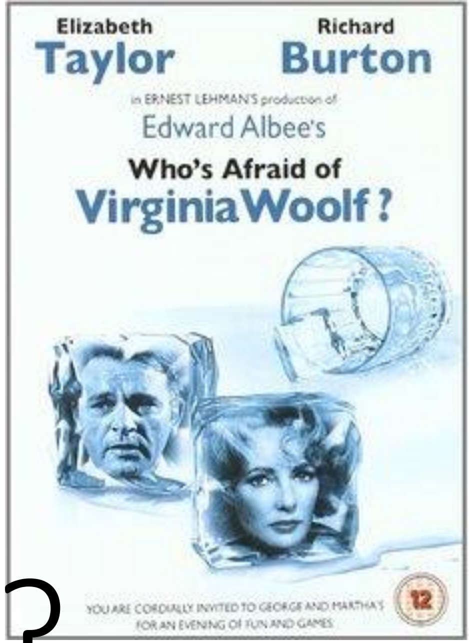
# Who‘s afraid of …
# Bayesian non-collapsibility?



timeout.com

Georg Heinze[1], Angelika Geroldinger[2], Sander Greenland[3]

[1]Medical University of Vienna, Austria

[2]INSERM, UMRS 1138, CRC, Paris, France

[3]University of California, Los Angeles, USA

*Supported by the Austrian Science Fund FWF, Project I2276-N33*

# An example

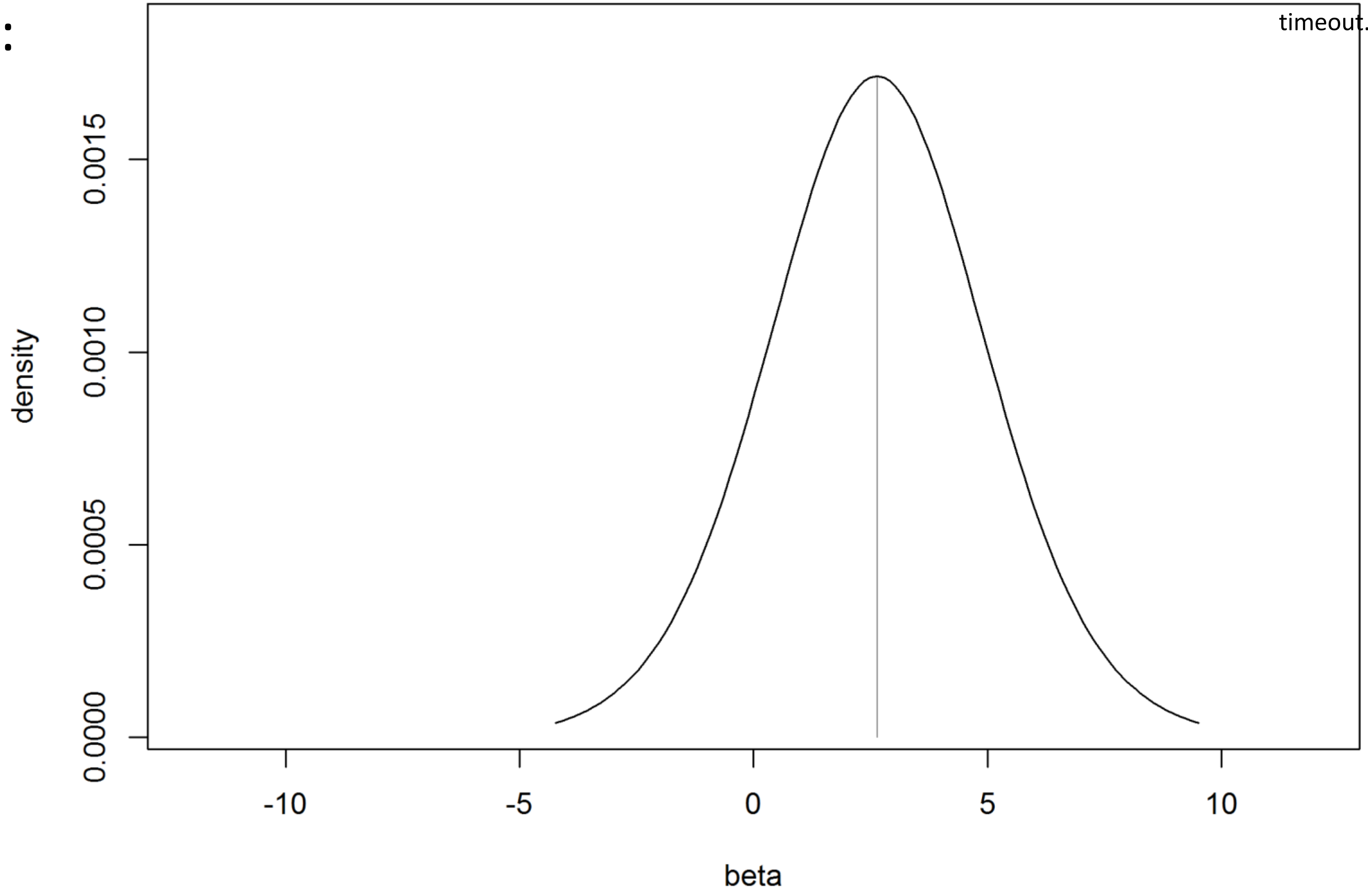

timeout.com

- Simple 2 x 2 table:

| | X=0 | X=1 | |
|---|---|---|---|
| Y=0 | 7 | 1 | *8* |
| Y=1 | 2 | 4 | *6* |
| | *9* | *5* | *14* |

- Suppose we are interested in the log odds ratio relating X to Y:
- $\beta_1 = \log \frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \log \frac{4/1}{2/7} = 2.6$


- The likelihood for the example:
- $L(\beta|x,y) =$
$$\pi(X=1)^{n_{11}}$$
$$\left(1-\pi(X=1)\right)^{n_{10}}$$
$$\pi(X=0)^{n_{01}}$$
$$(1-\pi(X=0))^{n_{00}}$$
- with
$$\pi(X=x) =$$
$$1/[1+\exp(-\beta_0-\beta_1 x)]$$



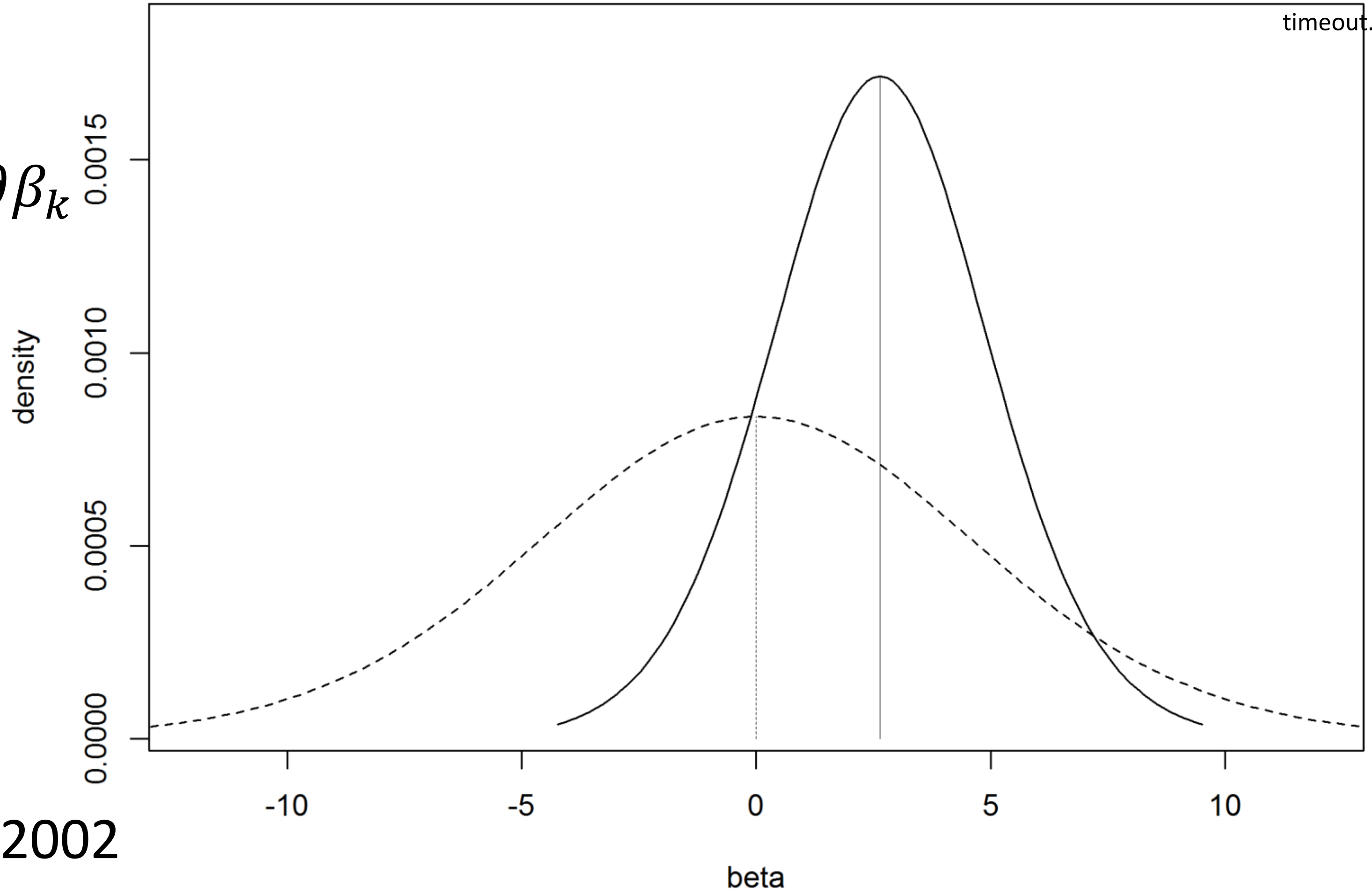
timeout.com

Jeffreys prior: $p(\beta) = |I(\beta)|^{1/2}$

$$I(\beta)_{jk} = -\partial^2 \log L(\beta) / \partial\beta_j \partial\beta_k$$

Weakly informative prior
Automatic solution
Nice properties

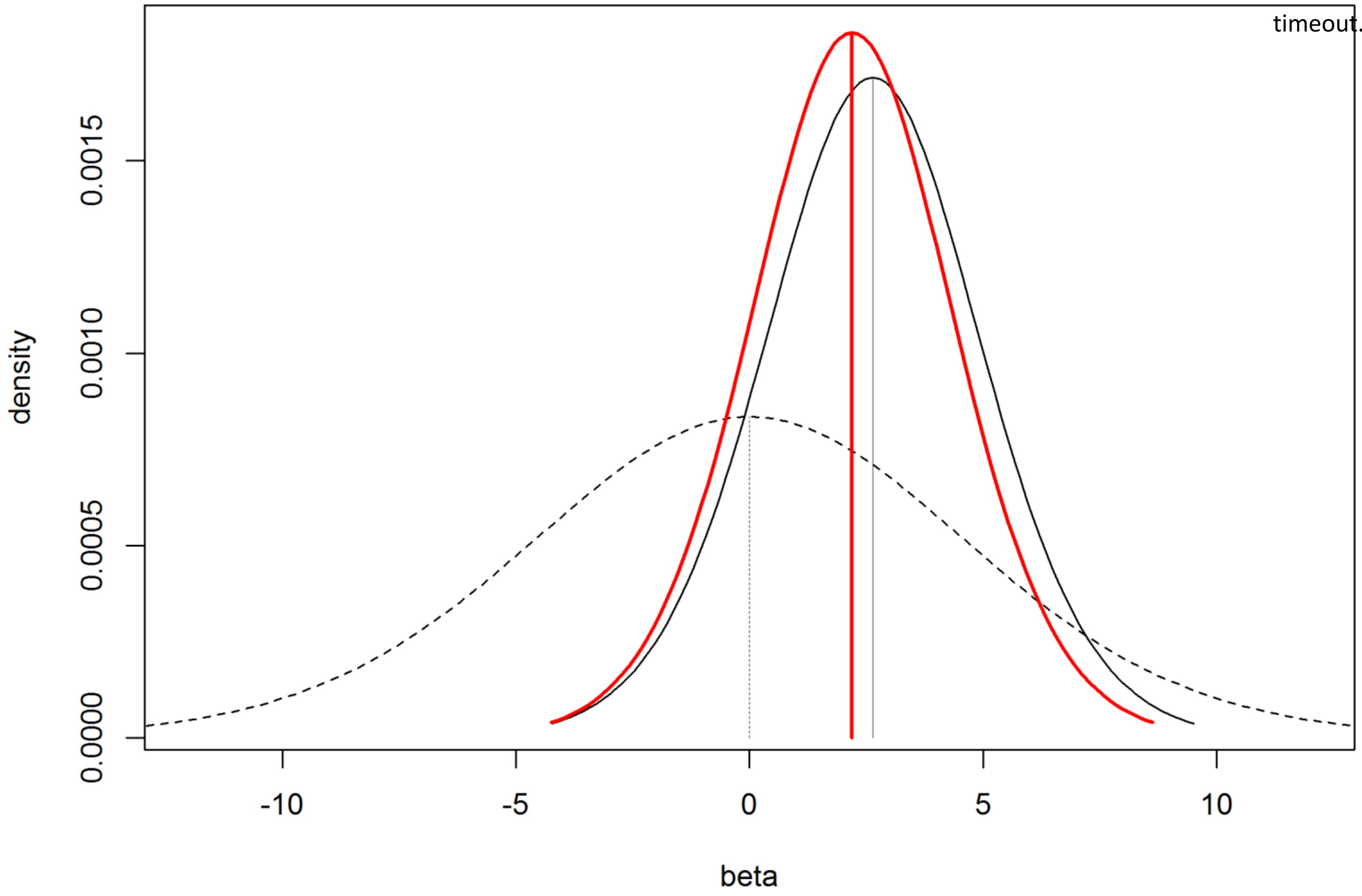
Chen et al JASA 2008,
Firth Biometrika 1993,
Heinze and Schemper StatMed 2002



- The posterior:

$$p(\beta|x,y) = p(\beta)L(\beta|x,y)$$

- As expected, the posterior is between the prior and likelihood





timeout.com

# Using priors in practice


timeout.com

| | |
|---|---|
| • General prior | Derive posterior by simulation (MCMC) |
| • Ridge regression<br>Firth's method/Jeffreys | Prior can be expressed as likelihood penalty |
| • Conjugate prior | Such that posterior has same algebraic form as prior,<br>can be expressed as pseudo-observations („prior data" or *data augmentation prior)* or as a penalty |

- In special cases Jeffreys prior reduces to data augmentation

Firth Biometrika 1993; Jeffreys JASA 1946

# An example


- Augmented 2 x 2 table:

| | X=0 | X=1 | |
|---|---|---|---|
| Y=0 | 7.5 | 1.5 | *9* |
| Y=1 | 2.5 | 4.5 | *7* |
| | *10* | *6* | *16* |

- Maximization of the likelihood of augmented table is now equivalent to finding the posterior mode with original data and Jeffreys prior

# Example of Greenland 2010


timeout.com

- 2x2 table

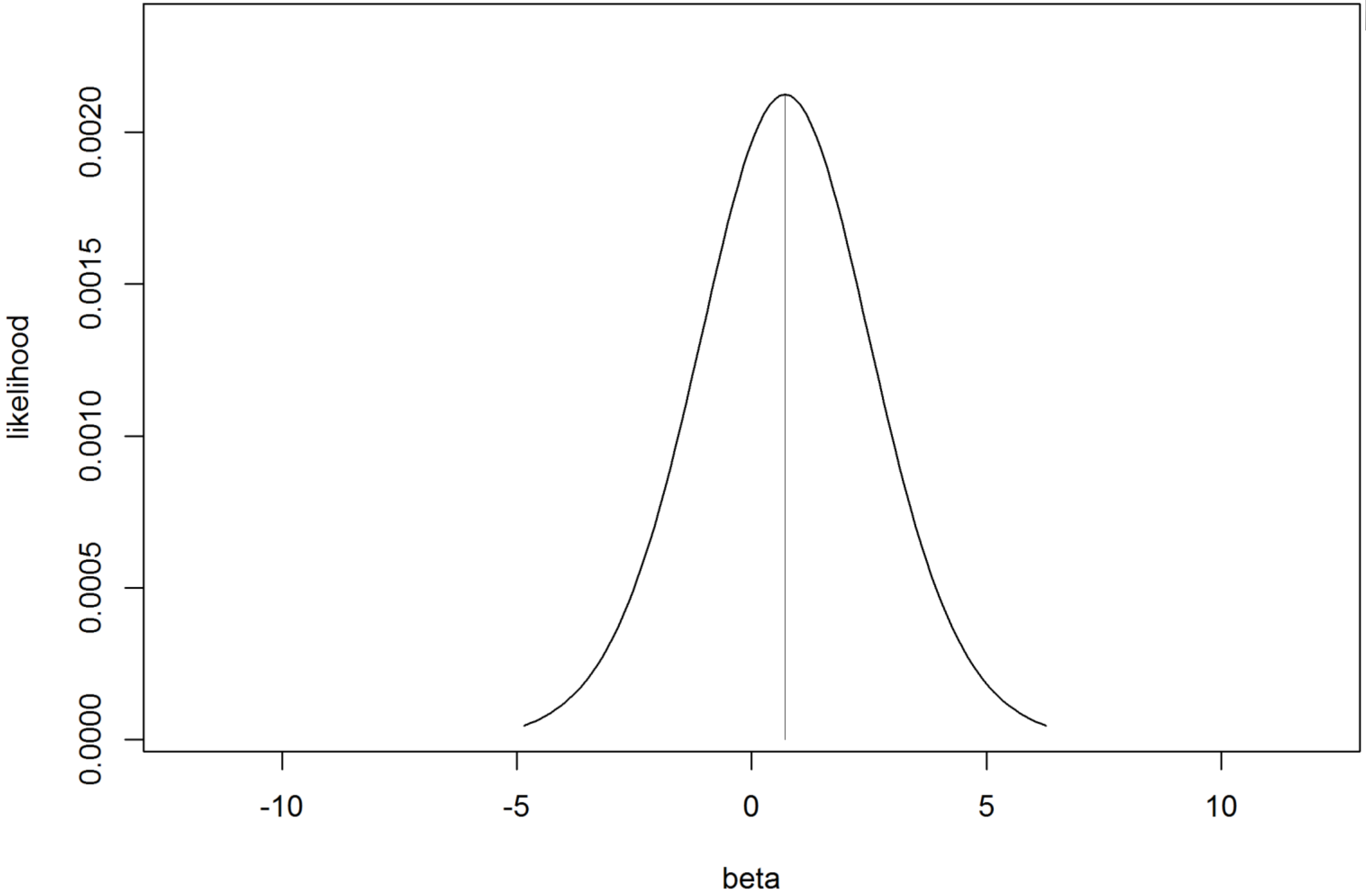
| | X=0 | X=1 | |
|---|---|---|---|
| Y=0 | 315 | 5 | *320* |
| Y=1 | 31 | 1 | *32* |
| | *346* | *6* | *352* |

← Rare outcome

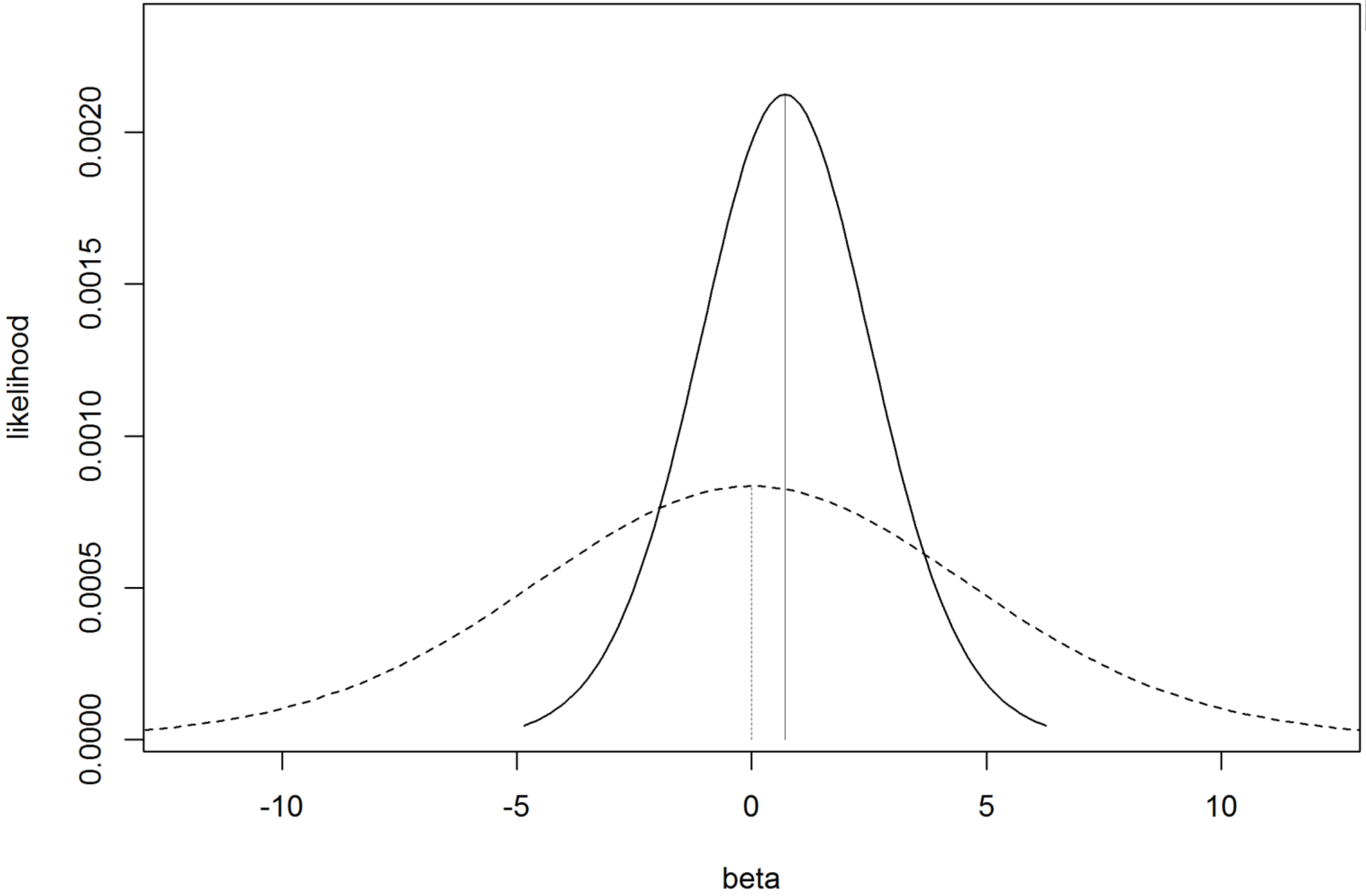
↑ Rare exposure

Greenland, AmStat 2010

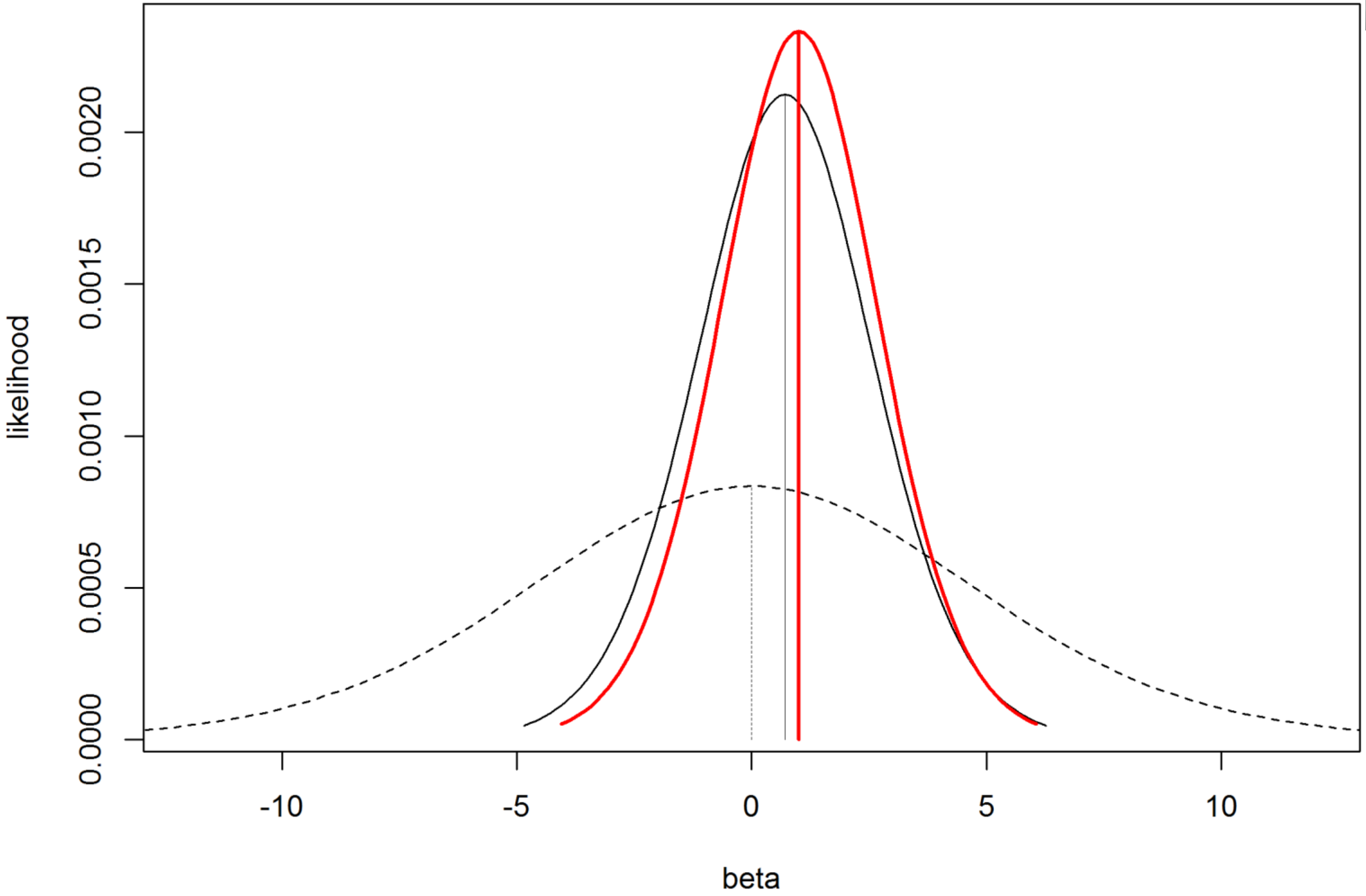
# Likelihood, *prior*, posterior

# Likelihood, *prior*, posterior

# Likelihood, *prior*, posterior

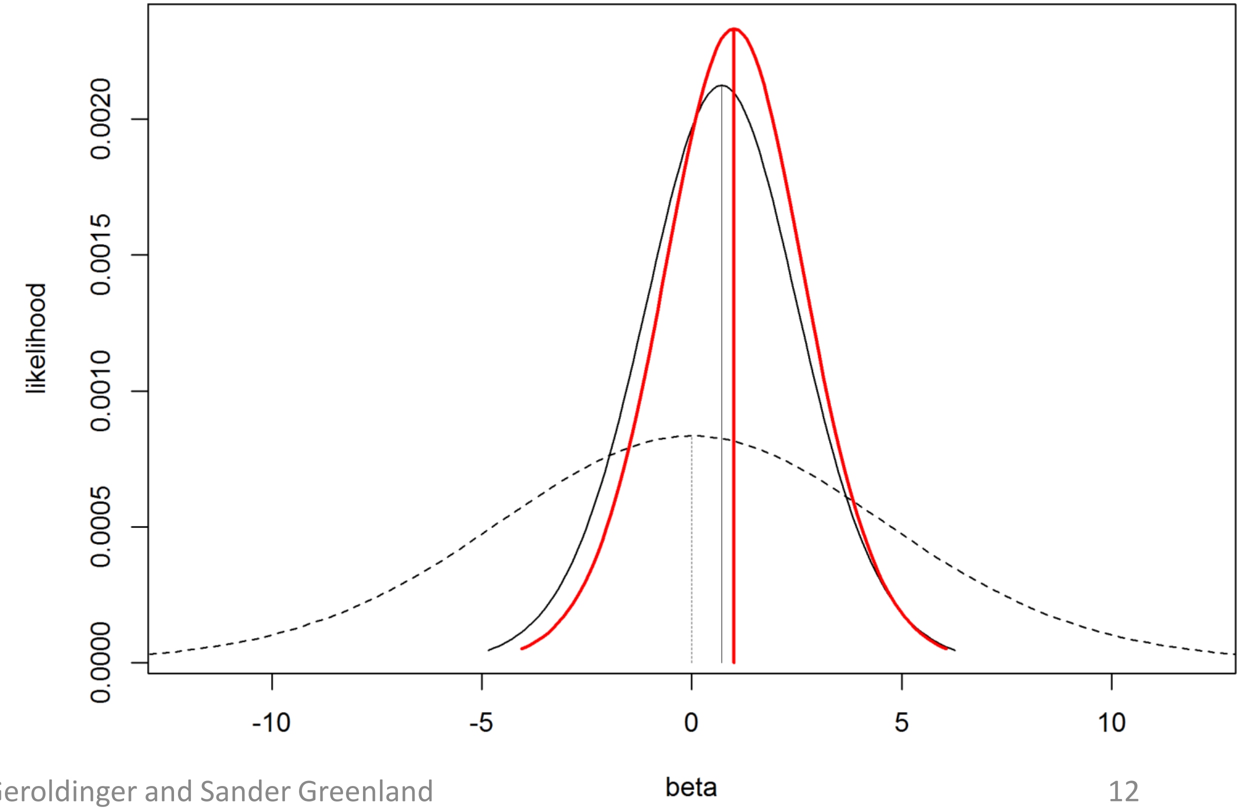



# Bayesian non-collapsibility:


timeout.com

- Prior and likelihood modes do not ‚collapse‘: posterior mode exceeds both
- The posterior mode is more extreme than the ML estimate (likelihood mode)

- How can that happen???



# An even more extreme example from Greenland 2010

timeout.com

- 2x2 table

| | X=0 | X=1 | |
|---|---|---|---|
| Y=0 | 25 | 5 | *30* |
| Y=1 | 5 | 1 | *6* |
| | *30* | *6* | *36* |

- Here we immediately see that the odds ratio = 1 ($\beta_1 = 0$)

- But the estimate from augmented data: odds ratio = 1.26 (try it out!)

Greenland, AmStat 2010

# Reason for Bayesian non-collapsilibity

timeout.com

- We look at the association of X and Y

- We could treat the source of data as a ‚ghost factor' G
- G=0 for original table
- G=1 for pseudo data

this is also the basic idea of
Puhr et al's (2017) FLAC method

- We ignore that the conditional association of X and Y given G is different from the marginal association

# Simulating the example of Greenland

timeout.com

- We should distinguish BNC in a single data set from a systematic increase in bias of a method (in simulations)

- (This is only of interest to frequentists)

- Simulation of the example:
- Fixed groups x=0 and x=1, P(Y=1|X) as observed in example
- True log OR=0.709

| | X=0 | X=1 | |
|---|---|---|---|
| Y=0 | 315 | 5 | *320* |
| Y=1 | 31 | 1 | *32* |
| | *346* | *6* | *352* |

# Simulating the example of Greenland

timeout.com

- True value: log OR = 0.709

| **Parameter** | **ML** | **Jeffreys-Firth** | |
|---|---|---|---|
| Bias $\beta_1$ | * | +18% | |
| RMSE $\beta_1$ | * | 0.86 | |
| **Bayesian non-collapsibility $\boldsymbol{\beta_1}$** | | **63.7%** | |

* Separation causes $\beta_1$ to be undefined ($-\infty$) in 31.7% of the cases (Mansournia et al AJE 2017)

timeout.com

# Simulating the example of Greenland

- To overcome Bayesian non-collapsibility,
  Greenland and Mansournia (2015)
  have proposed not to impose a prior on the intercept

- They suggest a log-F(1,1) prior for all other regression coefficients

- The method can be used with conventional frequentist software
  because it uses a data-augmentation prior (which can be imposed by adding pseudo-data and replacing the intercept with the source indicator G)

Greenland and Mansournia, StatMed 2015

# Simulating the example of Greenland


timeout.com

- Re-running the simulation with the log-F(1,1) method yields:

| Parameter | ML | Jeffreys-Firth | logF(1,1) |
|---|---|---|---|
| Bias $\beta_1$ | * | +18% | |
| RMSE $\beta_1$ | * | 0.86 | |
| **Bayesian non-collapsibility $\boldsymbol{\beta_1}$** | | **63.7%** | **0%** |

* Separation causes $\beta_1$ be undefined ($-\infty$) in 31.7% of the cases

# Simulating the example of Greenland

timeout.com

- Re-running the simulation with the log-F(1,1) method yields:

| Parameter | ML | Jeffreys-Firth | logF(1,1) |
|---|---|---|---|
| Bias $\beta_1$ | * | +18% | -52% |
| RMSE $\beta_1$ | * | 0.86 | 1.05 |
| **Bayesian non-collapsibility $\boldsymbol{\beta_1}$** | | **63.7%** | **0%** |

* Separation causes $\beta_1$ be undefined ($-\infty$) in 31.7% of the cases

# Other, more subtle occurrences of Bayesian non-collapsibility


timeout.com

- Ridge regression: normal prior around 0
- usually implies bias towards zero,
- But:
- With correlated predictors with different effect sizes, for some predictors the bias can be away from zero

# Simulation of bivariable log reg models


timeout.com

- $X_1, X_2 \sim \text{Bin}(0.5)$ with correlation $r = 0.8, n = 50$
- $\beta_1 = 1.5$, $\beta_2 = 0.1$, ridge parameter $\lambda$ was optimized by cross-validation

| Parameter | True value | ML | Ridge ($\lambda_{opt}$) | Log-F(1,1) | Jeffreys-Firth |
|---|---|---|---|---|---|
| Bias $\beta_1$ | 1.5 | +40% (+9%*) | -26% | -2.5% | +1.2% |
| RMSE $\beta_1$ | | 3.04 (1.02*) | 1.01 | 0.73 | 0.79 |
| Bias $\beta_2$ | 0.1 | -451% (+16%*) | +48% | +77% | +16% |
| RMSE $\beta_2$ | | 2.95 (0.81*) | 0.73 | 0.68 | 0.76 |
| **Bayesian non-collapsibility $\boldsymbol{\beta_2}$** | | | **25%** | **28%** | **23%** |

*excluding 2.7% separated samples

# Confidence intervals

timeout.com

- Appropriate coverage of Wald (-/+ 1.96SE) intervals?
  Needs unbiased estimators!
- Penalized profile-likelihood (PPL) intervals are advisable instead:
  - They do not depend on the point estimate
  - They provide at least good coverage averaged over the prior that produced the penalty. Gustafson and Greenland, 2009

- Penalty can be expressed as prior which does not depend on observed *Y* for:
  - log F method
  - Jeffreys/Firth in saturated models
- The prior depends on *Y* (directly or through a tuning parameter):
  - Jeffreys/Firth in non-saturated models: good coverage (by simulation)
  - Ridge: coverage levels violated Puhr et al, 2017

# Conclusion


timeout.com

Bayesian:

- Bayesian non-collapsibility is usually unintended
- Can be avoided in univariable models,
  but no general rule to avoid it in multivariable models

Frequentist:

- Frequentist looks at repeated-sampling properties (bias, RMSE)
- Likelihood penalization can often decrease RMSE (even with BNC)
- Likelihood penalization $\neq$ guaranteed shrinkage

# References


- Chen MH, Ibrahim JG, Kim S. Properties and Implementation of Jeffreys‘s Prior in Binomial Regression Models. JASA 2008; **103**:1659-1664.
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**(1):27–38.
- Greenland S. Simpson’s paradox from adding constants in contingency tables as an example of Bayesian noncollapsibility. *The American Statistician* 2010; **64**(4):340–344.
- Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine* 2015; **34**(23):3133–3143.
- Gustafson P, Greenland S. Interval Estimation for Messy Observational Data. *Statistical Science* 2009; **24**:328-342.
- Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002; **21**(16):2409–2419.
- Jeffreys H. An Invariant Form of the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London A* 1946; **196**:453-461.
- Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in Logistic Regression – Causes, Consequences, and Control. *American Journal of Epidemiology* 2017; early view
- Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth‘s logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine* 2017; **36**:2302-2317.