

A case study of the widely applicable Bayesian information criterion and its optimality

Tommi Mononen

Received: 28 June 2013 / Accepted: 22 March 2014 / Published online: 9 April 2014
© Springer Science+Business Media New York 2014

Abstract In Bayesian statistics, the marginal likelihood (evidence) is one of the key factors that can be used as a measure of model goodness. However, for many practical model families it cannot be computed analytically. An alternative solution is to use some approximation method or time-consuming sampling method. The widely applicable Bayesian information criterion (WBIC) was developed recently to have a marginal likelihood approximation that works also with singular models. The central idea of the approximation is to select a single thermodynamic integration term (power posterior) with the (approximated) optimal temperature $\beta^* = 1/\log(n)$, where n is the data size. We apply this new approximation to the analytically solvable Gaussian process regression case to show that the optimal temperature may depend also on data itself or other variables, such as the noise level. Moreover, we show that the steepness of a thermodynamic curve at the optimal temperature indicates the magnitude of the error that WBIC makes.

Keywords WBIC · Gaussian process regression · Thermodynamic integration · Log marginal likelihood

1 Introduction

In model selection the marginal likelihood is used as a model fitting criterion. It tells us, how well our set of observations (data) fits to our model. This is not an equivalent task with the predictive model selection, where one tries to find a model that predicts best the next observation. For numerical reasons,

usually the logarithm of the marginal likelihood is computed and the measure is called the negative log marginal likelihood or the Bayes free energy. It can be written rigorously as

$$\mathcal{F} = -\log \int \prod_{i=1}^n p(X_i|w) p(w) dw, \quad (1)$$

where X_i is a sample, n is the data size and w is a parameter or a parameter vector. The above formula also assumes that the likelihood can be partitioned. For many model families there do not exist analytical solutions for the criterion and therefore actual values have to be computed either by sampling or by approximating.

Gaussian process (GP) models can be used for regression and classification tasks (Rasmussen and Williams 2006). These models consist of latent variables as well as hyperparameters, which makes a situation more complicated. In the full Bayesian treatment, we integrate over all unknowns, which means in the GP setting analytically intractable computations. In general, a slow MCMC sampling approach is considered to give the most accurate value of the log marginal likelihood. However, with Gaussian process models this option has turned out not to be very viable (Kuss and Rasmussen 2005). The problem is that with the two levels of unknown variables, efficient chain mixing is very hard to achieve, because of the strong coupling between these two sets. In practice, accurate values can be computed using annealed importance sampling (AIS) that has a close connection to thermodynamic integration (Kuss and Rasmussen 2005; Filippone 2013). The thermodynamic integration method (defined later in this section) is itself considered to be one of the best ways to compute an accurate value of the log marginal likelihood (Calderhead and Girolami 2009), even though it is slow as it needs a lot of sampled estimates. In the GP framework also Hybrid Monte Carlo (HMC) (Filippone et al. 2012) and elliptical slice sam-

T. Mononen (✉)
Department of Biomedical Engineering and Computational Science,
Aalto University School of Science, P.O. Box 12200, 02150 Espoo,
Finland
e-mail: tommy.j.mononen@aalto.fi

pling (Filippone 2013) approaches are considered to work very well. The elliptical slice sampling method is automatic in a sense that it needs no tuning (Murray and Adams 2010). Yet another methodology with large datasets is to use additionally a sparse Gaussian process approximation to ease the computational burden (Quiñonero-Candela and Rasmussen 2005).

The other largely used main approach is to simplify the original problem by using fixed hyperparameter values. The hyperparameters are commonly optimized to the maximum of a marginal posterior (type II MAP) or to the maximum of a marginal likelihood (type II ML). The log marginal likelihood of the regression problem using the Gaussian observation model becomes analytically tractable and the other regression and classification models with non-Gaussian observation models can be solved time-efficiently with the Laplace approximation (LA) or the expectation propagation (EP) method. Although both approximations rely on the Gaussian assumption (a posterior is relatively close to a Gaussian distribution), the latter one that minimizes local KL-divergence, is shown to be much better (Kuss and Rasmussen 2005; Nickisch and Rasmussen 2008). It gives very accurate approximations for the log marginal likelihood. Notice that usability of LA and EP methods are not restricted to fixed hyperparameter settings as for example the INLA package has shown us with respect to the Laplace approximation (Rue et al. 2009). However, we bypass all the sampling difficulties in this paper by using the analytically solvable GP regression models (with fixed hyperparameters), as they do everything that is needed in our case.

In this paper we focus on the performance of a new sampling based approximation method called Widely Applicable Bayesian Information Criterion (WBIC) that Professor Sumio Watanabe recently introduced (Watanabe 2013):

Definition 1 WBIC can be written as

$$\text{WBIC} = -\frac{\int \log(p(\mathbf{X}|w))p(\mathbf{X}|w)^\beta p(w)dw}{\int p(\mathbf{X}|w)^\beta p(w)dw},$$

where $\beta = \beta^* = \frac{1}{\log(n)}$.

If β is not fixed, the above formula can be interpreted as a posterior expectation over the log likelihood function $\mathbb{E}_w^\beta(\log p(\mathbf{X}|w))$, where β is the temperature (related to the thermodynamic integration, see below) between zero and one and n is the data size. Watanabe calls β the inverse temperature, but we adopt a more straightforward term 'temperature' (used also in Friel and Pettitt 2008). The WBIC criterion has its roots in the singular learning theory and therefore it is designed to also work with singular models. In the case of regular (non-singular) statistical models the first terms of its series expansion coincide with the Bayesian Information Criterion (BIC) (Watanabe 2013).

Thermodynamic integration is generally used to obtain an accurate sampling estimate of the logarithm of the marginal likelihood (Friel and Pettitt 2008). It is based on the idea that a definite integral within the range from 0 to 1 can be estimated in a piecewise manner given a monotonically rising sequence of temperatures $0 = \beta_0 < \beta_1 < \dots < \beta_J = 1$. Integrals within each sub-range are then estimated using a linear approximation, the trapezoidal rule

$$\int_{\beta_i}^{\beta_{i+1}} g(v)dv = \frac{(\beta_{i+1} - \beta_i)}{2} (g(\beta_i) + g(\beta_{i+1})).$$

Now as

$$-\int_0^1 \mathbb{E}_w^\beta (\log p(\mathbf{X}|w)) d\beta = \mathcal{F},$$

we obtain an accurate estimate for the marginal likelihood as long as a β -sequence is dense enough in the areas where $g(v)$ changes rapidly. One rather good and simple scheme to select temperatures is to set $\beta_t = (t/n)^5$, where n is the data size and t is the index from 0 to n (Friel et al. 2013), but there is also a faster way to compute the log marginal likelihood. Watanabe shows in Watanabe (2013) that there exists a β^{opt} for which the following equivalence is true:

$$-\mathbb{E}_w^{\beta^{\text{opt}}} (\log p(\mathbf{X}|w)) = \mathcal{F}. \tag{2}$$

This intuitively says that we can compute the marginal likelihood just by computing one thermodynamic term using optimal temperature β^{opt} . Equation (2) is true because the

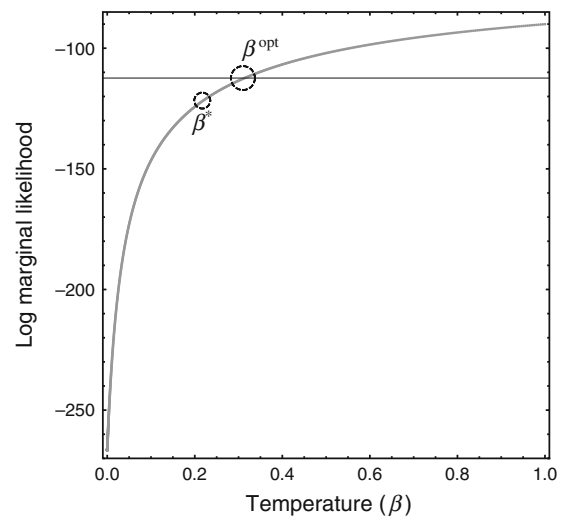


Fig. 1 A point where a thermodynamic curve meets the true log marginal likelihood value (horizontal solid line) gives the optimal temperature β^{opt} . In this example, WBIC uses the point $\beta^* \approx 0.217$, which is slightly off

left hand side is a continuous increasing function and with different values of β it gets both smaller and larger values than the right hand side (Fig. 1). The definition of WBIC arises from this observation and the criterion uses an approximated optimal temperature from Definition 1. WBIC can be further approximated with a BIC-looking formula (Watanabe 2013):

Definition 2 An approximation for WBIC is

$$\text{WBIC}_{\text{approx}} = -\log(p(\mathbf{X} | w_0)) + \lambda \log(n),$$

where w_0 is a parameter vector that minimizes the Kullback–Leibler divergence from a true distribution to a statistical model and λ is a learning coefficient.

The quantity w_0 is not necessary just a single parameter vector, but it can be also a set of parameter vectors (distributions) (Watanabe 2009). Singular learning theory uses a concept of the true distribution, but separates cases where the true distribution is included in a model family or excluded from it (Watanabe 2009). The historical burden of the true model concept is heavy, because many ad-hoc methods have not made this separation. However, one can also argue that in real life, distributions change in time and therefore the assumption of a static distribution is too unrealistic. Here we are examining the goodness of WBIC and therefore the validity of the theoretical foundation is irrelevant to this research, unless it explains some observed phenomenon.

The coefficient λ can be approximated in the following way:

$$\begin{aligned} \hat{\lambda} &= -\frac{\mathbb{E}_w^{\beta_1}(\log(p(\mathbf{X} | w))) - \mathbb{E}_w^{\beta_2}(\log(p(\mathbf{X} | w)))}{1/\beta_1 - 1/\beta_2} \\ &= \beta_1 \beta_2 \left(\frac{\mathbb{E}_w^{\beta_1}(\log(p(\mathbf{X} | w))) - \mathbb{E}_w^{\beta_2}(\log(p(\mathbf{X} | w)))}{\beta_1 - \beta_2} \right) \\ &= \beta_1 \beta_2 U(\beta_1, \beta_2), \end{aligned}$$

where $\beta_1 = a/\log(n)$ and $\beta_2 = b/\log(n)$ and a and b are positive constants. The first form was introduced in Watanabe (2013) without giving any real suggestion how the constants should be selected. However, the forms after it give more insight as function $U(\beta_1, \beta_2)$ can be seen as a derivative approximation for

$$g'(\beta) = -D_\beta(\mathbb{E}_w^\beta(\log(p(\mathbf{X} | w)))) \text{ at } \beta = \frac{1}{\log(n)},$$

while $a = 1 - \epsilon$ and $b = 1 + \epsilon$ and ϵ is a positive constant close to zero. Relying on this interpretation, it is evident that selection of the constants have some effect on the value of $\hat{\lambda}$. Otherwise $\beta_1 \beta_2$ and the derivative would cancel out each other to maintain $\hat{\lambda}$ as a fixed constant. Moreover, the difference between constants a and b must be greater, if the approximation is computed using a sampling method (due to

inaccuracy caused by stochasticity), but as we handle only analytically computable cases in this paper, we can use values that are very close to actual limits. In the experimental part, we discuss more closely of this derivative interpretation.

We aim to show that even if the temperature $\beta^* = 1/\log(n)$ is optimal in general (without knowledge of a used model family), it is non-optimal for example in the case of Gaussian process regression. Model family specific features can have huge effects on the selection of the optimal β and a training data size is not the only affecting feature.

2 Gaussian process regression

Let us assume that we have surfaces like in Fig. 2. The data is a sample of random locations (x, y) and the corresponding values $u_T(x, y)$ added with some Gaussian noise $\sigma_n \mathcal{N}(0, 1)$, where σ_n is a constant weight (noise level) and $T \in \{A, B\}$. Surface B is more resistant to higher noise levels (the structure of the surface does not disappear so easily) than surface A as the range in function values (z-axis) is larger and the polynomial is simpler. In real life a surface is not known, but it has to be estimated using data.

A Gaussian process is a finite set of random variables, which have a joint Gaussian distribution (Rasmussen and Williams 2006). The Gaussian process can be fully specified with mean and covariance functions. The first one is assumed here to be a constant function with value zero and the latter one is describing pairwise relationships between points. In this paper, we assume the familiar squared exponential covariance function:

$$k(\mathbf{z}_i, \mathbf{z}_j) = \sigma_s^2 \exp\left(-\frac{1}{2} \sum_{k=1}^d \frac{(\mathbf{z}_{i,k} - \mathbf{z}_{j,k})^2}{l_k^2}\right), \tag{3}$$

where σ_s^2 is a signal variance, l_k is a length scale and d is the dimensionality of the sample points. The GP prior is denoted as $p(\mathbf{f} | \mathbf{D}) = \mathcal{N}(\mathbf{f}; 0, \mathbf{K})$, where \mathbf{K} is a covariance matrix (a matrix of pairwise covariances over data points), \mathbf{f} is a vec-

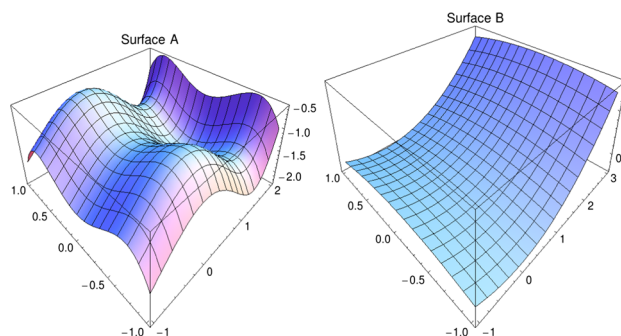


Fig. 2 Two polynomial surfaces: $u_A(x, y) = \frac{1}{4}x^4 - x^2 - 2y^4 + 2y^2 - 1$ and $u_B(x, y) = x^2 - y^2 - 1$

tor of latent function values and \mathbf{D} consists of covariate data. The prior defines initial probabilities of different latent functions living in the same space as our observations, where the covariance function determines the properties of these functions. The likelihood that incorporates data into the model is $p(\mathbf{y} | \mathbf{f}, \mathbf{D}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2 \mathbf{I})$, where \mathbf{y} is a vector of observed values. A set of hyperparameters $(l_1, l_2, \sigma_s, \sigma_n)$ is optimized to their MAP values to make computation easier (the Matlab toolbox GPstuff (Vanhatalo 2013) is used in computations). The posterior now gives probabilities over functions that take into account data, and the mean function is our regression function, the latent surface.

The model selection as well as computation of general GP models need a good and reliable approximation of the marginal likelihood. The Bayesian information criterion (BIC) cannot be used as GP models are non-parametric and therefore a parameter count does not tell the true dimensionality of a model. To bypass this problem, we investigate the applicability of WBIC in approximating the log marginal likelihood. The main advantage in WBIC is that we can compute the log marginal likelihood without making any Gaussian assumption (EP or Laplace approximation) for the posterior.

3 Exact form of WBIC for GP regression

We derive a WBIC formula for the GP regression only to compare the possible goodness and validity of WBIC for the Gaussian process regression. As the analytic marginal likelihood formula can be derived in the Gaussian observation model case (Rasmussen and Williams 2006), we can efficiently plot solution surfaces achieved by varying hyperparameters. Our main target is to recognize possible problems and failures of WBIC without any doubts of sampling accuracy or too sparse sampling. The approximation itself has only comparison significance, because exact solutions can be computed equally fast.

We start the derivation by transforming the general expressions into GP regression equivalents

$$\mathbb{E}_{\mathbf{f}}^{\beta}(\log \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2 \mathbf{I})) = \frac{\int_{-\infty}^{\infty} (\log \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2 \mathbf{I})) \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2 \mathbf{I})^{\beta} \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}) d\mathbf{f}}{\int_{-\infty}^{\infty} \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_n^2 \mathbf{I})^{\beta} \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}) d\mathbf{f}}$$

and as the final result we achieve the following theorem:

Theorem 1 *WBIC function with a free parameter β for Gaussian process regression is*

$$\text{WBIC}_{\text{GP}}(\beta) = \frac{n}{2} \left(\log(2\pi\sigma_n^2) + \frac{1}{\beta} \right) + \frac{\sigma_n^2}{2\beta^2} \left(\|\mathbf{A}\mathbf{y}\|^2 - \text{tr}(\mathbf{A}) \right),$$

where a matrix \mathbf{A} is

$$\mathbf{A} = \left(\left(\frac{\sigma_n^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1}.$$

Proof see “Appendix”. □

If we fix $\beta = \beta^* = 1/\log(n)$, we get the WBIC. Theorem 1 gives the form we are using in analyses and experiments. We tested the validity of the theorem also by performing full thermodynamic integrations for several different cases. It gave correct log marginal likelihood values, which gives an additional empirical confirmation that derivation of function WBIC_{GP} is correct.

4 Deriving the optimal temperature using a simplified covariance matrix

Let us derive a tailored optimal β in the GP regression case, which we will denote by $\hat{\beta}$. A derivation of the optimal temperature for the full case is too complicated, because β is not only in a role of a multiplier for matrix \mathbf{A} , but it stays also inside matrix elements. The Taylor expansion trick for the matrix inverse does not help either as the resulting series diverges in this case. Therefore, we consider a heavily pruned covariance matrix. We make a highly unrealistic assumption that non-diagonal covariances are zero, which leaves us only constant variance terms σ_s^2 on the diagonal (Formula 3). We do this derivation purely to show that this new $\hat{\beta}$ is a function of latent noise. After this, we compare the result to the full covariance matrix setting and show that both, a data size and magnitude of latent noise, is needed to infer the optimal temperature.

Our covariance matrix is now $\mathbf{K} = \sigma_s^2 \mathbf{I}$. The task is to derive expressions for the exact log marginal likelihood and for WBIC in this particular case. The resulting functions should both give the equivalent value when WBIC is using the optimal temperature. This of course, gives us possibility to solve the temperature analytically.

Lemma 1 *The Bayes free energy (negative log marginal likelihood) for the pruned GP regression model is*

$$\mathcal{F}_{\text{prunedGP}} = \frac{1}{2(\sigma_s^2 + \sigma_n^2)} \mathbf{y}^T \mathbf{y} + \frac{n}{2} \log(2\pi(\sigma_s^2 + \sigma_n^2)).$$

Proof We start from the analytic marginal likelihood formula (with the Gaussian observation model), which is

$$-\log p(\mathbf{y} | \mathbf{X}) = \frac{1}{2} \mathbf{y}^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} + \frac{1}{2} \log |\mathbf{K} + \sigma_n^2 \mathbf{I}| + \frac{n}{2} \log(2\pi).$$

By plugging in diagonal matrix $\mathbf{K} = \sigma_s^2 \mathbf{I}$ we obtain

$$\begin{aligned} & \frac{1}{2} \mathbf{y}^\top ((\sigma_s^2 + \sigma_n^2) \mathbf{I})^{-1} \mathbf{y} + \frac{1}{2} \log |(\sigma_s^2 + \sigma_n^2) \mathbf{I}| \\ & + \frac{n}{2} \log(2\pi) \\ & = \frac{1}{2} \mathbf{y}^\top \left(\frac{1}{(\sigma_s^2 + \sigma_n^2)} \mathbf{I} \right) \mathbf{y} + \frac{1}{2} \log \prod (\sigma_s^2 + \sigma_n^2) \\ & + \frac{n}{2} \log(2\pi) \\ & = \frac{1}{2(\sigma_s^2 + \sigma_n^2)} \mathbf{y}^\top \mathbf{y} + \frac{n}{2} \log(2\pi(\sigma_s^2 + \sigma_n^2)) \end{aligned}$$

□

The first part is data related as the latter part is a more general term without a direct data relationship. Next we do the same operation for the general WBIC formula.

Lemma 2 *The WBIC function for the pruned GP regression model is*

$$\begin{aligned} & \text{WBIC}_{\text{prunedGP}}(\beta) \\ & = \frac{n}{2} \log(2\pi \sigma_n^2) + \frac{n}{2\beta} - \frac{n\sigma_n^2}{2\beta^2 \sigma_s^2 + 2\beta \sigma_n^2} \\ & + \frac{\sigma_n^2}{2(\beta \sigma_s^2 + \sigma_n^2)^2} \mathbf{y}^\top \mathbf{y}. \end{aligned}$$

Proof We start by manipulating matrix \mathbf{A} :

$$\mathbf{A} = \left(\left(\frac{\sigma_n^2}{\beta} + \sigma_s^2 \right) \mathbf{I} \right)^{-1} = \frac{1}{\frac{\sigma_n^2}{\beta} + \sigma_s^2} \mathbf{I} = \frac{\beta}{\sigma_n^2 + \beta \sigma_s^2} \mathbf{I}.$$

After this we plug in diagonal matrix \mathbf{A} to the WBIC formula:

$$\begin{aligned} & \text{WBIC}_{\text{pruned}}(\beta) = \frac{n}{2} \left(\log(2\pi \sigma_n^2) + \frac{1}{\beta} \right) \\ & + \frac{\sigma_n^2}{2\beta^2} \left(\left\| \frac{\beta}{\sigma_n^2 + \beta \sigma_s^2} \mathbf{I} \mathbf{y} \right\|^2 - \text{tr} \left(\frac{\beta}{\sigma_n^2 + \beta \sigma_s^2} \mathbf{I} \right) \right) \\ & = \frac{n}{2} \left(\log(2\pi \sigma_n^2) + \frac{1}{\beta} \right) \\ & + \frac{\sigma_n^2}{2\beta^2} \left(\left(\frac{\beta}{\sigma_n^2 + \beta \sigma_s^2} \right)^2 \mathbf{y}^\top \mathbf{y} - \frac{n\beta}{\sigma_n^2 + \beta \sigma_s^2} \right) \\ & = \frac{n}{2} \log(2\pi \sigma_n^2) + \frac{n}{2\beta} - \frac{n\sigma_n^2}{2\beta^2 \sigma_s^2 + 2\beta \sigma_n^2} \\ & + \frac{\sigma_n^2}{2(\beta \sigma_s^2 + \sigma_n^2)^2} \mathbf{y}^\top \mathbf{y} \end{aligned}$$

□

The last term is directly data dependent and the rest of the terms are not. Now let us derive a result that does not depend

directly on data. This operation is done to make computations simpler and to derive a neat looking function for the optimal temperature, which also seems to work better than β^* with full covariance matrices (see Sect. 6). We discard the data depending terms from the log marginal likelihood and WBIC to have the equivalence

$$\begin{aligned} & \frac{n}{2} \log(2\pi(\sigma_s^2 + \sigma_n^2)) \\ & = \frac{n}{2} \log(2\pi \sigma_n^2) + \frac{n}{2\beta} - \frac{n\sigma_n^2}{2\beta^2 \sigma_s^2 + 2\beta \sigma_n^2}. \end{aligned}$$

The optimal temperature can be easily solved from this equation.

Theorem 2 *The optimal temperature $\hat{\beta}$ for the pruned GP regression case without data depending terms is*

$$\hat{\beta} = \frac{1}{\log \left(1 + \frac{\sigma_s^2}{\sigma_n^2} \right)} - \frac{\sigma_n^2}{\sigma_s^2}.$$

Proof

$$\begin{aligned} & \frac{n}{2} \log(2\pi(\sigma_s^2 + \sigma_n^2)) = \frac{n}{2} \log(2\pi \sigma_n^2) + \frac{n}{2} \\ & - \frac{n\sigma_n^2}{2\beta^2 \sigma_s^2 + 2\beta \sigma_n^2} \\ & \log \left(1 + \frac{\sigma_s^2}{\sigma_n^2} \right) = \frac{\beta \sigma_s^2 + \sigma_n^2 - \sigma_n^2}{\beta^2 \sigma_s^2 + \beta \sigma_n^2} \\ & (\beta^2 \sigma_s^2 + \beta \sigma_n^2) \log \left(1 + \frac{\sigma_s^2}{\sigma_n^2} \right) = \beta \sigma_s^2 \\ & \beta \left[\beta \sigma_s^2 \log \left(1 + \frac{\sigma_s^2}{\sigma_n^2} \right) + \sigma_n^2 \log \left(1 + \frac{\sigma_s^2}{\sigma_n^2} \right) - \sigma_s^2 \right] = 0 \end{aligned}$$

Solving the first order equation inside the square brackets with respect to β gives the answer. The other solution ($\beta = 0$) is not a valid one as it ignores the likelihood entirely. □

We can immediately notice that the solution does not depend on the data size at all. Increasing the data size does not help to predict the values of other points. This is clearly an unrealistic result, and it is caused by the simplifications we have made. However, notice that $\hat{\beta}$ seems to be a function related to latent noise and signal variance. This derivation is essentially equal to the one that with WBIC has been made in a more general setting by truncating series expansions to consist only of general data independent terms (Watanabe 2013). One problem is that the data dependent parts (actually \mathbf{y} independent) that we just discarded, differ from each other sometimes by a great amount. If the data dependent terms in the computation formulas are roughly of the same size in the optimal temperature, they do not have effect on its value. The derivation therefore implicitly expects that we

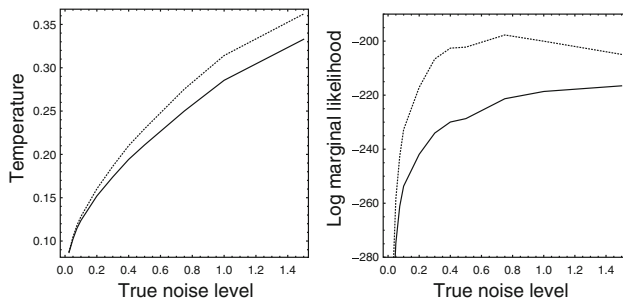


Fig. 3 On the left, a solid curve shows mean $\hat{\beta}^{\text{opt}}$ (mean of 50 runs) and a dotted line represents suboptimal mean temperatures ($\hat{\beta}$), where data dependent terms are ignored. On the right, WBIC function values are computed using the optimal (equals to the true log marginal likelihoods) and suboptimal temperatures with $n = 150$ (for surface B)

can have a data independent optimal temperature, but next we will investigate the difference of the cropped terms to show that the assumption is not always justified.

First, because we are using the standardized values of \mathbf{y} , it follows by substituting $\bar{y} = 0$ (mean) and $\text{std} = 1$ (standard deviation) to the formula of unbiased standard normal deviation that

$$\text{std} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

$$1 = \frac{1}{n-1} \sum_{i=1}^n (y_i - 0)^2$$

$$n - 1 = \mathbf{y}^T \mathbf{y}$$

Therefore, we do not need the actual data vector, only its size. Second, we look at the limits of \mathbf{y} dependent terms as $\sigma_n^2 \rightarrow 0$ (Lemmas 1 and 2). The term in the log marginal likelihood goes to $-(n-1)/2\sigma_s^2$ and the term in WBIC vanishes. They are not equal in this case, which means that discarding of those terms can be harmful. Last, let us look at Fig. 3, where data independent optimal temperatures are compared with data dependent optimal temperatures (the full solution $\hat{\beta}^{\text{opt}}$, which depends also on n is presented in the ‘‘Appendix’’). The noise level is affecting the goodness of our log marginal likelihood approximation via an incorrect optimal temperature. With greater noise levels the slope of a thermodynamic curve is getting milder, which leads to a smaller error, even if the error in temperatures is still increasing (this is explained in the next section more closely).

As the data dependent terms are not used in the derivation of the optimal temperature, it inevitably leads to estimation errors. Other message of our analysis is that a data size is not always the dominating feature, but it can happen that the optimal temperature may rather be a function of some other model family related variables. Next we do experi-

ments using full covariance matrices to gain more support for these findings.

5 Experiments

Our target is to show that the phenomenon of noise dependence, we discovered with the simplified covariance matrix, exists also in the general setting with full covariance matrices. The optimal temperatures in this case have been found using an optimization algorithm that finds the temperature, which produces minimum error between the log marginal likelihood and the WBIC function. We compare the optimized temperature (β^{opt}) and β^* against each other. In this section, we do not anymore use the results of Lemmas 1 and 2 were obtained using a rough oversimplification, but the correct WBIC formula (see Theorem 1) for GP regression models with full covariance matrices. Everything is therefore data dependent, except β^* itself. We use in experiments the example surfaces presented in Fig. 2 with the given ranges. All those cases, where covariance matrices turned out to be singular or the hyperparameter optimization failed (in the Matlab toolbox GPstuff), were excluded from experiments and replaced with new runs using different datasets.

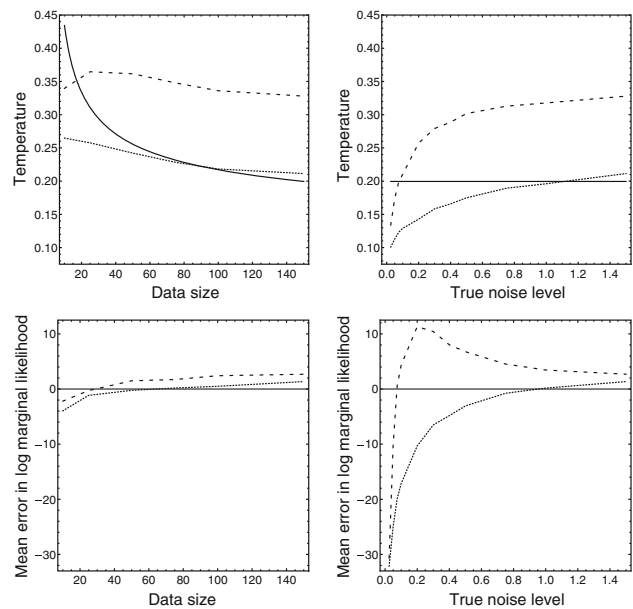


Fig. 4 On the top row a solid line is β^* , a dashed line is mean β^{opt} for surface A and a dotted line is mean β^{opt} for surface B. The corresponding mean errors between marginal likelihood and WBIC are on the bottom row. Dashed line represents errors computed for surface A and dotted line errors computed for surface B. In data size plots (left column) noise is fixed to 1.5 and in noise plots (right column) data size is fixed to 150. All the dotted and dashed lines represent the mean values of 50 runs with different data sets

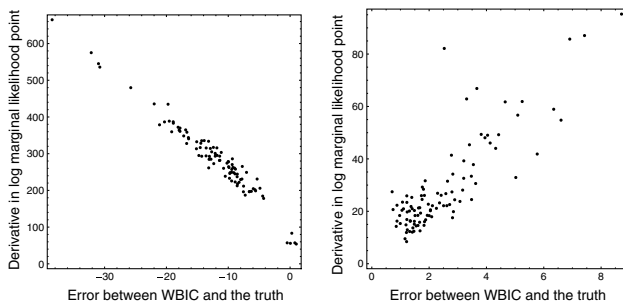


Fig. 5 On the *left* a low noise case $\sigma_n = 0.05$ and on the *right* a high noise case $\sigma_n = 0.4$. We have computed for 100 data sets an approximate derivative of a thermodynamic integration curve at the truth against the error that WBIC makes (surface A, $n = 50$)

Let us start with the left column in Fig. 4. The foremost observation is that the optimal temperatures do not seem to converge into β^* with respect to n . In addition, the optimal temperature curves are not on top of each other, but different polynomial surfaces seem to give temperature curves of different vertical shifts, although the general trend seems to match: larger temperatures for small data sets and smaller temperatures for larger data sets. However, despite the difference between β^* and β^{opt} , the marginal likelihood approximations seem to be quite good when the noise around the latent surface is large. The situation gets worse if we fix dataset size to 150 and look at the marginal likelihood approximation with respect to different levels of noise (right column). While there is not much noise, errors are much larger and the worst situation is when we have low noise (highly informative data) and tiny data sets, because then the relative error raises (the marginal likelihoods are close to zero, see Fig. 6). It is quite intuitive that, if the data is informative then data dependent terms in marginal likelihood and its approximations are likely to be important and they cannot be discarded. Even with an increased level of noise, the temperature is still non-stabilized to a constant value, which suggests that the optimal temperature should be a function of n as well as σ_n^2 (Fig. 6).

To track down a deeper cause of the failure, let us look at individual datasets instead of their mean behavior. If the temperature is not accurate and a thermodynamic integration curve is steep in the area of the true log marginal likelihood, then we obtain a bigger error as can be seen from Fig. 5. The plots are comparing approximate derivatives (difference quotients) of thermodynamic curves at the optimal temperatures (β^{opt}) against errors in the WBIC estimates and we can see that the errors are higher for larger derivative values. As noise increases the slopes of thermodynamic curves are not so steep anymore, which can be seen in the values of derivatives (on the right). The error changes sign, which causes a change in the slope direction between the

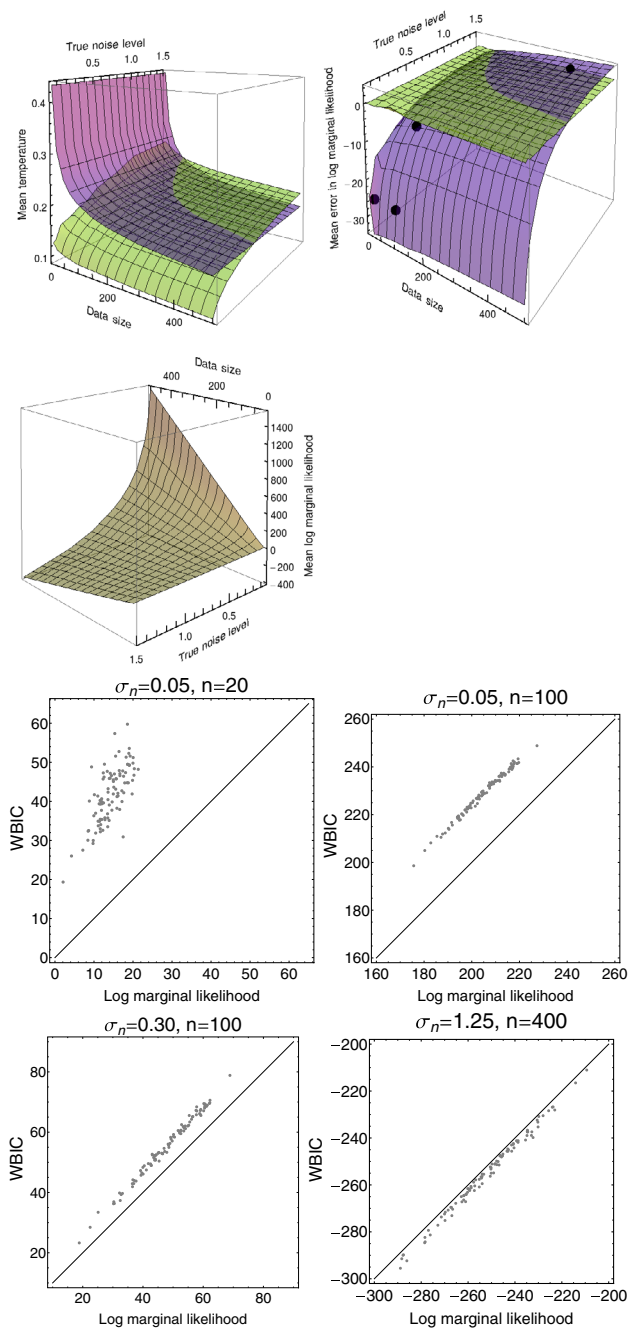


Fig. 6 Plots for surface B: The *purple* surfaces present WBIC’s optimal β^* and the corresponding mean errors in log marginal likelihood values (averages taken over 100 datasets per each hyperparameter setting). The *yellow* surfaces present the true optimal β (acquired by optimizing temperatures algorithmically). The *orange* surface represents the means of correct log marginal likelihood values. The *black dots* indicate places on the *purple* surface, from where example cross-plots are taken. *Points* in the cross-plots correspond different data sets generated with the given parameter values (Color figure online)

panels. To give more insight, we can write the approximation of WBIC (using the derivative interpretation) in the form

$$\begin{aligned}
 & -\log(p(\mathbf{X} | w_0)) + \hat{\lambda} \log(n) \\
 &= -\log(p(\mathbf{X} | w_0)) + \frac{a}{\log(n)} \frac{b}{\log(n)} U(\beta_1, \beta_2) \log(n) \\
 &\approx -\log(p(\mathbf{X} | w_0)) + \frac{g'(\beta^*)}{\log(n)},
 \end{aligned}$$

where the approximation penalizes overfitting using the derivative of the thermodynamic curve at point β^* . The plots are therefore partly showing residual effects caused by using β^* instead of the true optimal temperature β^{opt} .

Next, we look at the cross-plots in Fig. 6. The variance of the error seem to be very low: the estimates are only biased. The only exception is the case where we have small informative datasets (low noise). For a dataset of this kind, a WBIC estimate can be quite bad as it is biased and the variance is high.

In many cases, too accurate approximation formulas are overly sensitive to small inaccuracies in parameter estimation and it is better to use truncated series as Roos and Zou have noticed in the case of BIC and its expansions (Roos and Zou 2013). Therefore we also tested accuracy of $\text{WBIC}_{\text{approx}}$ using values $a = 0.9999$ and $b = 1.0001$. As we have ruled out any singular covariance matrices we know that there is unique hyperparameter setting w_0 minimizing Kullback–Leibler divergence and we expect it to be our optimized MAP hyperparameters. Then we can write

$$\begin{aligned}
 \text{WBIC}_{\text{GPapprox}} &= \frac{n}{2} \log(2\pi \sigma_n^2) \\
 &+ \frac{1}{2} (\mathbf{y} - \mathbb{E}(\mathbf{f}))^\top (\sigma_n^2 I)^{-1} (\mathbf{y} - \mathbb{E}(\mathbf{f})) + \hat{\lambda} \log(n),
 \end{aligned}$$

where we have replaced latent vectors \mathbf{f} with their expected values given by a GP model. The WBIC approximation seems to be better for surface A in a mean sense as it does not exhibit a positive error peak (Fig. 7). For surface B the full WBIC gives better results. Standard deviations of errors stay at the equal level, whether we are using WBIC or its approximation, which means that the approximation is truly working better with surface A.

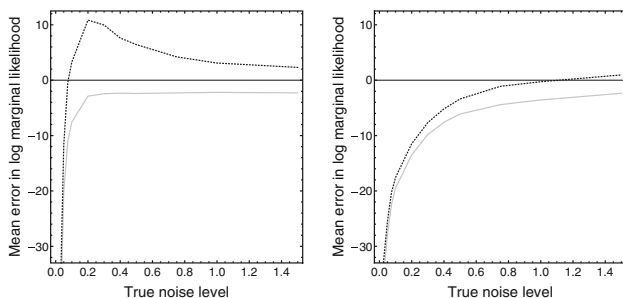


Fig. 7 The WBIC approximation (gray line) versus the full WBIC with respect to surfaces A (left) and B (right). The data size is 150

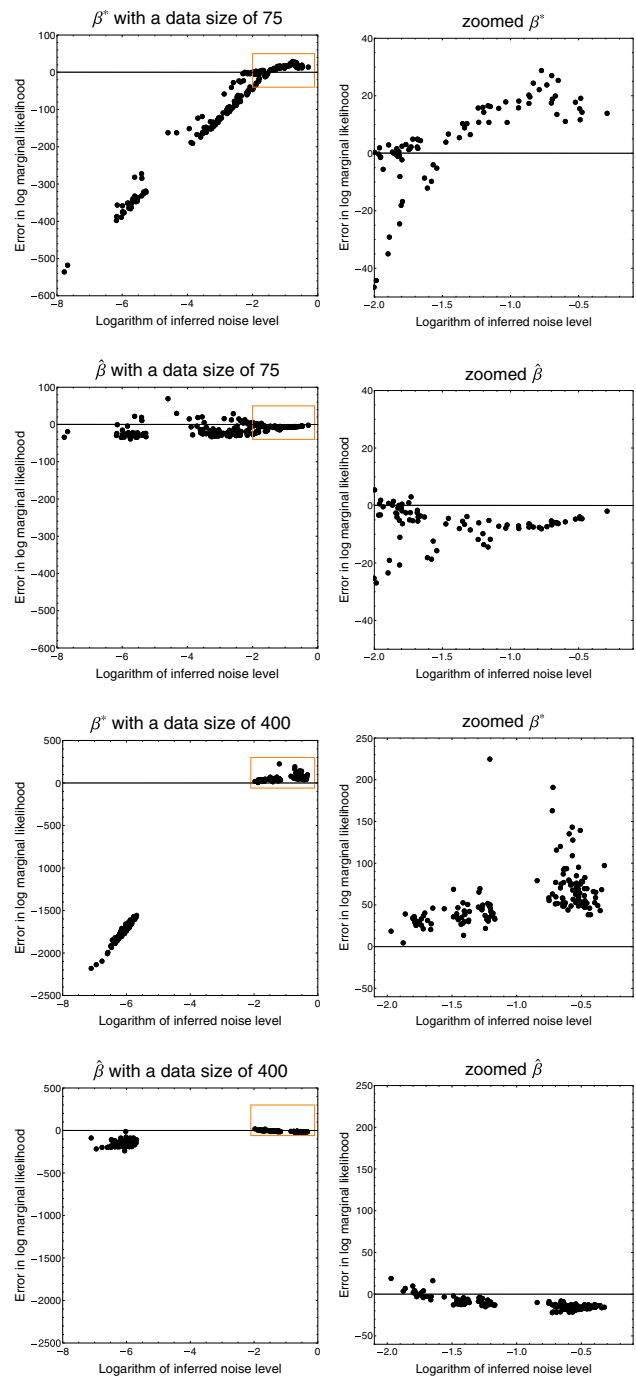


Fig. 8 β^* against $\hat{\beta}$ with real data (UCI: wine, random samples of 75 and 400 data points). Orange rectangles in the left column show zoomed areas represented in the right column. Notice the scale change in the y-axis between cases of 75 and 400 data points (Color figure online)

6 Experiment with real data

We show with a real dataset that the WBIC breakdown takes place also in real situations. Our data is the white wine dataset from the UCI machine learning repository (Cortez et al. 2009). It consists of 11 real valued variables that describe dif-

ferent properties of white wines. Many of these properties are quite correlated with each other, but not all. Our experiment is the following: we take one property in turn and build a GP regression model for it using the other properties as covariates. Moreover, we do not use the whole data, but select randomly sampled 75 (or 400) data points out of it and do this 20 times. Hence, we have 11 regression models (related to predicted variables) and as we use 20 different datasets, the total number of models is 220. Then we compare log marginal likelihoods against the WBIC formula results for temperatures β^* and $\hat{\beta}$ (Theorem 2). However, also the latter one is computed using full covariance matrices. All covariates as well as the target variable are standardized. We tested that the covariate standardization does not essentially change the results.

With real data, we do not know the true noise level, but we rely on model given noise estimate σ_n^2 . It is immediately evident that β^* is failing when the noise level is low (Fig. 8), hence in the cases where models fit really well. The phenomenon is not as strong when we are using temperature $\hat{\beta}$. Moreover, the artificial cases already showed that there is a sweet zone (top right panel), where β^* coincides with β^{opt} , but after that the error changes sign and starts rising again, until noise overwrites almost all the information from data and WBIC again approaches the log marginal likelihood.

Let us look at the case with more data (400 data points). The first observation is that points are in two main clusters. The points in a region of tiny noise correspond to the targets that can be described well with covariates and the points in a large noise region are the targets which are more difficult to describe. With smaller data, these two regions were more connected, because of stochasticity present in small sample sets. Notice that the other 11 targets start to separate from each other (notice the other visible clusters of ca. 20 points). The general behavior of failing β^* seems to be similar in 75 and 400 data point cases. However, in the region of realistic noise, temperature $\hat{\beta}$ performs much better than β^* . From this, we can draw the conclusion that at least in this randomly chosen real data example, our oversimple approximation for the optimal temperature is better than β^* . Hence, signal and noise variances approximate the optimal temperature better than the data size, albeit all of these variables are needed in the correct full solution.

7 Conclusions

We used the artificial setting, where the log marginal likelihood can be easily calculated to show that the WBIC approximation fails in some occasions. This paper suggests that there can also be other variables than data size n that have their contribution to the optimality of the temperature and therefore the optimal temperature should be derived the model family dependent way. Moreover, discarding data dependent terms,

while deriving this optimal temperature, can also make estimates worse. If the correct value exists in the steep part of a thermodynamic curve, the optimal temperature approximation should be more accurate to maintain the overall level of accuracy. Further studies are needed to investigate, if the results in this paper are applicable also to other model families. However, if lowered accuracy of WBIC will be observed with some other models, then one possible action is to check how dependent the optimal temperature is on other relevant variables.

Acknowledgments The author wishes to thank Daniel Simpson, Arno Solin, Aki Vehtari and the anonymous reviewers for providing valuable comments on this article.

Appendix

Proof of Theorem 1:

Computing the denominator of WBIC for GP

In the following, we have simplified notation by setting $\sigma_n^2 = \sigma^2$ as variable σ_s^2 is inside matrix \mathbf{K} and therefore it does not emerge in the proof. Let us start to compute β -weighted marginal likelihood:

$$\begin{aligned} & \int_{-\infty}^{\infty} p(\mathbf{y} | \mathbf{f}, \mathbf{D})^\beta p(\mathbf{f} | \mathbf{D}) \, d\mathbf{f} \\ &= \int_{-\infty}^{\infty} \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I})^\beta \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}) \, d\mathbf{f} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^{\beta n} |\sigma^2 \mathbf{I}|^\beta}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{f})^\top (\frac{\sigma^2}{\beta} \mathbf{I})^{-1}(\mathbf{y}-\mathbf{f})} \\ & \quad \times \frac{1}{\sqrt{(2\pi)^n |\mathbf{K}|}} e^{-\frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f}} \, d\mathbf{f} \\ &= (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \\ & \quad \times \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{f})^\top (\frac{\sigma^2}{\beta} \mathbf{I})^{-1}(\mathbf{y}-\mathbf{f}) - \frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f}} \, d\mathbf{f}. \end{aligned}$$

Then complete the sum of squared forms to square (Petersen and Pedersen 2008, 8.1.7 Sum of two squared forms)

$$\begin{aligned} & -\frac{1}{2}(\mathbf{f}-\mathbf{y})^\top \left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} (\mathbf{f}-\mathbf{y}) - \frac{1}{2}(\mathbf{f}-\mathbf{0})^\top \mathbf{K}^{-1} (\mathbf{f}-\mathbf{0}) \\ &= -\frac{1}{2} \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} \mathbf{y} \right)^\top \end{aligned}$$

$$\begin{aligned} & \times \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right) \\ & \times \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right) \\ & + \frac{1}{2} \left(\mathbf{y}^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right) \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \\ & \times \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right) - \frac{1}{2} \left(\mathbf{y}^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right) \end{aligned}$$

By using the matrix inversion lemma (Rasmussen and Williams 2006, A.3 Matrix Identities) to the second term we get

$$\begin{aligned} & - \frac{1}{2} \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right)^\top \\ & \times \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right) \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right. \\ & \left. \times \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right) - \frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y} \end{aligned}$$

Now we continue with the integral after substituting the previous results and using after that an integration rule of a Gaussian integral (Petersen and Pedersen 2008, 8.1.1 Density and normalization)

$$\begin{aligned} & (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \times \int_{-\infty}^{\infty} \exp \left(-\frac{1}{2} \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right. \right. \\ & \left. \left. \times \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right)^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right) \right. \\ & \left. \times \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} \right) \right) d\mathbf{f} \\ & = (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \times (2\pi)^{\frac{n}{2}} \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right|^{-\frac{1}{2}} \\ & = (2\pi)^{-\frac{\beta n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \times \left(\left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right| \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right| \left| \mathbf{K}^{-1} \right| \right)^{-\frac{1}{2}} \\ & = (2\pi)^{-\frac{\beta n}{2}} \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right|^{\frac{\beta}{2}} \left| \mathbf{K}^{-1} \right|^{\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y}} \end{aligned}$$

$$\begin{aligned} & \times \left(\left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right| \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right| \left| \mathbf{K}^{-1} \right| \right)^{-\frac{1}{2}} \\ & = \frac{1}{\sqrt{(2\pi)^{\beta n} \left| \mathbf{K} + \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) \right|}} \sqrt{\frac{\left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right|^{\beta}}{\left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \right|}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\mathbf{K} + \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) \right)^{-1} \mathbf{y}} \\ & = \text{DENOM} \end{aligned}$$

Notice that by setting $\beta = 1$ and taking the logarithm, we end up to the well-known log marginal likelihood formula.

Computing the numerator of WBIC for GP

$$\begin{aligned} & \int_{-\infty}^{\infty} \log p(\mathbf{y} | \mathbf{f}, \mathbf{D}) p(\mathbf{y} | \mathbf{f}, \mathbf{D})^\beta p(\mathbf{f} | \mathbf{D}) d\mathbf{f} \\ & = \int_{-\infty}^{\infty} \log \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I})^\beta \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}) d\mathbf{f} \\ & = \int_{-\infty}^{\infty} \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right. \\ & \quad \left. - \frac{1}{2} (\mathbf{y} - \mathbf{f})^\top (\sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{f}) \right) \\ & \quad \times \frac{1}{\sqrt{(2\pi)^{\beta n} |\sigma^2 \mathbf{I}|^\beta}} \mathbf{e}^{-\frac{1}{2} (\mathbf{y} - \mathbf{f})^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} (\mathbf{y} - \mathbf{f})} \\ & \quad \times \frac{1}{\sqrt{(2\pi)^n |\mathbf{K}|}} \mathbf{e}^{-\frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f}} d\mathbf{f} \\ & = \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right) (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} \\ & \quad \times |\mathbf{K}|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mathbf{e}^{-\frac{1}{2} (\mathbf{y} - \mathbf{f})^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} (\mathbf{y} - \mathbf{f}) - \frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f}} d\mathbf{f} \\ & \quad - \frac{1}{2} \int_{-\infty}^{\infty} (\mathbf{y} - \mathbf{f})^\top (\sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{f}) (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} \\ & \quad \times |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} (\mathbf{y} - \mathbf{f})^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} (\mathbf{y} - \mathbf{f}) - \frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f}} d\mathbf{f} \\ & = \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right) \times \text{DENOM} \\ & \quad - \frac{1}{2} (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \quad \times \int_{-\infty}^{\infty} (\mathbf{y} - \mathbf{f})^\top (\sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{f}) \\ & \quad \times \exp \left(-\frac{1}{2} \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right. \right. \end{aligned}$$

$$\begin{aligned} & \times \left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} \mathbf{y}^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} + \mathbf{K}^{-1} \right) \\ & \times \left(\mathbf{f} - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I}\right)^{-1} \mathbf{y} \right) d\mathbf{f} \end{aligned}$$

By using Matrix Cookbook formula 357 (Petersen and Pedersen 2008, 8.2.2 Mean and variance of square forms) and by compensating missing constants we will have

$$\begin{aligned} & \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right) \times \text{DENOM} \\ & - \frac{1}{2} (2\pi)^{-\frac{(\beta+1)n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} |\mathbf{K}|^{-\frac{1}{2}} \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \times (2\pi)^{\frac{n}{2}} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \Big|^{-\frac{1}{2}} \\ & \times \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right)^\top (\sigma^2 \mathbf{I})^{-1} \\ & \times \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right) \\ & + \text{tr} \left((\sigma^2 \mathbf{I})^{-1} \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right) \\ = & \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right) \times \text{DENOM} \\ & - \frac{1}{2} (2\pi)^{-\frac{\beta n}{2}} |\sigma^2 \mathbf{I}|^{-\frac{\beta}{2}} \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \Big|^{-\frac{1}{2}} \left| \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right|^{-\frac{1}{2}} \\ & \times \mathbf{e}^{-\frac{1}{2} \mathbf{y}^\top \left(\frac{\sigma^2}{\beta} \mathbf{I} + \mathbf{K} \right)^{-1} \mathbf{y}} \\ & \times \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right)^\top (\sigma^2 \mathbf{I})^{-1} \\ & \times \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right) \\ & + \text{tr} \left((\sigma^2 \mathbf{I})^{-1} \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right) \\ = & \text{DENOM} \times \left(-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma^2 \mathbf{I}| \right. \\ & - \frac{1}{2} \left[\text{tr} \left((\sigma^2 \mathbf{I})^{-1} \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \right) \right. \\ & + \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right)^\top (\sigma^2 \mathbf{I})^{-1} \\ & \left. \left. \times \left(\left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} + \mathbf{K}^{-1} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right)^{-1} \mathbf{y} - \mathbf{y} \right) \right] \right) \end{aligned}$$

$$= \text{DENOM} \times (-\mathbf{WBIC})$$

In the last line, we observe that as the numerator is divided by the denominator and the numerator has the denominator as a multiplier, then the rest of the numerator must be (negative) WBIC itself. After a manipulation (using the matrix inversion lemma to the all three appropriate terms) we achieve a numerically stable version

$$\begin{aligned} -\mathbf{WBIC} = & -\frac{n}{2} \log(2\pi\sigma^2) \\ & - \frac{1}{2\beta} \left[\text{tr} \left(1 - \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \left(\frac{\sigma^2}{\beta} \mathbf{I} \right) \right) \right. \\ & \left. + \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y} \right)^\top \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \mathbf{y} \right], \end{aligned}$$

which can be further manipulated by taking stuff outside of the trace and interpreting the latter term as a matrix norm. As a result we obtain a neat form

$$\begin{aligned} \mathbf{WBIC} = & \frac{n}{2} \left(\log(2\pi\sigma^2) + \frac{1}{\beta} \right) \\ & + \frac{\sigma^2}{2\beta^2} \left(\|\mathbf{A} \mathbf{y}\|^2 - \text{tr}(\mathbf{A}) \right), \text{ where} \\ \mathbf{A} = & \left(\left(\frac{\sigma^2}{\beta} \mathbf{I} \right) + \mathbf{K} \right)^{-1} \end{aligned}$$

□

The exact solution to the pruned GP case

The following exact solution is computed with Mathematica by equating the results of Lemmas 1 and 2

Lemma 3 *The optimal temperature for the pruned GP regression case with all the terms included is*

$$\begin{aligned} \hat{\beta}^{\text{opt}} = & - \left(-2 \mathbf{y}^\top \mathbf{y} \sigma_n^2 + \sigma_s^2 n W + 2n \sigma_n^2 W \log \frac{\sigma_n^2}{W} + \sqrt{W} \right. \\ & \times \sqrt{4(\mathbf{y}^\top \mathbf{y})^2 \sigma_n^2 + \sigma_s^4 n^2 W + 4 \mathbf{y}^\top \mathbf{y} n \sigma_n^2 W \log \frac{W}{\sigma_n^2}} \\ & \left. / \left(2\sigma_s^2 \left(-\mathbf{y}^\top \mathbf{y} + n W \log \frac{\sigma_n^2}{W} \right) \right), \right) \end{aligned}$$

$$\text{where } W = (\sigma_s^2 + \sigma_n^2).$$

References

Calderhead, B., Girolami, M.: Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data Anal.* **53**, 4028–4045 (2009)

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.* **47**, 547–553 (2009)
- Filippone, M.: Bayesian inference for Gaussian process classifiers with annealing and exact-approximate MCMC. [arXiv:1311.7320](https://arxiv.org/abs/1311.7320) (2013)
- Filippone, M., Zhong, M., Girolami, M.: A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Mach. Learn.* **93**, 93–114 (2012)
- Friel, N., Pettitt, A.N.: Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Series B (Stat. Methodol.)* **70**, 589–607 (2008)
- Friel, N., Hurn, M., Wyse, J.: Improving power posterior estimation of statistical evidence. *Stat. Comput.* (2013). doi:[10.1007/s11222-013-9397-1](https://doi.org/10.1007/s11222-013-9397-1)
- Kuss, M., Rasmussen, C.E.: Assessing approximate inference for binary Gaussian process classification. *J. Mach. Learn. Res.* **6**, 1679–1704 (2005)
- Murray, I., Adams, R.P.: Slice sampling covariance hyperparameters of latent Gaussian models. *Adv. Neural Inf. Process. Syst.* **23**, 1723–1731 (2010)
- Nickisch, H., Rasmussen, C.E.: Approximations for binary Gaussian process classification. *J. Mach. Learn. Res.* **9**, 2035–2078 (2008)
- Petersen, K.B., Pedersen, M.S.: *The Matrix Cookbook*. Version: November 14 (2008)
- Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6**, 1939–1959 (2005)
- Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA (2006)
- Roos, T., Zou, Y.: Keep it simple stupid—On the effect of lower-order terms in BIC-like criteria. *Information Theory and Applications Workshop*, February 2013, San Diego, USA (2013)
- Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Series B (Stat. Methodol.)* **71**, 319–392 (2009)
- Vanhatalo, J., et al.: GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **14**, 1175–1179 (2013)
- Watanabe, S.: *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, Cambridge (2009)
- Watanabe, S.: A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897 (2013)