Chapter 7

- 7.1 Measures of predictive accuracy
- 7.2 Information criteria and cross-validation
- 7.3 Model comparison based on predictive performance
- 7.4 Model comparison using Bayes factors
- 7.5 Continuous model expansion / sensitivity analysis
- 7.5 Example (may be skipped)

Predictive performance

- True predictive performance is found out by using it to make predictions and comparing predictions to true observations
  - external validation
- Expected predictive performance
  - approximates the external validation

Predictive performance

- We need to choose the utility/cost function
- Application specific utility/cost functions are important
  - eg. money, life years, quality adjusted life years, etc.

Predictive performance

- We need to choose the utility/cost function
- Application specific utility/cost functions are important
  - eg. money, life years, quality adjusted life years, etc.
- If are interested overall in the goodnes of the predictive distribution, or we don't know (yet) the application specific utility, then good information theoretically justified choice is log-score

$$\log p(y^{\mathrm{rep}}|y, M),$$

Ways to estimate the predictive performance

- Data estimate (within-sample)
  - use same data to form the posterior and to test the performance
  - corresponds to “training error”

Ways to estimate the predictive performance

- Data estimate (within-sample)
  - use same data to form the posterior and to test the performance
  - corresponds to "training error"

- Partial predictive
  - split data in two parts
  - use one part to form the posterior and the other part to test the performance
  - corresponds to "test error"

Ways to estimate the predictive performance

- Data estimate (within-sample)
  - use same data to form the posterior and to test the performance
  - corresponds to “training error”
- Partial predictive
  - split data in two parts
  - use one part to form the posterior and the other part to test the performance
  - corresponds to “test error”
- Cross-validation
  - improved version of partial approach
  - divide data in several parts, can be also n parts
  - use different parts to form posterior and to test performance

Ways to estimate the predictive performance

- Data estimate (within-sample)
  - use same data to form the posterior and to test the performance
  - corresponds to “training error”
- Partial predictive
  - split data in two parts
  - use one part to form the posterior and the other part to test the performance
  - corresponds to “test error”
- Cross-validation
  - improved version of partial approach
  - divide data in several parts, can be also *n* parts
  - use different parts to form posterior and to test performance
- Information criterion

Why 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

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

Why 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

Why 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
  - → 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?
    (e.g. Student-t vs. Gaussian, irrelevant covariates)

Why 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
  - → 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?
    (e.g. Student-t vs. Gaussian, irrelevant covariates)
  - → reduced measurement cost, simpler to explain

Why 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
  - → 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? (e.g. Student-*t* vs. Gaussian, irrelevant covariates)
  - → reduced measurement cost, simpler to explain (e.g. less biomarkers, and easier to explain to doctors)

Predictive 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 $\tilde{y}$

- 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
  - WAIC
  - reference predictive methods (* not in the course)

- See Aki Vehtari and Janne Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6:142-228, http://dx.doi.org/10.1214/12-SS102 for other methods.

$\mathcal{M}$-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

- 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

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

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

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

- LOO predictive density evaluated at $\mathbf{y}_i$

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

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

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

where w_i^s are importance weights

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

Truncated importance sampling

- The variance of the importance weights w^s in IS-LOO can be large or even infinite
- Truncated importance sampling with truncated weights

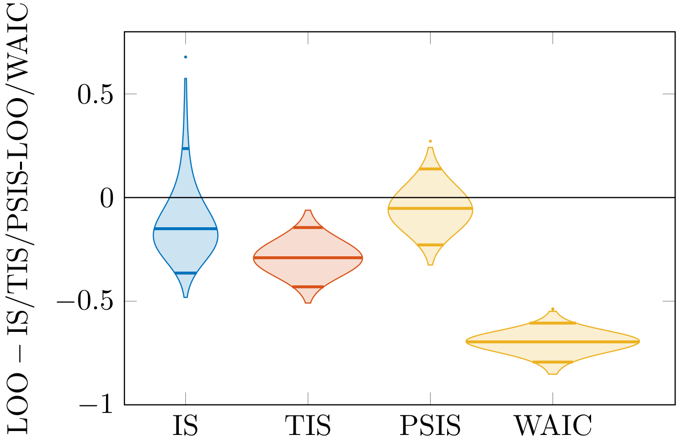
$$\tilde{w}^s = \min(\tilde{w}^s, \sqrt{S}\bar{w})$$

has a finite variance but also some optimistic bias

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

Pareto smoothed importance sampling



Aki Vehtari, Andrew Gelman and Jonah Gabry (2016). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544. http://arxiv.org/abs/1507.04544

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 shoud be used
  - if PSIS-LOO diagnostics indicate problems with importance sampling
  - if the prediction task is for groups

Widely applicable information criterion

- Bayes generalization utility

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

- Bayes training utility

$$BU_t = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i, D, M_k)$$

  - biased (overoptimistic) estimate of BU_g
- Information criteria approach considers a bias correction to this, to get unbiased estimate of
- Bias correction in information criteria is related to the effective number of parameters

Widely applicable information criterion

- Watanabe (2009,2010abc) proposed Widely applicable information criterion (WAIC)
  - WAIC has two alternative approximations

$$\mathrm{WAIC}_G = BU_t - 2(BU_t - GU_t)$$
$$\mathrm{WAIC}_V = BU_t - V/n$$

where GU_t is Gibbs utility

$$GU_t = \frac{1}{n}\sum_{i=1}^{n}\int p(\theta|D, M_k)\log p(y_i|x_i, \theta, M_k)d\theta$$

and V is functional variance

$$V = \sum_{i=1}^{n}\left\{ \mathrm{E}_{\theta|D,M_k}\left[(\log p(y_i|x_i, \theta, M_k))^2\right] - \left(\mathrm{E}_{\theta|D,M_k}[\log p(y_i|x_i, \theta, M_k)]\right)^2 \right\}$$

Widely applicable information criterion

- WAIC has two alternative approximations

$$\mathrm{WAIC}_G = BU_t - 2(BU_t - GU_t)$$
$$\mathrm{WAIC}_V = BU_t - V/n$$

  - these bias corrections are related to how much the model has fitted to the data, and thus thay have been considred as measures of effective number of parameters in the model

WAIC and effective number of parameters

- Widely applicable information criterion (WAIC)
  - only the full data posterior is needed
  - WAIC is asymptotically equal to BU_g and LOO

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

  - WAIC_G and WAIC_V are asymptotically equal, but the series expansion of WAIC_V has closer resemblance to the series expansion of LOO
  - in experiments WAIC_V has also shown to be better than WAIC_G

WAIC and Cross-validation

- WAIC and Bayesian cross-validation
  - both are Bayesian because the focus is on predictive distributions
  - even if the Bayesian word is dropped, if the focus is on predictive distributions, CV and LOO-CV are fully Bayesian

WAIC and Cross-validation

- WAIC and Bayesian cross-validation
  - both are Bayesian because the focus is on predictive distributions
  - even if the Bayesian word is dropped, if the focus is on predictive distributions, CV and LOO-CV are fully Bayesian
- WAIC and (PS)IS-LOO have same computation time
  - PSIS-LOO has better properties

WAIC and Cross-validation

- WAIC and Bayesian cross-validation
  - both are Bayesian because the focus is on predictive distributions
  - even if the Bayesian word is dropped, if the focus is on predictive distributions, CV and LOO-CV are fully Bayesian
- WAIC and (PS)IS-LOO have same computation time
  - PSIS-LOO has better properties
- Exact LOO or *k*-fold-CV more robust than WAIC or PSIS-LOO

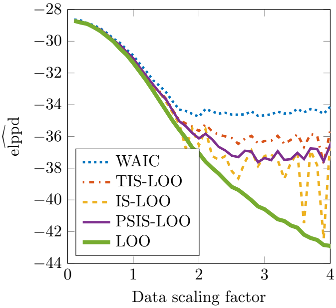
8 schools example

- A simple hierarchical model

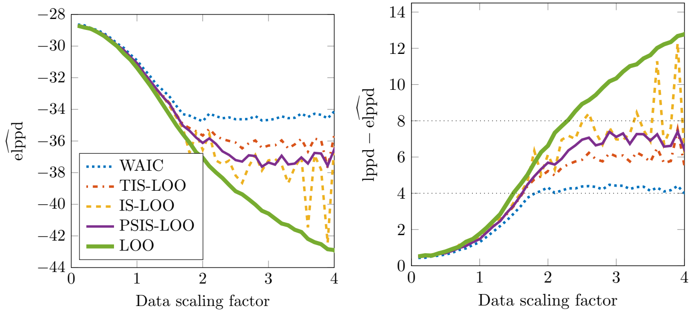
$$\begin{aligned} y_i &\sim \mathrm{N}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \mathrm{N}(\mu, \tau^2), \quad i = 1, \ldots, n = 8 \end{aligned}$$

with a uniform prior distribution on (μ, τ)

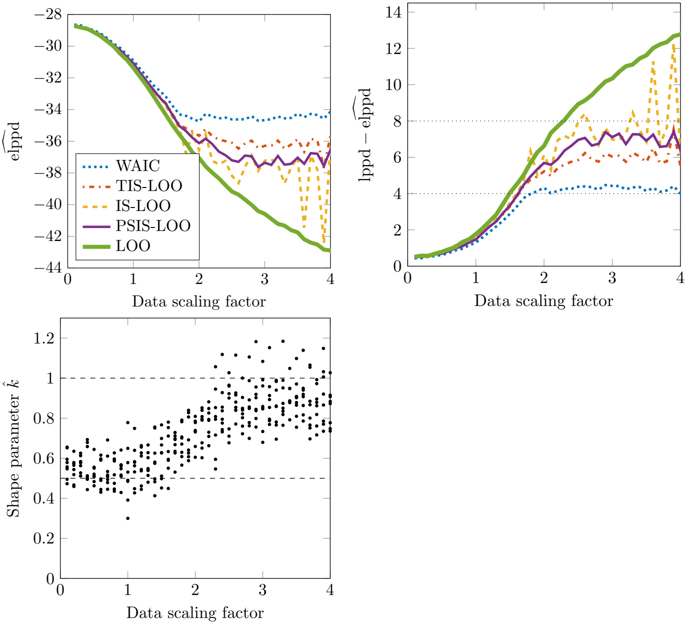
8 schools example



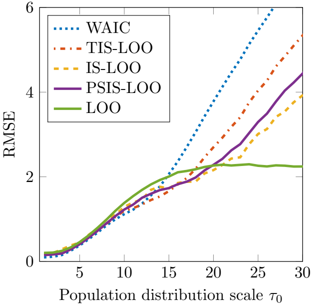
8 schools example



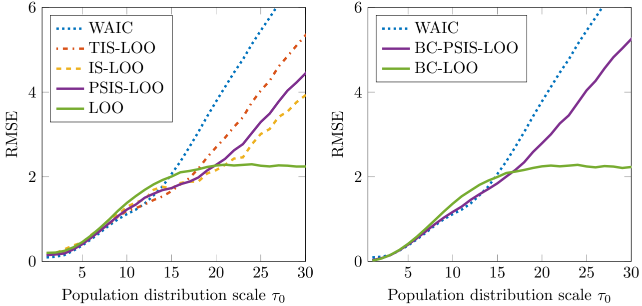
8 schools example



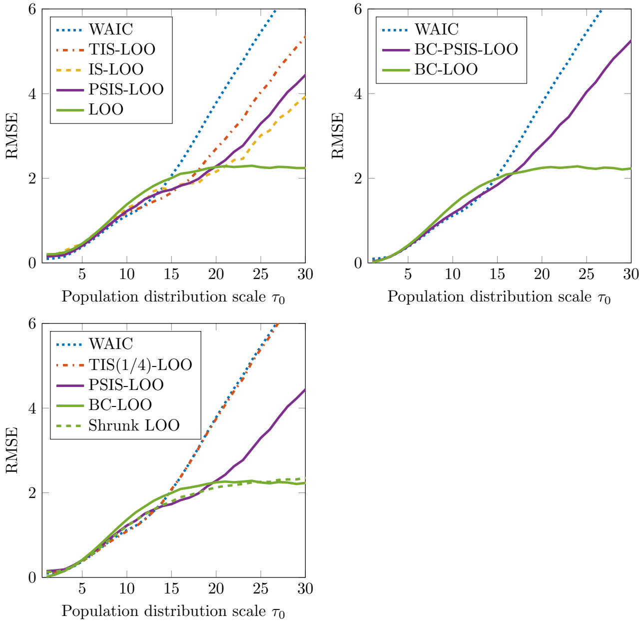
8 schools example



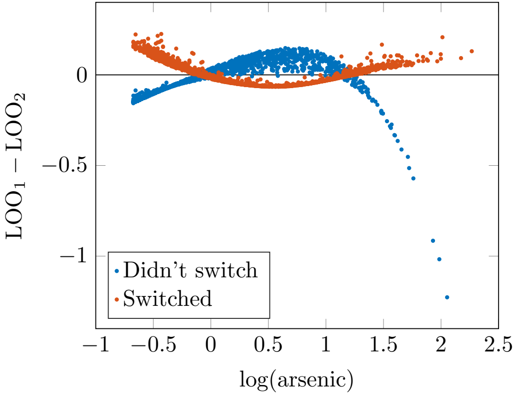
8 schools example



8 schools example



Arsenic well example – Model comparison

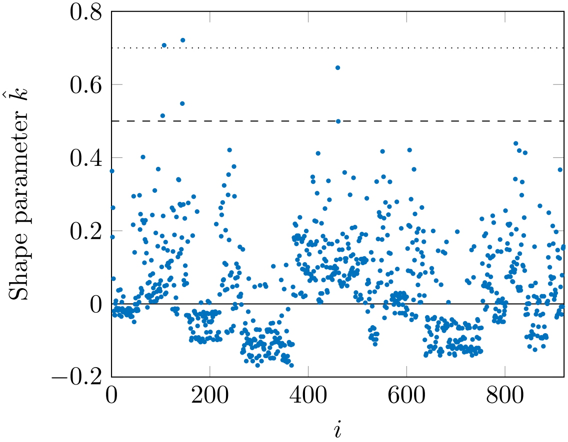


An estimated difference in $\mathrm{elpd}_{\mathrm{loo}}$ of 16.4 with a standard error of 4.4.

Aki Vehtari, Andrew Gelman and Jonah Gabry (2016). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544. http://arxiv.org/abs/1507.04544

Radon example

PSIS-LOO diagnostics



Aki Vehtari, Andrew Gelman and Jonah Gabry (2016). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544. http://arxiv.org/abs/1507.04544

AIC / DIC

- AIC: predictions using the maximum likelihood estimate
  - bias correction using full number of parameters
- DIC: predictions using the posterior mean estimate

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

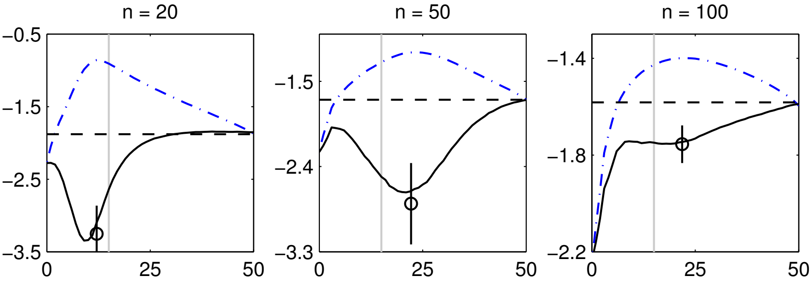
Selection induced bias

- 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

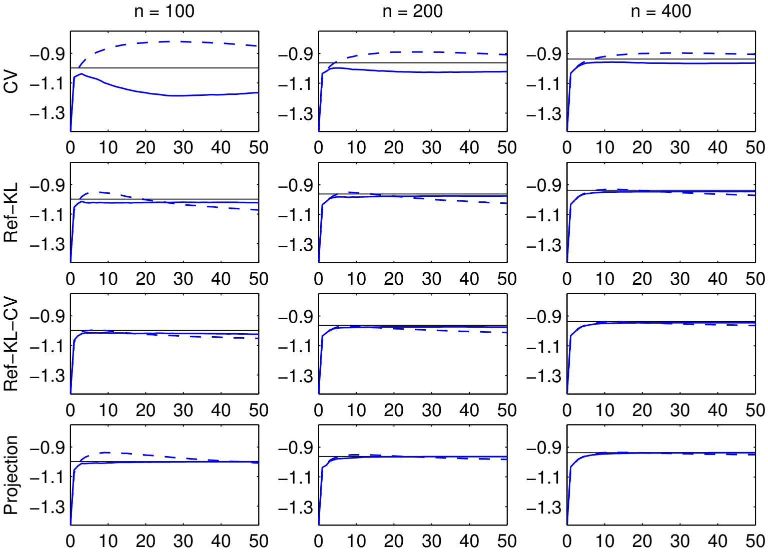
Selection induced bias

- 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

Selection induced bias – Toy example



Selection induced bias – Crime data



Other forms of model selection / hypothesis testing

- Marginal posterior probabilities and intervals
  - problems when posterior dependencies, e.g. due to correlation of covariates
- Bayes factor & evidence
  - sensitive to prior as seen from the predictive interpretation

Bayes factor

- Marginal likelihood in Bayes factor is also a predictive criterion
  - chain rule

$$p(y|M_k) = p(y_1|M_k)p(y_2|y_1, M_k), \ldots, p(y_n|y_1, \ldots, y_{n-1}, M_k)$$

- Sensitive to the first terms, and not defined if the prior is improper
  - especially problematic to use for models with large difference in the number of parameters