9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

HETEROSCEDASTIC GAUSSIAN PROCESS REGRESSION ON THE ALKENONE OVER SEA SURFACE TEMPERATURES

Taehee Lee[1], Charles E. Lawrence[1]

***Abstract*—To restore the historical sea surface temperatures (SSTs) better, it is important to construct a good calibration model for the associated proxies. In this paper, we introduce a new model for alkenone ($U_{37}^{K'}$) based on the heteroscedastic Gaussian process (GP) regression method. Our nonparametric approach not only deals with the variable pattern of noises over SSTs but also contains a Bayesian method of classifying potential outliers.**

I. INTRODUCTION

The alkenone is a widely used proxy for inferring sea surface temperatures (SSTs) in paleoceanography. Haplophytes make long-chain ketone lipids (alkenones) with 37 carbons changing in response to water temperature. Let $U_{37}^{K'}$ be the relative unsaturation index of these compounds as [1]:

$$U_{37}^{K'} = \frac{C_{37:2}}{C_{37:2} + C_{37:3}} \tag{1}$$

Because SSTs have other sources of inferences, what we are interested in is the distribution of $U_{37}^{K'}$ given SST, not the reverse: once the prior information (based on the latitude, for example) of SST is organized as a prior distribution and the distribution as a likelihood, it is possible to integrate those information by computing the posterior distribution of SST given the observed $U_{37}^{K'}$ by the Bayes' rule. Also, if SSTs are somehow correlated (with respect to their spatial pattern, for example) one another, the given proxies can be exploited better than a set of individual posteriors with an associated graphical model.

[1] has lead the way in the application of Bayesian statistics in paleoclimatology. The paper approaches to the problem with a Bayeian B-spline regression model (BAYSPLINE) on $U_{37}^{K'}$ data as well as considering their seasonality, and it shows improved performance to extant methods. However, the model lacks of a concrete rule of deciding the number of basis splines or their orders, and does not deal with heteroscedastic noises over SSTs neatly.

For the last decade, models based on the neural networks (NN) have been rapidly emerging and soon overwhelming all research fields requiring machine learning, including climatology, for the outstanding performance. Though some theoretical works such as [2], [3] and [4] guarantee the effectiveness of NNs in some senses, it remains as a potential problem in practice that parameters to be learned are often too many compared to the amount of available data. Bayesian statistics can give a possible solution of this problem: its philosophy that parameters are random variables allows to marginalize all such unknown values out to get the posterior predictive distributions.

In this point of view, among various nonparametric regression models, Gaussian processes (GP) ([5]) have some advantages which make them distinctive from all the others. Besides their explicit predictive posterior distributions, with some specific kernels corresponding to the choice of activation functions, it becomes a marginalized version of associated NN models: details can be found in [6] and [7]. The point is that only a very few hyperparameters (for example, only three hyperparameters are needed in the squared-exponential kernel) are now controlling the GP regression models, and let data explain themselves to make them free from the structure.

Though tuning hyperparameters can be done efficiently in the homoscedastic (i.e., noises are assumed to have a constant variance.) GP models ([5]), heteroscedastic GP models are not yet clear to learn. While [8] tries to model logarithms of variances from empirical variances based on another GP regression, [9] adapts a variational method which cannot avoid breaking up the completeness of the models. In this paper we suggest a heteroscedastic GP model which is not only intuitive but also easy to learn and apply it to

[1]Division of Applied Mathematics, Brown University, Rhode Island, USA

constructing a new calibration for $\mathrm{U}_{37}^{\mathrm{K}'}$.

There is one golden rule: every Gaussian model is sensitive to outliers in learning. Also, we should be aware of the model misspecification for the robustness, unless the model is guaranteed to follow a certain distribution based on the underlying theoretical arguments. GP models cannot be free from these limitations. Student's t-processes ([10]) can be an alternative, but they do not have explicit posterior predictive distributions if noises are also considered. Instead, classifying and excluding outliers in learning seems to be a better idea. We also introduce a Bayesian method of classifying and treating outliers automatically in the learning procedure.

II. METHOD

Let $\mathcal{X} = \{x_n\}_{n=1}^{\mathrm{N}}$ and $\mathcal{Y} = \{y_n\}_{n=1}^{\mathrm{N}}$ be the SSTs and $\mathrm{U}_{37}^{\mathrm{K}'}$ observations, respectively. What to construct is the regression model which returns the distribution $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ of $\mathrm{U}_{37}^{\mathrm{K}'}$, y, at an arbitrary query SST, x. Let β be the regression function of $\mathcal{Y}$ at $\mathcal{X}$. Then, what we assume are the following:

$$p\left(\mathcal{Y}|\beta,\mathcal{X}\right) = \mathcal{N}\left(\mathcal{Y}|\beta,\Lambda(\mathcal{X})\right) \tag{2}$$

$$p\left(\beta|\mathcal{X}\right) = \mathcal{GP}\left(\beta\middle|\vec{0},\mathbb{K}\left(\mathcal{F}(\mathcal{X}),\mathcal{F}(\mathcal{X})\right)+\xi^2\mathbb{I}\right) \tag{3}$$

, where Λ is a function returning variances of $\mathrm{U}_{37}^{\mathrm{K}'}$ observations in the form of a diagonal matrix at the query SSTs given their regression vector, $\mathcal{F}$ is a feature function, and $\mathbb{K}$ is the GP kernel function defined as follows: for a kernel function k,

$$\mathbb{K}(\mathcal{X}^{(1)},\mathcal{X}^{(2)}) \triangleq \left[k\left(\mathcal{F}(\mathcal{X}_m^{(1)}),\mathcal{F}(\mathcal{X}_n^{(2)})\right)\right]_{m,n} \tag{4}$$

, where $\mathrm{A} \triangleq \mathrm{B}$ means that A is defined by B.

Because feature functions may not be injective, the term $\xi^2\mathbb{I}$ is inevitable to guarantee (3) to be well-defined by making it nondegenerate.

Then, for a scalar regression b of y at x, the regression distribution $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ is derived as follows:

$$\begin{aligned} &p\left(y|x,\mathcal{X},\mathcal{Y}\right) \\ &= \int p\left(y|b,x\right)\int p\left(b|\beta,x,\mathcal{X}\right)p\left(\beta|\mathcal{X},\mathcal{Y}\right)d\beta db \end{aligned} \tag{5}$$

As consistent with (2), we have the likelihood of y given b and x as a Gaussian $p\left(y|b,x\right) = \mathcal{N}\left(y|b,\Lambda(x)\right)$.

Because the density of regression vector given SSTs is assumed to be a GP, we have the following extension:

$$\begin{aligned} &p\left(\beta,b|\mathcal{X},x\right) \\ &= \mathcal{GP}\left(\begin{pmatrix}\beta\\ b\end{pmatrix}\middle|0,\begin{pmatrix}\mathbb{K}_{11}+\xi^2\mathbb{I} & \mathbb{K}_{12}\\ \mathbb{K}_{21} & \mathbb{K}_{22}+\xi^2\end{pmatrix}\right) \end{aligned} \tag{6}$$

, $\mathbb{K}_{11} \triangleq \mathbb{K}\left(\mathcal{F}\left(\mathcal{X}\right),\mathcal{F}\left(\mathcal{X}\right)\right)$, $\mathbb{K}_{22} \triangleq \mathbb{K}\left(\mathcal{F}\left(x\right),\mathcal{F}\left(x\right)\right)$, $\mathbb{K}_{12} \triangleq \mathbb{K}\left(\mathcal{F}\left(\mathcal{X}\right),\mathcal{F}\left(x\right)\right)$ and $\mathbb{K}_{21} \triangleq \mathbb{K}\left(\mathcal{F}\left(x\right),\mathcal{F}\left(\mathcal{X}\right)\right)$ are the abbreviations. Thus, the conditional distribution of b given β, $\mathcal{X}$ and x can be computed explicitly as a Gaussian.

Finally, the posterior distribution of β given $\mathcal{X}$ and $\mathcal{Y}$ is derived from the likelihood (2) and prior (3), which is also a Gaussian.

I.e., three terms in (5) are all Gaussian, so $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ can be computed analytically as follows: note that $\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}$ is not a function of x or y, so only one matrix inversion is required to compute the following μ and ν.

$$\begin{aligned} &p\left(y|x,\mathcal{X},\mathcal{Y}\right) \\ &\quad= \mathcal{N}\left(y|\mu\left(x,\mathcal{X},\mathcal{Y}\right),\Lambda\left(x\right)+\nu\left(x,\mathcal{X},\mathcal{Y}\right)\right) \end{aligned} \tag{7}$$

$$\begin{aligned} \mu\left(x,\mathcal{X},\mathcal{Y}\right) &\triangleq \mathbb{K}_{21}\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}\mathcal{Y} \\ \nu\left(x,\mathcal{X},\mathcal{Y}\right) &\triangleq \mathbb{K}_{22}+\xi^2 \\ &\quad-\mathbb{K}_{21}\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}\mathbb{K}_{12} \end{aligned} \tag{8}$$

Heteroscedasticity comes from the choice of Λ: if it is defined to be a constant function, the model becomes homoscedastic. Some papers, such as [8], suggest modelling Λ as another GP regression on the logarithms of the residues. This approach assumes that such logarithms follow Gaussian distributions so symmetric, and always underestimates variances by the Jensens inequality applied to the concave log function: the average of logarithms is always smaller than or equal to the logarithm of average!

Here, we use a more intuitive and direct form of Λ inspired by Nadaraya-Watson kernel regression ([11], [12], [13] and [14]) as follows: let $\mu_n \triangleq \mu\left(x_n,\mathcal{X},\mathcal{Y}\right)$ and $\nu_n \triangleq \nu\left(x_n,\mathcal{X},\mathcal{Y}\right)$ as abbreviations.

$$\Lambda\left(x\right) \triangleq \frac{\sum_{n=1}^{\mathrm{N}}\left(\left(y_n-\mu_n\right)^2+\nu_n\right)\mathcal{K}_h\left(x-x_n\right)}{\sum_{n=1}^{\mathrm{N}}\mathcal{K}_h\left(x-x_n\right)} \tag{9}$$

, where $\mathcal{K}$ is a density kernel and h is a tuning parameter which is called the bandwidth. I.e., $\Lambda\left(x\right)$ is defined as a weighted average of squares of residues, where weights are determined by how much the query SST x is departed from the data. (9) is derived from the following expectation over $\beta|\mathcal{X},\mathcal{Y}$:

$$\mathbb{E}_{\beta|\mathcal{X},\mathcal{Y}}\left[\frac{\sum_{n=1}^{\mathrm{N}}\left(y_n-\beta_n\right)^2\mathcal{K}_h\left(x-x_n\right)}{\sum_{n=1}^{\mathrm{N}}\mathcal{K}_h\left(x-x_n\right)}\right] \tag{10}$$

The choice of $\mathcal{K}$ does not substantially affect to the regression model, but the model does substantially depend on the value of h. One suggestion is to adapt the K-nearest neighbor bandwidth ([15] and [16]).

A GP with arcsine kernel in (11) is interpretable as a one-layer neural network with infinite number of marginalized hidden nodes with a sigmoid function as the activation ([6]). Now we have a scalar η and a square matrix Σ as hyperparameters to be tuned. To avoid overfitting, it is common to assume in addition that Σ is a diagonal matrix. Tuning hyperparameters can be done by maximizing the marginal likelihood (ML) $p(\mathcal{Y}|\mathcal{X})$ from (2) and (3) or by the leave-one-out cross-validation (LOO-CV): more details can be found in [5].

$$k_{\mathrm{NN}}(x,\tilde{x}) \triangleq \eta^2 \sin^{-1}\left(\frac{2f^{\mathbb{T}}\Sigma\tilde{f}}{\sqrt{\left(1+2f^{\mathbb{T}}\Sigma f\right)\left(1+2\tilde{f}^{\mathbb{T}}\Sigma\tilde{f}\right)}}\right) \quad (11)$$

, where $f \triangleq \mathcal{F}(x)$ and $\tilde{f} \triangleq \mathcal{F}(\tilde{x})$.

In our model, outliers are represented as a hidden variable $\mathcal{H} = \{\mathrm{H_n}\}_{n=1}^{\mathrm{N}}$, where $\mathrm{H_n} = 0$ if y_n is an inlier and 1 an outlier. $\mathrm{H_n}$'s are assumed to be independent and each has the following prior distribution:

$$p(\mathrm{H_n}) = \mathrm{Bernoulli}(\mathrm{H_n}|q) \quad (12)$$

, where $q \in (0,1)$ is a hyperparameter not to be learned. A small q implies that most of the observations are believed not to be outliers. This reflects a point of view that outliers are in essence subjective.

Once $\mathcal{H}$ is given, we specified each $\mathrm{U}_{37}^{\mathrm{K}'}$ observation in $\mathcal{Y}$ as follows: let $\Lambda_n \triangleq \Lambda(x_n)$ as abbreviation.

$$p(y_n|x_n,\mathrm{H_n}=0) \triangleq \mathcal{N}(y_n|\mu_n,\nu_n+\Lambda_n) \quad (13)$$

$$p(y_n|x_n,\mathrm{H_n}=1) \triangleq \frac{1}{2}\mathcal{N}\left(y_n\middle|\mu_n+d\sqrt{\nu_n+\Lambda_n},\nu_n+\Lambda_n\right) + \frac{1}{2}\mathcal{N}\left(y_n\middle|\mu_n-d\sqrt{\nu_n+\Lambda_n},\nu_n+\Lambda_n\right) \quad (14)$$

, where $d > 0$ is a hyperparameter for how much outliers are deviated from the inlier distribution. Note that the above outlier classification is working because outputs are defined on the one-dimensional space in this problem.

By considering (12) as the prior and (13) and (14) as the likelihoods given H_n, respectively, we derive the posterior distribution of H_n given x_n and y_n by Bayes' rule:

$$p(\mathrm{H_n}|x_n,y_n) \propto p(\mathrm{H_n})\,p(y_n|x_n,\mathrm{H_n}) \quad (15)$$

Now we are prepared. Algorithm 1 summarizes the learning procedure of our heteroscedastic GP regression (HGPR) on $\mathrm{U}_{37}^{\mathrm{K}'}$ over SSTs.

Algorithm 1: HGPR on $\mathrm{U}_{37}^{\mathrm{K}'}$ over SSTs

```
initialize hyperparameters and H ≡ 0.
while convergence do
    tune kernel hyperparameters in (3) and (11) with X, Y|H = 0.
    compute μ and ν in (8) with X, Y|H = 0.
    choose the bandwidths in (9).
    update Λ in (9) with X, Y|H = 0.
    sample H|X, Y by (15).
end
return μ, ν and Λ.
```

One possible way of utilizing the obtained GP regression model in (7) is plugging it in the following Bayesian inversion:

$$p(\tilde{x}|\tilde{y}) \propto p(\tilde{y}|\tilde{x})\,p(\tilde{x}) \quad (16)$$

, where $\tilde{y}$ is the observed $\mathrm{U}_{37}^{\mathrm{K}'}$ data and $\tilde{x}$ is the associated SSTs to be inferred. The prior $p(\tilde{x})$ can be given as a distribution of SSTs given their geographical information, for instance.

In general, (16) does not have a closed form. Because SSTs are of one dimension, applying the Markov-chain Monte Carlo (MCMC) is enough to sample from the posterior $p(\tilde{x}|\tilde{y})$; if the event to infer is of high dimension, a variational inference to approximate the posterior with a known distribution must be considered.

III. DATA

We used the dataset same with [1], after discarding those in the locations that it excluded from their analysis. The rest of data with 1274 observations were used. We could discard more from them but did not do so for checking whether or not our model could classify apparent outliers desirably. Details about data are in [1].

For the density kernel and bandwidth in (9), we adapted a Gaussian kernel having an unbounded support and the K-nearest neighbor bandwidth where K is selected by the LOO-CV among 10 candidates, from 1% to 10% of the data per 10 iterations. Hyperparameters q in (12) and d in (13) and (14) were set to be 0.065 and 2.48, respectively: these values lead the marginal likelihood of y_n from (12), (13) and (14) to approximating the Student's t-distribution with 6 as the degree of freedom.

We also regularized the raw data $\mathcal{X}$ and $\mathcal{Y}$ by $x \leftarrow x/6.5656 - 3.0205$ and $y \leftarrow y/0.2104 - 3.3642$ so that they fit to the Student's t-distribution with 6 as the degree of freedom, and then took a feature function

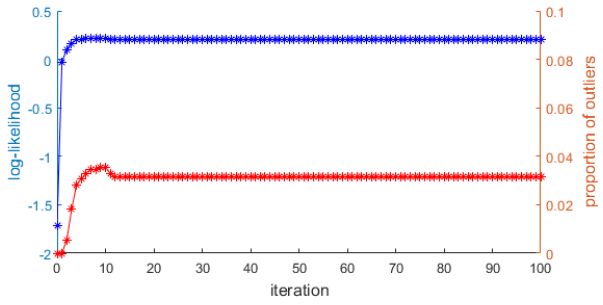

Fig. 1. Average log-likelihoods of inliers and proportions of outliers over iterations.

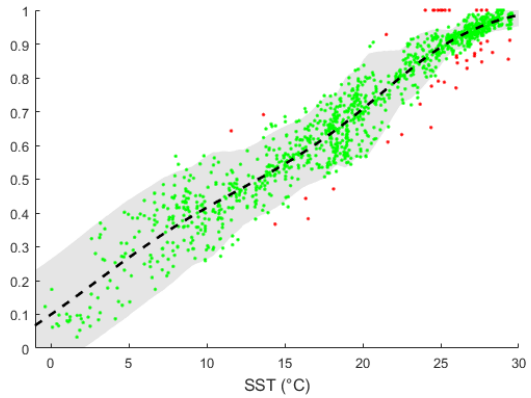

Fig. 2. The learned regression over SSTs. Green and red points are inliers and outliers, respectively. The dashed line shows the inferred means over SSTs and shaded region is the 95% confidence band.

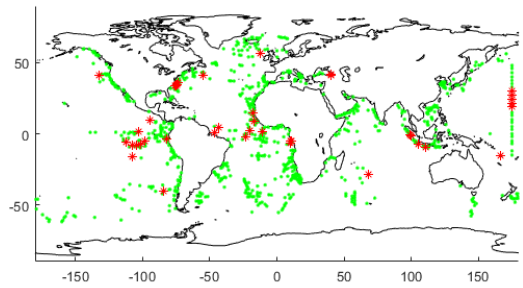

Fig. 3. The world map with data locations. Green and red points are inliers and outliers, respectively.

$\mathcal{F}(x) = (1, x)^{\mathbb{T}}$ on the regularized $\mathcal{X}$. To identify and regularize the model, we defined Σ in (11) to be a diagonal matrix, where the first entry is fixed to be 1.

IV. Results

After 100 iterations in about 7 minutes, we obtained the results shown in the eight figures, from 1 to 8. Finally selected K was 4%. Figure 1 shows the average log-likelihoods of inliers (blue) and proportions of inferred outliers (red). It shows convergence of the learning procedure after around the 15^{th} iteration. Only about 3% of the data were classifed as outliers. Learned kernel hyperparameters are the following:

$$\begin{aligned} \eta &= 2.1962 \\ \Sigma &= \begin{bmatrix} 1 & 0 \\ 0 & 0.4712^2 \end{bmatrix} \\ \xi &= 2.7253 \times 10^{-12} \end{aligned} \tag{17}$$

Figure 2 represents the inferred regression on SSTs with the classification of inliers (green) and outliers (red), and figure 3 those on the world map. It clearly shows the heteroscedasticity of observations and the associated model, as figure 4. Also, apparent outliers at 22-26°C of the upper boundary of the plot are classified as so. Classified outliers above 25°C below the model, however, could be from another cluster. This can be adjusted by tuning hyperparameters in section III. One advantage of GP regression is that the inferred mean at each input is expressed in a closed form and differentiable as much as the adopted kernel. Curvatures of the inferred means are shown in figure 5. [1] suggests that slopes of the means start to change at 23.4°C, which is roughly consistent with our inference, but some more changes are also captured in the inferred model.

To visually check how residuals are treated appropriately by our heteroscedastic model, figures 6 to 8 were also plotted. A normalized residual r_n at x_n is defined as follows:

$$r_n \triangleq \frac{y_n - \mu_n}{\sqrt{\nu_n + \Lambda_n}} \tag{18}$$

I.e., if means and variances of the GP regression model are properly inferred, normalized residuals at inliers must follow the standard normal distribution.

In figure 6, most of the normalized residuals seem to follow the standard normal distribution, and figure 7 supports that assertion as the Q-Q plot of inliers. In addition, figure 8 shows that normalized residuals are barely correlated with SSTs: the correlation between the pairs classified as inliers is 0.0019. These results strongly suggest that our GP regression model appropriately infer the heteroscedasticity of the model.

V. Conclusion

Our GP regression on $\mathrm{U}_{37}^{\mathrm{K}'}$ appropriately explains the heteroscedasticity of data and provides a probabilistic model with explicit distributions. It converges quickly,

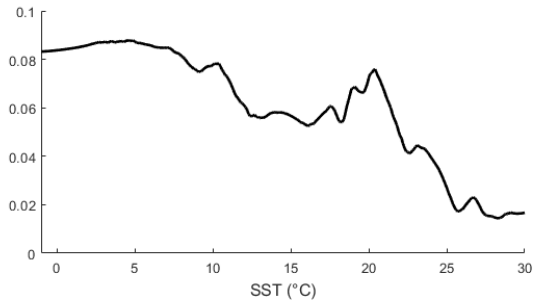

Fig. 4. Inferred standard deviations over SSTs.

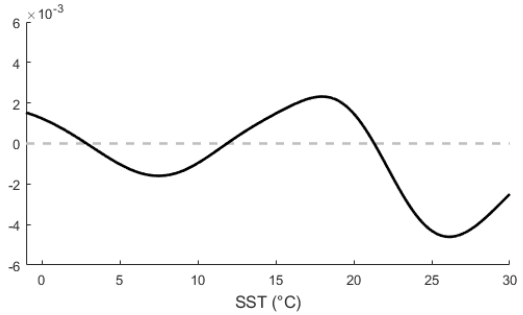

Fig. 5. Inferred curvatures over SSTs.

in about 7 minutes, even if we do not assume any informative prior structure (see $\vec{0}$ in (3)) on the regression. Thus, this work can be considered as a success of the nonparametric approach on the real data.

However, a GP has several disadvantages to overcome. Firstly, it requires at least one matrix inversion in constructing distributions, where its size is the same as the number of observations. There are only 1274 data, which is affordable in the currently available computing power, so it was not problematic in this case. We used two-dimensional features, which do not suffer from the curse of dimensionality.

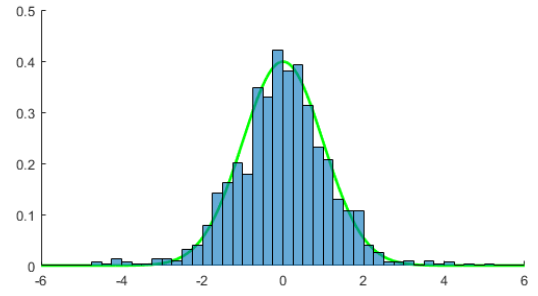

Fig. 6. The histogram of normalized residuals. The green graph is the standard normal distribution.

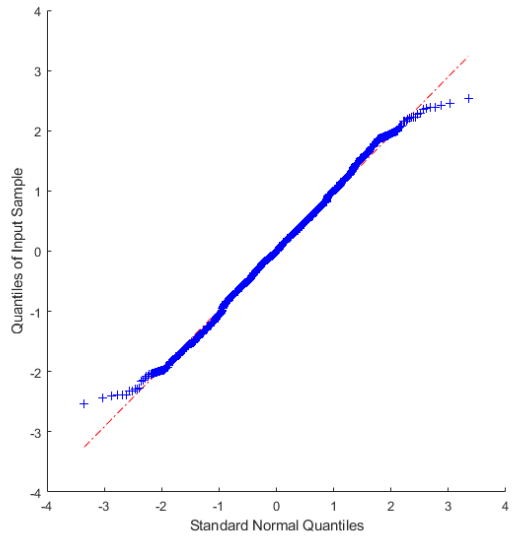

Fig. 7. The Q-Q plot of inliers.

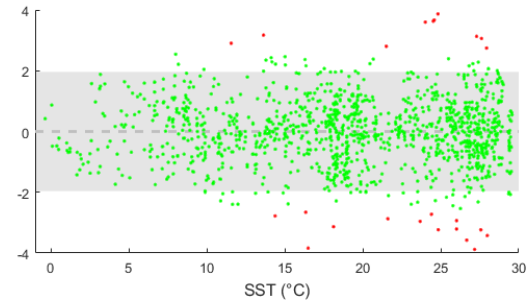

Fig. 8. A plot of normalized residuals over SSTs. Green and red points are inliers and outliers, respectively. The shaded region covers the 95% confidence intervals $[-1.96, 1.96]$.

Nonetheless, in this paper a model based on GPs shows its effectiveness on explaining $U_{37}^{K'}$ observations over SSTs, where relatively moderate amount of data are given and inputs are of low-dimensional. One more advantage of the nonparametric approaches is that it does not depend much on the specific characteristics of data: it is possible to adapt the same approach to any data to do regression.

Codes which run on MATLAB can be found in https://github.com/eilion/HGPR_SST_Proxy_Cal.

ACKNOWLEDGMENTS

This paper is based on the works supported by the Division of Applied Mathematics in Brown University, by the National Science Foundation (NSF) under a grant number OCE-1760838, and by the Kwanjeong Educational Foundation.

REFERENCES

[1] J. E. Tierney and M. P. Tingley, "Bayspline: A new calibration for the alkenone paleothermometer," *Paleoceanography and Paleoclimatology*, vol. 33, no. 3, pp. 281–301, 2018. 1, 3, 4

[2] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989. 1

[3] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991. 1

[4] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (USA), pp. 6232–6240, Curran Associates Inc., 2017. 1

[5] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. 1, 3

[6] C. K. I. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998. 1, 3

[7] Y. Cho and L. K. Saul, "Large-margin classification in infinite neural networks," *Neural Computation*, vol. 22, no. 10, pp. 2678–2697, 2010. 1

[8] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic gaussian process regression," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 393–400, 2007. 1, 2

[9] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 841–848, 2011. 1

[10] A. Shah, A. Wilson, and Z. Ghahramani, "Student-t Processes as Alternatives to Gaussian Processes," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, vol. 33, pp. 877–885, 2014. 2

[11] E. A. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 9, no. 1, pp. 141–142, 1964. 2

[12] G. S. Watson, "Smooth regression analysis," *Sankhya: The Indian Journal of Statistics, Series A*, vol. 26, pp. 359–372, 01 1964. 2

[13] H. J. Bierens, *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press, 1994. 2

[14] N. Langren and X. Warin, "Fast and stable multivariate kernel density estimation by fast sum updating," *Journal of Computational and Graphical Statistics*, vol. 28, no. 3, pp. 596–608, 2019. 2

[15] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1049–1051, 1965. 2

[16] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *Ann. Statist.*, vol. 20, no. 3, pp. 1236–1265, 1992. 2