age, family income to poverty ratio, BMI, nr of meals not prepared at home in last 7 days, nr of ready-to-eat foods in last 30 days, nr of frozen meals/pizza in last 30 days)

Group lasso selection

- ▶ $\widehat{\beta}_n \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \ell(X\beta; Y) + \sum_{g \in \mathcal{G}} \lambda_g \|\beta_g\|_2 \right\}$ (Yuan & Lin *JRSSB* 2006)
- ▶ Loss function to be minimized: $\ell(X\beta; Y) = \sum_{i=1}^n \rho(x_i^\top \beta; Y_i)$
- ▶ Regularization allowing for grouped variables: $\sum_{g \in \mathcal{G}} \lambda_g \|\beta_g\|_2$

Examples of loss functions:

- ▶ Negative log-likelihood, e.g. for binary data
 $\ell(X\beta; Y) = - \sum_{i=1}^n [y_i \log h(x_i^\top \beta) + (1 - y_i) \log \{1 - h(x_i^\top \beta)\}]$
with $h(\theta) = e^\theta / (1 + e^\theta)$ the inverse logit link function
- ▶ Quasi-likelihood estimation function (e.g. for overdispersed Poisson data)
- ▶ Least absolute deviation loss, Huber loss

More examples later in this talk.

	Post-GL		Naive	
Intercept	0.600		0.000	
Age	0.327		0.037	
Family income to poverty ratio	0.000		0.000	
BMI	0.616		0.093	
nr of meals not prepared at home last 7 days	0.223		0.773	
nr of ready-to-eat foods in last 30 days	0.166		0.171	
nr of frozen meals/ pizza in last 30 days	0.000		0.000	
Race (ref.: Other)		0.000		0.000
<i>Mexican american</i>	0.032		0.262	
<i>Other hispanic</i>	0.869		0.814	
<i>Non-hispanic white</i>	0.325		0.953	
<i>Non-hispanic black</i>	0.159		0.002	

	Post-GL		Naive	
Civil state (ref.: Never married)		0.000		0.000
<i>Married/Living with</i>	0.133		0.016	
<i>Widowed/Divorced/Separated</i>	0.315		0.639	
Healthy diet (ref.: Poor)		0.000		0.000
<i>Excellent</i>	0.023		0.000	
<i>Very good</i>	0.002		0.000	
<i>Good</i>	0.001		0.000	
<i>Fair</i>	0.005		0.000	
Male	0.822		0.000	
Daily 4/5 or more drinks	0.141		0.000	
High blood pressure/Hypertension	0.449		0.000	
Smoked at least 100 cigarettes in life	0.180		0.018	
Asthma	0.000		0.000	

Inference after group lasso

- ▶ Grouped variables: no polyhedral lemma as for lasso
(Lee, Sun, Sun, Taylor AoS 2016)
- ▶ General loss function – data not Gaussian
- ▶ Want to avoid infinite expected length for confidence intervals
(Kivaranovic & Leeb JASA 2021)
- ▶ Results need to be statistically valid

Randomized group lasso estimation:

$$\widehat{\beta}_n^{(\Lambda)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \ell(X\beta; Y) + \sum_{g \in \mathcal{G}} \lambda_g \|\beta_g\|_2 - \sqrt{n} \omega_n^\top \beta \right\}$$

with $\sqrt{n} \omega_n \sim N_p(0, \Omega)$

Selection set

Index set of selected coefficients

$$\widehat{E} = \{j \in \{1, \dots, p\} : \widehat{\beta}_{nj}^{(\Lambda)} \neq 0\}.$$

Complementary set $\widehat{E}' = \{1, \dots, p\} \setminus \widehat{E}$: indices that are not selected.

- Inference conditional on $\widehat{E}$

Target for inference when $\widehat{E} = E$

$$b_{n,E}^* = \operatorname{argmin}_{b \in \mathbb{R}^{|E|}} \frac{1}{\sqrt{n}} \mathbb{E} [\ell(X_E b; Y)].$$

E -restricted M-estimator $\widehat{\beta}_{n,E}$ satisfies $\frac{\partial}{\partial \beta_E} \ell(X_E \widehat{\beta}_E; Y) = 0_{|E|}$

Reparametrization

- ▶ $g \in \mathcal{G}_E$: set of selected groups

$$\widehat{\beta}_{n,g}^{(\Lambda)} = \widehat{\gamma}_{n,g} \widehat{u}_{n,g}, \text{ where } \widehat{\gamma}_{n,g} = \|\widehat{\beta}_{n,g}^{(\Lambda)}\|_2, \text{ and } \widehat{u}_{n,g} = \frac{\widehat{\beta}_{n,g}^{(\Lambda)}}{\|\widehat{\beta}_{n,g}^{(\Lambda)}\|_2}.$$

- ▶ $g \notin \mathcal{G}_E \Leftrightarrow g \in \mathcal{G}_{E'}$: unselected groups

$$\widehat{z}_n = (\widehat{z}_{n,g} : g \in \mathcal{G}_{E'})^\top$$

subgradient vector of the group lasso penalty $\sum_{g \in \mathcal{G}} \lambda_g \|\beta_g\|_2$ at the non-active (zero) groups.

$$\left\{ \widehat{E} = E, \widehat{u}_n = u_n, \widehat{z}_n = z_n \right\} = \left\{ \widehat{u}_n = u_n, \widehat{z}_n = z_n, \sqrt{n} \widehat{\gamma}_n \in \mathbb{R}_+^{|\mathcal{G}_E|} \right\}$$

Relate group lasso estimator to refitted one

Group lasso estimator $\widehat{\beta}_n^{(\Lambda)}$ satisfies

$$0_p = \frac{1}{\sqrt{n}} X^\top \nabla \ell(X \widehat{\beta}_n^{(\Lambda)}; Y) + \begin{pmatrix} (\lambda_g \widehat{u}_{n,g})_{g \in \mathcal{G}_E} \\ (\lambda_g \widehat{z}_{n,g})_{g \in \mathcal{G}_{E'}} \end{pmatrix} - \sqrt{n} \omega_n.$$

Equivalently,

$$\sqrt{n} \omega_n = \frac{1}{\sqrt{n}} X^\top \nabla \ell(X \widehat{\beta}_n^{(\Lambda)}; Y) + C_\mathcal{E} \Lambda_{E'} \widehat{z}_n + D_\mathcal{E} \widehat{u}_n.$$

Taylor series expansion of 1st term around refitted estimator gives

$$\sqrt{n} \omega_n = A_\mathcal{E} \sqrt{n} \widehat{\beta}_{n,E} + B_\mathcal{E} \sqrt{n} \widehat{U}_n \widehat{\gamma}_n + C_\mathcal{E} (\sqrt{n} \widehat{\beta}_{n,E}^\perp + \Lambda_{E'} \widehat{z}_n) + D_\mathcal{E} \widehat{u}_n + o_p(1).$$

Nuisance parameters

Aim: conditional distribution of $\sqrt{n}\widehat{\beta}_{n,E}$ given $\{\widehat{E} = E, \widehat{u}_n = u_n, \widehat{z}_n = z_n\}$

Eliminate nuisance parameters in the limit: condition on

$$\widehat{\beta}_{n,E}^\perp = \frac{1}{n} \frac{\partial}{\partial \beta_{E'}} \ell(X_E \widehat{\beta}_{n,E}; Y) - A_E \widehat{\beta}_{n,E} \in \mathbb{R}^{|E'|},$$

where $A_E = H_{E',E} - K_{E',E} K_{E,E}^{-1} H_{E,E}$.

$$\begin{aligned} H &= \frac{1}{n} \mathbb{E} \left[\frac{\partial^2}{\partial \beta \partial \beta^\top} \ell(X_E b_{n,E}^*; Y) \right] = \begin{bmatrix} H_{E,E} & H_{E,E'} \\ H_{E',E} & H_{E',E'} \end{bmatrix} \\ K &= \text{Var} \left(\frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta} \ell(X_E b_{n,E}^*; Y) \right) = \begin{bmatrix} K_{E,E} & K_{E,E'} \\ K_{E',E} & K_{E',E'} \end{bmatrix} \end{aligned}$$

Target nuisance parameter $b_{n,E}^\perp = \mathbb{E} \left[\frac{1}{n} \frac{\partial}{\partial \beta_{E'}} \ell(X_E b_{n,E}^*; Y) \right] - A_E b_{n,E}^*$

Marginal and conditional log-likelihood

Denote $L_{n,u_n,z_n,\beta_{n,E}^\perp}(b_{n,E}^*, b_{n,E}^\perp; \widehat{\beta}_{n,E}, \widehat{\gamma}_n)$
the **marginal** likelihood based on the distribution of

$$\sqrt{n}(\widehat{\beta}_{n,E}^\top, \widehat{\gamma}_n^\top)^\top \Big| \{ \widehat{\beta}_{n,E}^\perp = \beta_{n,E}^\perp, \widehat{u}_n = u_n, \widehat{z}_n = z_n \}$$

Now, truncate to the event $\{ \sqrt{n}\widehat{\gamma}_n \in \mathbb{R}_+^{|\mathcal{G}_E|} \}$.

Conditional likelihood:

$$\begin{aligned} & \frac{L_{n,u_n,z_n,\beta_{n,E}^\perp}(b_{n,E}^*, b_{n,E}^\perp; \widehat{\beta}_{n,E}, \widehat{\gamma}_n)}{\int L_{n,u_n,z_n,\beta_{n,E}^\perp}(b_{n,E}^*, b_{n,E}^\perp; \beta_{n,E}, \gamma_n) 1_{\mathbb{R}_+^{|\mathcal{G}_E|}}(\sqrt{n}\gamma_n) d\beta_{n,E} d\gamma_n} 1_{\mathbb{R}_+^{|\mathcal{G}_E|}}(\sqrt{n}\widehat{\gamma}_n) \\ &= \frac{L_{n,u_n,z_n,\beta_{n,E}^\perp}(b_{n,E}^*, b_{n,E}^\perp; \widehat{\beta}_{n,E}, \widehat{\gamma}_n)}{\mathbb{P}\left[\sqrt{n}\widehat{\gamma}_n \in \mathbb{R}_+^{|\mathcal{G}_E|} \Big| \widehat{\beta}_{n,E}^\perp = \beta_{n,E}^\perp, \widehat{u}_n = u_n, \widehat{z}_n = z_n\right]} 1_{\mathbb{R}_+^{|\mathcal{G}_E|}}(\sqrt{n}\widehat{\gamma}_n). \end{aligned}$$

Conditional log-likelihood

$$\log L_{n,u_n,z_n,\beta_{n,E}^\perp} \left(b_{n,E}^*, b_{n,E}^\perp; \widehat{\beta}_{n,E}, \widehat{\gamma}_n \right) \\ - \log \mathbb{P} \left[\sqrt{n} \widehat{\gamma}_n \in \mathbb{R}_+^{|\mathcal{G}_E|} \mid \widehat{\beta}_{n,E}^\perp = \beta_{n,E}^\perp, \widehat{u}_n = u_n, \widehat{z}_n = z_n \right]$$

1st term: asymptotic distribution of

$$\sqrt{n}(\widehat{\beta}_{n,E}^\top, \widehat{\gamma}_n^\top)^\top \Big| \{ \widehat{\beta}_{n,E}^\perp = \beta_{n,E}^\perp, \widehat{u}_n = u_n, \widehat{z}_n = z_n \}$$

is represented by the following likelihood

$$d(b_{n,E}^*) J(\sqrt{n} \widehat{\gamma}_n, U_n) \phi(\sqrt{n} \widehat{\beta}_{n,E}; \bar{R} \sqrt{n} b_{n,E}^* + \bar{s}, \bar{\Theta}) \phi(\sqrt{n} \widehat{\gamma}_n; \bar{A} \sqrt{n} \widehat{\beta}_{n,E} + \bar{b}, \bar{\Omega}),$$

with $d(b_{n,E}^*)$ normalizing constant

Large deviation result

- ▶ 2nd term. Positivity constraint: $\sqrt{n}\widehat{\gamma}_n \in \mathbb{R}_+^{|\mathcal{G}_E|}$
- ▶ Indicator function is non-differentiable $\hookrightarrow$ barrier function
 $\text{Barr}(v) = \sum_{k=1}^{|v|} \log(1 + 1/(v_k - c)).$



$$\lim_{n \rightarrow \infty} \frac{1}{a_n^2} \log \mathbb{P} \left[\sqrt{n}\widehat{\gamma}_n \in \mathcal{K} \mid \widehat{\beta}_{n,E}^\perp = b_E^\perp, \widehat{u}_n = u_n, \widehat{z}_n = z \right] - O_n = 0$$

- ▶ With $a_n = o(n^{1/2})$

$$\begin{aligned} O_n = - \inf_{b', g'} \left\{ \right. & \frac{1}{2} \left(b' - \bar{R}b_E - \frac{1}{a_n} \bar{s} \right)^\top \bar{\Theta}^{-1} \left(b' - \bar{R}b_E - \frac{1}{a_n} \bar{s} \right) \\ & + \frac{1}{2} \left(g' - \bar{A}b' - \frac{1}{a_n} \bar{b} \right)^\top \bar{\Omega}^{-1} \left(g' - \bar{A}b' - \frac{1}{a_n} \bar{b} \right) \\ & \left. - \frac{1}{a_n^2} \log J(a_n g'; U_n) + \frac{1}{a_n^2} \text{Barr}_{\mathcal{K}}(a_n g') \right\}, \end{aligned}$$

Asymptotic post-selection log-likelihood

- ▶ Substituting $a_n b' = \sqrt{n}b$, and $a_n g' = \sqrt{n}g$ gives the **asymptotic post-selection likelihood**.
- ▶ Ignoring constants, log-likelihood:

$$\begin{aligned} & \log \phi(\widehat{\sqrt{n}\beta_{n,E}}; \bar{R} \sqrt{n}b_{n,E}^* + \bar{s}, \bar{\Theta}) \\ & + \inf_{b,g} \left\{ \frac{1}{2} \left(\sqrt{n}b - \bar{R} \sqrt{n}b_{n,E}^* - \bar{s} \right)^\top \bar{\Theta}^{-1} \left(\sqrt{n}b - \bar{R} \sqrt{n}b_{n,E}^* - \bar{s} \right) \right. \\ & + \frac{1}{2} \left(\sqrt{n}g - \bar{A} \sqrt{n}b - \bar{b} \right)^\top \bar{\Omega}^{-1} \left(\sqrt{n}g - \bar{A} \sqrt{n}b - \bar{b} \right) \\ & \left. - \log J(\sqrt{n}g; U_n) + \text{Barr}_{\mathcal{K}}(\sqrt{n}g) \right\}. \end{aligned}$$

Conditional post-group lasso inference

► MLE

$$\sqrt{n}\widehat{b}_{n,E}^{mle} = \bar{R}^{-1} \sqrt{n}\widehat{\beta}_{n,E} - \bar{R}^{-1}\bar{s} + \Sigma_E \bar{A}^\top \bar{\Omega}^{-1} \left(\bar{A} \sqrt{n}\widehat{\beta}_{n,E} + \bar{b} - g_n^*(\sqrt{n}\widehat{\beta}_{n,E}) \right).$$

- Observed Fisher information matrix $I_{n,mle} = n\Sigma_E^{-1}M^{-1}\Sigma_E^{-1}$
 with $M = \bar{\Theta}^{-1} + \bar{A}^\top \bar{\Omega}^{-1} \bar{A} - \bar{A}^\top \bar{\Omega}^{-1} [\bar{\Omega}^{-1} - \nabla^2 \log J(g_n^*(\sqrt{n}\widehat{\beta}_{n,E}); U_n) + \nabla^2 \text{Barr}_{\mathcal{K}}(g_n^*(\sqrt{n}\widehat{\beta}_{n,E}))]^{-1} \bar{\Omega}^{-1} \bar{A}$.
- Valid post-selection confidence interval for $(b_{n,E}^*)_j$

$$\widehat{b}_{n,E}^{mle}[j] \pm z_{1-\alpha/2} \sqrt{I_{n,mle}^{-1}[j,j]}$$

- Valid post-selection confidence region for $b_{n,E}^*$ (or part of it)
- $$\left\{ b_{n,E}^* \in \mathbb{R}^{|E|} : (\widehat{b}_{n,E}^{mle} - b_{n,E}^*)^\top I_{n,mle} (\widehat{b}_{n,E}^{mle} - b_{n,E}^*) \leq \chi_{|E|,1-\alpha}^2 \right\}$$

Bounded confidence regions

For the inverse Fisher information matrix it holds that

$$\max_{i,j \in \{1, \dots, |E|\}} |(I_{n, \text{mle}}^{-1})_{ij}| \leq n^{-1} u_0 (1 + u_0^2),$$

where $u_0 = \max\{\lambda_{\max}(\Sigma_E), \lambda_{\max}(A_{\mathcal{E}}^{\top} \Omega^{-1} A_{\mathcal{E}})\}$
and for a square matrix A , $\lambda_{\max}(A)$ denotes its largest eigenvalue

Contrast with selective inference without randomization for lasso case
where the expected length of the confidence intervals is infinite
(Kivaranovic and Leeb, 2021)

Applications

- ▶ **GAMLSS**: generalized additive models for location, scale and shape (Rigby & Stasinopoulos 2005) **+randomization**
group lasso penalty: categorical covariates, smooth terms
 $\mu(x) = E(Y|x) = g_1(x_1)$ and $\sigma^2(x) = \text{Var}(Y|x) = g_2(x_2)$
 $g_1(x_1) = \beta_{10} + \sum_{j=1}^{q_1} \beta_{1j}x_{1j} + \sum_{j=q_1+1}^{p_1} s_{1j}(x_{1j})$
using spline function $s_{1j}(x_{1j}) = \sum_{k=1}^K \beta_{j,k}\psi_k(x_{1j})$
- ▶ Group regularized **quasi-likelihood estimation** **+randomization**
Overdispersed Poisson data
- ▶ **Smoothed quantile estimation** (He et al. *J. Econometrics* 2023)

$$\widehat{\beta}_{n,\tau}^{(\Lambda)} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ n^{-1/2} \sum_{i=1}^n (\rho_{\tau} * K_h)(Y_i - x_i^{\top} \beta) + \sum_{g \in \mathcal{G}} \lambda_g \|\beta_g\|_2 - \sqrt{n} \omega_n^{\top} \beta \right\},$$

Simulation comparison

Compare inference after randomized group lasso (**PostGL**) with:

- ▶ **naive**: refit after selection and classical inference *ignore selection*
- ▶ **Data splitting**: 33% for selection (n_1), 67% for inference (n_2)
- ▶ Only for Gaussian case: **Bayesian method** post-selection inference for group lasso (Panigrahi, MacDonald, Kessler, *JMLR* 2023)

Choice of randomisation matrix:

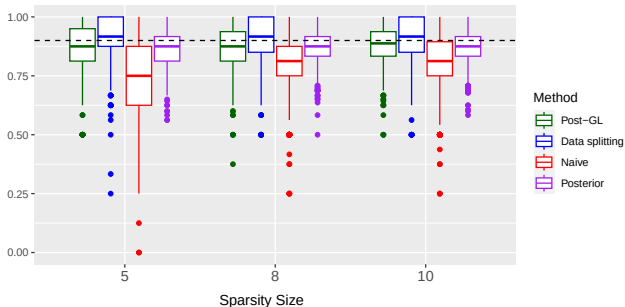
- ▶ ω_n from $N_p(0, \Omega)$, where $\Omega = fH, f = n_2/n_1 \approx 2$

Settings

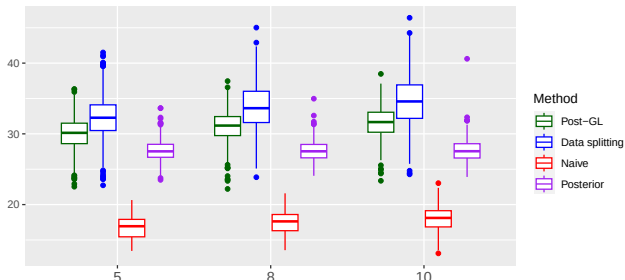
- ▶ $n = 500, p = 200$:
- ▶ 120 continuous variables $\sim N_{120}(0, \Sigma)$ with $\Sigma_{i,j} = 0.3^{|i-j|}$
- ▶ 20 discrete with 5 levels from discrete uniform on $[1, 5]$
- ▶ 500 replicates

sparsity: number of truly nonzero groups; 90% nominal coverage

Coverage rate – Gaussian data

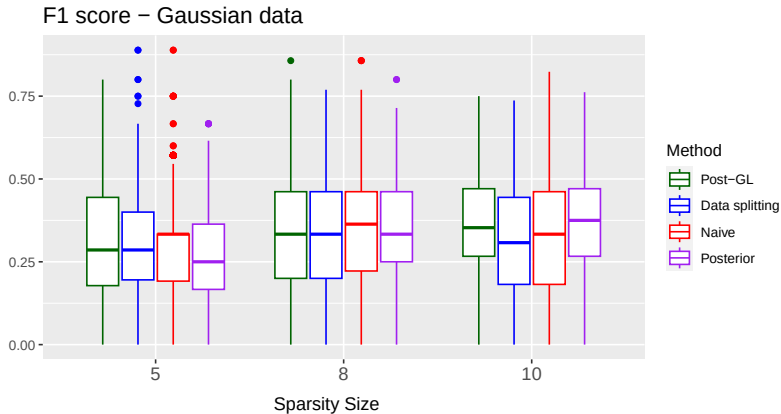


Length of intervals – Gaussian data

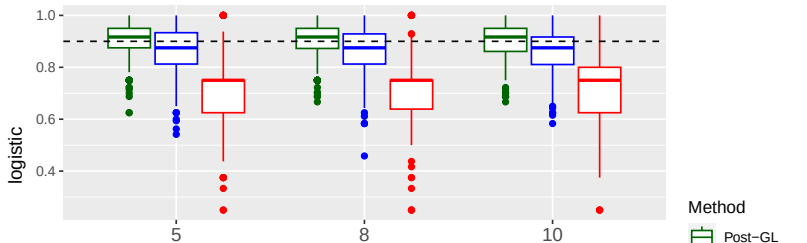


Randomization choice matches with data splitting

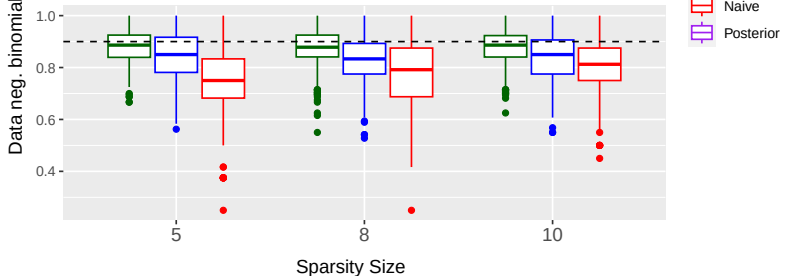
$$\text{F1-score} = \frac{\# \text{True Pos.}}{\# \text{True Pos.} + \frac{1}{2}(\# \text{False Pos.} + \# \text{False Neg.})}$$



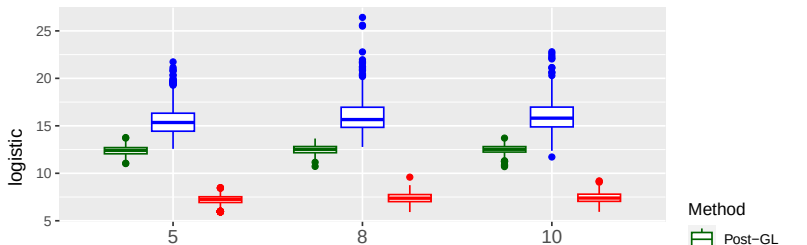
Coverage rate – logistic regression



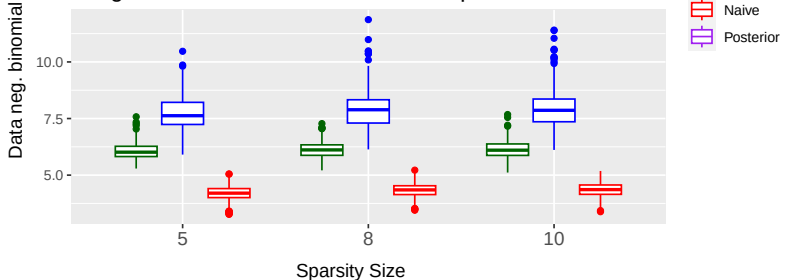
Coverage rate – fit with Poisson quasi-likelihood



Length of intervals – logistic regression



Length of intervals – fit with Poisson quasi-likelihood



Take home messages

- ▶ Method gets correct inference after selection by group lasso
 - ▶ ignoring selection doesn't work
- ▶ Python software available – working on R readable version
- ▶ Versatile method – applicable to:
 - ▶ different distributions for response
 - ▶ discrete and continuous covariates
 - ▶ different estimation methods

This talk was based on:

Y. Huang, S. Pirene, S. Panigrahi and G. Claeskens

Selective inference using randomized group lasso estimators for general models

<https://arxiv.org/pdf/2306.13829.pdf>



Thank you!