Reference models in variable selection

Aki Vehtari

Aalto University

Finnish Center for Artificial Intelligence FCAI

Stan

ArviZ

with Bürkner, Catalina, Le, Lindgren, McLatchie, Ojanen, Paasiniemi, Pavone, Piironen, Rögnvaldsson, Weber

Reference models in variable selection

Variable selection

1. is not needed to avoid overfitting
2. can be used to reduce costs and improve explainability

Reference models

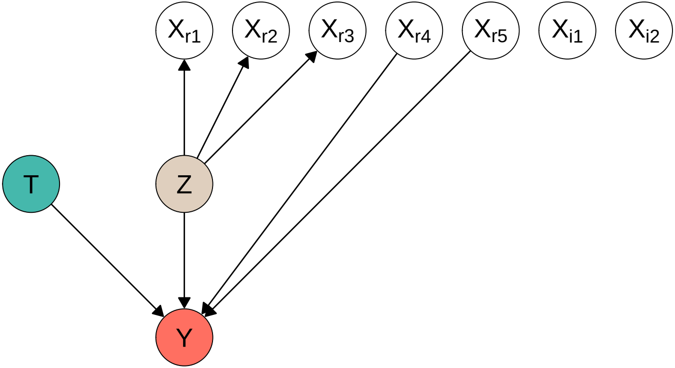
3. improve stability and reduces overfitting in selection
4. projection of the reference model is even better

Causal assumptions

- Causal assumptions affect which variables should be excluded from the selection (either always included or always excluded from the model)

Causal assumptions

- Causal assumptions affect which variables should be excluded from the selection (either always included or always excluded from the model)

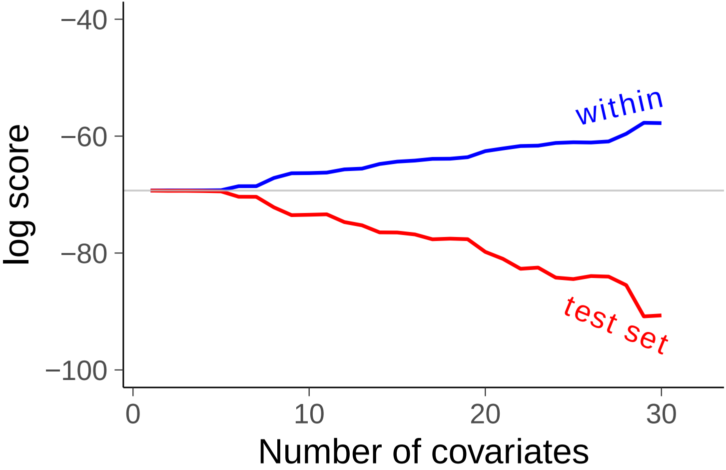


Model selection is needed to avoid overfitting?

logistic regression: 30 **completely irrelevant** variables, 100 observations

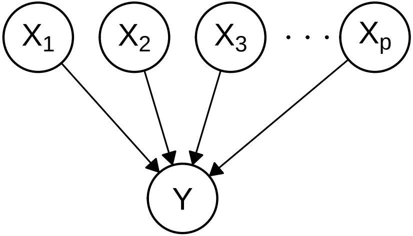
Model selection is needed to avoid overfitting?

logistic regression: 30 **completely irrelevant** variables, 100 observations



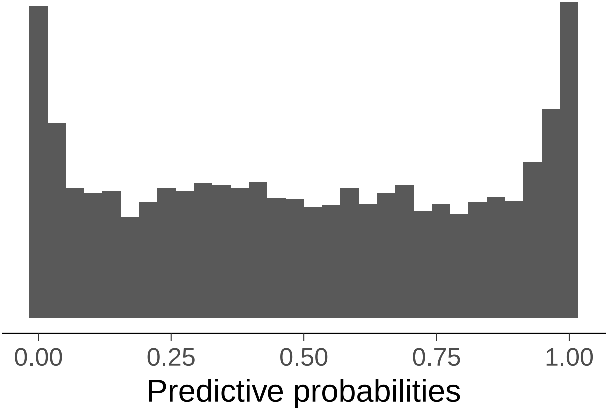
Prior on parameters vs predictions

N(0,3) prior on each coefficient



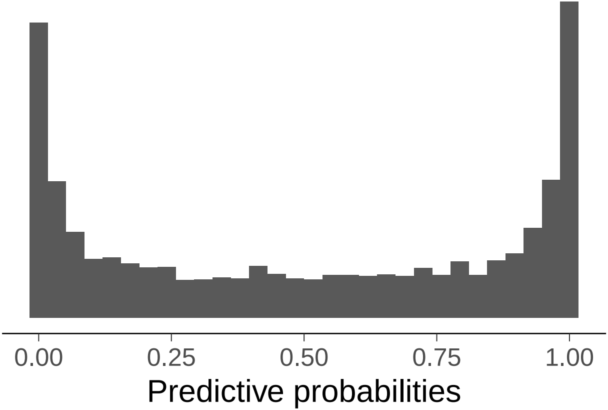
Prior on parameters vs predictions

N(0,3) prior on each coefficient
1 variable



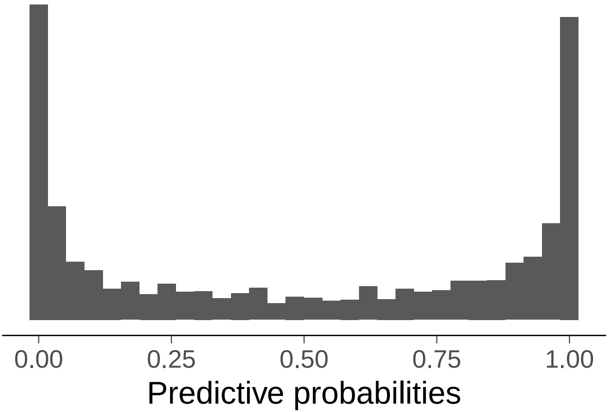
Prior on parameters vs predictions

N(0,3) prior on each coefficient
2 variables



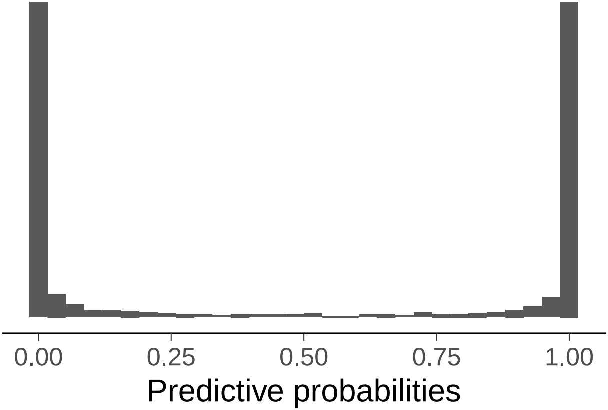
Prior on parameters vs predictions

N(0,3) prior on each coefficient
3 variables



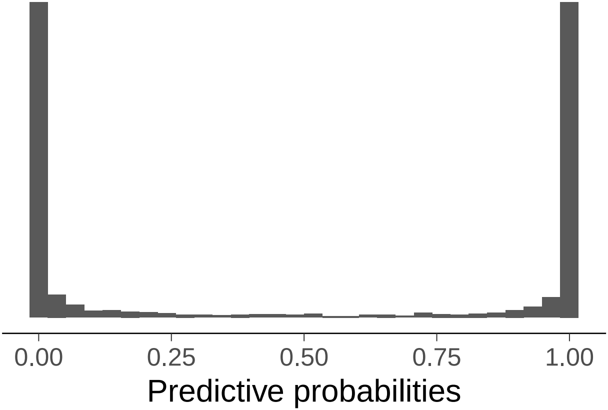
Prior on parameters vs predictions

N(0,3) prior on each coefficient
30 variables



Prior on parameters vs predictions

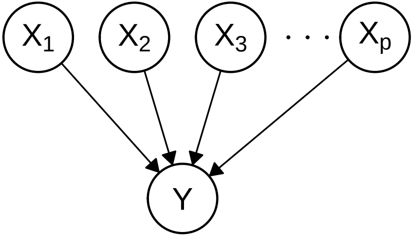
N(0,3) prior on each coefficient
30 variables



A weak prior on parameters can be a strong prior on predictions

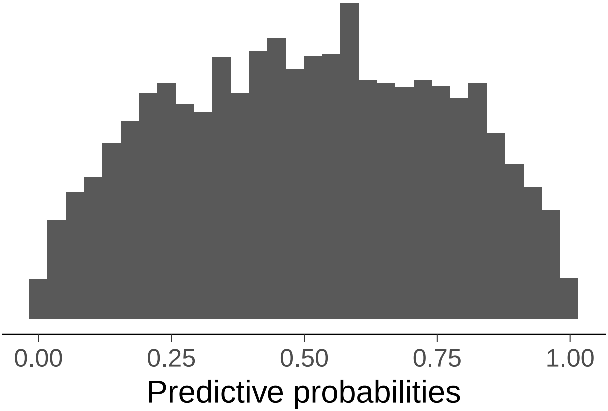
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient



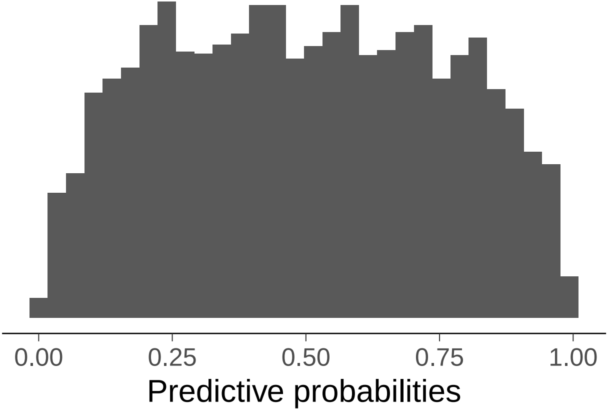
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
1 variable



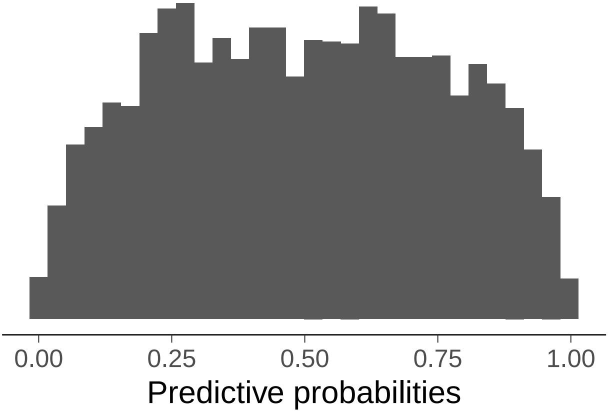
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
2 variables



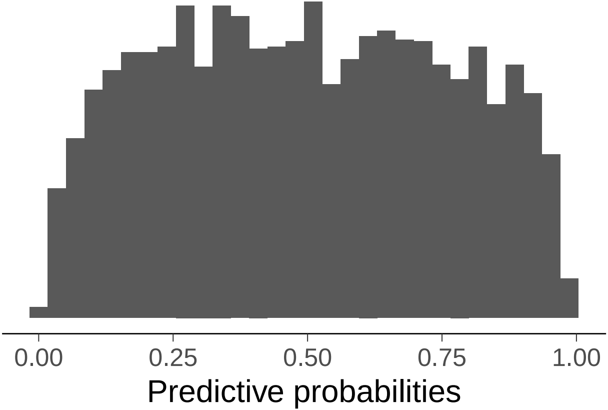
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
3 variables



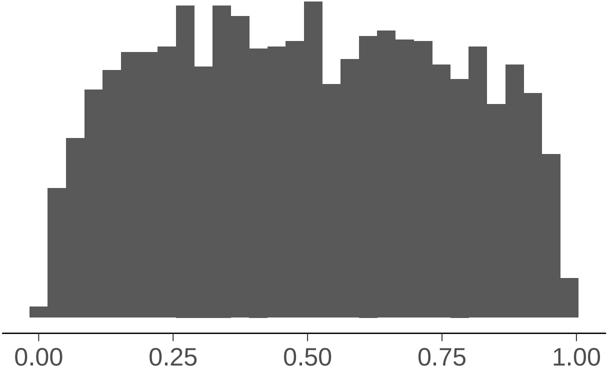
Better priors

$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
30 variables



Better priors

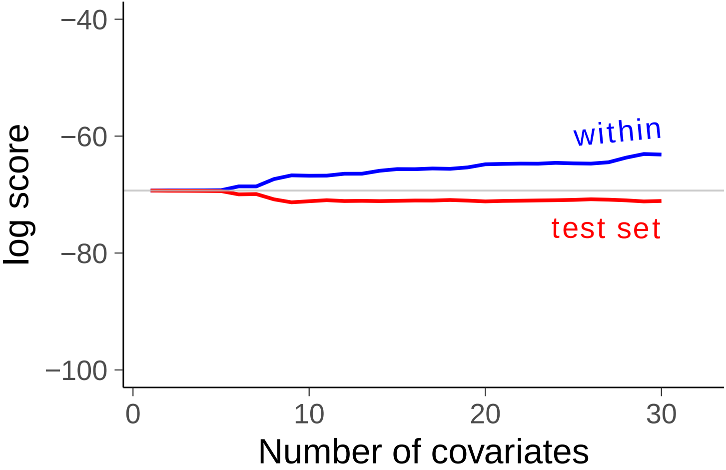
$N(0, \frac{1}{\sqrt{p}})$ prior on each coefficient
30 variables



Prior on predictions (almost) fixed when the model gets bigger

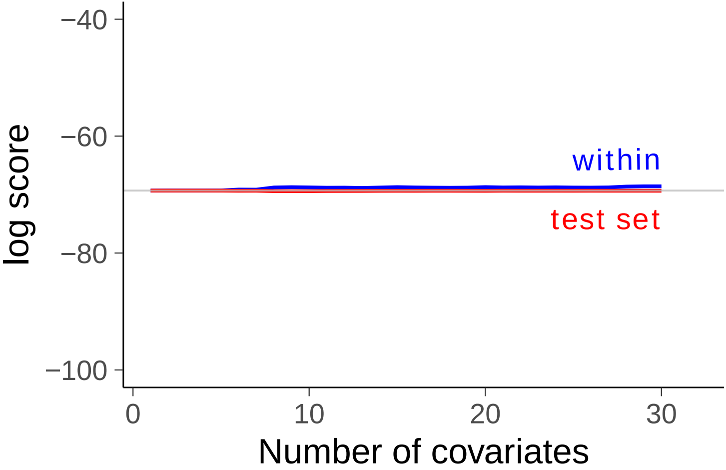
Better priors, no overfitting

logistic regression: 30 **completely irrelevant** variables,
100 observations



Better priors, no overfitting

logistic regression: 30 **completely irrelevant** variables, 100 observations

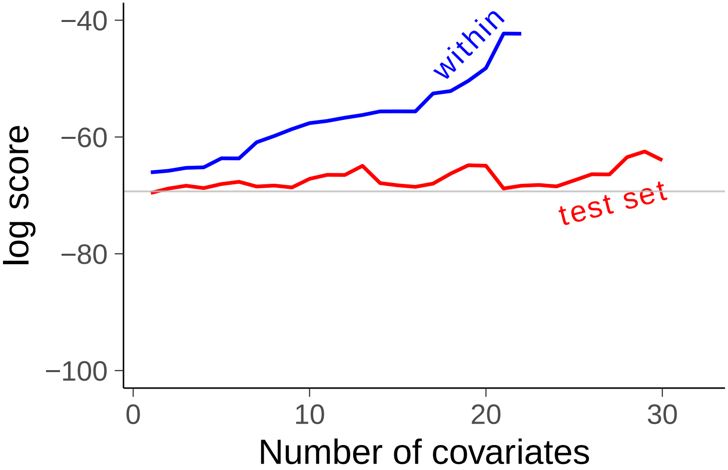


Many weak effects, wide prior on parameters

logistic regression: 30 **weakly relevant** variables,
100 observations, wide prior

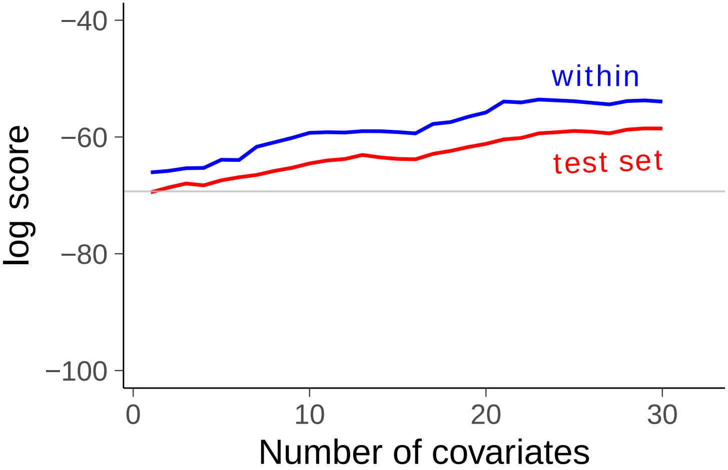
Many weak effects, wide prior on parameters

logistic regression: 30 **weakly relevant** variables,
100 observations, wide prior



Many weak effects, better prior

logistic regression: 30 **weakly relevant** variables,
100 observations, better prior

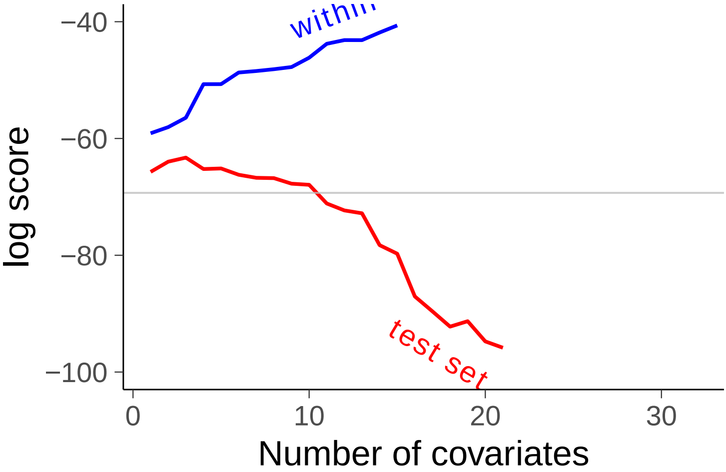


Correlating variables, wide prior on parameters

logistic regression: 30 **correlating relevant** variables, 100 observations

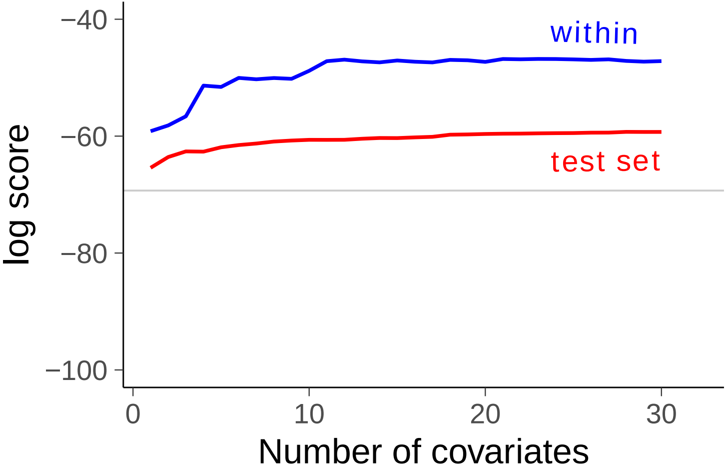
Correlating variables, wide prior on parameters

logistic regression: 30 **correlating relevant** variables,
100 observations



Correlating variables, better prior

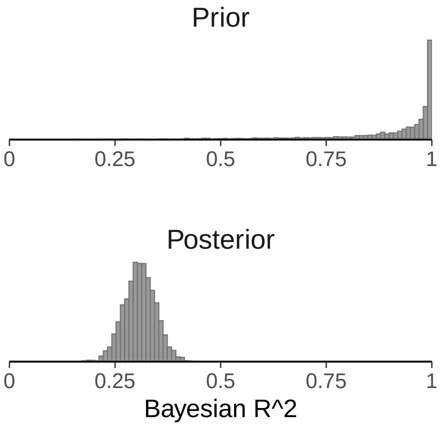
logistic regression: 30 **correlating relevant** variables,
100 observations



Prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

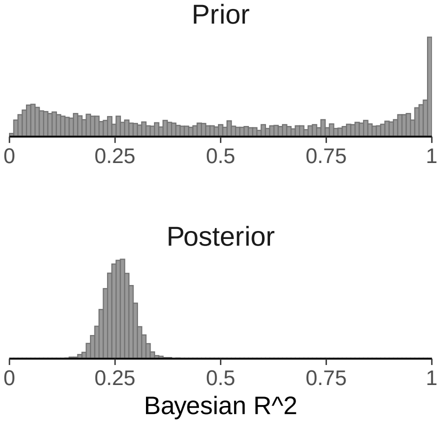
Wide prior



Prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

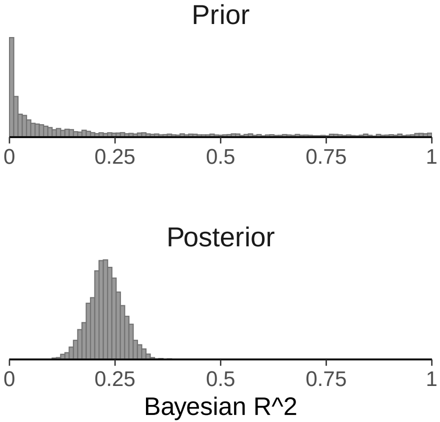
Scaled prior



Prior on R^2

Regression and Other Stories, Section 12.7 Models for regression coefficients:

Regularized horseshoe prior



Better priors

For example:

- scaled: many weak effects
- regularized horseshoe, R2-D2: only some relevant
- R2-D2: defined directly for R^2
- PCA-type: highly correlating variables

$p \gg n$

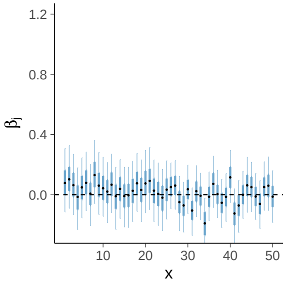
- With good priors, possible to have more variables than observations
- e.g. $p = 22283, n = 85$ demonstrated by Piironen, Paasiniemi, Vehtari (2020)

Variable selection

Variable selection

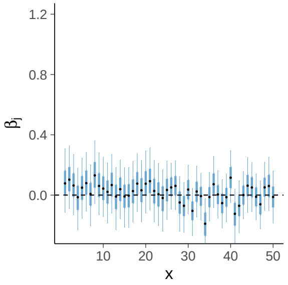
1. is not needed to avoid overfitting
2. can be used to reduce costs and improve explainability

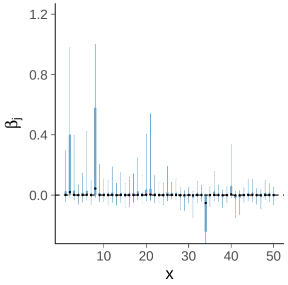
Variable selection by looking at the posterior



A) Gaussian prior, posterior median with 50% and 90% intervals

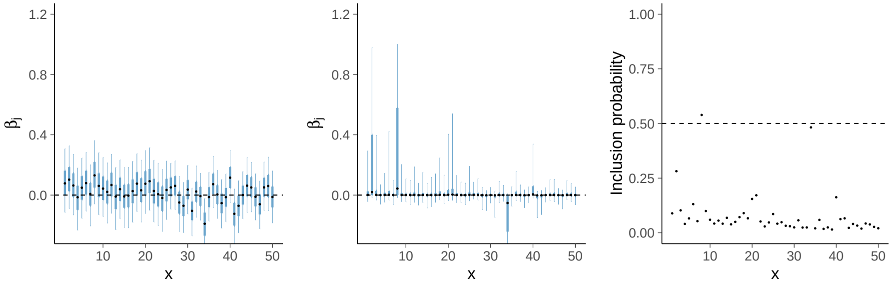
Variable selection by looking at the posterior





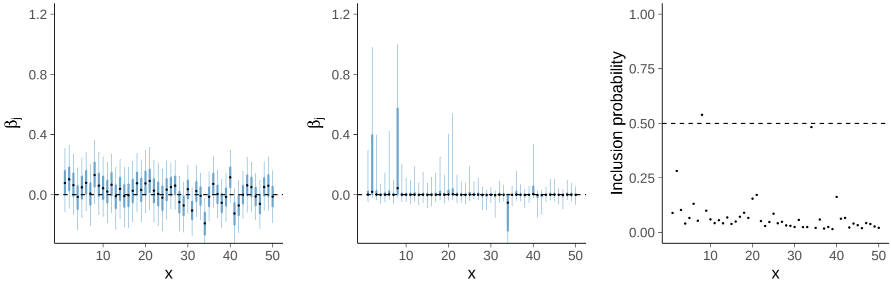
A) Gaussian prior, posterior median with 50% and 90% intervals
B) Horseshoe prior, same things

Variable selection by looking at the posterior



A) Gaussian prior, posterior median with 50% and 90% intervals
B) Horseshoe prior, same things
C) Spike-and-slab prior, posterior inclusion probabilities

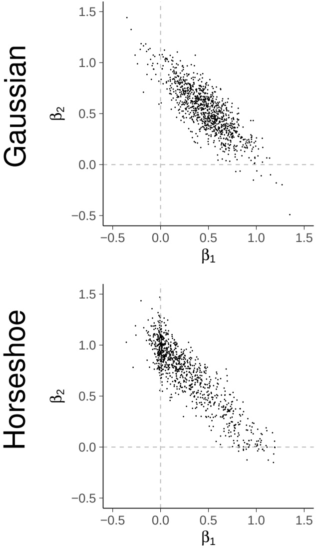
Variable selection by looking at the posterior



A) Gaussian prior, posterior median with 50% and 90% intervals
B) Horseshoe prior, same things
C) Spike-and-slab prior, posterior inclusion probabilities

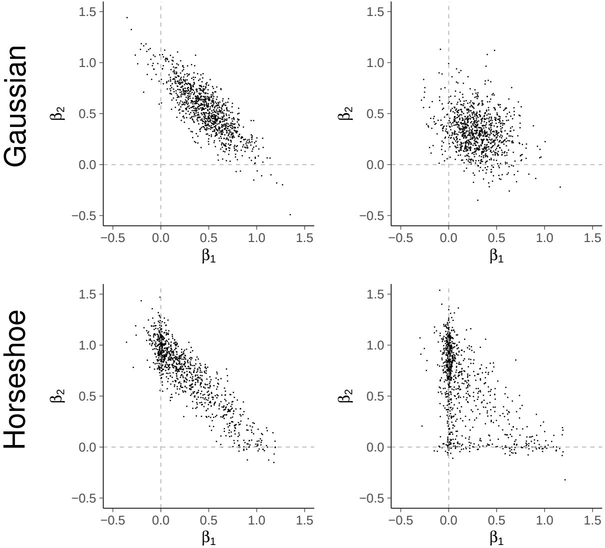
25 **correlating** variables, 25 **irrelevant** variables

What happens?



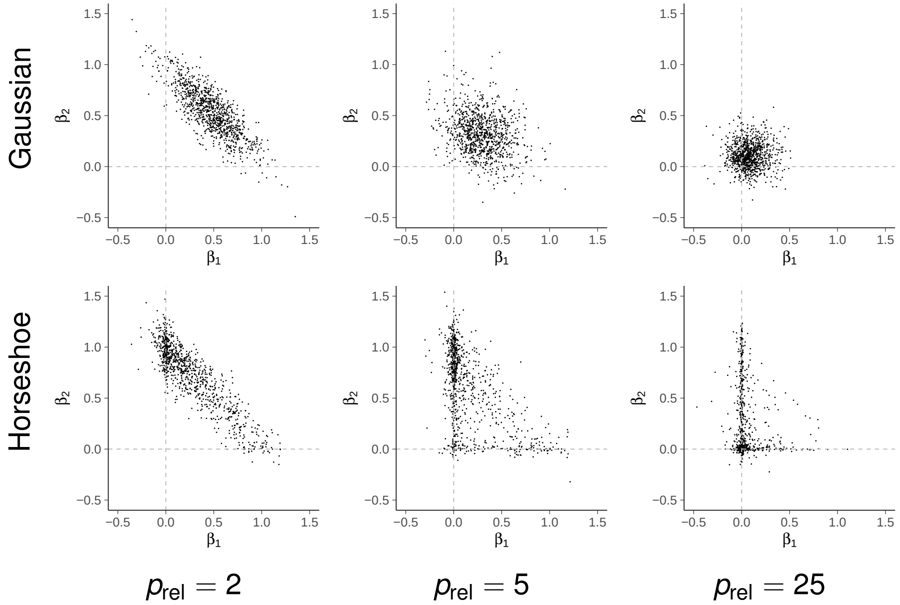
$p_{\text{rel}} = 2$

What happens?



$p_{\mathrm{rel}} = 2$ $\quad$ $p_{\mathrm{rel}} = 5$

What happens?



Variable selection

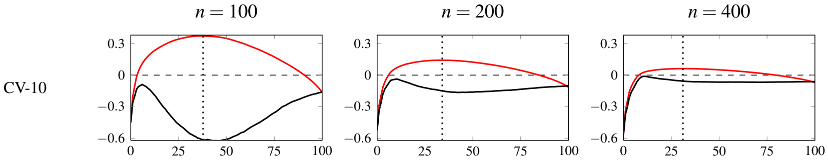
- We can do variable selection based on the predictive performance

Stepwise selection?

Mix of **correlating** and **irrelevant** variables

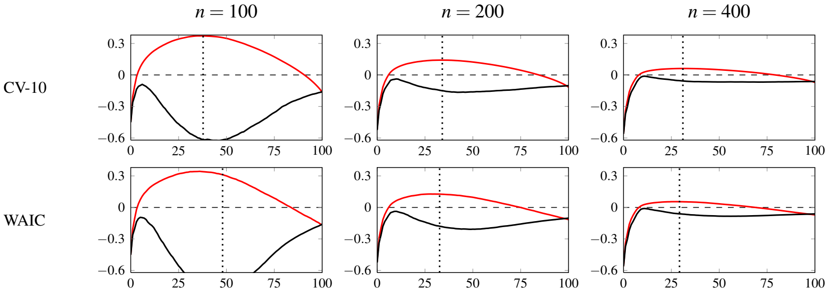
Stepwise selection?

Mix of **correlating** and **irrelevant** variables



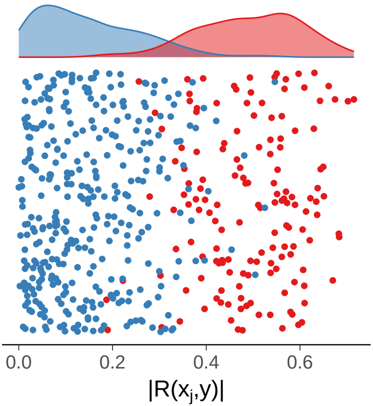
Stepwise selection?

Mix of **correlating** and **irrelevant** variables



Reference model improves variable selection

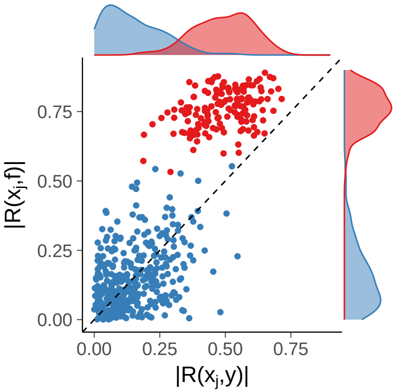
Mix of **correlating** and **irrelevant** variables
$n = 30$, $p = 500$, $p_{\text{rel}} = 150$



irrelevant x_j, relevant x_j

Sample correlation with y

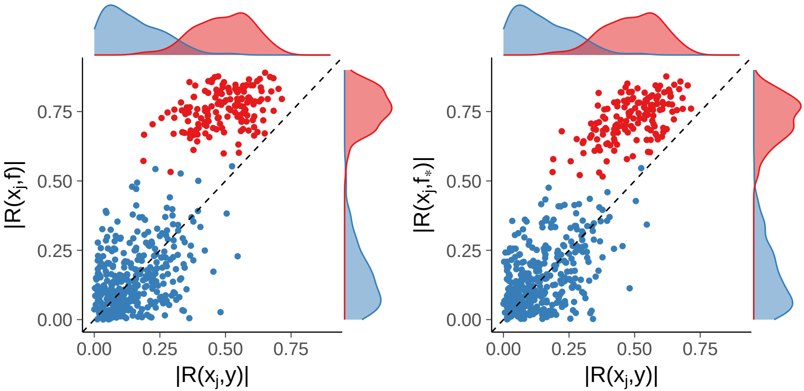
Reference model improves variable selection



irrelevant x_j, relevant x_j

A) Sample correlation with y vs. sample correlation with f

Reference model improves variable selection



irrelevant x_j, relevant x_j

A) Sample correlation with y vs. sample correlation with f

B) Sample correlation with y vs. sample correlation with f_*

$f_* =$ linear regression fit with 3 supervised principal components

Predictive projection, idea

- Model simplification technique

Predictive projection, idea

- Model simplification technique
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible

Predictive projection, idea

- Model simplification technique
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some θ_0
    $\Rightarrow$ "Optimal point estimates"

Predictive projection, idea

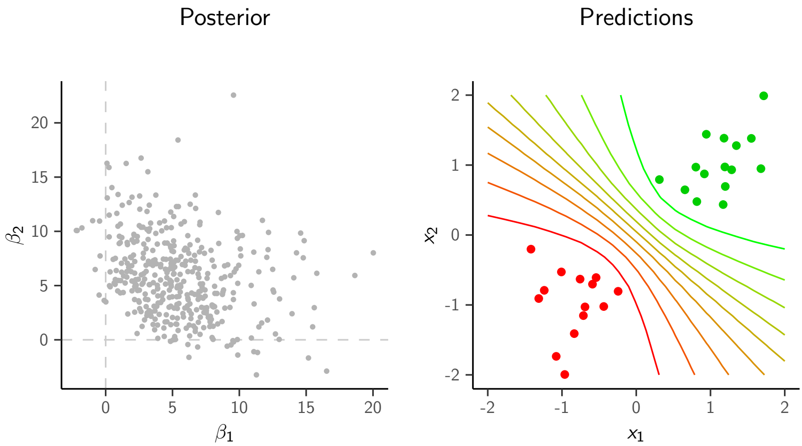
- Model simplification technique
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some θ_0
    $\Rightarrow$ "Optimal point estimates"
  - Some features must have exactly zero regression coefficient
    $\Rightarrow$ "Which features can be discarded"

Predictive projection, idea

- Model simplification technique
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some θ_0
    $\Rightarrow$ "Optimal point estimates"
  - Some features must have exactly zero regression coefficient
    $\Rightarrow$ "Which features can be discarded"

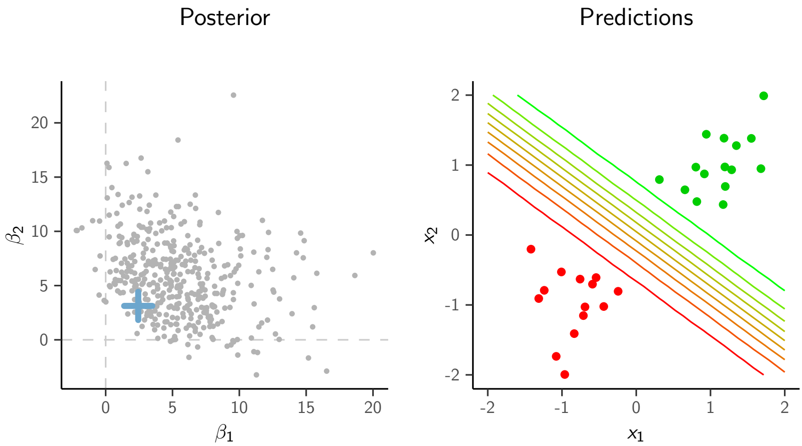
- The decision theoretic idea of conditioning the smaller model inference on the full model can be tracked to Lindley (1968)
  - draw by draw projection introduced by Goutis & Robert (1998), and Dupuis & Robert (2003)
  - see also many related references in a review by Vehtari & Ojanen (2012)

Logistic regression with two features



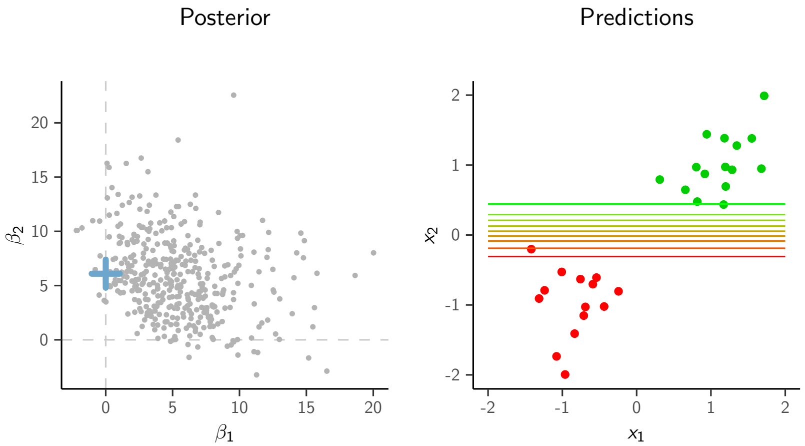
Full posterior for β_1 and β_2 and contours of predicted class probability

Logistic regression with two features



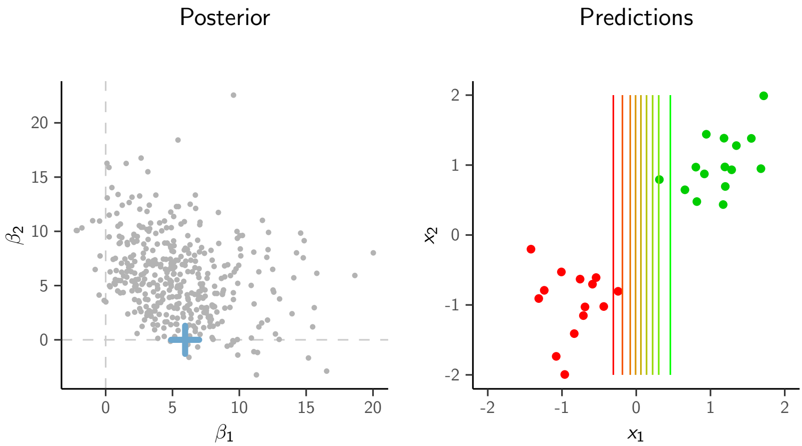
Projected point estimates for β_1 and β_2

Logistic regression with two features



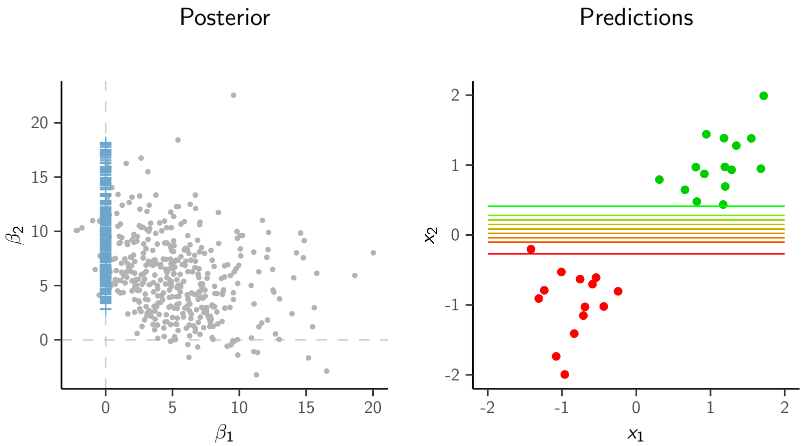
Projected point estimates, constraint $\beta_1 = 0$

Logistic regression with two features



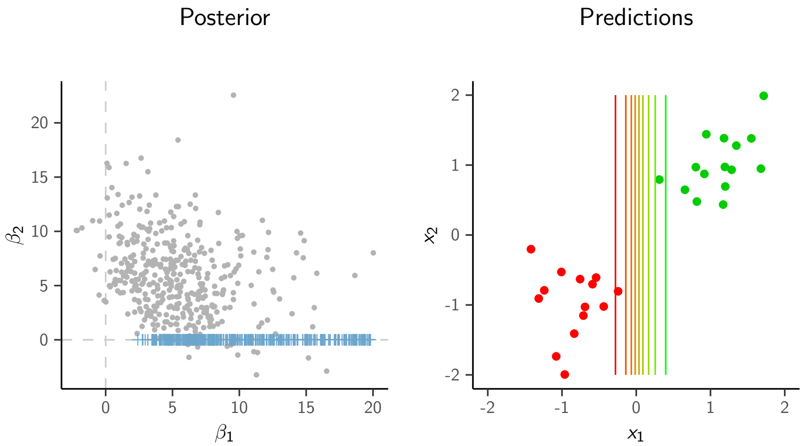
Projected point estimates, constraint $\beta_2 = 0$

Logistic regression with two features



Draw-by-draw projection, constraint $\beta_1 = 0$

Logistic regression with two features



Draw-by-draw projection, constraint $\beta_2 = 0$

Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible

Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- As the full posterior $p(\theta \mid D)$ is projected to $q(\theta)$
    - the prior is also projected and there is no need to define priors for submodels separately

Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- As the full posterior $p(\theta \mid D)$ is projected to $q(\theta)$
  - the prior is also projected and there is no need to define priors for submodels separately
  - even if we constrain some coefficients to be 0, the predictive inference is conditoned on the information related features contributed to the reference model

Projective selection

- How to select a feature combination?

Projective selection

- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss

Projective selection

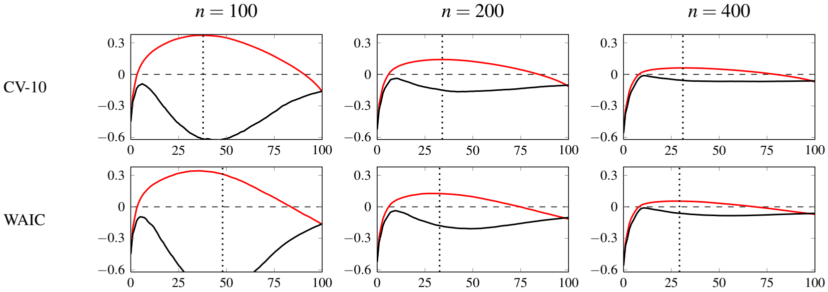
- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss
- Search heuristics, e.g.
  - Monte Carlo search
  - Forward search
  - L_1-penalization (as in Lasso)

Projective selection

- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss
- Search heuristics, e.g.
  - Monte Carlo search
  - Forward search
  - L_1-penalization (as in Lasso)
- Use cross-validation to select the appropriate model size
  - need to cross-validate over the search paths

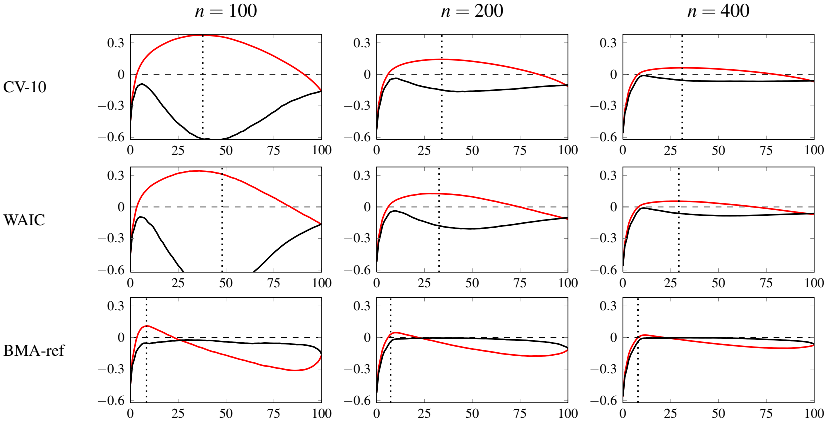
Stepwise selection

Mix of **correlating** and **irrelevant** variables



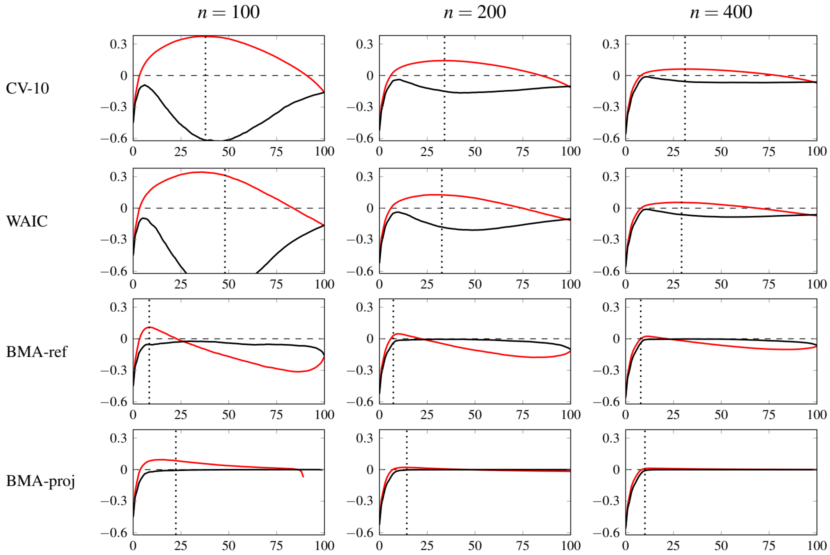
Stepwise selection

Mix of **correlating** and **irrelevant** variables



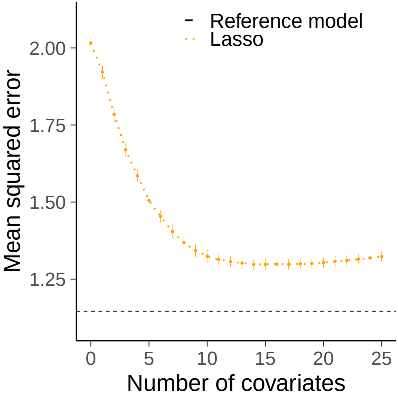
Stepwise selection

Mix of **correlating** and **irrelevant** variables



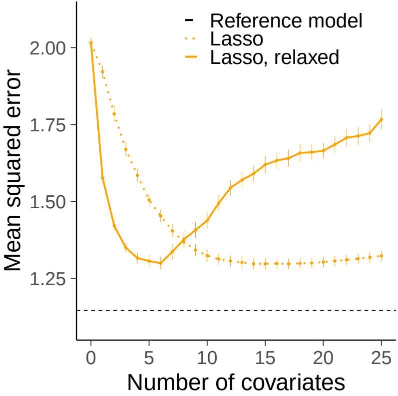
Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\mathrm{rel}} = 150$, $\rho = 0.5$



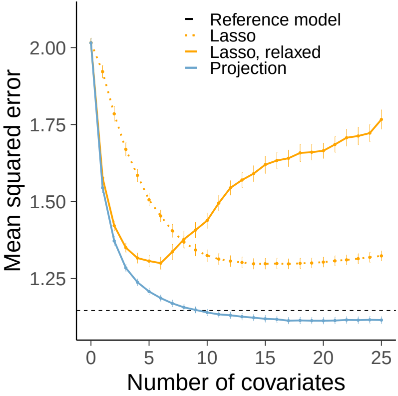
Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



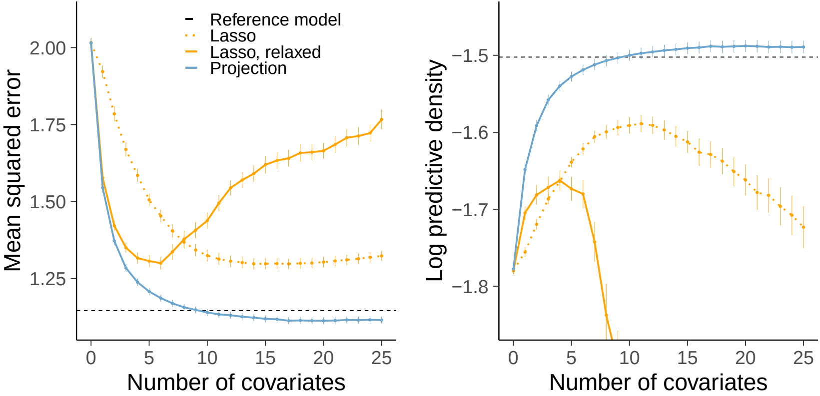
Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\mathrm{rel}} = 150$, $\rho = 0.5$



Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



Reference model improves variable selection

- Use of a reference model can improve even stepwise selection
- Projection predictive approach even better

Projection predictive inference

- Project the reference model posterior to the parameter space of a smaller model
- Choose the smallest model with similar predictive performance as the reference model

Projection predictive inference

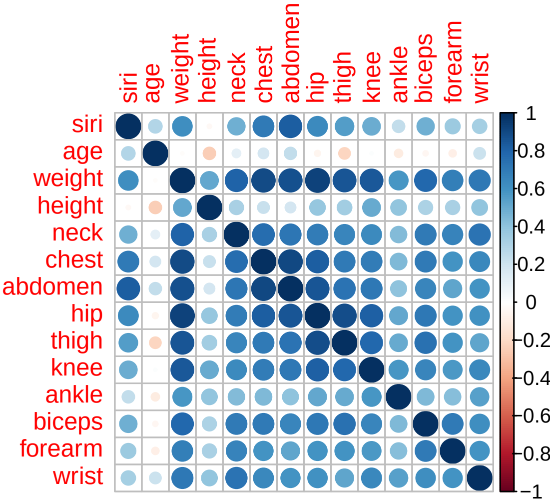
- Project the reference model posterior to the parameter space of a smaller model
- Choose the smallest model with similar predictive performance as the reference model
- Improves the selection process and provides good predictions after the selection

Bodyfat example

Predict bodyfat percentage. The reference value is obtained by immersing person in water. $n = 251$.

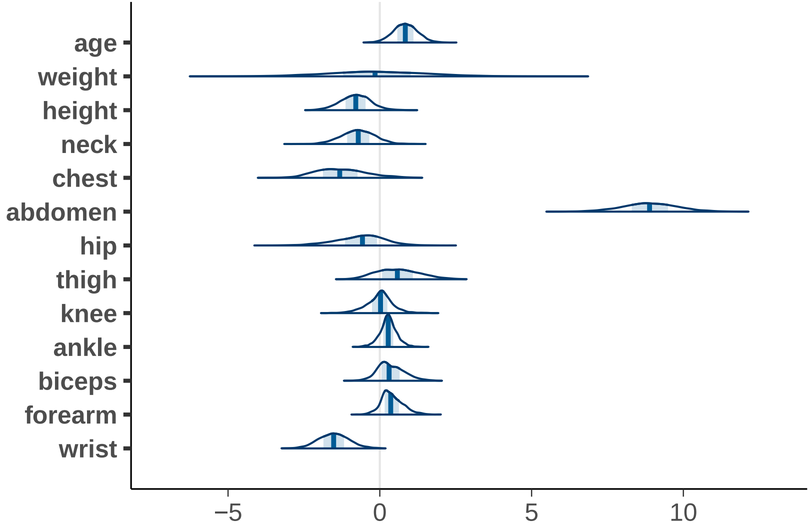
Bodyfat example

Predict bodyfat percentage. The reference value is obtained by immersing person in water. $n = 251$.



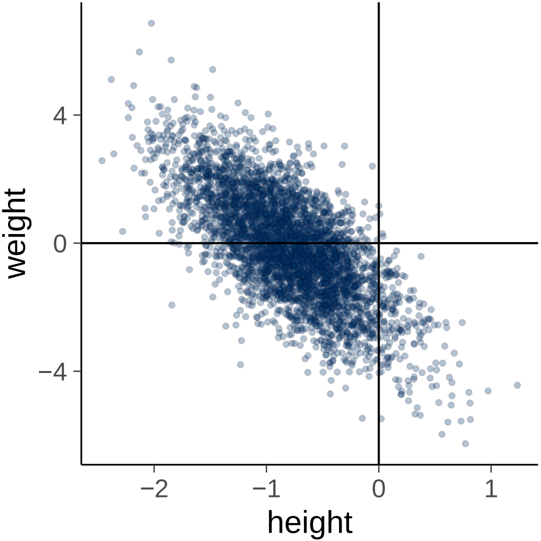
Bodyfat

Marginal posteriors of coefficients



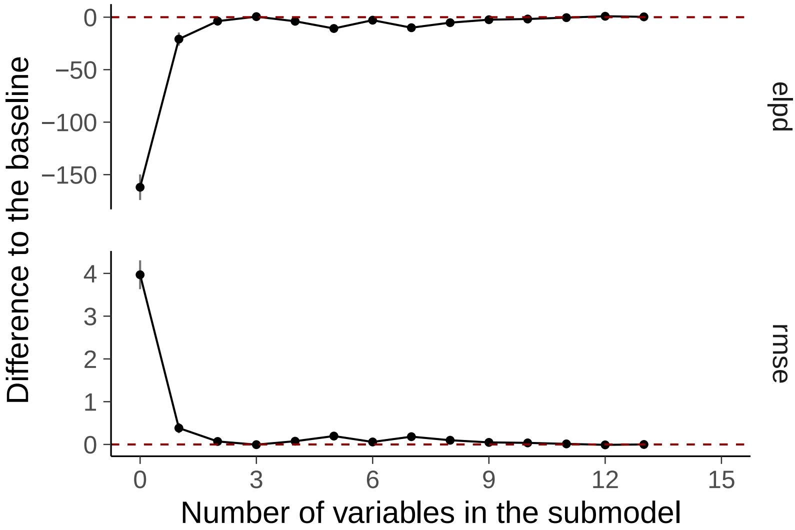
Bodyfat

Bivariate marginal of weight and height



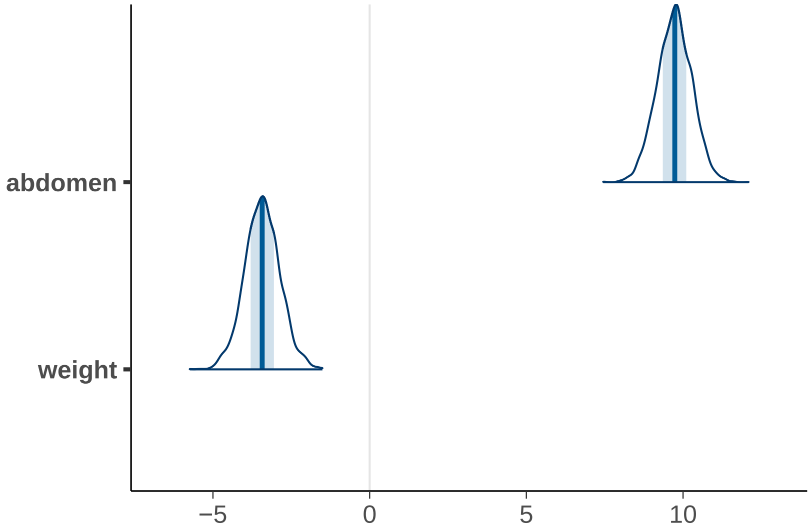
Bodyfat

The predictive performance of the full and submodels



Bodyfat

Marginals of projected posterior



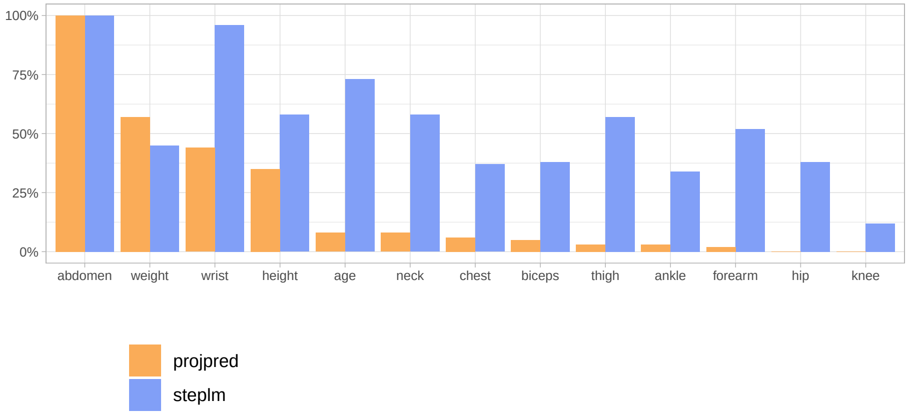
Bodyfat

Inclusion probabilities in bootstrap simulation

`projpred` vs `steplm`

Bodyfat

Inclusion probabilities in bootstrap simulation
`projpred` vs `steplm`



Bodyfat

Inclusion probabilities in bootstrap simulation
`projpred` vs `steplm`

- In case of highly correlating variables and finite data, there will be variation in the selected variables

Reference models in variable selection

Reference models

3. improve stability and reduces overfitting in selection
4. projection of the reference model is even better

Inference after selection?

- For example, for inference on treatment effect, it's best to use the big good model
- Under certain conditions, the projected posterior is also well calibrated, but we're still investigating more details

Beyond simple regression

- We have implemented projection predictive approach for
  - generalized linear models (also non-exponential family)
  - hierarchical models
  - splines
  - Gaussian processes

Beyond simple regression

- We have implemented projection predictive approach for
  - generalized linear models (also non-exponential family)
  - hierarchical models
  - splines
  - Gaussian processes
- The reference model approach can be used for any models and with any inference
  - e.g. trees and neural networks

Software for projection predictive variable/model selection

- `projpred` R package (in CRAN + github)
- `kulprit` Python package (github.com/yannmclatchie/kulprit)

Reference models in variable selection

Variable / model selection

1. is not needed to avoid overfitting
2. can be used to reduce costs and improve explainability

Reference models

3. improve stability and reduce overfitting in selection
4. projection of the reference model is even better

Calibrated causal inference

5. variable selection needs to take into account the causal assumptions

References and more results

- More results projpred vs. marginal posterior probabilities:
  Piironen and Vehtari (2017). Comparison of Bayesian predictive methods for model selection. Statistics and Computing, 27(3):711-735.
- More results projpred vs. Lasso and elastic net:
  Piironen, Paasiniemi, Vehtari (2020). Projective inference in high-dimensional problems: prediction and feature selection. *Electronic Journal of Statistics*, 14(1):2155–2197.
- More results on frequency properties and generic benefit of reference models:
  Pavone, Piironen, Bürkner, Vehtari (2022). Using reference models in variable selection. *Computational Statistics*, doi:10.1007/s00180-022-01231-6.
- Hierarchical and spline models:
  Catalina, Bürkner, and Vehtari (2022). Projection predictive inference for generalized linear and additive multilevel models. AISTATS 2022, PMLR 151:4446–4461.
- More references, and several case studies for small to moderate dimensional ($p = 4 \ldots 100$) small data:
  https://avehtari.github.io/modelselection/