On Bayes Cross Validation and Widely Applicable Information Criterion for Gaussian process models

The Second Workshop on Bayesian Inference for Latent Gaussian Models with Applications

Aki Vehtari[1]
with Ole Winther[2], Tommi Mononen[1], Ville Tolvanen[1]

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

[2]Informatics and Mathematical Modelling (IMM)
Technical University of Denmark (DTU)

Outline

- Goal: estimate the predictive performance of Gaussian process (GP)
  - useful for model assessment and selection
- Ideal criterion: Bayes generalization utility
  - can be estimated with LOO and WAIC
  - DIC is related to WAIC but estimates something else
- Comparison: LOO, approximated LOO, WAIC, DIC

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)
- If true future distribution $p_t(\tilde{y})$ would be known, 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

Predictive performance

- Future outcome $\tilde{y}$ is unknown (ignoring $\tilde{x}$ in this slide)
- If true future distribution $p_t(\tilde{y})$ would be known, 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

- 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 $\tilde{y}$

Predictive performance

- Bayes generalization utility

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

- Since $p_t(\tilde{x}, \tilde{y})$ is unknown, we have to estimate it
  - LOO and WAIC re-use observations $(x^{(i)}, y^{(i)})$ to approximate $p_t(\tilde{x}, \tilde{y})$

Estimating predictive performance

- Bayes training utility

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

  - biased (overoptimistic)

Estimating predictive performance

- Bayes training utility

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

  - biased (overoptimistic)

- Bayes leave-one-out cross-validation

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

  - almost unbiased (Watanabe 2010)

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

Estimating predictive performance

- Bayes training utility

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

  - biased (overoptimistic)

- Bayes leave-one-out cross-validation

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

  - almost unbiased (Watanabe 2010)

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

  - simplest approach requires computation of n LOO-posteriors

Widely applicable information criterion

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

$$\mathrm{WAIC}_G = BU_t - 2(BU_t - GU_t)$$

$$\mathrm{WAIC}_V = BU_t - V/n$$

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

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

Widely applicable information criterion

- 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 was better, and rest of results are using WAIC_V

LOO vs. WAIC

- Asymptotic equivalency of WAIC
  - does not tell how well it works for finite n
  - assumes infill (or fixed domain) asymptotics
- $\mathrm{LOO} \approx \mathrm{WAIC}$ only if

$$p(\tilde{y}|x_i, D_{-i}, M_k) \approx p(\tilde{y}|x_i, D, M_k)$$

LOO vs. WAIC

- Let's examine individual terms of LOO and WAIC

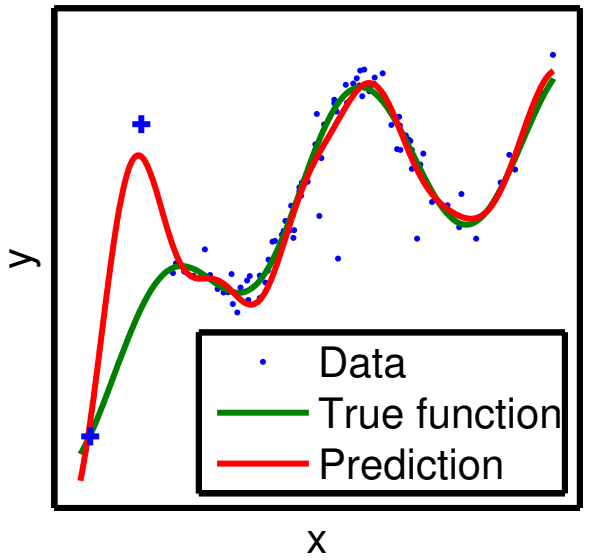
$$\mathrm{LOO}_i = \log p(y_i|x_i, D_{-i}, M_k)$$
$$\mathrm{WAIC}_i = \log p(y_i|x_i, D, M_k) - V_i/n$$

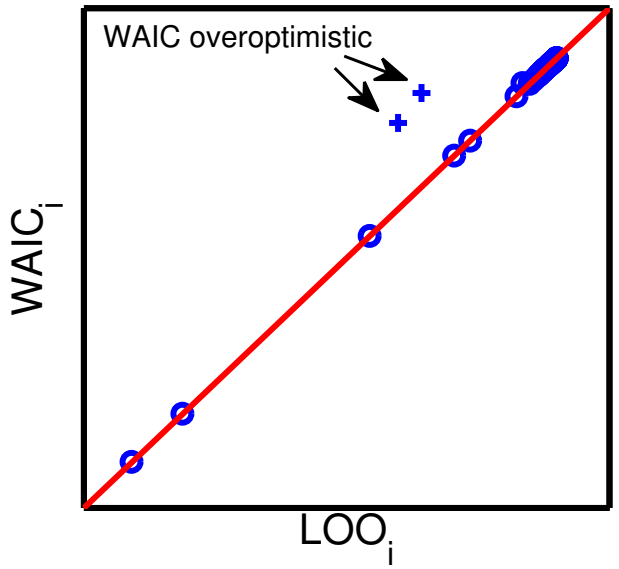
LOO vs. WAIC

- Let's examine individual terms of LOO and WAIC

$$\mathrm{LOO}_i = \log p(y_i|x_i, D_{-i}, M_k)$$
$$\mathrm{WAIC}_i = \log p(y_i|x_i, D, M_k) - V_i/n$$

- Example: Outliers and Gaussian observation model (model misspecification)



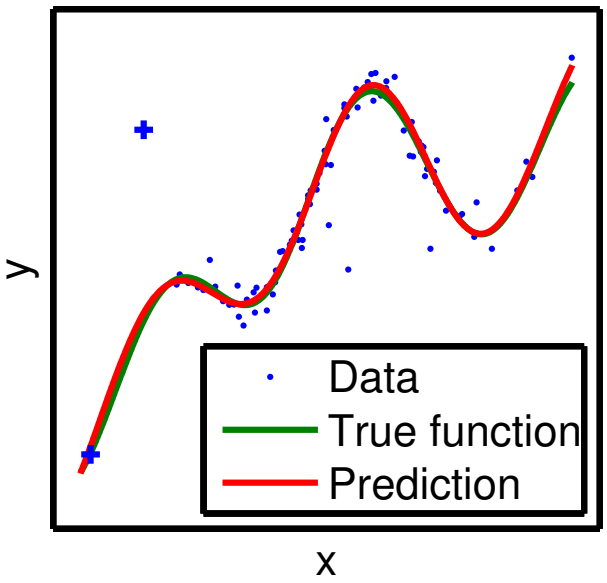


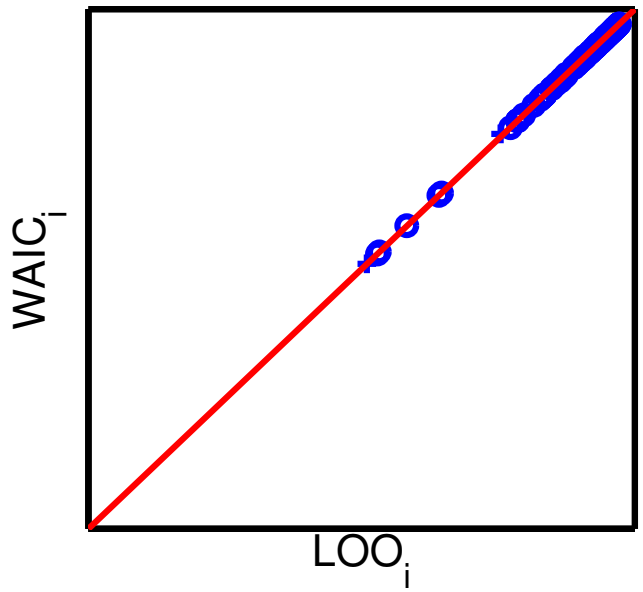
LOO vs. WAIC

- Let's examine terms of LOO and WAIC

$$\mathrm{LOO}_i = \log p(y_i|x_i, D_{-i}, M_k)$$
$$\mathrm{WAIC}_i = \log p(y_i|x_i, D, M_k) - V_i/n$$

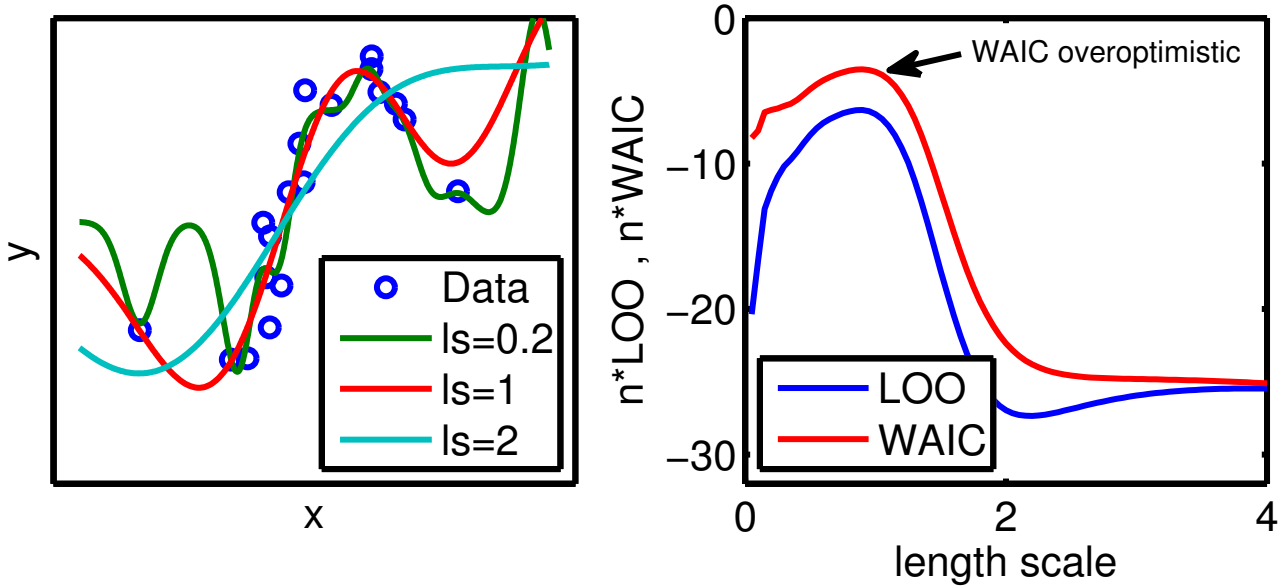
- Example: Outliers and Student's t observation model (with EP)





LOO vs. WAIC

- If length scale is small, WAIC differs from LOO
  - datapoints far from others almost independent (only little or no borrowing of information)
  - WAIC uses information from y_i, LOO does not
- Example: Gaussian noise and Gaussian model

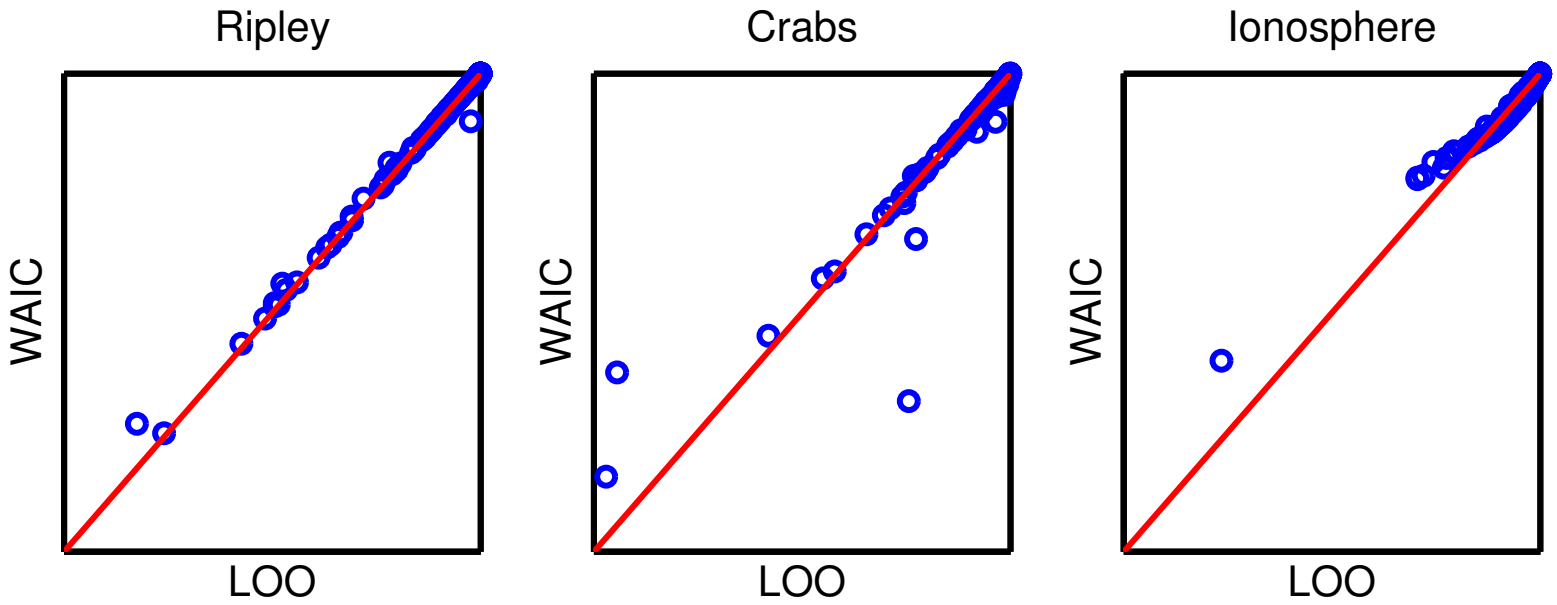


LOO vs. WAIC

- We made comparisons with 9 different data sets
  - brute-force LOO as baseline
  - classification, binomial, negative-Binomial, Student's-t
  - number of covariates 2–60, n=100–911
  - I show here results from 3 datasets, but other results are similar (or less interesting)
- Models
  - integration over latent values with Expectation propagation (EP) or Laplace (LA)
  - integration over the parameters with CCD

LOO vs. WAIC

- WAIC is not reliable replacement for LOO



WAIC vs. DIC

- Commonly used DIC can be written as

$$\mathrm{DIC} = PU_t - 2(PU_t - GU_t),$$

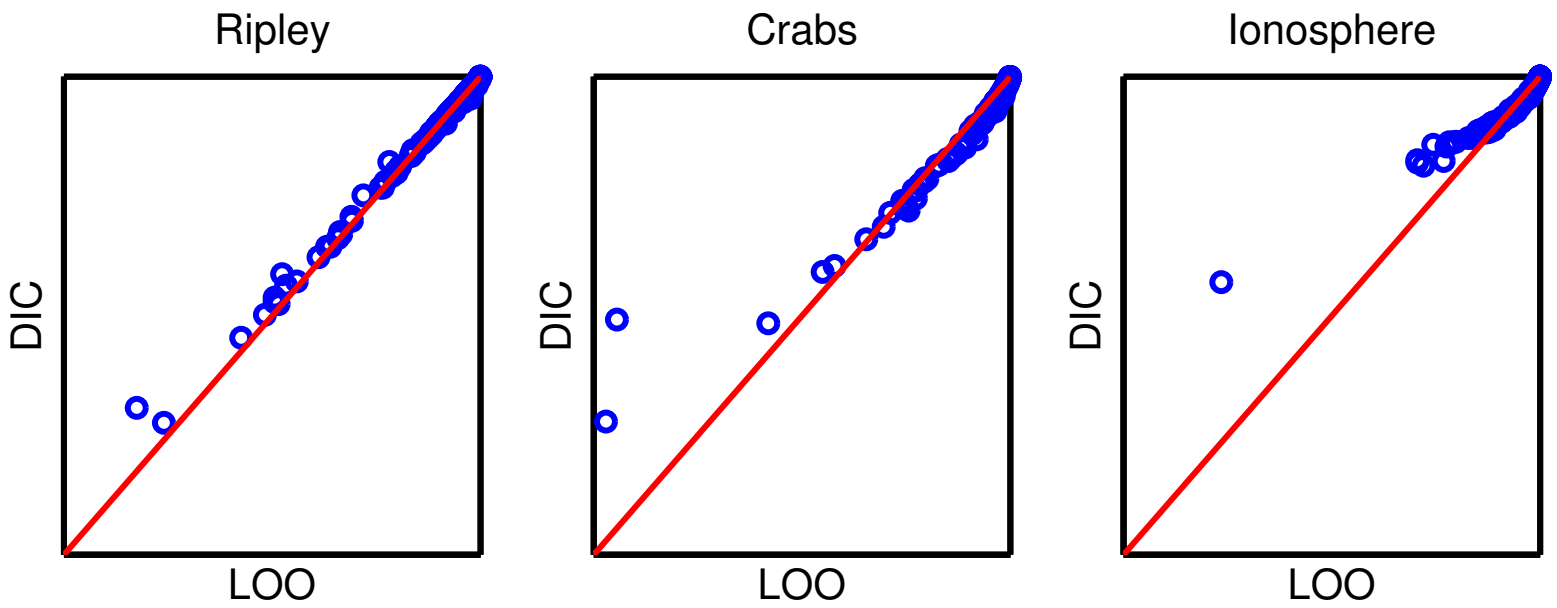
  where

$$PU_t = \frac{1}{n}\sum_{i=1}^{n} \log p(\dot{y}_i|\dot{x}_i, \bar{\theta}_k, M_k)$$

  is the plug-in training utility with point estimate $\bar{\theta}_k$
- DIC estimates plug-in generalization utility
- DIC works only for regular models (not for singular models)
- DIC is not Bayesian

LOO vs. DIC

- DIC is worse than WAIC



Approximations of Bayes LOO-CV

- k-fold-CV
- Mixed LOO
- Importance sampling LOO
- EP-LOO
- Laplace-LOO

LOO for GP

- For Gaussian process the LOO-CV density

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

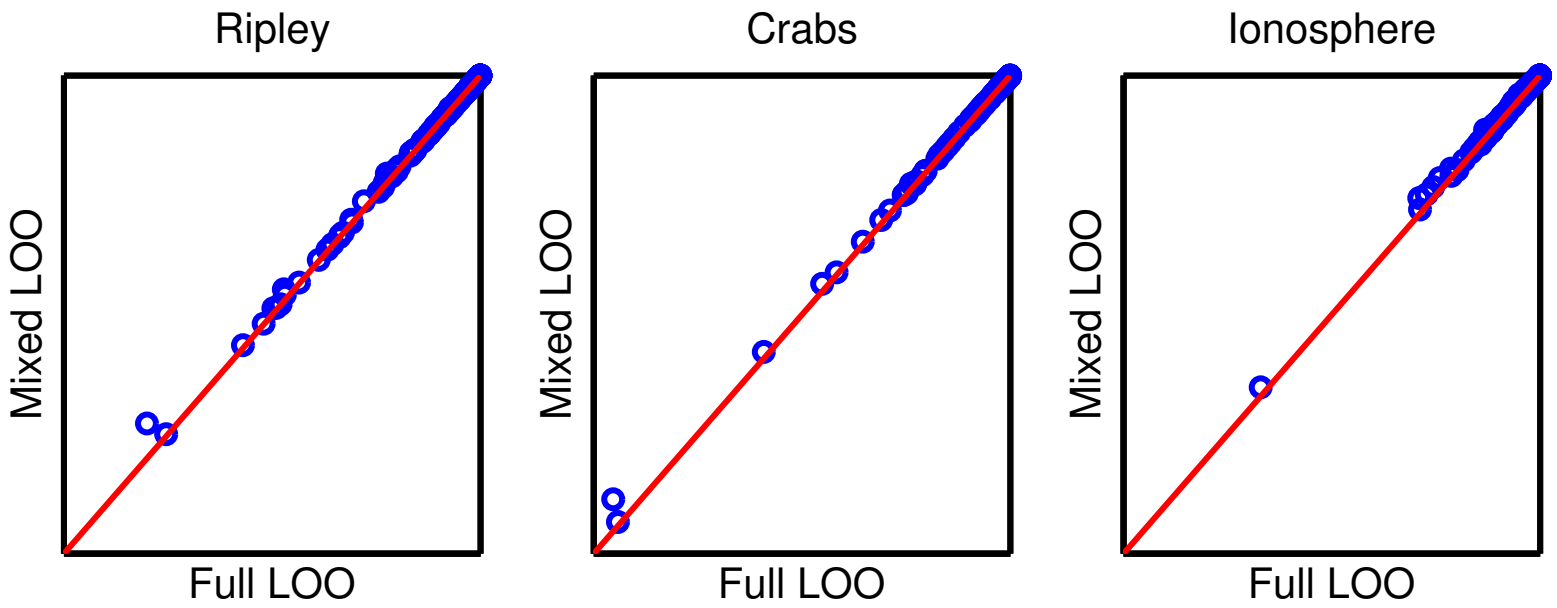
  conditioned on the hyperparameters can be either
  - computed analytically for Gaussian case
  - approximated with EP or Laplace approximation

Mixed LOO

- If the hyperparameter posterior is not sensitive to leaving one data point out

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

we can use the full posterior for hyperparameters



Importance sampling for hyperparameters

- LOO posterior for hyperparameters can be approximated using importance sampling (Gelfand et al, 1992)
  - for GP weights are inversely proportional to conditional LOO densities

$$\frac{p(\theta^t|D_{-i},M)}{p(\theta^t|D,M)} \propto \frac{1}{p(y_i|x_i,\theta^t,D_{-i},M)} = w^{(\backslash i),t}$$

Importance sampling for hyperparameters

- LOO posterior for hyperparameters can be approximated using importance sampling (Gelfand et al, 1992)
  - for GP weights are inversely proportional to conditional LOO densities

$$\frac{p(\theta^t|D_{-i},M)}{p(\theta^t|D,M)} \propto \frac{1}{p(y_i|x_i,\theta^t,D_{-i},M)} = w^{(\backslash i),t}$$

- For these datasets there was not much difference between mixed LOO and IS-LOO

Exact LOO

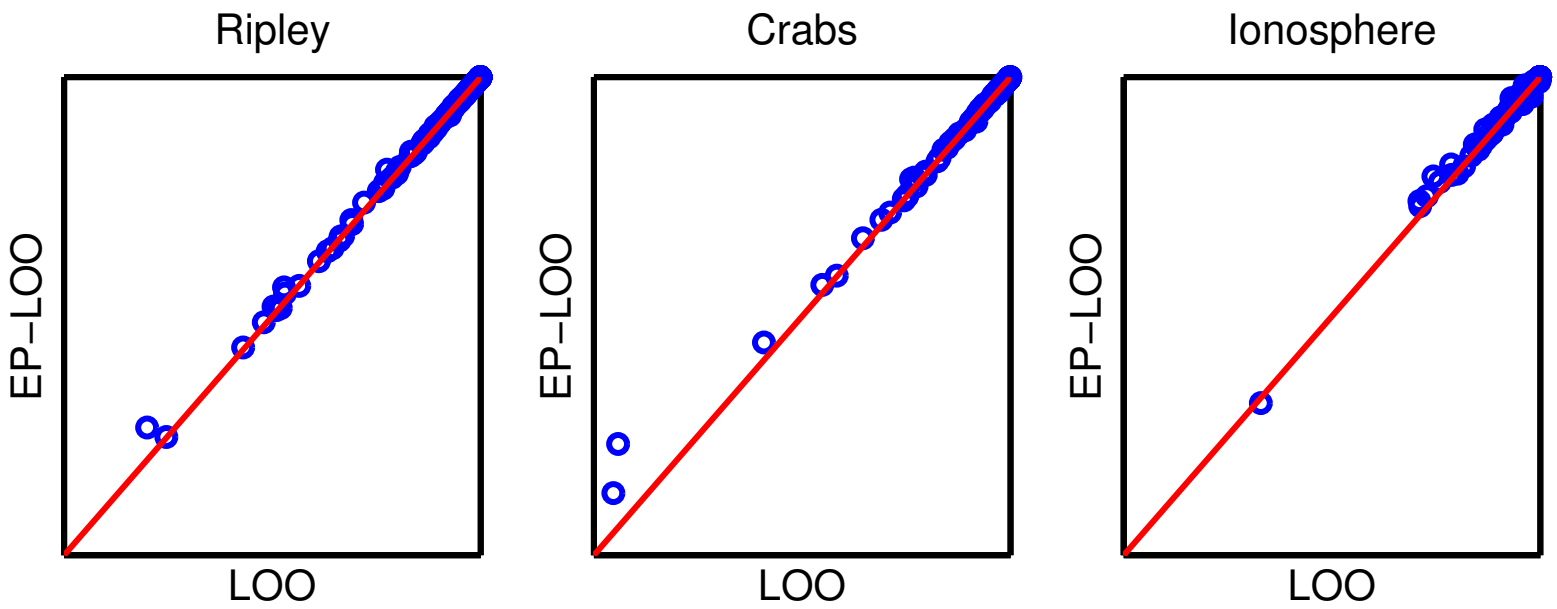
- With Gaussian observation model, exact LOO can be computed quickly analytically (Sundararajan & Keerthi, 2001)

EP-LOO

- Opper & Winther (2000) showed using linear response theory that cavity distributions can be used to approximate LOO distributions
  - EP-LOO is obtained as free byproduct of EP

EP-LOO

- Opper & Winther (2000) showed using linear response theory that cavity distributions can be used to approximate LOO distributions
  - EP-LOO is obtained as free byproduct of EP



LA-LOO INLA-style

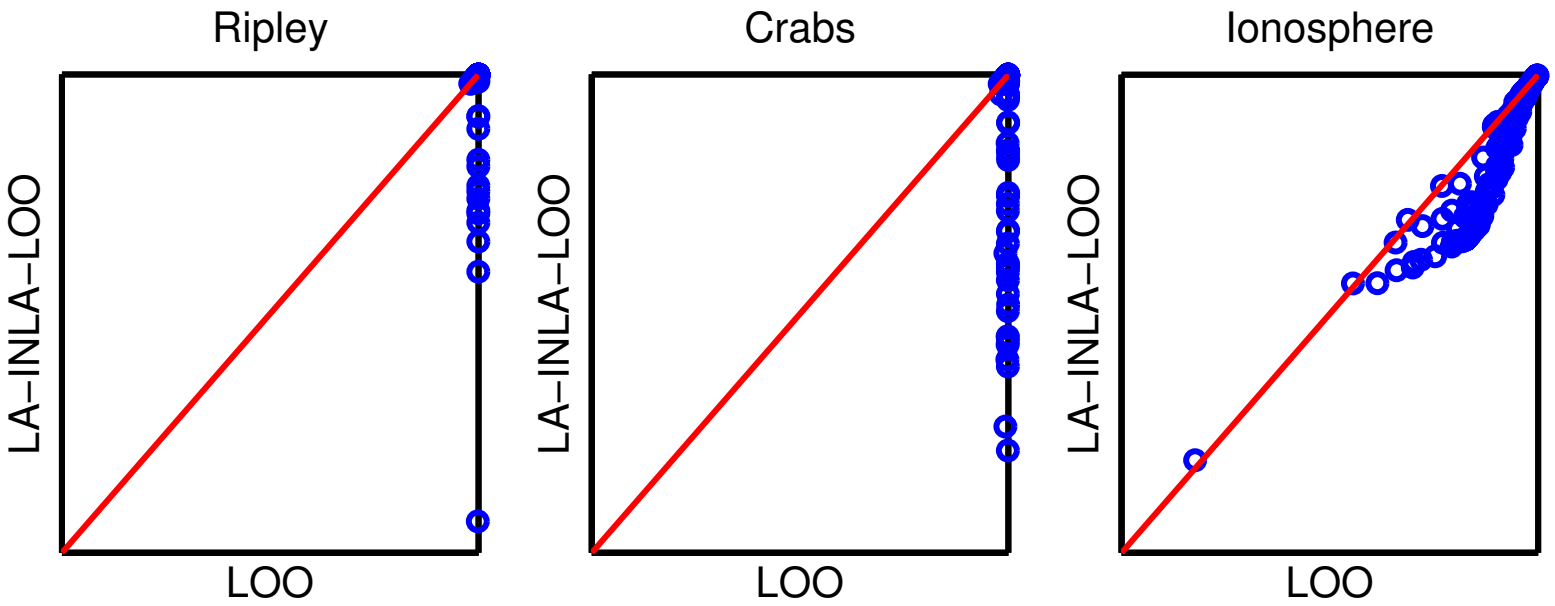
- Held et al (2010)

$$p(y_i|x_i, D_{-i}, \theta, M) = 1/\int \frac{p(f_i|D, \theta, M)}{p(y_i|f_i, \theta, M)} df_i$$

LA-LOO INLA-style

- Held et al (2010)

$$p(y_i|x_i, D_{-i}, \theta, M) = 1/\int \frac{p(f_i|D, \theta, M)}{p(y_i|f_i, \theta, M)} df_i$$

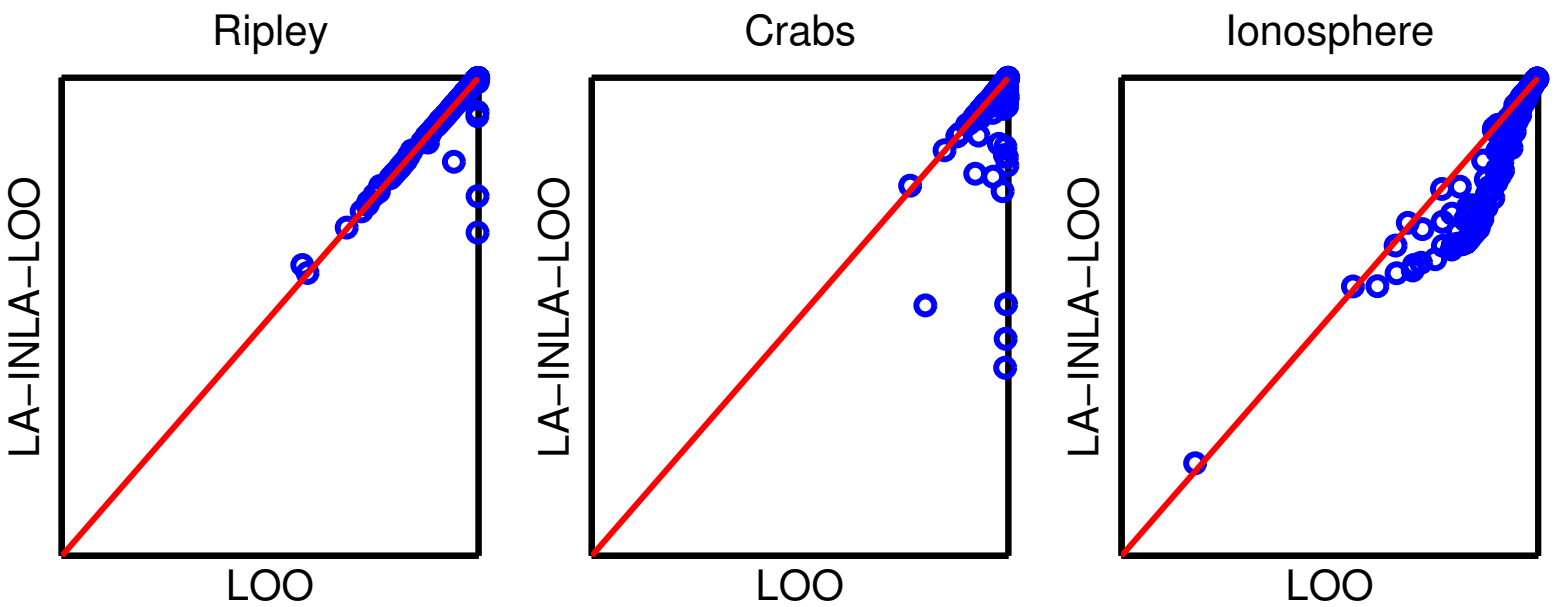


LA-LOO INLA-style

- Held et al (2010)

$$p(y_i|x_i, D_{-i}, \theta, M) = 1/\int \frac{p(f_i|D, \theta, M)}{p(y_i|f_i, \theta, M)} df_i$$

Zoomed to corner, we see that this works for easy predictions



LA-LOO Linear response -style

- New: linear response style for Laplace approximation

$$E[f_i|D_{-i},\theta] = E[f_i|D,\theta] - Var[f_i|D_{-i},\theta]g_i$$

$$Var[f_i|D_{-i},\theta] = \left[(K+\Lambda)^{-1}\right]^{-1}_{ii} - \Lambda_{ii}$$

where K is prior covariance and g and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ contain first and second derivates of the likelihood

  - obtained as free byproduct of Laplace approximation

LA-LOO Linear response -style

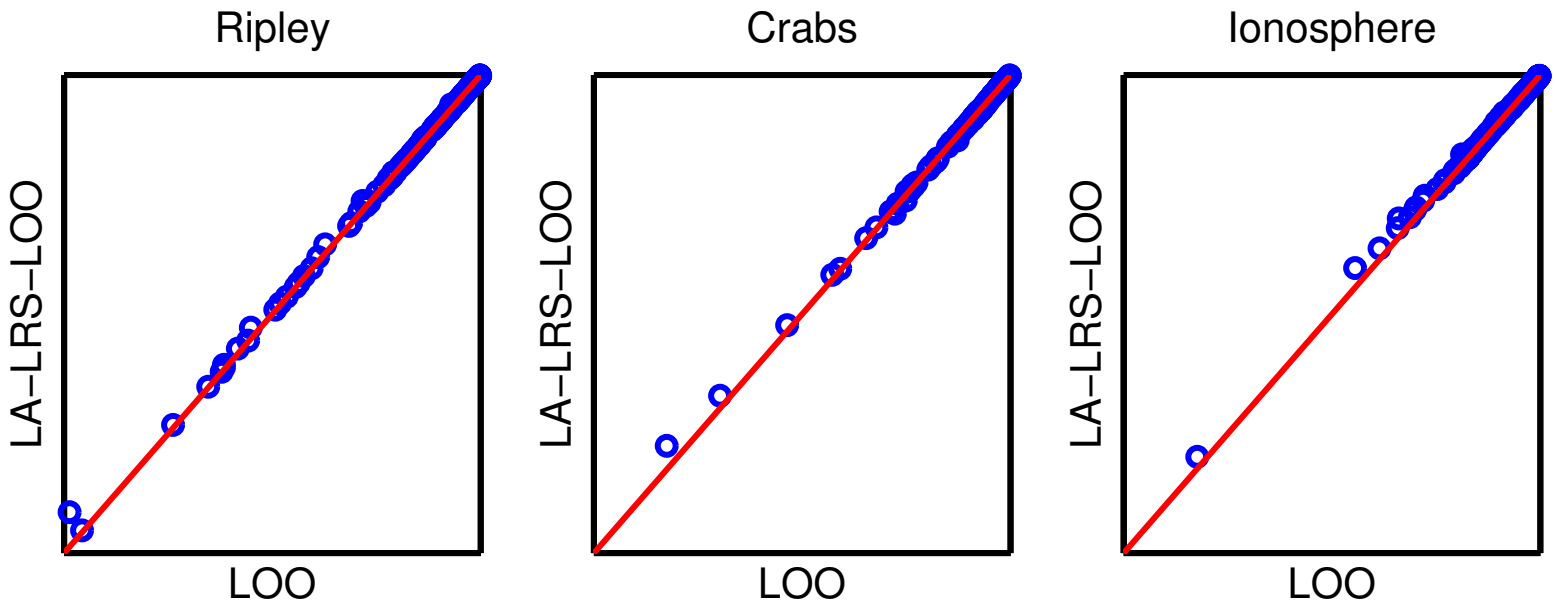
- New: linear response style for Laplace approximation

$$E[f_i|D_{-i},\theta] = E[f_i|D,\theta] - Var[f_i|D_{-i},\theta]g_i$$

$$Var[f_i|D_{-i},\theta] = \left[(K+\Lambda)^{-1}\right]^{-1}_{ii} - \Lambda_{ii}$$

where K is prior covariance and g and $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_n)$ contain first and second derivates of the likelihood

  - obtained as free byproduct of Laplace approximation



Conclusions

- Don't use DIC or WAIC
- New linear response style Laplace-LOO
- If posterior samples for hyperparameters
  - IS for hypers, EP-LOO or LA-LRS-LOO for latents
  - if IS weights bad $\rightarrow$ k-fold-CV
- If optimized hyperparameters
  - EP-LOO or LA-LRS-LOO for latents
  - if in doubt $\rightarrow$ k-fold-CV

Code available in free GPstuff toolbox (just Google it)

To be continued...

- In the next episode:
  - WAIC might be useful for fixed x and no outliers
  - Don't use LOO, WAIC, or DIC for model selection if there is a large number of models (e.g. in covariate selection)
    - because they have relatively large variance
    - because they are negatively correlated with BU_g