Outline

- Variable selection with projpred
- Bayesian software for Python users

Variable selection

- The process of identifying the most relevant variables in a model from a larger set of predictors.
- We assume variables contribute unevenly to the outcome.
  - We may want to identify the most "important" ones.
  - Sometimes we also want to rank them.

Motivation for variable selection

- We can always include all available variables in a model

Motivation for variable selection

- We can always include all available variables in a model
- Theory says this is a good idea, in particular if we use predictively consistent priors

Motivation for variable selection

- We can always include all available variables in a model
- Theory says this is a good idea, in particular if we use predictively consistent priors
- However, sometimes we need to reduce the number of variables
  - measurement cost in covariates
  - running cost of predictive model
  - easier explanation / learn from the model

The problem with variable selection

- The number of potential models is 2^p, where p is the number of variables
- Evaluating all models can be computationally infeasible even for moderate p
- The proccess is prone to overfitting

How to overcome the problem?

- We recommend to use a technique called projection predictive inference
- It can be easily done with `brms` + `projpred`

Variable selection with projpred

- The main advantage is that it reduces overfitting
- Other advantages are:
  - Automatic model building and fitting process.
  - Reduced number of models we need to fit.
  - Reduced time it takes to fit each model.

Main concepts

- Reference model:
- Search strategy:
- Projection:

Main concepts

- Reference model:
  - A model that includes all available variables and describes the data well.
- Search strategy:
- Projection:

Main concepts

- Reference model:
  - A model that includes all available variables and describes the data well.
- Search strategy:
  - A method for searching through the model space.
- Projection:

Main concepts

- Reference model:
  - A model that includes all available variables and describes the data well.
- Search strategy:
  - A method for searching through the model space.
- Projection:
  - A way to estimate the posterior distribution of a model given a reference model.

Using a reference model is not a novel idea

- Lindley (1968): *The choice of variables in multiple regression*
  - Bayesian and decision theoretical justification, but simplified model and computation

Using a reference model is not a novel idea

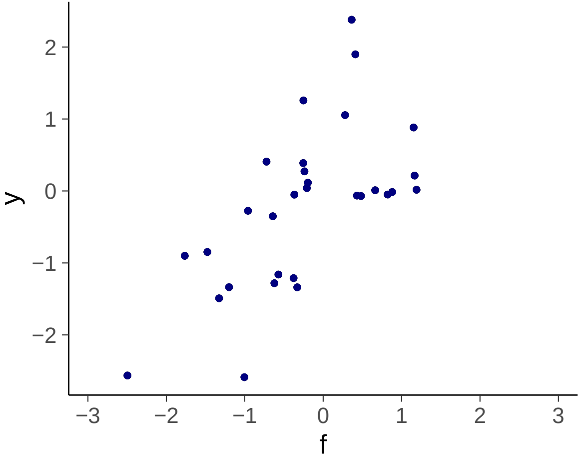
- Lindley (1968): *The choice of variables in multiple regression*
    - Bayesian and decision theoretical justification, but simplified model and computation
- Goutis & Robert (1998): *Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections*
    - one key part for practical computation

Using a reference model is not a novel idea

- Lindley (1968): *The choice of variables in multiple regression*
  - Bayesian and decision theoretical justification, but simplified model and computation
- Goutis & Robert (1998): *Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections*
  - one key part for practical computation
- Related approaches
  - gold standard, preconditioning, teacher and student, distilling, . . .

Example: Simulated regression

$$\begin{aligned} f &\sim \mathrm{N}(0,1), \\ y \mid f &\sim \mathrm{N}(f,1) \end{aligned}$$

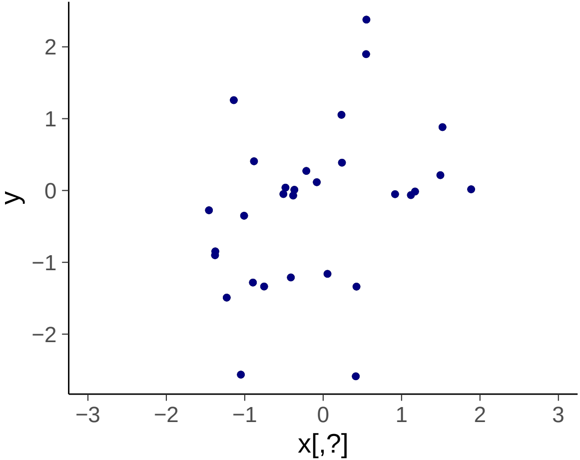


Example: Simulated regression

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1,\ldots,150\,, \\
y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151,\ldots,500\,.
\end{aligned}
$$

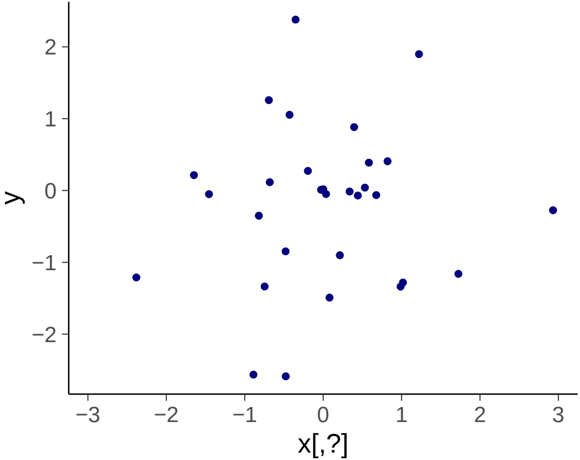
Example: Simulated regression

$$
\begin{array}{lll}
f \sim \mathrm{N}(0,1), & x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j = 1, \ldots, 150\,, \\
y \mid f \sim \mathrm{N}(f, 1) & x_j \mid f \sim \mathrm{N}(0,1), & j = 151, \ldots, 500\,.
\end{array}
$$



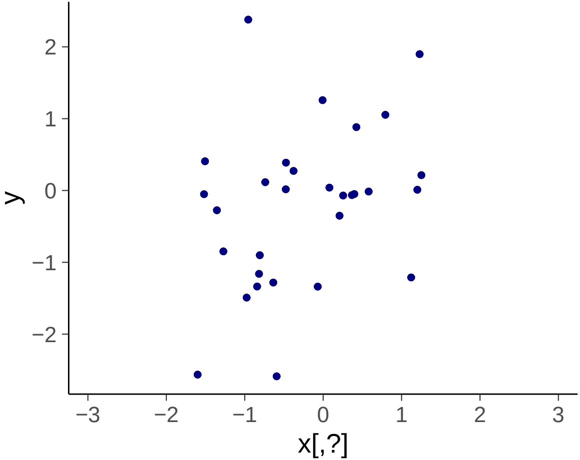
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f, 1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



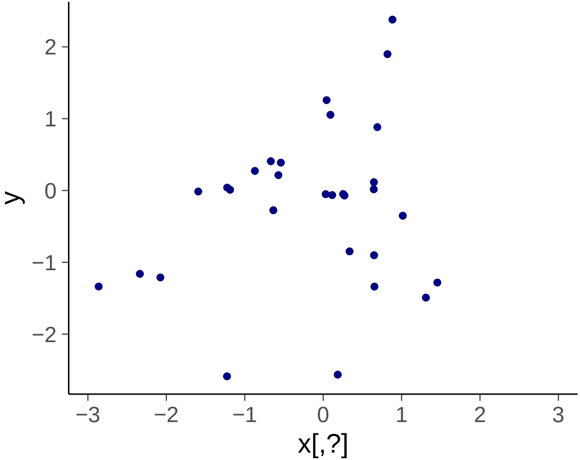
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



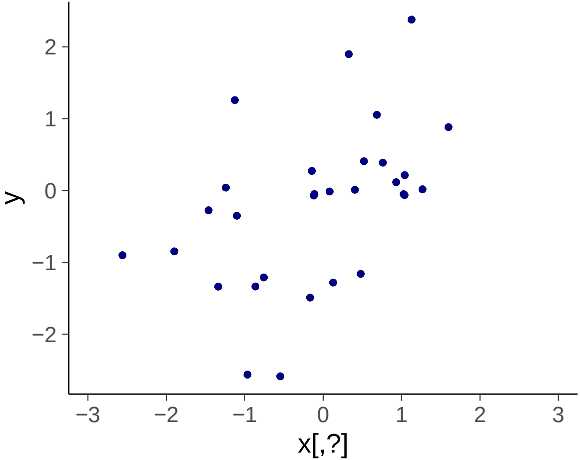
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



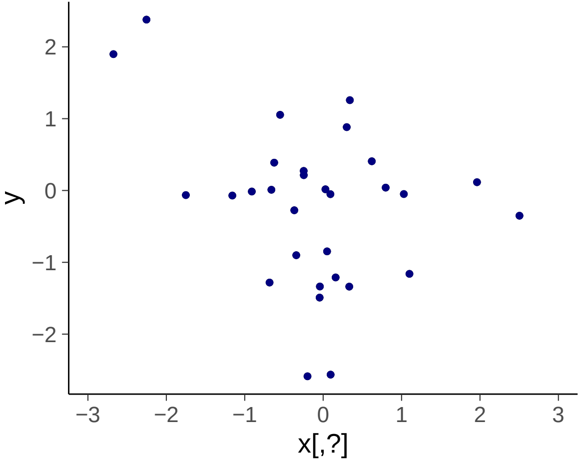
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



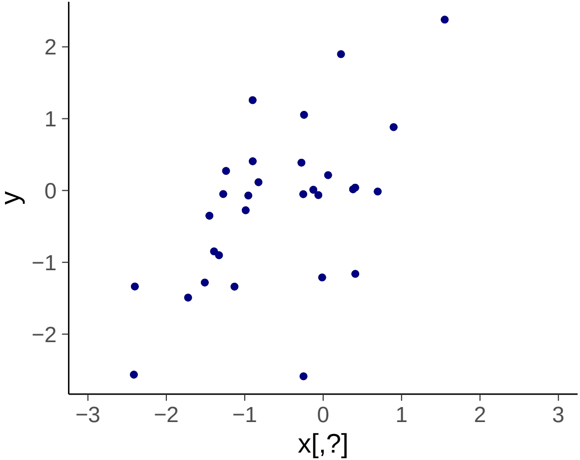
Example: Simulated regression

$$
\begin{aligned}
&f \sim \mathrm{N}(0,1), && x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), && j = 1, \ldots, 150\,,\\
&y \mid f \sim \mathrm{N}(f, 1) && x_j \mid f \sim \mathrm{N}(0,1), && j = 151, \ldots, 500\,.
\end{aligned}
$$



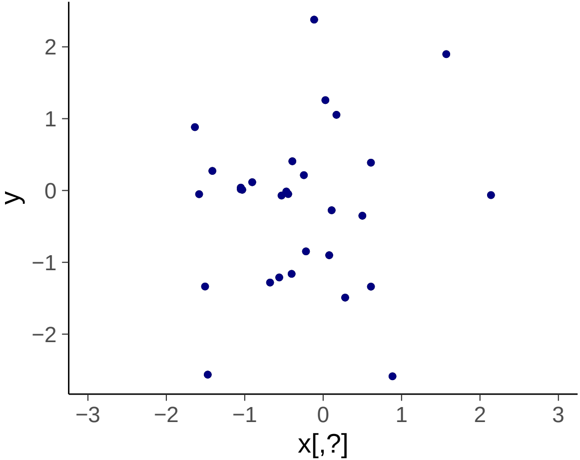
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \qquad j = 1,\ldots,150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151,\ldots,500\,.$$



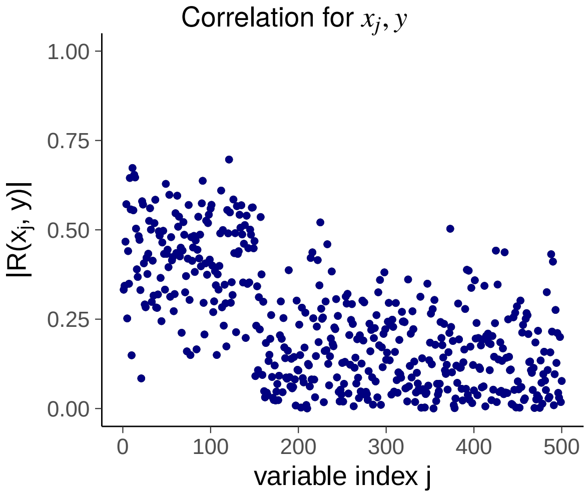
Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



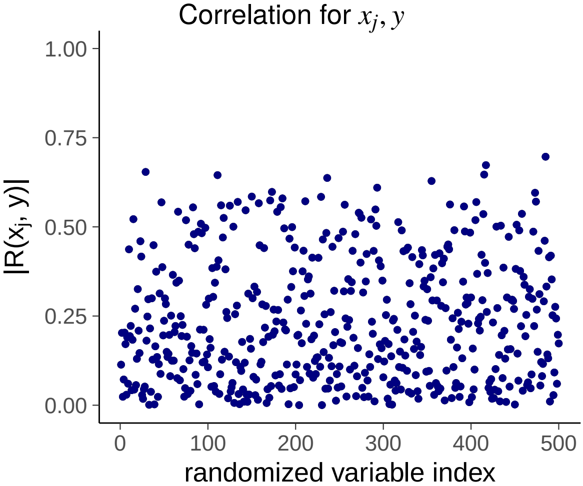
Example: Individual correlations

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \qquad j = 1, \dots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \dots, 500\,.$$



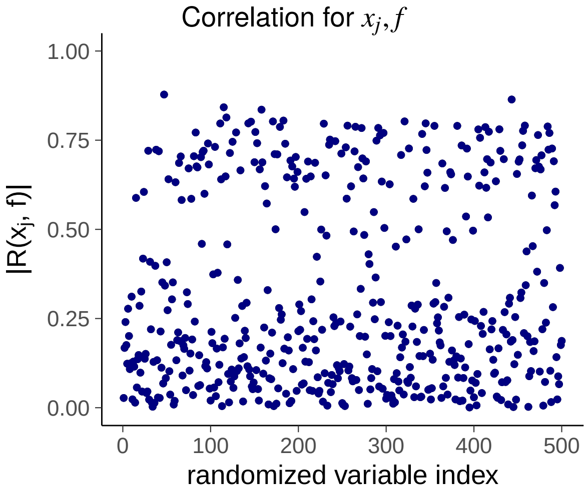
Example: Individual correlations

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1, \ldots, 150\,, \\
y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151, \ldots, 500\,.
\end{aligned}
$$



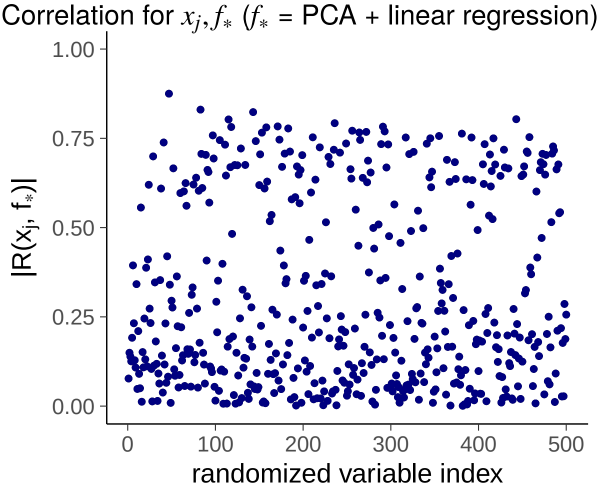
Example: Individual correlations

$$f \sim \mathrm{N}(0,1), \quad x_j \,|\, f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \quad j = 1,\ldots,150\,,$$
$$y \,|\, f \sim \mathrm{N}(f,1) \quad x_j \,|\, f \sim \mathrm{N}(0,1), \quad j = 151,\ldots,500\,.$$

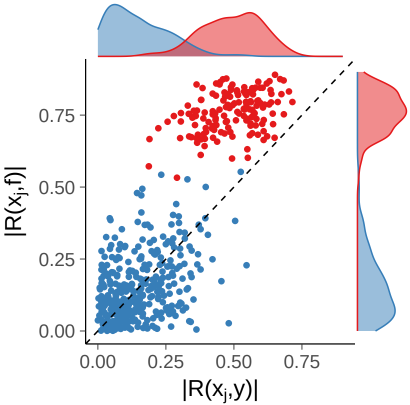


Example: Individual correlations

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), \qquad & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \qquad & j &= 1,\ldots,150\,,\\
y \mid f &\sim \mathrm{N}(f,1) \qquad & x_j \mid f &\sim \mathrm{N}(0,1), \qquad & j &= 151,\ldots,500\,.
\end{aligned}
$$



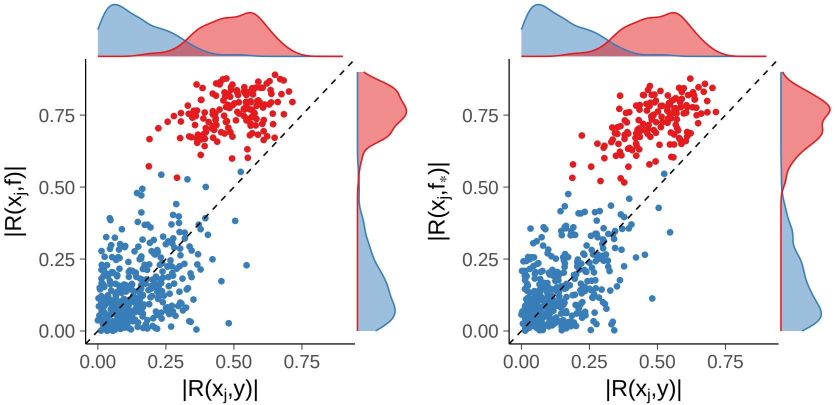
Knowing the latent values would help



irrelevant x_j, relevant x_j

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

Estimating the latent values with a reference model helps



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 principal components

Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model

Projection

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

Projection

- 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 $\theta_0 \Rightarrow$ "Optimal point estimates"

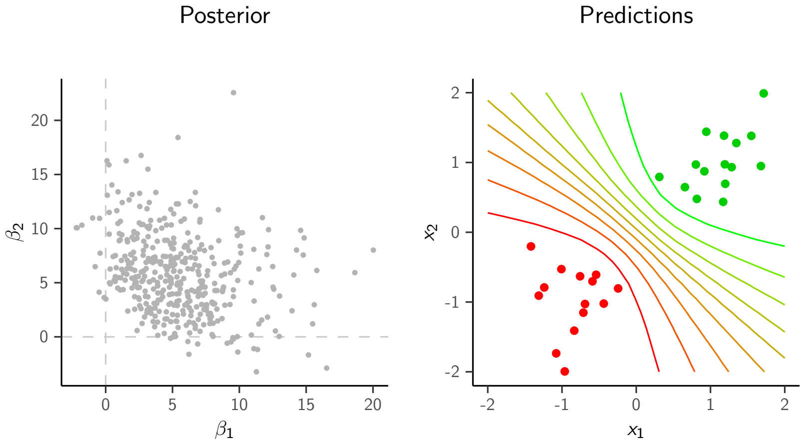
Projection

- 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 $\theta_0 \Rightarrow$ “Optimal point estimates”
  - Some covariates must have exactly zero regression coefficient $\Rightarrow$ “Which covariates can be discarded”

Projection

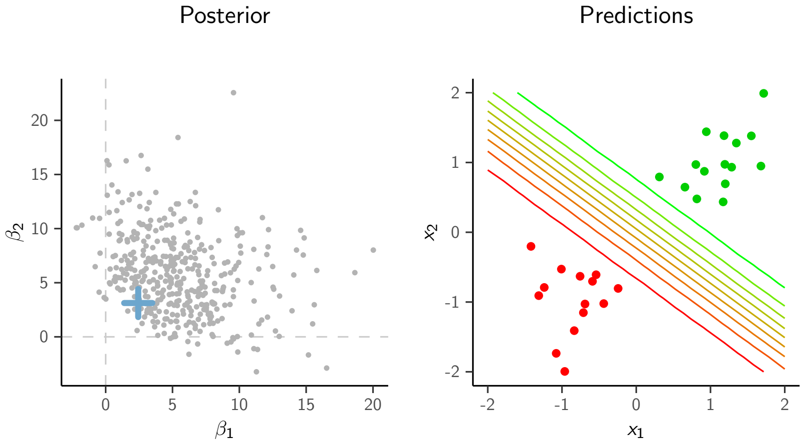
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some $\theta_0 \Rightarrow$ "Optimal point estimates"
  - Some covariates must have exactly zero regression coefficient $\Rightarrow$ "Which covariates can be discarded"
  - Much simpler model $\Rightarrow$ "Easier explanation"

Logistic regression with two covariates



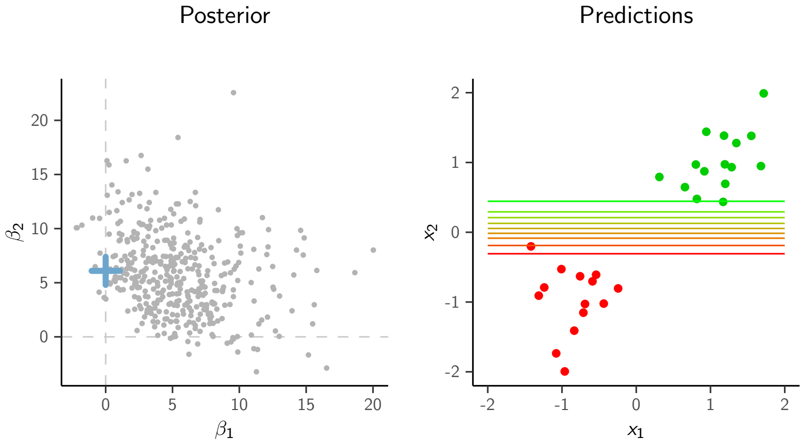
Full posterior for β_1 and β_2 and contours of predicted class probability

Logistic regression with two covariates



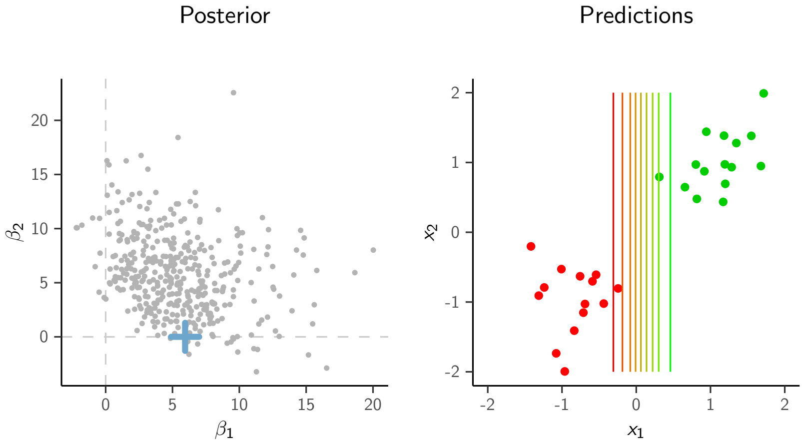
Projected point estimates for β_1 and β_2

Logistic regression with two covariates



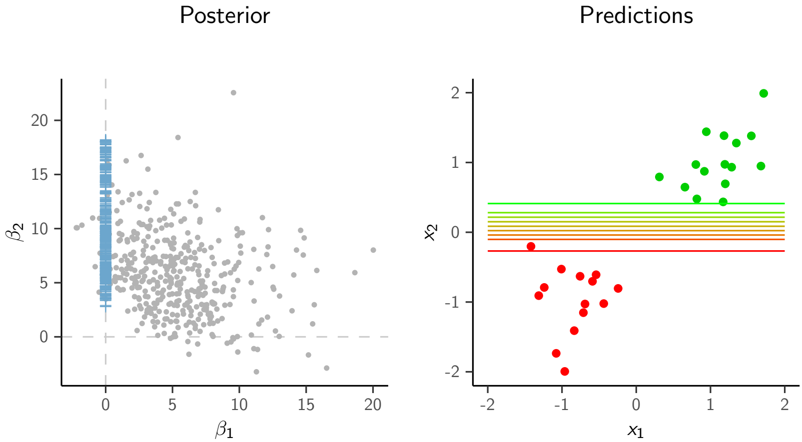
Projected point estimates, constraint $\beta_1 = 0$

Logistic regression with two covariates



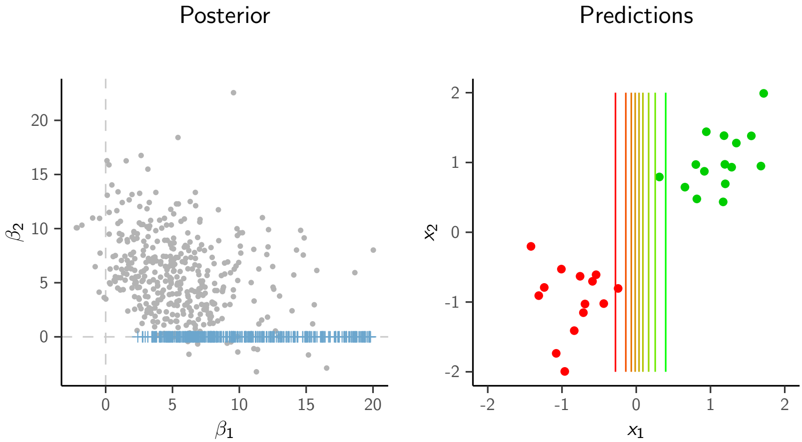
Projected point estimates, constraint $\beta_2 = 0$

Logistic regression with two covariates



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

Logistic regression with two covariates



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

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
  - solves the problem of how to do the inference after the model selection

Searching for submodels

- We may not be able to evaluate the 2^p combinations

Searching for submodels

- We may not be able to evaluate the 2^p combinations
- Search heuristics, e.g.
  - Forward search
  - L_1-penalization (as in Lasso)
  - ...

Searching for submodels

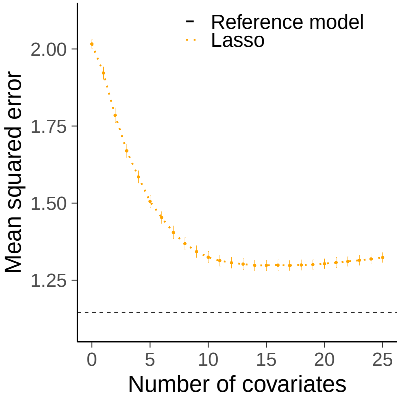
- We may not be able to evaluate the 2^p combinations
- Search heuristics, e.g.
  - Forward search
  - L_1-penalization (as in Lasso)
  - ...
- For a given model size, choose feature combination with minimal projective loss

Searching for submodels

- We may not be able to evaluate the 2^p combinations
- Search heuristics, e.g.
  - Forward search
  - L_1-penalization (as in Lasso)
  - ...
- For a given model size, choose feature combination with minimal projective loss
- Use cross-validation to select the appropriate model size
  - In some cases like, $p >> n$, we need to cross-validate over the search paths

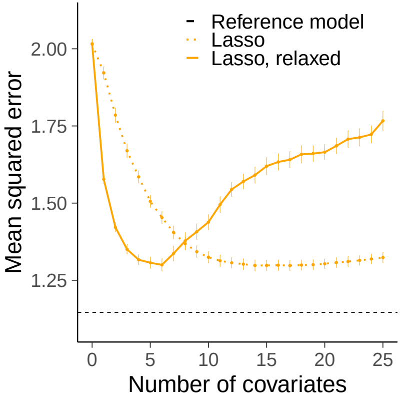
Projective selection vs. Lasso

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



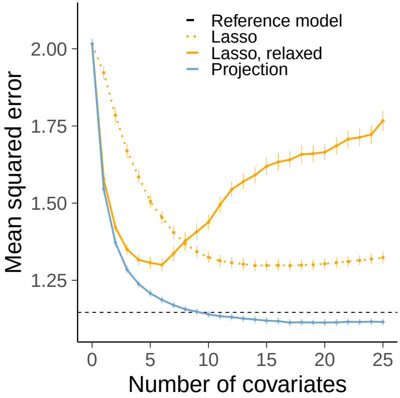
Projective selection vs. Lasso

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



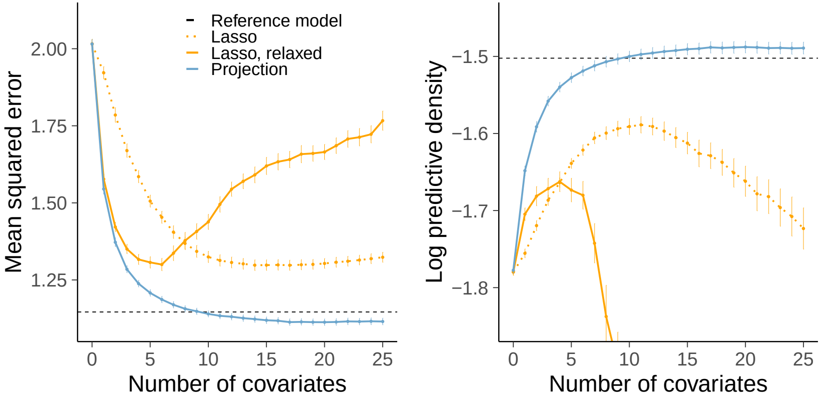
Projective selection vs. Lasso

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,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$

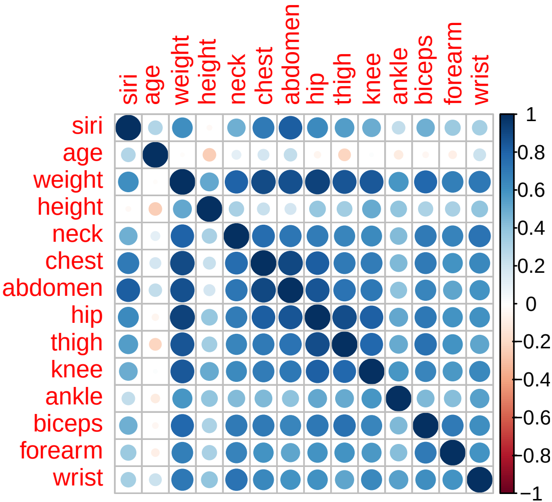


Bodyfat: small p example of projection predictive

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

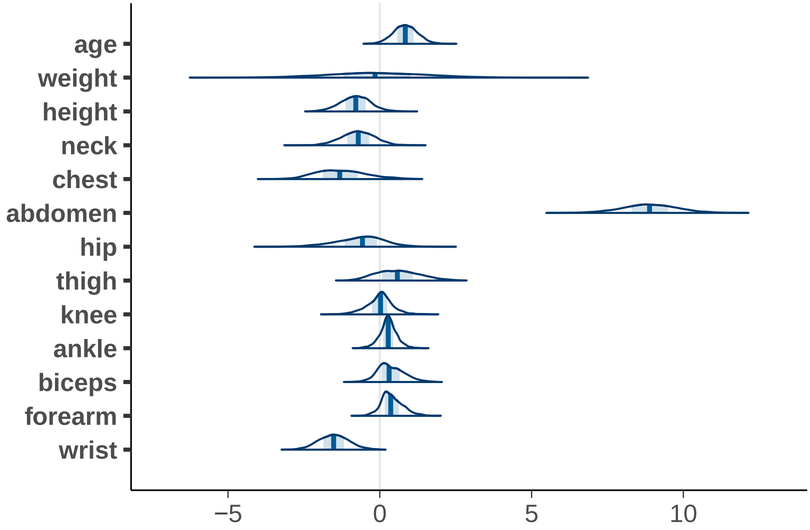
Bodyfat: small p example of projection predictive

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



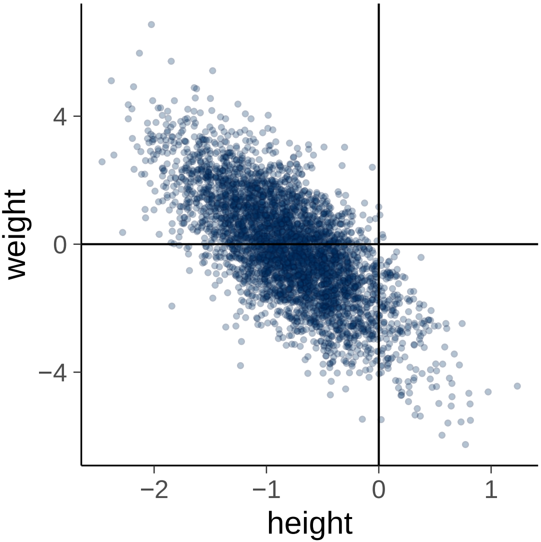
Bodyfat

Marginal posteriors of coefficients



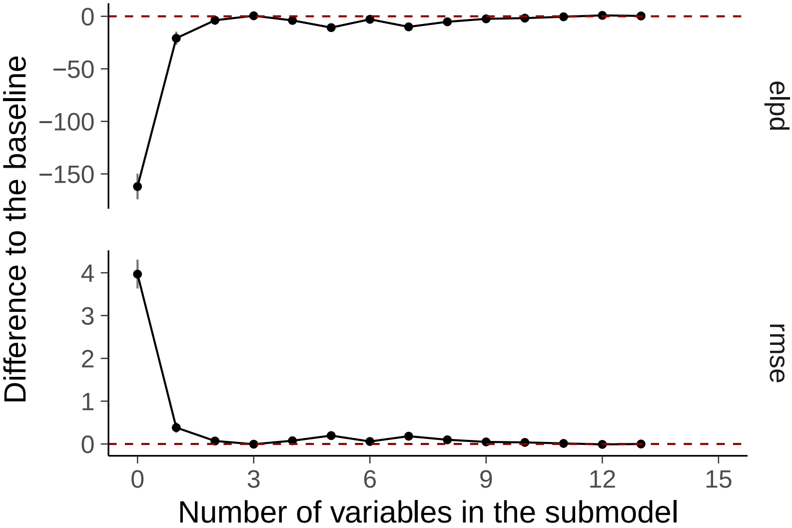
Bodyfat

Bivariate marginal of weight and height



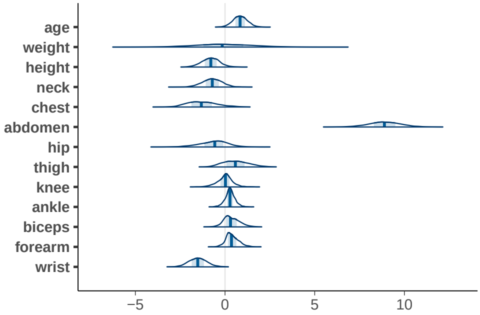
Bodyfat

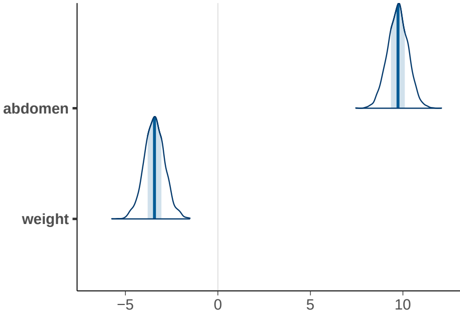
The predictive performance of the full and submodels



Bodyfat

Marginals of the reference and projected posterior





Predictive performance vs. selected variables

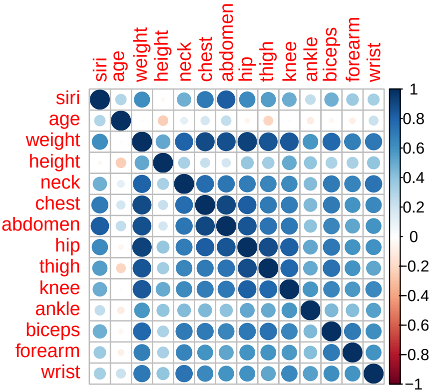
- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model

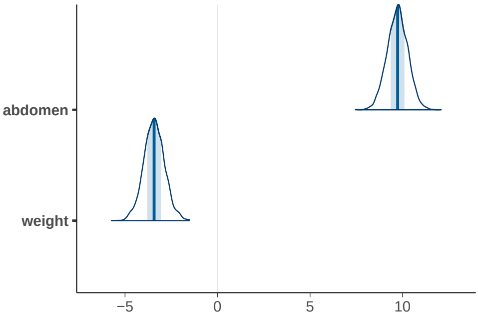
Predictive performance vs. selected variables

- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model
- Some keep asking can it find the true variables

Predictive performance vs. selected variables

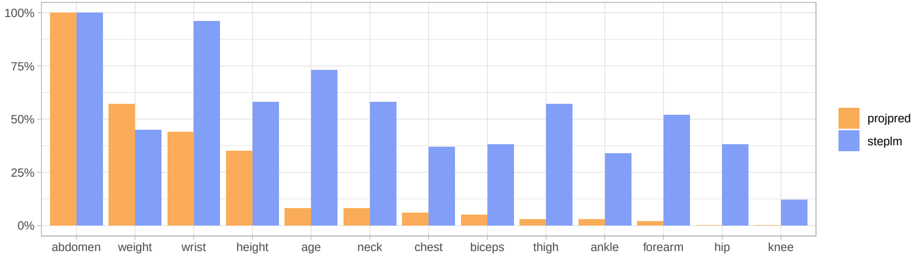
- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model
- Some keep asking can it find the true variables
  - What do you mean by true variables?





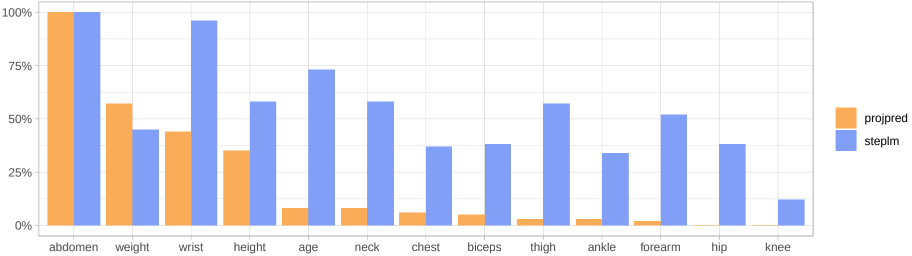
Variability under data perturbation

Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



Variability under data perturbation

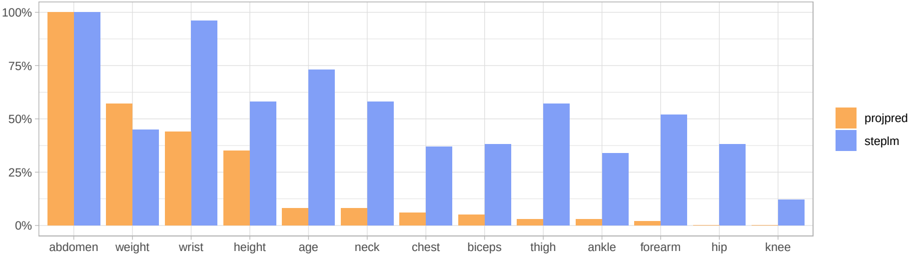
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



M	projpred	Freq %	steplm	Freq %
1	abdom., weight	39	abdom., age, forearm, height, hip, neck, thigh, wrist	4
2	abdom., wrist	10	abdom., age, chest, forearm, height, neck, thigh, wrist	4
3	abdom., height	10	abdom., forearm, height, neck, wrist	2
4	abdom., height, wrist	9	abdom., forearm, neck, weight, wrist	2
5	abdom., weight, wrist	8	abdom., age, height, hip, thigh, wrist	2
6	abdom., chest, height, wrist	2	abdom., age, height, hip, neck, thigh, wrist	2
7	abdom., biceps, weight, wrist	2	abdom., age, ankle, forearm, height, hip, neck, thigh, wrist	2
8	abdom., height, weight, wrist	2	abdom., age, biceps, chest, height, neck, wrist	2
9	abdom., age, wrist	2	abdom., age, biceps, chest, forearm, height, neck, thigh, wrist	2
10	abdom., age, height, neck, thigh, wrist	2	abdom., age, ankle, biceps, weight, wrist	2

Variability under data perturbation

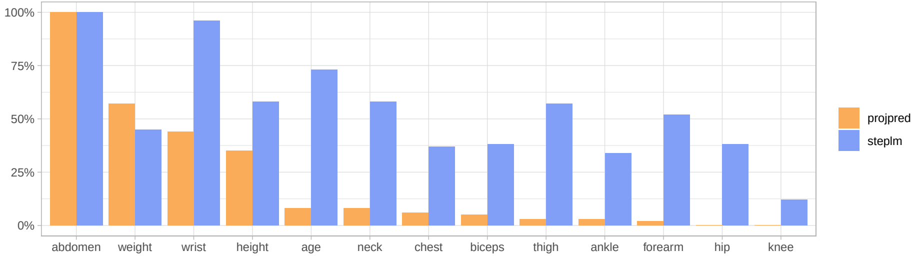
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation

Variability under data perturbation

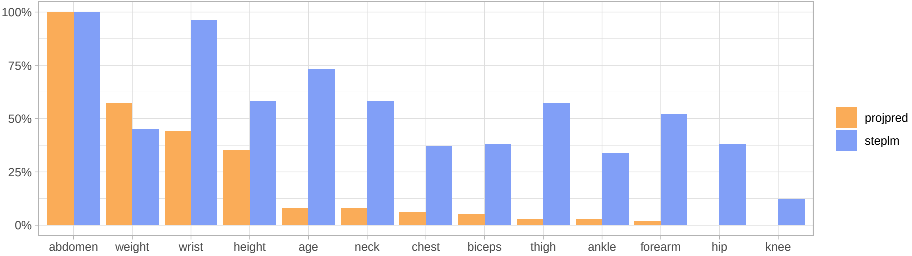
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation
- projpred uses
  - Bayesian inference for the reference
  - The reference model
  - Projection for submodel inference

Variability under data perturbation

Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation
- projpred uses
  - Bayesian inference for the reference
  - The reference model
  - Projection for submodel inference

Multilevel regression and GAMMs

- `projpred` supports also hierarchical models in `brms`
  Catalina, Bürkner, and Vehtari (2022). Projection predictive inference for generalized linear and additive multilevel models. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR* 151:4446–4461.
  https://proceedings.mlr.press/v151/catalina22a.html

Scaling

- So far the biggest number of variables we've tested is 22K
  - 96s for creating a reference model
  - 14s for projection predictive variable selection

Intro paper and brms and rstanarm + projpred examples

- McLatchie, Rögnvaldsson, Weber, and Aki Vehtari (2024). Advances in projection predictive inference. *Statistical Science*. https://arxiv.org/abs/2306.15581
- https://mc-stan.org/projpred/articles/projpred.html
- https://users.aalto.fi/~ave/casestudies.html
- Fast and often sufficient if $n \gg p$

```
varsel <- cv_varsel(fit, method='forward', cv_method='loo',
                    validate_search=FALSE)
```

- Slower but needed if not $n \gg p$

```
varsel <- cv_varsel(fit, method='forward', cv_method='kfold', K=10,
                    validate_search=TRUE)
```

- If p is very big use subsampling loo

```
# nloo should be a positive integer smaller than the number of observa
varsel <- cv_varsel(fit, cv_method='loo',
                    validate_search=TRUE, nloo=50)
```

Bayesian Python packages

- Probabilistic programming languages
  - Stan (via CmdStanPy)
  - PyMC
  - NumPyro
  - ...
- Workflow packages
  - ArviZ, MCMC diagnostics, model checking, model comparison, plotting, prior-sensitivity...
  - Bambi, BAyesian Model-Building Interface
  - Kulprit, projective inference

Download Jupyter Notebook with examples.