Rich model vs feature selection?

- If we care only about the predictive performance
  - Include all available prior information
  - Integrate over all uncertainties
  - No need for feature selection

Rich model vs feature selection?

- If we care only about the predictive performance
  - Include all available prior information
  - Integrate over all uncertainties
  - No need for feature selection
- Variable selection can be useful if
  - need to reduce measurement or computation cost in the future
  - improve explainability

Rich model vs feature selection?

- If we care only about the predictive performance
  - Include all available prior information
  - Integrate over all uncertainties
  - No need for feature selection
- Variable selection can be useful if
  - need to reduce measurement or computation cost in the future
  - improve explainability
- Two options for variable selection
  - Find a minimal subset of features that yield a good predictive model
  - Identify all features that have predictive information

Note on the terminology

- Two different problems
  - Find a minimal subset of features x_j that yield a good predictive model for y
  - Identify all features x_j that are statistically related to y

Note on the terminology

- Two different problems
  - Find a minimal subset of features x_j that yield a good predictive model for y
  - Identify all features x_j that are statistically related to y

I will focus ghere to the first case

Why shrinkage priors alone do not solve the variable selection problem

- A common strategy:
  - Fit model with a shrinkage prior
  - Select variables based on marginal posteriors (of the regression coefficients)

Why shrinkage priors alone do not solve the variable selection problem

- A common strategy:
  - Fit model with a shrinkage prior
  - Select variables based on marginal posteriors (of the regression coefficients)
- Problems
  - Marginal posteriors are difficult with correlated features
  - How to do post-selection inference correctly?

Example

Consider data

$$
\begin{aligned}
f &\sim \mathrm{N}(0, 1), \\
y \mid f &\sim \mathrm{N}(f, 1) \\
x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\, 1 - \rho), && j = 1, \ldots, 25\,, \\
x_j \mid f &\sim \mathrm{N}(0, 1), && j = 26, \ldots, 50\,.
\end{aligned}
$$

Example

Consider data

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), \\
y \mid f &\sim \mathrm{N}(f,1) \\
x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad & j = 1,\ldots,25\,, \\
x_j \mid f &\sim \mathrm{N}(0,1), & j = 26,\ldots,50\,.
\end{aligned}
$$

- y are noisy observations about latent f

Example

Consider data

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), \\
y \mid f &\sim \mathrm{N}(f,1) \\
x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), && j = 1,\ldots,25\,, \\
x_j \mid f &\sim \mathrm{N}(0,1), && j = 26,\ldots,50\,.
\end{aligned}
$$

- y are noisy observations about latent f
- First $p_{\mathrm{rel}} = 25$ features are correlated with ρ and predictive about y

Example

Consider data

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), \\
y \mid f &\sim \mathrm{N}(f,1) \\
x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), && j = 1,\ldots,25\,, \\
x_j \mid f &\sim \mathrm{N}(0,1), && j = 26,\ldots,50\,.
\end{aligned}
$$

- y are noisy observations about latent f
- First $p_{\mathrm{rel}} = 25$ features are correlated with ρ and predictive about y
- Remaining 25 features are irrelevant random noise

Example

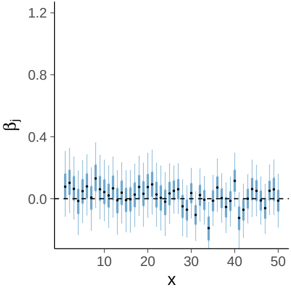
Consider data

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), \\
y \mid f &\sim \mathrm{N}(f,1) \\
x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad && j = 1,\ldots,25\,, \\
x_j \mid f &\sim \mathrm{N}(0,1), && j = 26,\ldots,50\,.
\end{aligned}
$$

- y are noisy observations about latent f
- First $p_{\mathrm{rel}} = 25$ features are correlated with ρ and predictive about y
- Remaining 25 features are irrelevant random noise

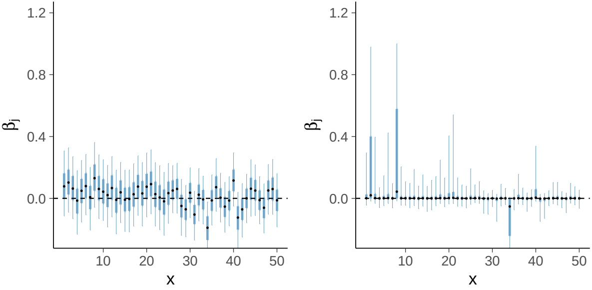
Generate one data set $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ with $n = 50$ and $\rho = 0.8$ and assess the feature relevances

Example



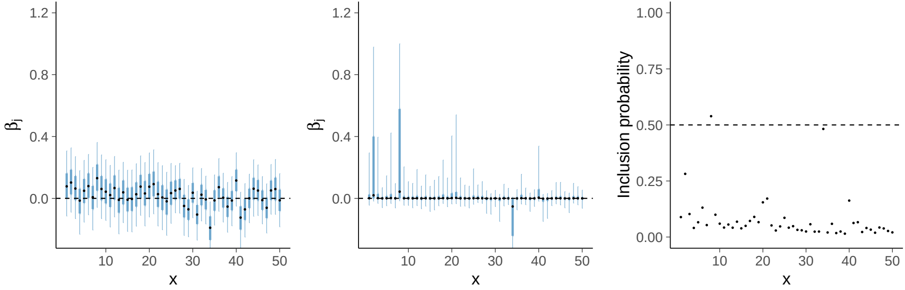
A) Gaussian prior, posterior median with 50% and 90% intervals

Example



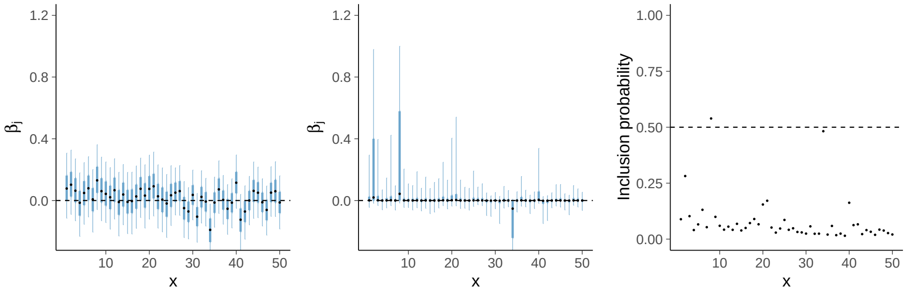
A) Gaussian prior, posterior median with 50% and 90% intervals
B) Horseshoe prior, same things

Example



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

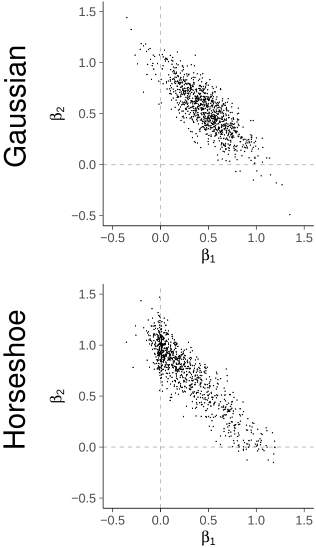
Example



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

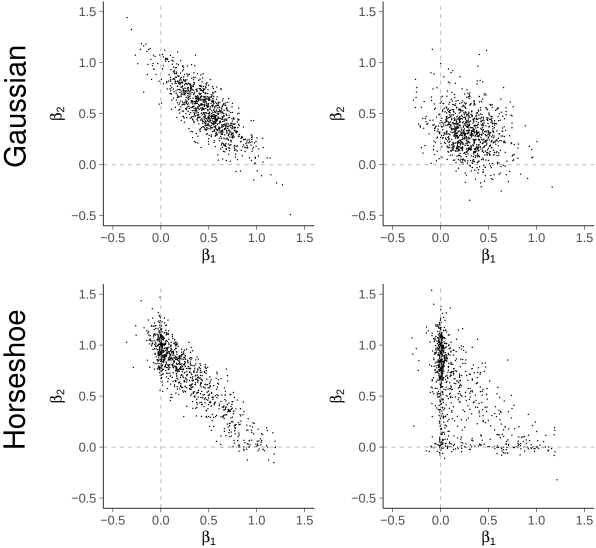
Half of the features relevant, but all marginals substantially overlapping with zero

What happens?



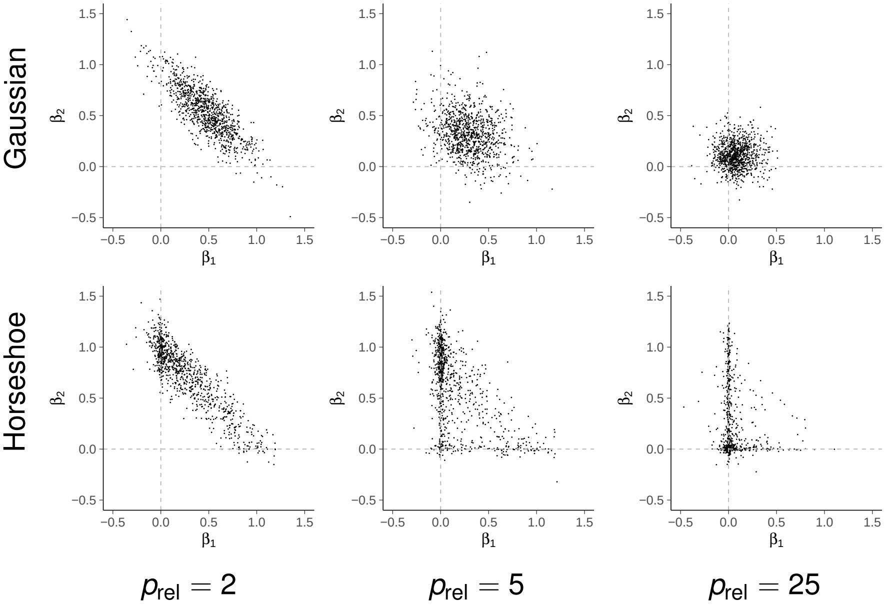
$p_{\text{rel}} = 2$

What happens?



$p_{\text{rel}} = 2$ $\qquad$ $p_{\text{rel}} = 5$

What happens?



Focus on predictive performance

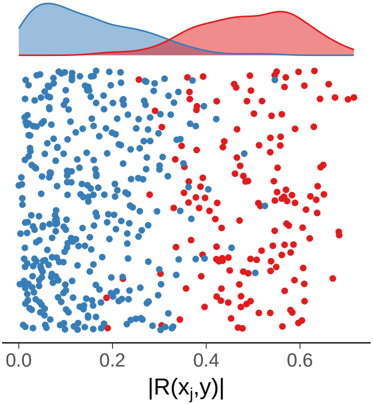
- Two stage approach
  - Construct a best predictive model you can
    $\Rightarrow$ *reference model*
  - Variable selection and post-selection inference
    $\Rightarrow$ *projection*

Focus on predictive performance

- Two stage approach
  - Construct a best predictive model you can
    $\Rightarrow$ *reference model*
  - Variable selection and post-selection inference
    $\Rightarrow$ *projection*
- Instead of looking at the marginals, find the minimal subset of features which have (almost) the same predictive performance as the reference model

Reference model improves variable selection

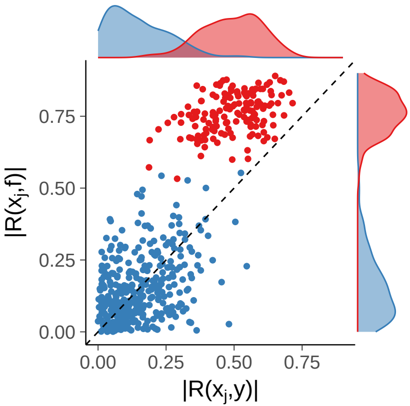
Same data generating mechanism, but
$n = 30$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$.



irrelevant x_j, relevant x_j

Sample correlation with y

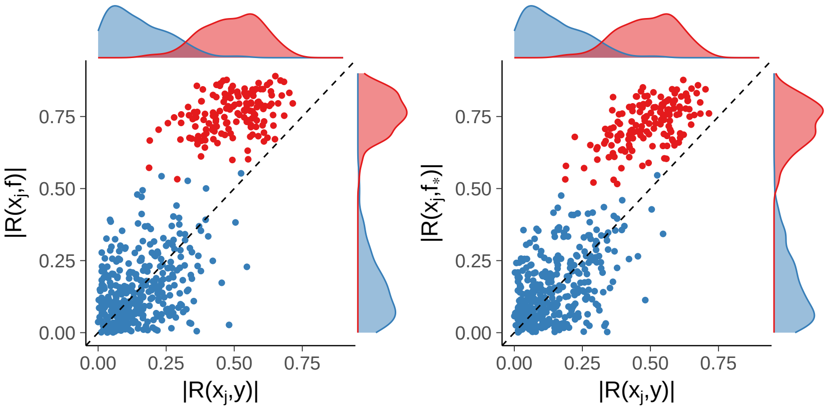
Reference model improves variable selection



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

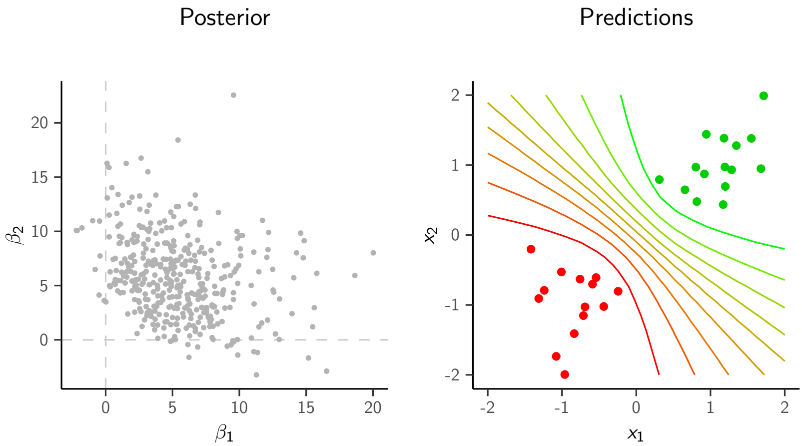
- Model simplification technique
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the predictive distribution changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some θ_0
    $\Rightarrow$ "Optimal point estimates"
  - Some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

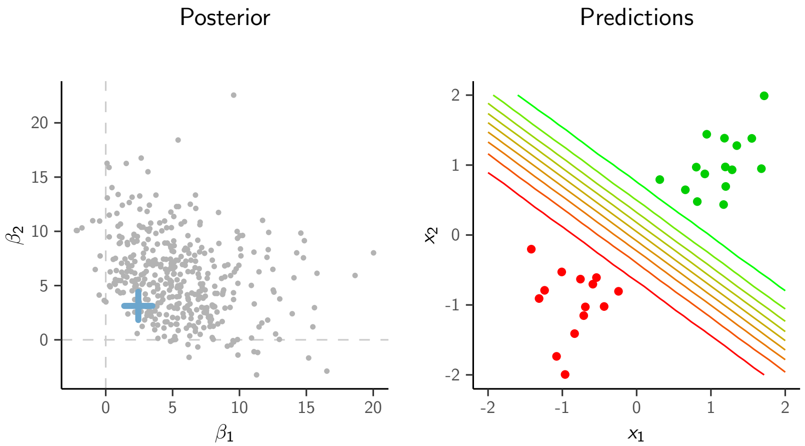
- The decision theoretic idea of conditioning the smaller model inference on the full model can be tracked to Lindley (1968)
  - draw by draw projection introduced by Goutis & Robert (1998), and Dupuis & Robert (2003)
  - see also many related references in a review by Vehtari & Ojanen (2012)

Logistic regression with two features



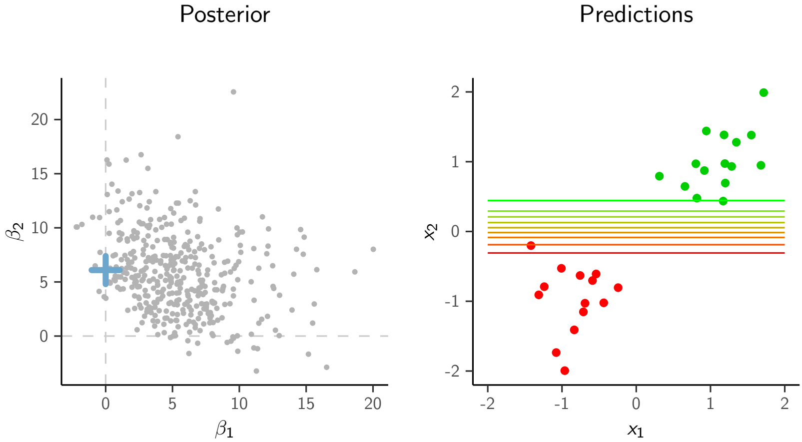
Full posterior for β_1 and β_2 and contours of predicted class probability

Logistic regression with two features



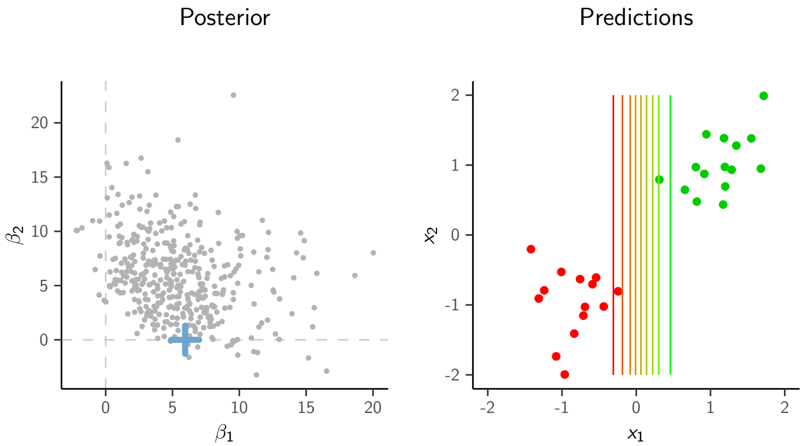
Projected point estimates for β_1 and β_2

Logistic regression with two features



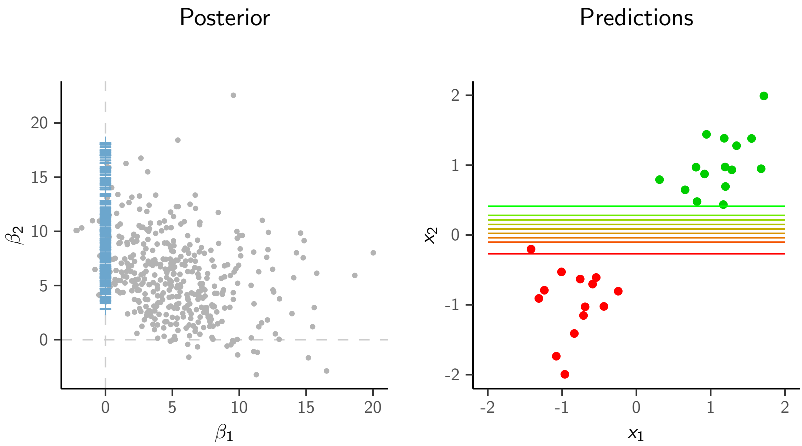
Projected point estimates, constraint $\beta_1 = 0$

Logistic regression with two features



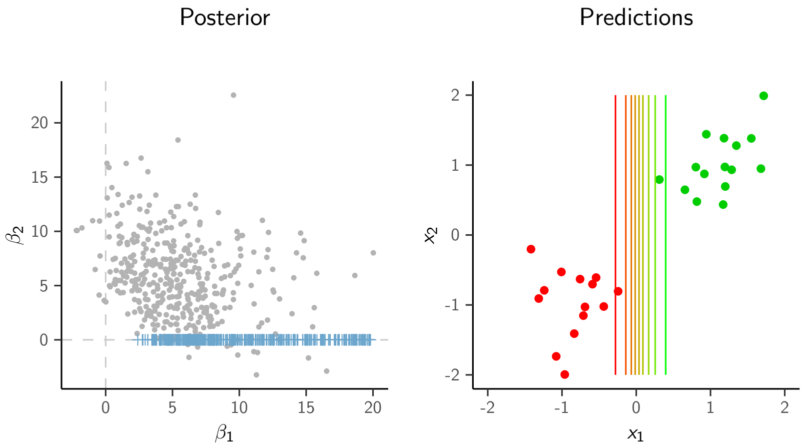
Projected point estimates, constraint $\beta_2 = 0$

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

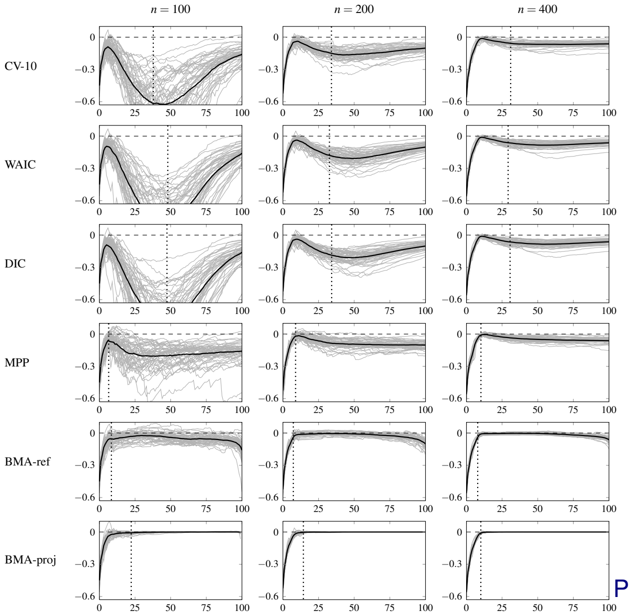
Projective selection

- 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

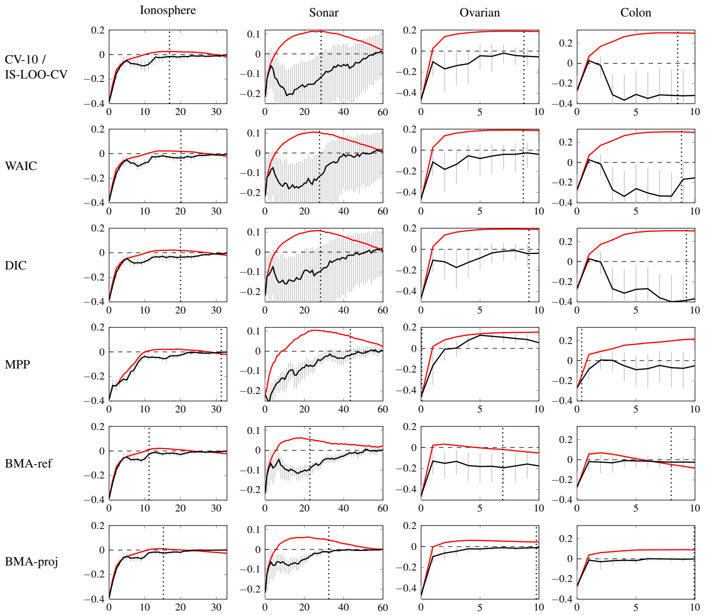
- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss
- Search heuristics, e.g.
    - Monte Carlo search
    - Forward search
    - L_1-penalization (as in Lasso)
- Use cross-validation to select the appropriate model size
    - need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

Selection induced bias in variable selection



Piironen & Vehtari (2017)

Selection induced bias in variable selection



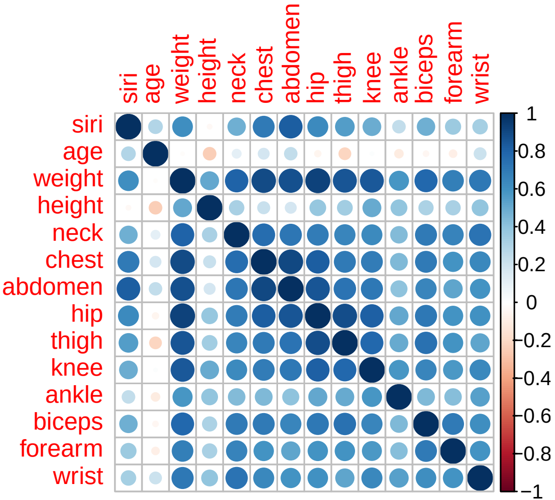
Piironen & Vehtari (2017)

Bodyfat: small p example of projection predictive

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

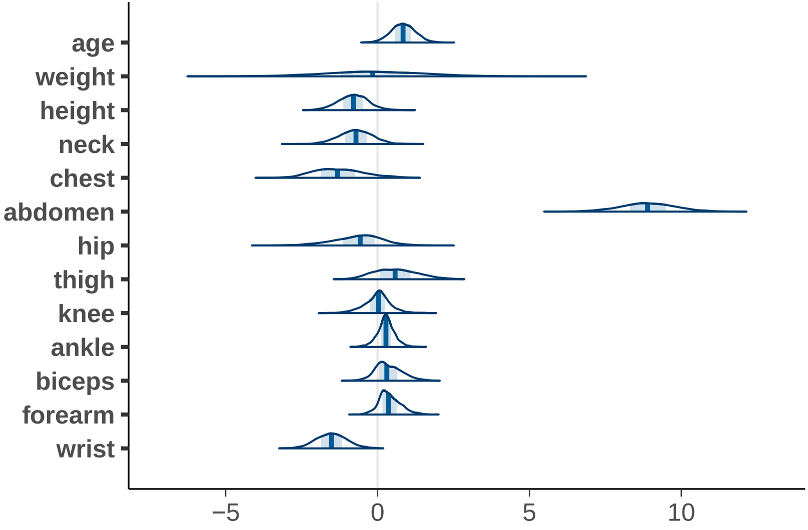
Bodyfat: small *p* example of projection predictive

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



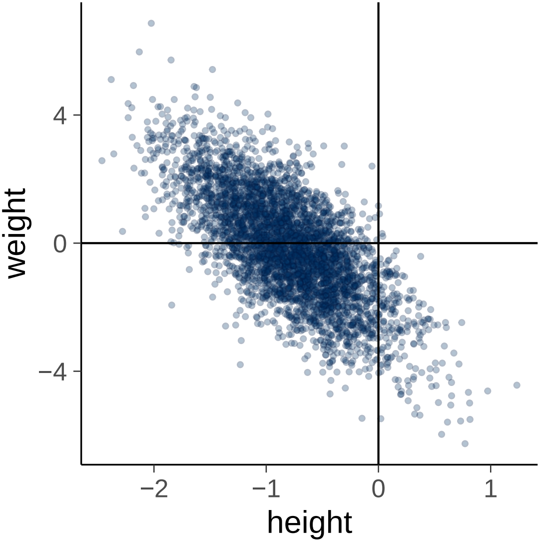
Bodyfat

Marginal posteriors of coefficients



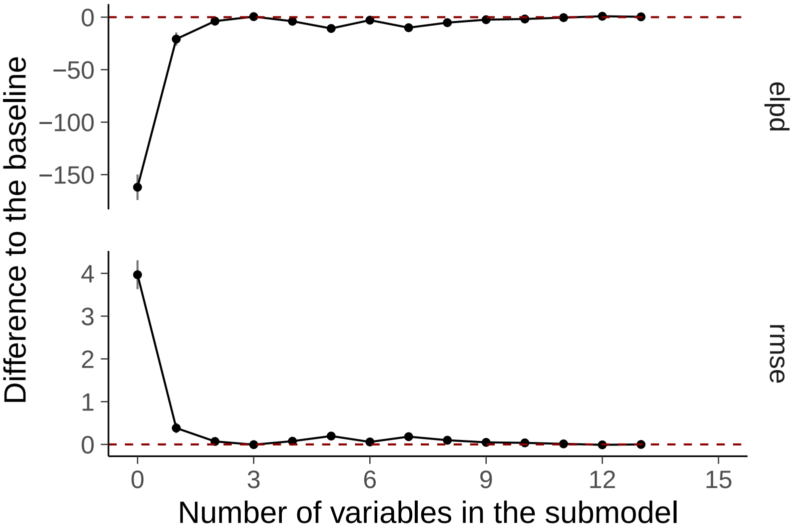
Bodyfat

Bivariate marginal of weight and height



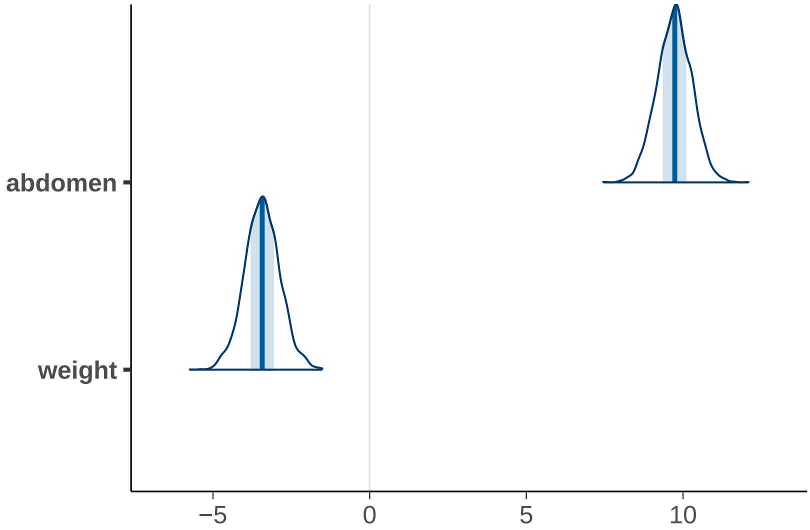
Bodyfat

The predictive performance of the full and submodels



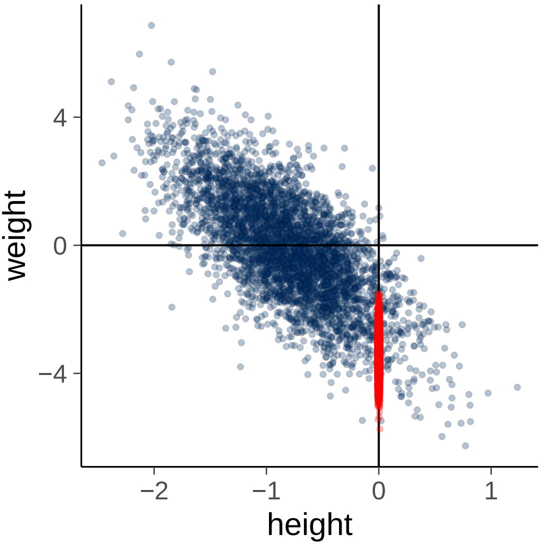
Bodyfat

Marginals of projected posterior



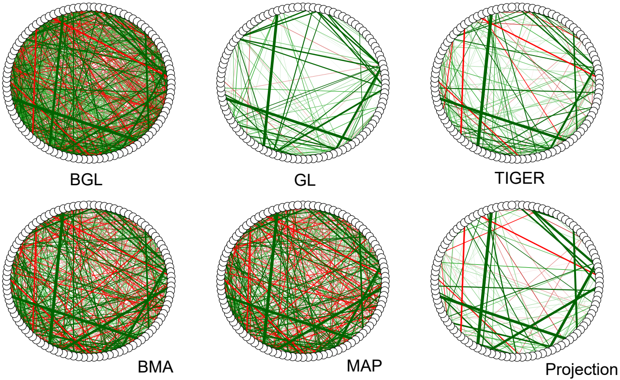
Bodyfat

Projected posterior is not just the conditional of joint



Projection of Gaussian graphical models

- Williams, Piironen, Vehtari, Rast (2018). Bayesian estimation of Gaussian graphical models with projection predictive selection. arXiv:1801.05725



CEU genetic network. BGL: Bayesian glasso; GL: glasso; TIGER: tuning insensitive graph estimation and regression; BMA: Bayesian model averaging; MAP: Maximum a posteriori; Projection: projection predictive selection.

More results

- More results projpred vs. Lasso and elastic net:
  Piironen, Paasiniemi, Vehtari (2018). Projective Inference in High-dimensional Problems: Prediction and Feature Selection. arXiv:1810.02406
- 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. doi:10.1007/s11222-016-9649-y.
- projpred for Gaussian graphical models:
  Williams, Piironen, Vehtari, Rast (2018). Bayesian estimation of Gaussian graphical models with projection predictive selection. arXiv:1801.05725
- More results for Bayes SPC:
  Piironen and Vehtari (2018). Iterative supervised principal components. 21st AISTATS, PMLR 84:106-114. Online.
- Several case studies for small to moderate dimensional ($p = 4 \ldots 100$) small data:
  Vehtari (2018). Model assesment, selection and inference after selection. https://avehtari.github.io/modelselection/

Take-home messages (part 2)

- Sparse priors do not automate variable selection
  - Don't trust marginal posteriors

Take-home messages (part 2)

- Sparse priors do not automate variable selection
  - Don't trust marginal posteriors
- Reference model + projection can improve feature selection
  - Excellent tradeoff between accuracy and model complexity
  - Useful also for identifying all the relevant features

Take-home messages (part 2)

- Sparse priors do not automate variable selection
  - Don't trust marginal posteriors
- Reference model + projection can improve feature selection
  - Excellent tradeoff between accuracy and model complexity
  - Useful also for identifying all the relevant features
- Well developed for GLMs, but can be used also with other model families

Take-home messages (part 2)

- Sparse priors do not automate variable selection
  - Don't trust marginal posteriors
- Reference model + projection can improve feature selection
  - Excellent tradeoff between accuracy and model complexity
  - Useful also for identifying all the relevant features
- Well developed for GLMs, but can be used also with other model families
- More details and results (+ some theoretical discussion) in the paper
  - Piironen, Paasiniemi, Vehtari (2018). Projective Inference in High-dimensional Problems: Prediction and Feature Selection. arXiv:1810.02406

Take-home messages (part 2)

- Sparse priors do not automate variable selection
  - Don't trust marginal posteriors
- Reference model + projection can improve feature selection
  - Excellent tradeoff between accuracy and model complexity
  - Useful also for identifying all the relevant features
- Well developed for GLMs, but can be used also with other model families
- More details and results (+ some theoretical discussion) in the paper
  - Piironen, Paasiniemi, Vehtari (2018). Projective Inference in High-dimensional Problems: Prediction and Feature Selection. arXiv:1810.02406
- R-package `projpred` in CRAN and github
  https://github.com/stan-dev/projpred
  (easy to use, e.g. with RStan, RStanARM, brms)

References

References and more at avehtari.github.io/masterclass/ and avehtari.github.io/modelselection//

- Model selection tutorial at StanCon 2018 Asilomar
  - more about projection predictive variable selection
- Regularized horseshoe talk at StanCon 2018 Asilomar
- Several case studies
- References with links to open access pdfs