Jeffreys-prior penalty in binomial-response generalized linear models

Ioannis Kosmidis
(𝕏) IKosmidis_
ikosmidis.com
ioannis.kosmidis@warwick.ac.uk
Professor of Statistics
University of Warwick

6 August 2023
JSM 2023

Collaborators: David Firth, Patrick Zietkiewicz
Slides at: ikosmidis.com/files/kosmidis_JSM2023.pdf

Outline

Outline

Logistic regression

Data

Responses $y_1, \ldots, y_n$ with $y_i \in \{0, 1\}$
Covariate vectors $x_1, \ldots, x_n$ with $x_i \in \Re^p$

Model

$Y_1, \ldots, Y_n$ conditionally independent with $Y_i | x_i \sim \text{Binomial}(m_i, \pi_i)$, $\log \dfrac{\pi_i}{1 - \pi_i} = \eta_i = \displaystyle\sum_{t=1}^{p} \beta_t x_{it}$

One of the most frequently applied generalized linear models for data-analytic tasks

Uses range from inference about covariate effects on binomial probabilities to probability calibration and prediction

Maximum likelihood estimation

Log-likelihood: $l(\beta) = \sum_{i=1}^{n} y_i \eta_i - \sum_{i=1}^{n} m_i \log\left(1 + e^{\eta_i}\right)$

Maximum likelihood (ML) estimator: $\hat{\beta} = \arg\max l(\beta)$

Iterative re-weighted least squares

$\hat{\beta} := \beta^{(\infty)}$, where

$$\beta^{(j+1)} = \left(X^T W^{(j)} X\right)^{-1} X^T W^{(j)} z^{(j)}$$

W is a diagonal matrix with ith diagonal element $m_i \pi_i (1 - \pi_i)$

$z_i = (y_i - m_i \pi_i)/\{m_i \pi_i (1 - \pi_i)\}$ is the working variate

Bradley-Terry model for team abilities

Data

Game outcomes from the first 262 games of the regular season of the 2014-2015 NBA conference[1]

Aim

Infer abilities of the basketball teams

[1] Data from `www.basketball-reference.com`

Bradley-Terry model

$y_{ij} = 1$ if team i beats team j, and $y_{ij} = 0$, otherwise

Outcomes are assumed independent

$Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$ with

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_i - \beta_j$$

Identifiability constraint

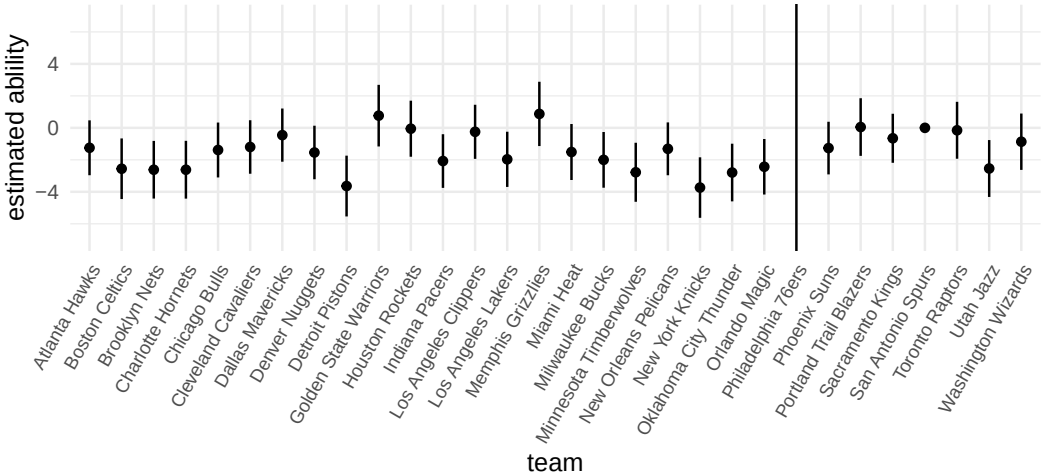
Set ability for San Antonio Spurs (champion of the 2013–2014 conference) to zero. Then, β_i represents the contrast of the ability of team i with that of the San Antonio Spurs

```
R> fit_ml


Call:  glm(formula = BT_formula, family = binomial(logit), data = nba_phil_BT)

Coefficients:
        Atlanta.Hawks         Boston.Celtics          Brooklyn.Nets       Charlotte.Hornets
              -1.24700               -2.55752               -2.61972               -2.61907
        Chicago.Bulls     Cleveland.Cavaliers       Dallas.Mavericks          Denver.Nuggets
              -1.38439               -1.19661               -0.45449               -1.54111
      Detroit.Pistons   Golden.State.Warriors        Houston.Rockets          Indiana.Pacers
              -3.63998                0.76292               -0.05369               -2.07574
  Los.Angeles.Clippers      Los.Angeles.Lakers       Memphis.Grizzlies              Miami.Heat
              -0.24765               -1.96871                0.87174               -1.51083
      Milwaukee.Bucks  Minnesota.Timberwolves    New.Orleans.Pelicans        New.York.Knicks
              -2.00144               -2.77836               -1.31219               -3.73422
 Oklahoma.City.Thunder           Orlando.Magic      Philadelphia.76ers             Phoenix.Suns
              -2.79467               -2.43380              -19.23826               -1.26507
Portland.Trail.Blazers       Sacramento.Kings        Toronto.Raptors              Utah.Jazz
               0.05028               -0.65240               -0.15207               -2.54268
    Washington.Wizards       San.Antonio.Spurs
              -0.86709                     NA

Degrees of Freedom: 262 Total (i.e. Null);  233 Residual
Null Deviance:     363.2
Residual Deviance: 238.9  AIC: 296.9
```



Philadelphia 76ers had 17 losses and no win; San Antonio Spurs had 13 wins in 17 games.

Wald test for difference in ability between the Philadelphia 76ers and the San Antonio Spurs results in no apparent evidence of a difference

```
R> update(fit_ml, method = "detect_infinite_estimates")

Implementation: ROI | Solver: lpsolve
Infinite estimates: TRUE
Existence of maximum likelihood estimates
         Atlanta.Hawks          Boston.Celtics          Brooklyn.Nets      Charlotte.Hornets
                     0                       0                      0                      0
         Chicago.Bulls     Cleveland.Cavaliers       Dallas.Mavericks       Denver.Nuggets
                     0                       0                      0                      0
       Detroit.Pistons  Golden.State.Warriors        Houston.Rockets        Indiana.Pacers
                     0                       0                      0                      0
  Los.Angeles.Clippers     Los.Angeles.Lakers      Memphis.Grizzlies           Miami.Heat
                     0                       0                      0                      0
       Milwaukee.Bucks Minnesota.Timberwolves   New.Orleans.Pelicans       New.York.Knicks
                     0                       0                      0                      0
 Oklahoma.City.Thunder          Orlando.Magic     Philadelphia.76ers          Phoenix.Suns
                     0                       0                   -Inf                      0
Portland.Trail.Blazers       Sacramento.Kings        Toronto.Raptors              Utah.Jazz
                     0                       0                      0                      0
    Washington.Wizards      San.Antonio.Spurs
                     0                      NA
0: finite value, Inf: infinity, -Inf: -infinity
```

[2]Using the detectseparation R package (Kosmidis et al., 2022)

Jeffreys' prior penalty

Penalized Log-likelihood: $\tilde{l}(\beta) = \sum_{i=1}^{n} y_i \eta_i - \sum_{i=1}^{n} m_i \log\left(1 + e^{\eta_i}\right) + \underbrace{\frac{1}{2} \log |X^\top W(\beta) X|}_{\text{log Jeffreys prior}}$

Maximum Jeffreys-prior penalized likelihood (mJPL) estimator: $\tilde{\beta} = \arg\max \tilde{l}(\beta)$

Key properties

1. If X is of full rank, $\tilde{\beta}$ has all of its components finite [3]
2. $\tilde{\beta}$ shrinks towards equi-probability across observations, relative to $\hat{\beta}$ [3]
3. Easy estimation (either through IWLS or repeated ML fits) [3] [4]
4. For usual fixed p asymptotics, $\tilde{\beta}$ is consistent and asymptotically normal
5. 2nd order bias: $E(\tilde{\beta} - \beta) = o(n^{-1})$ [5]
6. 1st order efficiency: $\mathrm{var}(\tilde{\beta}) = X^\top W(\beta) X + o(n^{-1})$ [5]

[3] see Kosmidis and Firth (2021)
[4] see Zietkiewicz and Kosmidis (2023)
[5] see Firth (1993)

```
R> update(fit_ml, method = "brglm_fit")


Call:  glm(formula = BT_formula, family = binomial(logit), data = nba_phil_BT,
    method = "brglm_fit")

Coefficients:
         Atlanta.Hawks          Boston.Celtics           Brooklyn.Nets       Charlotte.Hornets
               -1.07454               -2.19946               -2.26321               -2.26729
         Chicago.Bulls      Cleveland.Cavaliers        Dallas.Mavericks          Denver.Nuggets
               -1.19317               -1.02836               -0.38570               -1.33494
       Detroit.Pistons   Golden.State.Warriors         Houston.Rockets          Indiana.Pacers
               -3.14565                0.62436               -0.05221               -1.80126
  Los.Angeles.Clippers      Los.Angeles.Lakers       Memphis.Grizzlies             Miami.Heat
               -0.21008               -1.70415                0.72333               -1.30198
       Milwaukee.Bucks  Minnesota.Timberwolves    New.Orleans.Pelicans         New.York.Knicks
               -1.72878               -2.39847               -1.13360               -3.21771
 Oklahoma.City.Thunder          Orlando.Magic      Philadelphia.76ers            Phoenix.Suns
               -2.42025               -2.11217               -5.03883               -1.09020
Portland.Trail.Blazers        Sacramento.Kings         Toronto.Raptors              Utah.Jazz
                0.04853               -0.55887               -0.12174               -2.20417
    Washington.Wizards       San.Antonio.Spurs
               -0.73679                     NA

Degrees of Freedom: 262 Total (i.e. Null);  233 Residual
Null Deviance:     363.2
Residual Deviance: 241.1  AIC: 299.1
```

[6] Using the brglm2 R package (Kosmidis, 2023)

estimated ablility
4
0
−4
mJPL
ML
team
Atlanta Hawks
Boston Celtics
Brooklyn Nets
Charlotte Hornets
Chicago Bulls
Cleveland Cavaliers
Dallas Mavericks
Denver Nuggets
Detroit Pistons
Golden State Warriors
Houston Rockets
Indiana Pacers
Los Angeles Clippers
Los Angeles Lakers
Memphis Grizzlies
Miami Heat
Milwaukee Bucks
Minnesota Timberwolves
New Orleans Pelicans
New York Knicks
Oklahoma City Thunder
Orlando Magic
Philadelphia 76ers
Phoenix Suns
Portland Trail Blazers
Sacramento Kings
San Antonio Spurs
Toronto Raptors
Utah Jazz
Washington Wizards

Outline

$p/n \to \kappa \in (0, 1)$

Candès and Sur (2020):

sharp phase transition about when the ML estimate has infinite components, when $\eta_i = \alpha + x_i^\top \beta$, $x_i \sim N(0, \Sigma)$, $p/n \to \kappa \in (0, 1)$, $\mathrm{var}(x_i^\top \beta) \to \gamma_0^2$

Sur and Candès (2019):

a method based on approximate message passing that recovers estimation and inferential performance by appropriately rescaling $\hat{\beta}$, whenever that exists

Sur and Candès (2019) and Kosmidis and Firth (2021):

empirical evidence that mJPL achieves a substantial reduction in persistent bias when $p/n \to \kappa$, performing similarly to the rescaled ML estimator, whenever that exists

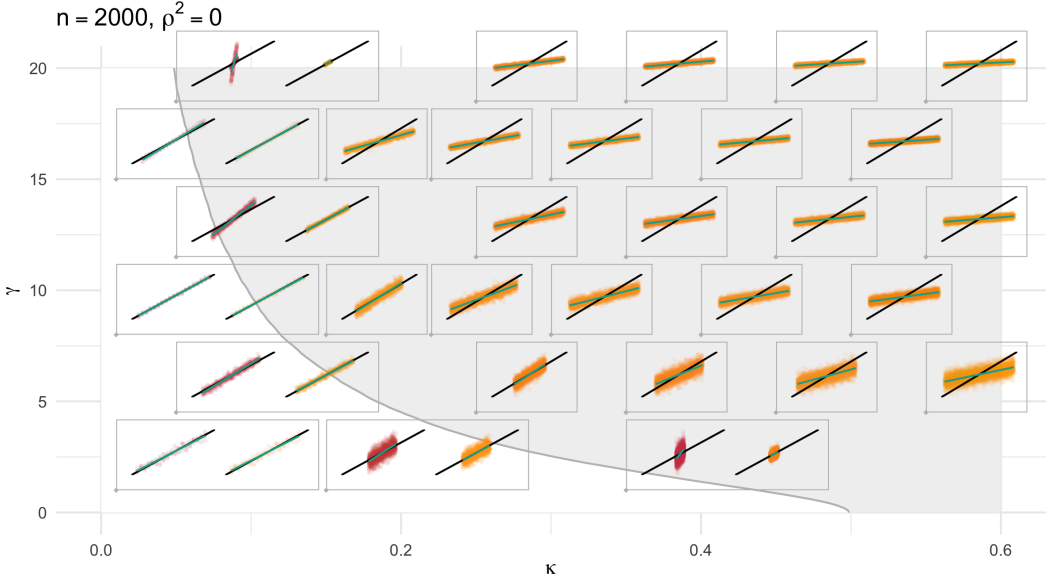
Experiment under the Candès and Sur (2020) asymptotic framework[7]

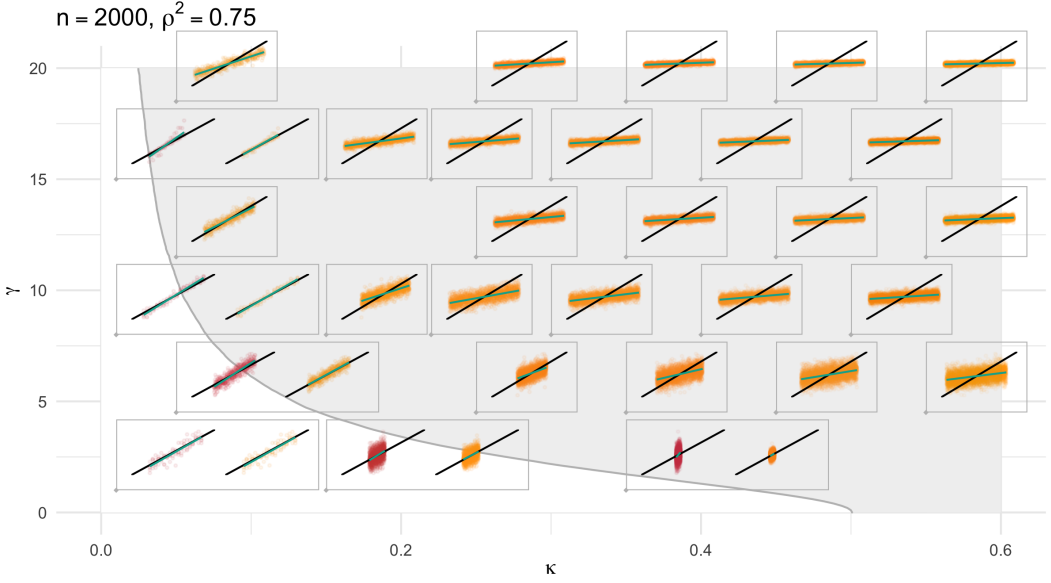
Fix $\kappa \in (0, 1)$, n, $\rho^2 \in [0, 1)$, and a p-vector $\beta*$, with $p = \lceil n\kappa \rceil$

Data generating process

1. Form an $n \times p$ matrix X with $N(0, 1)$ entries, sampled independently
2. Set $\alpha = \rho\gamma$ and $\gamma_0 = \gamma\sqrt{1 - \rho^2}$
3. Set $\beta = \gamma_0\beta^*/\|\beta^*\|_2$ (i.e. $\mathrm{var}(x_i^\top\beta) = \gamma_0^2$)
4. Generate $y_1, \ldots, y_n$ independently with $Y_i \sim$ Bernoulli$(1/(1 + e^{-\eta_i}))$, where $\eta_i = \alpha + x_i^\top\beta$

[7] see Zietkiewicz and Kosmidis (2023) for more details


n = 2000, ρ² = 0
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6


n = 2000, ρ² = 0.75
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

Conjecture: Adjusted mJPL estimator

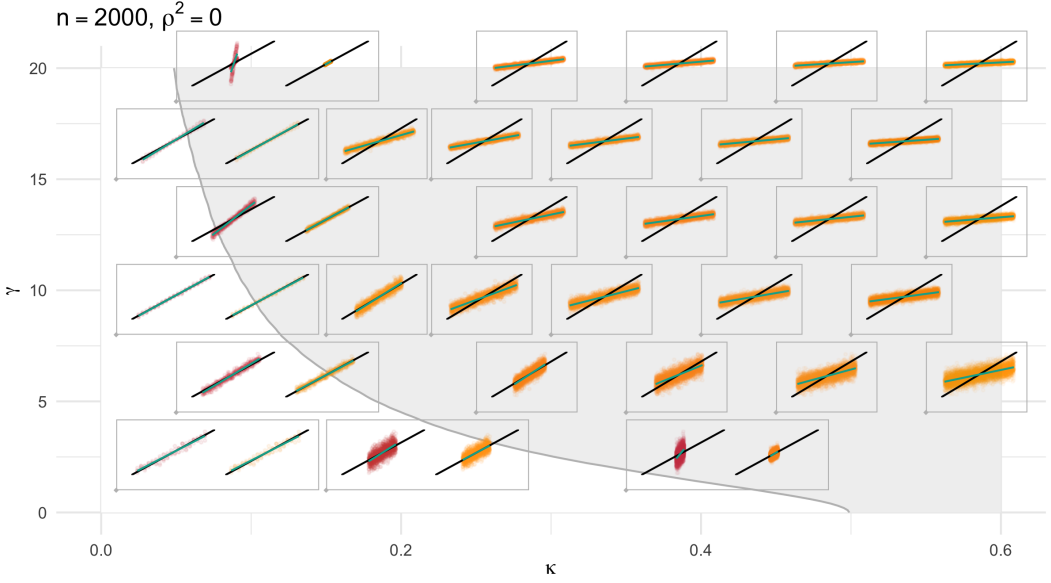
Exploratory linear regression analysis results in strong evidence for the following

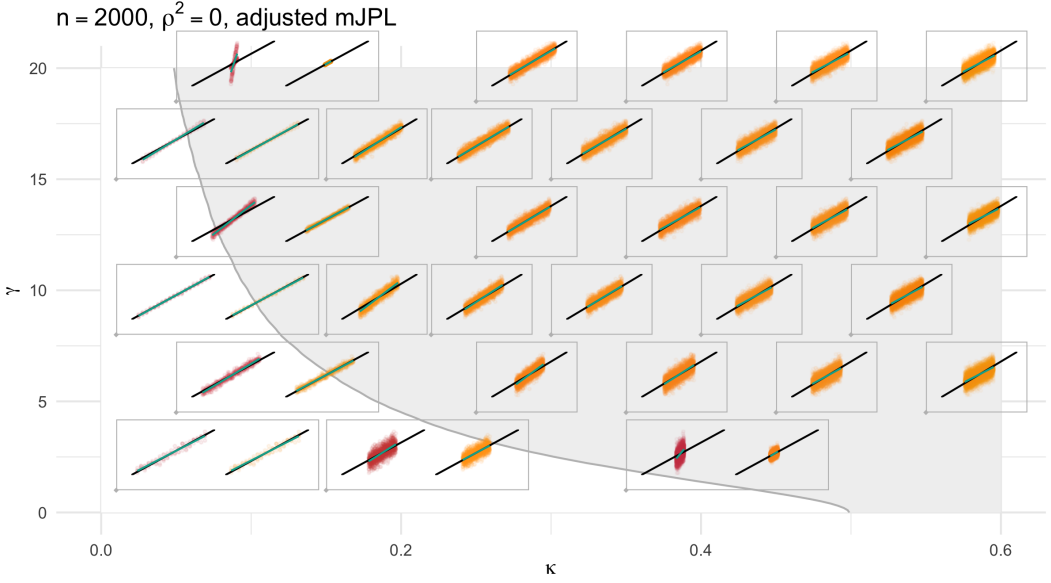
Conjecture

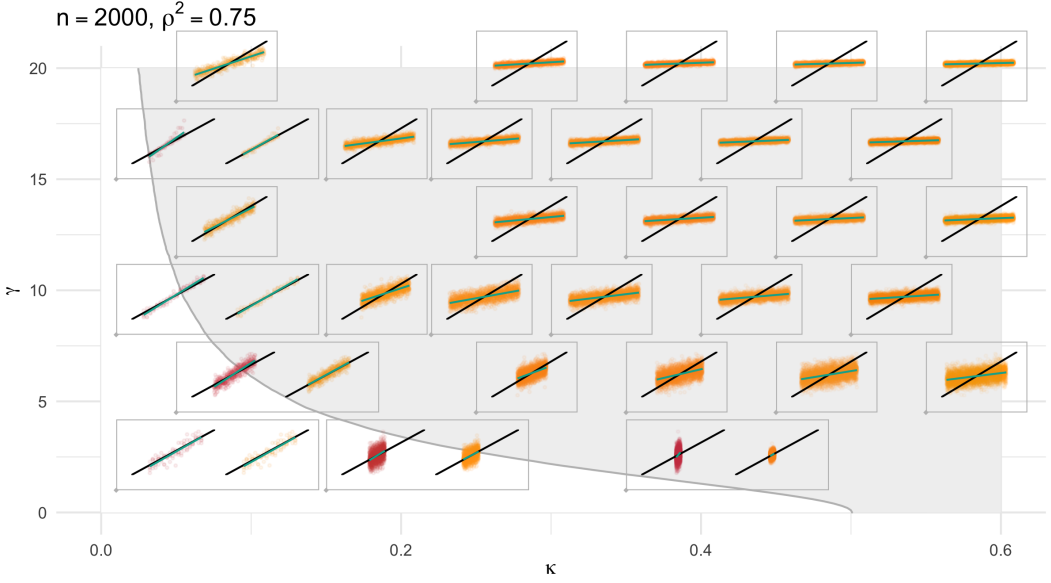
The estimator

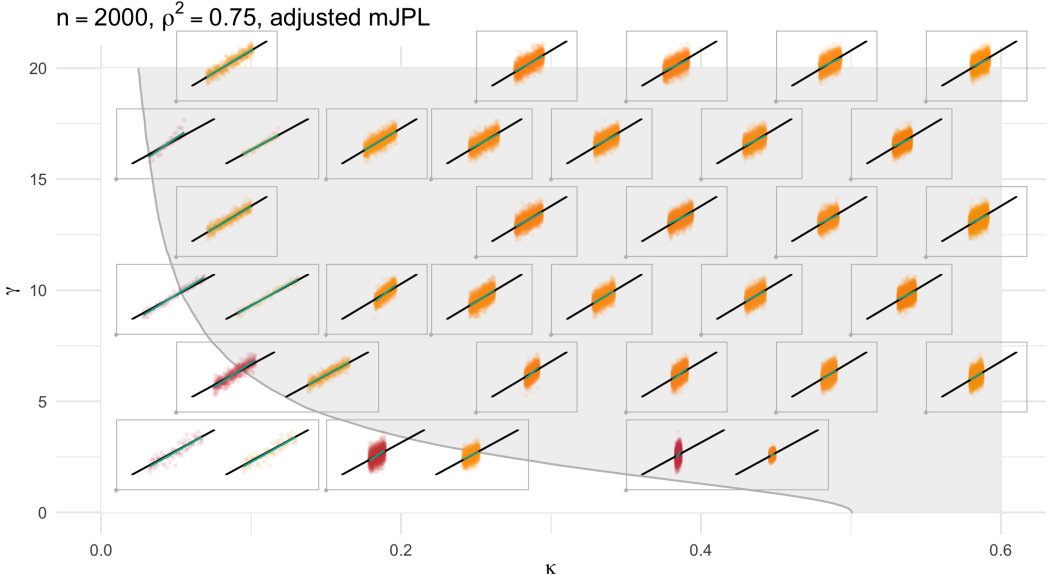
$$\beta^\dagger = \begin{cases} \tilde{\beta}\,, & \text{if } \hat{\beta} \text{ exists asymptotically} \\ \dfrac{\kappa\gamma^2}{\gamma_0}\tilde{\beta}\,, & \text{if } \hat{\beta} \text{ does not exist asymptotically} \end{cases}$$

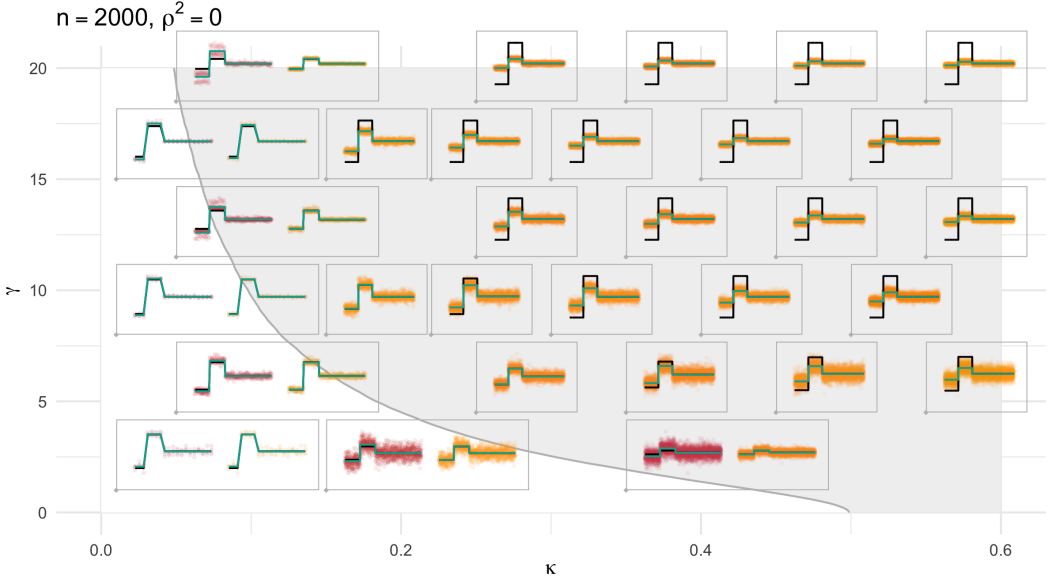
is effective in recovering the true signal in the Candès and Sur (2020) asymptotic framework ($\eta_i = \alpha + x_i^\top \beta$, $x_i \sim N(0, \Sigma)$, $p/n \to \kappa \in (0, 1)$, $\mathrm{var}(x_i^\top \beta) \to \gamma_0^2$) for a wide range of γ_0

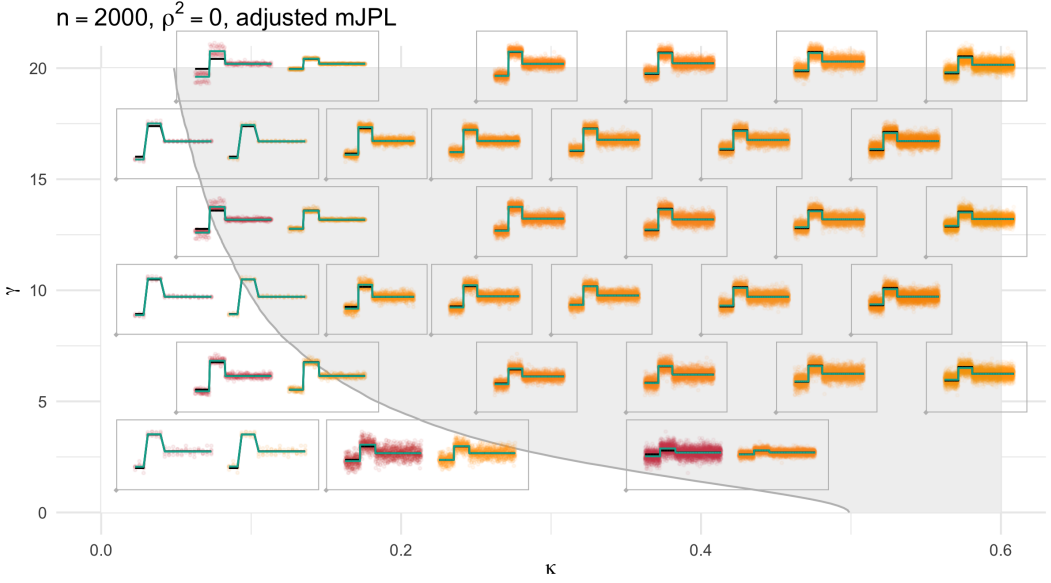

n = 2000, ρ² = 0
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

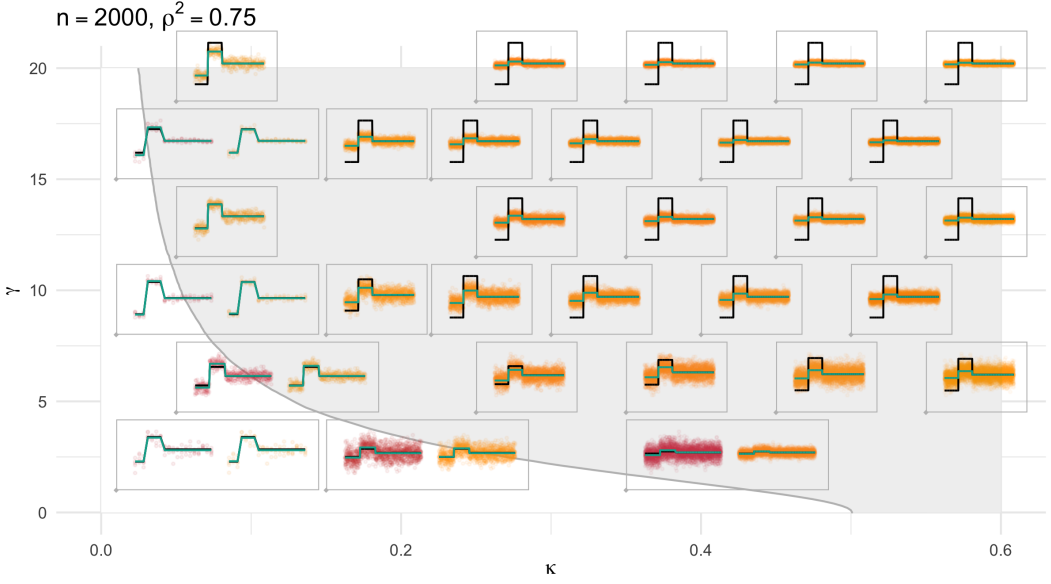

n = 2000, ρ² = 0, adjusted mJPL
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

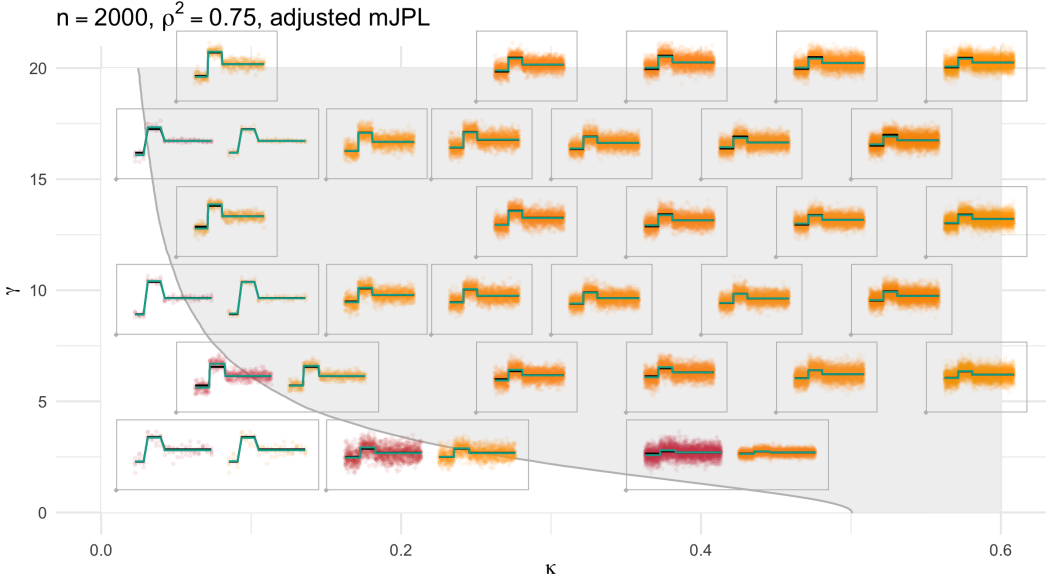

n = 2000, ρ² = 0.75
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

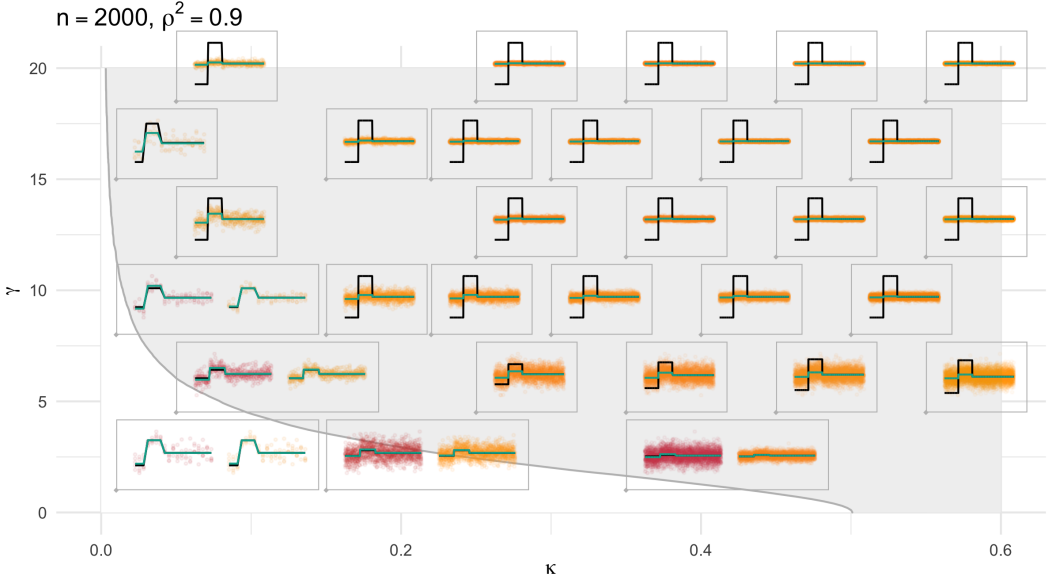

n = 2000, ρ² = 0.75, adjusted mJPL
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

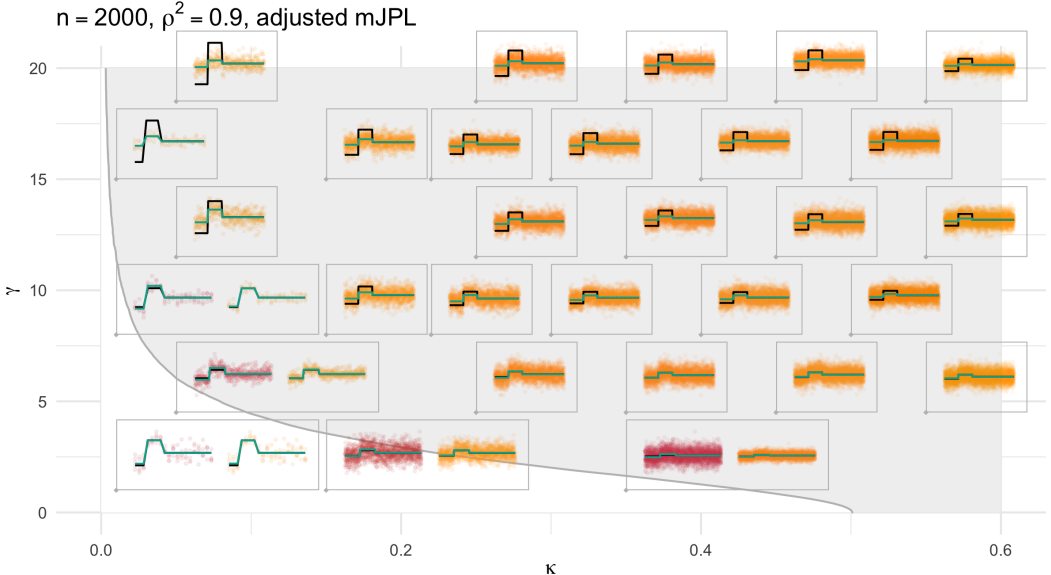

n = 2000, ρ² = 0
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6


n = 2000, ρ² = 0, adjusted mJPL
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6


n = 2000, ρ² = 0.75
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6


n = 2000, ρ² = 0.75, adjusted mJPL
γ
κ
20
15
10
5
0
0.0
0.2
0.4
0.6


n = 2000, ρ² = 0.9
γ
κ
0
5
10
15
20
0.0
0.2
0.4
0.6

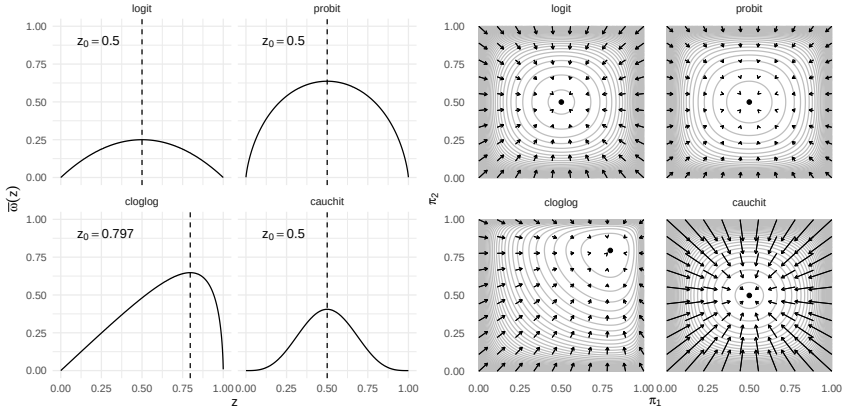


Effectiveness of $\beta^\dagger$ deteriorates as ρ^2 approaches 1 (i.e. α increases relative to $\|\beta\|_2$)

Outline

Discussion: Other links

The finiteness and shrinkage properties of mJPL extend also to many other well-used links, including probit, log-log, complementary log-log and cauchit



Discussion: Powers of Jeffreys-prior penalty

The finiteness and shrinkage properties of mJPL extend also apply for arbitrary powers of the Jeffreys-prior penalty

$$\tilde{l}(\beta) = l(\beta) + a \log |X^\top W(\beta) X| \quad (a > 0)$$

Discussion: Testing for extreme effects

- Finiteness: Estimates are finite $\longrightarrow$ Estimated variances are also finite
- Discreteness: For any given X the mJPL estimator takes only a finite number of values

There always exists a β with extreme-enough components that Wald-type confidence regions fail to cover regardless of the nominal level used

$Y_i \sim \text{Binomial}(m_i, \pi_i)$, $g(\pi_i) = x_i^\top \beta$

$\tilde{\beta} = \arg\max \left\{ l(\beta) + a \log |X^\top W(\beta) X| \right\} \quad (a > 0)$

Key works

Kosmidis I, Firth D (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, **108**, 71–82
DOI: 10.1093/biomet/asaa052

Zietkiewicz P, Kosmidis I (2023). Bounded-memory adjusted scores estimation in generalized linear models with large data sets.
ArXiV: `https://arxiv.org/abs/2307.07342`

Software

R packages: brglm2, detectseparation, a port of biglm [8]

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

[8] see `github.com/ikosmidis/bigbr-supplementary-material`

References

Candès, E. J. and P. Sur (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Annals of Statistics 48*(1), 27–42.

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

Kosmidis, I. (2023). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.9.

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

Kosmidis, I., D. Schumacher, and F. Schwendinger (2022). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.3.

Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*(29), 14516–14525.

Zietkiewicz, P. and I. Kosmidis (2023). Bounded-memory adjusted scores estimation in generalized linear models with large data sets. *arXiv preprint arXiv:2307.07342*.