

# INTEGRATION OVER HYPERPARAMETERS AND ESTIMATION OF PREDICTIVE PERFORMANCE

Aki Vehtari

Helsinki Institute for Information Technology HIIT,  
Department of Computer Science,  
Aalto University, Finland  
aki.vehtari@aalto.fi

# Outline

- ▶ GP hyperparameter inference
  - ▶ Priors on GP hyperparameters
  - ▶ Benefits of integration vs. point estimate
  - ▶ MCMC, CCD

# Gaussian processes and hyperparameters

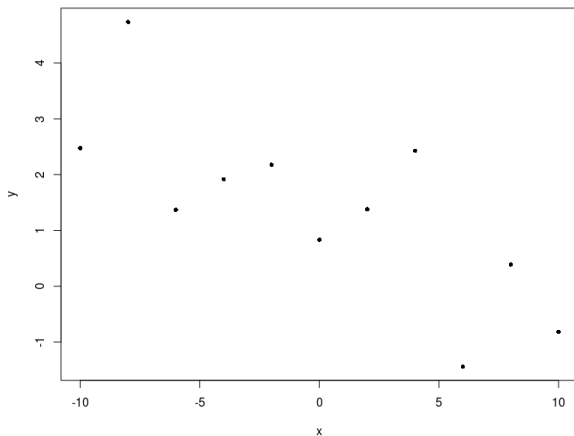
- ▶ Gaussian processes are priors on function space
- ▶ GPs are usually constructed with a parametric covariance function
  - ▶ we need to think about priors on those parameters

# Gaussian processes and hyperparameters

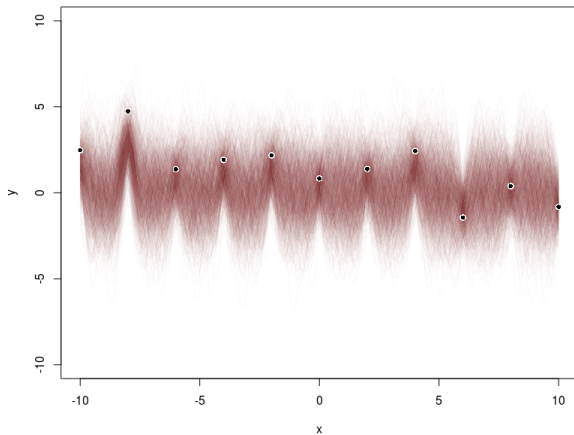
- ▶ Gaussian processes are priors on function space
- ▶ GPs are usually constructed with a parametric covariance function
  - ▶ we need to think about priors on those parameters
- ▶ If we have “big data” and small number of hyperparameters
  - ▶ priors and integration over the posterior is not so important
  - ▶ even more so when sparse approximations, which limit the complexity of the models, are used

# 1D demo

- ▶ 1D demo originally by Michael Betancourt



# 1D demo



# 1D demo summary

- ▶ Likelihood for lengthscale beyond the data scale is flat and non-identifiable because the functions look all the same
  - ▶ add prior making large lengthscale less likely
- ▶ If no repeated measurements non-identifiability between signal magnitude and noise magnitude when lengthscale short
  - ▶ add prior making short lengthscale less likely
  - ▶ add prior on measurement noise
  - ▶ make repeated measurements
- ▶ Nonidentifiability between lengthscale and magnitude

# Non-Gaussian likelihoods

- ▶ Poisson
  - ▶ variance is equal to mean, and thus can't overfit



# Non-Gaussian likelihoods

- ▶ Poisson
  - ▶ variance is equal to mean, and thus can't overfit
  - ▶ except if data is not conditionally Poisson distributed

# Non-Gaussian likelihoods

- ▶ Poisson
  - ▶ variance is equal to mean, and thus can't overfit
  - ▶ except if data is not conditionally Poisson distributed
- ▶ Binary classification (logit/probit)
  - ▶ unbounded likelihood if separable
  - ▶ with short if enough lengthscale separable

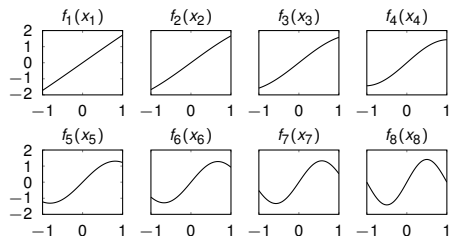
# Sparse approximations

- ▶ Sparse approximations limit the complexity
  - ▶ FITC type models work only with large lengthscale

# Higher dimensions

- ▶ Separate lengthscale for each dimension, aka ARD
  - ▶ lengthscale is related to non-linearity

# Toy example

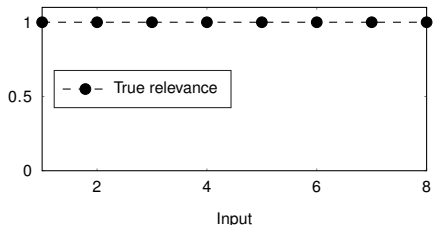


$$f(\mathbf{x}) = f_1(x_1) + \dots + f_8(x_8),$$

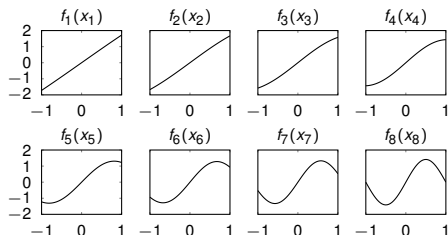
$$y \sim N(f, 0.3^2),$$

$$\text{Var}(f_j) = 1 \text{ for all } j.$$

$\Rightarrow$  All inputs equally relevant



# Toy example

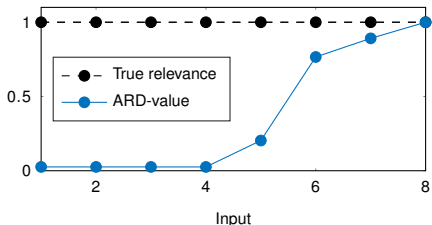


$$f(\mathbf{x}) = f_1(x_1) + \dots + f_8(x_8),$$

$$y \sim N(f, 0.3^2),$$

$$\text{Var}(f_j) = 1 \text{ for all } j.$$

$\Rightarrow$  All inputs equally relevant



Optimized ARD-values,  
 $\text{ARD}(j) = 1/\ell_j$  (averaged over  
100 data realizations,  $n = 200$ )

# Bayesian optimization

- ▶ GPs have been used too much as black boxes
- ▶ Bonus: use shape constrained GPs (see, e.g., Siivola et al., 2017)

# Periodic covariance function

- ▶ If you know the period fix it
- ▶ If you don't know, there can be serious identifiability problems unless informative priors are used



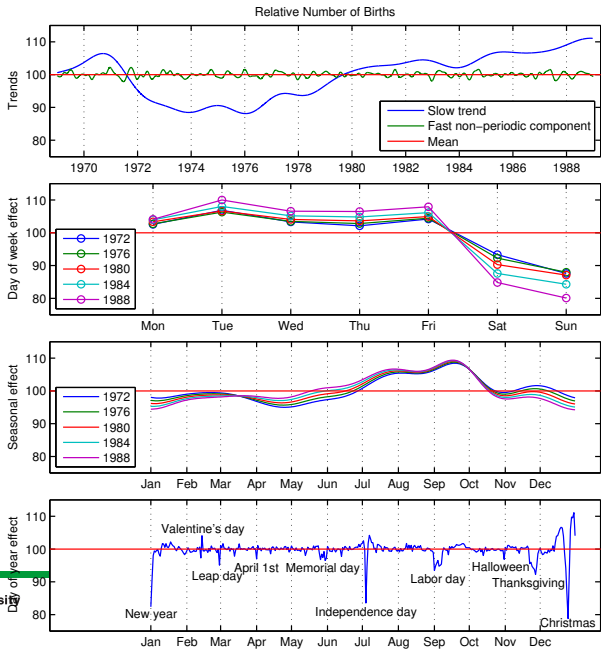
# Parametric model plus GP

- ▶ For example, linear model plus GP
  - ▶ with long lengthscale GP is like a linear model which causes non-identifiability and problems in interpretation

# Parametric model plus GP

- ▶ For example, linear model plus GP
  - ▶ with long lengthscale GP is like a linear model which causes non-identifiability and problems in interpretation
- ▶ Same for other parametric model + GP
  - ▶ need more informative priors

# GP plus GP



# GP plus GP

- ▶ Identifiability problems as different components are explaining same features in the data
  - ▶ priors which “encourage” specialization of the components

# Summary on priors and benefits of integration

- ▶ Specific prior recommendations for length scale
  - ▶ inverse gamma has a sharp left tail that puts negligible mass on small length-scales, but a generous right tail, allowing for large length-scales (but still reducing non-identifiability)
  - ▶ generalized inverse Gaussian has an inverse gamma left tail (if  $p \leq 0$ ) and a Gaussian right tail (avoids identifiability issue when combined with linear model)

# Summary on priors and benefits of integration

- ▶ Specific prior recommendations for length scale
  - ▶ inverse gamma has a sharp left tail that puts negligible mass on small length-scales, but a generous right tail, allowing for large length-scales (but still reducing non-identifiability)
  - ▶ generalized inverse Gaussian has an inverse gamma left tail (if  $\rho \leq 0$ ) and a Gaussian right tail (avoids identifiability issue when combined with linear model)
- ▶ Specific weakly informative prior recommendations for signal and noise magnitude
  - ▶ half-normals are often enough if length-scale has informative prior
  - ▶ if information about measurement accuracy is available, informative prior such as gamma or scaled inverse  $\text{Chi}^2$  for variance

# GPs in Stan

- ▶ Stan manual 2.16.0 (and later) chapter 16  
<http://mc-stan.org/users/documentation/index.html>
  - ▶ code and documentation by Rob Trangucci
  - ▶ prior recommendations by Rob Trangucci, Michael Betancourt, Aki Vehtari
- ▶ Code examples [https://github.com/rtrangucci/gps\\_in\\_stan](https://github.com/rtrangucci/gps_in_stan)
  - ▶ by Rob Trangucci

# Hamiltonian Monte Carlo + NUTS

- ▶ Uses gradient information for more efficient sampling
- ▶ Alternating dynamic simulation and sampling of the energy level
- ▶ Parameters
  - ▶ step size, number of steps in each chain



# Hamiltonian Monte Carlo + NUTS

- ▶ Uses gradient information for more efficient sampling
- ▶ Alternating dynamic simulation and sampling of the energy level
- ▶ Parameters
  - ▶ step size, number of steps in each chain
- ▶ No U-Turn Sampling
  - ▶ adaptively selects number of steps to improve robustness and efficiency

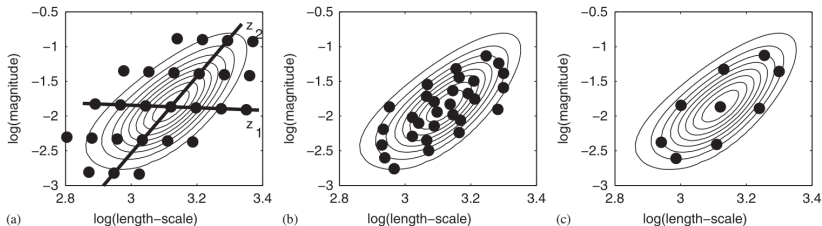
# Hamiltonian Monte Carlo + NUTS

- ▶ Uses gradient information for more efficient sampling
- ▶ Alternating dynamic simulation and sampling of the energy level
- ▶ Parameters
  - ▶ step size, number of steps in each chain
- ▶ No U-Turn Sampling
  - ▶ adaptively selects number of steps to improve robustness and efficiency
- ▶ Adaptation in Stan
  - ▶ Step size adjustment (mass matrix) is estimated during initial adaptation phase

# Hamiltonian Monte Carlo + NUTS

- ▶ Uses gradient information for more efficient sampling
- ▶ Alternating dynamic simulation and sampling of the energy level
- ▶ Parameters
  - ▶ step size, number of steps in each chain
- ▶ No U-Turn Sampling
  - ▶ adaptively selects number of steps to improve robustness and efficiency
- ▶ Adaptation in Stan
  - ▶ Step size adjustment (mass matrix) is estimated during initial adaptation phase
- ▶ Demo
  - ▶ <https://chi-feng.github.io/mcmc-demo/app.html#RandomWalkMH,donut>
  - ▶ note that HMC/NUTS in this demo is not exactly same as in Stan

► Deterministic placement of integration points



# Estimation of the predictive performance of GP

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

# Predictive distributions

- ▶ Posterior predictive distribution

$$p(\tilde{y}|\tilde{x}, D) \quad (1)$$

- ▶ LOO predictive distribution

$$p(y_i|x_i, D_{-i}) \quad (2)$$

# Hierarchical LOO computation

- Possible to compute first

$$p(y_i|x_i, D_{-i}, \theta, \phi) \quad (3)$$

and then

$$p(y_i|x_i, D_{-i}) = \int p(y_i|x_i, D_{-i}, \theta, \phi)p(\theta, \phi|D_{-i})d\theta d\phi \quad (4)$$

## 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

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \quad (5)$$



## 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

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \quad (5)$$

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

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

## 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

$$p(f_i|D) = \frac{p(y_i|f_i)p(f_i|x_i, D_{-i})}{p(y_i|x_i, D_{-i})} \quad (5)$$

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

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

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

$$p(y_i|x_i, D_{-i}) = 1 / \int \frac{p(f_i|D)}{p(y_i|f_i)} df_i \quad (7)$$

# 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,

$$p(y_i|x_i, D_{-i}) = \int p(f_i|x_i, D_{-i})p(y_i|f_i)df_i, \quad (8)$$

# Analytic

- ▶ With Gaussian likelihood and fixed hyperparameters analytic LOO equations for

$$\begin{aligned} p(f_i|x_i, D_{-i}, \theta, \phi) &\propto \frac{p(f_i|D, \theta)}{p(y_i|f_i, \phi)} \\ &= \mathbf{N}(f_i|\mu_{-i}, \mathbf{v}_{-i}), \end{aligned} \quad (9)$$

where

$$\begin{aligned} \mu_{-i} &= \mathbf{v}_{-i}(\Sigma_{ii}^{-1}\mu_i - \sigma^{-2}y_i) \\ \mathbf{v}_{-i} &= \left(\Sigma_{ii}^{-1} - \sigma^{-2}\right)^{-1} \end{aligned} \quad (10)$$

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

- ▶ 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  $q(\tilde{f}|\tilde{\mathbf{x}}, D, \theta, \phi)$  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  $q(\tilde{\mathbf{f}}|\tilde{\mathbf{x}}, D, \theta, \phi)$  we can use analytic equations given the Gaussian latent posterior approximation by Laplace approximation with site terms  $N(\mathbf{f}_i|\tilde{\mu}_i, \tilde{\Sigma}_i)$

$$\tilde{\Sigma}_i = -\frac{1}{\nabla_i \nabla_i \log p(\mathbf{y}_i|\mathbf{f}_i, \phi)|_{\mathbf{f}_i=\hat{\mathbf{f}}_i}} \quad (11)$$

$$\tilde{\mu}_i = \hat{\mathbf{f}} + \tilde{\Sigma}_i \nabla_i \log p(\mathbf{y}_i|\mathbf{f}_i, \phi)|_{\mathbf{f}_i=\hat{\mathbf{f}}_i} \quad (12)$$

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

- ▶ Likely that same holds for VB



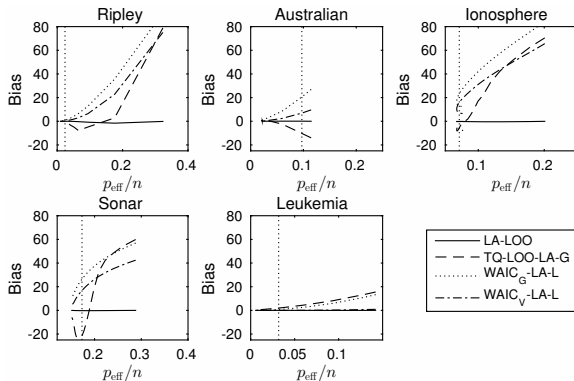
# Experimental results

- ▶ Small datasets, so that we can compute 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

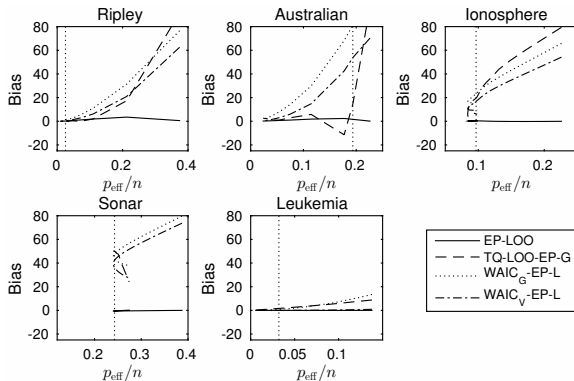
Table: Summary of datasets and models in our examples.

# LA results with fixed hyperparameters



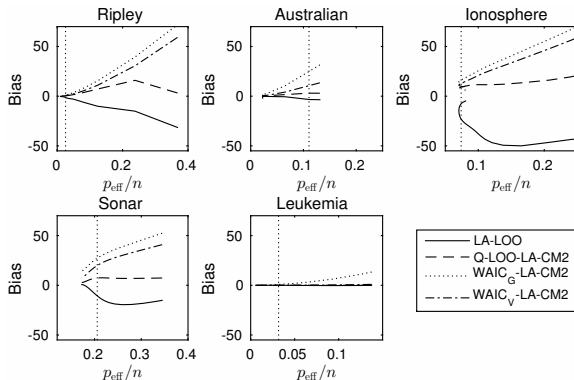
**Figure:** Bias when the target is brute-force-LOO with Laplace and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters  $p_{\text{eff}}/n$ . The flexibility of the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

# EP results with fixed hyperparameters



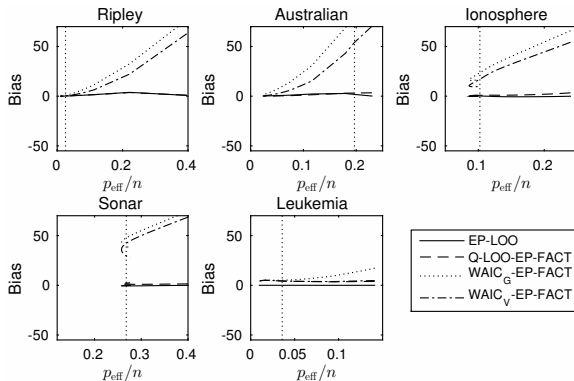
**Figure:** Bias when the target is brute-force-LOO with EP and varying flexibility of the model. Model flexibility was varied by rescaling the length scale(s) in the GP model. Model flexibility is measured by the relative effective number of parameters  $p_{\text{eff}}/n$ . The flexibility of the MAP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

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^S p(\tilde{y}_i|D_{-i}, \phi^s) w_i^s}{\sum_{s=1}^S w_i^s}, \quad (13)$$

where  $w_i^s$  are importance weights and

$$w_i^s \propto \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}, \quad (14)$$

# 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

$$p(\tilde{y}_i|x_i, D_{-i}) \approx \frac{\sum_{s=1}^S p(\tilde{y}_i|D_{-i}, \phi^s) w_i^s}{\sum_{s=1}^S w_i^s}, \quad (13)$$

where  $w_i^s$  are importance weights and

$$w_i^s \propto \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}, \quad (14)$$

- ▶ The LOO predictive density simplifies to

$$p(y_i|x_i, D_{-i}) \approx \frac{1}{\frac{1}{S} \sum_{s=1}^S \frac{1}{p(y_i|x_i, D_{-i}, \theta^s, \phi^s)}} \quad (15)$$

# 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	<b>0.2</b> (0.1)	<b>3.4</b> (0.4)	<b>-0.1</b> (0.1)	<b>-0.13</b> (0.06)	<b>0.56</b> (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	<b>0.2</b> (0.1)	<b>3.4</b> (0.4)	<b>-0.1</b> (0.1)	<b>-0.13</b> (0.06)	<b>0.56</b> (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	<b>0.42</b> (0.14)	<b>7.3</b> (1.4)	<b>0.8</b> (0.6)	<b>-0.24</b> (0.14)	<b>0.49</b> (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.

# 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

# 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 improvement methods don't fix this problem

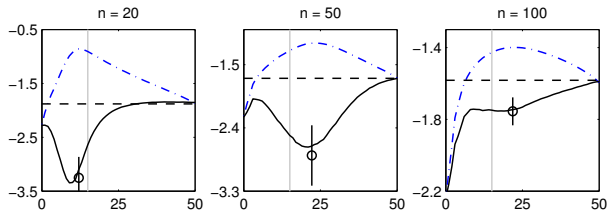


# Summary

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

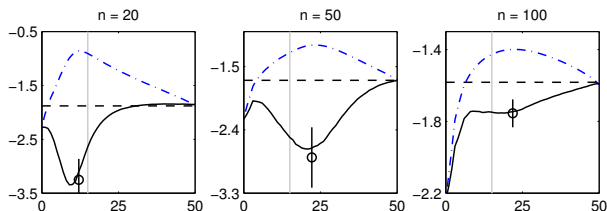
# Warning

- ▶ LOO-CV can be used to compare a small set of models
- ▶ For a large number of models
  - ▶ the selection process will cause overfitting
  - ▶ the inference conditional on the selected model is wrong



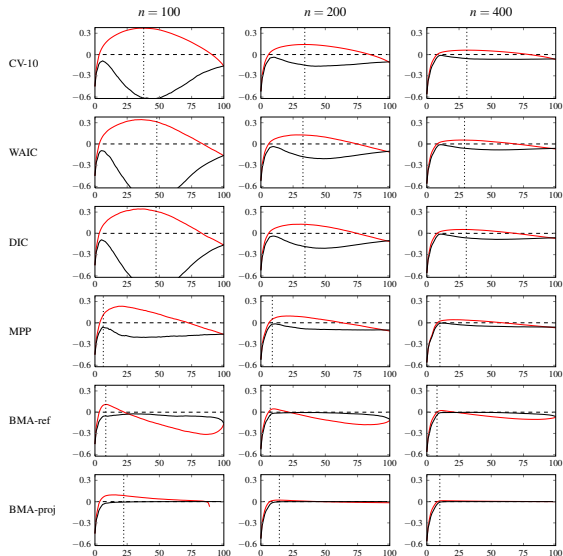
# Warning

- ▶ LOO-CV can be used to compare a small set of models
- ▶ For a large number of models
  - ▶ the selection process will cause overfitting
  - ▶ the inference conditional on the selected model is wrong



- ▶ Use instead a projection predictive approach  
Piironen, J., and Vehtari, A. (2016b). Projection predictive input variable selection for Gaussian process models. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE International Workshop on*, doi:10.1109/MLSP.2016.7738829. arXiv preprint arXiv:1510.04813.

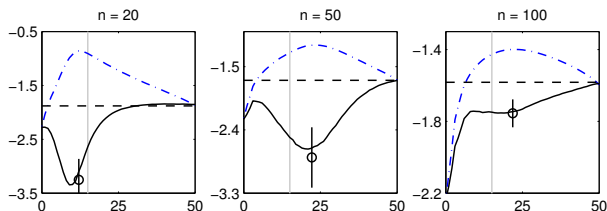
# Selection induced bias in variable selection



Piironen & Vehtari (2016)

# Warning

- ▶ LOO-CV can be used to compare a small set of models
- ▶ For a large number of models
  - ▶ the selection process will cause overfitting
  - ▶ the inference conditional on the selected model is wrong



- ▶ Use instead a projection predictive approach  
Piironen, J., and Vehtari, A. (2016b). Projection predictive input variable selection for Gaussian process models. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE International Workshop on*, doi:10.1109/MLSP.2016.7738829. arXiv preprint arXiv:1510.04813.

# References

- Piironen, J. and Vehtari, A. (2016a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.
- Piironen, J. and Vehtari, A. (2016b). Projection predictive input variable selection for gaussian process models. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE International Workshop on*.
- Siivola, E., Vehtari, A., Vanhatalo, J., and González, J. (2017). Bayesian optimization with virtual derivative sign observations. *arXiv:1704.00963*.
- Stan Development Team (2017). Stan: A C++ library for probability and sampling, version 2.16.
- Vehtari, A., Gelman, A., and Gabry, J. (2016a). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv:1507.04544*.
- Vehtari, A., Mononen, T., Tolvanen, V., and Winther, O. (2016b). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research*, 17(103):1–38.