On the reference predictive approach for covariate selection

ISBA Regional Meeting & International Workshop/Conference on Bayesian Theory and Applications, Varanasi, India, 2013

Aki Vehtari
Aki.Vehtari@aalto.fi

with Janne Ojanen

Department of Biomedical Engineering and Computational Science (BECS)
Aalto University

Outline

- Rich models and model selection
- Assessing predictive performance of models
- Bayesian predictive methods
- Selection induced bias
- Reference predictive approaches
- Comparison

Rich model vs. model selection

- Assume a model rich enough capturing lot of uncertainties
  - e.g. Bayesian model average (BMA) or non-parametric
  - model criticism and predictive assessment done

  $\rightarrow$ if we are happy with the model, no need for model selection

Rich model vs. model selection

- Assume a model rich enough capturing lot of uncertainties
  - e.g. Bayesian model average (BMA) or non-parametric
  - model criticism and predictive assessment done

  $\rightarrow$ if we are happy with the model, no need for model selection
  - Box: "All models are wrong, but some are useful"

Rich model vs. model selection

- Assume a model rich enough capturing lot of uncertainties
  - e.g. Bayesian model average (BMA) or non-parametric
  - model criticism and predictive assessment done

  $\rightarrow$ if we are happy with the model, no need for model selection
  - Box: "All models are wrong, but some are useful"
  - there are known unknowns and unknown unknowns

Rich model vs. model selection

- Assume a model rich enough capturing lot of uncertainties
  - e.g. Bayesian model average (BMA) or non-parametric
  - model criticism and predictive assessment done

  $\rightarrow$ if we are happy with the model, no need for model selection
  - Box: "All models are wrong, but some are useful"
  - there are known unknowns and unknown unknowns
- Model selection
  - what if some smaller (or more sparse) or parametric model is practically as good?
  - which uncertainties can be ignored?

  $\rightarrow$ reduced measurement cost, simpler to explain

Rich model vs. model selection

- Goodnes of the model is evaluated by its predictive performance
- Select a simpler model whose predictive performance is similar to the rich model

Predictive model

- $p(\tilde{y}|\tilde{x}, D, M_k)$ is the posterior predictive distribution
  - $p(\tilde{y}|\tilde{x}, D, M_k) = \int p(\tilde{y}|\tilde{x}, \theta, M_k)p(\theta|D, \tilde{x}, M_k)d\theta$
  - $\tilde{y}$ is a future observation
  - $\tilde{x}$ is a future random or controlled covariate value
  - $D = \{(x^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$
  - M_k is a model
  - θ denotes parameters

Predictive performance

- Future outcome $\tilde{y}$ is unknown (ignoring $\tilde{x}$ in this slide)
- With a known true distribution $p_t(\tilde{y})$, the expected utility would be

$$\bar{u}(a) = \int p_t(\tilde{y}) u(a; \tilde{y}) d\tilde{y}$$

where u is utility and a is action (in our case, a prediction)

Predictive performance

- Future outcome $\tilde{y}$ is unknown (ignoring $\tilde{x}$ in this slide)
- With a known true distribution $p_t(\tilde{y})$, the expected utility would be

$$\bar{u}(a) = \int p_t(\tilde{y}) u(a; \tilde{y}) d\tilde{y}$$

where u is utility and a is action (in our case, a prediction)

- Bayes generalization utility

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

where $a = p(\cdot|D, M_k)$ and $u(a; \tilde{y}) = \log(a(\tilde{y}))$

  - a is to report the whole predictive distribution
  - utility is the log-density evaluated at $\tilde{y}$

Bayesian predictive methods

- Many ways to approximate

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

  for example
  - Bayesian cross-validation
  - reference predictive methods

Bayesian predictive methods

- Many ways to approximate

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

  for example
  - Bayesian cross-validation
  - reference predictive methods
- Many other Bayesian predictive methods estimating something else, e.g.,
  - DIC
  - L-criterion, posterior predictive criterion
  - projection methods

Aki Vehtari and Janne Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. In Statistics Surveys, 6:142-228.

$\mathcal{M}$-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 $p_t(\tilde{y})$ is avoided by re-using the observed data D as a pseudo Monte Carlo samples from the distribution of future data
- For example, Bayes leave-one-out cross-validation

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

Cross-validation

- Bayes leave-one-out cross-validation

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

  - different part of the data is used to update the posterior and assess the performance
  - almost unbiased estimate for a single model

$$\mathrm{E}[\mathrm{LOO}(n)] = \mathrm{E}[BU_g(n-1)]$$

expectation is taken over all the possible training sets

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)
- Same holds for many other methods, e.g., DIC/WAIC

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)
- Same holds for many other methods, e.g., DIC/WAIC
- 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)
- Same holds for many other methods, e.g., DIC/WAIC
- 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

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Explicit model for $p_t(\tilde{y})$

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Explicit model for $p_t(\tilde{y})$
- $\mathcal{M}$-closed
  - possible to enumerate all possible model candidates $\{M_k\}_{k=1}^K$
  - belief that one of the candidate models is "true"
  - set a prior distribution $p(M_k)$ and compute $p_{\mathrm{BMA}}(\tilde{y}|D)$

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Explicit model for $p_t(\tilde{y})$
- $\mathcal{M}$-closed
  - possible to enumerate all possible model candidates $\{M_k\}_{k=1}^K$
  - belief that one of the candidate models is "true"
  - set a prior distribution $p(M_k)$ and compute $p_{\mathrm{BMA}}(\tilde{y}|D)$
- $\mathcal{M}$-completed
  - suitable when $\mathcal{M}$-closed can not be assumed
  - rich enough model M_* whose predictions are considered to best reflect the uncertainty in the prediction task

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Actual belief model M_*
  - a rich enough model, describing well the knowledge about the modeling problem and capturing the essential prior uncertainties
  - could be, for example
    - encompassing model
    - Bayesian model averaging model
    - flexible non-parametric model

$\mathcal{M}$-closed and $\mathcal{M}$-completed

- Actual belief model M_*
  - a rich enough model, describing well the knowledge about the modeling problem and capturing the essential prior uncertainties
  - could be, for example
    - encompassing model
    - Bayesian model averaging model
    - flexible non-parametric model
  - the predictive distribution of the actual belief model $p(\tilde{y}|\tilde{x}, D, M_*)$ is a quantitatively coherent representation of our subjective beliefs about the unobserved future data

Reference model and reference predictive approach

- Reference model
  - a model used to asses the predictive performance of other models
  - natural choice is the actual belief model M_*

Reference model and reference predictive approach

- Reference model
  - a model used to asses the predictive performance of other models
  - natural choice is the actual belief model M_*
- Reference predictive approach
  - predictive model assessment using a reference model

Unknown $\tilde{x}$

- $\mathcal{M}$-open for both $p(\tilde{y}|\tilde{x})$ and $p(\tilde{x})$
- Reference model for both $p(\tilde{y}|\tilde{x})$ and $p(\tilde{x})$
- Reference model for $p(\tilde{y}|\tilde{x})$ and $\mathcal{M}$-open for $p(\tilde{x})$

see our survey for discussion about fixed and deterministic x

Reference predictive approach

- Reference model for both $p(\tilde{y}|\tilde{x}, D, M_*)$ and $p(\tilde{x}|D, M_*)$
  - good model for $\tilde{x}$ may often be difficult to construct

Reference predictive approach

- Reference model for both $p(\tilde{y}|\tilde{x}, D, M_*)$ and $p(\tilde{x}|D, M_*)$
  - good model for $\tilde{x}$ may often be difficult to construct
- Lindley (1968)
  - use of linear Gaussian model for $y|x$ and squared error cost function made computations simpler
  - only first moments of x were needed

Reference predictive approach

- Reference model for $p(\tilde{y}|\tilde{x})$ and simple $\mathcal{M}$-open for $p(\tilde{x})$

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n} \int \log p(\tilde{y}|\dot{x}_i, D, M_k) p(\tilde{y}|\dot{x}_i, D, M_*) d\tilde{y}$$

Reference predictive approach

- Reference model for $p(\tilde{y}|\tilde{x})$ and simple $\mathcal{M}$-open for $p(\tilde{x})$

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n} \int \log p(\tilde{y}|\dot{x}_i, D, M_k) p(\tilde{y}|\dot{x}_i, D, M_*) d\tilde{y}$$

- San Martini & Spezzaferri (1984) used BMA model as the reference model

Reference predictive approach

- Reference model for $p(\tilde{y}|\tilde{x})$ and CV for $p(\tilde{x})$

$$\bar{u} \approx \bar{u}_*(M_k) = \frac{1}{n}\sum_{i=1}^{n} \int \log p(\tilde{y}|x_i, D_{-i}, M_k) p(\tilde{y}|x_i, D_{-i}, M_*) d\tilde{y}$$

  - better assessment of the out-of-sample predictive performance

Reference predictive approach

- Reference predictive model selection using log-score corresponds to minimizing the KL-divergence from the reference predictive distr. to the submodel predictive distr.

Reference predictive approach

- Reference predictive model selection using log-score corresponds to minimizing the KL-divergence from the reference predictive distr. to the submodel predictive distr.
  - divergence is minimized by the reference model itself
  - requires additional cost term or calibration of acceptable divergence from the reference

Reference predictive approach

- Reference predictive model selection using log-score corresponds to minimizing the KL-divergence from the reference predictive distr. to the submodel predictive distr.
  - divergence is minimized by the reference model itself
  - requires additional cost term or calibration of acceptable divergence from the reference
  - no selection induced bias, since data has been used only once to create the reference model, and selection process fits towards the reference model
  - bias depends on the reference model and is generally unknown

Reference predictive approach

- Reference predictive model selection using log-score corresponds to minimizing the KL-divergence from the reference predictive distr. to the submodel predictive distr.
  - divergence is minimized by the reference model itself
  - requires additional cost term or calibration of acceptable divergence from the reference
  - no selection induced bias, since data has been used only once to create the reference model, and selection process fits towards the reference model
  - bias depends on the reference model and is generally unknown
  - variance is reduced as model is used for $p(\tilde{y})$ instead of n pseudo Monte carlo samples
  - reduced variance helps discriminating good models from the others

Experimental results

Toy data with $n = 20$, 200 replications

$$
\begin{aligned}
z_1, z_2, z_3, z_4 &\sim \mathrm{U}(-1.73, 1.73) \\
x_{1,2,3,4} &\sim \mathrm{N}(z_1, .5^2) \\
x_{5,6,7,8} &\sim \mathrm{N}(z_2, .5^2) \\
x_{9,10,11,12} &\sim \mathrm{N}(z_3, .5^2) \\
x_{13,14,15,16} &\sim \mathrm{N}(z_4, .5^2) \\
y &= z_1 + .5z_2 + .25z_3 + \epsilon \\
\epsilon &\sim t_4(0, 0.5^2),
\end{aligned}
$$

that is, x's are noisy observations of z so that there are four groups of correlated covariates and four of the covariates have no effect on y

Experimental results

- Model

$$\tilde{y} = \sum_{j=1}^{16} \gamma_j \alpha_j x_j + e$$
$$\alpha_j \sim \mathrm{N}(0, \sigma_\alpha^2)$$
$$\sigma_\alpha \sim \text{Inv-}\chi^2(0.5, 0.5^2)$$
$$e_i \sim \mathrm{N}(0, \sigma_{e_i}^2)$$
$$\sigma_{e_i} \sim \text{Inv-}\chi^2(\nu, \sigma_e^2),$$

  where $\gamma_j = 1$ if covariate is included in the model and otherwise $\gamma_j = 0$

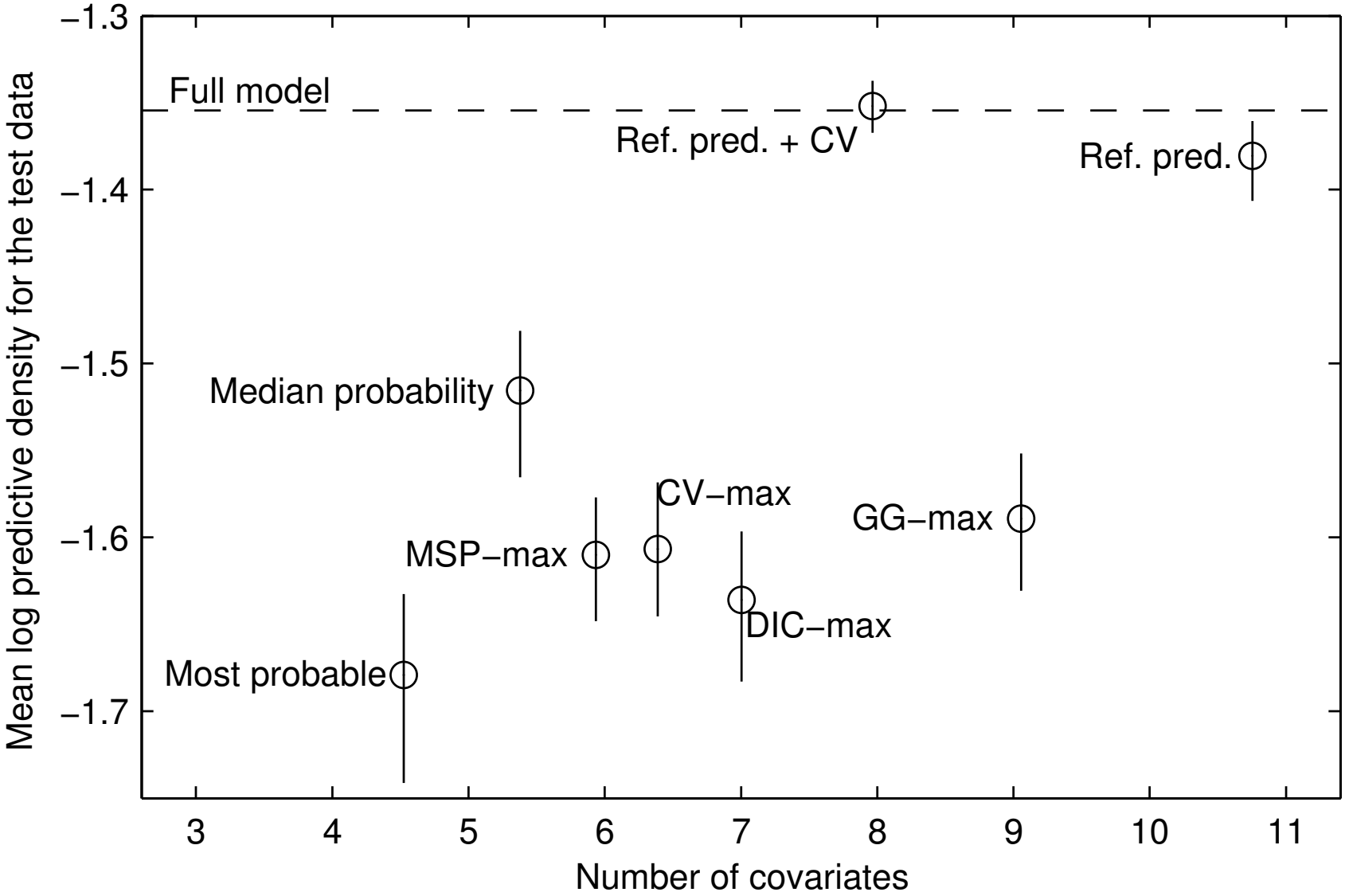
- The reference model is the full BMA model

Experimental results

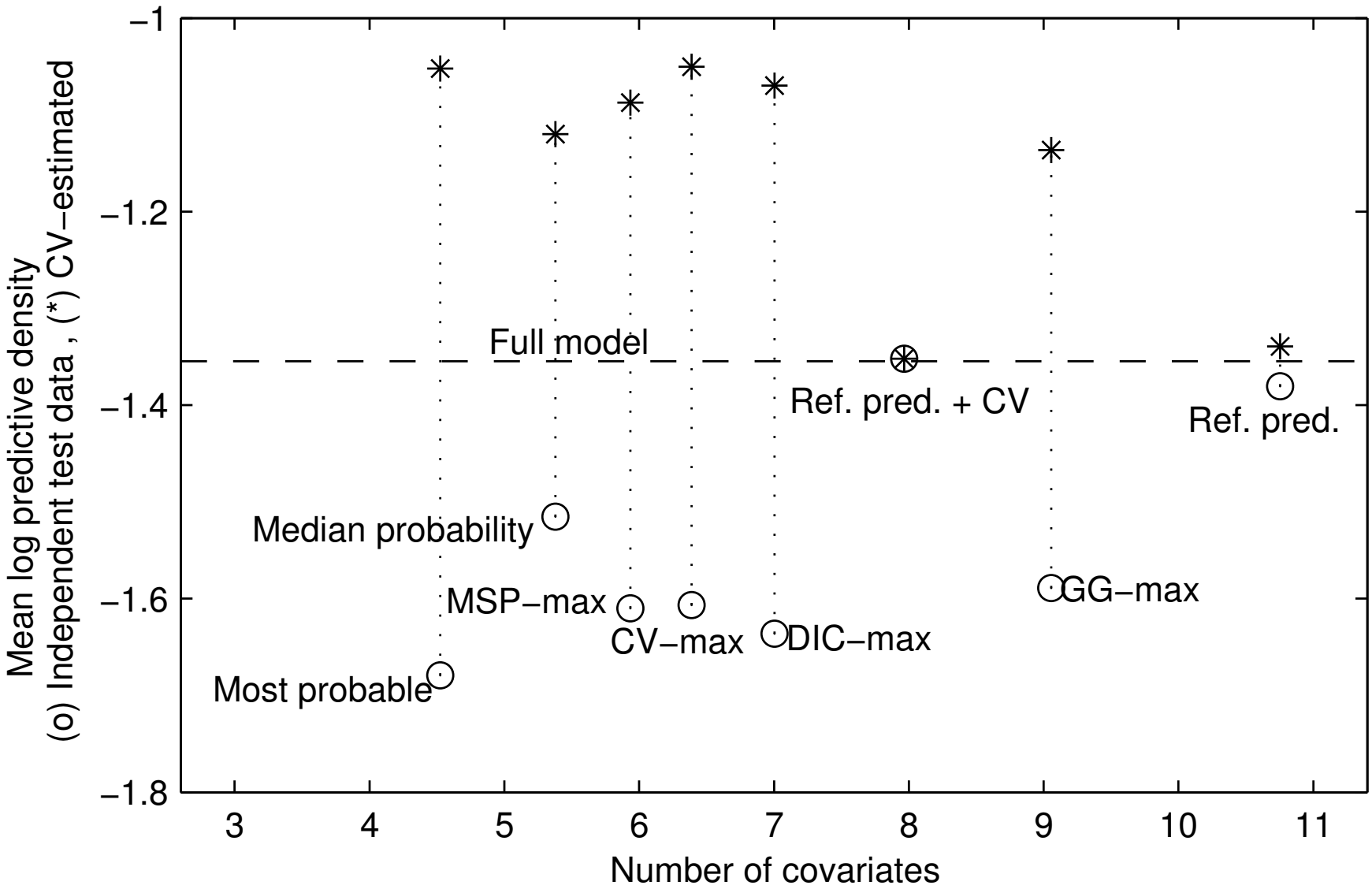
Methods compared

- reference predictive for $y|x$ (1% inform. loss)
- reference predictive for $y|x$ + CV for x (1% inform. loss)
- LOO-CV
- DIC (Spiegelhalter et al, 2002)
- posterior predictive loss = GG (Gelfand & Ghosh, 1998)
- cross-validation predictive loss = MSP (Marriot et al, 2001)
- most probable model, i.e. Bayes factor
- median probability model (Barbieri & Berger, 2004)

Experimental results



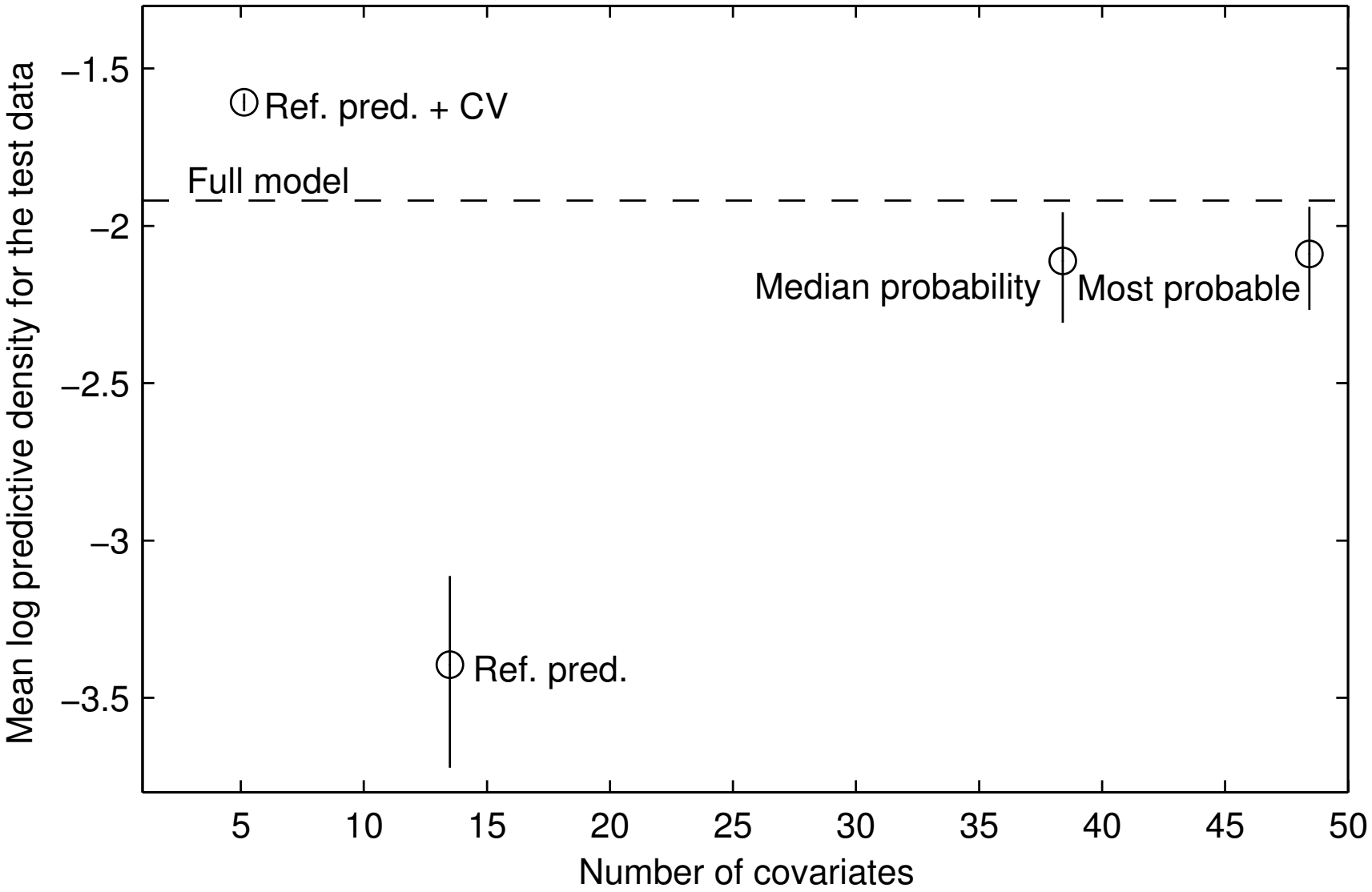
Experimental results



Experimental results

- Extended toy data with additional 84 irrelevant covariates

Experimental results



Conclusions

- Selection induced bias is a problem when there are many models (e.g. in covariate selection)
- Reference predictive approach with CV for x avoids selection induced bias

To be continued...

- Projection methods
- Gibbs utility
- Joint predictions
- Computational issues

See also

Aki Vehtari and Janne Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. In Statistics Surveys, 6:142-228.

Part 2: projection methods

Sheffield, 6th November 2013

Projection

- In projection predictive model selection the optimal prediction $\hat{a}_k$ under a candidate model M_k is the M_*-optimal prediction over $\mathcal{A}_k$, which is a set of possible predictions restricted by the model structure M_k
- The optimal prediction $\hat{a}_k$ is obtained by maximizing the expected utility

$$\hat{a}_k = \arg \max_{a_k \in \mathcal{A}_k} \int u(M_k, a_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}$$

Projection

- In projection predictive model selection the optimal prediction $\hat{a}_k$ under a candidate model M_k is the M_*-optimal prediction over $\mathcal{A}_k$, which is a set of possible predictions restricted by the model structure M_k
- The optimal prediction $\hat{a}_k$ is obtained by maximizing the expected utility

$$\hat{a}_k = \arg \max_{a_k \in \mathcal{A}_k} \int u(M_k, a_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}$$

- The resulting maximized expected utility is given by

$$\bar{u}(M_k, \hat{a}_k) = \int u(M_k, \hat{a}_k, \tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}$$

Projection

- The key component in the projection predictive approach is the definition of $\mathcal{A}_k$
- For example, in probabilistic prediction the space $\mathcal{A}_k$ can be restricted to parametric probability distributions $\{p(\tilde{y}|\theta_k, M_k) : \theta_k \in \Theta_k\}$, so that selecting the optimal prediction $\hat{a}(\tilde{y})$ becomes equal to selecting the optimal point estimate $\hat{\theta}$
- A major difference to the reference predictive approach is the possibility to avoid defining priors $p(\theta_k|M_k)$ for the candidate models M_k by treating the parameters of the candidate model as decision variables.

Projection

- Given a logarithmic utility function the optimal prediction $\hat{a}_k(\tilde{y})$ for the model M_k is determined by maximizing the expected utility

$$\bar{u}(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}$$

- The maximization is performed over the set of parametric models defined by M_k so that $a_k(\tilde{y}) \in \mathcal{A}_k = \{p(\tilde{y}|\theta_k, M_k) : \theta_k \in \Theta_k\}$

Projection

- Given a logarithmic utility function the optimal prediction $\hat{a}_k(\tilde{y})$ for the model M_k is determined by maximizing the expected utility

$$\bar{u}(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y}$$

- The maximization is performed over the set of parametric models defined by M_k so that
  $a_k(\tilde{y}) \in \mathcal{A}_k = \{p(\tilde{y}|\theta_k, M_k) : \theta_k \in \Theta_k\}$
- The maximization can be written equivalently in terms of the parameter as

$$\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} \int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}.$$

Projection

- The maximization can be written equivalently in terms of the parameter as

$$\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} \int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}.$$

- Given the optimal prediction $\hat{a}_k(\tilde{y}) = p(\tilde{y}|\hat{\theta}_k, M_k)$ the expected utility for the model M_k is

$$\bar{u}(M_k, \hat{\theta}_k) = \int \log p(\tilde{y}|\hat{\theta}_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}$$

Projection

- The maximization can be written equivalently in terms of the parameter as

$$\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} \int \log p(\tilde{y}|\theta_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}.$$

- Given the optimal prediction $\hat{a}_k(\tilde{y}) = p(\tilde{y}|\hat{\theta}_k, M_k)$ the expected utility for the model M_k is

$$\bar{u}(M_k, \hat{\theta}_k) = \int \log p(\tilde{y}|\hat{\theta}_k, M_k) p(\tilde{y}|D, M_*) d\tilde{y}$$

- In other words, the point estimate $\hat{\theta}_k$ is such that the parametric distribution $p(\tilde{y}|\hat{\theta}_k, M_k)$ is as close as possible to the posterior predictive distribution of the actual belief model in the KL divergence sense

Projection

- The M_*-optimal prediction is determined by maximizing the expected utility

$$\bar{u}(M_k, a_k) = \int \log a_k(\tilde{y}) p(\tilde{y}|D, M_*) d\tilde{y},$$

  where $a_k(\tilde{y}) \in \mathcal{A}_k$
- The definition $\mathcal{A}_k = \{\int p(\tilde{y}|\theta_k, M_k) q(\theta_k) d\theta_k : q \in \mathcal{Q}\}$ requires specifying a set of posterior projections $q(\theta_k)$ belonging to a suitable restricted family of probability distributions $\mathcal{Q}$
- For example, $\mathcal{Q}$ could consist of Gaussian distributions

Projection

- The expected utility maximization can be written in terms of $q(\theta_k)$, so that the optimal posterior projection is given by

$$\hat{q}(\theta_k) = \arg \max_{q(\theta_k) \in \mathcal{Q}} \int \log \left(\int p(\tilde{y}|\theta_k, M_k) q(\theta_k) d\theta_k \right) p(\tilde{y}|D, M_*) d\tilde{y}$$

- the corresponding maximized expected utility is defined as

$$\bar{u}(M_k, \hat{q}) = \int \log \left(\int p(\tilde{y}|\theta_k, M_k) \hat{q}(\theta_k) d\theta_k \right) p(\tilde{y}|D, M_*) d\tilde{y},$$

where $\hat{a}_k(\tilde{y}) = \int p(\tilde{y}|\theta_k, M_k) \hat{q}(\theta_k) d\theta_k$ is the optimal prediction

Projection

- The M_*-optimal posterior projection $\hat{q}(\theta_k)$ is not an approximation for the posterior distribution $p(\theta_k|D, M_k)$ of the model M_k.
- Instead, $\hat{q}(\theta_k)$ contains properties that are important in producing a prediction approximating the properties of the predictive distribution of the actual belief model.

Projection

- The M_*-optimal posterior projection $\hat{q}(\theta_k)$ is not an approximation for the posterior distribution $p(\theta_k|D, M_k)$ of the model M_k.
- Instead, $\hat{q}(\theta_k)$ contains properties that are important in producing a prediction approximating the properties of the predictive distribution of the actual belief model.
- For example, in input variable selection the prediction $\hat{a}_k(\tilde{y})$ may contain information about input variables included in the structure of M_* but not M_k; the standard Bayesian treatment of the model M_k would disregard all information about variables not included in the model M_k.

Projection

- Maximization with respect to the posterior projection $q(\theta_k)$ is not trivial, and we are not aware of any successful applications of this principle.
- See Vehtari & Ojanen (2012) for additional discussion.

Projection in GPs

- Instead of predictive distribution $p(\tilde{y}|\cdot)$ consider simpler predictive distribution $p(\tilde{f}|\cdot)$
  - joint predictive approach possible as $p(f|\cdot)$ is a multivariate Gaussian

Projection in GPs

- Instead of predictive distribution $p(\tilde{y}|\cdot)$ consider simpler predictive distribution $p(\tilde{f}|\cdot)$
  - joint predictive approach possible as $p(f|\cdot)$ is a multivariate Gaussian
- SVI-GP can be considered as a projection, where the model M_k is constrained to have a finite number of inducing points

Projection in GPs

- Instead of predictive distribution $p(\tilde{y}|\cdot)$ consider simpler predictive distribution $p(\tilde{f}|\cdot)$
  - joint predictive approach possible as $p(f|\cdot)$ is a multivariate Gaussian
- SVI-GP can be considered as a projection, where the model M_k is constrained to have a finite number of inducing points
- SVI-GP for covariate selection: constraint Z to live in a subspace of space where X live

Projection in GPs

- Instead of predictive distribution $p(\tilde{y}|\cdot)$ consider simpler predictive distribution $p(\tilde{f}|\cdot)$
  - joint predictive approach possible as $p(f|\cdot)$ is a multivariate Gaussian
- SVI-GP can be considered as a projection, where the model M_k is constrained to have a finite number of inducing points
- SVI-GP for covariate selection: constraint Z to live in a subspace of space where X live
- I guess that you would get similar results to what I showed above, but even better
  - the uncertainty related to the removed covariates is included in the projection

To be continued...

- Experiments to be done