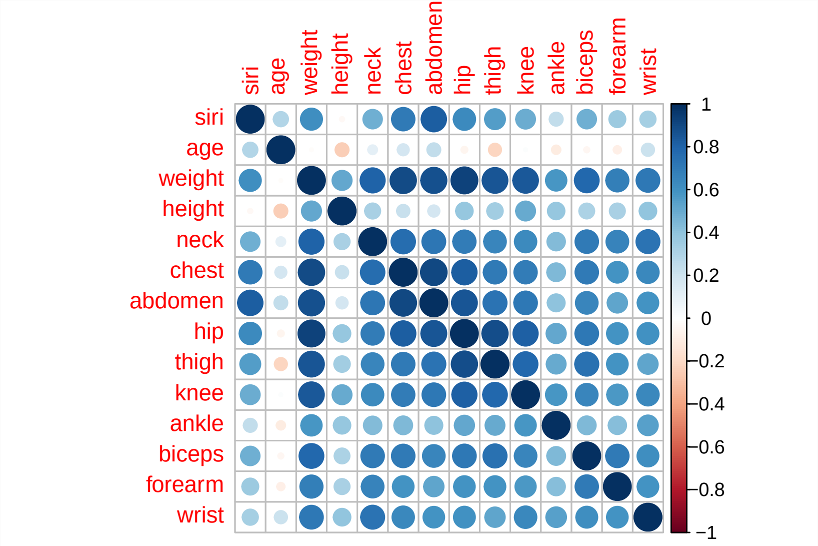
Bayesian model and variable selection using approximate cross-validation and projective predictions

Paul Bürkner

Case Study: Predictors of Body Fat



Part 1: PSIS-LOO-CV

Body Fat: Example Models

General form of likelihood:

$$\text{siri}_i \sim \text{normal}(\mu_i, \sigma)$$

Model 1:

$$\mu_i = b_0 + b_1 \text{age}_i + b_2 \text{weight}_i$$

Model 2:

$$\mu_i = b_0 + b_1 \text{age}_i + b_2 \text{weight}_i + b_3 \text{height}_i$$

Model with Age and Weight

```r
library(brms)
model1 <- brm(
  formula = siri ~ age + weight,
  data = bodyfat,
  family = gaussian()
)
```

Summary of the regression coefficients:

	Estimate	Est.Error	Q2.5	Q97.5
b_Intercept	19.08	0.37	18.37	19.79
b_age	2.52	0.39	1.77	3.26
b_weight	5.19	0.38	4.43	5.94

Model with Age, Weight, and Height

```r
model2 <- brm(
  formula = siri ~ age + weight + height,
  data = bodyfat,
  family = gaussian()
)
```

Summary of the regression coefficients:

	Estimate	Est.Error	Q2.5	Q97.5
b_Intercept	19.10	0.35	18.43	19.78
b_age	1.73	0.36	1.05	2.43
b_weight	6.90	0.40	6.10	7.70
b_height	-3.36	0.42	-4.19	-2.55

Does including ‘height’ improve model fit?

What exactly is model fit?

In-sample vs. out-of-sample fit

In-sample fit:

- How close are the model's predictions to the data it was estimated on?
- Problem: High danger of overfitting

Out-of-sample fit:

- How close are the model's predictions to new data?
- Balances under- and overfitting
- Problem: How do we evaluate predictions on new data without actual new data?

Cross-Validation

Steps in cross-validation:

(1) Split the data into two Subsets: training data and test data
(2) Fit the model on the training data
(3) Evaluate the predictions on the test data
(4) Repeat (1) to (3) with multiple data splits
(5) Summarize the results of all splits

Types of cross-validation (selection):

- Leave-one-out cross-validation (LOO-CV)
- K-fold cross-validation (K-fold-CV)
- Leave-group-out cross-validation (LGO-CV)
- Leave-future-out cross-validation (LFO-CV)

Measures of Predictive Accuracy / Utility

Example measures for a single data split:

$$\text{ELPD} = \log\ p(y|y_{\text{Tr}}) = \log \int p(y|\theta)\ p(\theta|y_{\text{Tr}})\ d\theta \approx \log\ \frac{1}{S}\sum_{s=1}^{S} p(y|\theta^{(s)})$$

$$\text{RMSE} = \sqrt{\int (y-\hat{y})^2\ p(\hat{y}|y_{\text{Tr}})\ d\hat{y}} \approx \sqrt{\frac{1}{S}\sum_{s=1}^{S}(y-\hat{y}^{(s)})^2}$$

$$\text{MAE} = \int |y-\hat{y}|\ p(\hat{y}|y_{\text{Tr}})\ d\hat{y} = \frac{1}{S}\sum_{s=1}^{S}|y-\hat{y}^{(s)}|$$

Leave-One-Out Cross-Validation

Leave out a single observation y_i and predict by all other observations y_{-i} using the ELPD:

$$\text{ELPD} = \sum_{i=1}^{N} \log\ p(y_i | y_{-i})$$

(other measures are possible as well)

Important properties of LOO-CV:

- All possible N splits can be evaluated
- Can be approximated using the full model

Importance Sampling

Approximate expectations over a target distribution $f(\theta)$ using an approximating proposal distribution $g(\theta)$:

$$\mathbb{E}_f[h(\theta)] = \int h(\theta) f(\theta)\, d\theta = \frac{\int h(\theta) f(\theta)\, d\theta}{\int f(\theta)\, d\theta} = \frac{\int h(\theta) r(\theta) g(\theta)\, d\theta}{\int r(\theta) g(\theta)\, d\theta}$$

Raw importance ratios:

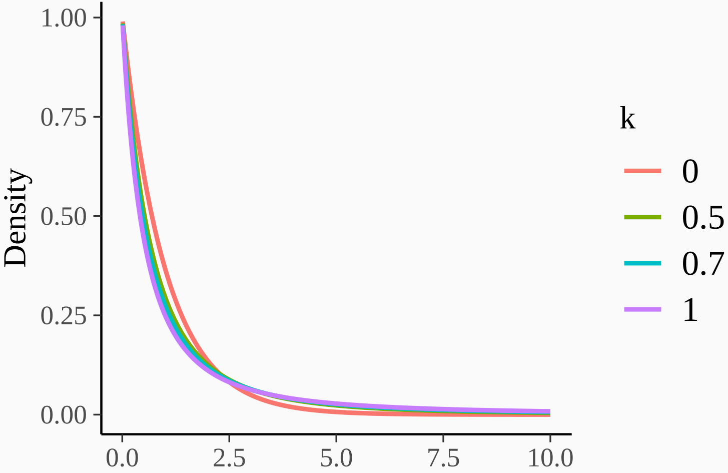
$$r(\theta) = \frac{f(\theta)}{g(\theta)}$$

Approximation via $\theta^{(s)} \sim g(\theta)$:

$$\mathbb{E}_f[h(\theta)] \approx \frac{\sum_{s=1}^{S} h(\theta^{(s)}) r(\theta^{(s)})}{\sum_{s=1}^{S} r(\theta^{(s)})}$$

Pareto Smoothed Importance Sampling (PSIS)

Replace the largest importance ratios with quantiles of the generalized Pareto distribution (GPD)



The $\hat{k}$-Diagnostic

The number of existing moments of the GPD is

$$\#\text{moments} = \begin{cases} \text{if } k > 0: \ \text{floor}\left(\frac{1}{k}\right) \\ \text{else:} \ \infty \end{cases}$$

Relevant thresholds:

- $k < 0.5$: Finite variance and fast convergence rate
- $0.5 \leq k \leq 0.7$: Convergence rate is still ok
- $k > 0.7$: Preasymptotic behavior gets in your way
- $k > 1$: All is lost

PSIS-LOO-CV

Compute the raw LOO importance ratios:

$$r_i^{(s)} = \frac{f_i(\theta^{(s)})}{g(\theta^{(s)})} \propto \frac{1}{p(y_i \mid \theta^{(s)})}$$

Obtain smoothed importance weights $w_i^{(s)}$ via PSIS

Approximate the *i*th posterior preditive density (PPD):

$$p(y_i \,|\, y_{-i}) \approx \frac{\sum_{s=1}^{S} w_i^{(s)} \, p(y_i \,|\, \theta^{(s)})}{\sum_{s=1}^{S} w_i^{(s)}}$$

Sum over the log pointwise contributions:

$$\text{ELPD} = \sum_{i=1}^{N} \log \; p(y_i | y_{-i})$$

Body Fat: PSIS-LOO-CV for Model 1

```r
loo1 <- loo(model1)
print(loo1)
```

```
##
## Computed from 4000 by 251 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -810.5 11.1
## p_loo         3.7  0.5
## looic      1621.0 22.2
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

Body Fat: PSIS-LOO-CV for Model 2

```r
loo2 <- loo(model2)
print(loo2)
```

```
##
## Computed from 4000 by 251 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -781.7  9.4
## p_loo         4.8  0.5
## looic      1563.4 18.7
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

Body Fat: PSIS-LOO-CV Model Comparison

```
loo_compare(loo1, loo2)
```

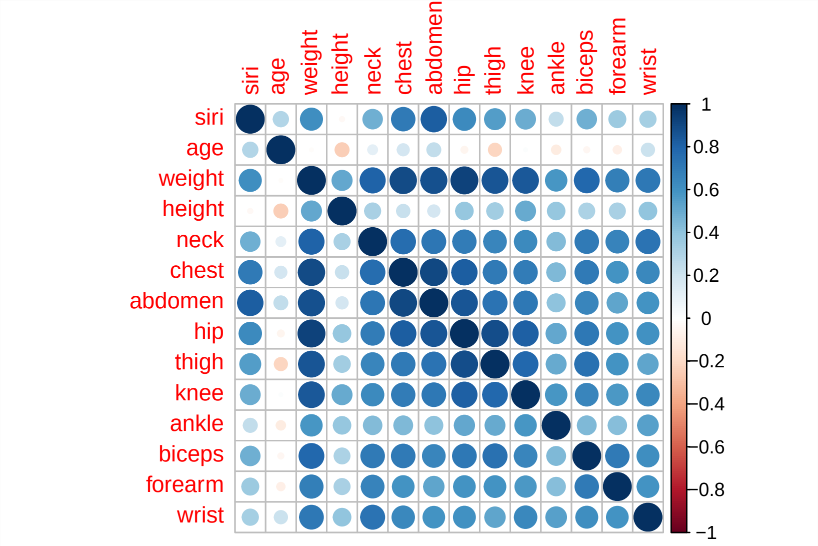
```
##        elpd_diff se_diff
## model2   0.0       0.0
## model1 -28.8       8.3
```

More detailed summary available via

```
print(loo_compare(loo1, loo2), simplify = FALSE)
```

Part 2: Projpred

Body Fat: Feature Selection



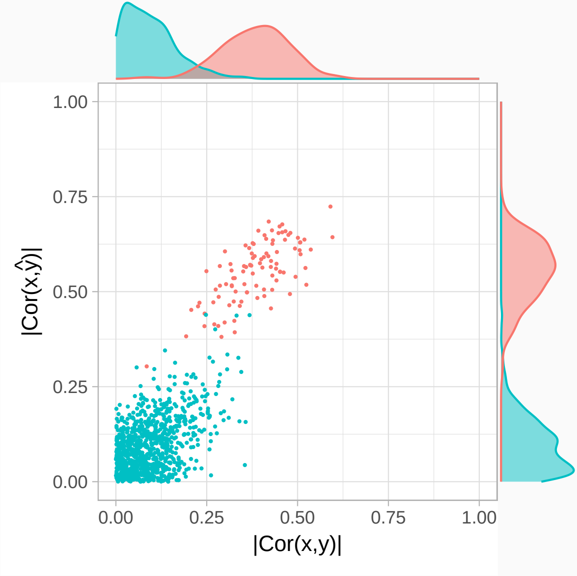
The Projpred approach for variable selection

Goal: Select a minimally sufficient set of predictors/features

Relevant aspects:

- What is the reference to compare to?
- How do we compare a model to this reference?
- How to incorporate uncertainty correctly?
- How to do all of this efficienctly?

Reference Models



Kullback–Leibler Divergence

The KL divergence measures how much one distribution q differs from another distribution p:

$$KL(p \,||\, q) = \int \log\left(\frac{p(x)}{q(x)}\right)\, p(x)\, dx$$

Application in projective predictions:

$$\text{Maximize} \quad KL(p(\hat{y}|y)||q(\hat{y}|y))$$

- $p(\hat{y}|y)$: PPD of the reference model
- $q(\hat{y}|y)$: PPD of a sub model

Incorporating Uncertainty

For each posterior draw $\theta_p^{(s)}$ from p, find $\theta_q^{(s)}$ that maximizes

$$KL(p(\hat{y}|\theta_p^{(s)}) \,||\, q(\hat{y}|\theta_q^{(s)}))$$

Easy to compute for generalized linear models (GLMs):

- Replace the actual responses by the reference predictions
- Perform maximum likelihood estimation

We can further improve efficiency by clustering posterior draws

Ongoing research: Extend projpred to more complex models

Feature Inclusion and Decision Strategies

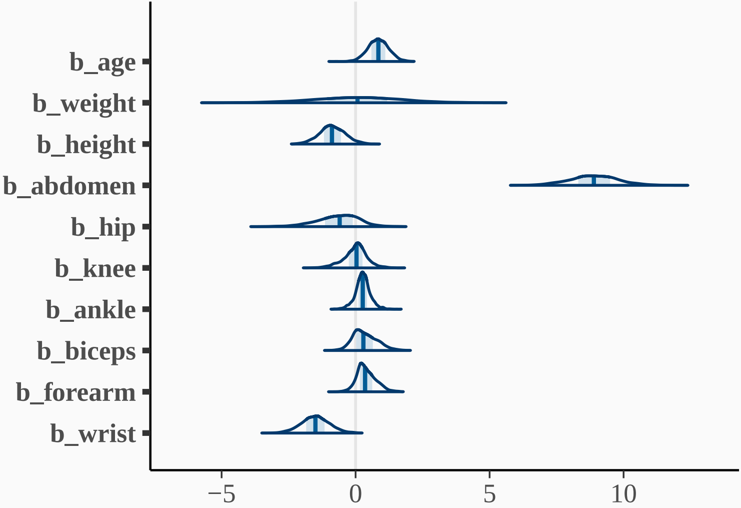
Feature inclusion strategies:

- Check all possible sub models: $\#\text{models} = 2^K$
- Forward stepwise inclusion: $\#\text{models} = \frac{K(K+1)}{2}$
- Penalized regression such as Lasso or Elastic Net: $\#\text{models} = K$

Decision strategies:

- Choose a measure of predictive accuracy u
- Choose a cross-validation procedure
- Order promising sub models according to their complexity
- Compute u_q for a sub model
- Compare u_q to the u_p of the reference model
- Stop once u_q of the current sub model is close enough to u_p

Body Fat: Fitting the Reference Model



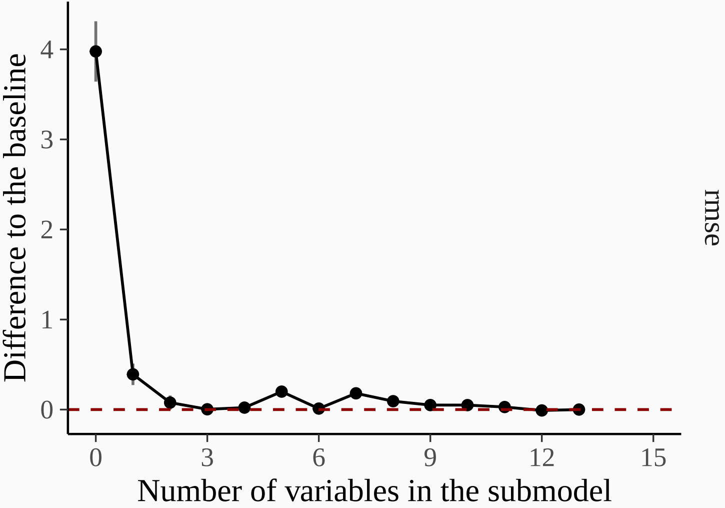
Running projpred in R

```r
library(projpred)
cvvs <- cv_varsel(
  fit_ref, method = 'forward', cv_method = 'LOO',
  nloo = N, verbose = FALSE
)
```

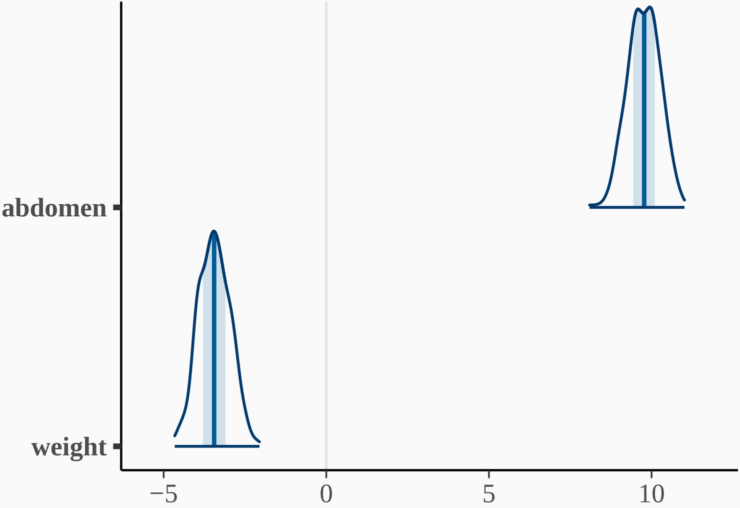
Summarize the results

	size	solution_terms	elpd	elpd.se
2	0	NA	-888.2021	10.341620
3	1	abdomen	-747.0721	9.144744
4	2	weight	-729.9316	8.879844
5	3	wrist	-725.6908	8.791947
6	4	height	-726.6267	8.908563
7	5	chest	-736.6859	9.270678
8	6	age	-725.9468	8.933133
9	7	biceps	-735.7083	9.356540
10	8	neck	-730.6905	9.237834
11	9	forearm	-728.2471	9.276072
12	10	ankle	-728.2443	9.278398

Summarize the results



Summarize the results



References

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1). 1–8.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2019). Pareto smoothed importance sampling. *arXiv preprint*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.

Piironen, J., Paasiniemi, M., & Vehtari, A. (2018). Projective inference in high-dimensional problems: prediction and feature selection. *arXiv preprint*.

Catalina A., Bürkner P. C., & Vehtari A. (2020). Projection Predictive Inference for Generalized Linear and Additive Multilevel Models. *arXiv preprint*.