Lossy compression of observations for Gaussian process regression

Emile Visser[1*], *Corné E.* van Daalen[1], and *J. C.* Schoeman[1]

[1]Department of Electrical & Electronic Engineering, Stellenbosch University, South Africa

Abstract. This paper proposes a novel approach of Gaussian process observation set compression based on a squared difference measure. It is used to discard observations to speed up Gaussian process prediction while retaining the information encoded in the full set of observations. Furthermore, this paper compares the regression performance of a compressed Gaussian process to its uncompressed version and to a randomly downsampled Gaussian process for a standard two-dimensional test function. The empirical results of this paper show that this is an effective algorithm for Gaussian process compression, speeding up prediction while maintaining predictive accuracy with regards to the predicted means.

1 Introduction

A Gaussian process (GP) is an extension of the multidimensional Gaussian distribution to infinitely many dimensions [1]. Using an appropriate choice of covariance function (or kernel) and a set of observed function values for given inputs (i.e., set of input-output pairs), a GP can be used for regression [2]. In addition, a GP can then produce univariate Gaussian predictions (reflecting the uncertainty of the prediction) for previously unobserved inputs.

This GP regression was first used in the work of Danie G. Krige (as such, GP regression is also known as Kriging) where he sought to determine the distribution of gold in the Witwatersrand reef complex using a small set of gold concentration measurements from boreholes. To do this, he fitted a GP model to these measurements and calculated points for which the GP model predicted that there would likely be a high concentration of gold [3].

Calculating these predictions, however, are computationally expensive (each prediction has a computational complexity of $O(n^2)$, where n is the number of observed points [1]). It would therefore be of interest to reduce the number of observations in a GP model to increase prediction speed while preserving predictive accuracy. Since there will always be a trade-off between predictive accuracy and computational efficiency of the GP model, a good strategy to determine which observations to discard is required (which must be superior to naïve random dropout). This is analogous to lossy compression in image processing [4], where data is discarded while maintaining image quality as much as possible, rather than randomly discarding pixels.

To the best of our knowledge, existing literature does not address GP compression. The closest work related to our work consists either of speculative, computationally expensive

* Corresponding author: 21595240@sun.ac.za

numerical approximations or measures between GPs and target functions to quantify regression performance [5] which are of limited use if the target function is not known. Other studies use difference metrics such as the Bhattacharyya distance that is not applied to GP regression (notably, in spectral [6] or time series analysis [7]).

This paper derives a novel closed-form expression for the squared difference of two predicted GP means. This measure is used to develop a compression algorithm that discards points such that the surface of the predicted mean is changed the least between compression iterations. This reduces the set of observations for GP regression while preserving the predictive accuracy of the original model.

The rest of the paper is organised as follows: Sec. 2 contains the preliminary knowledge regarding GP regression and KL-divergence necessary to understand the novel contributions of the squared difference measure and compression algorithm derived in Sec. 3. This algorithm is then qualitatively evaluated by applying it to the Himmelblau test function and these results are discussed, followed by conclusions regarding the performance/restrictions of this algorithm and future work.

2 Preliminaries

This section contains the prerequisite knowledge upon which the novel contributions of this paper will be based. Specifically, the theory of GP regression that is used to approximate a target function with a set of observations and the theory of the KL-divergence that quantifies the difference between the predictions of two GPs.

2.1 Gaussian process regression

To construct a GP regression model to approximate a general target function $f\colon \mathbb{R}^{\boldsymbol{n}} \to \mathbb{R}$, the following components are required: an appropriate choice of kernel function k, observations $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$ with vector of corresponding observed values $\mathbf{y} = [y_1, y_2, \dots, y_p]^\top \in f(\mathbf{x}_i) = y_i$, matrix $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j)) \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{k}_* = (k(\mathbf{x}_*, \mathbf{x}_i)) \in \mathbb{R}^{p\times 1}$. The mean and variance equations that describe a predictive univariate Gaussian distribution from this GP at a test point $\mathbf{x}_* \in \mathbb{R}^{\boldsymbol{n}}$ are given by [1]:

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{y} \tag{1}$$

and

$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_*^\top. \tag{2}$$

In this paper, the commonly used radial basis function (RBF, also known as the squared exponential) is chosen as the kernel,

$$k(\mathbf{x}_*, \mathbf{x}) = \exp\left(-\frac{(\mathbf{x}_* - \mathbf{x})^\top \cdot (\mathbf{x}_* - \mathbf{x})}{2l}\right). \tag{3}$$

This function has the useful property that the closed form for the definite integral of this function exists. This property will be used in Sec. 2.3.

2.2 Kullback–Leibler divergence

The Kullback–Leibler (KL) divergence describes the information loss when one probability distribution is used to approximate another. While not a true metric due to being asymmetric,

it is non-negative and is zero for identical distributions. For the case of two univariate Gaussian distributions, the KL-divergence is given by [8]

$$D_{\mathrm{KL}}(p||q) = \ln(\sigma_2) - \ln(\sigma_1) + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \quad \text{where} \quad p \sim N(\mu_1,\sigma_1) \quad \text{and} \quad q \sim N(\mu_2,\sigma_2). \tag{4}$$

This can be used in a simple calculation for the difference in information between the two GP predictions at a single queried point (which are also defined by two univariate Gaussian distributions). However, for a very large or potentially infinite set of test points, this computation becomes prohibitively expensive.

3 Derivation of GP compression approach

This section consists of the novel contributions of this paper: the mathematical derivation of the squared difference measure quantifying the relative importance of each GP observation and the compression algorithm that uses this measure to reduce the set of observations used during GP regression.

3.1 Squared difference measure derivation

Quantifying the effect that a single observation has on the GP regression model involves comparing the predictions of a model that contains the observation and one that does not. In this framework, if the predictions are very similar the observation does not have a very important effect on the GP model. As mentioned in Sec. 2.2, the naïve method to compare these two GPs would be to take a large set of test points and calculate the KL-divergence between the predictions obtained from the two GPs.

For a more ideal method, the GP predictive equations (Eqs. 1,2) are substituted into Eq. 4. Taking the integral of this formula should yield the KL divergence (and by extension, the information loss) between the surfaces defined by the predictions of two GPs over region $\boldsymbol{R}$:

$$D_{\mathrm{KL}}(\mathrm{GP}_1||\mathrm{GP}_2) = \int_{\boldsymbol{R}} \ln\left(\sqrt{\sigma_2^2(\mathbf{x}_*)}\right) d^n\mathbf{x}_* - \int_{\boldsymbol{R}} \ln\left(\sqrt{\sigma_1^2(\mathbf{x}_*)}\right) d^n\mathbf{x}_* + \int_{\boldsymbol{R}} \frac{\sigma_1^2(\mathbf{x}_*)+\left(\mu_1(\mathbf{x}_*)-\mu_2(\mathbf{x}_*)\right)^2}{2\sigma_2^2(\mathbf{x}_*)} d^n\mathbf{x}_* - \frac{1}{2}\int_{\boldsymbol{R}} d^n\mathbf{x}_*. \tag{5}$$

This integration seems to have (to the extent of the authors' knowledge) no analytic solution due to, among others, the root in the logarithmic terms. However, note that the squared difference of predicted means component seems to be a measure that could be utilized. This is also a widely used measure of fit for statistical regression models, such as in the method of least squares [9]. This measure (defined as D_{SD}) and its expansion becomes

$$D_{\mathrm{SD}}(\mathrm{GP}_1||\mathrm{GP}_2) = \int_{\boldsymbol{R}} \left(\mu_1(\mathbf{x}_*) - \mu_2(\mathbf{x}_*)\right)^2 d^n\mathbf{x}_* \tag{6}$$

$$D_{\mathrm{SD}}(\mathrm{GP}_1||\mathrm{GP}_2) = \int_{\boldsymbol{R}} \left(\mu_1(\mathbf{x}_*)\right)^2 d^n\mathbf{x}_* - 2\int_{\boldsymbol{R}} \mu_1(\mathbf{x}_*)\mu_2(\mathbf{x}_*) d^n\mathbf{x}_* + \int_{\boldsymbol{R}} \left(\mu_1(\mathbf{x}_*)\right)^2 d^n\mathbf{x}_*. \tag{7}$$

Note in Eq. 1 that $\mathbf{K}^{-1}\mathbf{y}$ is a constant product and, for notational brevity, redefine the predictive mean equations for two different GPs (with p observations respectively) as:

$$\mu_1(\mathbf{x}_*) = \mathbf{k}_1^{\mathsf{T}}\boldsymbol{\alpha}_1 \tag{8}$$

and

$$\mu_2(\mathbf{x}_*) = \mathbf{k}_2^\top \boldsymbol{\alpha}_2. \quad (9)$$

After taking the integral of the inner products in Eqs. 8 and 9 and exploiting linearity, the general forms of the integration terms in Eq.7 are given below. Note that α_{1i} refers to the i-th element of the vector $\boldsymbol{\alpha}_1$ and likewise for the vectors $\boldsymbol{\alpha}_2$, $\mathbf{k}_1$ and $\mathbf{k}_2$.

$$\int((\mu_a(\mathbf{x}_*))^2)\, d\mathbf{x}_* = \sum_{i=1}^{p}\sum_{j=1}^{i}\begin{cases}\alpha_{1i}^2 \int_{\boldsymbol{R}} (k_{ai})^2 d^n\mathbf{x}_*, & i = j \\ 2\alpha_{1i}\alpha_{1j}\int_{\boldsymbol{R}} k_{ai}k_{aj} d^n\mathbf{x}_*, & i \neq j\end{cases} \text{ where } a \in \{1,2\} \quad (10)$$

$$\int \mu_1(\mathbf{x}_*)\cdot\mu_2(\mathbf{x}_*)\, d\mathbf{x}_* = \sum_{i=1}^{p}\sum_{j=1}^{i}\alpha_{1i}\alpha_{2j}\int_{\boldsymbol{R}} k_{1i}k_{2j} d^n\mathbf{x}_* \quad (11)$$

Note that these equations require the integrals of the GP kernel functions to exist. As stated in Sec. 2.1, the radial basis function (Eq. 3) is chosen for this paper, since the definite integral of this function over a domain can be written as a product of the definite integrals over each of the dimensions of $\mathbf{x}_* \in \mathbb{R}^{\boldsymbol{n}}$:

$$\int_{\boldsymbol{R}} k(\mathbf{x}_*, \mathbf{x}) d^n\mathbf{x}_* = \int_a^b k(x_{*1}, x_1)\, dx_{*1} \int_c^d k(x_{*2}, x_2)\, dx_{*2} \ldots \int_y^z k(x_{*k}, x_k)\, dx_{*n}. \quad (12)$$

Each of these integral terms can readily be expressed in terms of the error function using the same kernel integral as Bayesian quadrature [10]:

$$\int_a^b k(x_*, x)\, dx_* = \frac{\sqrt{\pi}}{2}[\operatorname{erf}(u)]\Big|_{\frac{1}{\sqrt{2}l}(a-x)}^{\frac{1}{\sqrt{2}l}(b-x)} \quad (13)$$

In summary, the definite integrals of the chosen kernel function (in this case Eq. 13) are substituted into Eqs. 10 and 11, which are in turn substituted into Eq. 7 to yield the full squared difference measure D_{SD} that is used in the next section to rank the relative information content of each observation in a GP model.

3.2 Compression algorithm definition

With the squared difference measure in hand that can determine the difference between the predicted means of two GPs, the relative importance of each observation with regards to this predicted surface can be determined. This is done for each observation by applying the measure to a GP with a full set of observed values and a GP with the subset of observations that excludes the observation in question.

Intuitively, this gives a measure that describes how much the predicted mean of the original GP will change if an observation is discarded, with observations containing important information having a higher score than those that are redundant. Given this ranking of observations, we can simply discard the observation with the lowest score and repeat the process until a suitable termination condition is reached. Concisely, one iteration of this novel compression algorithm is defined as

$$X_{\mathrm{iter}+1} = X_{\mathrm{iter}} \backslash \mathbf{x}_{\min}$$
$$\text{where } \mathbf{x}_{\min} = \underset{\mathbf{x}_{\min} \in X_{\mathrm{iter}}}{\operatorname{argmin}} D_{\mathrm{SD}}\left(\mathrm{GP}_{X_{\mathrm{iter}}} || \mathrm{GP}_{X_{\mathrm{iter}} \backslash \mathbf{x}_{\mathrm{i}}}\right). \quad (14)$$

An illustrative example of this algorithm after halving the observation set can be seen in Fig. 1. These examples show the behaviour that this algorithm seeks to achieve, a predicted mean that consistently approximates the uncompressed GP as much as possible with fewer observations where, in contrast, discarding randomly may yield much poorer approximations.

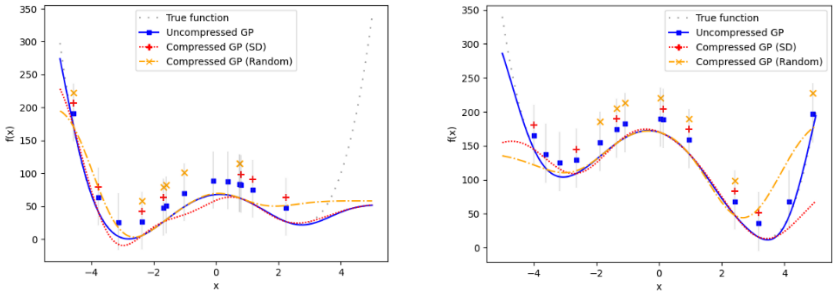

Fig. 1. 1-D examples of predicted means after discarding half of the original observations with squared difference measure (SD) and discarding observations randomly.

4 Performance evaluation

To evaluate the compression algorithm, a GP is used to approximate the target function known as Himmelblau's function ($f(x,y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$) [11], a well-known test function, with 100 observed points randomly chosen over this function's domain. Three compressed GPs are generated by discarding observations using the squared difference measure, as previously described, a numerical approximation of the KL-divergence (Eq. 3) or discarding observations randomly.

Predictive performance of the three GPs is quantified using the root mean squared error (RMSE) between the target function value y_i and the maximum likelihood estimate (MLE) of the predicted Gaussian distribution (i.e., the mean) $\overline{y_{*i}}$ for test point $\mathbf{x}_{*i}$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_{*i}})^2}. \tag{15}$$

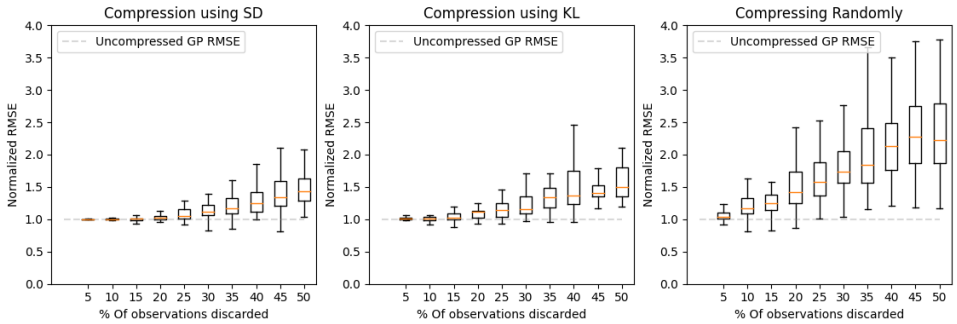

Fig. 2. Box-and-whisker plots of normalized RMSE distributions after compressing GPs with 100 observations using the squared difference measure (left), numerical KL-divergence (middle) and discarding points randomly (right).

These test points are evenly distributed over a 2-D grid over the domain of the Himmelblau function ($x \in [-5,5]$ and $y \in [-5,5]$).

The RMSE of the three new compressed GPs are normalized by the RMSE of the uncompressed GP to obtain the relative predictive performance of the three compressed GPs compared to the uncompressed GP.

The results in Fig. 2 are very promising and clearly show that compression using the squared difference measure performs strictly better than random compression and similarly to the ideal method of compression using the KL-divergence. However, the RMSE performance measure may overestimate the ability of the squared difference compression, since the squared difference measure and the RMSE are based on the squared difference between the prediction and true values. Therefore, for a more general indicator of performance, the experiment is repeated using the mean absolute error. This measure is more robust to outliers in the data and several studies have advocated for its use to supersede the RMSE [12, 13].

The mean absolute error (MAE) is evaluated on the same grid of test points over the domain of the Himmelblau function ($x \in [-5,5]$ and $y \in [-5,5]$), with the MAE given as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y_{*i}}|. \tag{16}$$

Fig. 3 is obtained in the same manner as Fig. 2; the MAE of three new compressed GPs are normalized by the MAE of the uncompressed GP to obtain the relative predictive performance of the three compressed GPs compared to the uncompressed GP.

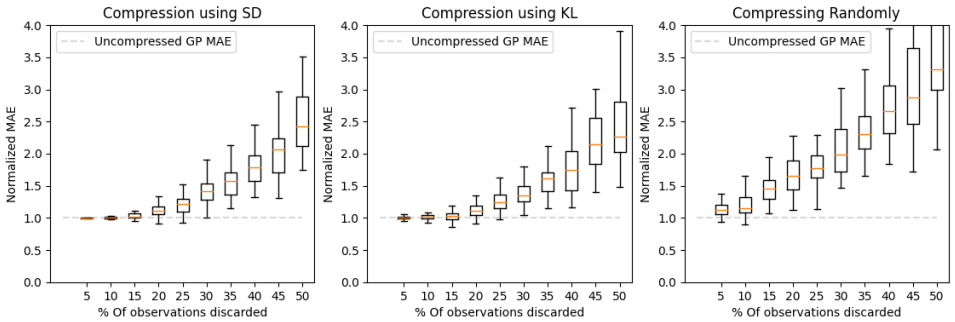

Fig. 3. Box-and-whisker plots of normalized MAE distributions after compressing GPs with 100 observations using the squared difference measure (left), numerical KL-divergence (middle) and discarding points randomly (right).

The results in Fig. 3 agree with those in Fig. 2, with the compression using the squared difference performing better than random compression and similarly to compression using the numerically approximated KL-divergence. In this case, using the squared difference measure, the observation set can be reduced by 25% with only a slight MAE increase.

In the previous evaluations, the MLE of the predicted Gaussian distributions was compared to the target function. This is good enough for most situations where predictions are made with GP regression, however, this method of prediction does not take the variance of the GP prediction into account. Therefore, for a more complete measure of performance, the likelihoods of the target function given the GP models are evaluated.

Defining the negative predictive log likelihood (to obtain a loss function) of the target function y given the set of observations X at test point $\mathbf{x}_*$ as [1]

$$-\log p(y|X,\ \mathbf{x}_*) = \frac{1}{2}\log\left(2\pi\sigma^2(\mathbf{x}_*)\right) + \frac{(y-\mu(\mathbf{x}_*))^2}{2\sigma^2(\mathbf{x}_*)}. \tag{17}$$

This likelihood can be standardized as the standardized log loss (SLL) by subtracting the negative log likelihood obtained from the original GP (X_{full}) from the negative log likelihood

of the compressed GP (X_{compr}), effectively an inverse likelihood ratio. The mean of this value over the test points denoted as the MSLL:

$$\text{MSLL} = \frac{1}{n}\sum_{i=1}^{n}(-\log p(y_i|X_{\text{compr}},\ \mathbf{x}_{*i}) + \log p(y_i|X_{\text{full}},\ \mathbf{x}_{*i})). \quad (18)$$

In the same fashion as the previous experiments, the MSLL is evaluated over the same grid of test points over the domain of the Himmelblau function, with the results plotted in Fig. 4.

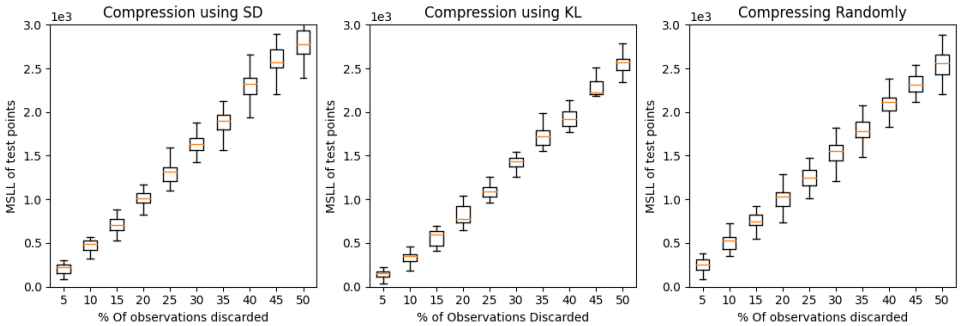

Fig. 4. Box-and-whisker plots of MSLL distributions after compressing GPs with 100 observations using the squared difference measure (left), numerical KL-divergence (middle) and discarding points randomly (right).

For clarity, the means of each distribution in Fig. 4 are superimposed and given in Fig. 5.

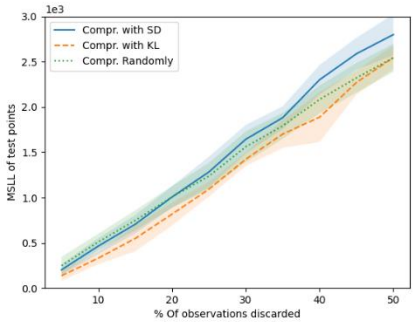

Fig. 5. Superimposed means of each distribution in Fig. 4 with shaded 1-σ boundaries.

Fig. 4 and 5 show that the squared difference and random compression have approximately the same performance degradation under the MSLL measure, with the approximate KL measure having slightly better performance than both. This result is expected, since the KL measure also takes the variance of the GP prediction into account, preserving more of the information in the model and being able to make predictions with more confidence (which leads to a lower MSLL).

In summary, discarding observations in a GP according to the squared difference measure is an effective strategy when MAE estimates (i.e., the means) of the GP predictions are of interest when compared to the approximate KL divergence and random compression. If variance information preservation is desired, the approximate numerical KL-divergence should be used if the computational cost is acceptable.

5 Conclusion

In this paper, a novel definition for the squared difference of the predicted means of two Gaussian processes (derived from the KL-divergence) is used to construct a lossy compression algorithm for Gaussian process observations that preserves regression performance. Considering the results in this paper, it can be concluded that this is an effective method of lossy GP observation compression that can be used to speed up GP prediction while retaining prediction accuracy with regards to the mean of the GP predictions.

This approach of compressing GP observation sets is a new avenue of research, with potential for extension to more general kernel functions, derivation of new measures and compression heuristics.

This method should be compatible with most uses of GPs in practice, but this method can potentially excel in applications with severe memory constraints (i.e., a limit on the observation set) or very fast prediction speed requirements, such as real-time embedded systems. It may, however, struggle when used in applications that place an emphasis on the variance (or certainty) of the GP predictions, such the expected improvement (EI) acquisition function [14] or cases where high confidence thresholds are used. In these cases, the additional computational costs of the approximate KL divergence may be justified.

5.1 Future work

It should be possible to use the minimum value of D_{SD} relative to the overall distribution of D_{SD} at each iteration for early stopping if the compression ratio is not predefined. Intuitively, if the variance in the distribution of D_{SD} is very low (all the scores for each point are similar), there is no longer any point that would be "better" to discard than any other. This, however, has not been tested extensively and remains an avenue for further research.

References

1. C. E. Rasmussen, *Gaussian Processes for Machine Learning*, MIT Press (2006)
2. G. Matheron, *Principles of Geostatistics*, Economic Geology **58**, 1246-1266 (1963)
3. D. G. Krige, *A statistical approach to some basic mine valuation problems on the Witwatersrand*, J. of the Chem., Metal. and Mining Soc. of South Africa **52**, 119-139 (1951)
4. K. Sayood, *Introduction to Data Compression*, Morgan Kaufmann Publishers (1996)
5. X. Hong, J. Gao, X. Jiang, C. J. Harris, *Estimation of Gaussian process regression model using probability distance measures*, Systems Science & Control Engineering **2**, 655-663 (2014)
6. D. Kazakos and P. Papantoni-Kazakos, *Spectral distance measures between Gaussian processes*, IEEE Transactions on Automatic Control **25**, 950-959 (1980)
7. F. C. Schweppe, *On the Bhattacharyya distance and the divergence between Gaussian processes*, Information and Control **11**, 373-395 (1967)
8. S. J. Roberts, W. D. Penny, *Variational Bayes for generalized autoregressive models*, IEEE Trans. Signal Process. **50**, 2245-2257 (2002)
9. M. Mansfield, *A List Of Writings Relating To The Method Of Least Squares: With Historical And Critical Notes (1877)*, Kessinger Publishing (2009)

10. A. O'Hagan, *Bayes-Hermite Quadrature*, Journal of Statistical Planning and Inference, **29**, 245–260 (1991)
11. D. Himmelblau, *Applied Nonlinear Programming*, McGraw-Hill (1972)
12. C. J. Willmott, K. Matsuura. *Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance*, Climate Research **30**, 79-82 (2005)
13. R. G. Pontius., O. Thontteh, H. Chen, *Components of information for multiple resolution comparison between maps that share a real variable*. Environ Ecol. Stat. **15**, 111–142 (2008)
14. D. Zhan, H. Xing, *Expected improvement for expensive optimization: a review*, J Glob Optim ***78***, 507–544 (2020)