Cost effective prediction of bodyfat

An example of project presentation slides

Aki Vehtari
Aalto University

Cost effective prediction of bodyfat

An example of project presentation slides

Aki Vehtari
Aalto University

Introduce yourself

Measuring bodyfat percentage

- Bodyfat percentage is related to many health outcomes

[Nice figures here]

Measuring bodyfat percentage

- Bodyfat percentage is related to many health outcomes
- Relatively accurate way to measure bodyfat is to weight a person in air and immersed in water
  - proportion of body fat can be derived from body density with Siri's (1956) formula
  - water immersion requires a big tub for the water and harness system for lowering a person to water

[Nice figures here]

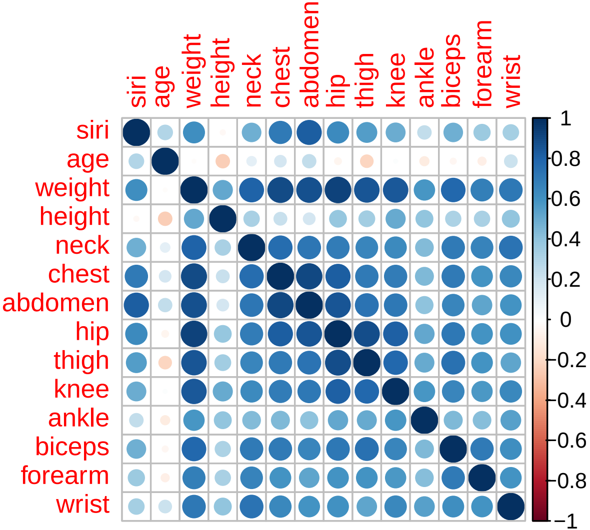
Measuring bodyfat percentage

- Bodyfat percentage is related to many health outcomes
- Relatively accurate way to measure bodyfat is to weight a person in air and immersed in water
  - proportion of body fat can be derived from body density with Siri's (1956) formula
  - water immersion requires a big tub for the water and harness system for lowering a person to water
- Can we estimate the bodyfat percentage with faster and a smaller equipment?
  - with just a scale and measure tape?
  - 252 subjects

[Nice figures here]

Measuring bodyfat percentage

- With just a scale and measure tape?



Bodyfat predictive model

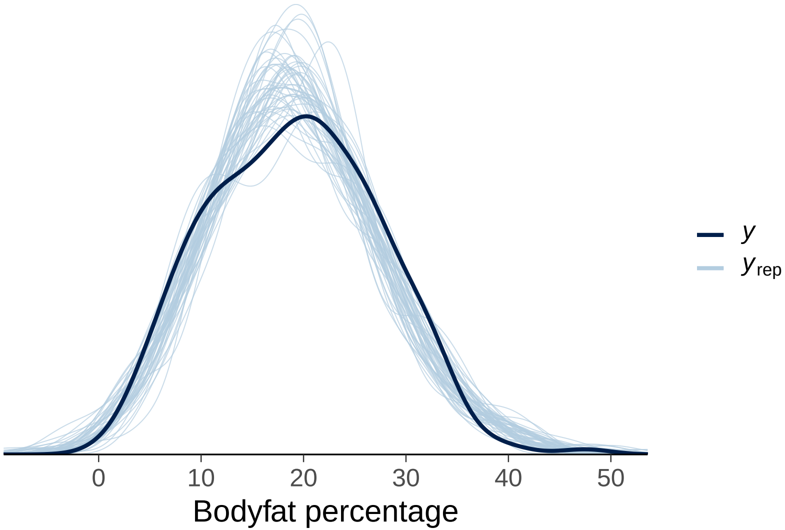
- Gaussian linear regression model with normal vs. regularized horseshoe prior ($p_0 = 5$) on coefficients

Bodyfat predictive model

- Gaussian linear regression model with normal vs. regularized horseshoe prior ($p_0 = 5$) on coefficients
- Model build with `rstanarm` and inference run with Stan
  - all convergence diagnostics were good

Bodyfat model checking

Posterior predictive checking



Bodyfat model comparison

- Leave-one-out cross-validation comparison
  - no difference

	elpd_diff	se_diff
RHS prior	0.0	0.0
Gaussian prior	−1.1	2.2

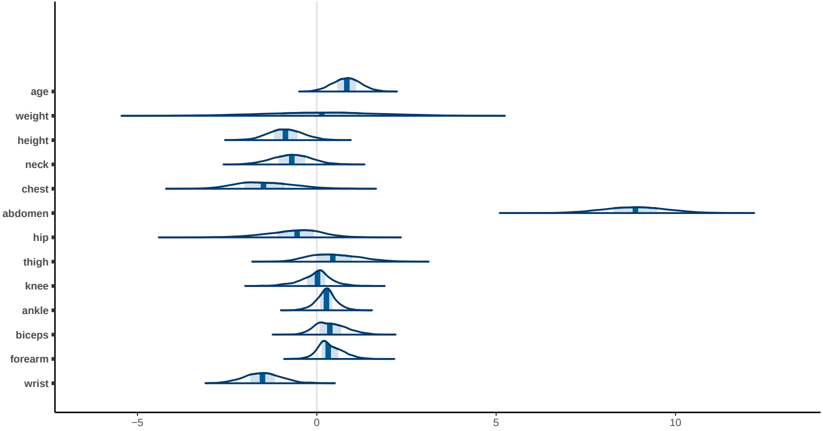
Bodyfat model comparison

- Leave-one-out cross-validation comparison
  - no difference

```
                  elpd_diff  se_diff
RHS prior            0.0       0.0
Gaussian prior      -1.1       2.2

Computed from 4000 by 250 log-likelihood matrix

          Estimate   SE
elpd_loo    -723.9  9.4
p_loo         13.4  1.2
looic       1447.9 18.8
------
Monte Carlo SE of elpd_loo is 0.1.

Pareto k diagnostic values:
                         Count  Pct.    Min. n_eff
(-Inf, 0.5]   (good)     249    99.6%   1374
 (0.5, 0.7]   (ok)         1     0.4%   724
   (0.7, 1]   (bad)        0     0.0%   <NA>
   (1, Inf)   (very bad)   0     0.0%   <NA>
```

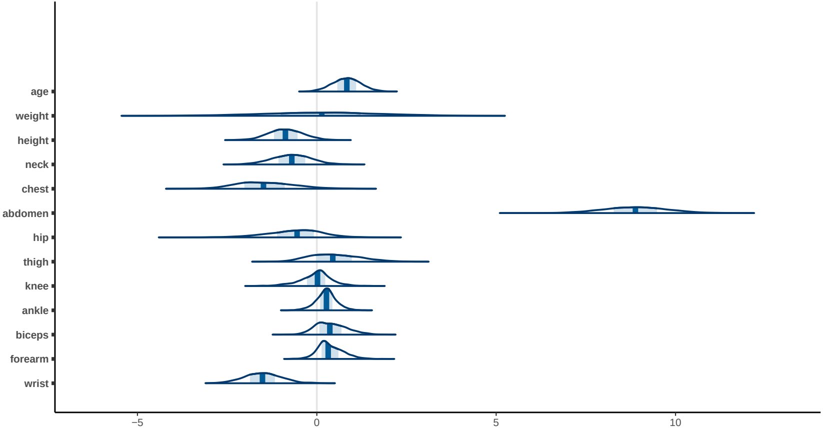
Bodyfat

Marginal posteriors of coefficients



Bodyfat

Check that the font in all figures is big enough!



Bodyfat

Marginal posteriors of coefficients (Much better!)

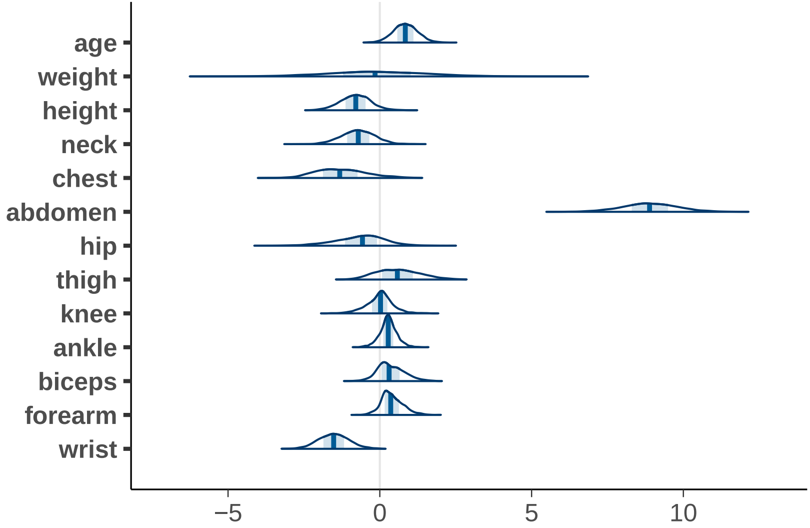


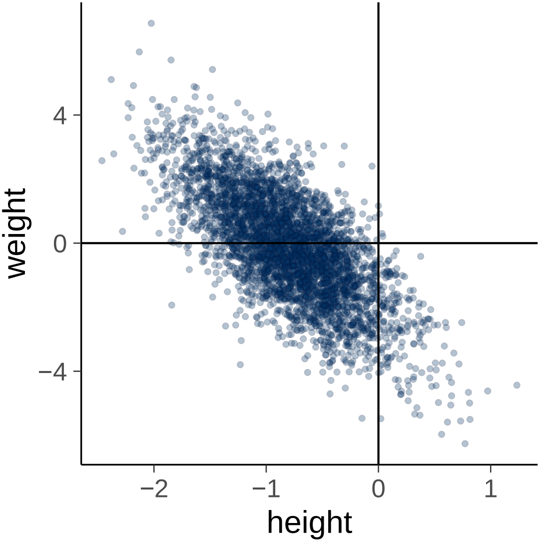
Figure font size

For example:

```
theme_set(bayesplot::theme_default(base_family = "sans",
                                   base_size=16))
```

Bodyfat

Bivariate marginal of weight and height



Bodyfat variable selection

- Do we need all the measurements?
- We find the model with a minimal set of variables which have similar predictive performance as the model with all variables

Bodyfat variable selection

- Do we need all the measurements?
- We find the model with a minimal set of variables which have similar predictive performance as the model with all variables
- We use projection predictive variable selection implemented in `projpred` package

Projective predictive covariate selection

- The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y}|D) = \int p(\tilde{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta},$$

  where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$) and $\boldsymbol{\beta}$ is in general non-sparse (all $\boldsymbol{\beta}_j \neq 0$)
- What is the best distribution $q_\perp(\boldsymbol{\theta})$ given a constraint that only selected covariates have nonzero coefficient
- Optimization problem:

$$q_\perp = \arg\min_q \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y}_i \mid D) \parallel \int p(\tilde{y}_i \mid \boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}\right)$$

- Optimal projection from the full posterior to a sparse posterior (with minimal predictive loss)

For 10min presentation, too much information

- The full model predictive distribution represents our best knowledge about future $\tilde{y}$

$$p(\tilde{y}|D) = \int p(\tilde{y}|\theta)p(\theta|D)d\theta,$$

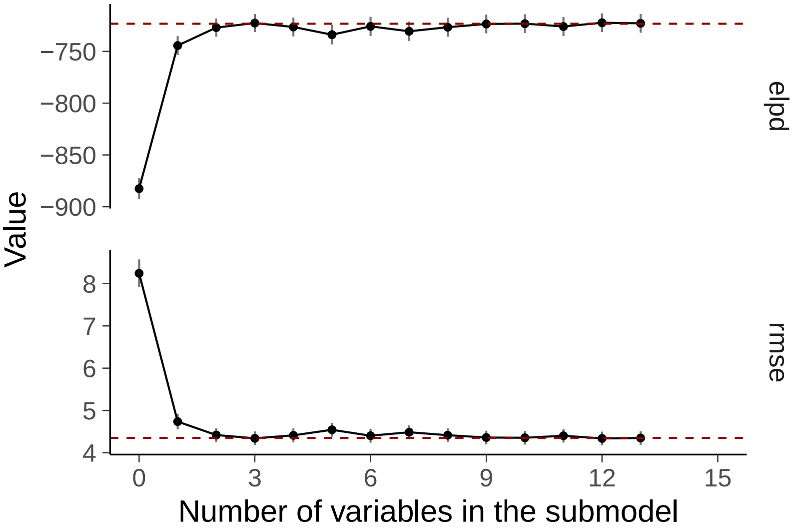
  where $\theta = (\beta, \sigma^2)$) and β is in general non-sparse (all $\beta_j \neq 0$)
- What is the best distribution $q_\perp(\theta)$ given a constraint that only selected covariates have nonzero coefficient
- Optimization problem:

$$q_\perp = \arg\min_q \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y}_i \mid D) \parallel \int p(\tilde{y}_i \mid \theta)q(\theta)d\theta\right)$$

- Optimal projection from the full posterior to a sparse posterior (with minimal predictive loss)

Bodyfat

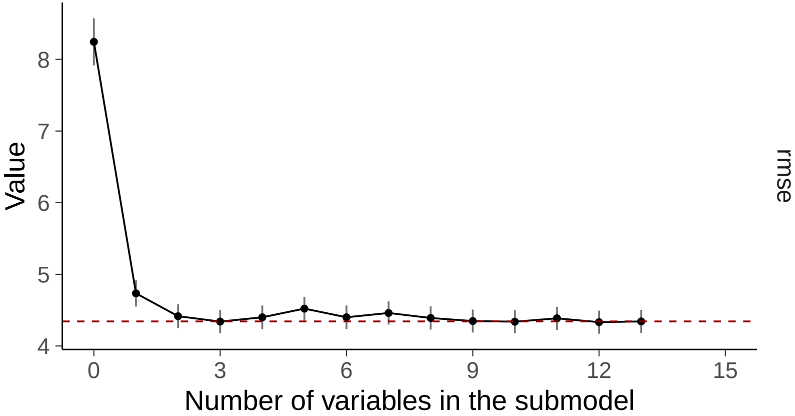
The predictive performance of the full and submodels



Bodyfat

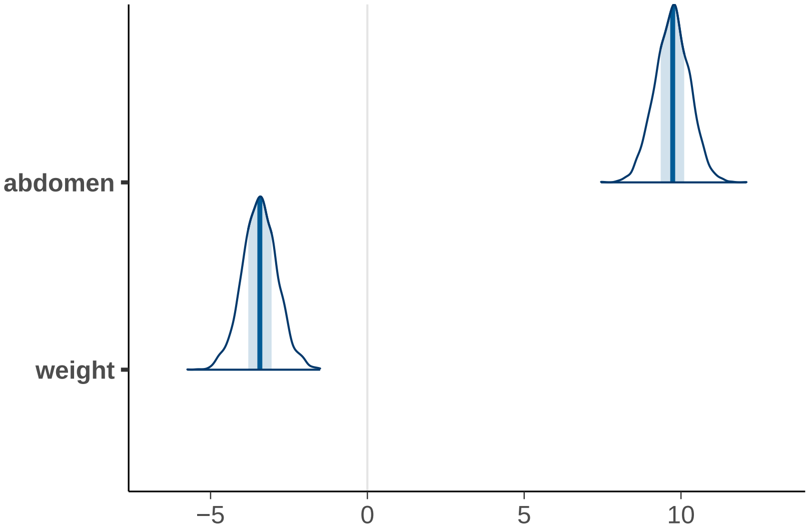
The predictive performance of the full and submodels

One of these plots is probably sufficient



Bodyfat

Marginals of projected posterior



Bodyfat – Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure

Bodyfat – Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure
- The accuracy using mean of data is 16%-units (95% interval)

Bodyfat – Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure
- The accuracy using mean of data is 16%-units (95% interval)
- The accuracy using all anthropometric measures is 8.6%-units (95% interval)

Bodyfat – Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure
- The accuracy using mean of data is 16%-units (95% interval)
- The accuracy using all anthropometric measures is 8.6%-units (95% interval)
- The same accuracy can be obtained using just abdomen circumference and weight

Bodyfat – Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure
- The accuracy using mean of data is 16%-units (95% interval)
- The accuracy using all anthropometric measures is 8.6%-units (95% interval)
- The same accuracy can be obtained using just abdomen circumference and weight
- More results at avehtari.github.io/modelselection/bodyfat.html

THANKS!

NO “THANKS”!

NO “THANKS”!

- Don’t ever end with a slide having just “THANKS”

NO “THANKS”!

- Don’t ever end with a slide having just “THANKS”
- “THANKS” slide has zero information content

NO “THANKS”!

- Don’t ever end with a slide having just “THANKS”
- “THANKS” slide has zero information content
- Leave the conclusion slide or contact information slide

Conclusion

- Bodyfat percentage estimated using water immersion can be predicted using scale and tape measure
- The accuracy using mean of data is 16%-units (95% interval)
- The accuracy using all anthropometric measures is 8.6%-units (95% interval)
- The same accuracy can be obtained using just abdomen circumference and weight
- More results at avehtari.github.io/modelselection/bodyfat.html

Additional information

- You can have additional slides after the conclusion for supporting material to answer questions
  - for example, in this course, include Stan code and additional convergence and model checking results

Gaussian linear model with regularized horseshoe prior

```
// generated with brms 2.14.4
functions {
  vector horseshoe(vector z, vector lambda, real tau, real c2) {
    int K = rows(z);
    vector[K] lambda2 = square(lambda);
    vector[K] lambda_tilde = sqrt(c2 * lambda2 ./ (c2 + tau^2 * lambda2));
    return z .* lambda_tilde * tau;
  }
}
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  // data for the horseshoe prior
  real<lower=0> hs_df;  // local degrees of freedom
  real<lower=0> hs_df_global;  // global degrees of freedom
  real<lower=0> hs_df_slab;  // slab degrees of freedom
  real<lower=0> hs_scale_global;  // global prior scale
  real<lower=0> hs_scale_slab;  // slab prior scale
  int prior_only;  // should the likelihood be ignored?
}
transformed data {
```

Classification example: Pima Indians Diabetes

Predict diabetes based on

- Pregnancies
- Glucose
- Blood pressure
- Skin thickness
- Insulin
- BMI
- Diabetes Pedigree
- Age

768 observations

https://avehtari.github.io/modelselection/diabetes.html

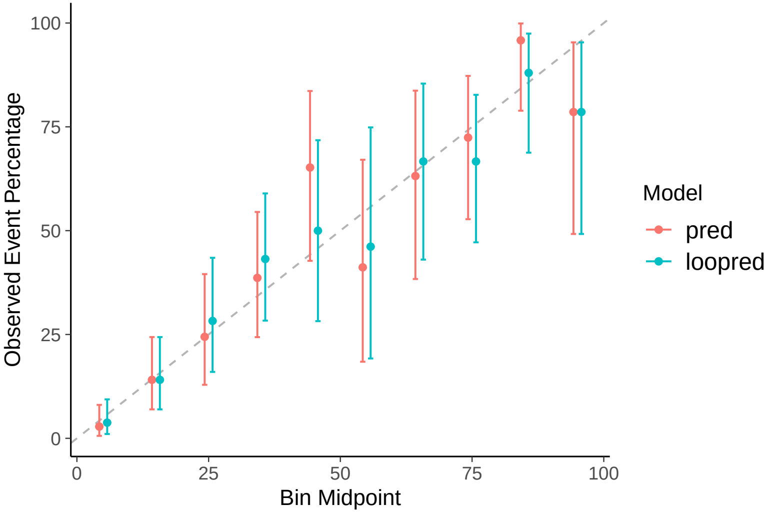
Classification example: Pima Indians Diabetes

Leave-one-out cross-validation classification accuracy 78%

Classification example: Pima Indians Diabetes

Leave-one-out cross-validation classification accuracy 78%

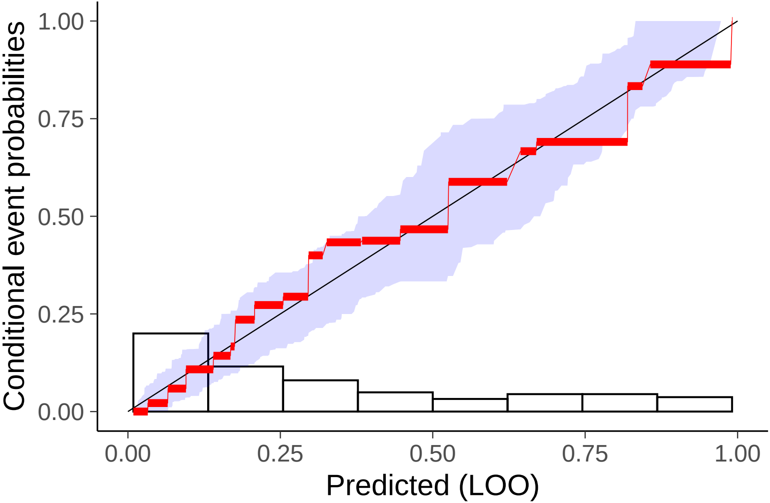
Calibration:



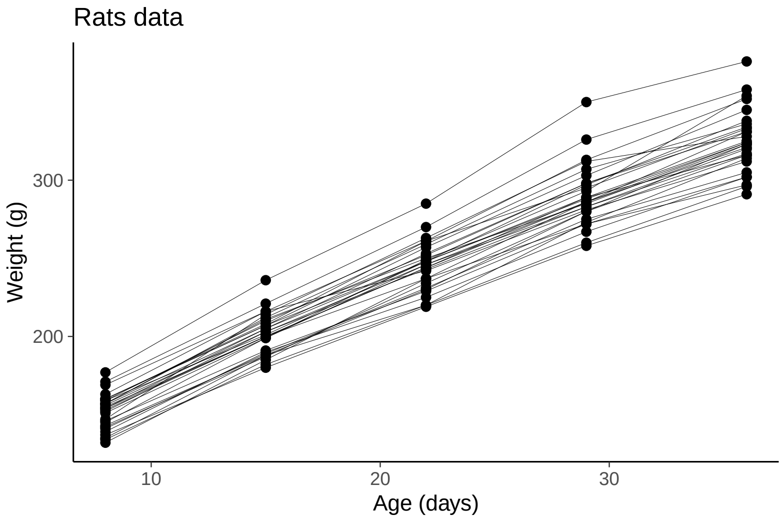
Classification example: Pima Indians Diabetes

Leave-one-out cross-validation classification accuracy 78%

Calibration:



Hierarchical example: Rats growth curves



https://avehtari.github.io/modelselection/rats_kcv.html

Hierarchical example: Rats growth curves

Simple linear model

```
fit_1 <- stan_glm(weight ~ age, data=dfrats)
```

Linear model with hierarchical intercept

```
fit_2 <- stan_glmer(weight ~ age + (1 | rat), data=dfrats)
```

Linear model with hierarchical intercept and slope

```
fit_3 <- stan_glmer(weight ~ age + (age | rat), data=dfrats)
```

Hierarchical example: Rats growth curves

Simple linear model

```
fit_1 <- stan_glm(weight ~ age, data=dfrats)
```

Linear model with hierarchical intercept

```
fit_2 <- stan_glmer(weight ~ age + (1 | rat), data=dfrats)
```

Linear model with hierarchical intercept and slope

```
fit_3 <- stan_glmer(weight ~ age + (age | rat), data=dfrats)
```

Instead of `stan_glm(er)`, use `brm` to get the Stan code, too.

Hierarchical example: Rats growth curves

Leave-one-out cross-validation

	elpd_diff	se_diff
hierarchical intercept and slope	0.0	0.0
hierarchical intercept	−23.6	9.3
simple linear model	−109.6	13.3

Hierarchical example: Rats growth curves

Leave-one-out cross-validation

	elpd_diff	se_diff
hierarchical intercept and slope	0.0	0.0
hierarchical intercept	−23.6	9.3
simple linear model	−109.6	13.3

Example analyses

- Time series with various ARMA models or Gaussian processes
- Spatial data with CAR or Gaussian processes
- Survival analyses with various hazard functions
- Linear vs non-linear regression
- Linear vs hierarchical model
- Ranking models