Improved estimation through additive adjustment of estimating functions

Ioannis Kosmidis

IKosmidis_

ioannis.kosmidis@warwick.ac.uk

https://www.ikosmidis.com

Professor of Statistics
University of Warwick

20 July 2022

IWSM 2022

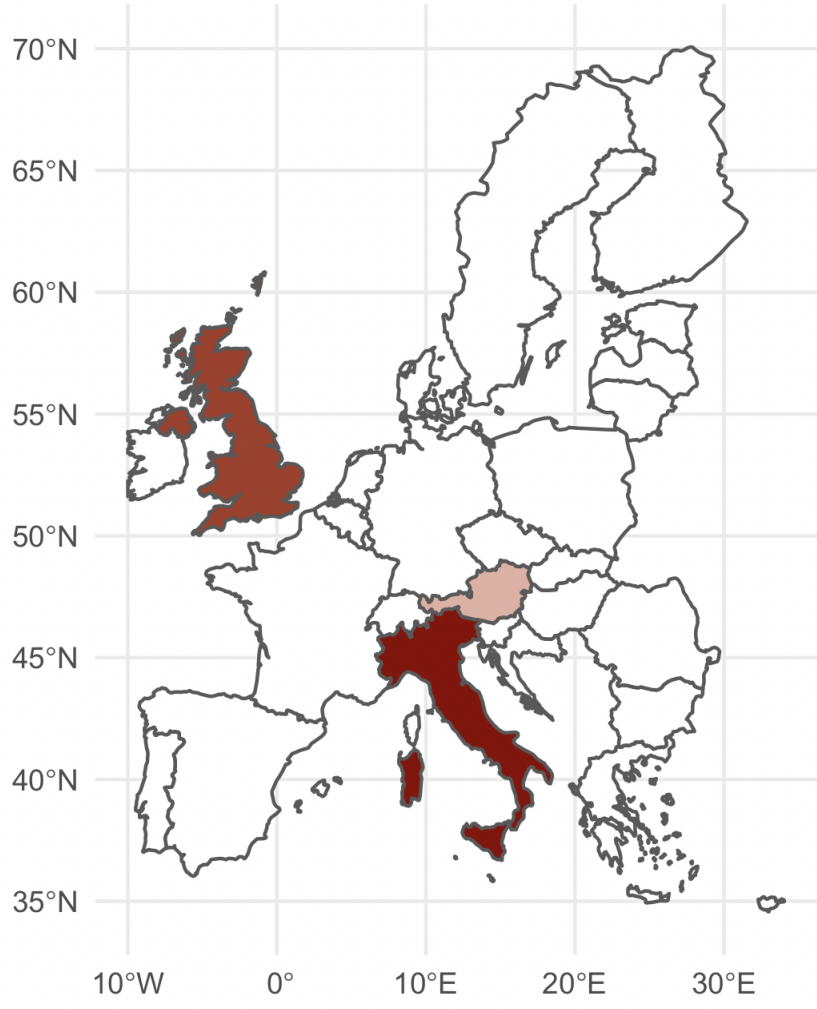
Slides available at:
ikosmidis.com/files/kosmidis_IWSM2022.pdf

Collaborators

C Di Caterina
B Grün
D Firth
A Guolo
CE Kenne Pagui
S Kyriakou
N Lundaron
A Saleh
N Sartori
P Sterzinger
C Varin
A Zeileis
P Zietkiewicz

Collaborators

C Di Caterina
B Grün
D Firth
A Guolo
CE Kenne Pagui
S Kyriakou
N Lundaron
A Saleh
N Sartori
P Sterzinger
C Varin
A Zeileis
P Zietkiewicz



Outline

Outline

Maximum likelihood estimation

Data

$(y_1, x_1^\top), \ldots, (y_k, x_k^\top)$

$x_i = (x_{i1}, \ldots, x_{ik})^\top \in \Re^k$ is a vector of explanatory variables for y_i

Model

Independent random variables $Y_1, \ldots, Y_k$ with pmf/pdf $f_{Y_i}(y_i|x_i; \theta)$

Parameter $\theta \in \Theta \subset \Re^p$

Task

Estimate θ

Key quantities

Log-likelihood[1]

$$\ell(\theta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i|x_i;\theta)$$

Score function

$$s(\theta) = \nabla\ell(\theta)$$

Maximum likelihood (ML) estimator

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell(\theta)$$

Information

Observed: $j(\theta) = -\nabla\nabla^\top\ell(\theta)$

Expected: $i(\theta) = \mathrm{E}_F\left(s(\theta)s(\theta)^\top\right)$

[1] subject to usual regularity conditions; see, Pace and Salvan (1997, §4.3)

Good properties

Equivariance

$g(\hat{\theta})$ is the ML estimator of $g(\theta)$ for general transformations $g(.)$

Asymptotic properties

If the model is adequate, then, under fairly general conditions

Consistency	$\hat{\theta} \xrightarrow{p} \theta$
Asy. mean unbiasedness	$\mathrm{E}_F(\hat{\theta} - \theta) = b(\theta) + O(n^{-2})$, $b(\theta) = O(n^{-1})$
Asy. median unbiasedness	$P(\hat{\theta}_t < \theta_t) = 1/2 + O(n^{-1/2})$
Asy. efficiency	$\mathrm{var}_F(\hat{\theta}) = i(\theta) + O(n^{-2})$, i.e. the approximate variance of $\hat{\theta}$ is the Cramér-Rao bound for unbiased estimators[2]
Asy. normality	$i(\theta)^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N_p(0, I_p)$

[2] "bound" in the sense that $\mathrm{var}_\theta(\tilde{\theta}) - i(\theta)^{-1}$ is a positive semidefinite matrix for any unbiased estimator $\tilde{\theta}$.

Beta regression: A fully-specified model

$Y_1, \ldots, Y_n$, independently distributed, each with density

$$f_i(y; \mu_i, \phi_i) = \frac{\Gamma(\mu_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y)^{(1-\mu_i)\phi_i - 1}$$

with $0 < y < 1$, $0 < \mu_i < 1$, $\phi_i > 0$

Then,

$$\begin{aligned} \mathrm{E}(Y_i; \mu_i, \phi_i) &= \mu_i \\ \mathrm{var}(Y_i; \mu_i, \phi_i) &= \mu_i(1-\mu_i)/(1+\phi_i) \quad (i = 1, \ldots, n) \end{aligned}$$

So, ϕ_i are *precision* parameters.

Beta regression

Link μ_i and ϕ_i to covariates as

$$g_1(\mu_i) = x_i^\top \beta$$
$$g_2(\phi_i) = z_i^\top \gamma$$

where g_1 and g_2 are monotone functions with appropriate domain and range $\Re$

Gasoline yield

A beta regression model with

$$\log \frac{\mu_i}{1-\mu_i} = \beta_0 + \sum_{u=1}^{9} \beta_u s_{iu} + \beta_{10} z_i \quad (i = 1, \ldots, n)$$

is fitted to $n = 32$ observations on gasoline yield (Prater, 1956)

- Response is the proportion of crude oil converted to gasoline after distillation and fractionation
- $s_{i1}, \ldots, s_{i9}$ are the values of 9 binary covariates which represent 10 distinct experimental settings
- z_i is temperature (in F) at which all gasoline has vaporized

Parameters: $\theta^\top = (\beta_0, \beta_1, \ldots, \beta_{10}, \phi)$

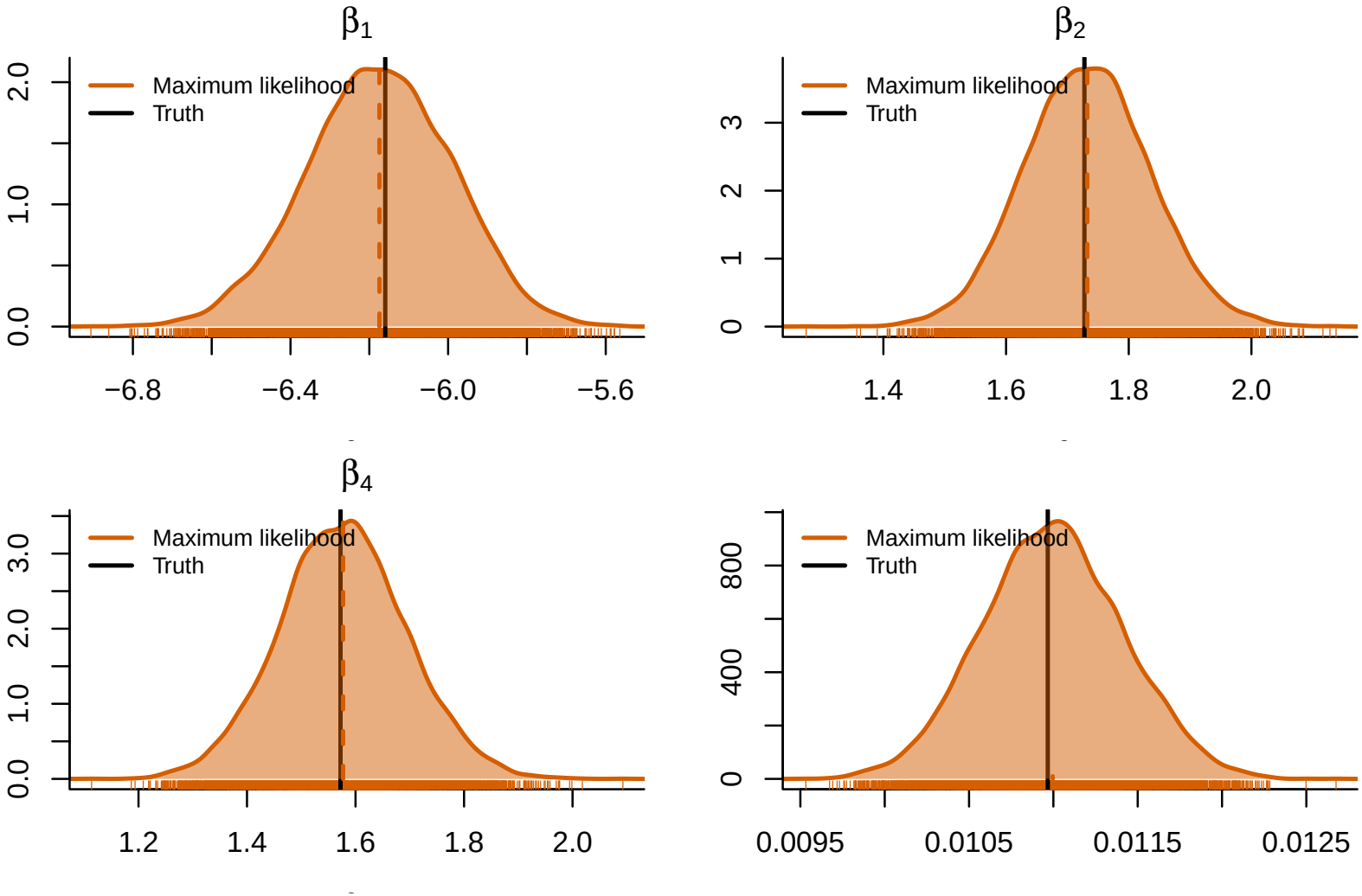
ML estimates

Parameter	ML Estimate		95% Wald CI (Est $\pm$ 1.96 StdErr)	
β_0	-6.160	(0.182)	-6.517	-5.802
β_1	1.728	(0.101)	1.529	1.926
β_2	1.323	(0.118)	1.092	1.554
β_3	1.572	(0.116)	1.345	1.800
β_4	1.060	(0.102)	0.859	1.260
β_5	1.134	(0.104)	0.931	1.337
β_6	1.040	(0.106)	0.832	1.248
β_7	0.544	(0.109)	0.330	0.758
β_8	0.496	(0.109)	0.282	0.709
β_9	0.386	(0.119)	0.153	0.618
β_{10}	0.011	(0.001)	0.010	0.012
ϕ	440.278			

Parenthesized quantities are the est. standard errors based on the diagonal of inverse expected information evaluated at the estimates

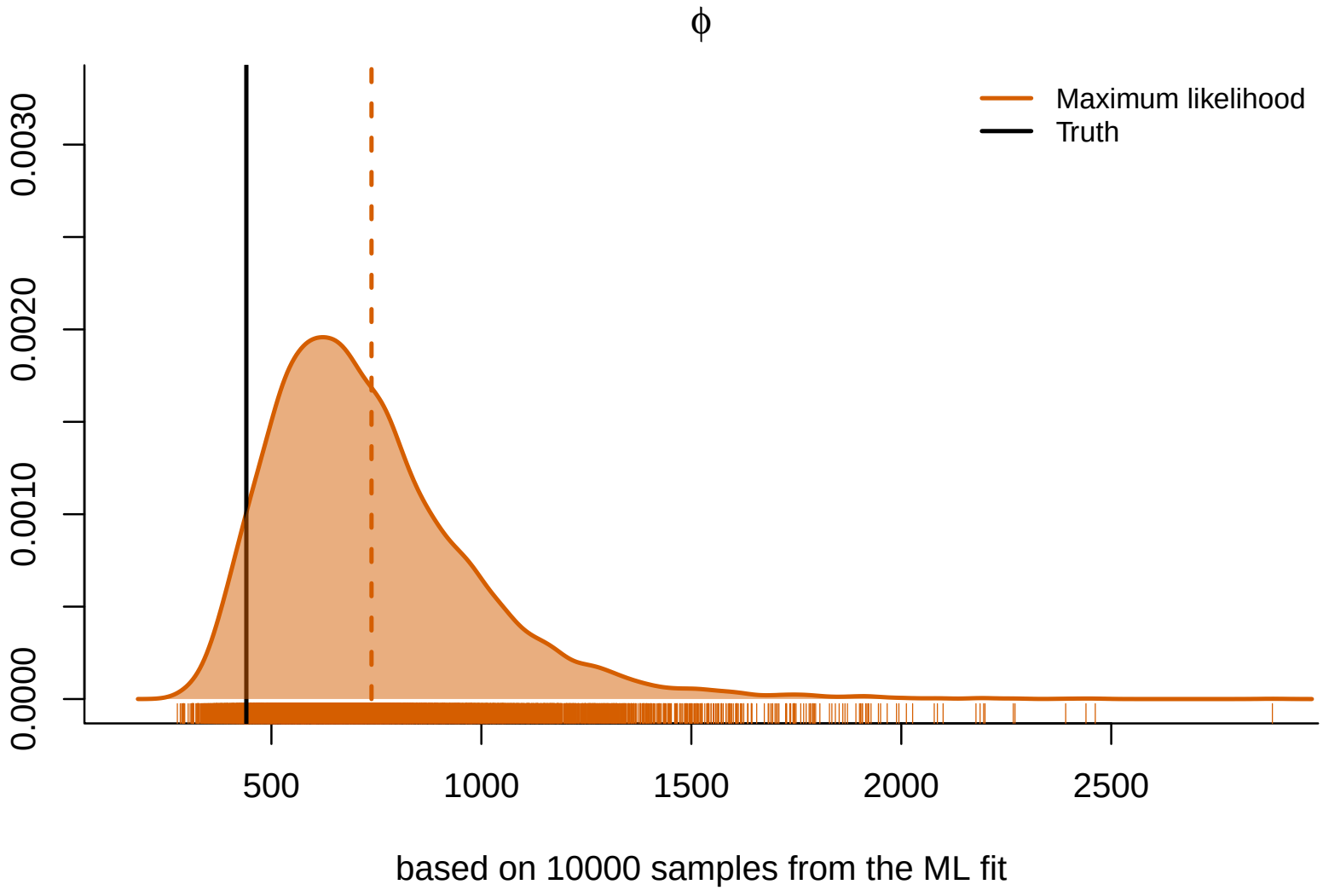
[3] ML using `betareg` (Grün, Kosmidis, and Zeileis, 2012)

Densities of the mean regression estimators

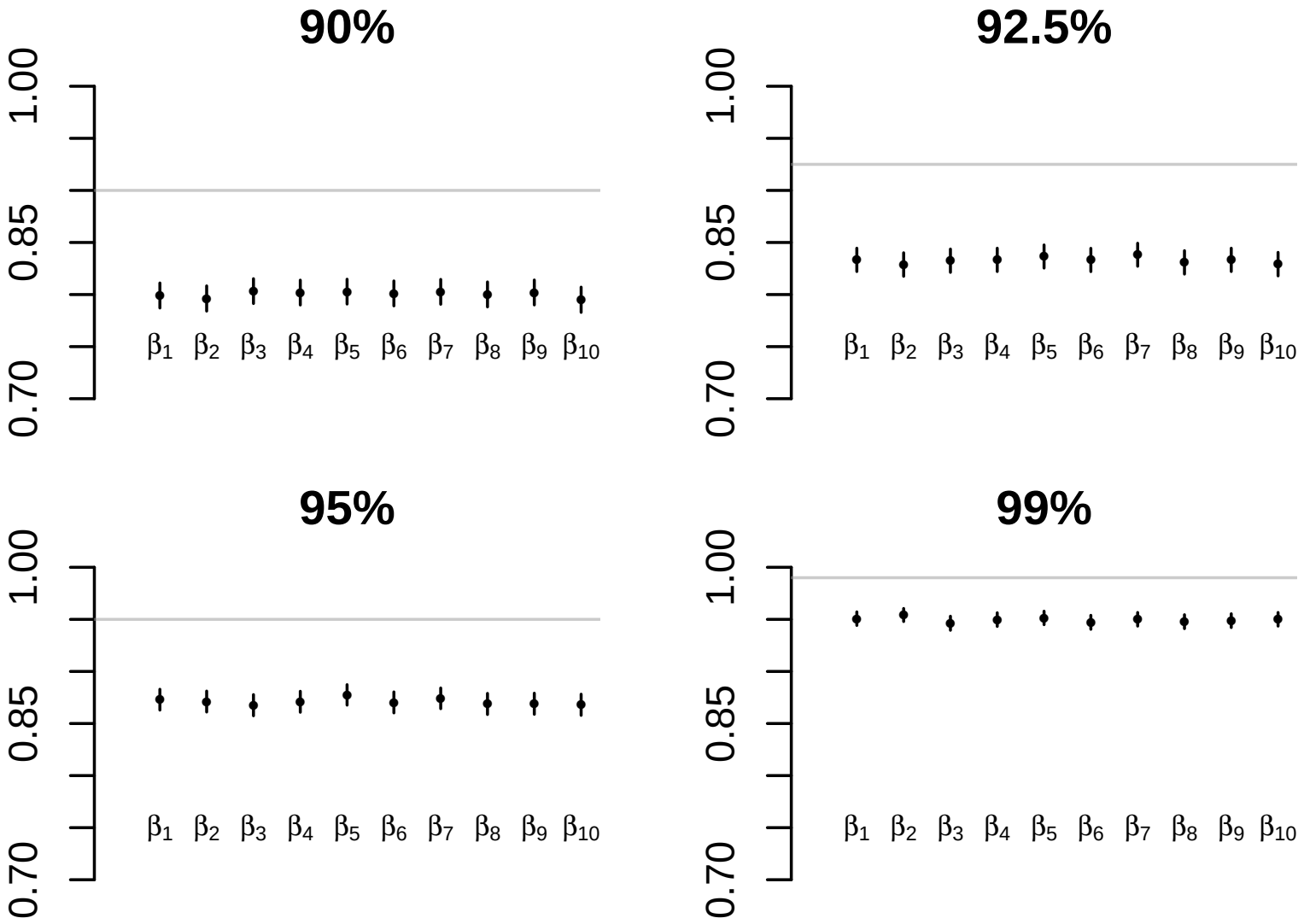


based on 10000 samples from the ML fit

Density of the estimator for ϕ

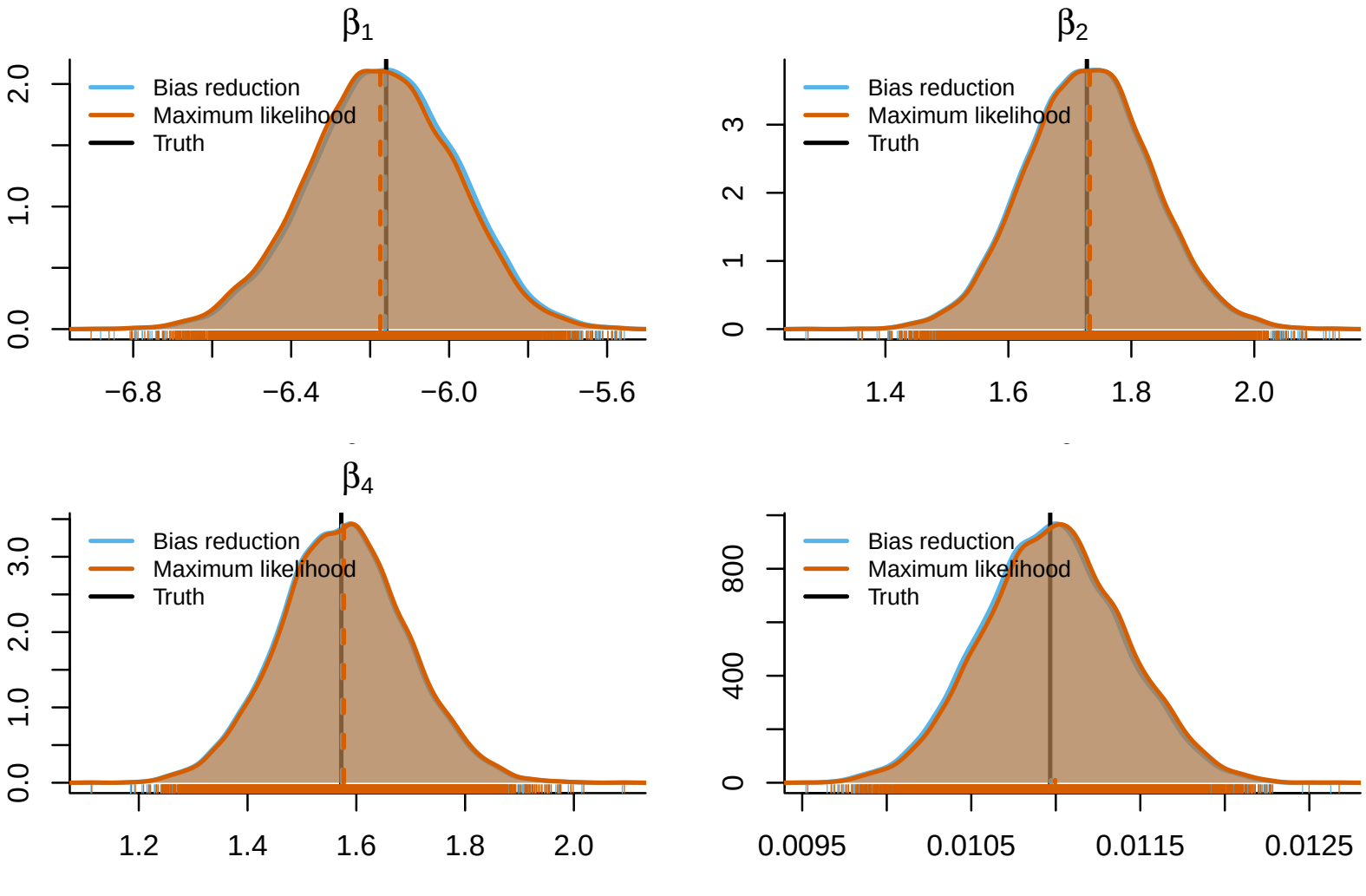


Coverage of Wald-type confidence intervals



10000 samples under the ML fit

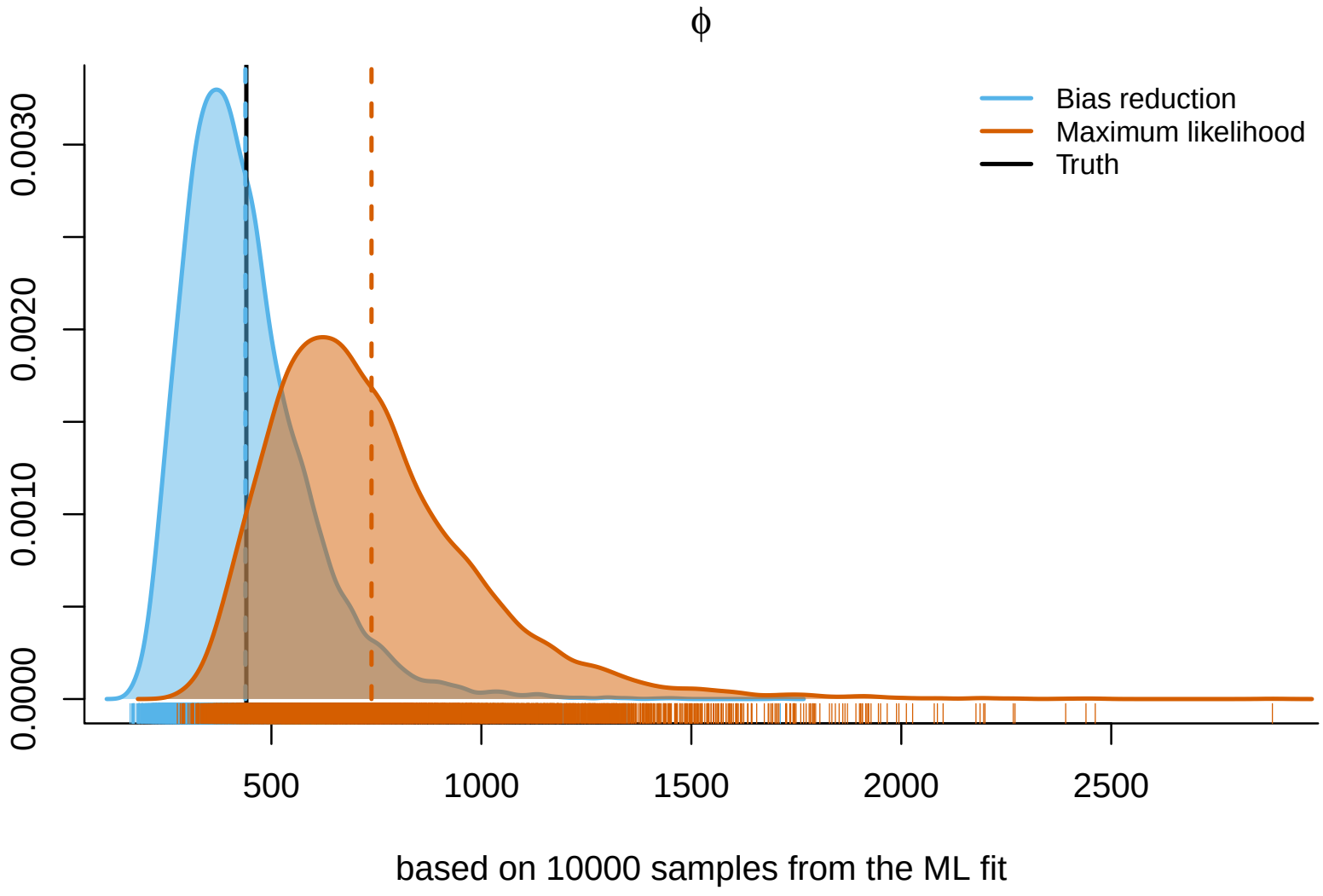
Densities of the mean regression estimators



based on 10000 samples from the ML fit

[4] Bias reduction using `betareg` (Grün, Kosmidis, and Zeileis, 2012)

Density of the estimator for ϕ

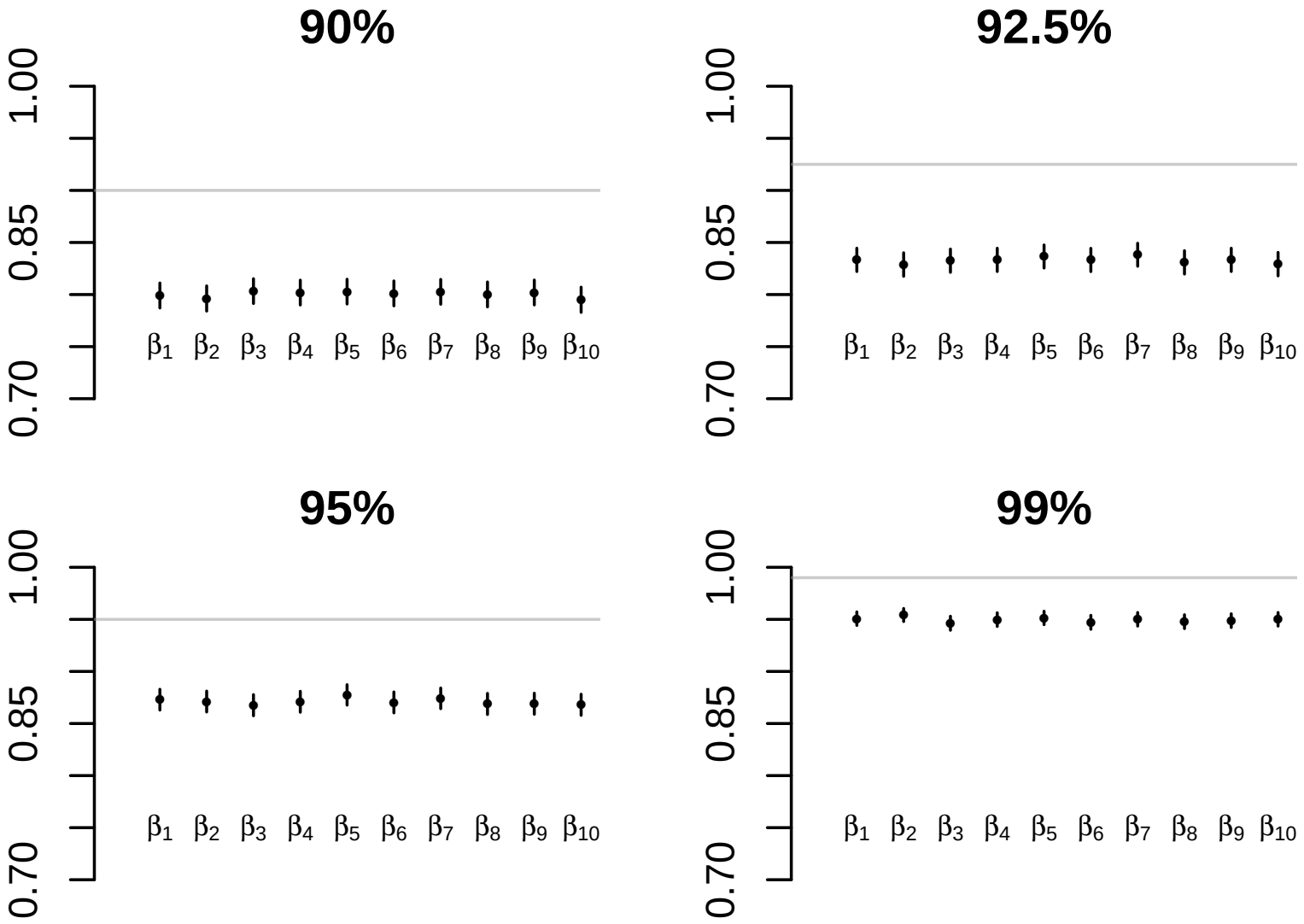


Gasoline yield

	Maximum likelihood		Bias reduction	
β_0	-6.160	(0.182)	-6.142	(0.236)
β_1	1.727	(0.101)	1.723	(0.131)
β_2	1.323	(0.118)	1.319	(0.153)
β_3	1.572	(0.116)	1.567	(0.150)
β_4	1.060	(0.102)	1.057	(0.132)
β_5	1.134	(0.104)	1.130	(0.134)
β_6	1.040	(0.106)	1.037	(0.137)
β_7	0.544	(0.109)	0.542	(0.141)
β_8	0.496	(0.109)	0.494	(0.141)
β_9	0.386	(0.119)	0.385	(0.154)
β_{10}	0.011	(< 0.001)	0.011	(< 0.001)
ϕ	440.278		261.038	

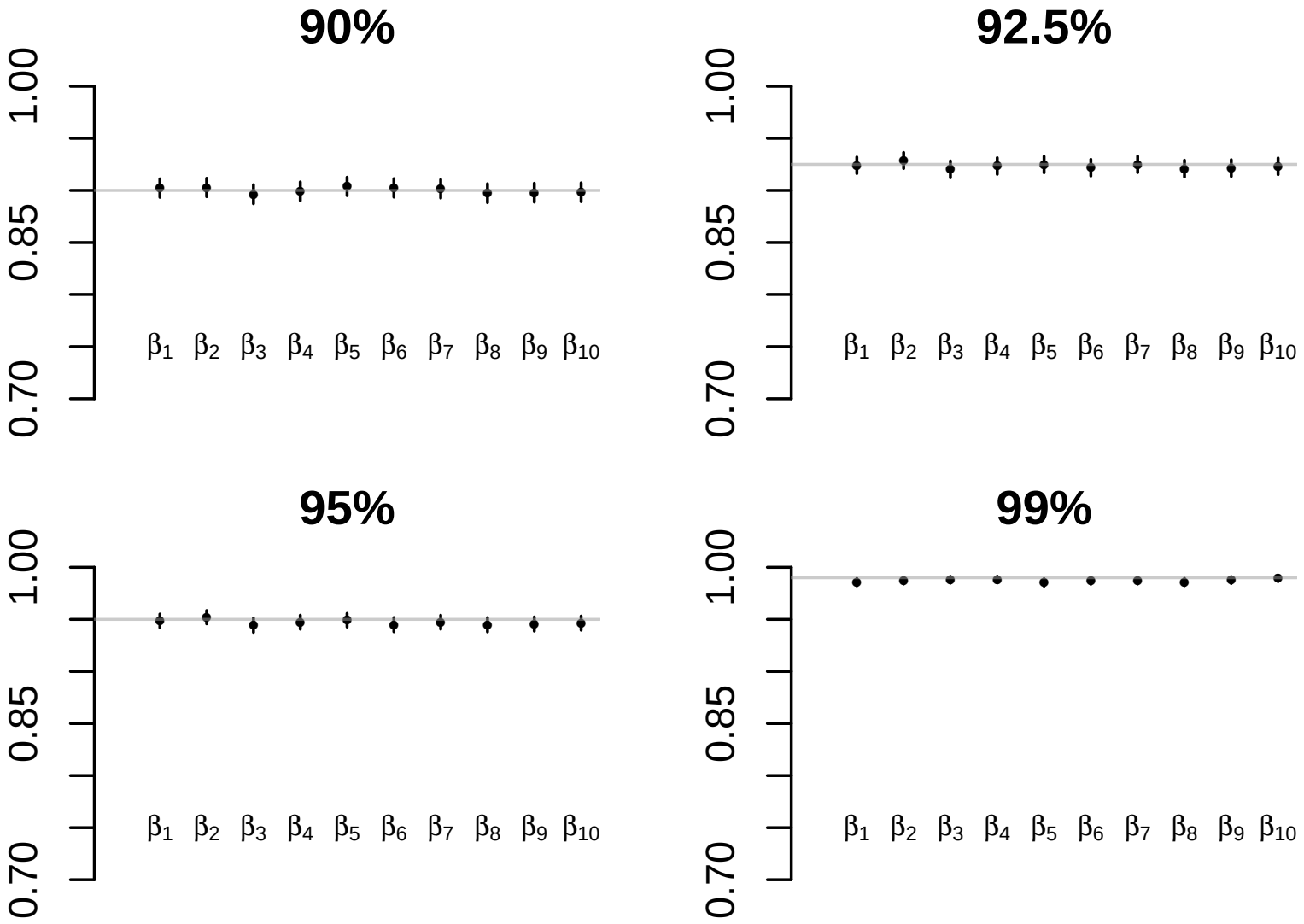
Parenthesized quantities are the est. standard errors based on $\{i(\tilde{\theta})\}^{-1}$

Performance of Wald-type confidence intervals (ML)



10000 samples under the ML fit

Performance of Wald-type confidence intervals (BR)



10000 samples under the ML fit

Adjusted score functions

Key idea[5]

For $A(\theta) = O_p(1)$, define

$$\tilde{\theta} \longleftarrow s(\theta) + A(\theta) = 0$$

Find $A(\theta)$ such that $\tilde{\theta}$ has "better" asymptotic properties than $\hat{\theta}$, aiming to improve its finite sample properties

Instances of better asymptotic properties

Property	$\hat{\theta}$	$\tilde{\theta}$
Mean bias	$O(n^{-1})$	$O(n^{-2})$
Median bias	$1/2 + O(n^{-1/2})$	$1/2 + O(n^{-3/2})$
Efficiency	$i(\theta) + O(n^{-2})$	$i(\theta) + O(n^{-3})$

[5] Used in Firth (1993) for mean bias reduction in ML estimation, and ealrier (see, e.g. Warm, 1989) in more special contexts

Inference using adjusted scores estimators

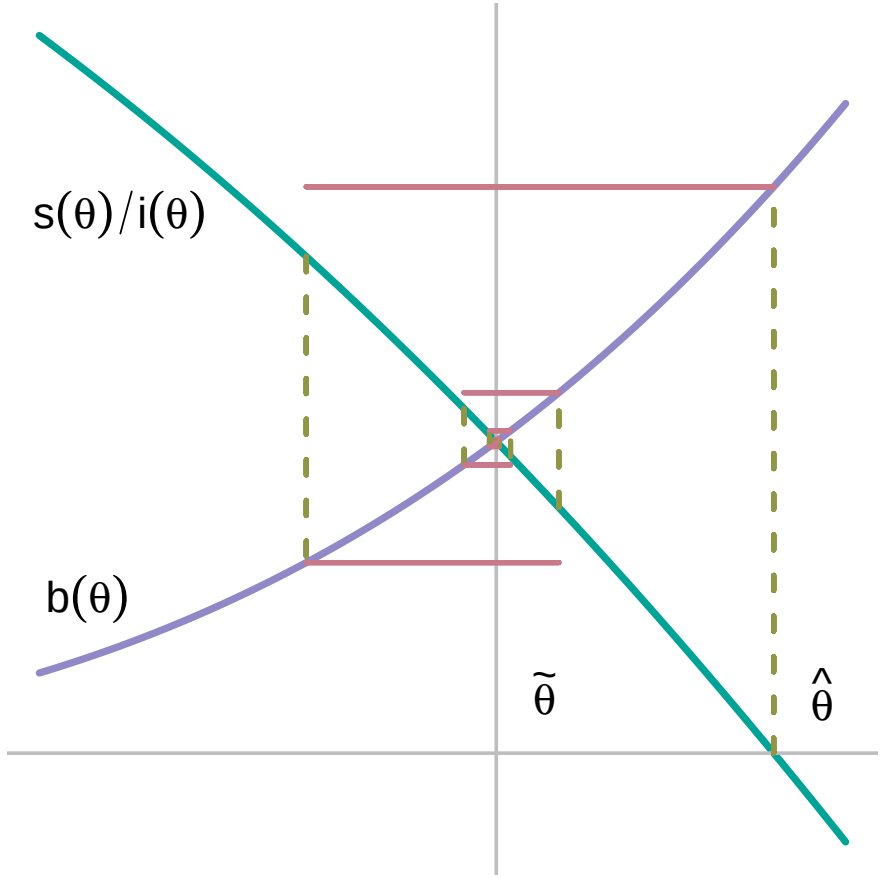
Since $A(\theta) = O_p(1)$ and $s(\theta) = O_p(n^{1/2})$, $\tilde{\theta}$ is consistent and has the same asymptotic distribution as the ML estimator.

$$\Downarrow$$

Usual first-order inferential procedures can be used out of the box:

e.g. Wald-type intervals (with estimated variance $i(\tilde{\theta})$), (adjusted) score tests, model selection using frequentist information-criteria, etc.

Quasi-Fisher scoring for solving adjusted score equations[6]



$$\theta_{(m+1)} := \theta_{(m)} + i\left(\theta_{(m)}\right)^{-1} s\left(\theta_{(m)}\right) - b\left(\theta_{(m)}\right)$$

where $b(\theta) = -i(\theta)^{-1}A(\theta)$

[6] see, Kosmidis and Firth (2010), for quasi-Newton Raphson and quasi-Fisher scoringe in mean bias reduction

Outline

Mean-bias reducing adjusted score equations[8]

Adjustment

The adjusted scores estimator has $\mathrm{E}_F(\tilde{\theta} - \theta) = o(n^{-1})$ if

$$A_t(\theta) = \frac{1}{2}\mathrm{trace}\left[i(\theta)^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right]$$

with $P_t(\theta) = \mathrm{E}_F\left\{s(\theta)s(\theta)^\top s_t(\theta)\right\}$ and $Q_t(\theta) = -\mathrm{E}_F\left\{j(\theta)s_t(\theta)\right\}$

Quasi-Fisher scoring iteration

$$\theta_{(m+1)} := \theta_{(m)} + i\left(\theta_{(m)}\right)^{-1} s\left(\theta_{(m)}\right) - b\left(\theta_{(m)}\right)$$

where $b(\theta) = -i(\theta)^{-1}A(\theta)$ is the $O(n^{-1})$ term in $\hat{\theta}$'s bias expansion[7]

Equivariance

$g(\tilde{\theta})$ is a mean-bias reduced estimator of $g(\theta)$ only if $g(\theta) = C\theta$, for any known real matrix C

[7] Mean-bias reduction can also be achieved using:
i) other bias function estimators in $B^\dagger(\theta)$: iterated bootstrap (Kuk, 1995)
ii) single-step from $\hat{\theta}$: asymptotic bias correction (Efron, 1975), bootstrap, jackknife

[8] see, Firth (1993) and Kosmidis and Firth (2009)

Bias-reducing penalized likelihoods for GLMs

Data

Response vector $y = (y_1, \ldots, y_n)^\top$ and $n \times p$ model matrix X of full rank.

Model

$Y_1, \ldots, Y_n$ are independent from an exponential family of distributions with the ith mean μ_i linked to predictors $\eta_i = x_i^\top \beta$

Theorem[9]

For univariate GLMs with fixed dispersion, there exists a penalized log-likelihood $\ell^*(\beta)$ such that $\nabla \ell^*(\beta) \equiv s(\beta) - i(\beta)b(\beta)$, for all possible specifications of model matrix X, if and only if

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\mathrm{E}(Y_i) \propto \mathrm{var}(Y_i)^k \tag{1}$$

where $k \in \Re$ does not depend on the model parameters.

e.g. canonical links (logistic regression, log-linear models, inverse Gamma regression) and poisson regression with $\eta = \{\mathrm{E}(Y)^\nu - 1\}/\nu$

[9] Kosmidis and Firth (2009)

For logistic regression ($Y_i \sim \text{Bernoulli}(\mu_i)$, $\mu_i = 1/(1 + e^{-\eta_i})$), the reduced bias estimates are

$$\tilde{\beta} = \text{arg max} \left\{ \ell(\beta) + \frac{1}{2} \log |X^T W(\beta) X| \right\} ,$$

where $W(\beta) = \text{diag}\{\pi_1(1 - \pi_1), \ldots, \pi_n(1 - \pi_n)\}$.

Theorem[10]

If the model matrix X is of full rank, the estimates $\tilde{\beta}$ are always finite

[10] see Kosmidis and Firth (2021) for proof and an account of the shrinkage properties of maximum Jeffreys-penalized likelihood in binomial response GLMs

```
R> data("endometrial", package = "brglm2")
R> endo_ML <- glm(HG ~ NV + PI + EH, family = binomial(), data = endometrial)
R> summary(endo_ML)


Call:
glm(formula = HG ~ NV + PI + EH, family = binomial(), data = endometrial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50137  -0.64108  -0.29432   0.00016   2.72777

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.30452    1.63730   2.629 0.008563 **
NV           18.18556 1715.75089   0.011 0.991543
PI           -0.04218    0.04433  -0.952 0.341333
EH           -2.90261    0.84555  -3.433 0.000597 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
AIC: 63.393

Number of Fisher Scoring iterations: 17
```

```
R> library("brglm2")
R> endo_mBR <- update(endo_ML, method = "brglm_fit")
R> summary(endo_mBR)


Call:
glm(formula = HG ~ NV + PI + EH, family = binomial(), data = endometrial,
    method = "brglm_fit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4740  -0.6706  -0.3411   0.3252   2.6123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.77456    1.48869   2.535 0.011229 *
NV           2.92927    1.55076   1.889 0.058902 .
PI          -0.03475    0.03958  -0.878 0.379914
EH          -2.60416    0.77602  -3.356 0.000791 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  56.575  on 75  degrees of freedom
AIC:  64.575

Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
Number of Fisher Scoring iterations: 6
```

$$s(\theta) - i(\theta)b(\theta) = 0$$

Widespread applied usage: in a range of of models (GLMs, GNMs, meta-regression, beta regression, cox regression) and disciplines, mainly because of side-effects of bias reduction in categorical data models

Avoids replication — typically required for bootstrap/jackknife/indirect inference — at the expense of relying heavily on the chosen model

Ample availability of software for specific model classes, e.g. in R

package	model
`brglm`	logistic, probit, cloglog, cauchit regression
`logistf`	logistic regression
`brglm2`	all GLMs
`betareg`	beta regression
`coxphf`	Cox regression
...	

Theoretical extensions: e.g. median bias reduction (Kenne Pagui et al., 2017; Kosmidis et al., 2020)

Median-bias reducing adjusted score equations[11]

Adjustment

The adjusted scores estimator has $P(\hat{\theta}_t < \theta_t) = 1/2 + o(n^{-1})$ if

$$A_t(\theta) = \frac{1}{2}\text{trace}\left[i(\theta)^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right] - i(\theta)R(\theta)$$

where $R_t(\theta) = [i(\theta)^{-1}]_t^\top \tilde{R}_t(\theta)$, with

$$\tilde{R}_{tu}(\theta) = \text{trace}\left[\tilde{i}_u(\theta)\left\{\frac{1}{3}P_t(\theta) + \frac{1}{2}Q_t(\theta)\right\}\right]$$

and $\tilde{i}_u(\theta) = [i(\theta)^{-1}]_u[i(\theta)^{-1}]_u^\top / [i(\theta)^{-1}]_{uu}$

Equivariance

$g(\tilde{\theta}_j)$ is the median-bias reduced estimator of $g(\theta_j)$ for any one-to-one function $g(\cdot)$ $(j = 1, \ldots, p)$

[11] see, Kenne Pagui et al. (2017)

IWLS for mean and median bias reduction[12]

Quasi-Fisher scoring is equivalent to an IWLS step for β

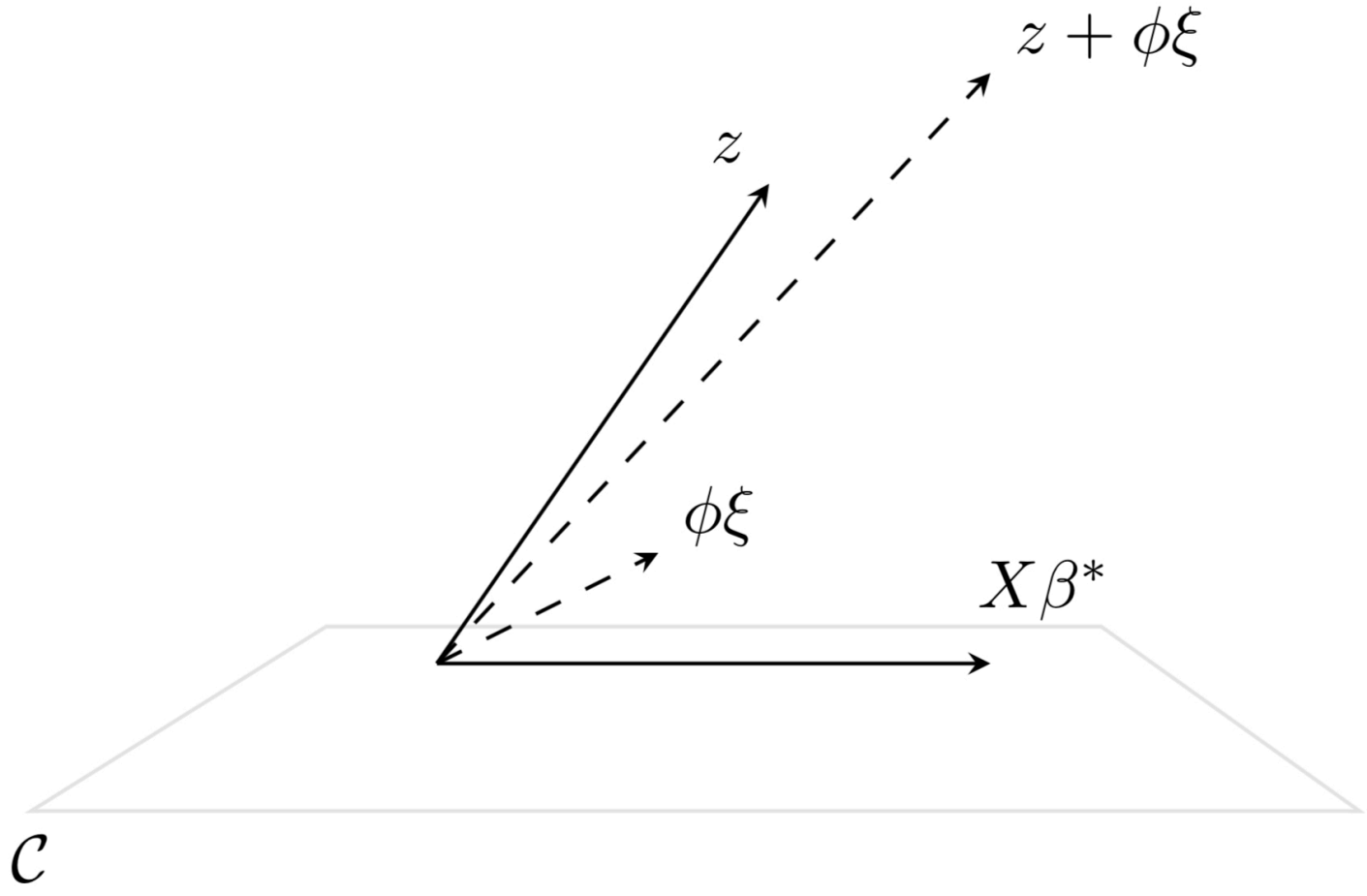
$$\beta^{(j+1)} \leftarrow \left(X^\top W^{(j)} X\right)^{-1} X^\top W^{(j)} \left(z^{(j)} + \phi^{(j)}\xi^{(j)} + \phi^{(j)} X u^{(j)}\right)$$

In contrast with the ML estimator, β and ϕ are updated simultaneously

Kosmidis et al. (2020) give closed-form expressions for ξ and u, and the mean and median BR updates in ϕ

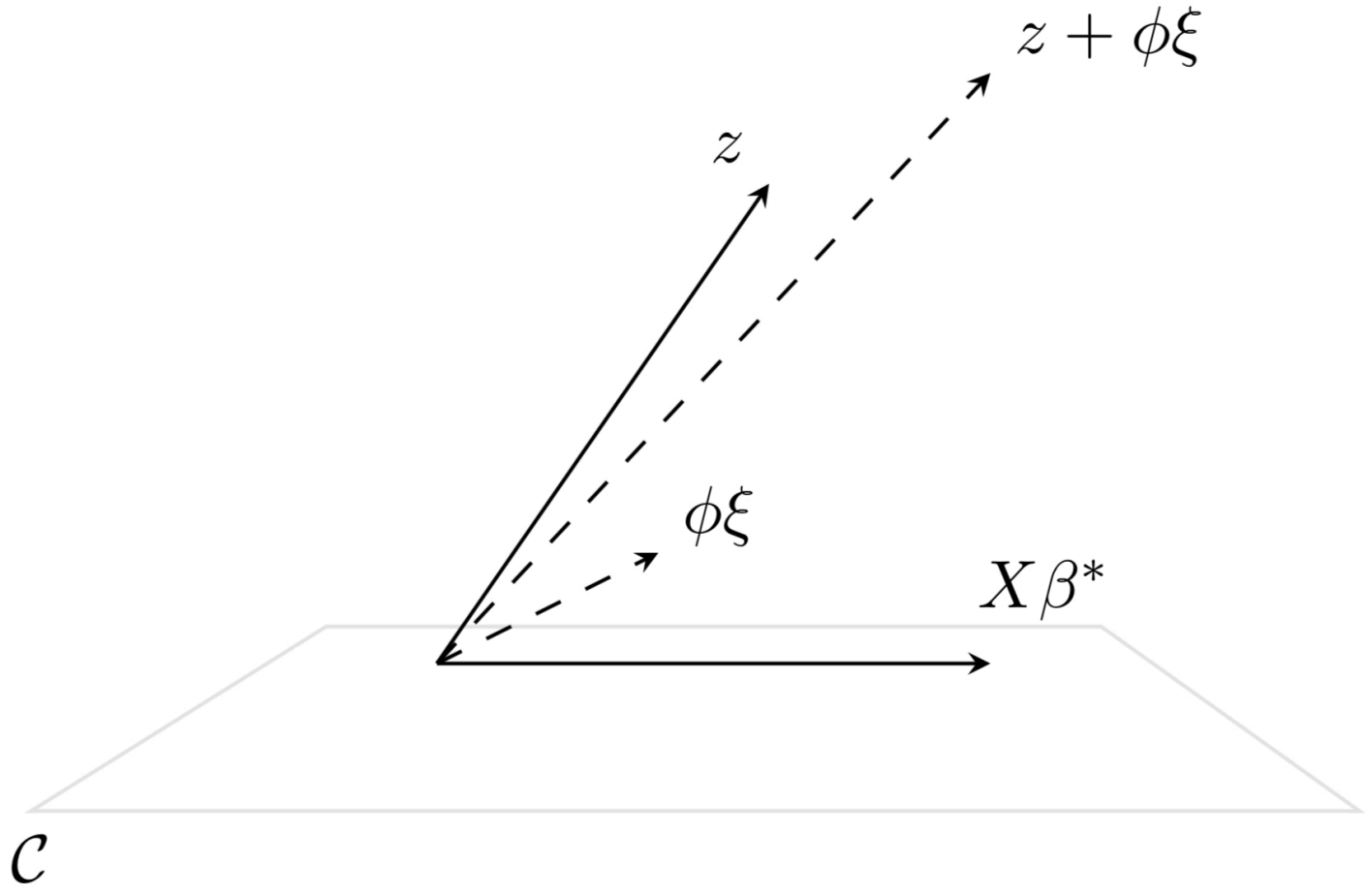
[12] see, Kosmidis et al. (2020)

IWLS step for mean bias reduction



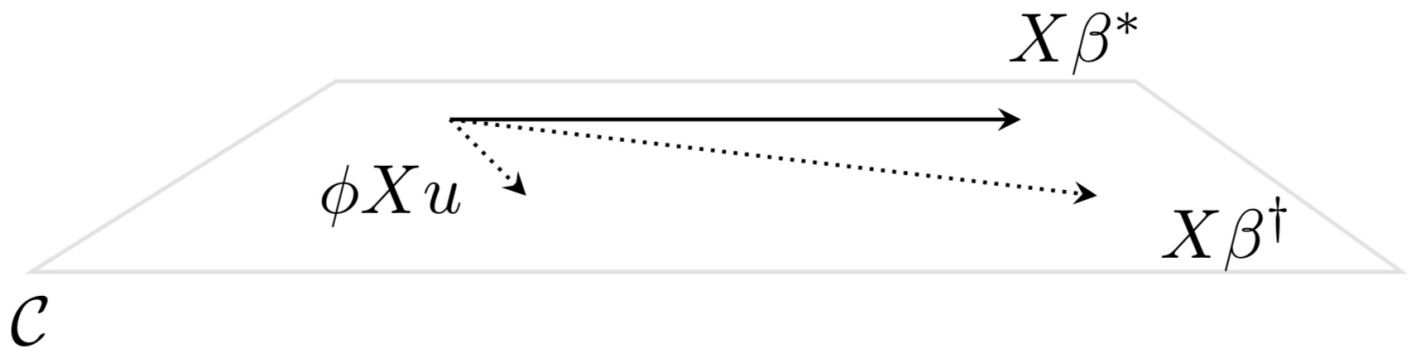
(every quantity is pre-multiplied by $W^{1/2}$)

IWLS step for median bias reduction



(every quantity is pre-multiplied by $W^{1/2}$)

IWLS step for median Bias reduction



(every quantity is pre-multiplied by $W^{1/2}$)

```
R> library("brglm2")
R> endo_mBR <- update(endo_ML, method = "brglm_fit")
R> summary(endo_mBR)


Call:
glm(formula = HG ~ NV + PI + EH, family = binomial(), data = endometrial,
    method = "brglm_fit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4740  -0.6706  -0.3411   0.3252   2.6123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.77456    1.48869   2.535 0.011229 *
NV           2.92927    1.55076   1.889 0.058902 .
PI          -0.03475    0.03958  -0.878 0.379914
EH          -2.60416    0.77602  -3.356 0.000791 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  56.575  on 75  degrees of freedom
AIC:  64.575

Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
Number of Fisher Scoring iterations: 6
```

```
R> library("brglm2")
R> endo_mdBR <- update(endo_ML, method = "brglm_fit", type = "AS_median")
R> summary(endo_mdBR)


Call:
glm(formula = HG ~ NV + PI + EH, family = binomial(), data = endometrial,
    method = "brglm_fit", type = "AS_median")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4683  -0.6588  -0.3204   0.2083   2.6530

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.96936    1.55232   2.557 0.010557 *
NV           3.86921    2.29824   1.684 0.092269 .
PI          -0.03868    0.04187  -0.924 0.355569
EH          -2.70793    0.80301  -3.372 0.000746 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.868  on 75  degrees of freedom
AIC:  63.868

Type of estimator: AS_median (median bias-reducing adjusted score equations)
Number of Fisher Scoring iterations: 8
```

Outline

$$s(\theta) - i(\theta)b(\theta) = 0$$

Intractable or infeasible likelihoods

Impractical due to the requirement to compute expectations under the chosen model

Other objectives (e.g. composite likelihoods), estimating functions

Theory does not apply; requires fully-specified models

$$s(\theta) - i(\theta)b(\theta) = 0$$

Intractable or infeasible likelihoods

Impractical due to the requirement to compute expectations under the chosen model

Other objectives (e.g. composite likelihoods), estimating functions

Theory does not apply; requires fully-specified models

$b(\theta)$**: first-term in the bias expansion of the MLE**

Typically, requires substantial algebraic/implementation effort

Beta regression

Detailed expressions for $s(\theta)$, $i(\theta)$, $P_t(\theta) + Q_t(\theta)$ for beta regressions are in Grün, Kosmidis, and Zeileis (2012)

$$P_t(\theta) + Q_t(\theta) = \begin{bmatrix} V_{\beta\beta,t} & V_{\beta\gamma,t} \\ V_{\beta\gamma,t}^\top & V_{\gamma\gamma,t} \end{bmatrix} \quad (t = 1, \ldots, p)\,,$$

$$P_{p+s}(\theta) + Q_{p+s}(\theta) = \begin{bmatrix} W_{\beta\beta,s} & W_{\beta\gamma,s} \\ W_{\beta\gamma,s}^\top & W_{\gamma\gamma,s} \end{bmatrix} \quad (s = 1, \ldots, q)\,,$$

$$\begin{aligned}
V_{\beta\beta,t} &= X^\top \Phi^{\mathbf{2}} D_{\mathbf{1}} \left(\Phi D_{\mathbf{1}}^{\mathbf{2}} K_{\mathbf{3}} + D_{\mathbf{1}}' K_{\mathbf{2}}\right) X_t^{\mathrm{D}} X\,,\\
V_{\beta\gamma,t} &= X^\top \Phi D_{\mathbf{1}}^{\mathbf{2}} D_{\mathbf{2}} \left\{\Phi \left(M K_{\mathbf{3}} + \Psi_{\mathbf{2}}\right) + K_{\mathbf{2}}\right) X_t^{\mathrm{D}} Z\,,\\
V_{\gamma\gamma,t} &= Z^\top \Phi D_{\mathbf{1}} \left\{D_{\mathbf{2}}^{\mathbf{2}} \left(M^{\mathbf{2}} K_{\mathbf{3}} + 2M\Psi_{\mathbf{2}} - \Psi_{\mathbf{2}}\right) + D_{\mathbf{2}}' \left(M K_{\mathbf{2}} - \Psi_{\mathbf{1}}\right)\right\} X_t^{\mathrm{D}} Z\\
W_{\beta\beta,s} &= X^\top \Phi D_{\mathbf{2}} \left\{\Phi D_{\mathbf{1}}^{\mathbf{2}} \left(M K_{\mathbf{3}} + \Psi_{\mathbf{2}}\right) + D_{\mathbf{1}}' \left(M K_{\mathbf{2}} - \Psi_{\mathbf{1}}\right)\right\} Z_s^{\mathrm{D}} X\,,\\
W_{\beta\gamma,s} &= X^\top D_{\mathbf{1}} D_{\mathbf{2}}^{\mathbf{2}} \left\{\Phi \left(M^{\mathbf{2}} K_{\mathbf{3}} + 2M\Psi_{\mathbf{2}} - \Psi_{\mathbf{2}}\right) + M K_{\mathbf{2}} - \Psi_{\mathbf{1}}\right\} Z_s^{\mathrm{D}} Z\,,\\
W_{\gamma\gamma,s} &= Z^\top D_{\mathbf{2}}^{\mathbf{3}} \left\{M^{\mathbf{3}} K_{\mathbf{3}} + \left(3M^{\mathbf{2}} - 3M + 1_n\right) \Psi_{\mathbf{2}} - \Omega_{\mathbf{2}}\right\} Z_s^{\mathrm{D}} Z\\
&\qquad + Z^\top D_{\mathbf{2}} D_{\mathbf{2}}' \left\{M^{\mathbf{2}} K_{\mathbf{2}} + \Psi_{\mathbf{1}} - 2M\Psi_{\mathbf{1}} - \Omega_{\mathbf{1}}\right\} Z_s^{\mathrm{D}} Z\,.
\end{aligned}$$

$$s(\theta) - i(\theta)b(\theta) = 0$$

Intractable or infeasible likelihoods

Impractical due to the requirement to compute expectations under the chosen model

Other objectives (e.g. composite likelihoods), estimating functions

Theory does not apply

$b(\theta)$**: first-term in the bias expansion of the MLE**

Typically, requires substantial algebraic/implementation effort

[13] see, Firth (1993) and Kosmidis and Firth (2009) for iff for GLMs

$$s(\theta) - i(\theta)b(\theta) = 0$$

Intractable or infeasible likelihoods

Impractical due to the requirement to compute expectations under the chosen model

Other objectives (e.g. composite likelihoods), estimating functions

Theory does not apply

$b(\theta)$**: first-term in the bias expansion of the MLE**

Typically, requires substantial algebraic/implementation effort

Bias-reducing penalized likelihoods

A bias-reducing penalized likelihood with gradient $s(\theta) - i(\theta)b(\theta)$ exists only for special models[13]

e.g. for exponential families, bias reduction through Jeffreys penalties

[13] see, Firth (1993) and Kosmidis and Firth (2009) for iff for GLMs

Setting

M-estimator of a parameter vector $\theta \in \Theta \subset \Re^p$

$$\hat{\theta} \leftarrow \sum_{i=1}^{k} \psi^i(\hat{\theta}) = 0$$

where $\psi^i(\theta) = \psi(\theta, Y_i, x_i)$ and $\psi^i(\theta) = (\psi^i_1(\theta), \ldots, \psi^i_p(\theta))^\top$

Data

Realizations of random vectors $Y_1, \ldots, Y_k$ with $Y_i = (Y_{i1}, \ldots, Y_{ic_i})^\top \in \mathcal{Y} \subset \Re^{c_i}$

Possibly, a sequence of covariate vectors $x_1, \ldots, x_k$ with $x_i = (x_{i1}, \ldots, x_{iq_i})^\top \in \mathcal{X} \subset \Re^{q_i}$

$Y_i | x_i$ from an unknown data-generating process G

Modelling regimes covered

Fully-parametric models (e.g. GLMs, GLMMs, Beta regression, ...)

Quasi-likelihoods[14]

Generalized estimating equations[15]

Composite likelihoods[16]

...

[14] see Wedderburn (1974) and McCullagh (1983)

[15] see Liang and Zeger (1986)

[16] see Varin et al. (2011) for an overview

Empirical mean-bias reducing adjustments

Adjustment

Under fairly standard regularity assumptions for M-esitmation[17], $\tilde{\theta} \leftarrow \sum_{i=1}^{k} \psi^i(\hat{\theta}) + A(\theta) = 0$ has $\mathrm{E}_G(\tilde{\theta} - \bar{\theta}) = O(n^{-3/2})$ if

$$A_t(\theta) = -\mathrm{trace}\left\{j(\theta)^{-1} d_t(\theta)\right\} - \frac{1}{2}\mathrm{trace}\left[j(\theta)^{-1} e(\theta)\left\{j(\theta)^{-1}\right\}^\top u_t(\theta)\right]$$

$j(\theta)$ is the matrix with sth row $-\sum_{i=1}^{k} \nabla \psi_s^i(\theta)$, assumed invertible

$u_t(\theta) = \sum_{i=1}^{k} \nabla\nabla^\top \psi_t^i(\theta)$

$e(\theta) = \sum_{i=1}^{k} \psi^i(\theta)\{\psi^i(\theta)\}^\top$

$d_r(\theta) = \sum_{i=1}^{k} \nabla \psi_r^i(\theta)\psi^i(\theta)$

[17] see, Kosmidis and Lunardon (2021) for RBM-estimation and assumptions

Ratio of two means

Estimate $\theta = \mu_Y/\mu_X$ from pairs $(x_1, y_1)^\top, \ldots, (x_n, y_n)^\top$ where $\mu_X = E_G(X_i) \neq 0$ and $\mu_Y = E_G(Y_i)$

M-estimator (intuitive by WLLN + continuous mapping theorem)

$$\hat{\theta} = \frac{s_Y}{s_X} \leftarrow \sum_{i=1}^{n}(y_i - \theta x_i) = 0 \quad \text{where} \quad s_Z = \sum_{i=1}^{n} Z_i$$

$\hat{\theta}$ is biased for finite samples; bias, typically, depends on $g(x_i, y_i)$[18]

RBM-estimator (RB = Reduced-Bias)

$$\tilde{\theta} = \frac{s_Y + s_{XY}/s_X}{s_X + s_{XX}/s_X} \quad \text{where} \quad s_{ZW} = \sum_{i=1}^{n} Z_i W_i$$

Robustness side-effects

As $s_x \to 0$, $\hat{\theta}$ diverges, while $\tilde{\theta} \to s_{XY}/s_{XX}$

[18] e.g. Durbin (1959)

Asymptotic distribution

Empirical bias-reducing adjustment is small

$$-\text{trace}\left\{j(\theta)^{-1}d_t(\theta)\right\} - \frac{1}{2}\text{trace}\left[j(\theta)^{-1}e(\theta)\left\{j(\theta)^{-1}\right\}^\top u_t(\theta)\right] = O_p(1)$$

$$\Downarrow$$

Asymptotic normality

$$Q(\bar{\theta})^{1/2}(\tilde{\theta} - \bar{\theta}) \xrightarrow{d} N_p(0, I)$$

with

$Q(\theta)^{-1} = B(\theta)^{-1}M(\theta)\{B(\theta)^{-1}\}^\top$

$M(\theta) = E_G(e(\theta))$

$B(\theta) = E_G(j(\theta))$

Asymptotic distribution

Empirical bias-reducing adjustment is small

$$-\text{trace}\left\{j(\theta)^{-1}d_t(\theta)\right\} - \frac{1}{2}\text{trace}\left[j(\theta)^{-1}e(\theta)\left\{j(\theta)^{-1}\right\}^\top u_t(\theta)\right] = O_p(1)$$

$$\Downarrow$$

Asymptotic normality

$$Q(\bar{\theta})^{1/2}(\tilde{\theta} - \bar{\theta}) \xrightarrow{d} N_p(0, I)$$

with

$Q(\theta)^{-1} = B(\theta)^{-1}M(\theta)\{B(\theta)^{-1}\}^\top$

$M(\theta) = E_G(e(\theta))$

$B(\theta) = E_G(j(\theta))$

Empirical variance-covariance matrix for $\tilde{\theta}$

$$\hat{V}(\tilde{\theta}) = j(\tilde{\theta})^{-1}e(\tilde{\theta})\left\{j(\tilde{\theta})^{-1}\right\}^\top$$

Bias-reducing penalties to objective functions

Suppose that $\sum_{i=1}^{k} \psi^i(\theta) = \nabla \ell(\theta)$

Empirical bias-reducing penalized objective functions

Then, the adjusted score functions $\sum_{i=1}^{k} \psi^i(\theta) + D(\theta)$ are the gradient of

$$\ell(\theta) - \frac{1}{2}\mathrm{trace}\left\{j(\theta)^{-1}e(\theta)\right\}$$

whose maximizer is $\tilde{\theta}$

Bias reduction and model selection

$\hat{\theta}$ is the maximum (composite) likelihood estimator and $\ell(\theta)$ is the logogarith of the (composite) likelihood[19]

Takeuchi Information criterion (TIC)

$$-2\ell(\hat{\theta}) + 2\mathrm{trace}\left\{j(\hat{\theta})^{-1}e(\hat{\theta})\right\}$$

TIC model selection by selecting the model with largest

$$\ell(\hat{\theta}) - \mathrm{trace}\left\{j(\hat{\theta})^{-1}e(\hat{\theta})\right\}$$

Empirical bias-reducing penalized objective functions

$$\ell(\theta) - \frac{1}{2}\mathrm{trace}\left\{j(\theta)^{-1}e(\theta)\right\}$$

TIC with reduced-bias estimates is still consistent

$$\ell(\tilde{\theta}) - \mathrm{trace}\left\{j(\tilde{\theta})^{-1}e(\tilde{\theta})\right\}$$

[19] See Varin and Vidoni (2005) for when $\ell(\theta)$ is a composite likelihood and CLIC

Easy implementation

Calculation of $A_t(\theta)$ or of the penalty to the objective requires only the contributions to the estimating functions / objective and their derivatives

Ingredients for general implementation[20]

Implementation of the estimating functions / objective

Software for automatic differentiation[21]

e.g. `RcppEigenAD` in R or `ForwardDiff`, `ReverseDiff` in Julia

Matrix multiplication

A general method for solving nonlinear systems of equations

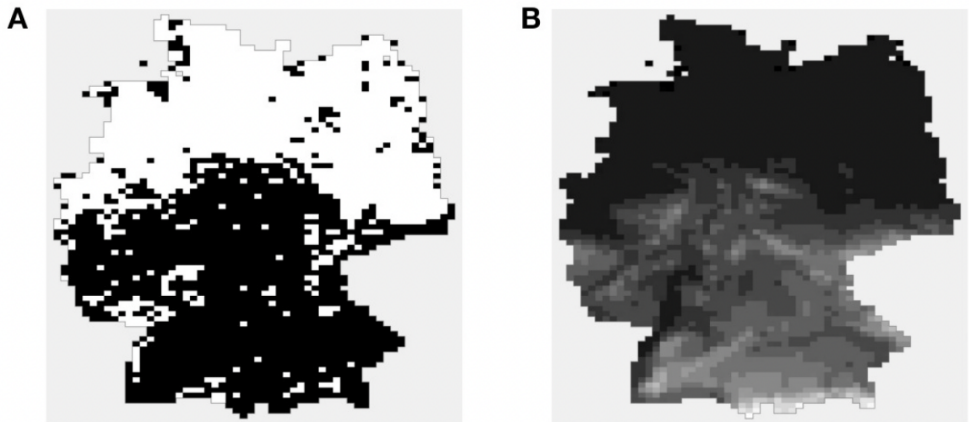
e.g. `nleqslv` in R, or `NLsolve` in Julia

[20] See the `MEstimation` Julia package at `github.com/ikosmidis/MEstimation.jl`

[21] see, e.g. Griewank and Walther (2008)

Autologistic regression

Data: $y(s) \in \{-1, 1\}$ and covariates $x(s) \in \Re^p$ at locations $s \in \mathcal{S} = \{s_{11}, \ldots, s_{1c_1}, \ldots, s_{k1}, \ldots, s_{kc_k}\}$



Hydrocotyle vulgaris data: (A) Observed presence/absence data (white indicates presence, which is taken to be the "high" level). (B) Value of the covariate, altitude.[22]

[22] Picture from Wolters (2017), who also provides an authoritative overview of autologistic regression

Autologistic regression[24]

Data: $y(s) \in \{-1, 1\}$ and covariates $x(s) \in \Re^p$ at locations $s \in \mathcal{S} = \{s_{11}, \ldots, s_{1c_1}, \ldots, s_{k1}, \ldots, s_{kc_k}\}$

Model: $f(y \mid \{y(u) : u \in G(s)\}, x(s), \theta) = \dfrac{e^{y\zeta(s)}}{e^{-\zeta(s)} + e^{\zeta(s)}}$

with $\zeta(s) = x(s)^\top \beta + \lambda \sum_{u \in G(s)} y(u)$

Composite conditional likelihood[23]

$$\exp\{\ell(\theta)\} = \prod_{i=1}^{k} \prod_{j=1}^{c_i} f(y(s_{ij}) \mid \{Y(u) : u \in G(s_{ij})\}, x(s_{ij}), \theta)$$

[23] aka Besag's pseudo-likelihood (Besag, 1975)

[24] See, Wolters (2017) for an authoritative overview on autologistic regression

Autologistic regression[24]

Data: $y(s) \in \{-1, 1\}$ and covariates $x(s) \in \Re^p$ at locations $s \in \mathcal{S} = \{s_{11}, \ldots, s_{1c_1}, \ldots, s_{k1}, \ldots, s_{kc_k}\}$

Model: $f(y \mid \{y(u) : u \in G(s)\}, x(s), \theta) = \dfrac{e^{y\zeta(s)}}{e^{-\zeta(s)} + e^{\zeta(s)}}$

with $\zeta(s) = x(s)^\top \beta + \lambda \sum_{u \in G(s)} y(u)$

Composite conditional likelihood[23]

$$\exp\{\ell(\theta)\} = \prod_{i=1}^{k} \prod_{j=1}^{c_i} f(y(s_{ij}) \mid \{Y(u) : u \in G(s_{ij})\}, x(s_{ij}), \theta)$$

Bias-reducing penalized objective with plug-in penalty

$$\ell(\theta) - \frac{1}{2}\mathrm{trace}\left\{j(\theta)^{-1} e(\theta)\right\} + \frac{1}{\sum_{j=1}^{k} c_i} \log|\tilde{X}^\top W(\theta) \tilde{X}|$$

where $\tilde{x}(s) = (2x(s)^\top, 2\sum_{u \in G(s)} y(u))^\top$, $W(\theta)$ has diagonal elements $\pi(s_{11})\{1 - \pi(s_{11})\}, , \ldots, \pi(s_{kc_k})\{1 - \pi(s_{kc_k})\}$, $\pi(s) = 1/\{1 + e^{-2\zeta(s)}\}$

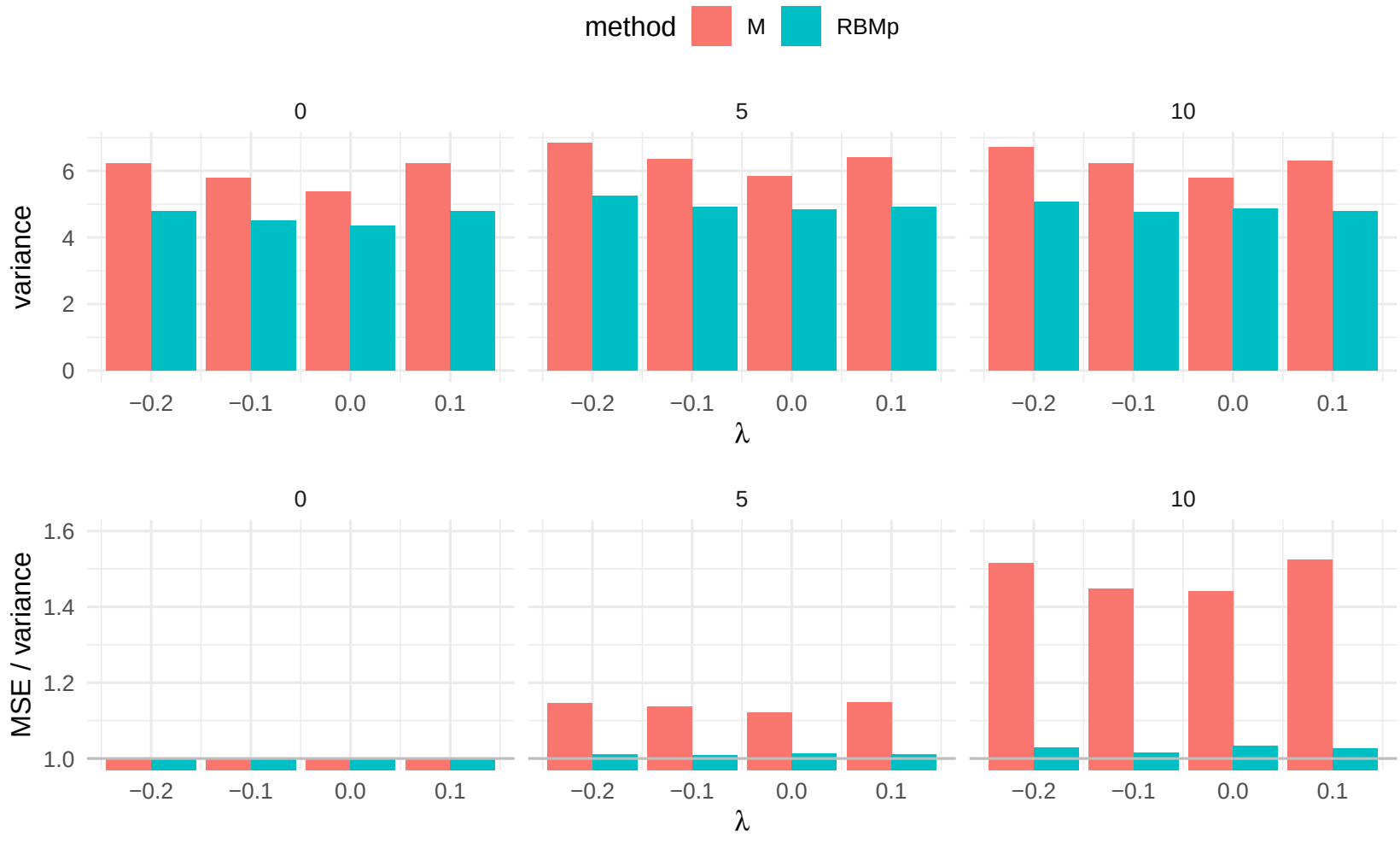
[23] aka Besag's pseudo-likelihood (Besag, 1975)

[24] See, Wolters (2017) for an authoritative overview on autologistic regression

High-dimensional autologistic regression

$k = 100$ fully-connected groups, $c = 10$, $x(s) \sim N_{100}(0, (2kc)^{-1})$

$\beta = (\underbrace{10, \ldots, 10}_{\times 20}, \underbrace{5, \ldots, 5}_{\times 20}, 0, \ldots, 0)^\top$, $\lambda \in \{-0.2, -0.1, 0, 0.1\}$



Outline

```
Call:
betareg(formula = accuracy ~ dyslexia * iq | dyslexia + iq, data = ReadingSkills,
    type = "ML")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.3900 -0.6416  0.1572  0.8524  1.6446

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1232     0.1428   7.864 3.73e-15 ***
dyslexia     -0.7416     0.1428  -5.195 2.04e-07 ***
iq            0.4864     0.1331   3.653 0.000259 ***
dyslexia:iq  -0.5813     0.1327  -4.381 1.18e-05 ***

Phi coefficients (precision model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.3044     0.2227  14.835  < 2e-16 ***
dyslexia      1.7466     0.2623   6.658 2.77e-11 ***
iq            1.2291     0.2672   4.600 4.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  65.9 on 7 Df
Pseudo R-squared: 0.5756
Number of iterations: 25 (BFGS) + 1 (Fisher scoring)
```

Finite sample variance of the ML estimator

For $p = 1$

$$\mathrm{var}_F(\hat{\theta}) = \underbrace{\frac{1}{i(\theta)}}_{O(n^{-1})} + \underbrace{\frac{2b'(\theta)}{i(\theta)} + 2b(\theta)^2 + \frac{\gamma(\theta)^2}{i(\theta)}}_{O(n^{-2})} + O\left(n^{-3}\right) ,$$

where $\gamma(\theta)$ is the statistical curvature (Efron, 1975), which is zero for full exponential families and positive, otherwise

$$\underbrace{\mathrm{var}_F(\hat{\theta})}_{\text{target}} = \overbrace{\frac{1}{i(\theta)}}^{\substack{\text{typically} \\ \text{used at} \\ \theta := \hat{\theta}}} + \frac{2b'(\theta)}{i(\theta)} + 2b(\theta)^2 + \frac{\gamma(\theta)^2}{i(\theta)} + O\left(n^{-3}\right) ,$$

$$\underbrace{\mathrm{var}_F(\hat{\theta})}_{\text{target}} = \overbrace{\frac{1}{i(\theta)} + \frac{2b'(\theta)}{i(\theta)} + 2b(\theta)^2 + \frac{\gamma(\theta)^2}{i(\theta)}}^{\text{use all terms at } \theta := \hat{\theta}?} + O\left(n^{-3}\right) \ ,$$

$$\underbrace{\mathrm{var}_F(\hat{\theta})}_{\text{target}} = \overbrace{\frac{1}{i(\theta)} + \frac{2b'(\theta)}{i(\theta)} + 2b(\theta)^2 + \frac{\gamma(\theta)^2}{i(\theta)}}^{\text{use all terms at } \theta := \hat{\theta}?} + O\left(n^{-3}\right) ,$$

Not necessarily non-negative due to $2b'(\theta)/i(\theta)$ [25]

[25] Examples in the PhD thesis by Claudia Di Caterina (Di Caterina, 2017)

$$\underbrace{\mathrm{var}_F(\hat{\theta})}_{\text{target; bootstrap?}} = \frac{1}{i(\theta)} + \frac{2b'(\theta)}{i(\theta)} + 2b(\theta)^2 + \frac{\gamma(\theta)^2}{i(\theta)} + O\left(n^{-3}\right) ,$$

Bootstrap CDF of $\hat{\theta}$ is not necessarily available in closed form so refitting is required, in general

Choice of bootstrap (parametric, non-parametric, semi-parametric?)

Can result in instabilities, e.g. when estimating the variance of the log-odds, where there is positive probability of infinite estimates

Variance corrected estimators

Let $\ell_r(\theta) = d^r\ell(\theta)/d\theta^r$; $v(\theta) = \mathrm{E}_F(\ell_r(\theta))$; $v_{r,s}(\theta) = \mathrm{E}_\theta(\ell_r(\theta)\ell_s(\theta))$, ...

Adjustment

The adjusted scores estimator has $\mathrm{var}_F(\tilde{\theta}) = 1/i(\theta) + O(n^{-3})$ if $A(\theta)$ is the solution of the first-order linear differential equation

$$\frac{dA}{d\theta} - \frac{v_{1,2}(\theta) + v_3(\theta)}{v_2(\theta)} A = q(\theta) \tag{2}$$

where

$$q = \frac{2v_{1,1,2} + 3v_{2,2} - v_2^2 + 3v_{1,3} + v_4}{2v_2} - \frac{5v_3^2 + 16v_3v_{1,2} + 10v_{1,2}^2}{4v_2^2}$$

Solution of (2) for initial condition $A(t) = a$, $t \in \Theta$ and $a \in \Re$

$$A(\theta) = \nu_2(\theta)\left\{\int_t^\theta \frac{q(s)}{\nu_2(s)} ds + \frac{a}{\nu_2(t)}\right\}$$

Choice of initial conditions

$\theta^\dagger \leftarrow s(\theta) + A(\theta) = 0$ attains the Cramér-Rao lower bound to 2nd order for any $t \in \Theta$ and $a \in \Re$

Choice of t and a can be based on bias considerations

Binomial odds

Random variable Y with pmf

$$p_Y(y|m,\theta) = \binom{m}{y}\pi(\theta)^y(1-\pi(\theta))^{m-y}\,, \quad y \in \{0,1,\ldots,m\}, \theta > 0$$

where $\pi(\theta) = \theta/(1+\theta)$, i.e. θ is the odds of success.

Score function: $s(\theta) = \dfrac{y}{\theta} - \dfrac{m}{\theta+1}$

ML estimator: $\hat{\theta} = y/(m-y)$

Bias reducing score adjustment: $A(\theta) = -\dfrac{1}{1+\theta}$

RB estimator: $\tilde{\theta} = y/(m+1-y)$

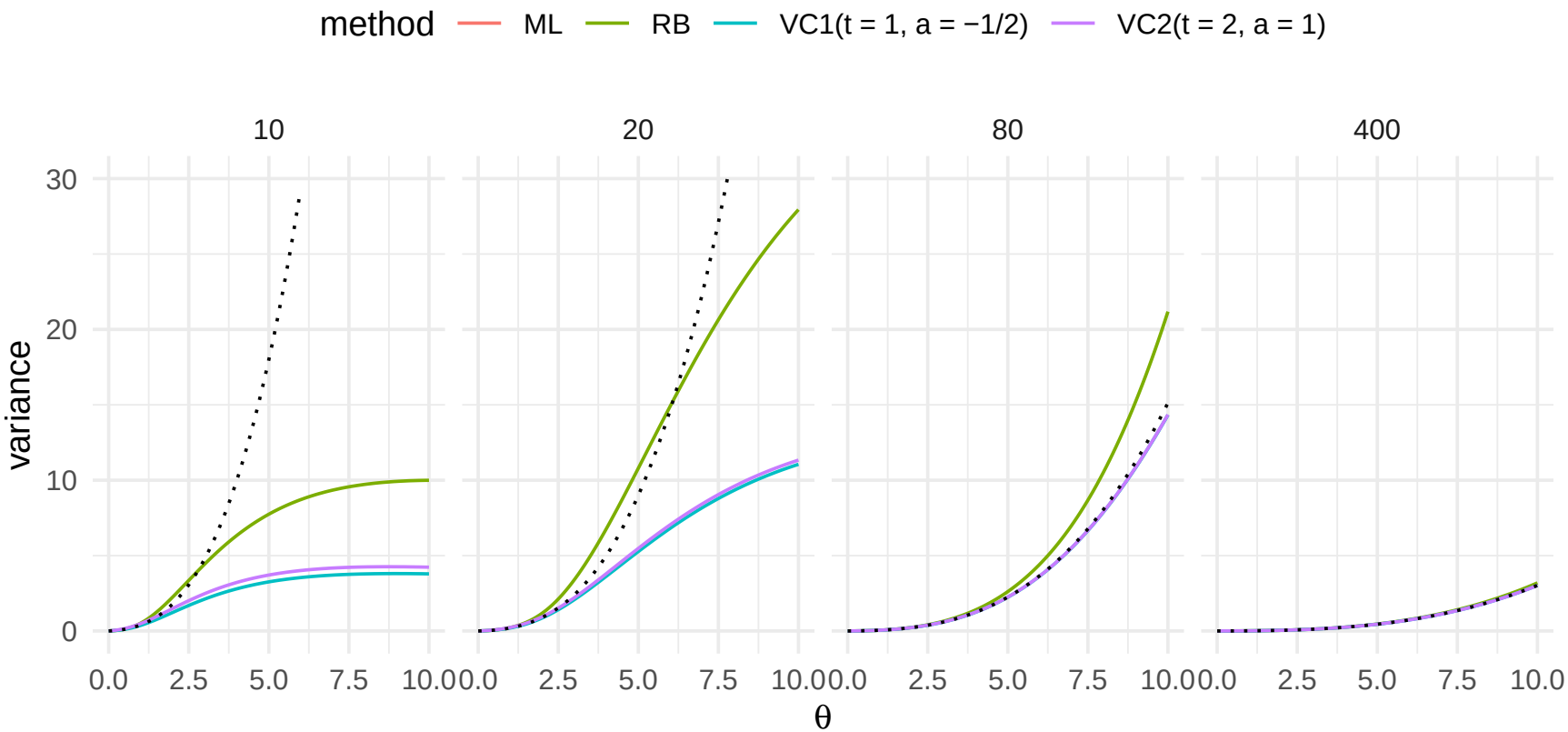
Variance correcting score adjustments

$$A(\theta) = \frac{t\left(2a(t+1)^2 + 3t + 2\right) - \theta(3\theta+2)}{2\theta(\theta+1)^2}$$

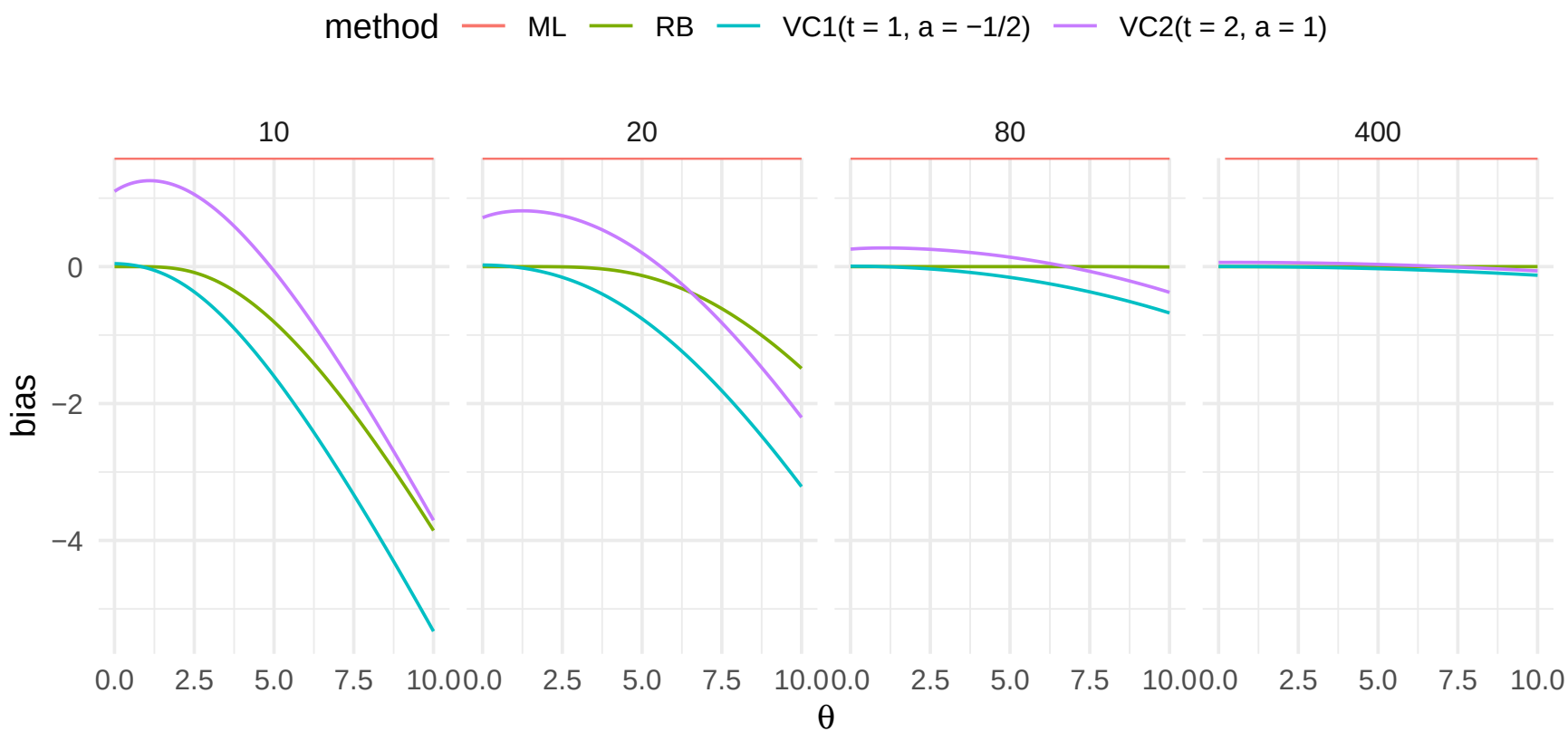
Binomial odds

y	ML	RB	VC1(t = 1, a = -1/2)	VC2(t = 2, a = 1)
0	0.000	0.000	0.043	1.100
1	0.111	0.100	0.143	1.231
2	0.250	0.222	0.263	1.387
3	0.429	0.375	0.412	1.576
4	0.667	0.571	0.600	1.810
5	1.000	0.833	0.846	2.108
6	1.500	1.200	1.182	2.505
7	2.333	1.750	1.667	3.062
8	4.000	2.667	2.429	3.912
9	9.000	4.500	3.800	5.395
10	∞	10.000	7.000	8.745

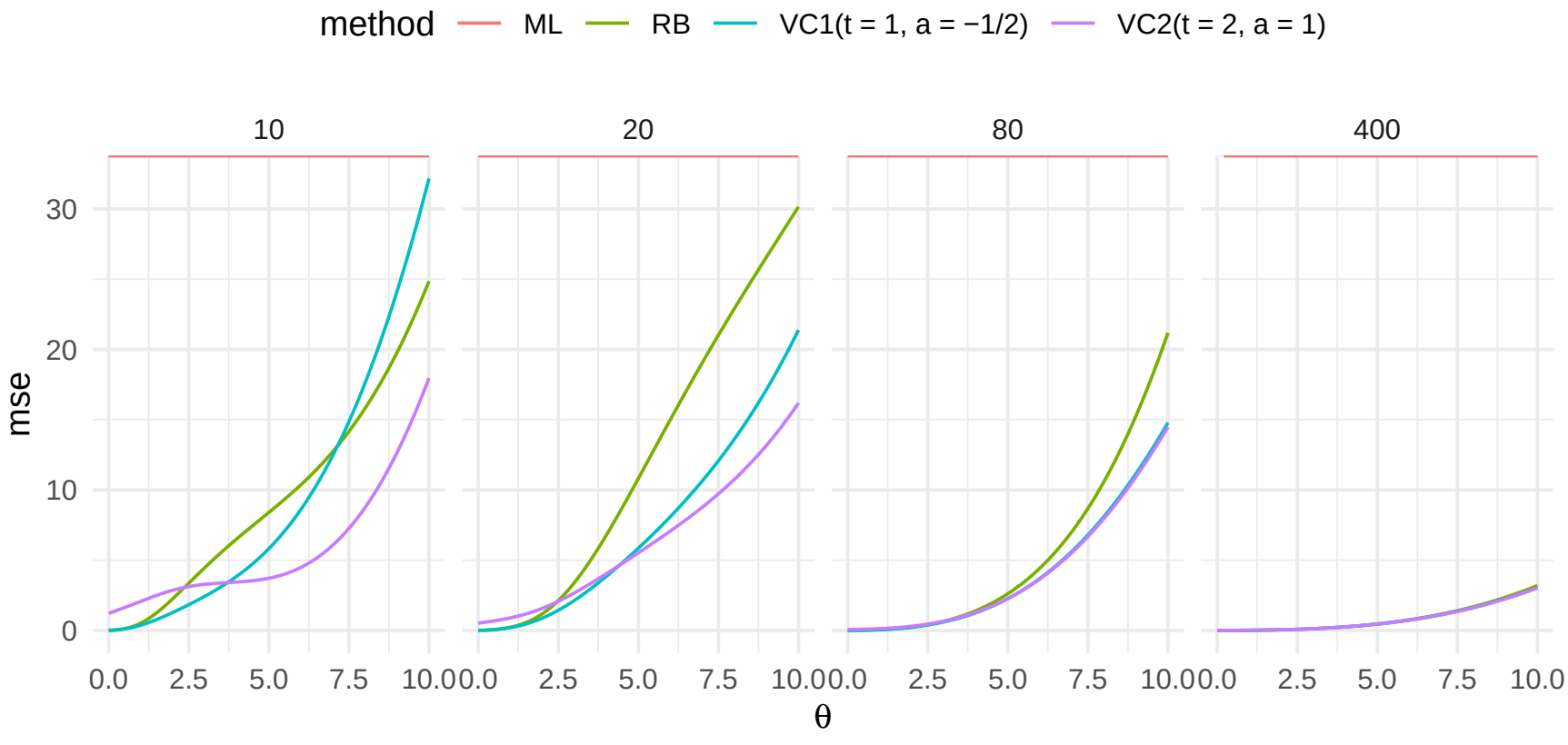
Binomial odds: Variance



Binomial odds: Bias



Binomial odds: Mean squared error



Outline

Discussion I

RBM-estimation

Relies only on derivatives of estimating functions or objectives

Penalized bias-reducing objectives if an objective is available (e.g. composite likelihoods, partial likelihoods, etc)

Variance correction

Second-order efficiency considerations

Estimating functions for stratified settings[25]

Bias reduction of a low-dimensional parameter of interest in the presence of strata-specific nuisance parameters

[25] see Sartori (2003) and Lunardon (2018)

Discussion II

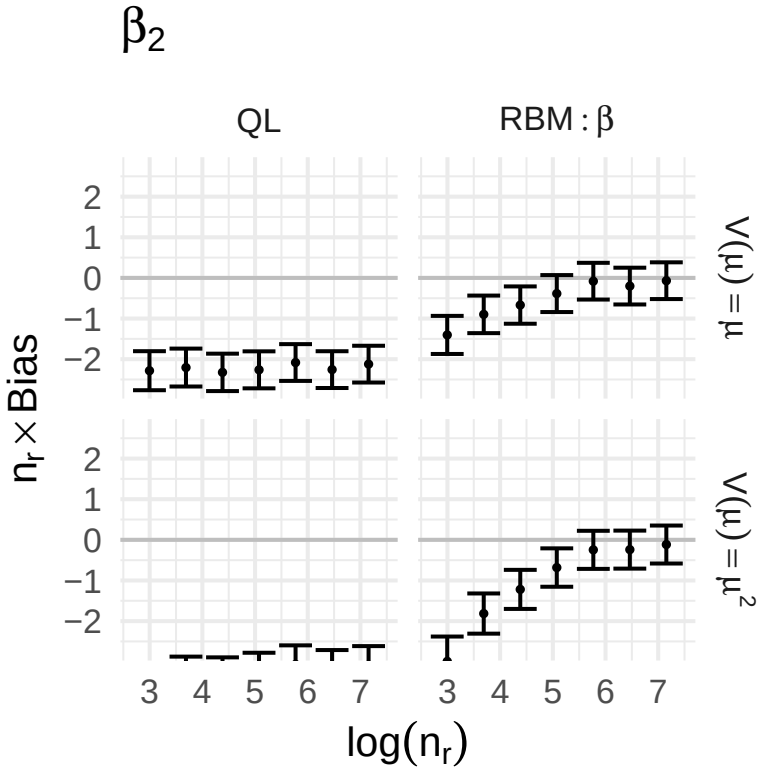
Dependent series

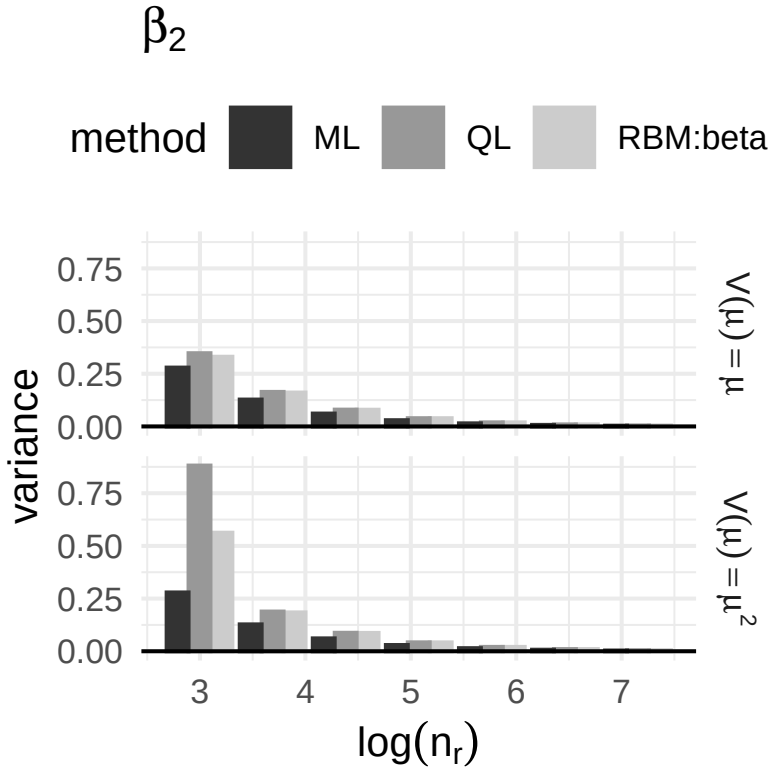
Time-series modelling under stationarity assumptions[26]

$$\Downarrow$$

Adjacent non-overlapping sub-series for the definition of $e(\theta)$ and $d_r(\theta)$

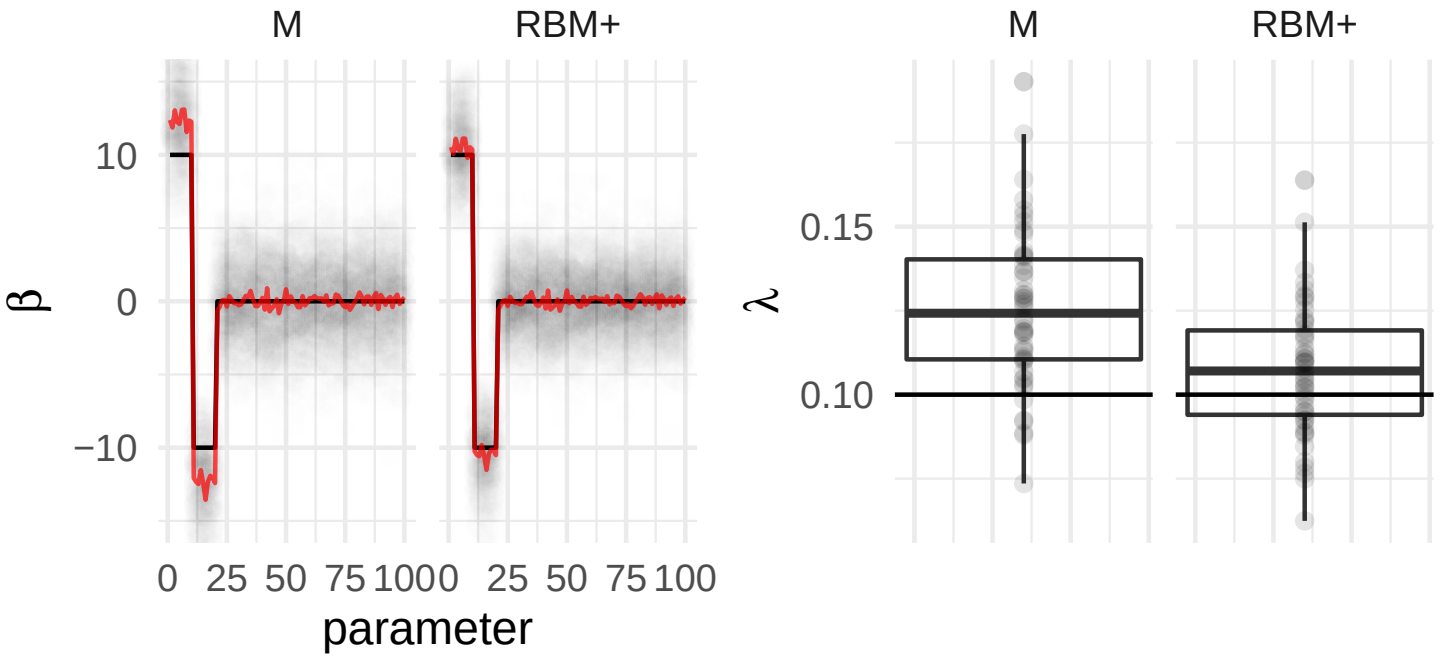
Efficiency gains





[26] see Carlstein (1986)

High-dimensional regression settings



$n = 800$, $p = 100$, $\{x(s_{ij})\}$ are iid $N(0, 1/(2n))$, 100 fully-connected groups

Mean bias reduction in ML estimation

$A_t = \frac{1}{2}\text{trace}\left[i^{-1}\left\{P_t + Q_t\right\}\right]$

Median bias reduction in ML estimation

$A_t = \frac{1}{2}\text{trace}\left[i^{-1}\left\{P_t + Q_t\right\}\right] - iR$

RBM-estimation

$A_t = -\text{trace}\left\{j^{-1}d_t\right\} - \frac{1}{2}\text{trace}\left[j^{-1}e\left\{j^{-1}\right\}^{\top}u_t\right]$

Bias-reducing penalized objectives $\ell - \frac{1}{2}\text{trace}\left\{j^{-1}e\right\}$

Variance correction

$A' - A(v_{1,2} + v_3)/v_2 = q$

Software

brglm2	R	github.com/ikosmidis/brglm2
betareg	R	r-forge.r-project.org/projects/betareg
MEstimation	julia	github.com/ikosmidis/MEstimation.jl
brquasi	R	github.com/ikosmidis/brquasi
...		

IKosmidis_ ioannis.kosmidis@warwick.ac.uk www.ikosmidis.com

References I

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician) 24*(3), 179–195.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics 14*(3), 1171–1179.

Di Caterina, C. (2017). *Reducing the Impact of Bias in Likelihood Inference for Prominent Model Settings*. Ph. D. thesis, University of Padova.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika 46*(3), 477–480.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics 3*(6), 1189–1242.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Griewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation* (2nd ed.). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Grün, B., I. Kosmidis, and A. Zeileis (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software 48*(11), 1–25.

Kenne Pagui, E. C., A. Salvan, and N. Sartori (2017). Median bias reduction of maximum likelihood estimates. *Biometrika 104*(4), 923–938.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804.

Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics 4*, 1097–1112.

Kosmidis, I. and D. Firth (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika 108*(1), 71–82.

References II

Kosmidis, I., E. C. Kenne Pagui, and N. Sartori (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing (to appear) 30*, 43–59.

Kosmidis, I. and N. Lunardon (2021). Empirical bias-reducing adjustments to estimating functions.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 395–407.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lunardon, N. (2018). On bias reduction and incidental parameters. *Biometrika 105*(1), 233–238.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics 11*(1), 59–67.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*. London, UK: World Scientific.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika 90*(3), 533–549.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika 92*(3), 519–528.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika 54*(3), 427–450.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika 61*(3), 439–447.

Wolters, M. A. (2017). Better autologistic regression. *Frontiers in Applied Mathematics and Statistics 3*, 24.