Toward Statistical Emulators of Agricultural Computer Models

Jarad Niemi

Iowa State University

November 4, 2022

Funded, in part, by

- the Iowa State University Presidential Interdisciplinary Research Initiative on C-CHANGE: Science for a Changing Agriculture
- USDA NIFA: Consortium for Cultivating Human And Naturally reGenerative Enterprises (C-CHANGE Grass2Gas)
- Foundation for Food and Agriculture Research: Prairie Strips for Healthy Soils and Thriving Farms

Collaborators

Prairie STRIPS Collaborators: `http://prairiestrips.org/people`

Gaussian Process Emulators:




Agriculture

Corn Belt

https://www.rma.usda.gov/en/RMALocal/Iowa

Iowa Agricultural Production

https://www.iadg.com/iowa-advantages/target-industries/

Iowa is the largest producer of corn, pork and eggs in the United States and second in soybeans and red meat production.

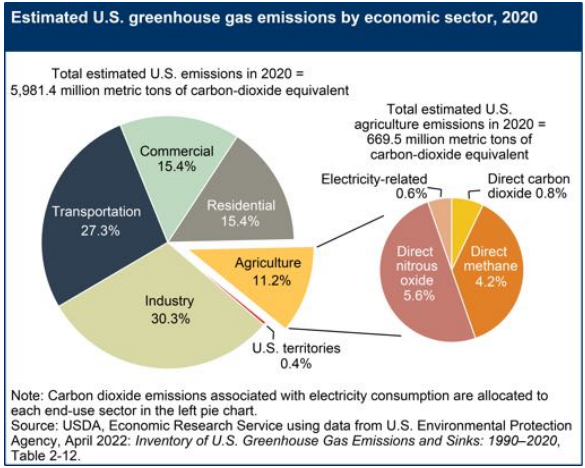




https://www.britannica.com/plant/corn-plant
https://www.nationalhogfarmer.com/marketing/total-pork-production-2014-down-slightly
https://www.medicalnewstoday.com/articles/283659
https://www.midwestfarmreport.com/2019/12/11/state-soybean-yield-contest-entries-announced/

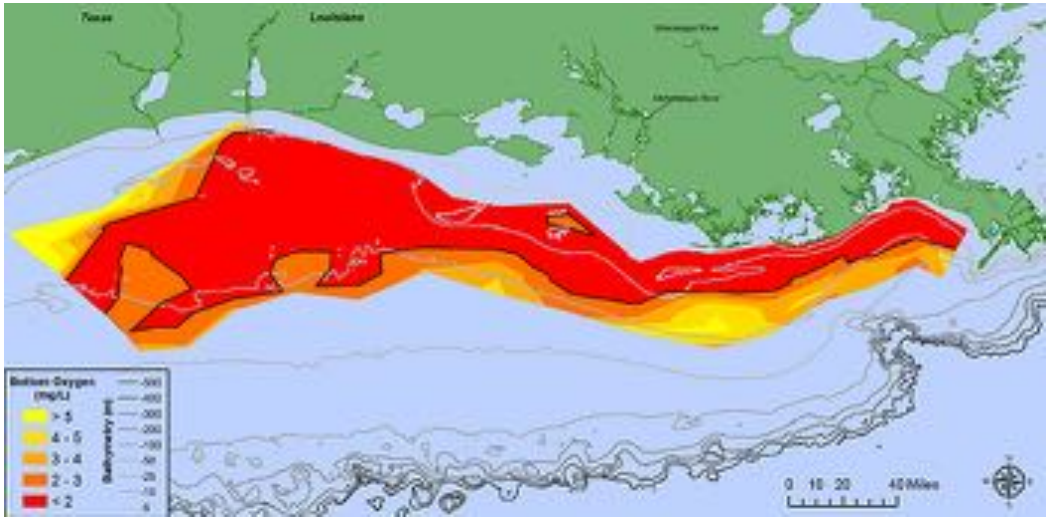
https://www.scientificamerican.com/article/meat-and-environment/

Agriculture Impact on Climate Change



https://www.ers.usda.gov/topics/natural-resources-environment/climate-change/

Gulf of Mexico Dead Zone



https://www.noaa.gov/media-release/gulf-of-mexico-dead-zone-is-largest-ever-measured

Soil loss

Iowa loses $1,000,000,000/year in soil

https://www.desmoinesregister.com/story/money/agriculture/2014/05/03/erosion-estimated-cost-iowa-billion-yield/8682651/

C-CHANGE

C-CHANGE: Science for a changing agriculture

http://agchange.org

Prairie STRIPS
C-CHANGE PIRI
C-CHANGE Grass2Gas
Climate Smart Ag

Prairie STRIPS in Conservation Reserve Program

Prairie strips improve biodiversity and the delivery of multiple ecosystem services from corn–soybean croplands

Lisa A. Schulte, Jarad Niemi, Matthew J. Helmers, +11, and Chris Witte Authors Info & Affiliations

USDA
United States Department of Agriculture

Farm Service Agency

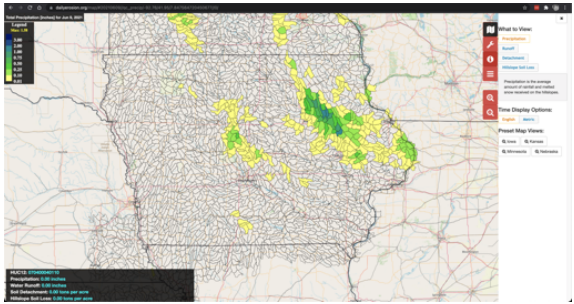
Conservation Reserve Program

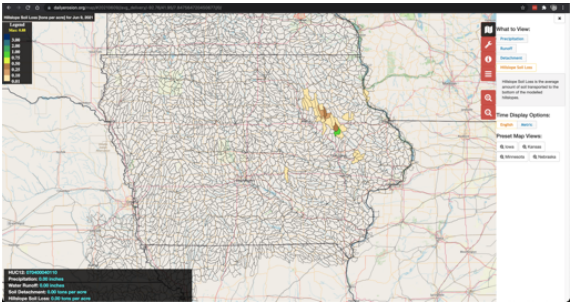
CLEAN LAKES, ESTUARIES AND RIVERS (CLEAR) INITIATIVE
Prairie Strip Practice (CP-43)

FACT SHEET
December 2019

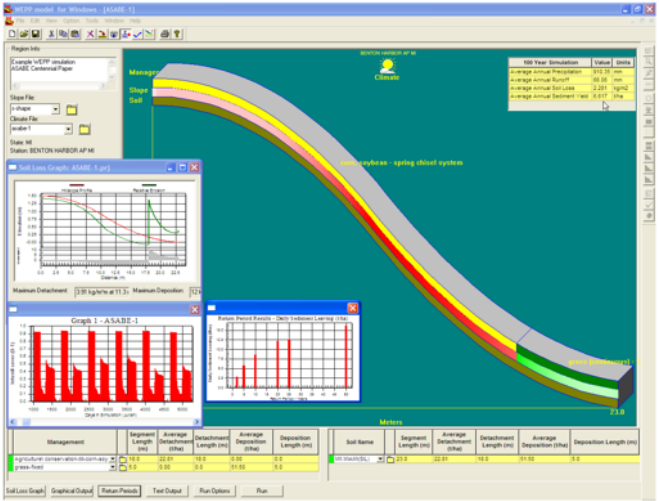
Daily Erosion Project (DEP)

using Water Erosion Prediction Project (WEPP)





Water Erosion Prediction Project (WEPP)



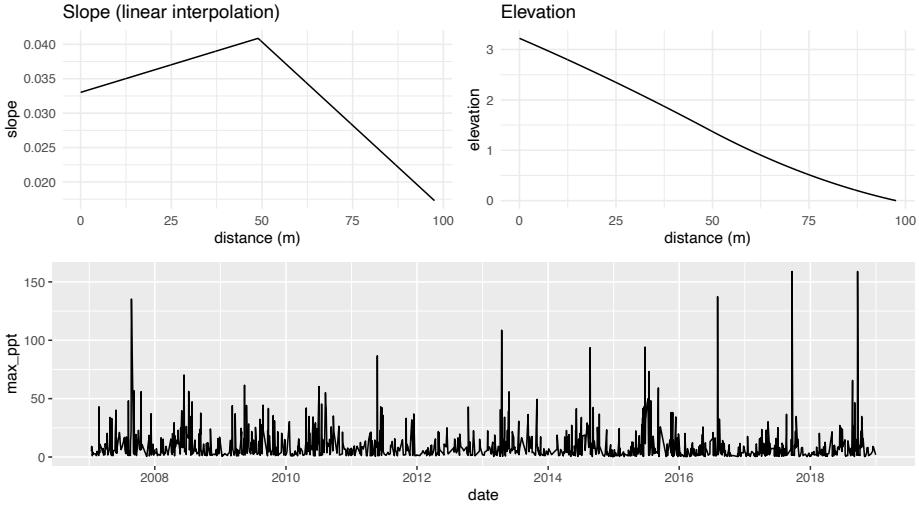
WEPPR

R packages:

- `WEPPR`: R interface for WEPP computer model
- `DEPR`: R interface for DEP/WEPP computer model
- `WEPPemulator`: utility features for emulator construction

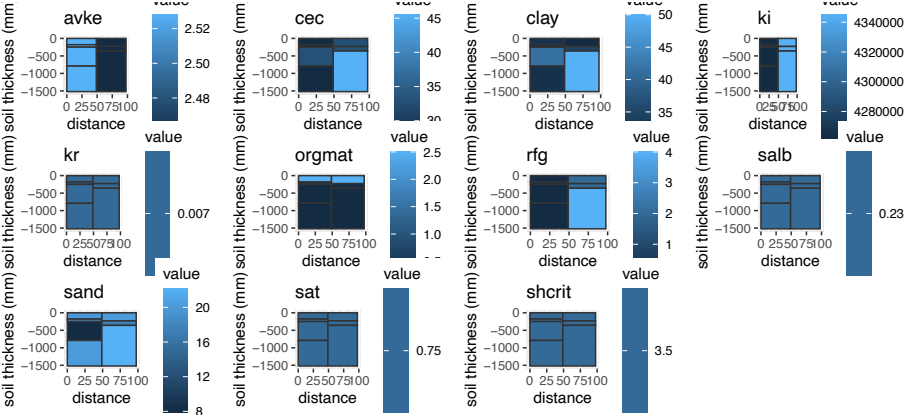
```
library("WEPPR")
library("tidyverse")
library("gridExtra")

slp <- read_slp(system.file("extdata", "071000090603_2.slp", package="WEPPR"))
cli <- read_cli(system.file("extdata", "092.63x040.90.cli", package="WEPPR"))

grid.arrange(plot(slp, plots="slope"),
             plot(slp, plots="elevation"),
             plot(cli),
             layout_matrix = rbind(1:2,3))
```

WEPPR Plots


Slope (linear interpolation)
slope
0.040
0.035
0.030
0.025
0.020
0
25
50
75
100
distance (m)
Elevation
elevation
3
2
1
0
0
25
50
75
100
distance (m)
max_ppt
150
100
50
0
2008
2010
2012
2014
2016
2018
date

WEPPR Plots

```
sol <- read_sol(system.file("extdata", "071000090603_2.sol", package="WEPPR"))
plot(merge_slp_sol(slp,sol))
```



GP Emulators

Gaussian Process (GPs)

Consider a deterministic computer model $f(\cdot)$ with

$$Y_i = f(X_i), \quad i = 1, \ldots, N$$

where

- X_i are inputs and
- Y_i are outputs with $Y = (Y_1, \ldots, Y_N)$.

We assume a Gaussian Process prior

$$f \sim \mathcal{GP}(m, k) \implies Y \sim N(m_x, \Sigma)$$

where (often) $m_x = 0$ and, for scalar inputs, $\Sigma_{ij} = k(x_i, x_j) = \sigma^2 e^{-(x_i - x_j)^2/\phi}$.

Two research directions:

- Computational tractability for large N
- Dealing with functional inputs

Training a GP

Find the maximum likelihood estimator (MLE) for $\theta = (\sigma^2, \phi)$,

$$\hat{\theta} = \text{argmax}_\theta \, p(y|\theta) = \text{argmax}_\theta \, N\left(y; 0, \Sigma(\theta)\right)$$

where $y = (y_1, \ldots, y_N)$.

The log-likelihood is

$$\log \mathcal{N}(y; 0, \Sigma(\theta)) \propto C - \log |\Sigma(\theta)| - y^\top \Sigma(\theta)^{-1} y$$

If there are N observations, $\Sigma(\theta)$ is an $N \times N$ covariance matrix and thus the computational time scales as $\mathcal{O}(N^3)$.

This is doable if $N \approx 1,000$ but not when you start getting larger and larger data sets.

Fully Independent Conditional (FIC) Approximation

Introduce a set of knots $x^\dagger = \left\{x_1^\dagger, \ldots, x_K^\dagger\right\}$ with $K << N$, such that

$$p(f_x, f_{x^\dagger}|\theta) = p(f_x|f_{x^\dagger}, \theta)p(f_{x^\dagger}|\theta)$$

where

$$\begin{aligned} f_{x^\dagger}|\theta &\sim \mathcal{N}(0, \Sigma_{x^\dagger x^\dagger}) \\ f_x|f_{x^\dagger}, \theta &\sim \mathcal{N}\left(\Sigma_{xx^\dagger}\Sigma_{x^\dagger x^\dagger}^{-1}(f_{x^\dagger}), \Lambda\right) \end{aligned}$$

with $\Lambda = \text{diag}\left(\Sigma_{xx} - \Sigma_{xx^\dagger}\Sigma_{x^\dagger x^\dagger}^{-1}\Sigma_{x^\dagger x}\right)$.

This joint implies the following marginal distribution for f_x:

$$f_x|\theta \sim \mathcal{N}(0, \Lambda + \Sigma_{xx^\dagger}\Sigma_{x^\dagger x^\dagger}^{-1}\Sigma_{x^\dagger x})$$

which has the correct marginal means and variances, but the covariances are controlled by the knots.

[Seeger et al., 2003, Quiñonero-Candela and Rasmussen, 2005, Snelson and Ghahramani, 2006, Banerjee et al., 2008, Finley et al., 2009, Titsias, 2009, Cao et al., 2013]

Train FIC Model

Let $\Psi(x^\dagger, \theta) \equiv \Lambda(\theta) + \Sigma_{xx^\dagger}(\theta)\Sigma_{x^\dagger x^\dagger}(\theta)^{-1}\Sigma_{x^\dagger x}(\theta)$, then

$$Y|x^\dagger, \theta \sim \mathcal{N}(0, \Psi(x^\dagger, \theta)).$$

Train the model by finding

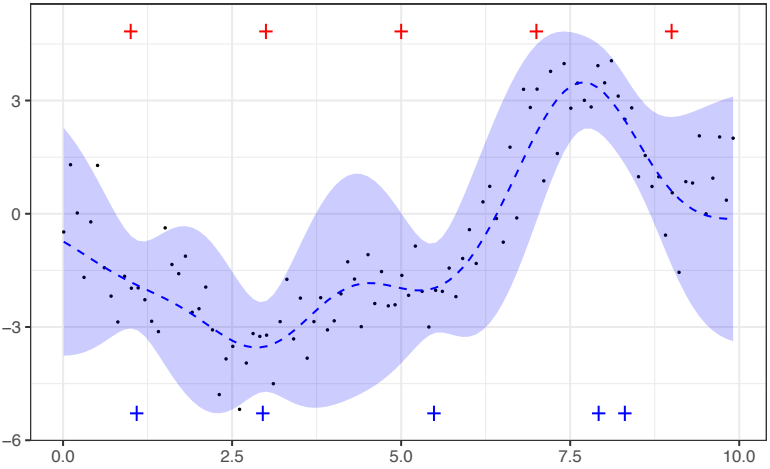
$$\hat{x}^\dagger, \hat{\theta} = \text{argmax}_{x^\dagger, \theta}\mathcal{N}(y; 0, \Psi(x^\dagger, \theta)).$$

which has computational complexity of $\mathcal{O}(NK^2)$.

There are a number of questions:

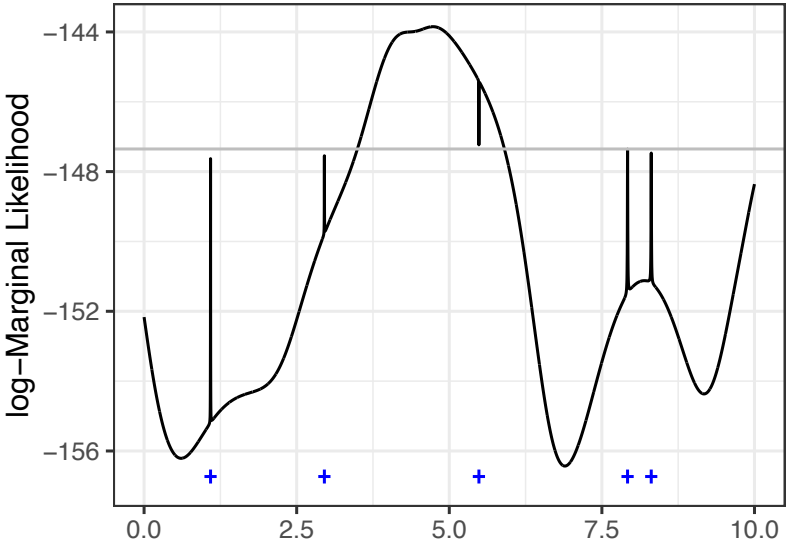
- how many knots are needed?
- where should the knots be?

Simultaneous knot optimization



(this has a nugget)

Adding another knot



OAT Algorithm

Knot selection algorithm

Algorithm 1. OAT knot selection algorithm. Convergence in the repeat loop is declared when the change in the objective function, the log-marginal likelihood, falls below a threshold. Set initial number of knots (K_I).

Initialize: $x^\dagger = \{x_i^\dagger\}_{i=1}^{K_I}$;
$\hat{\theta} = \text{argmax}_\theta p(y|x, x^\dagger, \theta)$;
repeat
  propose new knot $x^{\dagger^*} \leftarrow J(y, x, x^\dagger, \hat{\theta})$;
  $(\hat{x}^{\dagger^*}, \hat{\theta}) = \text{argmax}_{(x^{\dagger^*}, \theta)} p(y|x, \{x^\dagger, x^{\dagger^*}\}, \theta)$;
  $x^\dagger = \{x^\dagger, \hat{x}^{\dagger^*}\}$;
until $|x^\dagger| = K_{max}$ *or convergence*;

Bayesian optimization

Let

- $w_{1:t-1}$ be the vector of log-marginal likelihood values at the candidates for the knot proposal which have thus far been explored at time t and
- $w^+ = \max(w_{1:t-1})$.

Let $W(z)$ be the unknown marginal likelihood at input location z and model it using a GP, for clarity call this the meta GP.

Then expected improvement is

$$\alpha\left(z; x^{\dagger}_{1:t-1}, w_{1:t-1}\right) = E[\max(W(z) - w^+, 0)] = \begin{aligned}&(E\left[W(z)|w_{1:t-1}\right] - w^+)\Phi\left(\frac{E\left[W(z)|w_{1:t-1}\right]-w^+}{\sqrt{V\left[W(z)|w_{1:t-1}\right]}}\right)\\ &+\sqrt{V\left[W(z)|w_{1:t-1}\right]}\phi\left(\frac{E\left[W(z)|w_{1:t-1}\right]-w^+}{\sqrt{V\left[W(z)|w_{1:t-1}\right]}}\right).\end{aligned}$$

where ϕ and Φ are the pdf and cdf of a standard normal, respectively.

[Jones, 2001, Shahriari et al., 2016]

Knot proposal algorithm

Algorithm 2. Knot proposal algorithm. Set the minimum (T_{min}) and maximum (T_{max}) number of marginal likelihood evaluations.

---

set the mean of the meta GP equal to $\log p\left(y \middle| x, \{x^\dagger, \cdot\}, \hat{\theta}\right)$;

sample $x_1^\dagger, ..., x_{T_{min}}^\dagger$ without replacement from x ;

augment known marginal likelihood values $w_j = \log p\left(y \middle| x, \left\{x^\dagger, x_j^\dagger\right\}, \hat{\theta}\right)$ for $j = 1, \ldots, k$ with evaluations of the marginal likelihood at the new knots, that is $w_{k+j} = \log p\left(y \middle| x, \left\{x^\dagger, x_j^\dagger\right\}, \hat{\theta}\right)$ for $j = 1, ..., T_{min}$;

for $t = T_{min} + 1, ..., T_{max}$ **do**

  update covariance parameters in meta GP ;

  $x_t^* = \text{argmax}_{z \in x \setminus x_{1:t-1}^\dagger} \alpha\left(z; x_{1:t-1}^\dagger, w_{1:t-1}\right)$;

  $w_t = \log p\left(y \middle| x, \{x^\dagger, x_t^*\}, \hat{\theta}\right)$;

end

return x_j^* such that $j = \text{argmax}_t\, w_t$

Knot selection algorithm

Algorithm 1. OAT knot selection algorithm. Convergence in the repeat loop is declared when the change in the objective function, the log-marginal likelihood, falls below a threshold. Set initial number of knots (K_I).

Initialize: $x^\dagger = \{x_i^\dagger\}_{i=1}^{K_I}$;
$\hat{\theta} = \text{argmax}_\theta p(y|x, x^\dagger, \theta)$;
repeat
  propose new knot $x^{\dagger^*} \leftarrow J(y, x, x^\dagger, \hat{\theta})$;
  $(\hat{x}^{\dagger^*}, \hat{\theta}) = \text{argmax}_{(x^{\dagger^*}, \theta)} p(y|x, \{x^\dagger, x^{\dagger^*}\}, \theta)$;
  $x^\dagger = \{x^\dagger, \hat{x}^{\dagger^*}\}$;
until $|x^\dagger| = K_{max}$ *or convergence*;

Performance metrics

All data models:

$$MNLP = \text{median}_{i \in 1,\ldots,N_{test}}\{-\log p(\tilde{y}_i|x^\dagger, \hat{\theta}, y)\}.$$

$$AUKL = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \int p_{full}(f(\tilde{x}_i)|\hat{\theta}, y) \log \frac{p_{full}(f(\tilde{x}_i)|\hat{\theta}, y)}{p_{sparse}(f(\tilde{x}_i)|x^\dagger, \hat{\theta}, y)} df(\tilde{x}_i).$$

Gaussian:

$$SRMSE = \sigma_{\tilde{y}}^{-1} \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (E\left[f(\tilde{x}_i)|Y\right] - \tilde{y}_i)^2},$$

where $\sigma_{\tilde{y}}^2 = \frac{1}{N_{test}-1} \sum_{i=1}^{N_{test}} (\tilde{y}_i - \bar{\tilde{y}})^2$, $\bar{\tilde{y}} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \tilde{y}_i$, and $\tilde{y}$ is the vector of test set target values.

Boston Housing

490 observations (random 80% for training) with $d = 3$

Method	Runtime	K	Tmax	SRMSE	MNLP	AUKL
Full	394	–	–	0.359	2.500	0.000
OAT-BO	545	13	25	0.366	2.466	0.045
OAT-RS	356	12	25	0.366	2.464	0.039
OAT-RS	339	15	50	0.364	2.469	0.047
Simult.	25831	50	–	0.378	2.291	0.356
Simult.	3945	13	–	0.356	2.313	0.242

Banana Data

5,300 binary observations (10% training) with $d = 2$

Method	Runtime	Tmax	K	MNLP	AUKL
Full	26795	–	–	0.038	0.000
OAT-BO	3150	25	50	0.037	0.061
OAT-RS	2954	25	50	0.038	0.051
OAT-RS	3471	50	50	0.038	0.039
Simult.	6219	–	50	0.069	3.265

Lansing Woods Hickory Data

2,251 Poisson process observations with $d = 2$

Type	Runtime	K	Tmax	MNLP	AUKL
Full	5892	–	–	1.040	0.000
OAT-BO	846	28	25	1.058	0.321
OAT-RS	1083	33	25	1.066	0.276
OAT-RS	989	30	50	1.055	0.279
Simult.	18456	50	–	1.068	0.597
Simult.	15046	28	–	1.060	0.426

One-at-a-time (OAT) selection

We developed a one-at-a-time (OAT) knot selection that

- Begins with a small number of knots
- Optimizes the knot locations according to the marginal likelihood or variational objective function
- Iteratively adds knots until no improvement is seen in the objective function

Summary of results:

- Prediction is equivalent to Full GP and simultaneous knot optimization
- Formally, OAT has computational complexity of $\mathcal{O}(NK^2)$
- Practically, OAT is much faster than Full GP (on large data sets) and simultaneous knot optimization

Manuscripts:

- Nate Garton, Jarad Niemi, and Alicia Carriquiry. (2020) "Knot Selection in Sparse Gaussian Processes with a Variational Objective." Statistical Analysis and Data Mining. 13(4): 324-336.
- Nate Garton, Jarad Niemi, and Alicia Carriquiry. "Knot Selection in Sparse Gaussian processes." arXiv:2002.09538

Functional inputs

Vector-input Gaussian Process (viGP)

For observation i, we have response $Y_i \in \mathbb{R}$ and input $X_i = (X_{i,1}, \ldots, X_{i,D})$. Our deterministic computer model is $f()$ with $Y_i = f(X_i)$.

Assume f is a zero-mean Gaussian process such that

$$Cor(Y_i, Y_j) = e^{-\frac{1}{2}D(X_i, X_j, \omega)}$$

and

$$D(X_i, X_j, \omega) = \sum_{d=1}^{D} \omega_d (x_{i,d} - x_{j,d})^2.$$

Some refer to this as automatic relevance determination.

Functional-input Gaussian Process

For observation i, we have $Y_i \in \mathbb{R}$ and $X_i(t)$ for $t \in [0, T]$. Our computer model is $f : \mathcal{X} \to \mathbb{R}$ with $Y_i = f(X_i(t))$.

Following Morris [2012, 2018], the functional-input Gaussian Process has

$$Cor(Y_i, Y_j) = e^{-\frac{1}{2} D(X_i(t), X_j(t), \omega)}$$

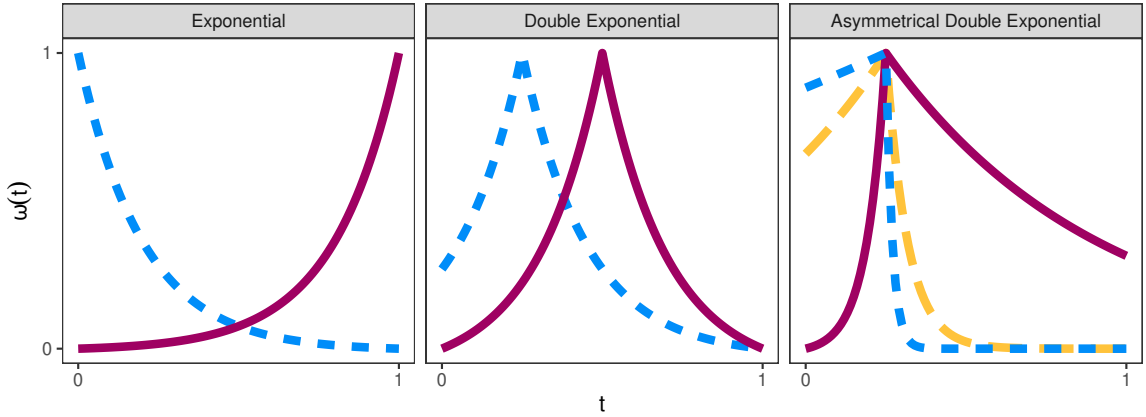
$$D(X_i, X_j, \omega) = \int_0^T \omega(t)(X_i(t) - X_j(t))^2 dt \approx \sum_{d=1}^{D} \omega(t_d)(X_i(t_d) - X_j(t_d))^2.$$

We'll refer to this as fiGP with automatic dynamic relevance determination.

For the functional length-scale, one option is the asymmetric double exponential function

$$\omega(t) = \begin{cases} \exp\left(-\lambda_1 \ |t - \tau|\right) & \text{for } t \leq \tau \\ \exp\left(-\lambda_2 \ |t - \tau|\right) & \text{for } t > \tau \end{cases}$$

Theoretical fiGP functional length-scale



Case study

Implementation

- 8 training and 8 test complementary sets JPL MLS data with 1,000 soundings each
- 7 plausible models
  - viGP SE, ARD, FPCA, FFPCA
  - fiGP Edn, SDE, ADE
  - One model fit separately per input-output pair
- Fully Bayesian inference
  - Hamiltonian Monte Carlo [Brooks, 2011, ch. 5]
  - NUTS algorithm [Hoffman and Gelman, 2014] via Stan [Team, 2021]
  - 1 long chain [Raftery and Lewis, 1992]
  - Extensive search for an initial value
  - 500 post-warmup iterations
  - 1,500 posterior samples
- Several out-of-sample validation statistics

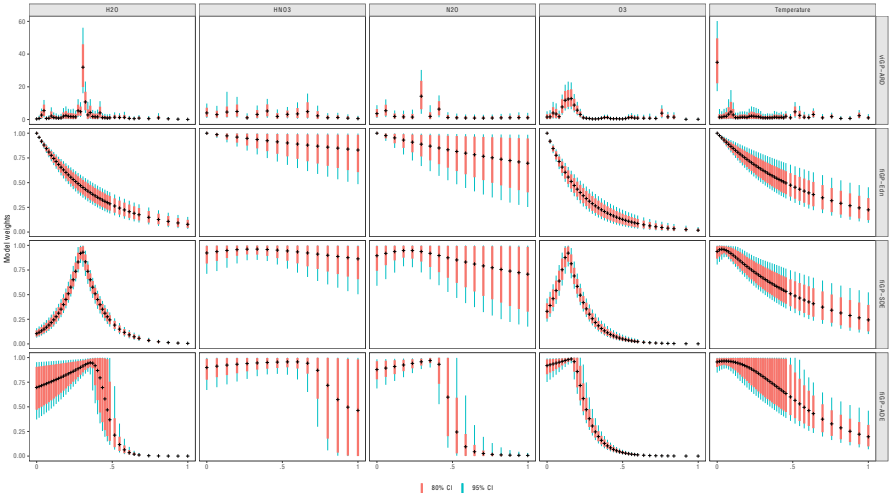
Validation statistics

	H2O	HNO3	N2O	O3	Temp	Mean
SE	.34	**.48**	**.44**	.32	**.25**	.37
ARD	**.31**	**.47**	**.43**	**.30**	**.25**	.35
FPCA	.67	.91	.99	.46	.54	.71
FFPCA	.46	.54	**.46**	.38	.33	.44
EDN	.33	**.47**	**.44**	**.29**	**.25**	.36
SDE	**.31**	**.47**	**.44**	**.29**	**.25**	.35
ADE	**.31**	**.47**	**.43**	**.29**	**.25**	.35
Mean	.39	.55	.52	.33	.31	.42

	H2O	HNO3	N2O	O3	Temp	Mean
SE	273	**614**	**585**	138	**-7**	323
ARD	**196**	**619**	**581**	**92**	**-13**	295
FPCA	1024	1320	1406	637	802	1038
FFPCA	535	**646**	**630**	295	268	475
EDN	261	**623**	**585**	**90**	**4**	312
SDE	**202**	**623**	**585**	**85**	**4**	300
ADE	**202**	**610**	**581**	**87**	**2**	297
Mean	385	722	708	204	152	434

Table: Mean validation statistics $\bar{v}^{(p,q)}$: RMSE (left) and negPPLD (right). Smaller values are better. Bold is best in class.

Estimated fiGP functional length-scale



Summary

Research/policy advances

- Prairie strips impact on agroecosystem services and inclusion in CRP as CP-43
- OAT algorithm for knot selection to deal with computational intractability due to large n
- fiGP/ADRD model for functional inputs

These slides are available at

- `https://github.com/jarad/VirginiaTech2022`
- `http://www.jarad.me/research/presentations.html`

Thank you!

Other links:

- `http://www.jarad.me/`
- `http://www.youtube.com/jaradniemi`

References

Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

Steve Brooks. *Handbook of Markov Chain Monte Carlo*. 2011. ISBN 9781420079425 9780429138508 9781283257282 9786613257284. URL https://doi.org/10.1201/b10905.

Yanshuai Cao, Marcus A Brubaker, David J Fleet, and Aaron Hertzmann. Efficient optimization for sparse gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2013.

Andrew O. Finley, Huiyan Sang, Sudipto Banerjee, and Alan E. Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884, June 2009. doi: https://doi.org/10.1016/j.csda.2008.09.008.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, 2014.

Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

Max D. Morris. Gaussian surrogates for computer models with time-varying inputs and outputs. *Technometrics*, 54(1):42–50, February 2012. ISSN 0040-1706. doi: 10/ghbnds. URL https://doi.org/10.1080/00401706.2012.648870.

Max D. Morris. Decomposing functional model inputs for variance-based sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1584–1599, January 2018. doi: 10/ghbnd7. URL https://epubs-siam-org.eu1.proxy.openathens.net/doi/abs/10.1137/18M1173058.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 2005.

Adrian E. Raftery and Steven M. Lewis. [Practical Markov chain Monte Carlo]: Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7(4), November 1992. ISSN 0883-4237. doi: 10.1214/ss/1177011143. URL https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Practical-Markov-Chain-Monte-Carlo--Comment--One-Long/10.1214/ss/1177011143.full.

Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *Artificial Intelligence and Statistics*, 2003.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.