Bayesian LOO approximations for GPs

Aki Vehtari
Aki.Vehtari@aalto.fi

Aalto University
School of Science

Submitted paper

- Based on a submitted paper
  Aki Vehtari, Tommi Mononen, Ville Tolvanen and Ole Winther: Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models
  http://arxiv.org/abs/1412.7461

Motivation

- Disease risk prediction, survival analysis
  - noisy data, small amount of events
- Model selection, reporting the estimated predictive performance

Bayesian predictions

- Posterior predictive distribution

$$p(\tilde{y}|\tilde{x}, D) = \int p(\tilde{y}|\tilde{f}, \phi)p(\tilde{f}|D, \theta)p(\theta, \phi|D)d\theta d\phi \qquad (1)$$

Estimation of the predictive performance of GP

- How to avoid naïve *k*-fold-CV?
  - leave-one-out (LOO) approximations
- Approximations depend on how the predictions are made
  - anlytically, Laplace, EP, VB, MCMC for latents?
  - marginal posterior improvements?
  - integration over the hyperparameters?

Predictive distributions

- Posterior predictive distribution

$$p(\tilde{y}|\tilde{x}, D) \tag{2}$$

- LOO predictive distribution

$$p(y_i|x_i, D_{-i}) \tag{3}$$

Predictive distributions

- Posterior predictive distribution

$$p(\tilde{y}|\tilde{x}, D) \tag{2}$$

- LOO predictive distribution

$$p(y_i|x_i, D_{-i}) \tag{3}$$

- Sloppy notation and distribution vs density

Hierarchical LOO computation

- Possible to compute first

$$p(y_i|x_i, D_{-i}, \theta, \phi) \tag{4}$$

and then

$$p(y_i|x_i, D_{-i}) = \int p(y_i|x_i, D_{-i}, \theta, \phi)p(\theta, \phi|D_{-i})d\theta d\phi \tag{5}$$

Generic approach

- Consider the case where we have not yet seen the *i*th observation. Then using the Bayes' rule we can add information from the *i*th observation

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \tag{6}$$

Generic approach

- Consider the case where we have not yet seen the *i*th observation. Then using the Bayes' rule we can add information from the *i*th observation

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \tag{6}$$

- Correspondingly we can remove the effect of the *i*th observation from the full posterior:

$$p(f_i|x_i, D_{-i}) = \frac{p(f_i|D)p(y_i|x_i, D_{-i})}{p(y_i|f_i)} \tag{7}$$

Generic approach

- Consider the case where we have not yet seen the *i*th observation. Then using the Bayes' rule we can add information from the *i*th observation

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \tag{6}$$

- Correspondingly we can remove the effect of the *i*th observation from the full posterior:

$$p(f_i|x_i, D_{-i}) = \frac{p(f_i|D)p(y_i|x_i, D_{-i})}{p(y_i|f_i)} \tag{7}$$

  If we now integrate both sides over f_i and rearrange the terms we get

$$p(y_i|x_i, D_{-i}) = 1 / \int \frac{p(f_i|D)}{p(y_i|f_i)} df_i \tag{8}$$

Generic approach

- In some cases, we can compute $p(f_i|x_i, D_{-i})$ exactly or approximate it efficiently and then we can compute the LOO predictive density,

$$p(y_i|x_i, D_{-i}) = \int p(f_i|x_i, D_{-i})p(y_i|f_i)df_i, \tag{9}$$

Analytic

- With Gaussian likelihood and fixed hyperparameters analytic LOO equations for

$$\begin{aligned} p(f_i|x_i, D_{-i}, \theta, \phi) &\propto \frac{p(f_i|D, \theta)}{p(y_i|f_i, \phi)} \\ &= \mathrm{N}(f_i|\mu_{-i}, v_{-i}), \end{aligned} \tag{10}$$

where

$$\begin{aligned} \mu_{-i} &= v_{-i}(\Sigma_{ii}^{-1}\mu_i - \sigma^{-2}y_i) \\ v_{-i} &= \left(\Sigma_{ii}^{-1} - \sigma^{-2}\right)^{-1} \end{aligned} \tag{11}$$

which removes the effect of observation y_i from the marginal $p(f_i|x_i, D, \theta, \phi)$

EP

- Opper & Winther (2000) showed that EP cavity distribution is up to first order LOO consistent
  - this means that if we are going to use EP approximated predictive distribution of the latent $q(\tilde{f}|\tilde{x}, D, \theta, \phi)$ we can use analytic equations given the Gaussian latent posterior approximation by EP
  - LOO distributions are cavity distributions, which are obtained as a byproduct of the method

Laplace

- First order LOO consistency of the Laplace approximation was shown by Vehtari, Mononen, Tolvanen, Winther (2014)
    - this means that if we are going to use Laplace approximated predictive distribution of the latent $q(\tilde{f}|\tilde{x}, D, \theta, \phi)$ we can use analytic equations given the Gaussian latent posterior approximation by Laplace approximation

Laplace

- First order LOO consistency of the Laplace approximation was shown by Vehtari, Mononen, Tolvanen, Winther (2014)
    - this means that if we are going to use Laplace approximated predictive distribution of the latent $q(\tilde{f}|\tilde{x}, D, \theta, \phi)$ we can use analytic equations given the Gaussian latent posterior approximation by Laplace approximation with site terms N$(f_i|\tilde{\mu}_i, \tilde{\Sigma}_i)$

$$\tilde{\Sigma}_i = -\frac{1}{\nabla_i \nabla_i \log p(y_i|f_i, \phi)|_{f_i=\hat{f}_i}} \tag{12}$$

$$\tilde{\mu}_i = \hat{f} + \tilde{\Sigma}_i \nabla_i \log p(y_i|\,\mathbf{f}_i, \phi)|_{f_i=\hat{f}_i} \tag{13}$$

    - computation of LOO takes same time as in case of Gaussian likelihood

VB

- Likely that same holds for VB

Marginal improvements

- If various marginal improvements (LA/EP-L,LA-TK,EP-FULL,LA-CM(2),EP-1STEP,LA/EP-FACT) are used the marginal is not any more Gaussian

Marginal improvements

- If various marginal improvements (LA/EP-L,LA-TK,EP-FULL,LA-CM(2),EP-1STEP,LA/EP-FACT) are used the marginal is not any more Gaussian
- With local approximations corresponding to using tilted distributions (LA-L,EP-L), the predictive latent distribution is still Gaussian

Marginal improvements

- If various marginal improvements (LA/EP-L,LA-TK,EP-FULL,LA-CM(2),EP-1STEP,LA/EP-FACT) are used the marginal is not any more Gaussian
- With local approximations corresponding to using tilted distributions (LA-L,EP-L), the predictive latent distribution is still Gaussian
- With global approximations LA-TK,EP-FULL,LA-CM(2),EP-1STEP,LA/EP-FACT), the predictive latent distribution is not Gaussian
  - other approximations for integral in

$$p(y_i|x_i, D_{-i}) \approx 1 / \int \frac{q(f_i|D)}{p(y_i|f_i)} df_i \qquad (14)$$

LOO with global marginal approximations

- Quadrature
  - used, e.g., in INLA software
- WAIC
- Monte Carlo

Quadrature LOO

- Quadrature integration for

$$p(y_i|x_i, D_{-i}) = 1 / \int \frac{q(f_i|D)}{p(y_i|f_i)} df_i \qquad (15)$$

  - problems if tail of $p(y_i|f_i)$ goes down faster than $q(f_i|D)$
  - depends on the accuracy of non-Gaussian approximation $q(f_i|D)$
  - we propose a truncation approach which makes the quadrature more robust

WAIC

- WAIC uses Taylor series approximation for

$$p(y_i|x_i, D_{-i}) = 1/\int \frac{q(f_i|D)}{p(y_i|f_i)} df_i \qquad (16)$$

  - does not help compared to direct quadrature

Monte Carlo

- Monte Carlo approximation for

$$p(y_i|x_i, D_{-i}) = 1 \Big/ \int \frac{q(f_i|D)}{p(y_i|f_i)} df_i \tag{17}$$

  - e.g., importance sampling
  - does not help compared to direct quadrature

Experimental results

- Small datasets, so that we can compoute brute-force LOO
- Accuracy of the approximations improves for larger datasets

Data set	n	d	observation model
Ripley	250	2	probit
Australian	690	14	probit
Ionosphere	351	33	probit
Sonar	208	60	probit
Leukemia	1043	4	log-logistic with censoring

Table : Summary of datasets and models in our examples.

LA results with fixed hyperparameters

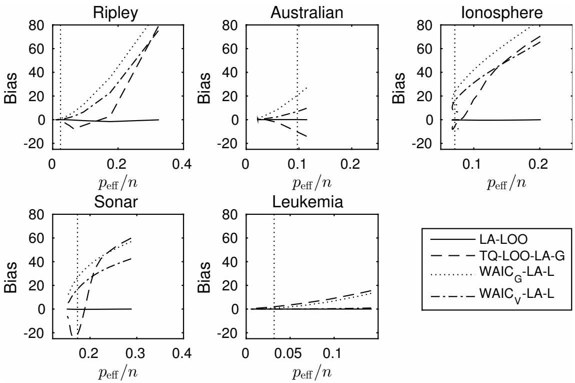

Figure : Bias when the target is brute-force-LOO with Laplace and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters p_{eff}/n. The flexibility of the MAP model is shown with a vertical dashed line.

EP results with fixed hyperparameters

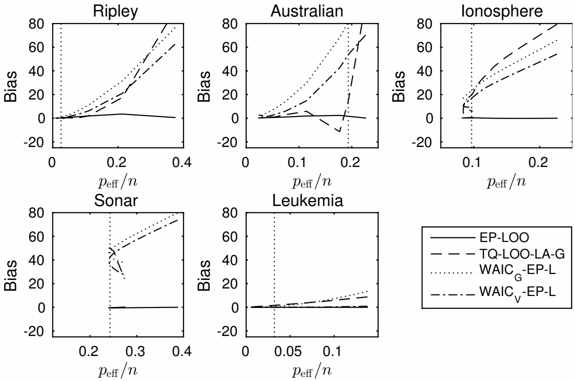

Figure : Bias when the target is brute-force-LOO with EP and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters p_{eff}/n. The flexibility of the MAP model is shown with a vertical dashed line.

LA-CM2 results with fixed hyperparameters

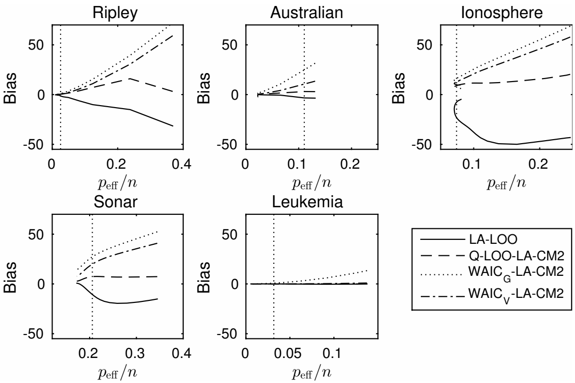

Figure : Bias when the target is brute-force-LOO with Laplace-CM2 and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters p_{eff}/n. The flexibility of the MAP model is shown with a vertical dashed line.

EP-FACT results with fixed hyperparameters

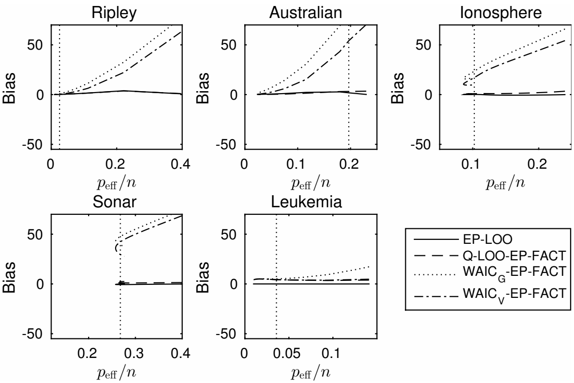

Figure : Bias when the target is brute-force-LOO with EP-FACT and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters p_{eff}/n. The flexibility of the MAP model is shown with a vertical dashed line.

Unknown hyperparameters

- If hyperparameters are unknown and optimised, the above estimates are optimistic
  - bias can be negligible, if big data and the number of hyperparameters is small

Unknown hyperparameters

- If hyperparameters are unknown and optimised, the above estimates are optimistic
    - bias can be negligible, if big data and the number of hyperparameters is small
- Better to integrate over the hyperparameters
    - deterministic samples, e.g., CCD
    - stochastic samples, e.g. importance sampling, MCMC

Hierarchical approximation using IS

- Using above results for the conditional part $p(y_i|x_i, D_{-i}, \theta, \phi)$, the LOO predictive distribution can be approximated using IS for hyperparameters

Hierarchical approximation using IS

- Using above results for the conditional part $p(y_i|x_i, D_{-i}, \theta, \phi)$, the LOO predictive distribution can be approximated using IS for hyperparameters

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^{S} p(\tilde{y}_i|D_{-i}, \phi^s) w_i^s}{\sum_{s=1}^{S} w_i^s}, \tag{18}$$

where w_i^s are importance weights and

$$w_i^s \propto \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}, \tag{19}$$

Hierarchical approximation using IS

- Using above results for the conditional part $p(y_i|x_i, D_{-i}, \theta, \phi)$, the LOO predictive distribution can be approximated using IS for hyperparameters

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^{S} p(\tilde{y}_i|D_{-i}, \phi^s) w_i^s}{\sum_{s=1}^{S} w_i^s}, \qquad (18)$$

where w_i^s are importance weights and

$$w_i^s \propto \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}, \qquad (19)$$

- The LOO predictive density simplifies to

$$p(y_i|x_i, D_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}} \qquad (20)$$

Improving IS

- Variance of IS can be reduced by using truncated importance sampling
- “Very Good Importance Sampling” (work in progress)

Hierarchical approximation using IS

- Importance weighting works also for deterministic CCD method

LA/EP results with unknown hyperparameters

Method	Ripley	Australian	Ionosphere	Sonar	Leukemia
LA-LOO+CCD+IS	**0.2** (0.1)	**3.4** (0.4)	**-0.1** (0.1)	**-0.13** (0.06)	**0.56** (0.05)
LA-LOO+CCD	0.8 (0.2)	7.2 (0.9)	0.6 (0.2)	0.5 (0.2)	4.8 (0.2)
LA-LOO+MAP	1.0 (0.2)	9.2 (1.8)	1.3 (0.2)	1.3 (0.3)	4.9 (0.6)

Table : Bias and standard deviation when the target is brute-force-LOO with Laplace and CCD.

LA/EP results with unknown hyperparameters

Method	Ripley	Australian	Ionosphere	Sonar	Leukemia
LA-LOO+CCD+IS	**0.2** (0.1)	**3.4** (0.4)	**-0.1** (0.1)	**-0.13** (0.06)	**0.56** (0.05)
LA-LOO+CCD	0.8 (0.2)	7.2 (0.9)	0.6 (0.2)	0.5 (0.2)	4.8 (0.2)
LA-LOO+MAP	1.0 (0.2)	9.2 (1.8)	1.3 (0.2)	1.3 (0.3)	4.9 (0.6)

Table : Bias and standard deviation when the target is brute-force-LOO with Laplace and CCD.

Method	Ripley	Australian	Ionosphere	Sonar	Leukemia
EP-LOO+CCD+IS	**0.42** (0.14)	**7.3** (1.4)	**0.8** (0.6)	**-0.24** (0.14)	**0.49** (0.04)
EP-LOO+CCD	1.3 (0.4)	15 (2)	2.8 (1.3)	0.6 (0.3)	4.8 (0.2)
EP-LOO+MAP	1.4 (0.3)	17 (2)	2.8 (0.7)	0.9 (0.3)	4.9 (0.6)

Table : Bias and standard deviation when the target is brute-force-LOO with EP and CCD.

Sparse GPs

- LOO approximations work well with fixed inducing points
- Naïve optimisiation of inducing points locations would produce optimistic estimates
- VB?

Non-log-concave likelihoods

- Above nice results are with log-concave likelihoods
- Does not work so well with non-log-concave likelihoods
  - first order consistency proof assumes log-concave likelihoods
  - posterior can be multimodal → unimodal approximation bad
  - pseudo observations may have repulsive effect

Non-log-concave likelihoods

- Above nice results are with log-concave likelihoods
- Does not work so well with non-log-concave likelihoods
  - first order consistency proof assumes log-concave likelihoods
  - posterior can be multimodal $\rightarrow$ unimodal approximation bad
  - pseudo observations may have repulsive effect
  - (current) marginal improvment methods don't fix this problem

Summary

- LOO with LA or EP, log-concave likelihoods and fixed hyperparameters is fast and reliable
- IS can be used to handle unknown hyperparameters