Projection Predictive Model Selection For Gaussian Processes

Juho Piironen, Aki Vehtari

Helsinki Institute for Information Technology HIIT,
Department of Computer Science,
Aalto University, Finland
juho.piironen@aalto.fi, aki.vehtari@aalto.fi

Contents

Introduction

- Model target y with several input variables $\mathbf{x}$
- Only some of the inputs $\mathbf{x}$ relevant
  - Bayesian approach: use a relevant prior and integrate over all uncertainties

Introduction

- Model target y with several input variables $\mathbf{x}$
- Only some of the inputs $\mathbf{x}$ relevant
  - Bayesian approach: use a relevant prior and integrate over all uncertainties
  - Radford Neal won the NIPS 2003 feature selection competition using Bayesian methods with all the features (500 – 100 000)

Introduction

- Model target y with several input variables $\mathbf{x}$
- Only some of the inputs $\mathbf{x}$ relevant
  - Bayesian approach: use a relevant prior and integrate over all uncertainties
  - Radford Neal won the NIPS 2003 feature selection competition using Bayesian methods with all the features (500 – 100 000)
- Sometimes we want to select a minimal subset from $\mathbf{x}$ with a good predictive performance
  - improved model interpretability
  - reduced measurement costs in the future
  - reduced prediction time

Gaussian process (GP) regression

- GP-prior

$$f(\mathbf{x}) \sim \mathcal{GP}\big(0, k(\mathbf{x}, \mathbf{x}')\big)$$

- Observation model

$$\mathbf{y} \,|\, \mathbf{f} \sim \mathrm{N}\Big(\mathbf{y} \,|\, \mathbf{f}, \sigma^2 \mathbf{I}\Big)$$

- Predictive distribution

$$\begin{aligned}
\mathbf{f}_* \,|\, \mathbf{y} &\sim \mathrm{N}(\mathbf{f}_* \,|\, \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \\
\boldsymbol{\mu}_* &= \mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \\
\boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_*^{\mathsf{T}}.
\end{aligned}$$

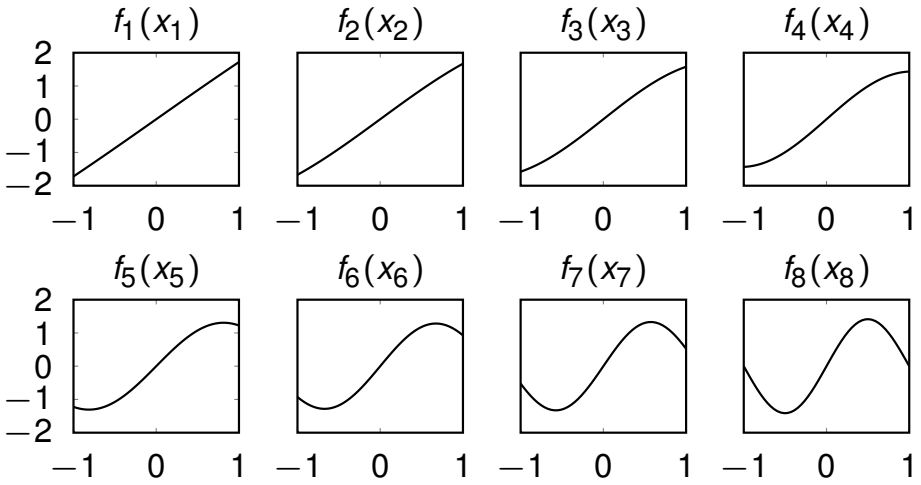
“Automatic relevance determination”

- Squared exponential (SE) or exponentiated quadratic covariance function

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{j=1}^{D}\frac{(x_j - x_j')^2}{\ell_j^2}\right).$$

- Use of separate length-scales ℓ_j for each input referred to as *automatic relevance determination* (ARD)
  - Idea: Optimizing marginal likelihood will yield large values ℓ_j for irrelevant inputs
  - Problem: Large length-scale may simply mean linearity w.r.t. the input (not irrelevance)

Toy example

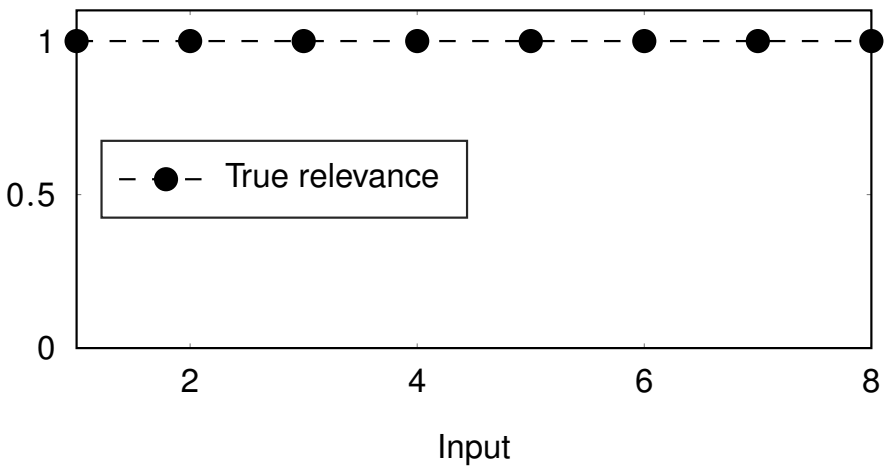


$$f(\mathbf{x}) = f_1(x_1) + \cdots + f_8(x_8),$$

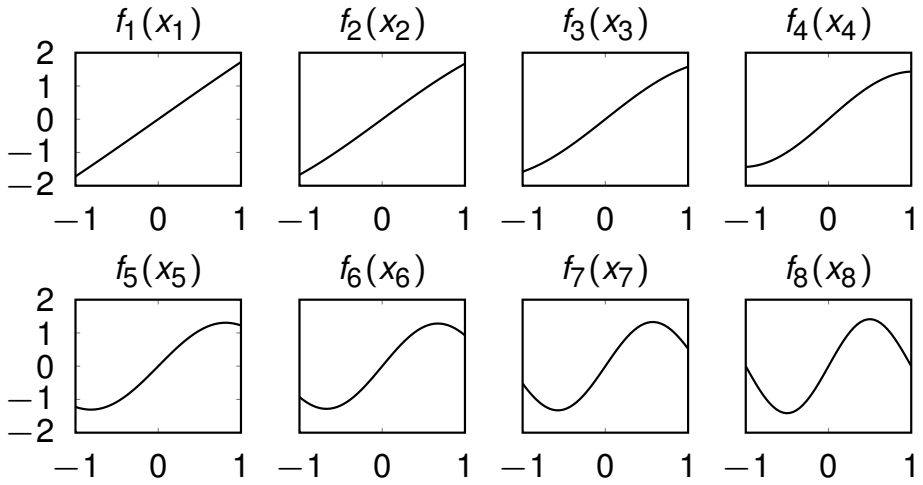
$$y \sim \mathrm{N}\left(f, 0.3^2\right),$$

$$\mathrm{Var}\left(f_j\right) = 1 \text{ for all } j.$$

$\Rightarrow$ All inputs equally relevant



Toy example

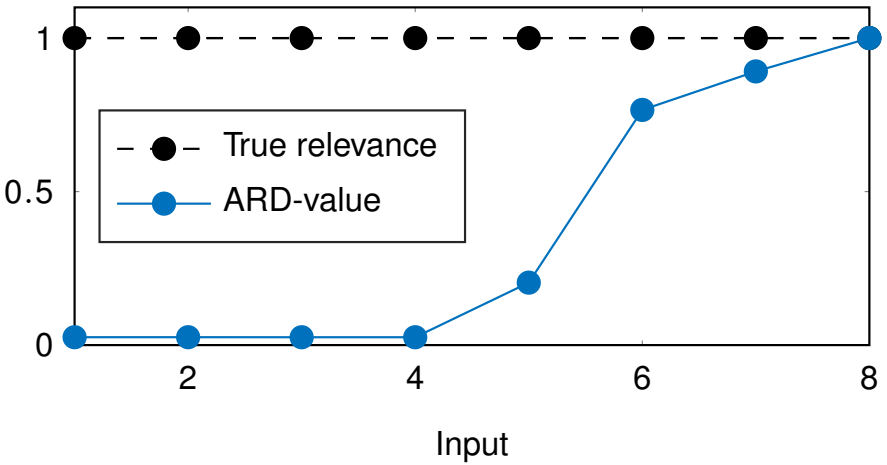


$$f(\mathbf{x}) = f_1(x_1) + \cdots + f_8(x_8),$$

$$y \sim \mathrm{N}\left(f, 0.3^2\right),$$

$$\mathrm{Var}\left(f_j\right) = 1 \text{ for all } j.$$

$\Rightarrow$ All inputs equally relevant



Optimized ARD-values, $\mathrm{ARD}(j) = 1/\ell_j$ (averaged over 100 data realizations, $n = 200$)

How about estimating the predictive performance?

- Cross-validation gives an (almost) unbiased estimate of the predictive performance
  - Fast LOO-CV approximations in
    Vehtari, Mononen, Tolvanen, Sivula, and Winther (2017). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. JMLR 17(103):1-38.

How about estimating the predictive performance?

- Cross-validation gives an (almost) unbiased estimate of the predictive performance
  - Fast LOO-CV approximations in
    Vehtari, Mononen, Tolvanen, Sivula, and Winther (2017). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. JMLR 17(103):1-38.
- But...

Selection induced bias in variable selection

- Even if the model performance estimate is unbiased (like LOO-CV), but it's noisy (like LOO-CV), then using it for model selection introduces additional fitting to the data

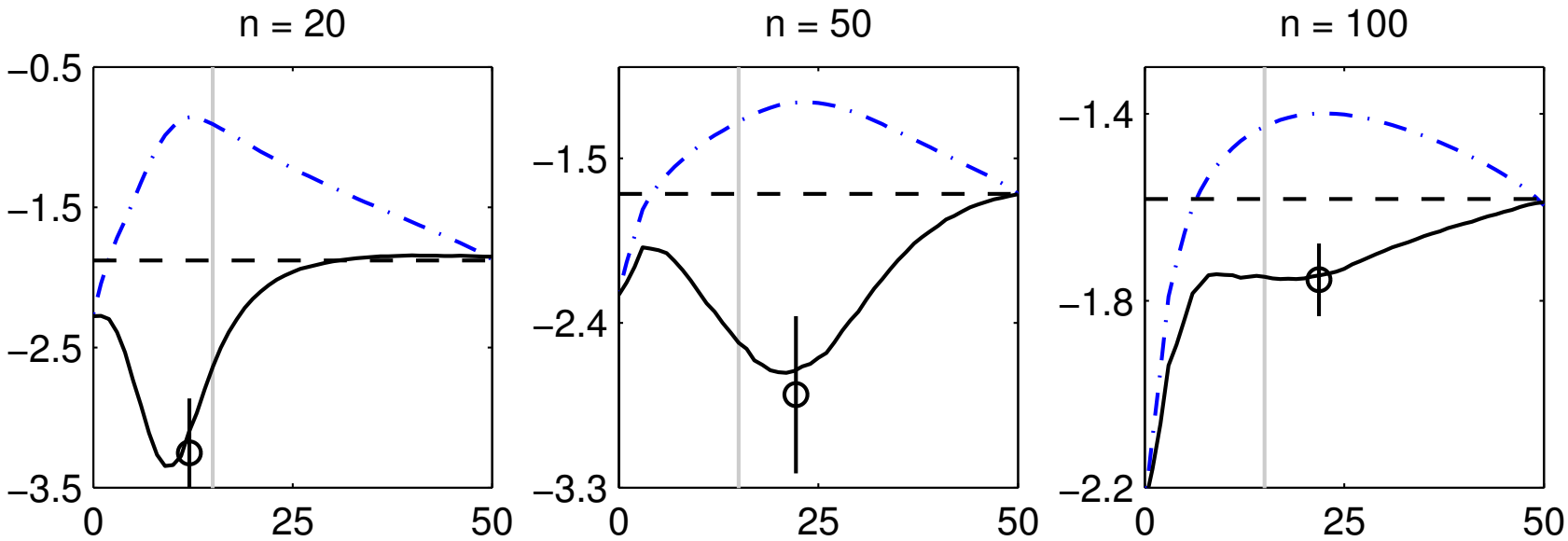
Selection induced bias in variable selection

- Even if the model performance estimate is unbiased (like LOO-CV), but it's noisy (like LOO-CV), then using it for model selection introduces additional fitting to the data
- Performance of the selection process itself can be assessed using two level cross-validation, but it does not help choosing better models

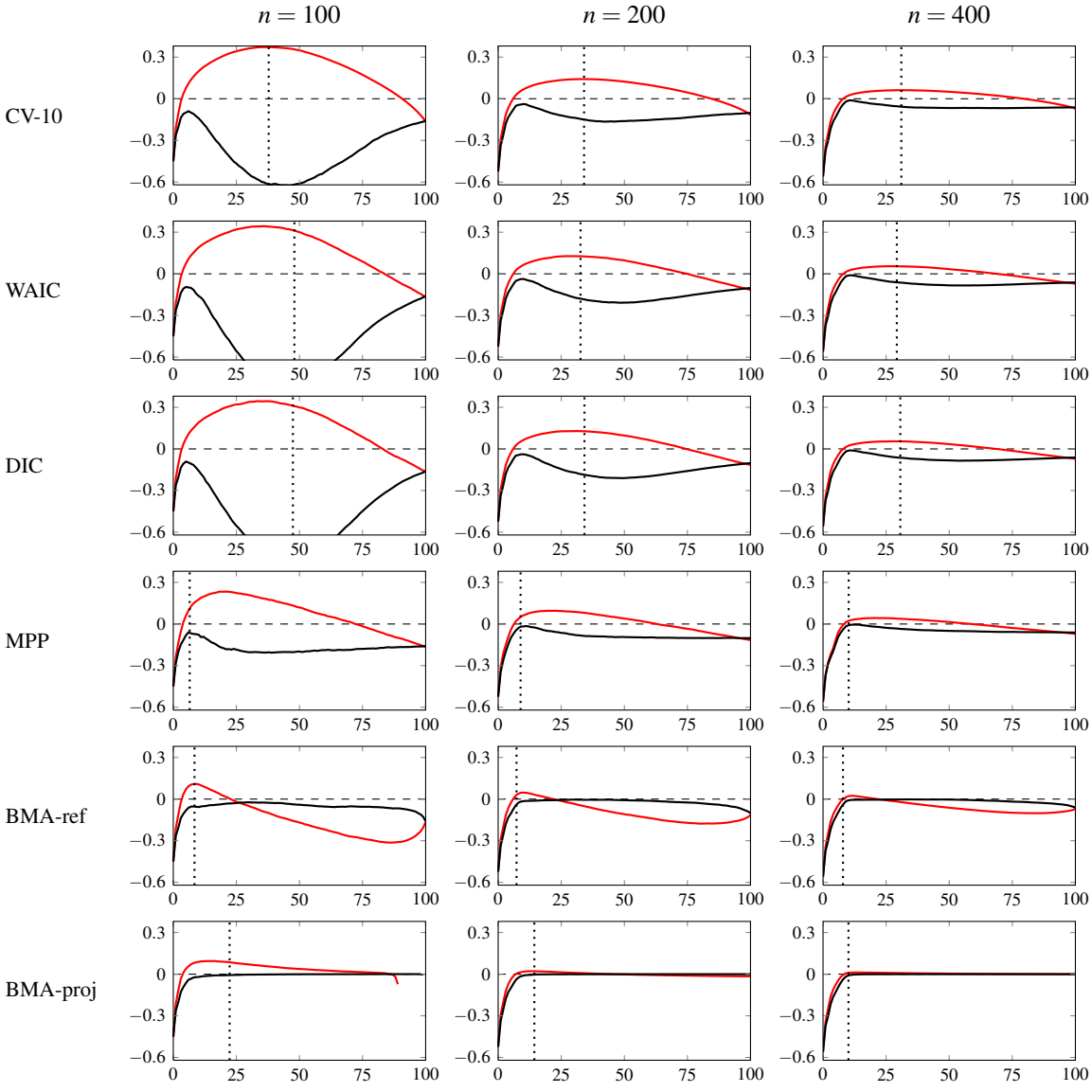
Selection induced bias in variable selection

- Even if the model performance estimate is unbiased (like LOO-CV), but it's noisy (like LOO-CV), then using it for model selection introduces additional fitting to the data
- 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
- Juho Piironen and Aki Vehtari (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711-735. doi:10.1007/s11222-016-9649-y. arXiv preprint arXiv:1503.08650.

Selection induced bias in variable selection

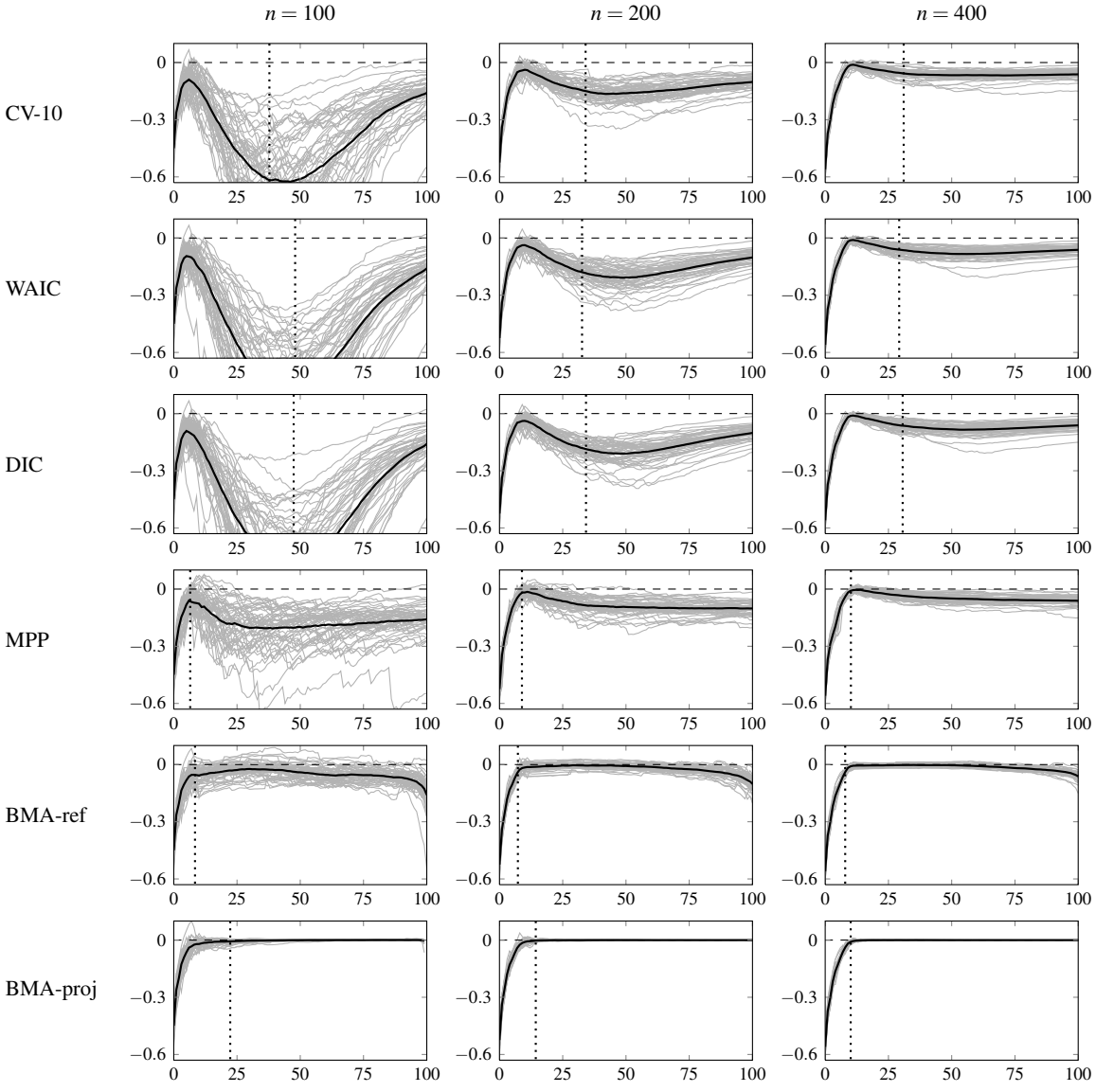


Selection induced bias in variable selection



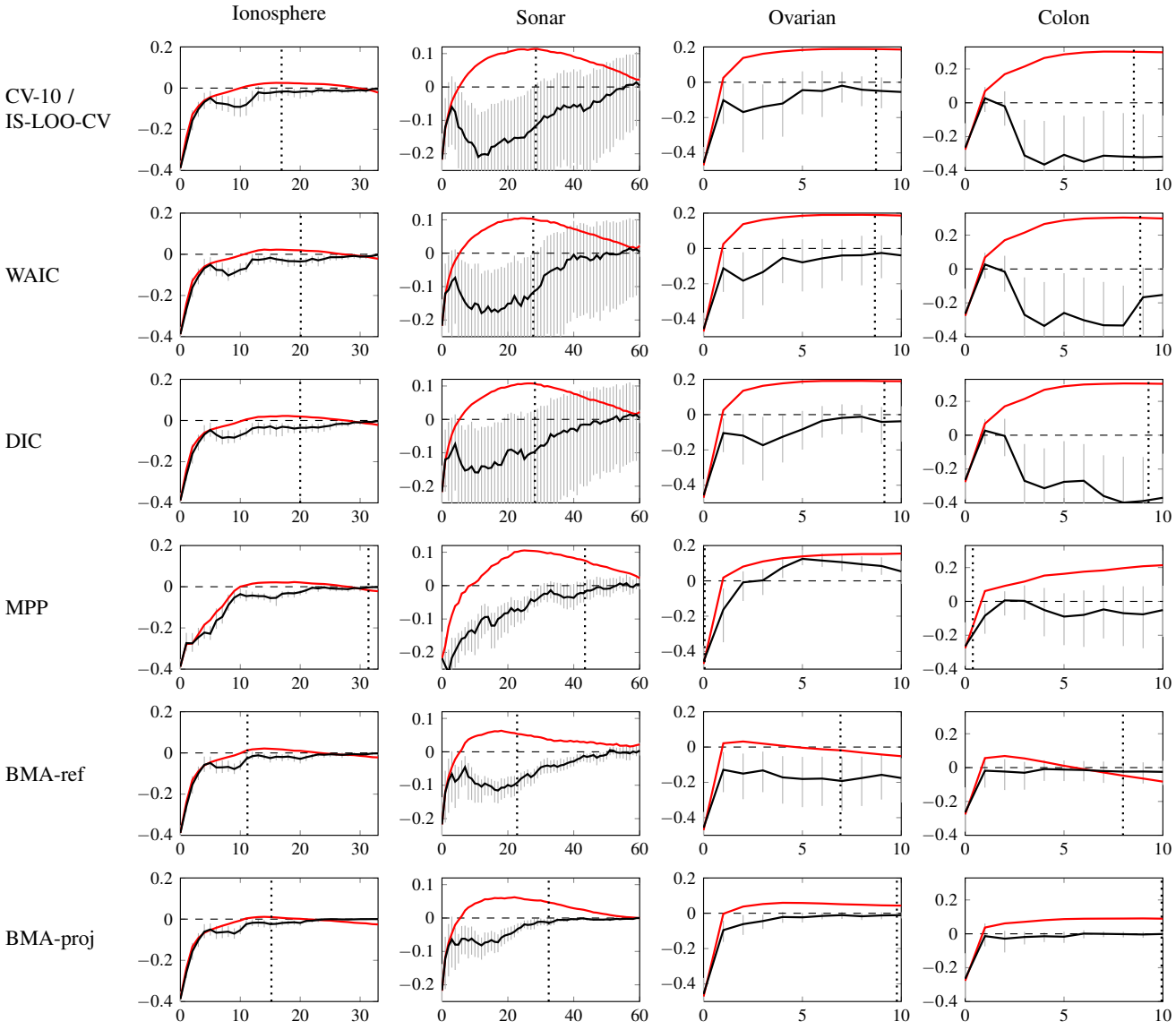
Piironen & Vehtari (2017)

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
  1. 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
  1. Fit the full encompassing model (with all the inputs) with best possible prior information
  2. 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):

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

- $\theta_\perp^s$ 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)

Projection predictive method for GPs

- The parameters of the GP are essentially the latent values $\mathbf{f}$ (and likelihood parameters like σ)
- Without constraints for the latent values in the submodel, the solution to the minimization problem is $\mathbf{f}_\perp = \mathbf{f}$

Projection predictive method for GPs

- The parameters of the GP are essentially the latent values $\mathbf{f}$ (and likelihood parameters like σ)
- Without constraints for the latent values in the submodel, the solution to the minimization problem is $\mathbf{f}_\perp = \mathbf{f}$
- We require constraint that that the submodel prediction satisfies the usual GP predictive equations

Projection predictive method for GPs

- Fit the full model M by learning the hyperparameters θ to obtain the latent fit $\mathbf{f} \mid \mathbf{y}, \theta \sim \mathrm{N}(\mathbf{f} \mid \mu_\theta, \mathbf{\Sigma}_\theta)$

Projection predictive method for GPs

- Fit the full model M by learning the hyperparameters θ to obtain the latent fit $\mathbf{f} \mid \mathbf{y}, \theta \sim \mathrm{N}(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\theta})$
- The projection to a submodel $M_{\perp}$ with fewer number of variables $D_{\perp}$ is obtained by solving

$$\delta(M||M_{\perp}) = \min_{\boldsymbol{\theta}_{\perp}}. \quad \mathrm{KL}\big(\mathrm{N}(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \,\big\|\, \mathrm{N}\big(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}_{\perp}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\perp}}\big)\big) \quad (1)$$

Projection predictive method for GPs

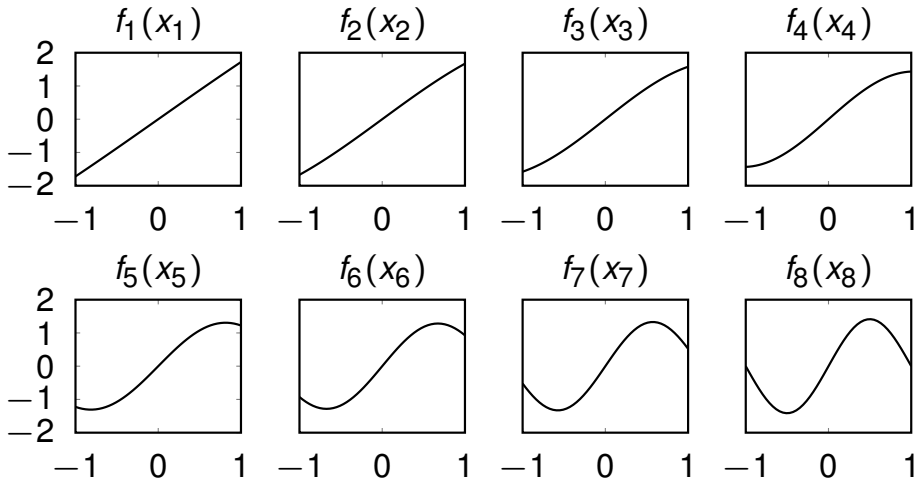
- Fit the full model M by learning the hyperparameters θ to obtain the latent fit $\mathbf{f} \mid \mathbf{y}, \theta \sim \mathrm{N}(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$
- The projection to a submodel $M_\perp$ with fewer number of variables $D_\perp$ is obtained by solving

$$\delta(M||M_\perp) = \min_{\boldsymbol{\theta}_\perp} . \quad \mathrm{KL}\big(\mathrm{N}(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \,\big\|\, \mathrm{N}\big(\mathbf{f} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}_\perp}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_\perp}\big)\big) \quad (1)$$

where

$$\mu_\perp = \mathbf{K}_\perp(\mathbf{K}_\perp + \sigma^2\mathbf{I})^{-1}\mathbf{y},$$
$$\boldsymbol{\Sigma}_\perp = \mathbf{K}_\perp - \mathbf{K}_\perp(\mathbf{K}_\perp + \sigma^2\mathbf{I})^{-1}\mathbf{K}_\perp,$$

Toy example

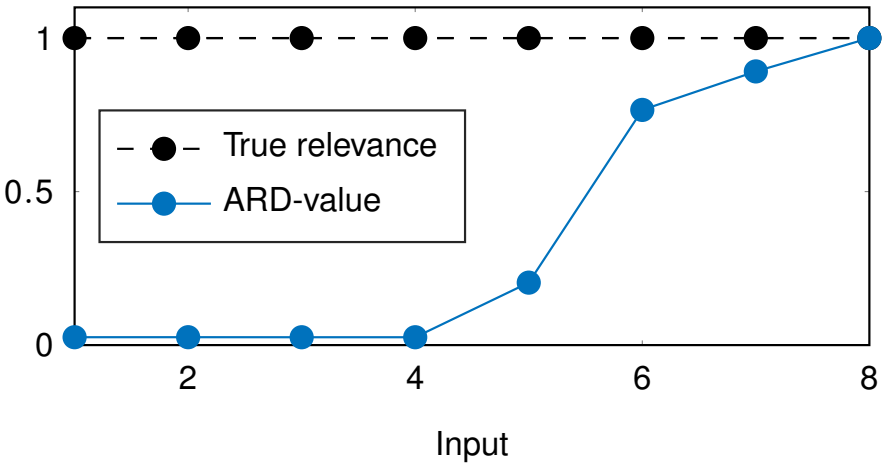


$$f(\mathbf{x}) = f_1(x_1) + \cdots + f_8(x_8),$$

$$y \sim \mathrm{N}\left(f, 0.3^2\right),$$

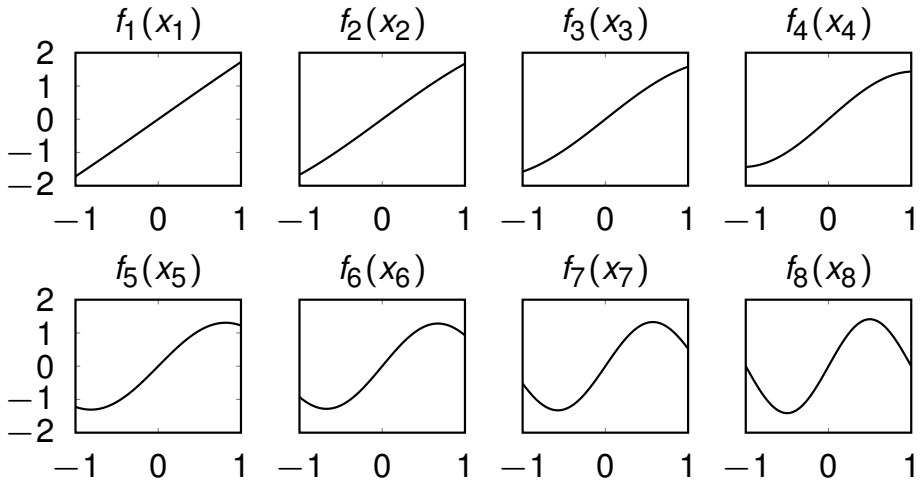
$$\mathrm{Var}\left(f_j\right) = 1 \text{ for all } j.$$

$\Rightarrow$ All inputs equally relevant



Optimized ARD-values, $\mathrm{ARD}(j) = 1/\ell_j$ (averaged over 100 data realizations, $n = 200$)

Toy example

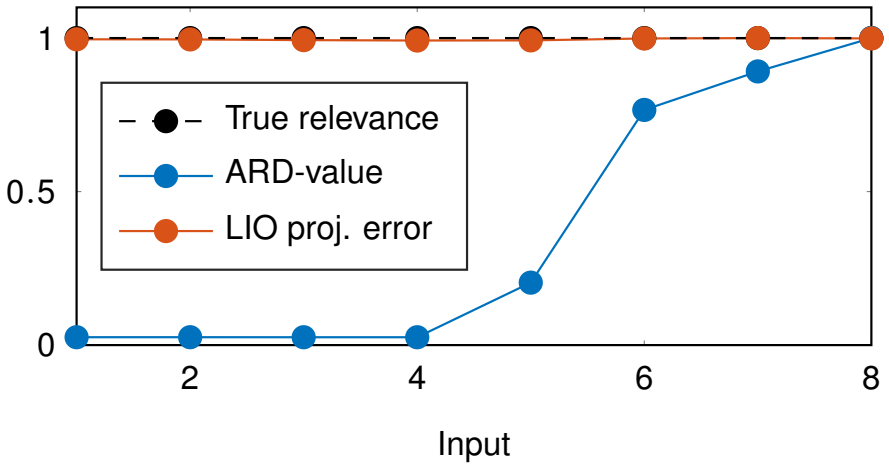


$$f(\mathbf{x}) = f_1(x_1) + \cdots + f_8(x_8),$$

$$y \sim \mathrm{N}\left(f, 0.3^2\right),$$

$$\mathrm{Var}\left(f_j\right) = 1 \text{ for all } j.$$

$\Rightarrow$ All inputs equally relevant



Leave-input-out (LIO) projection errors (averaged over 100 data realizations, $n = 200$)

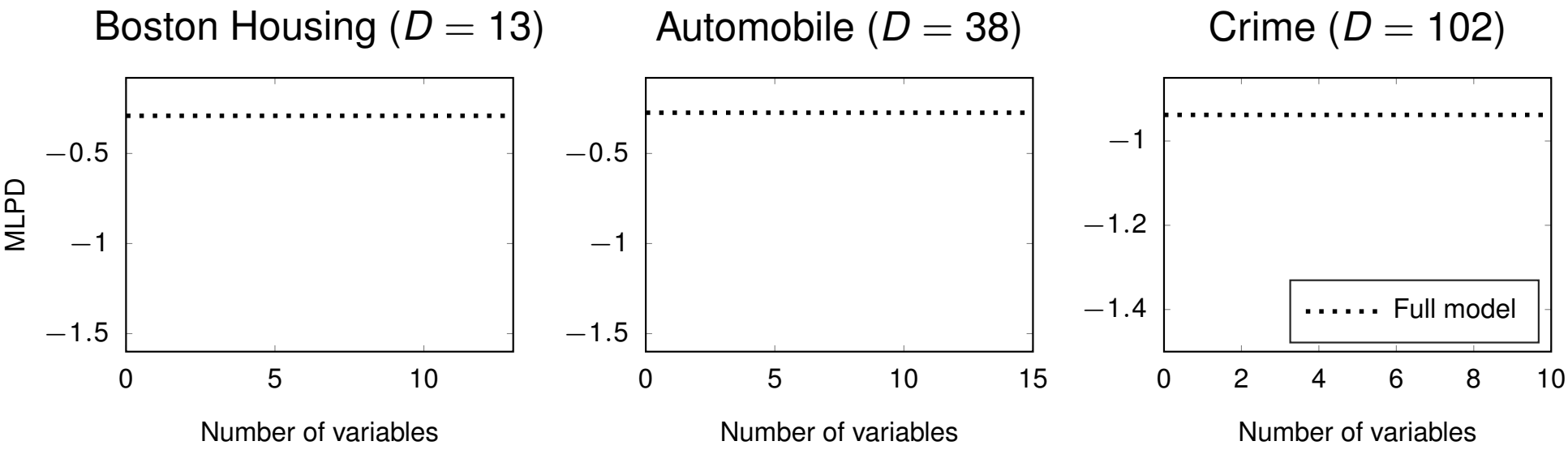
Projection predictive variable selection

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

Projection predictive variable selection

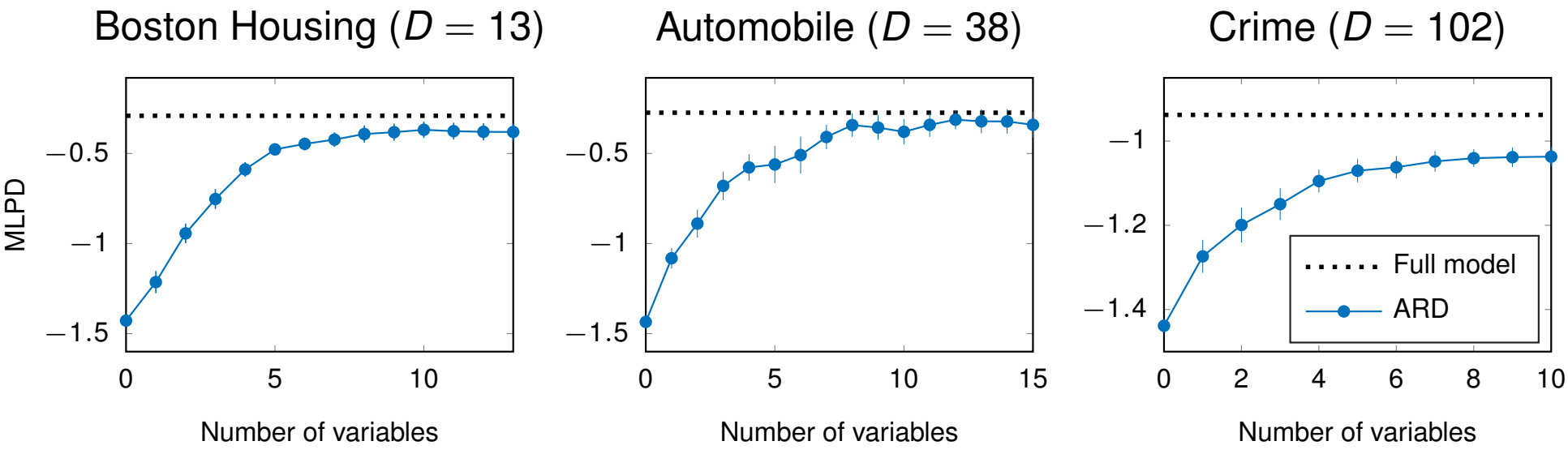
- In variable selection usually not feasible to go through all variable combinations
- Use e.g. forward search to explore promising combinations
  - start from the empty model, at each step add the variable that reduces the objective (1) the most
  - stop when the performance similar to the full model

Real world examples



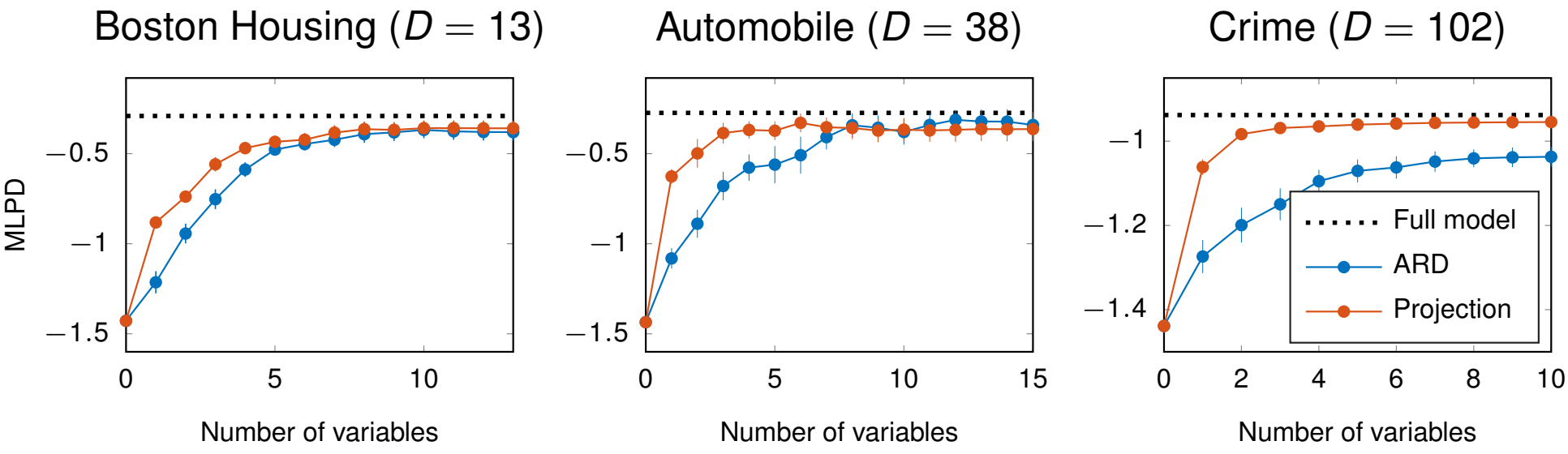
Mean log predictive density (MLPD) on test data for full model (all inputs) with sampled hyperparameters.

Real world examples



Accuracy for each submodel size, variables sorted by ARD (length-scales), hyperparameters optimized to maximum marginal likelihood.

Real world examples



Accuracy for each submodel size, variables sorted by stepwise minimization of projection error (forward search), hyperparameters learned via the projection.

Non-Gaussian likelihood

- Given Gaussian posterior approximation (e.g. obtained using EP), we can make the projection conditional on Gaussian likelihood approximations

Projection predictive method, pros and cons

- Advantage:
  - Discrepancy to the full model much more reliable indicator of submodel's performance than the length-scales
- Disadvantage:
  - Computational complexity for the projection is $O(n^3)$ (unless sparse approximations are used) $\Rightarrow$ slow if several submodels (e.g. variable combinations) are explored

Summary

- Carry out inference for the full model for best performance, select only if necessary
- ARD-values (length-scales) are unreliable for input relevance assessment
- Projection discrepancy to the full model is a more robust indicator
  - However, the forward search requires substantial amount of additional computations (in addition to fitting the full model)

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*. First online.

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