Variable selection for Gaussian processes using Kullback-Leibler projections

Jaakko Riihimäki* and Aki Vehtari

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Biomedical Engineering and Computational Science

*mailto: jaakko.riihimaki@tkk.fi

Outline

- Introduction
- Variable selection
  - Kullback-Leibler projections
- Gaussian process regression
  - Projections in regression
- Gaussian process binary classification and Expectation Propagation
  - Projections in classification
- Experiments
- Conclusions and future work

Introduction - motivation for variable selection

- In Bayesian inference, it is feasible to keep all the observed input variables in the model
  - Control the effects of variables with priors
  - Automatic Relevance Determination (ARD) prior: less relevant variables have smaller effects

Introduction - motivation for variable selection

- In Bayesian inference, it is feasible to keep all the observed input variables in the model
  - Control the effects of variables with priors
  - Automatic Relevance Determination (ARD) prior: less relevant variables have smaller effects

  HOWEVER:

- If a smaller set of variables is used in the model
  - Model easier to analyse
  - Measurement costs lower in the future
  - Savings in the computation time
  - Knowledge of variable relevances

Introduction - variable selection

- Model selection where nested models are considered

- A full (encompassing) probability model
  - Uses all available input variables
  - Assumption: reasonable and sufficient for the modelling problem under study

- Our goal is to find a submodel having predictive performance as close to the full model as possible
  - Using only a set of the most necessary input variables
  - If correlating variables, is there redundancy?
  - How to find a useful subset of variables?

Introduction - variable relevances

- The choice of variables based on variable relevances
  - Predictive comparison
  - Reversible jump Markov chain Monte Carlo (RJMCMC)
  - ARD prior?

- Advices from an application specialist?

- Using some relevance measure $\longrightarrow$ proposal for a subset

- Further inference conditioned only on a selected subset?

Introduction - variable relevances

- The choice of variables based on variable relevances
  - Predictive comparison
  - Reversible jump Markov chain Monte Carlo (RJMCMC)
  - ARD prior?

- Advices from an application specialist?

- Using some relevance measure $\longrightarrow$ proposal for a subset

- Further inference conditioned only on a selected subset?

UNFORTUNATELY:

- Omitting a set of variables may introduce a selection bias
  - Removal ignores the uncertainty related to the removed parts of data (using data twice)

Introduction - avoiding the selection bias

- Further inference conditioned also to the information that a set of variables are omitted

- Inference via joint distribution
  - (Lindley 1968)* presented how to do a choice of variables for Gaussian linear models by infering via joint distribution of input variables

- Inference via projections
  - Here the bias is avoided by doing inference via Kullback-Leibler (KL) projections introduced in (Goutis et al. 1998)†
  - In (Goutis et al. 1998) the projections for GLM, we apply the method for GP
  - We do a small modification in the classification models

---

*Lindley, D. V.: The choice of variables in multiple regression

†Goutis, C. and Robert, C.P.: Model choice in generalised linear models: A Bayesian approach via KL projections

Variable selection - definitions

- Training data: $D = \{(\mathbf{x}^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$
- A matrix $\mathbf{X}$ the entire set of observations(n rows, d columns), the targets $\mathbf{y}$ continuous in regression and binary valued in classification
- Model $p(\mathbf{y} \mid \mathbf{X})$
- Interested in the predictive distribution of y given a new input $\mathbf{x}^{(n+1)}$
- $\mathbf{X}_I$ the chosen variables, $\mathbf{X}_J$ the omitted variables
- Try to find a subset I that preserves the performance of the submodel $p(\mathbf{y} \mid \mathbf{X}_I)$ close to the performance of the full model $p(\mathbf{y} \mid \mathbf{X})$

Variable selection - inference via joint distribution

- The effect of removed variables $\mathbf{X}_J$ in further inference?
- Using the conditional distribution $p(\mathbf{X}_J \mid \mathbf{X}_I)$ the model can be written

$$p(\mathbf{y} \mid \mathbf{X}_I) = \int p(\mathbf{y} \mid \mathbf{X}_I, \mathbf{X}_J) p(\mathbf{X}_J \mid \mathbf{X}_I) d\,\mathbf{X}_J,$$

  where the uncertainty related to $\mathbf{X}_J$ explicitly modelled with the variables $\mathbf{X}_I$
- In (Lindley 1968), the choice of variables analytically for Gaussian linear models
  - suitable assumptions of $p(\mathbf{X}_J \mid \mathbf{X}_I) \rightarrow$ closed form computations
- Model $p(\mathbf{X}_J \mid \mathbf{X}_I)$ with Gaussian processes?

Variable selection - inference via projections

- Variable selection for GP using KL divergences is motivated by a Bayesian model choice method proposed in (Goutis et al. 1998)
  - No need to assume (or know) the conditional model $p(\mathbf{X}_J \mid \mathbf{X}_I)$

- The choice of variables is based on the evaluation of the Kullback-Leibler divergence between the full model and a submodel
  - Uncertainty related to the removed variables $\mathbf{X}_J$
  - Information from the removed variables

- By considering two distributions f_0 and f_1, the Kullback-Leibler divergence from f_0 to f_1 is given by

$$D_{\mathsf{KL}}(f_0 \| f_1) = \int f_0(x) \log \frac{f_0(x)}{f_1(x)} dx$$

Variable selection - inference via projections

- By writing a general probability model as $f(\cdot|\theta)$ where $\theta \in \Theta$, the projection of the parameter θ is defined as a point $\theta^{\perp}$ in Θ_0, where

$$D_{\mathrm{KL}}\{f(\cdot|\theta)\|f(\cdot|\theta^{\perp})\} = \inf_{\theta_0 \in \Theta_0} D_{\mathrm{KL}}\{f(\cdot|\theta)\|f(\cdot|\theta_0)\}$$

  is achieved (if nested: $\Theta_0 \subset \Theta$)

- Prior only needed for the full model parameters

- Samples from the posterior distribution of θ
  - For each parameter value, a projection with corresponding minimum divergence is computed
  - Average minimum divergence

Variable selection - inference via projections

- The choice of variables is then done by finding the most parsimonious submodel with an acceptable distance from the full model

- Where to set the acceptable distances?
  - Explanatory power of the full model (Dupuis et al. 2003)* $\longrightarrow$ changes the problem into a more interpretable form

- In (Goutis et al. 1998) and (Dupuis et al. 2003) the projection method is shown for generalised linear models

- We apply the projection method for models where a Gaussian process (GP) prior is set for latent function values

*Dupuis, J. A. and Robert, C. P.: Variable selection in qualitative models via an entropic explanatory power

Gaussian process

- Assuming the full model to be sufficient for modelling practices, a flexible GP model is a viable choice
  - No need to assume any functional forms in advance
  - Nonlinear effects
  - Implicit interactions between input variables

- The prior is set directly over functions of one or more input variables

- Mean and covariance functions define the nonparametric Gaussian process completely

- Given the training input points $\mathbf{X}$ and the targets $\mathbf{y}$, we associate latent values $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$ for each case

Gaussian process

- A zero-mean Gaussian process: a finite set of latent variables have the multivariate Gaussian distribution $p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$ as a prior

- The elements of a size n by n covariance matrix $\mathbf{K}$ are given by a covariance function
  - Covariances depend on the values of $\mathbf{X}$
  - The parameters $\boldsymbol{\theta}$ are hyperparameters of the GP model

- We use e.g. a squared exponential covariance function

$$\text{Cov}[f(\mathbf{x}_j), f(\mathbf{x}_k)] = \mathbf{K}_{jk} = \eta^2 \exp\left(-\sum_{i=1}^{d} \rho_i^2 (x_i^{(j)} - x_i^{(k)})^2\right),$$

  where highly correlated outputs are resulted if the input space distances between two cases are close (gives smooth solutions)

Gaussian process regression

- The marginal likelihood is given by the integral

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f} | \mathbf{X}, \boldsymbol{\theta}) d\mathbf{f}$$

- With Gaussian likelihood $p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2 \mathbf{I})$ an analytic solution is found: $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} \,|\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I})$ where $\boldsymbol{\theta} = (\eta^2, \boldsymbol{\rho}, \sigma^2)$

- The values of hyperparameters?
  - A point estimate
  - Integration numerically over the distribution with hybrid Monte Carlo (HMC) simulation (uses gradient information and a momentum parameter to avoid random walk behaviour)

- The predictive distribution for a new target $y^{(n+1)}$ is also Gaussian

KL projections for GP regression

- Latent variables can be integrated out analytically $\longrightarrow$ the marginal model

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} \,|\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I})$$

- We use Gaussian processes $\longrightarrow$ all the KL divergences needed in projections are between multivariate Gaussian distributions

- Denote the full marginal model as $\mathcal{N}_{\text{full}} = \mathcal{N}(\mathbf{0}, \mathbf{C}_{\text{full}})$ where $\mathbf{C}_{\text{full}} = \mathbf{K}_{\text{full}} + \sigma^2 \mathbf{I}$, and respectively the marginal submodel as $\mathcal{N}_{\text{sub}} = \mathcal{N}(\mathbf{0}, \mathbf{C}_{\text{sub}})$

- The KL divergence between two models simplifies to

$$D_{\text{KL}}(\mathcal{N}_{\text{full}} \| \mathcal{N}_{\text{sub}}) = \frac{1}{2} \left(\log_e \left(\frac{|\mathbf{C}_{\text{sub}}|}{|\mathbf{C}_{\text{full}}|} \right) + \text{tr}(\mathbf{C}_{\text{sub}}^{-1} \, \mathbf{C}_{\text{full}}) - n \right)$$

Gaussian process binary classification

- Inference with Gaussian processes becomes analytically intractable when likelihood is non-Gaussian

- We use a probit likelihood in binary classification $\longrightarrow$ approximations
  - Markov chain Monte Carlo (MCMC) sampling
  - Analytic approximations: Laplace or Expectation Propagation (EP)

- EP is used to approximate the distribution of latent values
  - Better than the Laplace's approximation
  - Solution close to the accuracy of MCMC integration

GP binary classification - Expectation Propagation

- An iterative algorithm for approximative inference
- Approximates the intractable posterior $p(\mathbf{f} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$ by a Gaussian distribution $q(\mathbf{f} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \rightarrow$ analytic treatment of the latent variables
- The posterior is approximated by

$$q(\mathbf{f} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\theta}) \prod_{i=1}^{n} \tilde{t}_i(f_i | \tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2),$$

  where $\tilde{t}_i(f_i | \tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \tilde{Z}_i \mathcal{N}(f_i | \tilde{\mu}_i, \tilde{\sigma}_i^2)$ are local likelihood approximations whose parameters $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ are site parameters
- The site parameters are successively updated until convergence
- An approximation for the marginal likelihood is given by EP
  - Hyperparameters for the full model (a point estimate/integration)

KL projections for GP binary classification

- The observations $\mathbf{y}$ are assumed to come from a latent phenomenon for which a Gaussian process prior is placed
- The distribution for latents given by the EP is $q(\mathbf{f} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma} = (\mathbf{K}^{-1} \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$
- $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are local likelihood approximation terms (site parameters)
- The KL divergence needed in projection is computed in latent space between the full model and the submodel assumed for the latent function values $\mathbf{f}$
- The KL divergence again is between two Gaussian distributions

KL projections for GP binary classification

- In order to allow an increase in uncertainty, an additional noise parameter for the GP classification model is introduced

- We assume an independent noise model between the latent variables in the submodel
  - To assure we do not underestimate the uncertainty caused by the removal

- One model for classification, another for uncertainty

- In regression there is a noise parameter in the submodel that allows an increase in uncertainty between latent variables
  - The noise term we use in classification corresponds this

Sparse approximation for large data sets

- The computation time with GP scales as $\mathcal{O}(n^3)$ (due to inversion)
- Approximate sparse GP method based on pseudo-inputs
  - the Fully Independent Conditional (FIC) approximation
- Based on a small set of additional latent variables $\mathbf{u}$ that induce the dependencies between all the other latent values
- The approximate prior becomes

$$q_{\mathrm{FIC}}(\mathbf{f} \mid \mathbf{X}_{\mathrm{u}}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\mathrm{f,f}} + \boldsymbol{\Lambda}),$$

  where $\boldsymbol{\Lambda} = \mathrm{diag}(\mathbf{K}_{\mathrm{f,f}} - \mathbf{Q}_{\mathrm{f,f}})$ and $\mathbf{Q}_{\mathrm{f,f}} = \mathbf{K}_{\mathrm{f,u}} \mathbf{K}_{\mathrm{u,u}}^{-1} \mathbf{K}_{\mathrm{u,f}}$
- In variable selection using projections, the FIC approximation changes the matrices $\mathbf{K}$ in the prior distributions

Searching input combinations

- How many input variables are chosen in the submodel?
- We do downward excursion starting from the full model
  - Remove one covariate at a time
  - The one that increases the divergence measured from the full model least
- The algorithm may produce a suboptimal solution
  - Try upward steps?
  - Use stochastic search?
  - ’Branch and bound’ algorithm
- The computation time vs. accuracy

Experiments - regression problem

- Friedman regression problem*
- Input variables $x_1, \ldots, x_{10} \sim U(0, 1)$, and targets are generated according to

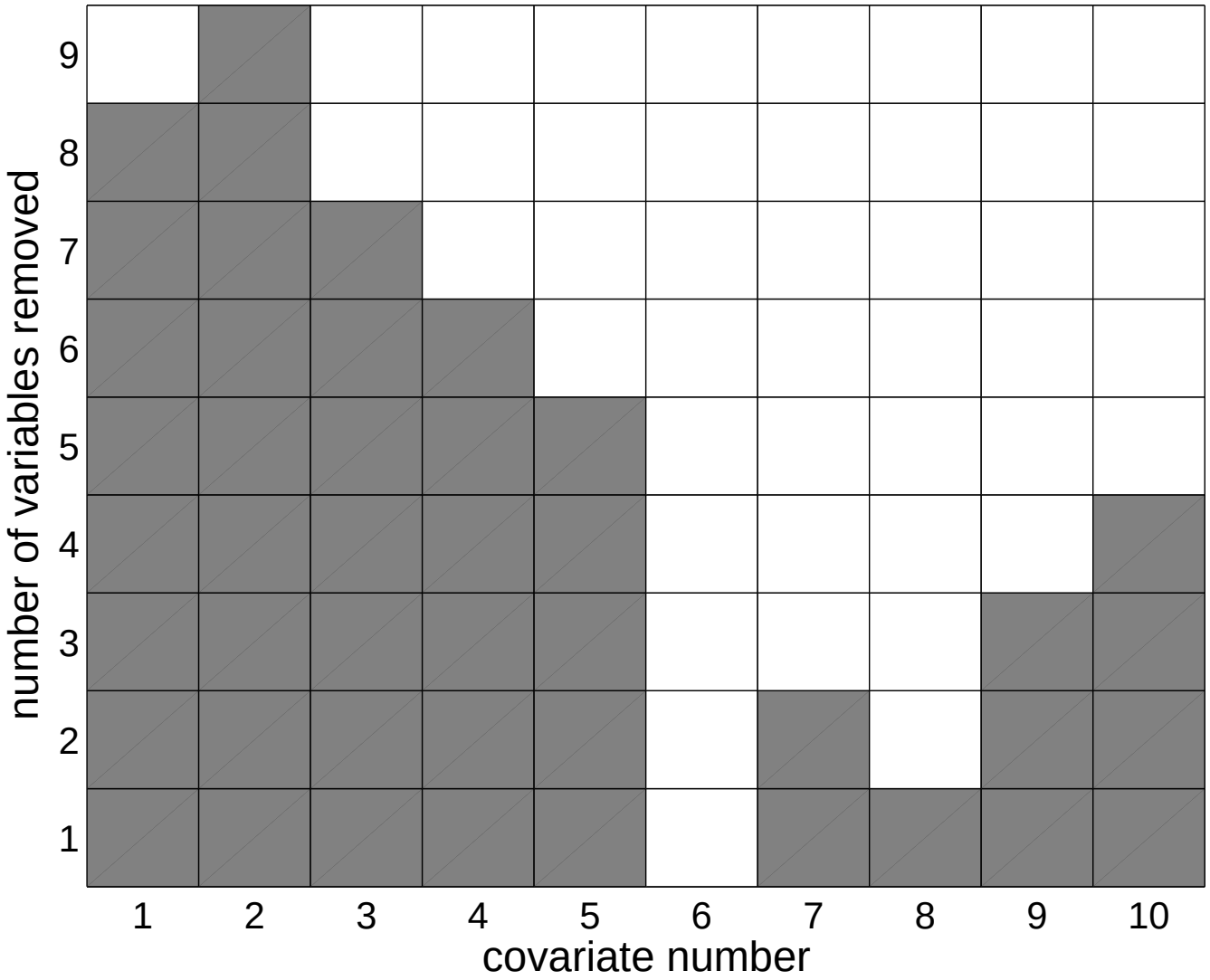
$$y = 10 \sin(\pi x_1 x_2) + 20 \left(x_3 - \frac{1}{2} \right)^2 + 10 x_4 + 5 x_5 + \varepsilon$$

  where $\varepsilon \sim \mathcal{N}(0, 1)$
- Only variables $x_1, \ldots, x_5$ are relevant
- We use 300 training points

* Friedman, J. H.: Multivariate adaptive regression splines

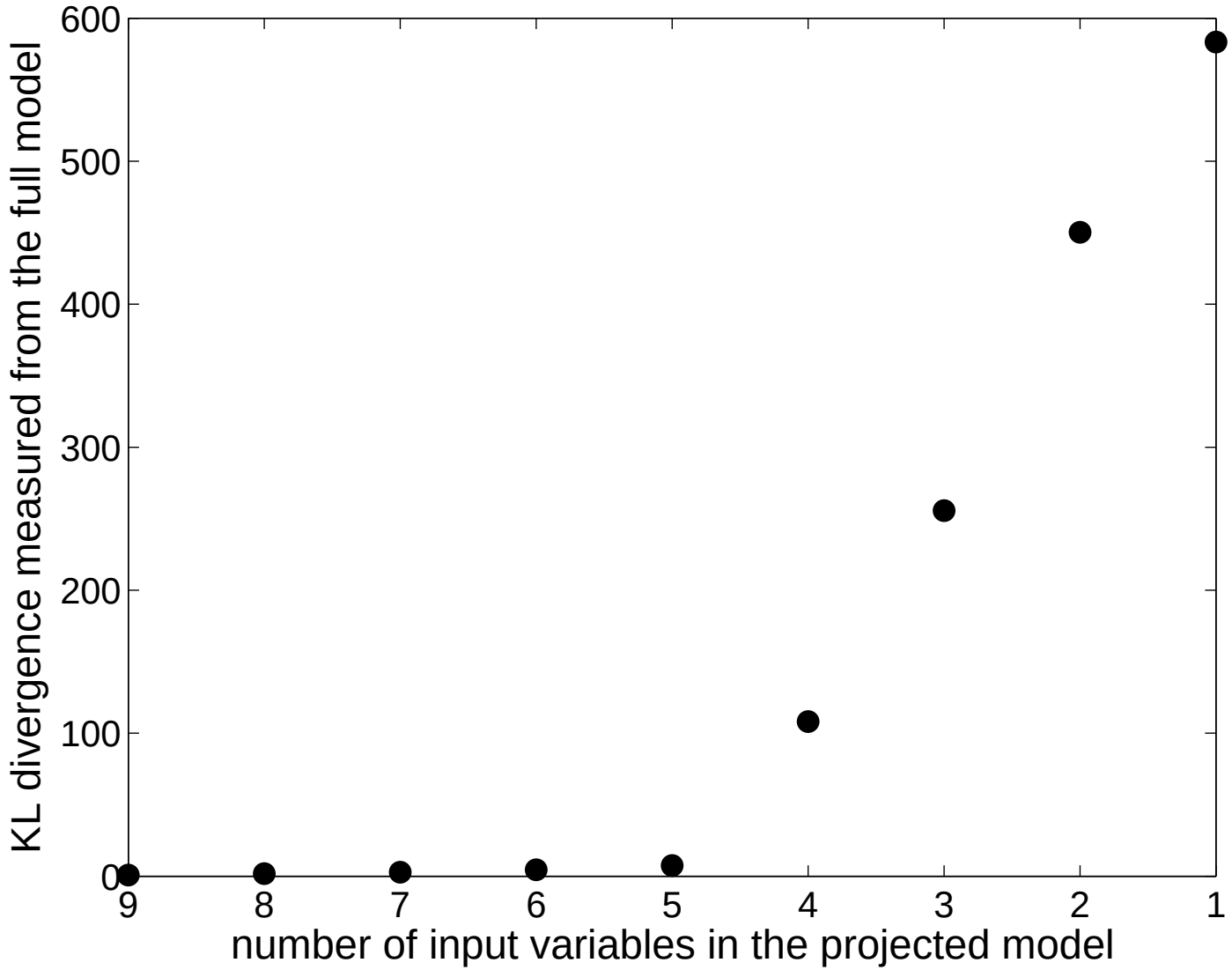
Experiments - regression problem

The result of the downward excursion (grey: chosen, white: removed)



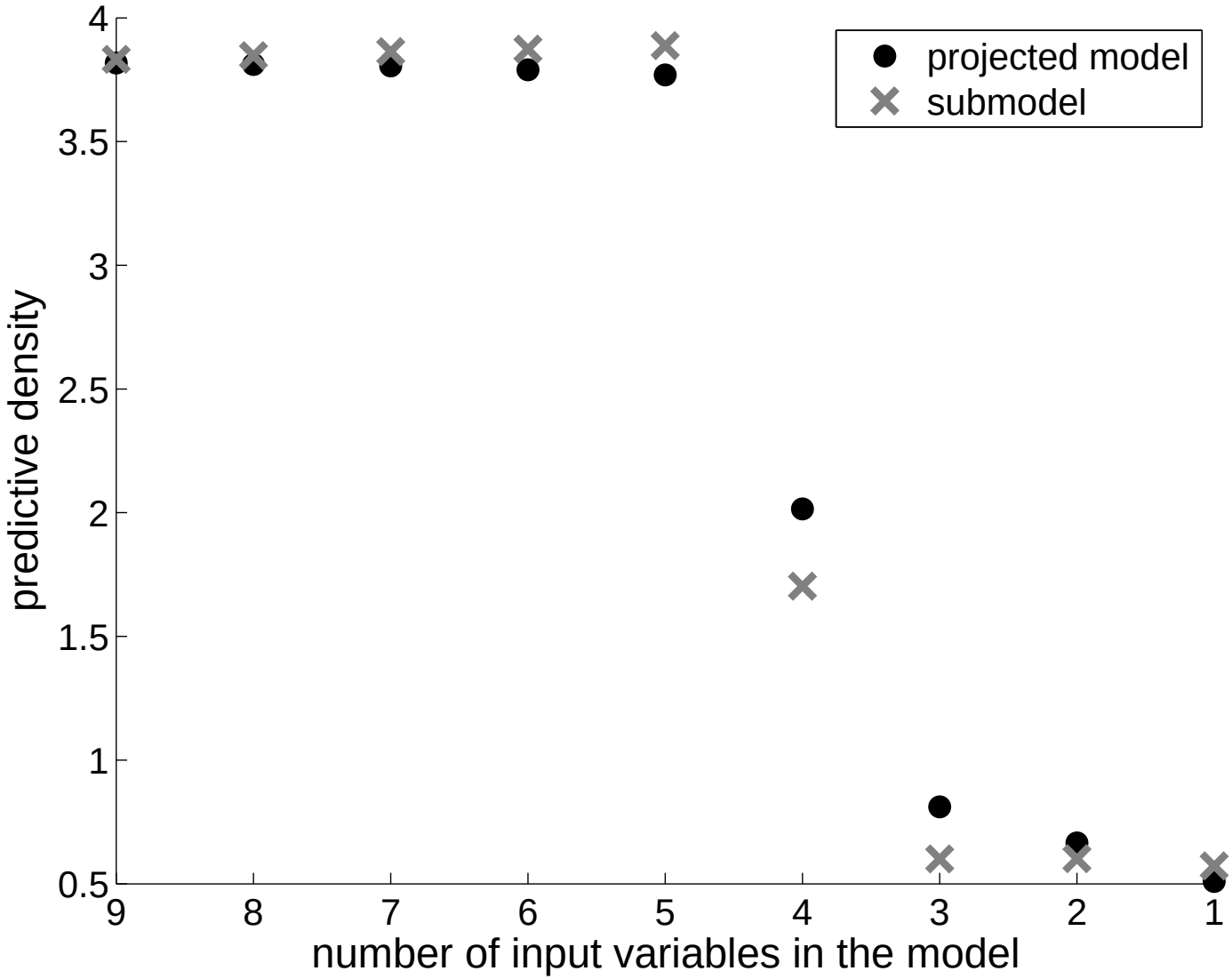
Experiments - regression problem

Kullback-Leibler divergences measured from the full model



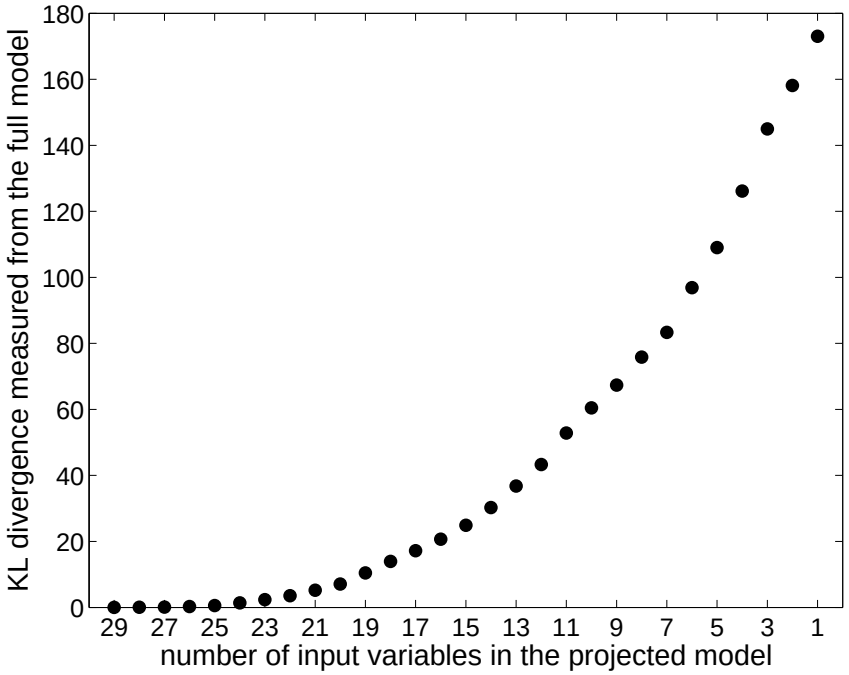
Experiments - regression problem

Predictive densities for a test data set



Experiments - classification problem

- Wisconsin Diagnostic Breast Cancer*
  - 569 observations, 30 input features, diagnosis: benign or malignant



Predictive densities

Covariates	Projected	Submodel
23	0.929	0.929
13	0.926	0.930
3	0.918	0.903

Predictive density for the full model: 0.930

* Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Conclusions and future work

- Preliminary experiments seem promising
  - Need to do more experiments with various data sets
- What is the practical significance of results?
- How to finally choose the number of input variables?
- In the projection method we need to form the full model
  - Problematic if large number of input variables

Conclusions and future work

- Preliminary experiments seem promising
  - Need to do more experiments with various data sets

- What is the practical significance of results?

- How to finally choose the number of input variables?

- In the projection method we need to form the full model
  - Problematic if large number of input variables

  HOWEVER:

- If at all possible use all available variables in the model and control the effects of them with priors!