Tutorial on model assesment, selection and inference after selection

Aki Vehtari

Department of Computer Science,
Aalto University, Finland
aki.vehtari@aalto.fi

Outline

- Basics of predictive performance estimation
- When cross-validation is not needed
  - Simple model we trust
- When cross-validation is useful
  - We don't trust the model
  - Complex model with posterior dependencies
- On accuracy of cross-validation
- Cross-validation and hierarchical models
- When cross-validation is not enough
  - large number of models
- loo 2.0

LOO vs. WAIC

- LOO and WAIC estimate the same predictive performance citerion and are asymptotically equal
  - some of the discussion holds for WAIC, too
  - WAIC doesn't have as good diagnostics and fails earlier than PSIS-LOO used in loo package.

LOO vs. WAIC

- LOO and WAIC estimate the same predictive performance citerion and are asymptotically equal
  - some of the discussion holds for WAIC, too
  - WAIC doesn't have as good diagnostics and fails earlier than PSIS-LOO used in loo package.
- See more in
  - Decision theoretic review and more methods in Vehtari, A., and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6**, 142–228.
  - Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016. arXiv preprint arXiv:xxxxxxxxx. `http://arxiv.org/abs/xxxxx`
  - Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.. arXiv preprint arXiv:1507.04544. `http://arxiv.org/abs/1507.04544`

Predictive performance

- Ideal predictive performance with log score

$$\text{elpd} = \int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y},$$

where $p_t(\tilde{y})$ is unknown true future distribution

$\mathcal{M}$-open,-closed,-completed

- Following Bernardo & Smith (1994), there are three different approaches for dealing with the unknown p_t
  - $\mathcal{M}$-open
  - $\mathcal{M}$-closed
  - $\mathcal{M}$-completed

$\mathcal{M}$-open

- Explicit specification of $p_t(\tilde{y})$ in

$$\int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y},$$

  is avoided by re-using the observed data D as a pseudo Monte Carlo samples from the distribution of future data
- Bayesian leave-one-out cross-validation

$$\widehat{\mathrm{elpd}}_{\mathrm{LOO}} = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i, D_{-i}, M_k)$$

  - almost unbiased estimate for a single model

Cross-validation

- Naïve computation requires computation of n posteriors
- Less computation with
  - analytic solutions and approximations available for some models
  - importance sampling using the full posterior as the proposal (easy to use with Stan)
  - k-fold cross-validation
    - most robust

Leave-one-out cross-validation

- Special case is if we leave only one data point out (LOO-CV)
- LOO predictive density evaluated at $\mathbf{y}_i$

$$p(y_i|x_i, D_{-i}) = \int p(y_i|x_i, \theta)p(\theta|D_{-i})d\theta,$$

where D_{-i} is all the data except (y_i, x_i)

  - leave-one-out posterior $p(\theta|D_{-i})$ is close to full posterior $p(\theta|D)$, but we still avoid the double use of data
  - naïve implementation requires to do the posterior inference n times

Importance sampling

- Having samples θ^s from $p(\theta^s|D)$

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^{S} p(\tilde{y}_i|\theta^s) w_i^s}{\sum_{s=1}^{S} w_i^s},$$

where w_i^s are importance weights and

$$w_i^s = \frac{p(\theta^s|x_i, D_{-i})}{p(\theta^s|D)} \propto \frac{1}{p(y_i|\theta^s)}.$$

```
...
model {
  vector[N] eta;
  eta <- beta0 + z*beta;
  beta ~ normal(0, phi);
  phi ~ double_exponential(0, 10);
  y ~ bernoulli_logit(eta);
}
generated quantities {
  vector[N] log_lik;
  vector[N] eta;
  eta <- beta0 + z*beta;
  for (n in 1:N)
    log_lik[n] <- bernoulli_logit_lpdf(y[n],eta[n]);
}
```

Pareto smoothed importance sampling

- The variance of the importance weights in IS-LOO can be large or even infinite
- By fitting a generalized Pareto distribution to the tail of the weight distribution
  - obtain an estimate of the shape parameter k
  - if $k < \frac{1}{2}$ variance is finite, the central limit theorem holds
  - if $\frac{1}{2} \leq k < 1$ variance is infinite but mean exists, the generalized central limit theorem holds
  - if $k \geq 1$ variance and mean do not exist, the truncated estimate will have a finite variance but considerable bias
  - variance of the IS estimate can be reduced by Pareto smoothing the weights $\rightarrow$ PSIS-LOO
  - for $k < 0.7$ finite sample convergence rates practical

PSIS-LOO

- loo package in CRAN implements PSIS-LOO
  - loo 2.0 is using new version of Pareto smoothing
- rstanarm has integrated support
- References
  - 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. arXiv preprint `http://arxiv.org/abs/1507.04544`.
  - Vehtari, A., Gelman, A., Gabry, J. (2017). Pareto smoothed importance sampling. arXiv preprint `http://arxiv.org/abs/1507.02646`.

Model comparison with LOO

- Pairwise comparison of individual elpd's

$$\widehat{\text{elpd}}_{\text{diff}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\text{elpd}}_{\text{LOO},i,M_2} - \widehat{\text{elpd}}_{\text{LOO},i,M_1}\right]$$

- Compute also se for accuracy of the comparison

Example where cross-validation is not needed

- Simple model: we can look at the posterior directly
  - treatment effect of beta-blockers on mortality – betablockers.Rmd

$\mathcal{M}$-completed

- Ideal predictive performance with log score

$$\text{elpd} = \int p_t(\tilde{y}) \log p(\tilde{y}|D, M_k) d\tilde{y},$$

- Reference predictive approach

$$\widehat{\text{elpd}}_{\text{ref}} = \int p(\tilde{y}|D, M_*) \log p(\tilde{y}|D, M_k) d\tilde{y},$$

where M_* is a reference model we trust

  - using a model decreases variance, but may introduce bias
  - smaller error more useful than unbiasednes, but need to be careful as bias can be very large

Examples where cross-validation is useful

- We don't trust the model: possible model misspecification
  - treatment effect on number of roaches – roaches.Rmd
- Complex model with posterior dependencies: difficult to analyse posterior
  - colinearity in covariates – colinear.Rmd

On Accuracy of cross-validation

- se for elpd_loo is slightly underestimated and the distribution is often highly skewed
- se for elpd_diff is also underestimated and true distribution might be skewed
- If using elpd_diff and se to compute the probability that one model is better than other ($\Phi(0|\text{elpd\_diff}, \text{se})$), these probabilities are not calibrated
  - be cautious when interpreting or reporting these
- We know how to slightly improve the calibration, and we'll report results on the effects of miscalibration later this year

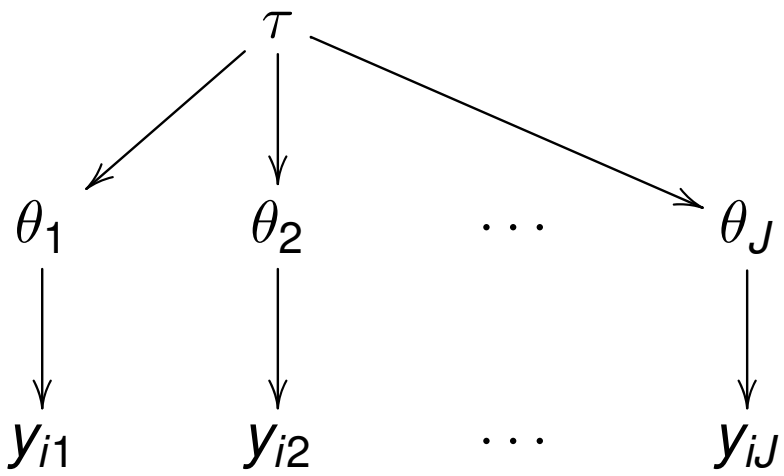
PSIS-LOO+

- Compute leave-one-out posterior exactly for those observations for which $\hat{k} > 0.7$
  - For rstanarm models: loo(rstanarmfit, k_threshold=0.7)

K-fold-CV

- Instead of leaving one observation out, leave a block of observations
- When data is divided in K blocks the approach is called K-fold-CV
- If, for example, $K = 10$, then 90% of data is used to form the posterior, which often produces similar posterior as full data
- k-fold-CV should be used
  - if PSIS-LOO diagnostics indicate problems with importance sampling and PSIS-LOO+ would compute many more than K posteriors
    For rstanarm models: kfold(rstanarmfit, K = 10)
  - if the prediction task is for groups

Hierarchical models



1) Predicting new y_{ij} given an existing group $j \in (1, \ldots, J)$
   - LOO or randomized/stratified k-fold-CV
2) Predicting new y_{ij} given a new group $j = J + 1$
   - grouped K-fold-CV

Leave-one-group-out cross-validation

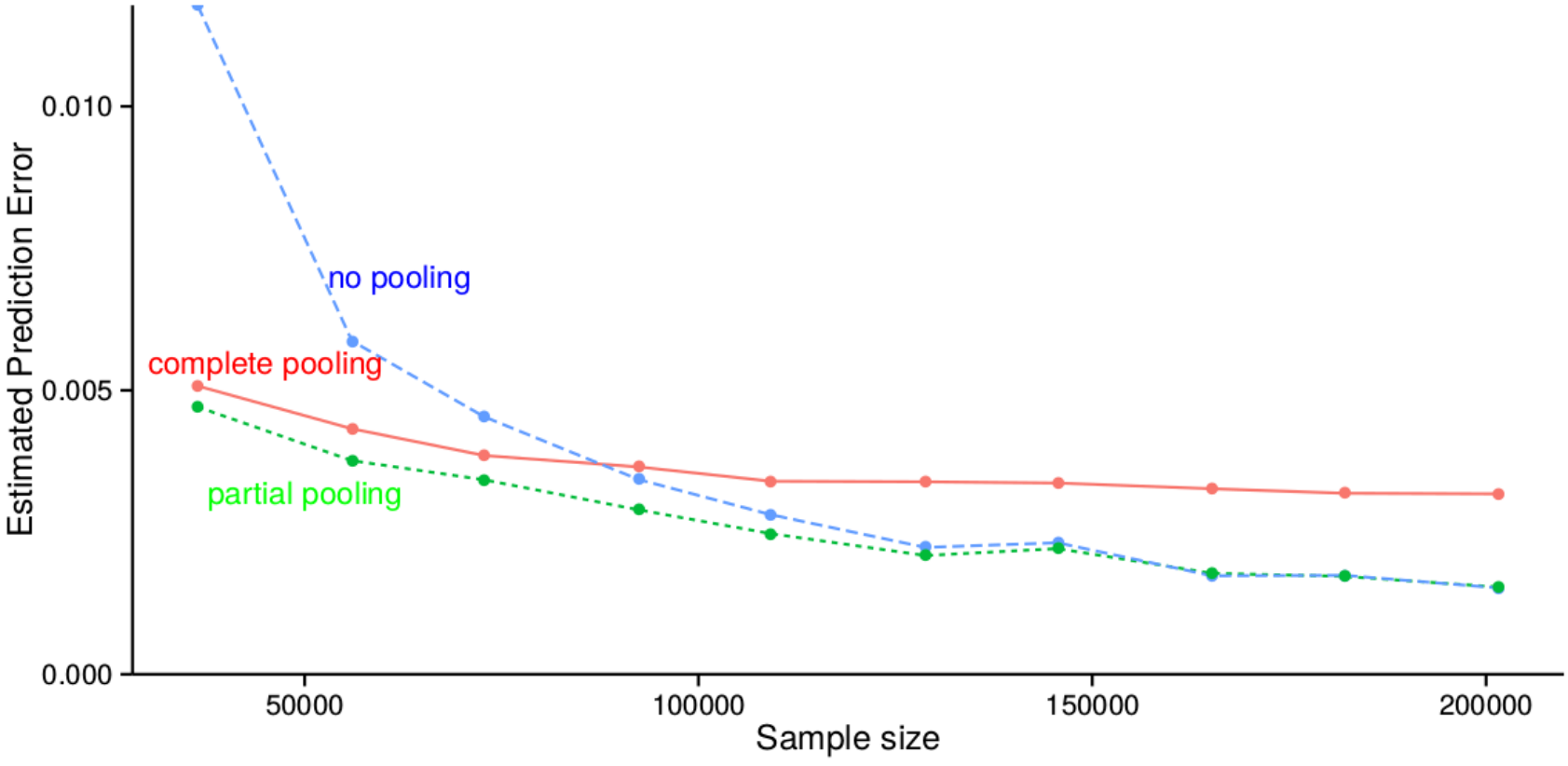
- Importance sampling is less likely to work, as a group of observations is likely to be more influential than just one (and thus full posterior and loo posterior can be doo different)

Leave-one-group-out cross-validation

- Importance sampling is less likely to work, as a group of observations is likely to be more influential than just one (and thus full posterior and loo posterior can be doo different)
- Marginalization in style of Rabe-Hesketh and Furr (Invited talk Wednesday morning) can be used (currently with qudrature implemented in additional software)

Hierarchical models

- Hierarchical model for polling results in different states
  - Predicting new y_{ij} given an existing group $j \in (1, \ldots, J)$
  - Wang, W., and Gelman, A. (2014). Difficulty of selecting among multilevel models using predictive accuracy, Statistics and Its Inference, 7:1.



Cross-validation for hierarchical models

- rstanarm support for leave-one-out-group cross-validation in progress
- Hierarchical model comparison examples in progress

Selection induced bias

- Selection induced bias in LOO-CV
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the LOO-CV estimate for the selected model is biased
  - recognised already, e.g., by Stone (1974)

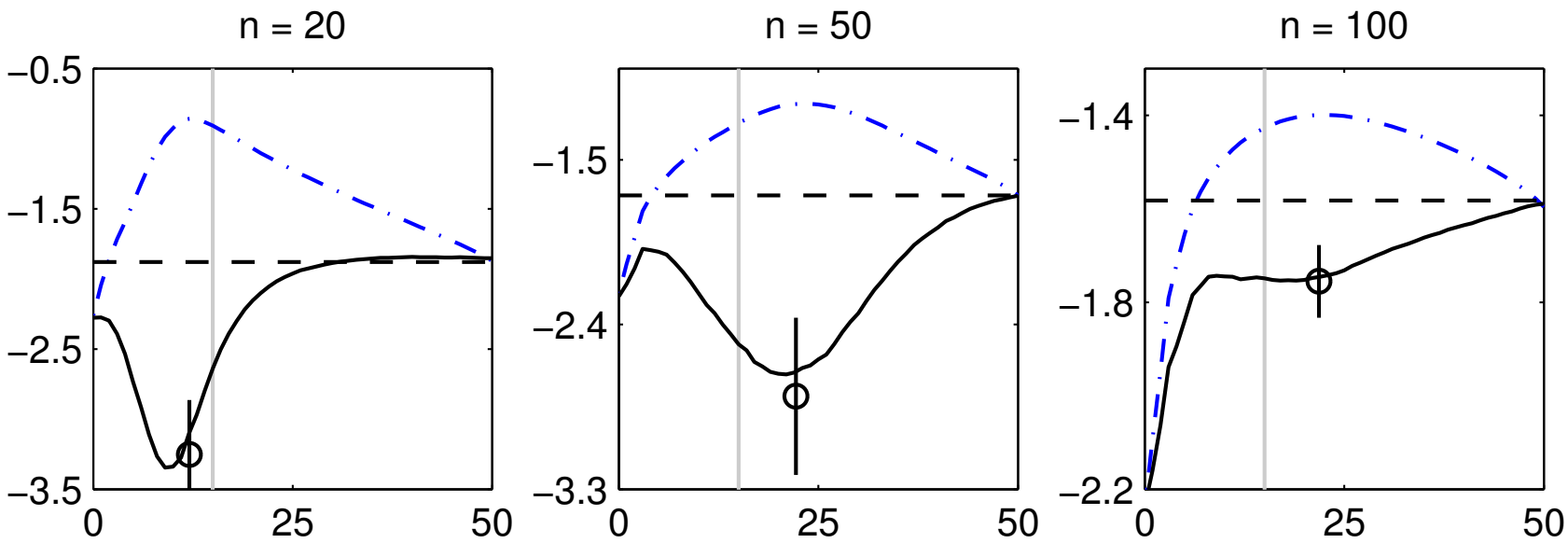
Selection induced bias

- Selection induced bias in LOO-CV
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the LOO-CV estimate for the selected model is biased
  - recognised already, e.g., by Stone (1974)
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models

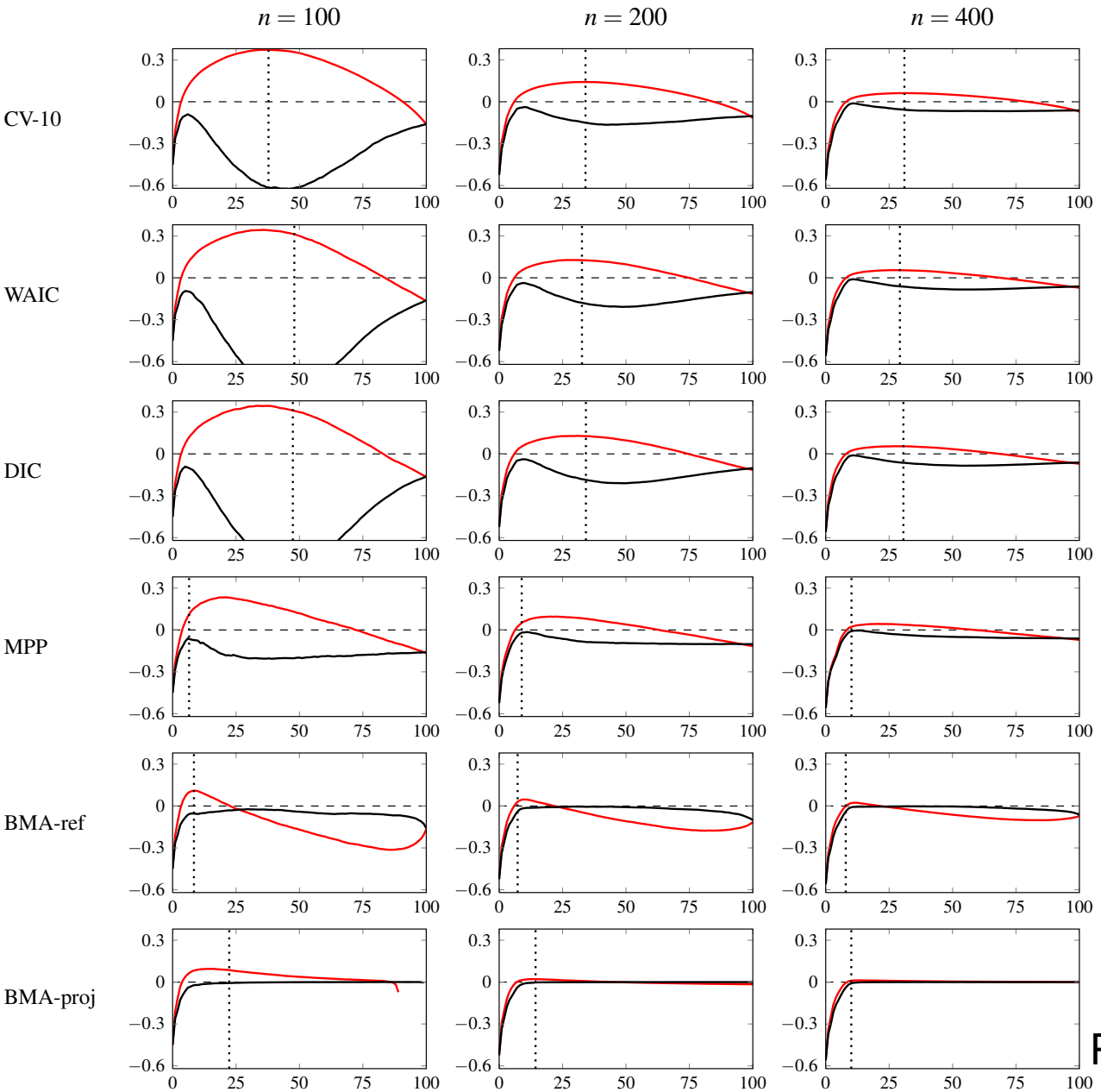
Selection induced bias

- Selection induced bias in LOO-CV
  - same data is used to assess the performance and make the selection
  - the selected model fits more to the data
  - the LOO-CV estimate for the selected model is biased
  - recognised already, e.g., by Stone (1974)
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models
- Bigger problem if there is a large number of models as in covariate selection

Selection induced bias in variable selection

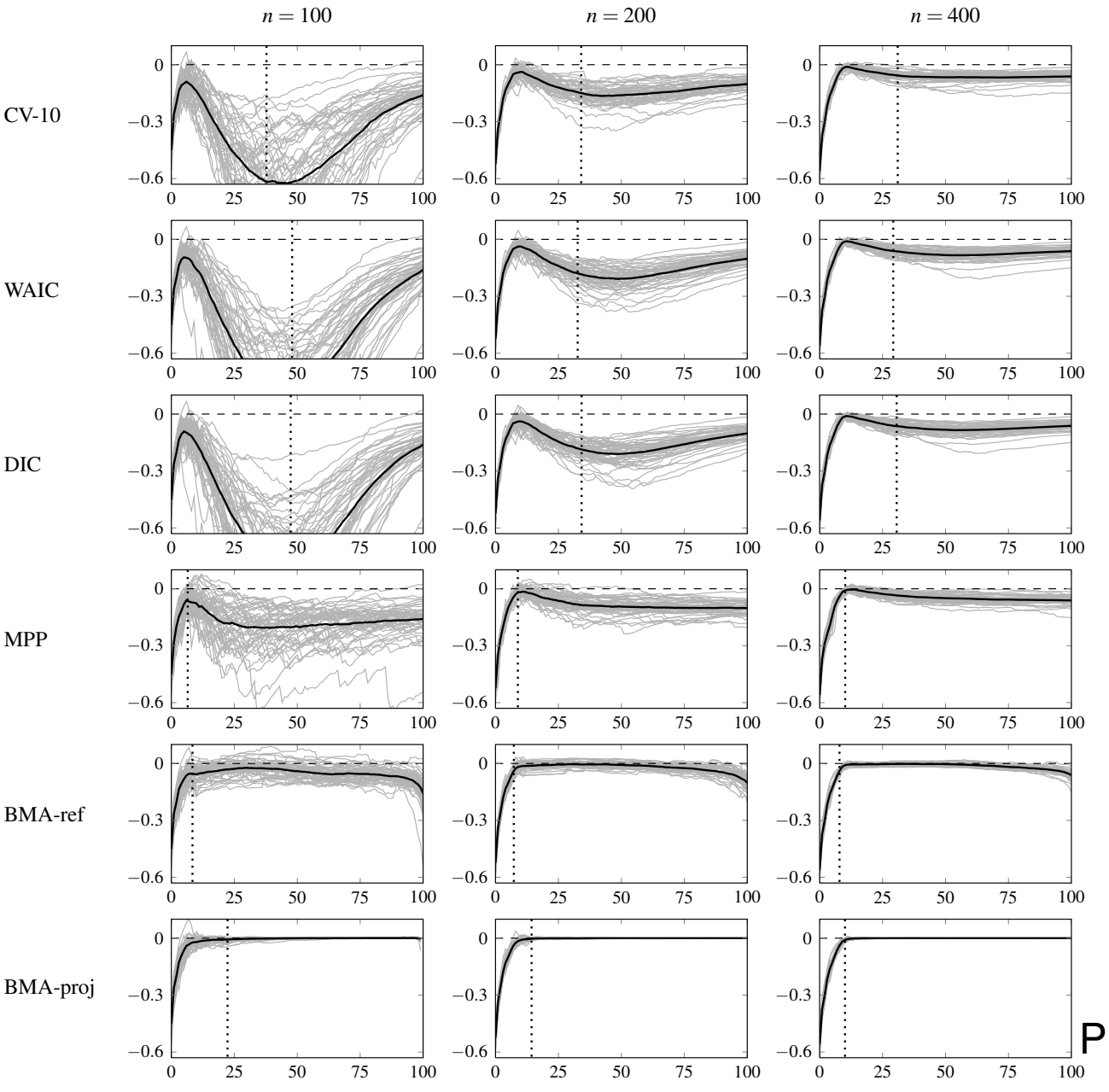


Selection induced bias in variable selection



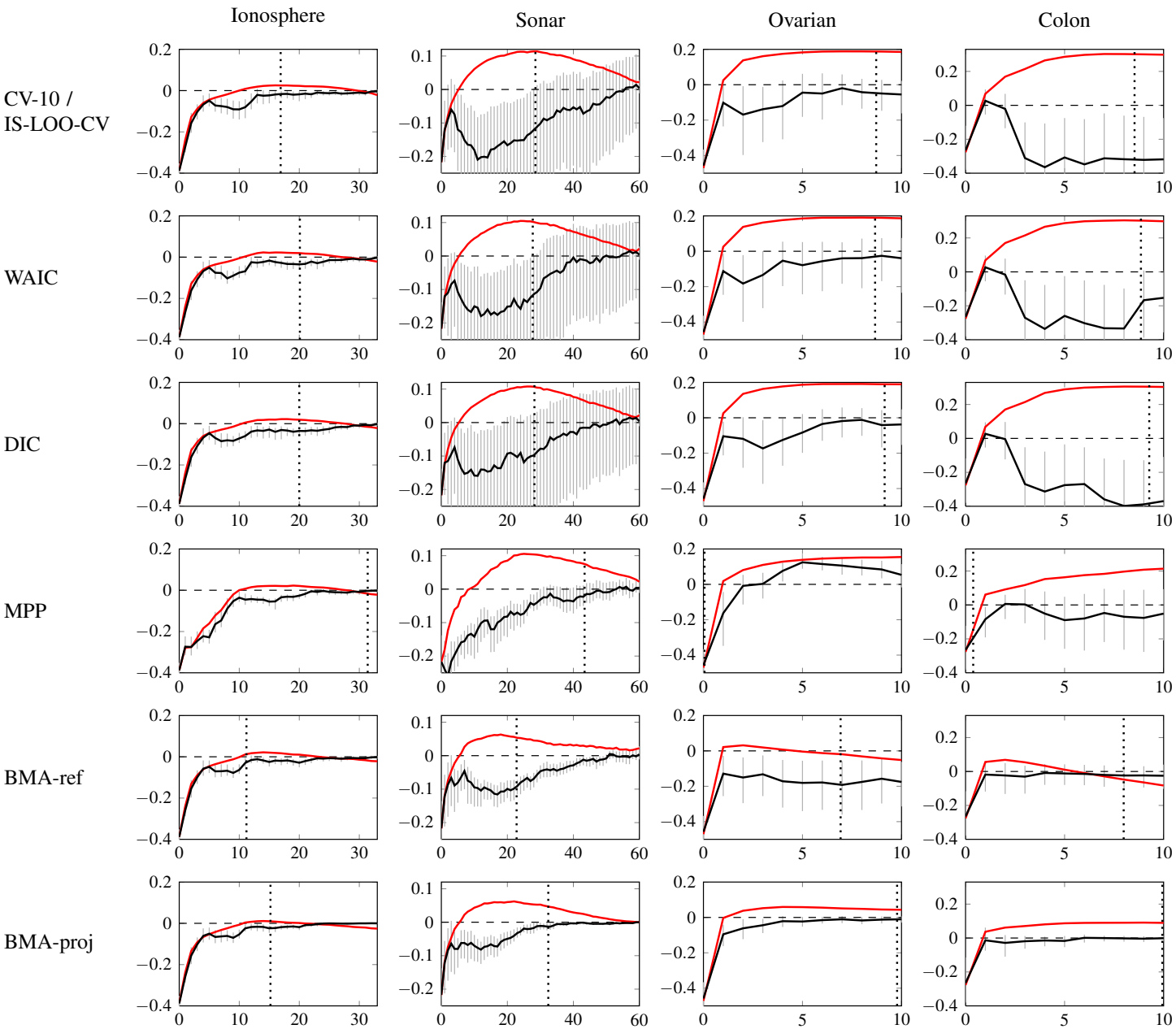
Piironen & Vehtari (2017)

Selection induced bias in variable selection



Piironen & Vehtari (2017)

Selection induced bias in variable selection



Piironen & Vehtari (2017)

Projection predictive method, general idea

- Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))
- Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)
  - has low variance
  - able to preserve information from the full model

Projection predictive method, general idea

- Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))
- Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)
  - has low variance
  - able to preserve information from the full model
- General idea
  - Fit the full encompassing model (with all the inputs) with best possible prior information

Projection predictive method, general idea

- Originally proposed for generalized linear models by Goutis and Robert (1998), Dupuis and Robert (2003) (the decision theoretic idea of using the full model can be tracked to Lindley (1968), see also many related references in Vehtari and Ojanen (2012))
- Performs well in practice in comparison to many other methods (Piironen and Vehtari 2016)
  - has low variance
  - able to preserve information from the full model
- General idea
  - Fit the full encompassing model (with all the inputs) with best possible prior information
  - Any submodel (reduced number of inputs) is trained by minimizing predictive Kullback-Leibler (KL) divergence to the full model (= projection)
    - For a given number of variables, choose the model with minimal projection discrepancy

Projective predictive covariate selection, idea

- 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 $\beta_j \neq 0$)

Projective predictive covariate selection, idea

- 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

Projective predictive covariate selection, idea

- 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:

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

Projective predictive covariate selection, idea

- 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:

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

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

Projective predictive feature selection, computation

- We have posterior draws $\{\boldsymbol{\theta}^s\}_{s=1}^S$, for the full model ($\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$) and $\boldsymbol{\beta}$ is in general non-sparse (all $\boldsymbol{\beta}_j \neq 0$)

Projective predictive feature selection, computation

- We have posterior draws $\{\theta^s\}_{s=1}^S$, for the full model ($\theta = (\beta, \sigma^2)$) and β is in general non-sparse (all $\beta_j \neq 0$)
- The predictive distribution $p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s)$ represents our best knowledge about future $\tilde{y}$

Projective predictive feature selection, computation

- We have posterior draws $\{\theta^s\}_{s=1}^S$, for the full model ($\theta = (\beta, \sigma^2)$) and β is in general non-sparse (all $\beta_j \neq 0$)
- The predictive distribution $p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s)$ represents our best knowledge about future $\tilde{y}$
- Easier optimization problem by changing the order of integration and optimization (Goutis & Robert, 1998):

$$\theta_\perp^s = \arg\min_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y}_i \mid \theta^s) \,\|\, p(\tilde{y}_i \mid \hat{\theta}) \right)$$

Projective predictive feature selection, computation

- We have posterior draws $\{\theta^s\}_{s=1}^S$, for the full model ($\theta = (\beta, \sigma^2)$) and β is in general non-sparse (all $\beta_j \neq 0$)
- The predictive distribution $p(\tilde{y} \mid D) \approx \frac{1}{S} \sum_s p(\tilde{y} \mid \theta^s)$ represents our best knowledge about future $\tilde{y}$
- Easier optimization problem by changing the order of integration and optimization (Goutis & Robert, 1998):

$$\theta_\perp^s = \arg\min_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\Big(p(\tilde{y}_i \mid \theta^s) \,\|\, p(\tilde{y}_i \mid \hat{\theta})\Big)$$

- $\theta_\perp^s$ are now (approximate) draws from the projected distribution

Projection by draws

- Projection of one Monte Carlo sample can be solved
  - Gaussian case: analytically

$$\mathbf{w}_\perp = (\mathbf{X}_\perp{}^\mathsf{T}\mathbf{X}_\perp)^{-1}\mathbf{X}_\perp{}^\mathsf{T}\,\mathbf{f}$$

$$\sigma_\perp^2 = \sigma^2 + \frac{1}{n}(\mathbf{f}-\mathbf{f}_\perp)^\mathsf{T}(\mathbf{f}-\mathbf{f}_\perp)$$

Projection by draws

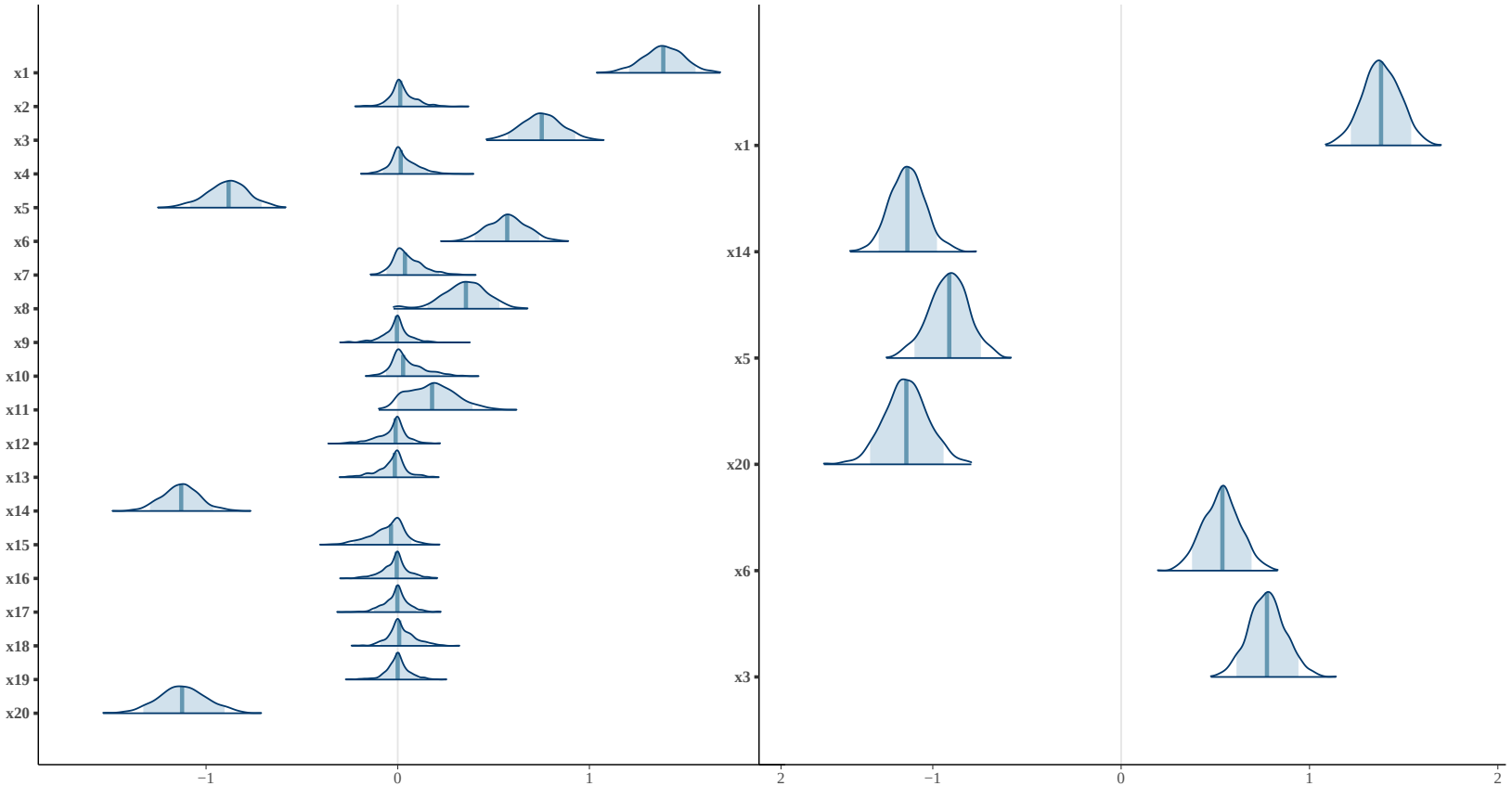
- Projection of one Monte Carlo sample can be solved
  - Gaussian case: analytically

$$\mathbf{w}_\perp = (\mathbf{X}_\perp{}^\mathsf{T}\mathbf{X}_\perp)^{-1}\mathbf{X}_\perp{}^\mathsf{T}\,\mathbf{f}$$

$$\sigma_\perp^2 = \sigma^2 + \frac{1}{n}(\mathbf{f}-\mathbf{f}_\perp)^\mathsf{T}(\mathbf{f}-\mathbf{f}_\perp)$$

  - Exponential family case: equivalent to finding the maximum likelihood parameters for the submodel with the observations replaced by the fit of the reference model (Goutis & Robert, 1998; Dupuis & Robert, 2003)

Example



The full model

A projected model (with variables ordered in relevance)

rstanarm + projpred + bayesplot

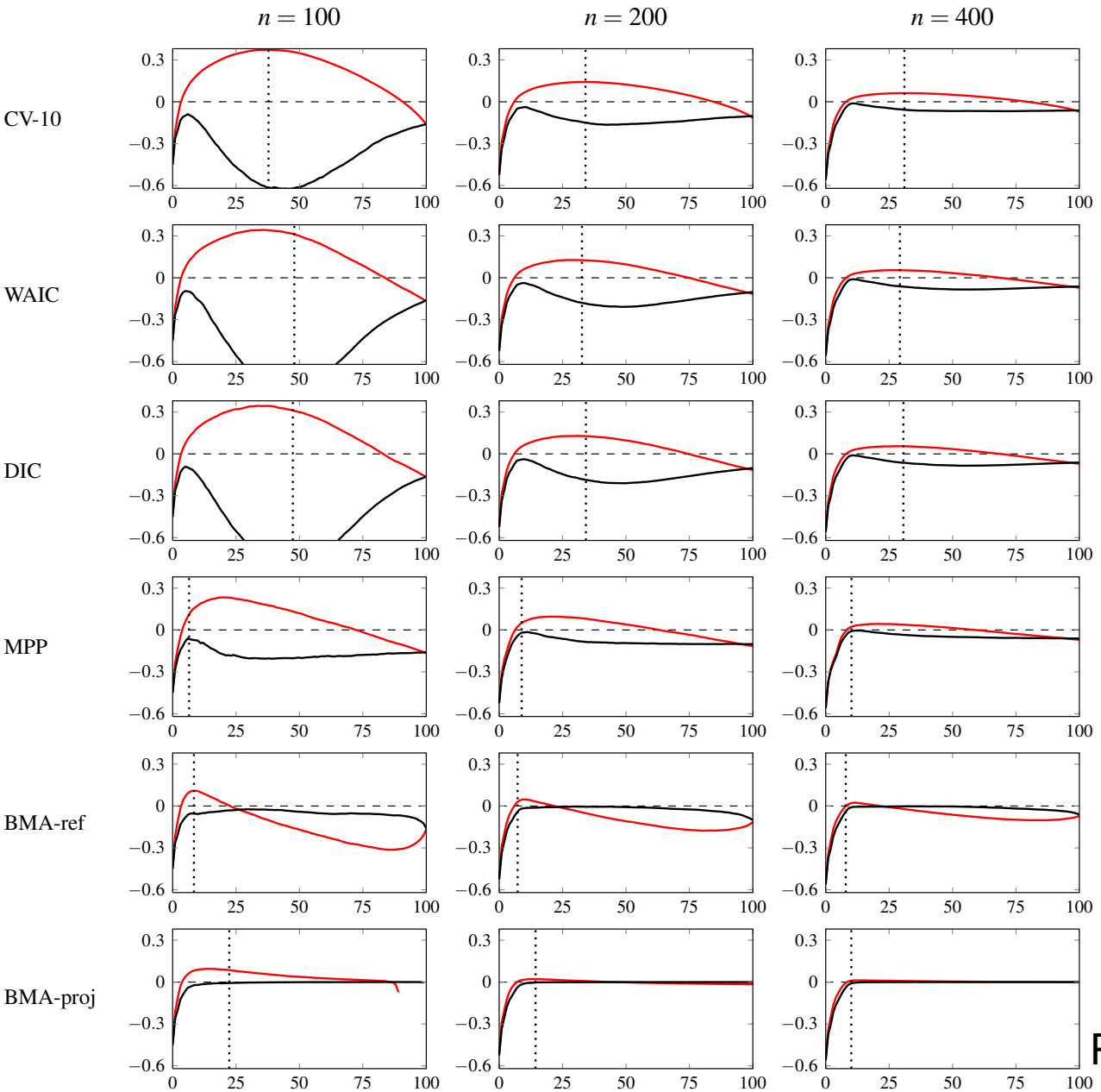
Projection predictive variable selection

- In variable selection usually not feasible to go through all variable combinations

Projection predictive variable selection

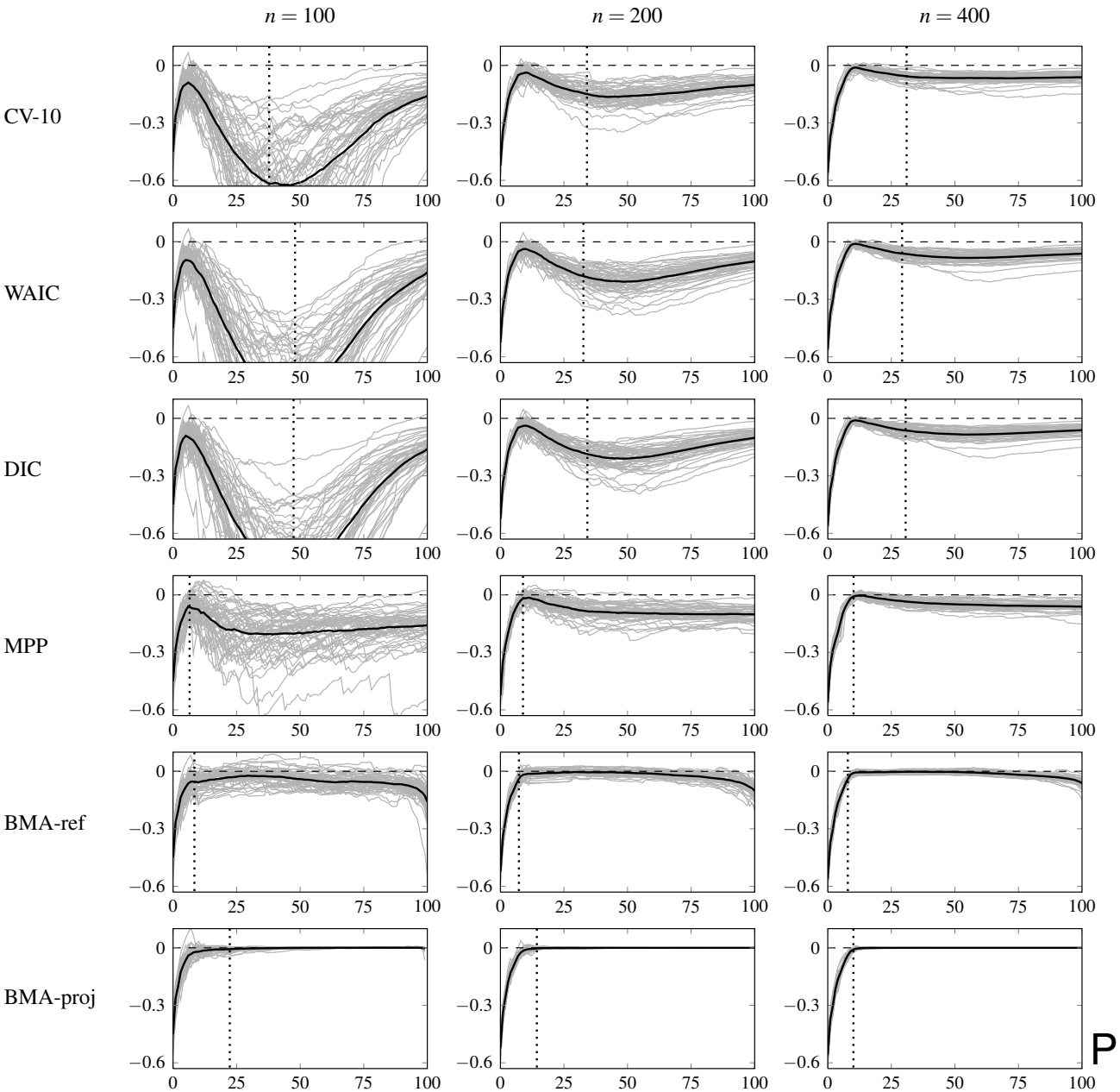
- In variable selection usually not feasible to go through all variable combinations
- Use e.g. forward search to explore promising combinations
    - start from the empty model, at each step add the variable that reduces the objective the most
    - stop when the performance similar to the full model
        - can use PSIS-LOO to estimate the performance and to choose the model size

Selection induced bias in variable selection



Piironen & Vehtari (2017)

Selection induced bias in variable selection



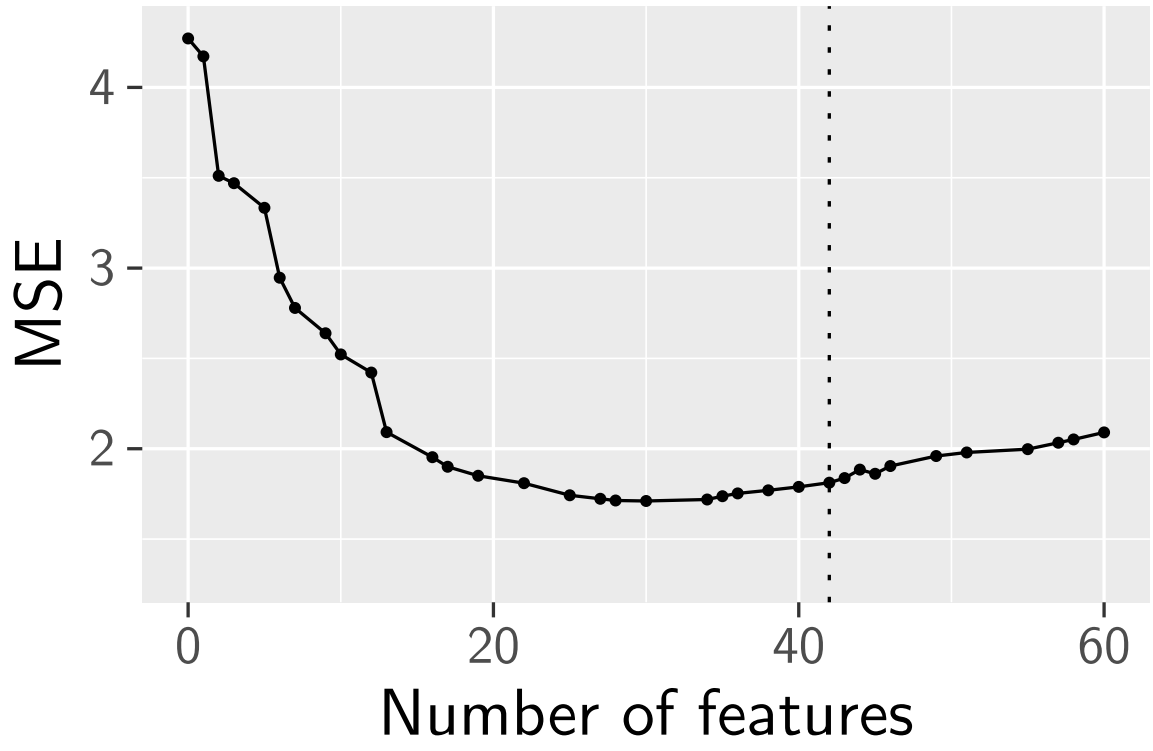
Piironen & Vehtari (2017)

Simulated example

$n = 80, p = 200$, only 7 features are relevant

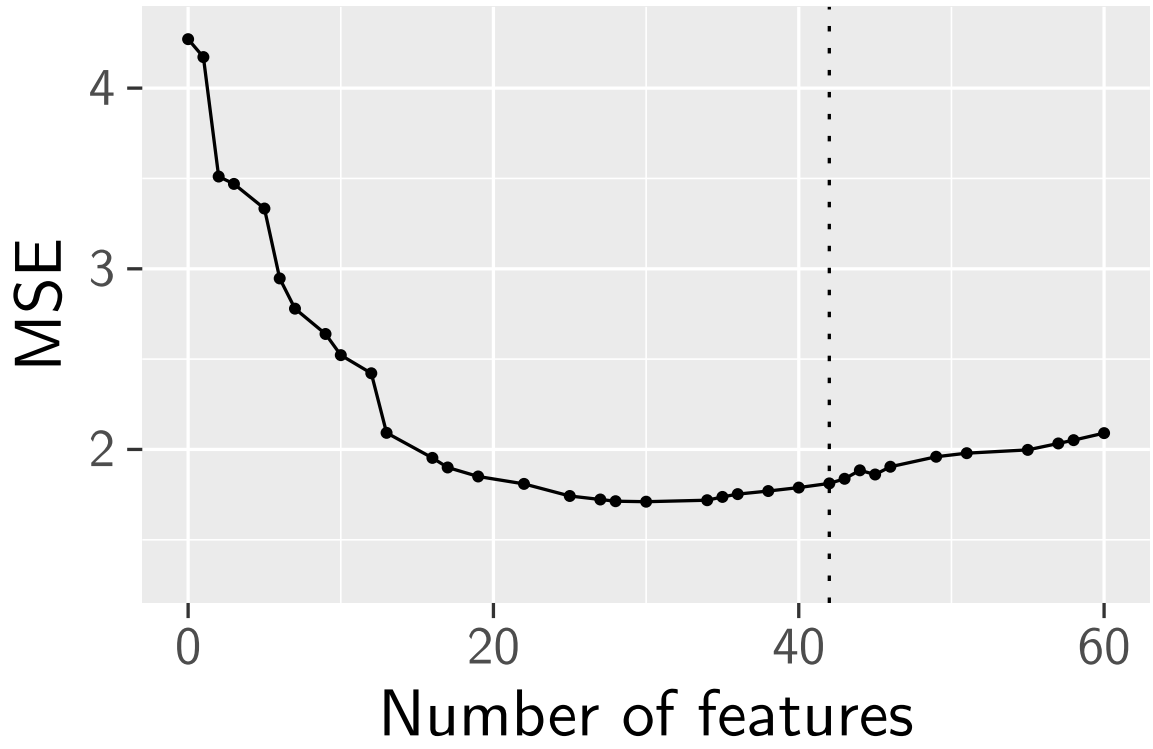
Simulated example

$n = 80, p = 200$, only 7 features are relevant



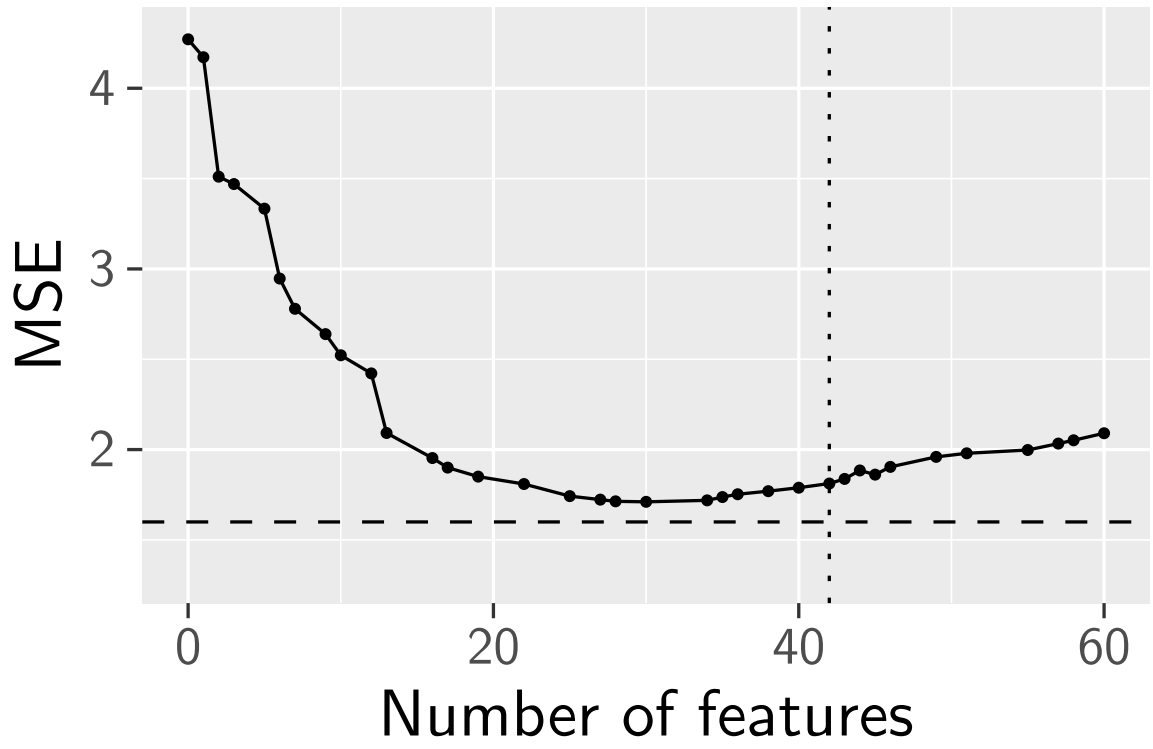
- Lasso-path when λ is varied, optimal model size by cross-validation (dotted) vertical axis shows the test error

Simulated example



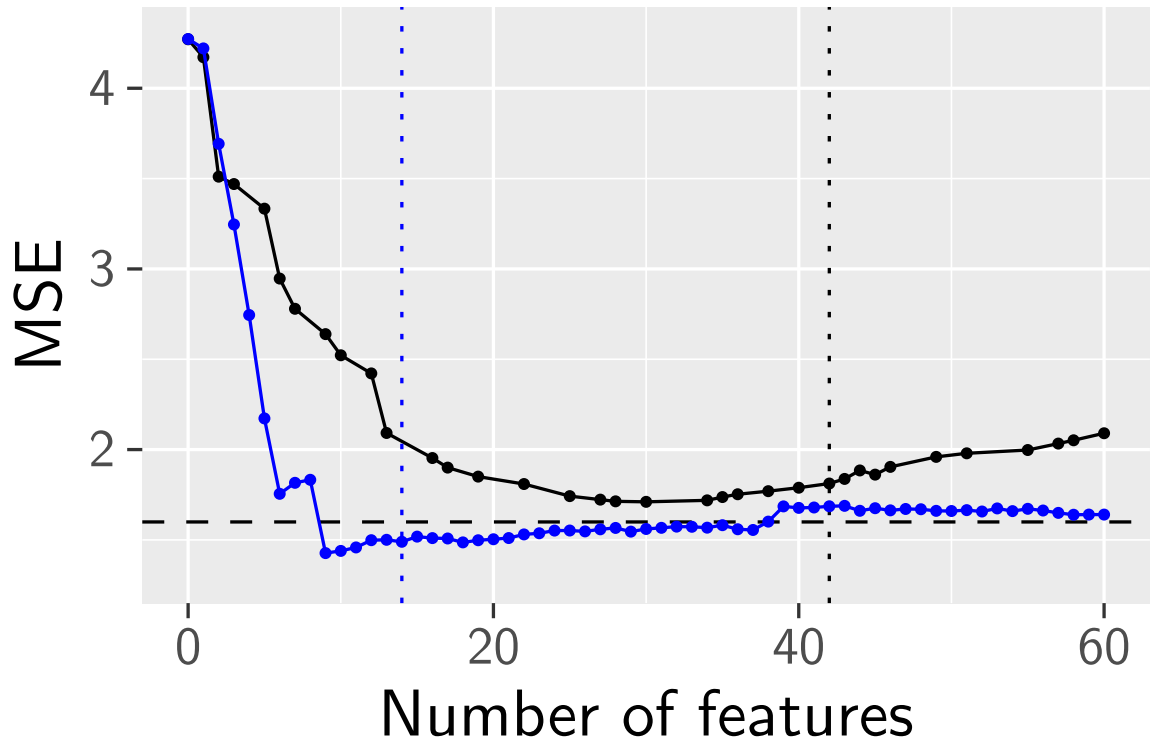
- Lasso-path (black)

Simulated example



- Lasso-path (black), full model Bayes with HS-prior (dashed)

Simulated example



- Lasso-path (black), full model Bayes with HS-prior (dashed) and the L_1-projection (blue)

Example

- Predicting disease risk with logistic regression – diabetes.Rmd

projpred

- Not yet in CRAN, but hopefully this spring

loo 2.0

- Improved PSIS-LOO
  - Improved Pareto diagnostics and smoothing
  - Effective sample size and se estimates
  - Vehtari, A., Gelman, A., Gabry, J. (2017). Pareto smoothed importance sampling. arXiv preprint `http://arxiv.org/abs/1507.02646`
- Model averaging
  - Pseudo-BMA+ weights
  - Stacking weights
  - Yao, Y., Vehtari, A., Simpson, D. and Gelman, A.: 2017, Using stacking to average Bayesian predictive distributions, Bayesian analysis . Accepted for publication, preprint arXiv:1704.02030.
- Helper functions for k-fold-CV

Model averaging

- How to combine different models?
  - BMA: integrate over the model space
  - Akaike type weighting
  - Stacking

Akaike type weighting

- Replace marginal likelihoods with exp(-AIC/2)
  - one vague sentence about asymptotics by Akaike, no any theoretical proof that it would work

Akaike type weighting

- Replace marginal likelihoods with exp(-AIC/2)
  - one vague sentence about asymptotics by Akaike, no any theoretical proof that it would work
- Replace marginal likelihoods with exp(-WAIC/2)
  - WAIC is the fully Bayesian version of AIC

Akaike type weighting

- Replace marginal likelihoods with exp(-AIC/2)
  - one vague sentence about asymptotics by Akaike, no any theoretical proof that it would work
- Replace marginal likelihoods with exp(-WAIC/2)
  - WAIC is the fully Bayesian version of AIC
- Replace marginal likelihoods with exp(LOO)
  - Bayesian LOO is asymptotically equal to WAIC, but more reliable

Akaike type weighting

- Replace marginal likelihoods with exp(LOO)
  - Pseudo BMA
  - Pseudo BMA+
    - take into account the uncertainty in LOO esimate

Bayesian stacking

- Consider the model averaging as a decision problem with aim of maximizing the predictive performance

Bayesian stacking

- Consider the model averaging as a decision problem with aim of maximizing the predictive performance
- Maximize the scoring rule of the predictive distribution,

$$\max_{w} S\Big(\sum_{k=1}^{K} w_k p(\tilde{y}|D, M_k), p_t(\tilde{y}|D)\Big),$$

- As we don't know $p_t(\tilde{y}|D)$, let's do the usual M-open trick and use LOO

Bayesian stacking

- Consider the model averaging as a decision problem with aim of maximizing the predictive performance
- Maximize the scoring rule of the predictive distribution,

$$\max_{w} S\Big(\sum_{k=1}^{K} w_k p(\tilde{y}|D, M_k), p_t(\tilde{y}|D)\Big),$$

- As we don't know $p_t(\tilde{y}|D)$, let's do the usual M-open trick and use LOO
- We define the stacking weights as the solution to the following optimization problem:

$$\max_{w} \frac{1}{n} \sum_{i=1}^{n} S\Big(\sum_{k=1}^{K} w_k \hat{p}(y_i|D_{-i}, M_k)\Big),$$

$$s.t. \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1.$$

Bayesian stacking

- The combined estimation of the predictive density is

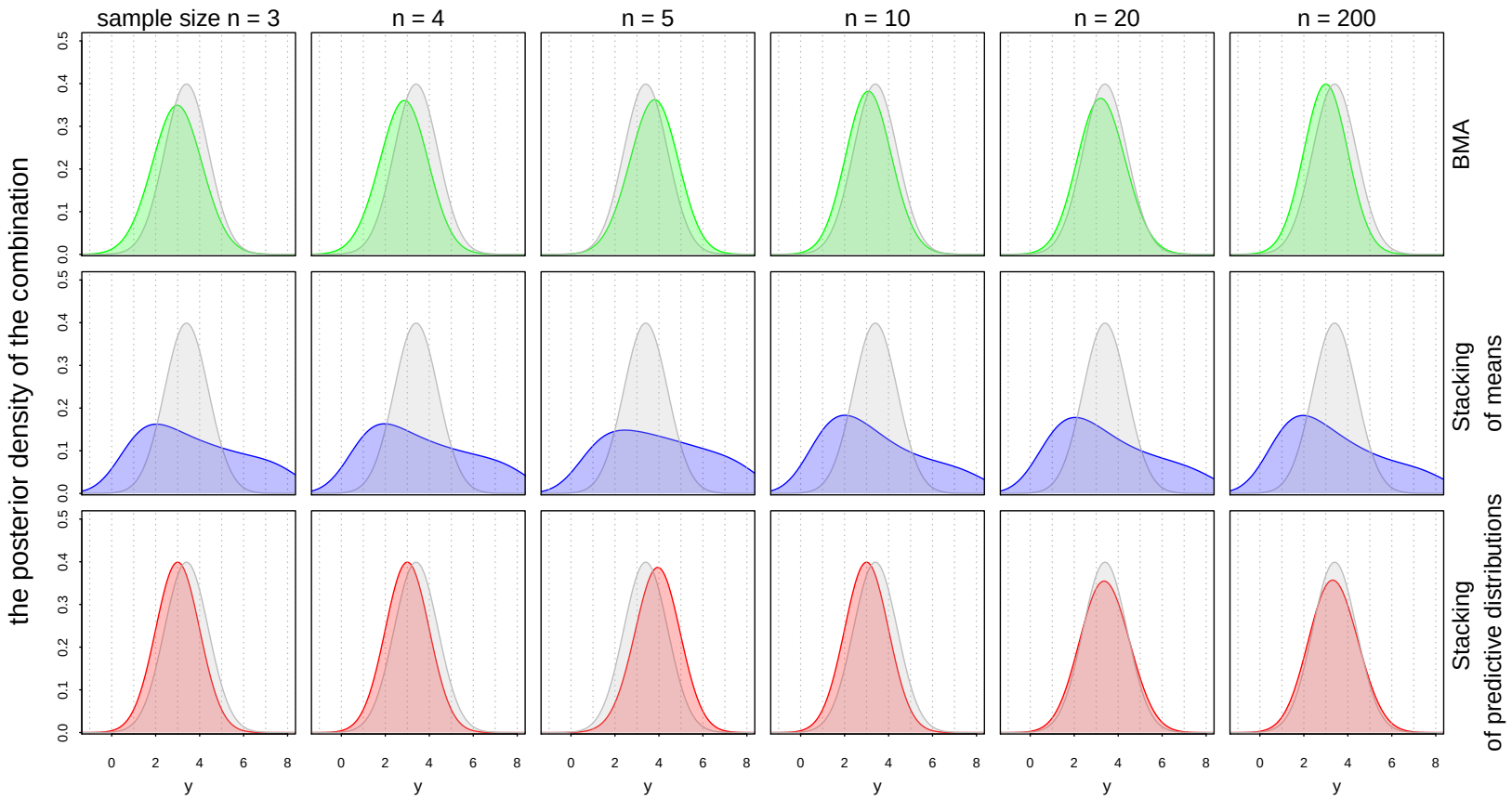
$$\hat{p}(\tilde{y}|D) = \sum_{k=1}^{K} \hat{w}_k p(\tilde{y}|D, M_k).$$

- When using log-score (corresponding to Kullback-Leibler divergence), we call this *stacking of predictive distributions*:

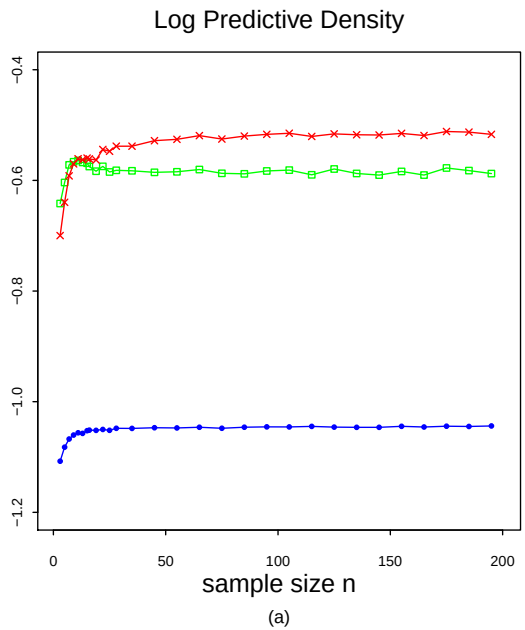
$$\max_{w} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k p(y_i|D_{-i}, M_k),$$

$$s.t. \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1.$$

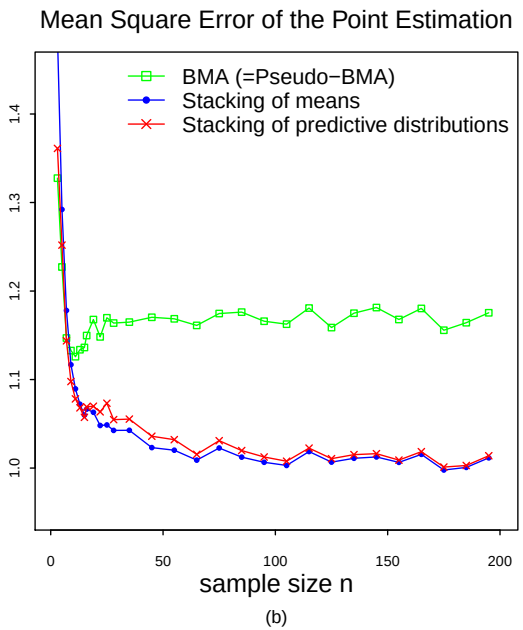
Gaussian mixture example



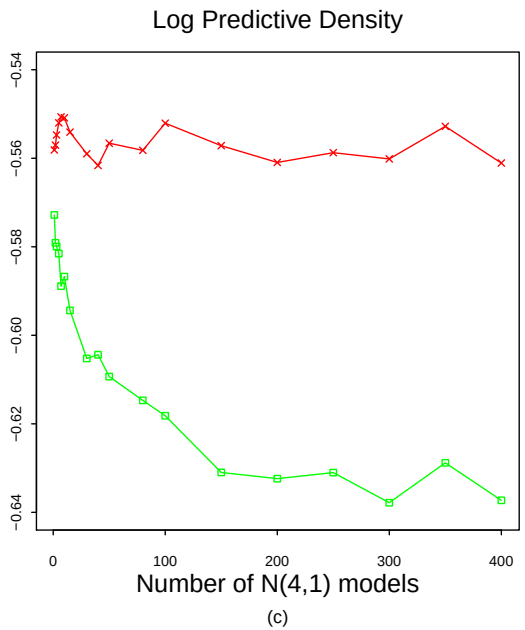
Gaussian mixture example



(a)

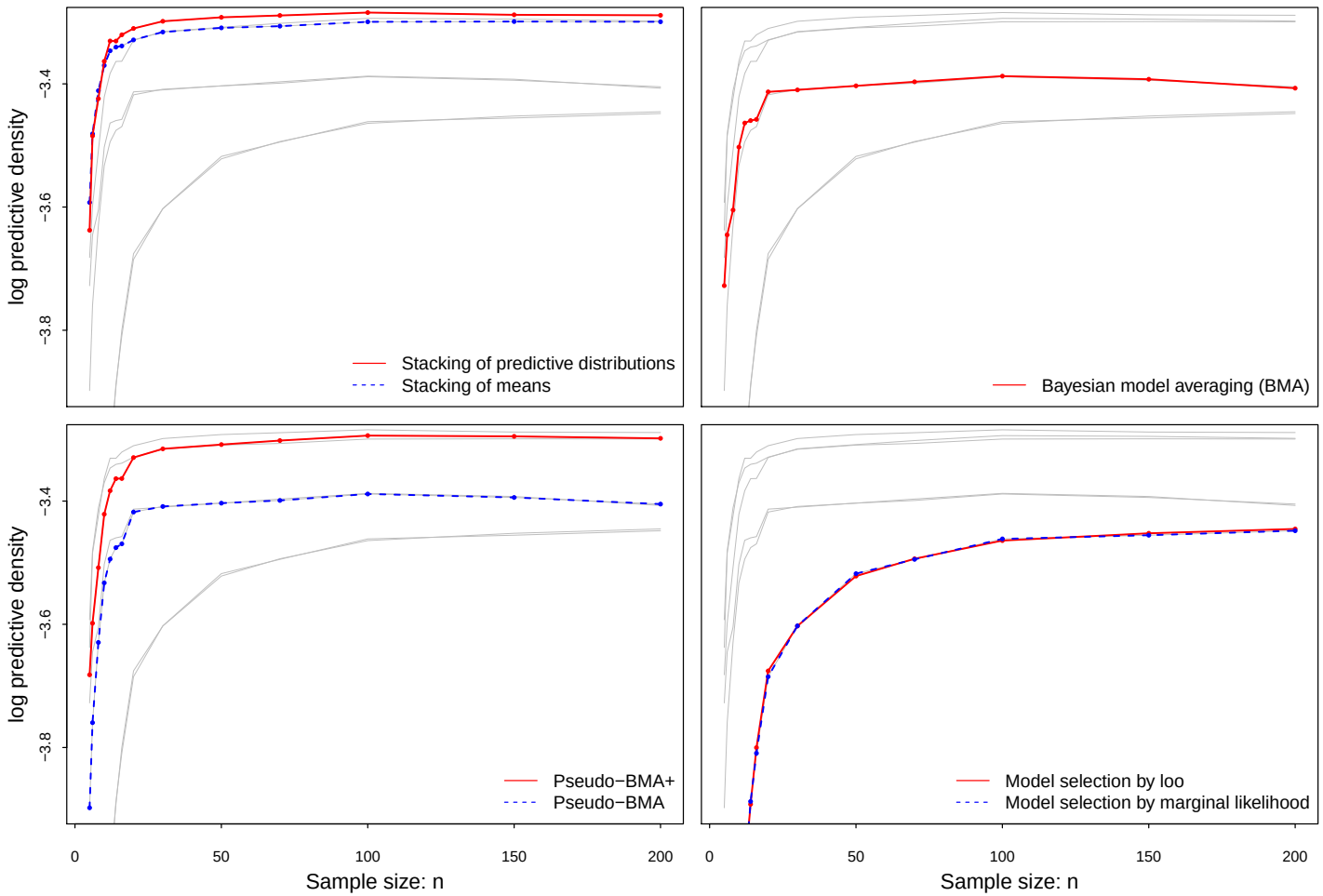


(b)

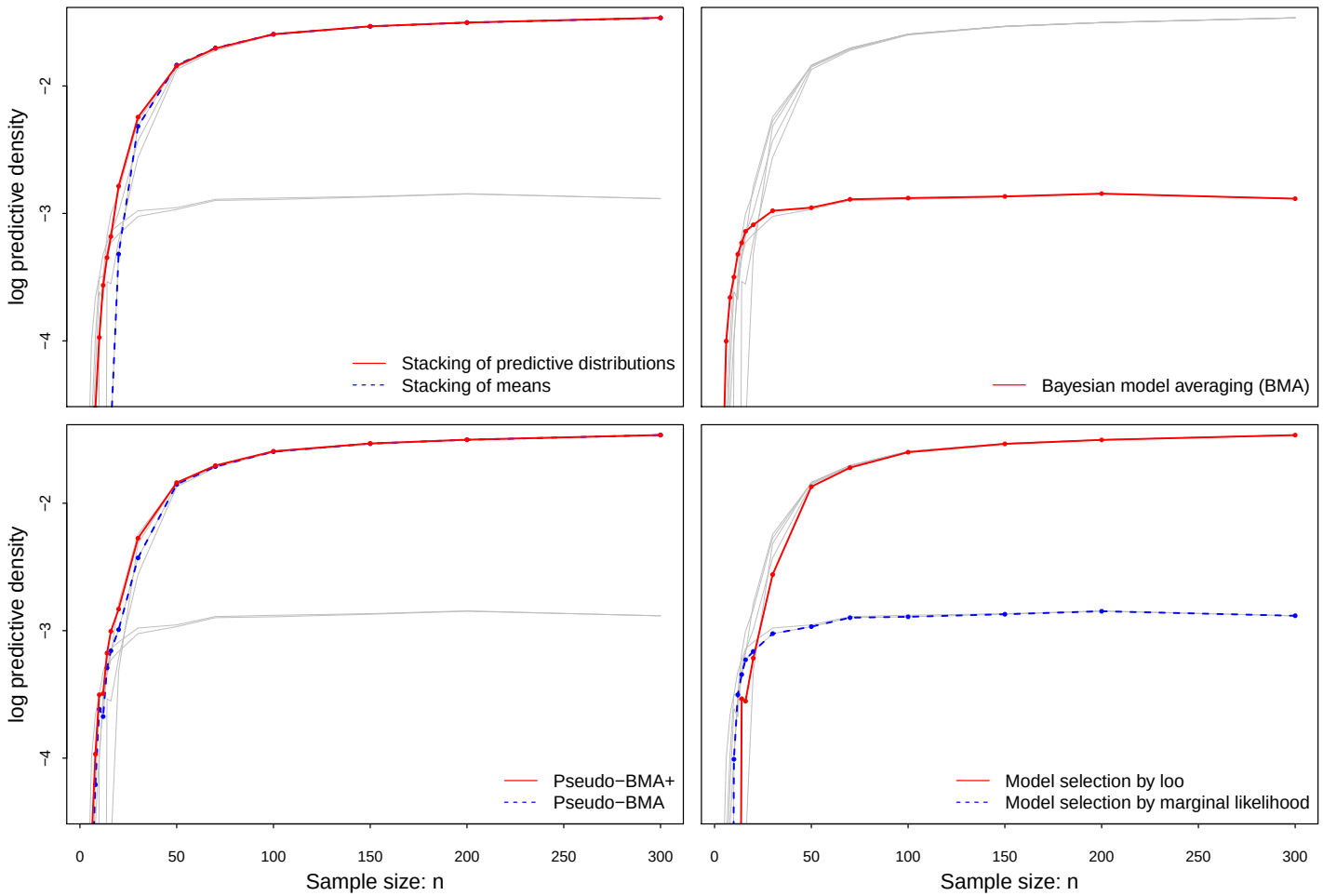


(c)

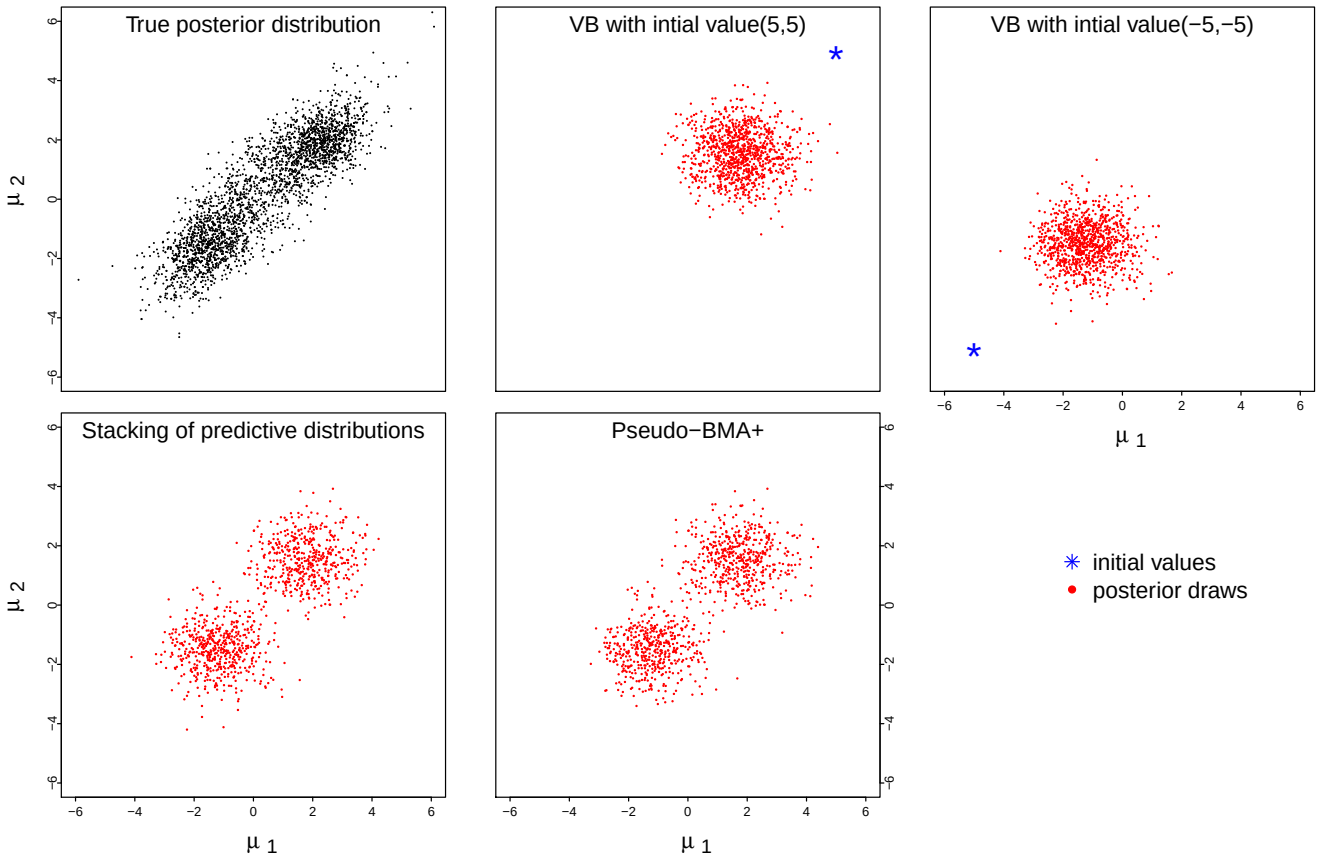
Linear subset regression example *k*



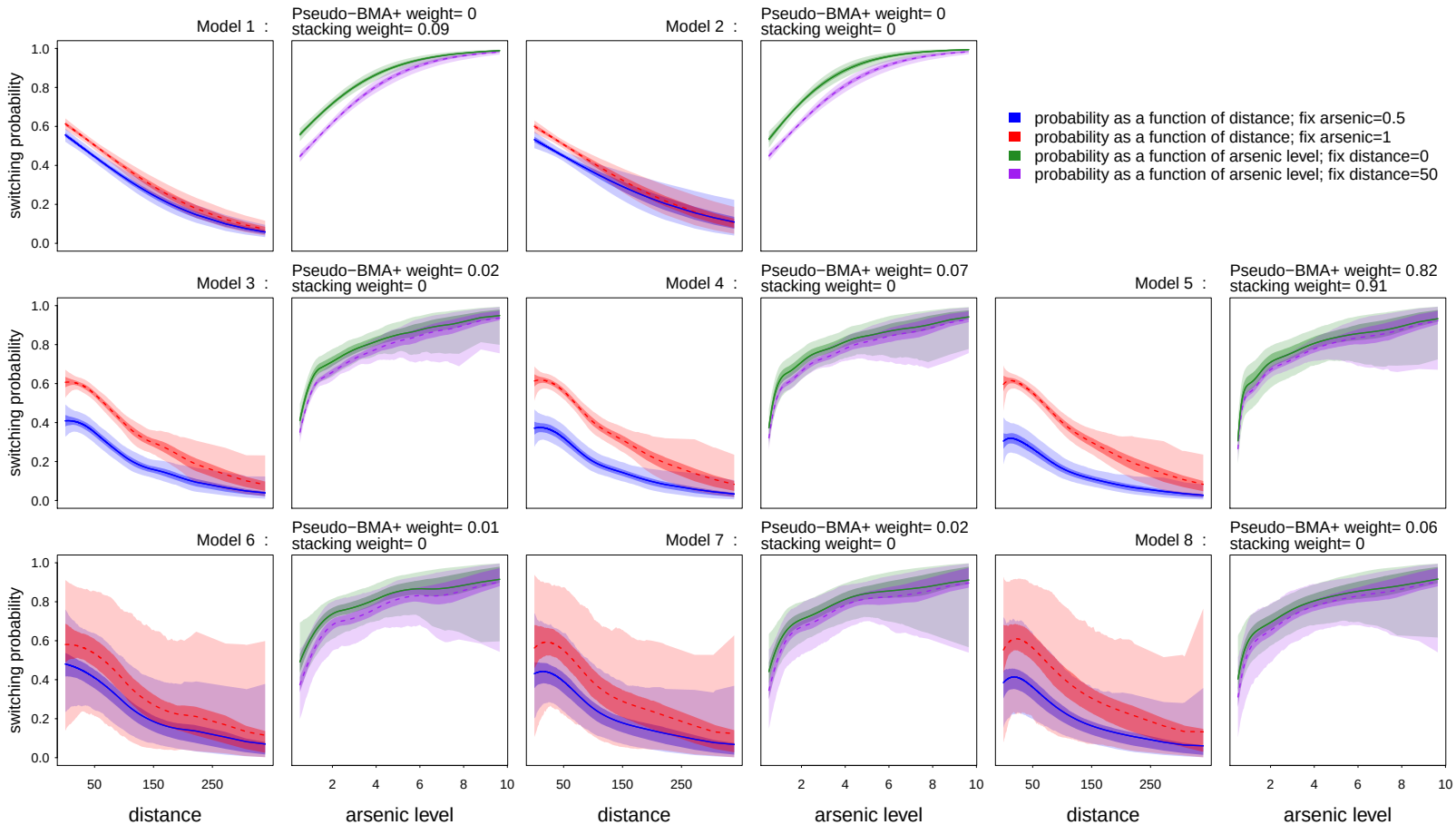
Linear subset regression example 1 : k



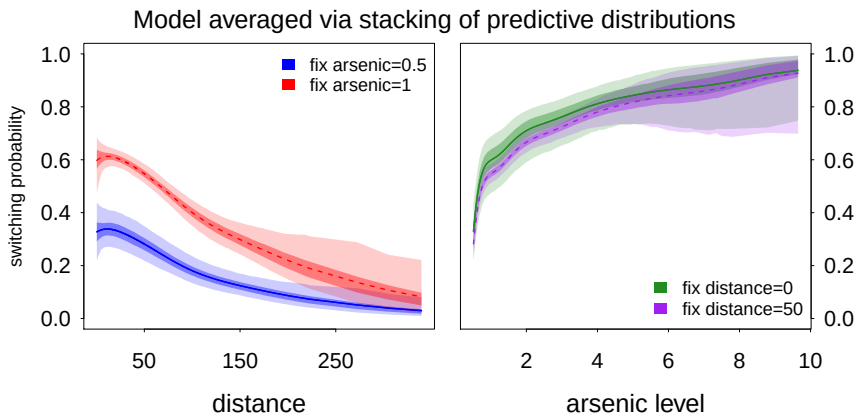
Variational multimodal example



Non-linear model example



Non-linear model example



Bayesian stacking

- In $\mathcal{M}$-open case works even better than BMA
- Yao, Vehtari, Simpson and Gelman (2017)

Additional references

Dupuis, J. A. and Robert, C. P.: 2003, Variable selection in qualitative models via an entropic explanatory power, *Journal of Statistical Planning and Inference* **111**(1-2), 77–94.

Goutis, C. and Robert, C. P.: 1998, Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections, *Biometrika* **85**(1), 29–37.

Lindley, D. V.: 1968, The choice of variables in multiple regression, *Journal of the Royal Statistical Society. Series B (Methodological)* **30**, 31–66.

Piironen, J. and Vehtari, A.: 2016, Comparison of Bayesian predictive methods for model selection, *Statistics and Computing* **27**(3), 711–735.

Vehtari, A. and Ojanen, J.: 2012, A survey of Bayesian predictive methods for model assessment, selection and comparison, *Statistics Surveys* **6**, 142–228.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A.: 2017, Using stacking to average Bayesian predictive distributions, *Bayesian analysis* . Accepted for publication, preprint arXiv:1704.02030.