Computer Model Emulation and Uncertainty Quantification Using a Deep Gaussian Process

Derek Bingham
Department of Statistics and Actuarial Science
Simon Fraser University

Faezeh Yazdi
Department of Statistics and Actuarial Science
Simon Fraser University

Danny Williamson
Department of Mathematics
University of Exeter

Ilya Mandel
School of Physics and Astronomy
Monash University

SFU
Department of Statistics and Actuarial Science
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

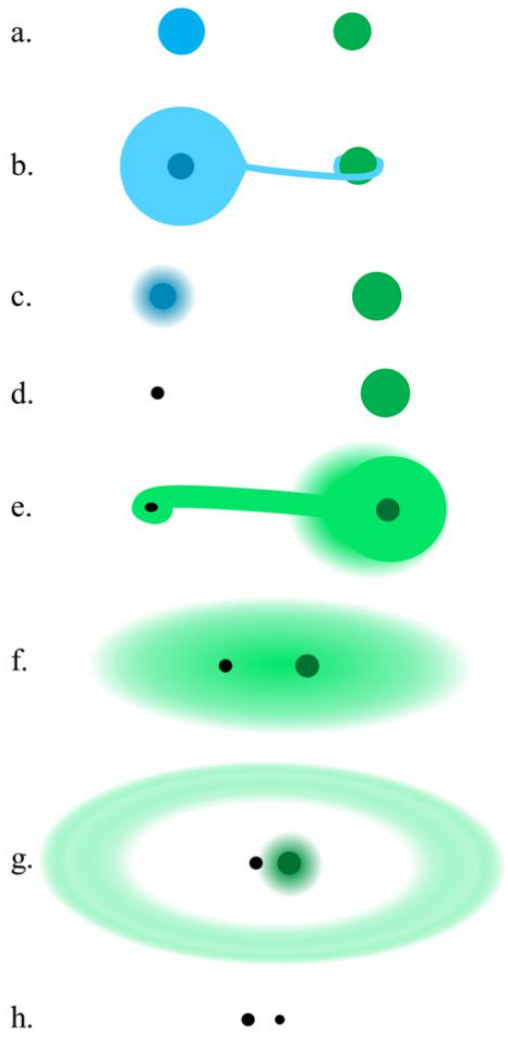
Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Application

- Interested in emulating the behavior of merging binary black holes
- Binary black holes are systems with two black holes orbiting around one another
- When binary black holes (BBH) merge, energy is released in the form of gravitational waves (predicted as the result of general relativity)
- As the BBHs spin closer together, rotational frequency increases as the objects merge, the gravitational waves can be observed/heard in the form of a *chirp*
- Laser Interferometer Gravitational-Wave Observatory (LIGO) detected this

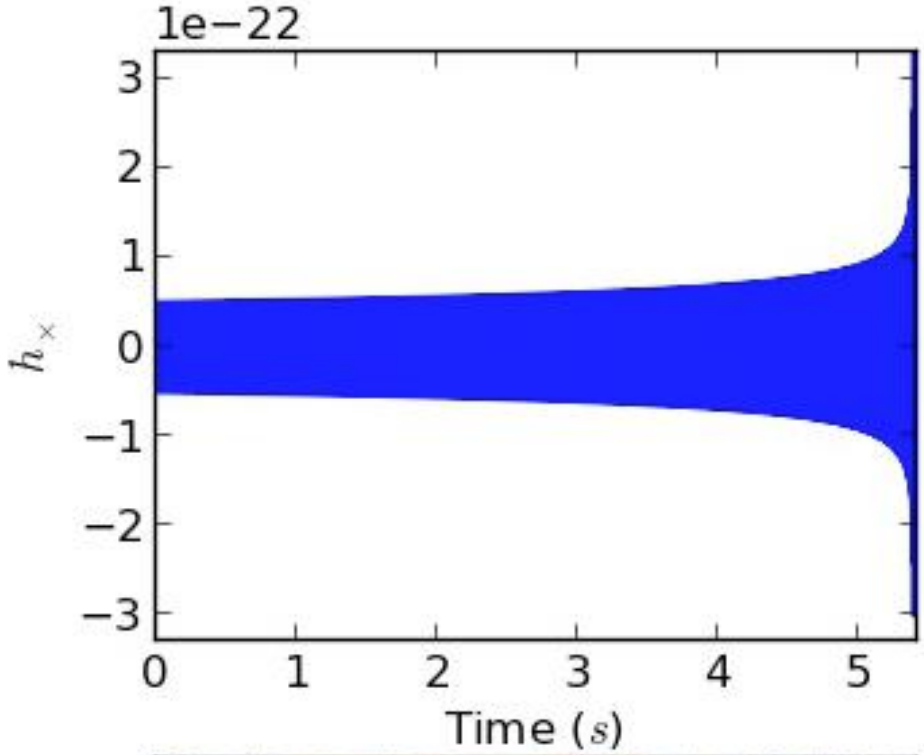
Application

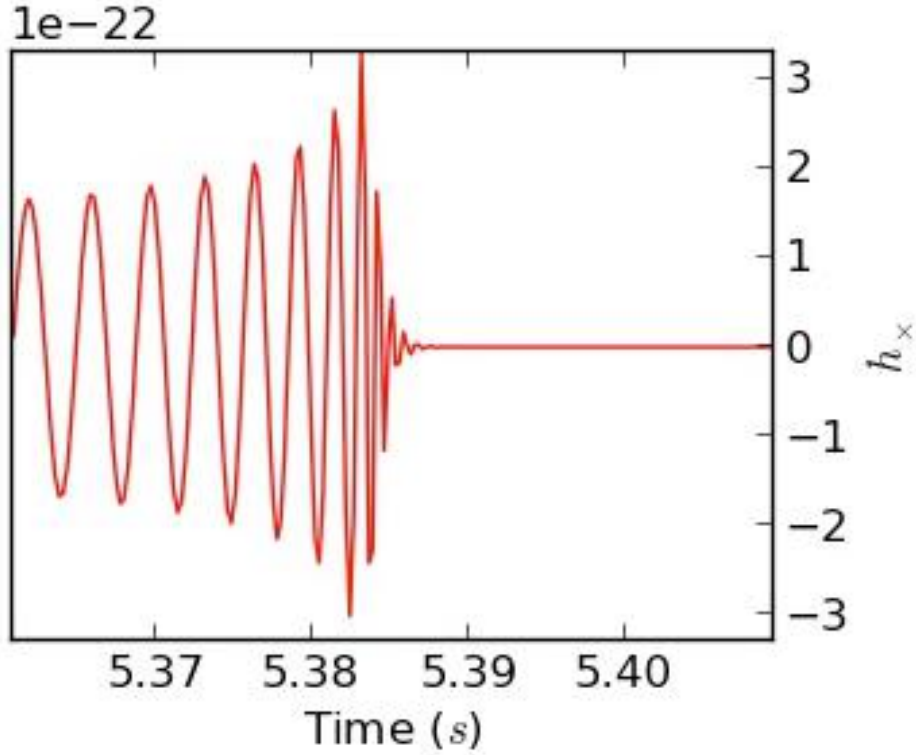


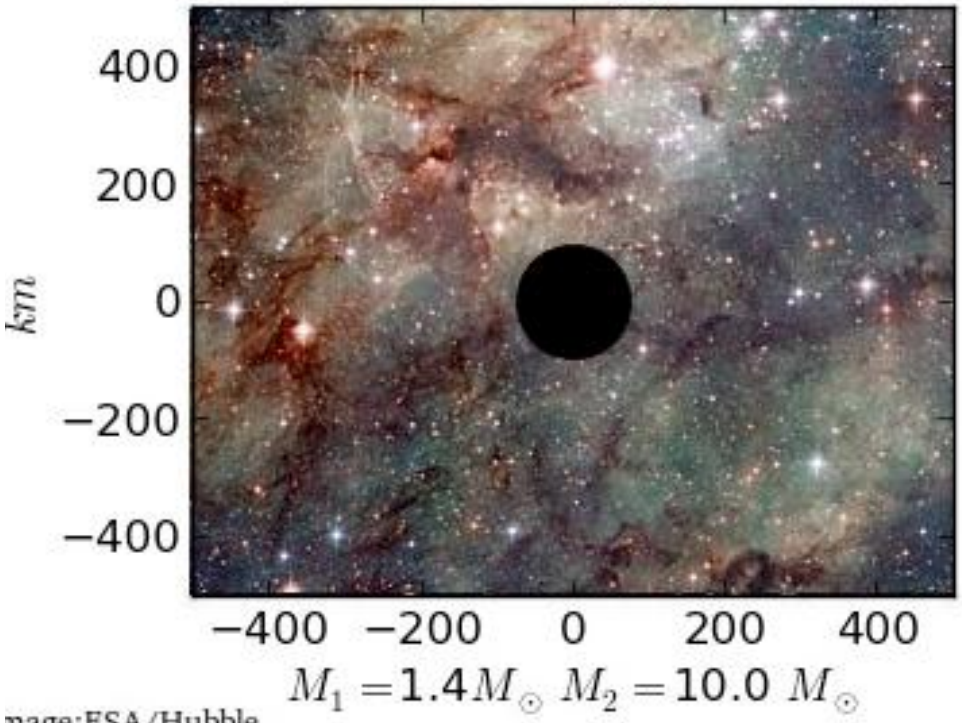
Mandel and Farmer, 2018

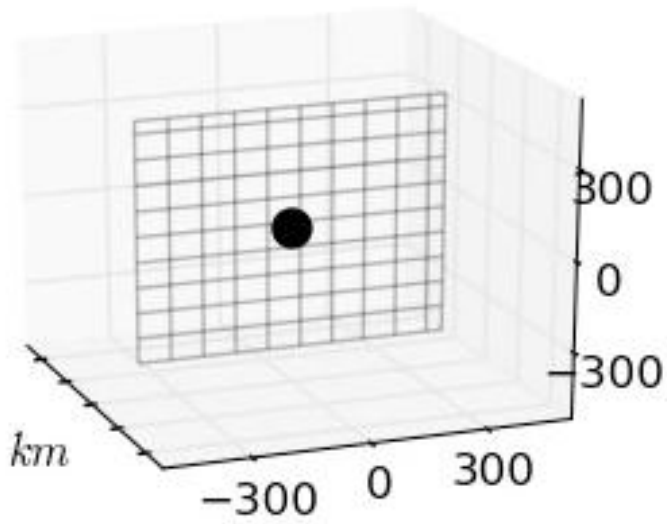
Goal: Want to construct an emulator of binary population synthesis simulation codes … interest lies in characteristics of binary black hole (BBH) mergers

a. Initial stellar binary
b. Mass transfer
c. Primary loses its entire envelope
d. Primary collapses into a black hole
e. Mass transfer from the initially less massive star onto the black hole which leads to …
f. The formation of a common envelope
g. Ejection of the common envelope
h. Collapse of the companion into a black hole
i. Merger









nage:ESA/Hubble

Jason Tye, University of Birmingham

Have a computational model

- **COMPAS** (Compact Object Mergers: Population Astrophysics and Statistics)

- **Inputs:** initial conditions of the binary system at birth (e.g., mass of the primary binary, initial orbital separation) and population parameters (e.g., mass loss rate during luminous blue variable phase)

- **Output:** chirp mass of the formed binary

- **Short-term goal:** emulate the chirp mass for BBH

- **Long-term goal:** using the observed distribution of chirp masses to constrain (i.e., *calibrate* in the sense of Kennedy and O'Hagan, 2001) population parameters that govern BBH formation

There are some challenges

- COMPAS is fast, but not readily available, nor fast enough
- Would like to exercise the code, potentially, billions of times
- We have a lot of code evaluations (about $m = 2$ x 10^6)
- While COMPAS is deterministic, the set of inputs does not always result in BBH formation... have regions of activity, otherwise, no chirp mass

Challenges

- More simply,

  1. Emulator has to be fast

  2. Have many deterministic simulations
     - Previous work on emulators with large simulation suites: e.g., Kaufman et al. (2011), Gramacy and Apley (2015) ...

  3. For many of the simulations runs, the output is unobserved (we replace with zero)

Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Often Gaussian processes are used to emulate response surface with estimates of uncertainty

Notation:

- Have m evaluations of the computer model with d-dimensional inputs

- Design matrix:

$$\boldsymbol{X} = (x_1, x_2, \ldots, x_m)'$$

- Outputs:

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_m)'$$

Often Gaussian processes are used to emulate response surface with estimates of uncertainty

- View the computer code as a single realization of a Gaussian process (GP):

$$y(\mathbf{x}) = \mu + z(\mathbf{x})$$

where,

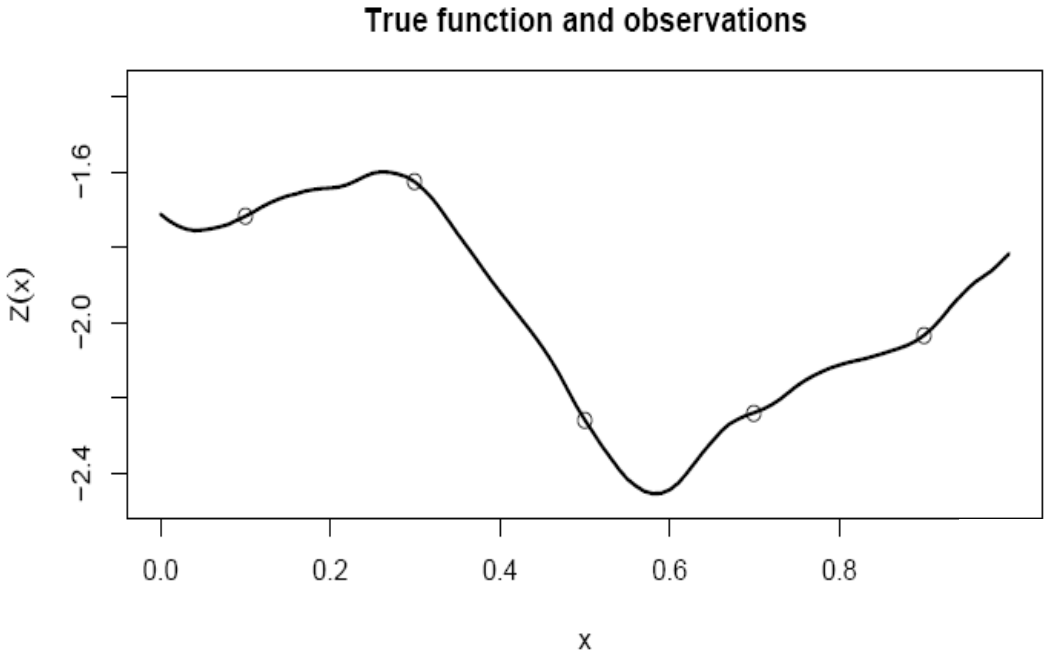
$$E\left(z(\mathbf{x})\right) = 0$$

$$Var\left(z(\mathbf{x})\right) = \sigma^2$$

$$Corr\left(z(\mathbf{x}), z(\mathbf{x}')\right) = \prod_{i=1}^{d} e^{-\frac{(x_i - x_i')^2}{\phi_i}}$$

- For m observations, will have the covariance matrix, $C = \sigma^2 R$
- Vector of responses follow a multivariate normal, N(μ, C)

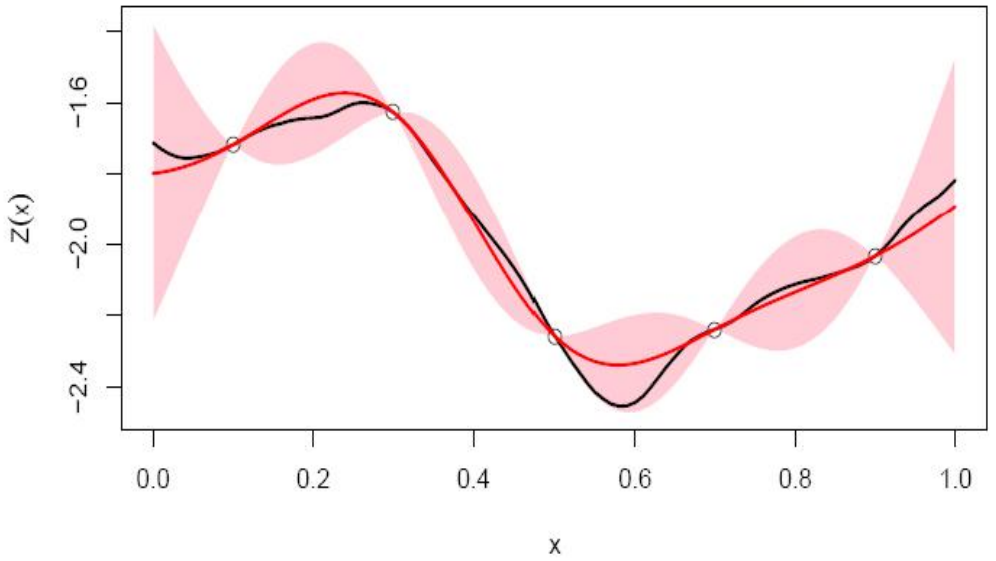
Gaussian process emulates response surface with estimates of uncertainty



$$\hat{y}(x^*) = \hat{\mu} + r'\hat{R}^{-1}(y - \hat{\mu})$$

$$s^2(x^*) = \hat{\sigma}^2\left(1 - r'\hat{R}^{-1}r + \frac{\left(1 - 1'\hat{R}^{-1}r\right)^2}{1'\hat{R}^{-1}1}\right)$$

GPs tend to do well for small-moderate sample sizes and smoothly varying functions

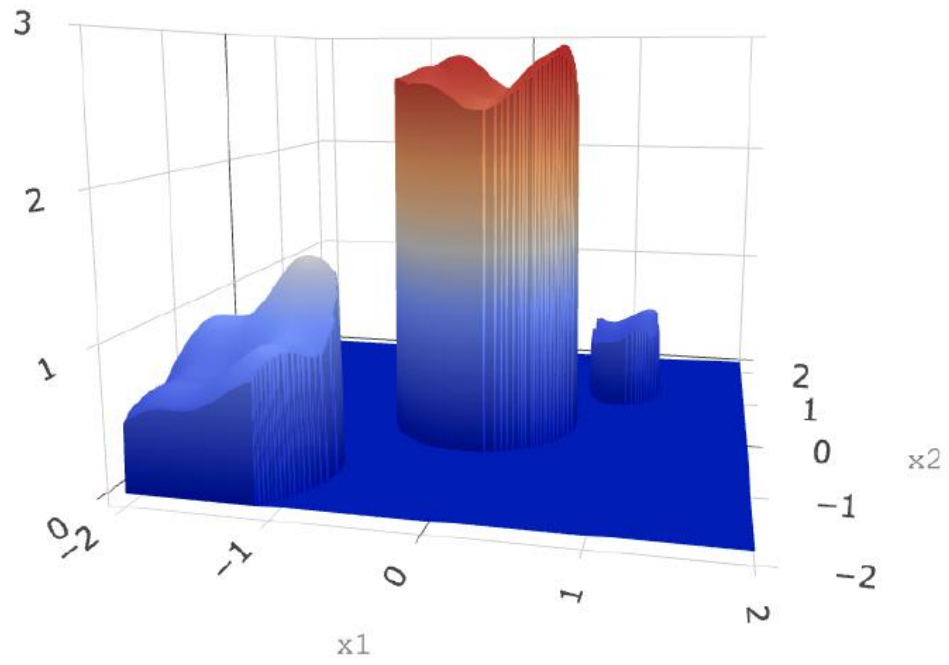


SFU
Department of Statistics and Actuarial Science
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Need an emulator that will adjust for discontinuities

- The usual GP specification will struggle with simulators like COMPAS

- COMPAS is smoothly varying in active regions, but only get response in some areas of the input space



Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Deep Gaussian processes (DGPs)

- There are broadly two formulations to DGPs
  (i) Damianou and Lawrence (2013)
  (ii) Dunlop et al. (2018)

- Recent interest in DGPs and emulation (Dutordoir et al., 2017, Sauer et al., 2020, Ming et al., 2021, …)

- **Our work:**
  - Adapts Dunlop et al. (2018) DGP to incorporate prior information about the smoothness of the computer model
  - Develops some theoretical results
  - Variational inference

DGP - Damianou and Lawrence

- A DGP with N hidden layers in this form is defined by composition of functions

$$u_1 : D \subseteq \mathbb{R}^d \to \mathbb{R}^{d_1'} \text{ and } u_n : \mathbb{R}^{d_{n-1}'} \to \mathbb{R}^{d_n'}$$

that are conditionally Gaussian

$$u_1(\mathbf{x}) \sim GP(\mathbf{0}, k_1(\mathbf{x}; \boldsymbol{\theta}_1))$$

$$u_n(\mathbf{x})|u_{n-1}(\mathbf{x}) \sim GP(\mathbf{0}, k_n(\boldsymbol{\theta}_n(u_{n-1}(\mathbf{x})))) \text{ for } n = 2, \ldots, N+1$$

- **Comments:**
  - Is an example of space warping
  - Final layer is a stationary GP, with design points $u_N(X)$
  - Is a generalization of the non-stationary GP model of Schmidt and O'Hagan (2003), which is builds on Sampson and Guttorp (1992)

DGP - Dunlop, Stuart, Girolami, and Teckentrup

- A DGP with N hidden layers in this form is defined by sequences of length-scale functions that are conditionally Gaussian

  $$u_n : D \subseteq \mathbb{R}^d \to \mathbb{R}$$

  that are conditionally Gaussian

  $$u_1(\mathbf{x}) \sim GP(\mathbf{0}, k_1(\mathbf{x}; \boldsymbol{\theta}_1))$$

  $$u_n(\mathbf{x}) | u_{n-1}(\mathbf{x}) \sim GP(\mathbf{0}, k_n(\mathbf{x}; \boldsymbol{\theta}_n(u_{n-1}(\mathbf{x})))) \text{ for } n = 2, \ldots, N+1$$

- **Comments:**
  - Covariance still a function of the original inputs, but the covariance between observations varies across the input space
  - Builds on non-stationary modeling of Paciorek and Schervish (2004)

DGP - Dunlop, Stuart, Girolami, and Teckentrup

- Specify covariance as (Paciorek and Schervish, 2004):

$$k_1(\mathbf{x}, \mathbf{x}') = \sigma_1^2 \rho_s(\| \mathbf{x} - \mathbf{x}' \|_2)$$

For n = 2, …, N+1

$$k_n(\mathbf{x}, \mathbf{x}'; [(u_{n-1}(\mathbf{x}), u_{n-1}(\mathbf{x}')]) =$$

$$= \sigma_n^2 \frac{|\Sigma(u_{n-1}(\mathbf{x}))|^{1/4} |\Sigma(u_{n-1}(\mathbf{x}'))|^{1/4}}{|(\Sigma(u_{n-1}(\mathbf{x}) + \Sigma(u_{n-1}(\mathbf{x}'))/2|^{1/2}} \rho_s \left(\sqrt{Q(\mathbf{x}, \mathbf{x}', \Sigma(u_{n-1}(\mathbf{x})), \Sigma(u_{n-1}(\mathbf{x}'))} \right)$$

where

$$Q(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \left(\frac{\Sigma(u_{n-1}(\mathbf{x})) + \Sigma(u_{n-1}(\mathbf{x}'))}{2} \right)^{-1} (\mathbf{x} - \mathbf{x}'), \qquad \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$$

$$\Sigma(u_{n-1}(\mathbf{x})) = F(u_{n-1}(\mathbf{x}))\mathbf{I}_d = e^{\alpha u(\mathbf{x})} \mathbf{I}_d$$

Adapts to local anisotropy

DGP - Dunlop, Stuart, Girolami, and Teckentrup

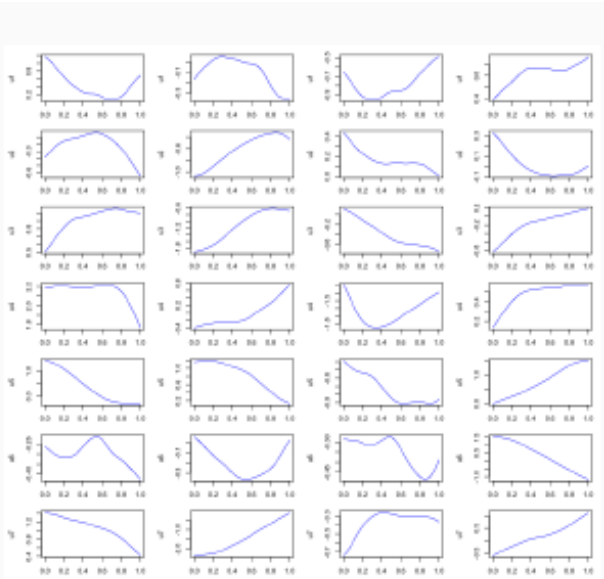
- **Comments:**
  - The u's in this setting adjust the correlation length-scale as a function of $\boldsymbol{x}$ and $\boldsymbol{x'}$
  - Dunlop et al. suggested $\alpha = 1$ in $F(\ )$, but this parameter controls smoothness of the response surface... we estimate α

$$= \sigma_n^2 \frac{|\Sigma(u_{n-1}(\mathbf{x}))|^{1/4}|\Sigma(u_{n-1}(\mathbf{x}'))|^{1/4}}{|(\Sigma(u_{n-1}(\mathbf{x}) + \Sigma(u_{n-1}(\mathbf{x}'))/2|^{1/2}} \rho_s \left(\sqrt{Q(\mathbf{x}, \mathbf{x}', \Sigma(u_{n-1}(\mathbf{x})), \Sigma(u_{n-1}(\mathbf{x}'))}\right)$$

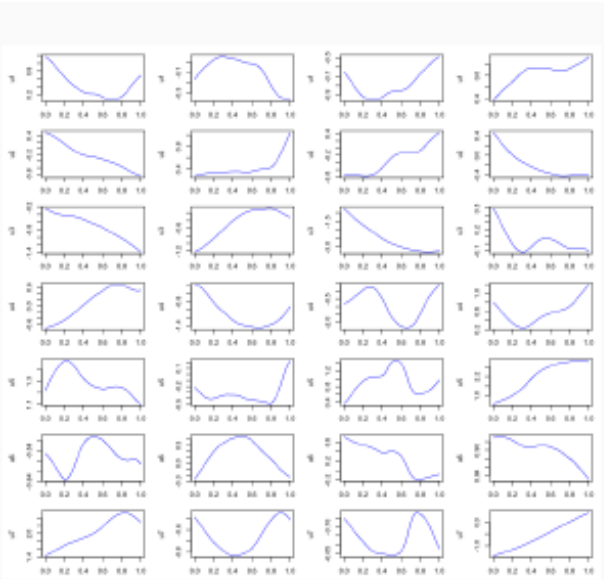
$$\Sigma(u_{n-1}(\mathbf{x})) = F(u_{n-1}(\mathbf{x}))\mathbf{I}_d \quad = e^{\alpha u(\mathbf{x})}\mathbf{I}_d$$

$$\sqrt{Q(\mathbf{x}, \mathbf{x}')} = \frac{||(\mathbf{x} - \mathbf{x}')||_2}{\sqrt{[F(u_{n-1}(\mathbf{x})) + F(u_{n-1}(\mathbf{x}'))]/2}}$$

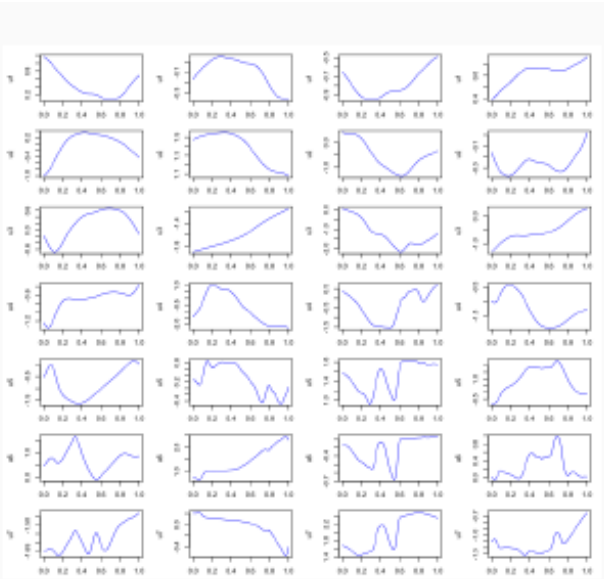
We have inputs and outputs



(a) $F(u) = exp(0.1u)$



(b) $F(u) = exp(u)$



(c) $F(u) = exp(2u)$

Look at final (data layer)

A useful proposition

- **Proposition 1:** $u_{n\text{-}1}(x)$ is constant iff $u_n(x)|u_{n\text{-}1}(x)$ is a stationary GP
- Gives an idea if it is necessary to use a DGP at all
- Can prove a similar result for Damianou and Lawrence (2013)
- Can also derive equivalence conditions between the two specifications for 1-hidden layer

Model

- A DGP with N hidden layers in this form is defined by sequences of length-scale functions that are conditionally Gaussian

  $$u_n : D \subseteq \mathbb{R}^d \to \mathbb{R}$$

  that are conditionally Gaussian

$$u_1(\mathbf{x}) \sim GP(\mathbf{0}, k_1(\mathbf{x}; \boldsymbol{\theta}_1))$$
$$u_n(\mathbf{x}) | u_{n-1}(\mathbf{x}) \sim GP(\mathbf{0}, k_n(\mathbf{x}; \boldsymbol{\theta}_n(u_{n-1}(\mathbf{x})))) \text{ for } n = 2, \ldots, N+1$$

- **Comment:**
  - $\boldsymbol{\theta}_n(\ )$ is a parameter vector
  - Can view as a Bayesian hierarchical model with a prior, hyper-prior, … on the length scale functions

Inference

- We have a lot of things to estimate
- Need to estimate $\boldsymbol{u}$'s for each $\boldsymbol{x}$ at each layer
- If interested in complex function, like the COMPAS model, need **a lot of data**
- Found MCMC not suitable for our settings, though could be useful for smaller dimensions and less complex functions
- Used variational methods instead

Variational inference

- **Idea:** Choose a family of distributions, $q(\ )$, for the unknowns at each layer (e.g., the u's) and estimate the parameters of $q(\ )$ so that it as close a possible to the true posterior distribution

- Closeness is measure via the KL divergence between the simple model and the posterior that is hard to evaluate(cannot be done directly

- Instead minimize evidence lower bound (ELBO) that function that is equal to it up to a constant

$$ELBO = E_{\text{variational posterior}}\left[log(\frac{\text{joint distribution of data and parameters}}{\text{variational posterior}})\right]$$

Doubly stochastic variational inference (DVSI)

- DSVI was proposed for inference in the DGP for the other formulation of DGPs (Salimbeni et al. (2017))

- Fast Inference using DSVI
  - Employ a sparse inducing point variational framework (Matthews et al., (2016))
  - Using two sources of stochasticity in evaluation of the ELBO
  - Using *Tensorflow* (Abadi et al., 2015)
  - Using *GPflow*, Python package for GPs (Matthews et al., 2017)

- Need to adapt to our model and ELBO

Key modifications

- Well, the model is different
- Include inference on the smoothness parameter
- Have to adapt the variational inference accordingly
- Need to derive ELBO... lots of marginalization since choose variational distributions to be Gaussian

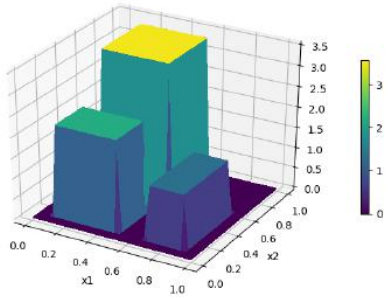
$$\mathcal{L}_{DGP} = \mathbb{E}_{q(\{\mathbf{u}_n, \tilde{\mathbf{u}}_n\}_{n=1}^N, \alpha)}\left[\log \frac{\mathbb{P}(\mathbf{y}, \{\mathbf{u}_n, \tilde{\mathbf{u}}_n\}_{n=1}^N, \alpha)}{q(\{\mathbf{u}_n, \tilde{\mathbf{u}}_n\}_{n=1}^N, \alpha)}\right]$$

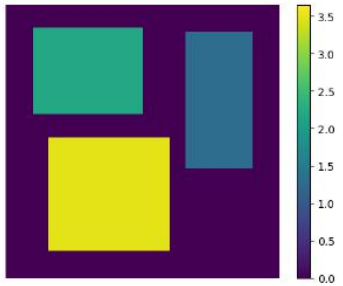
Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Simple example

$$\eta(x_1, x_2) = \begin{cases} 1.3 & x_1 \in [0.66, 0.91] \text{ and } x_2 \in [0.4, 0.91] \\ 2.2 & x_1 \in [0.1, 0.5] \text{ and } x_2 \in [0.6, 0.92] \\ 3.5 & x_1 \in [0.15, 0.6] \text{ and } x_2 \in [0.1, 0.52] \\ 0 & \text{o.w.} \end{cases}$$



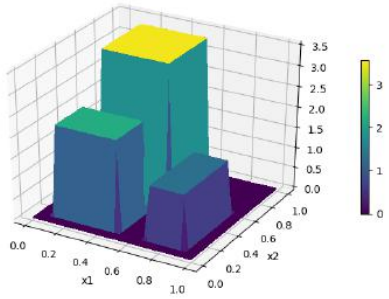


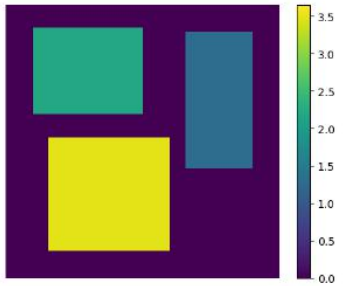
- Design: 25 x 25 grid in $[0,1]^2$
- Use 200 inducing points
- Validation on 70 x 70 grid in $[0,1]^2$
- Measure goodness using Nash-Sutcliffe efficiency

$$\tilde{R}^2 = 1 - \frac{MSPE_{validation}}{Var\left(Y(\mathbf{X}_{validation})\right)}$$

Simple example

$$\eta(x_1, x_2) = \begin{cases} 1.3 & x_1 \in [0.66, 0.91] \text{ and } x_2 \in [0.4, 0.91] \\ 2.2 & x_1 \in [0.1, 0.5] \text{ and } x_2 \in [0.6, 0.92] \\ 3.5 & x_1 \in [0.15, 0.6] \text{ and } x_2 \in [0.1, 0.52] \\ 0 & \text{o.w.} \end{cases}$$





DGP	$\widetilde{R}^2$	Coverage probability (95%)
Variational inference on α	0.92	94%
Optimized α	0.91	92%
α = 1	0.88	90%

Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Back to COMPAS

- Has 12-dimensional input

- Have about 2,000,000 simulations... about 552,010 giving a BBH

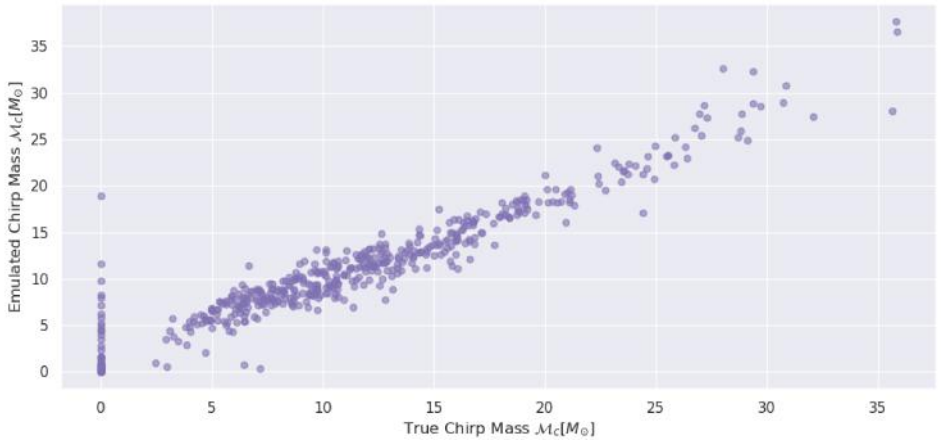
Input	Description
α	the common envelope ejection efficiency
σ	parameter describing the supernova kick distribution
$flbv$	a parametrisation of the mass loss rate during the luminous blue variable phase
m_1	the mass of the initially more massive star at the start of its life, in solar masses
m_2	the mass of the initially less massive star at the start of its life, in solar masses.
separation	the initial orbital separation, in astronomical units
kickSize1	the first supernova kick
kickSize2	the second supernova kick
kick θ_1,θ_2	angles defining the direction of the kick
kick ϕ_1, ϕ_2	angles defining the direction of the kick

Building an emulator

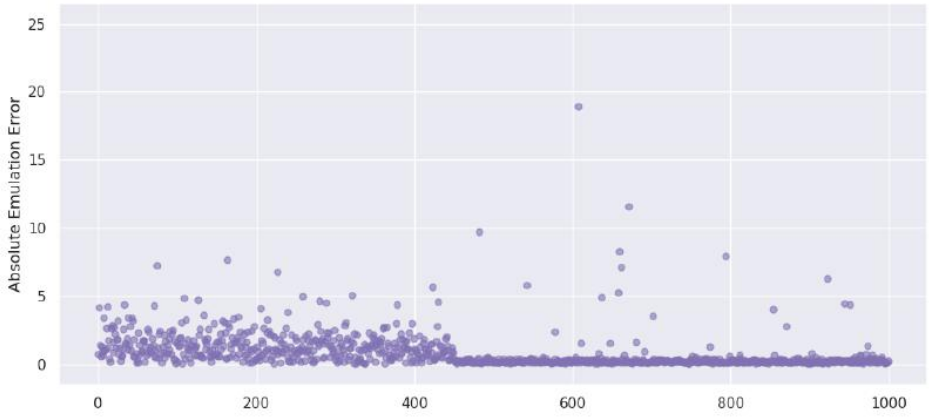
- Have about 2,000,000 simulations... about 552,010 giving a BBH
- Kept 1,000 validation runs as validation (450 active active and 550 inert)
- Fit a three layer DGP using mini-batch size of 5,000 samples and 100 inducing points
- Priors on all parameters were Gaussian, in particular, $\pi(\alpha) \sim N(4, 1.5)$

Predictions from the emulator

Predicted vs actual - validation



Absolute prediction error - validation



DGP	$\widetilde{R}^2$	Coverage probability (95%)
Variational inference on α	0.96	92%

Lin et al. (2021) used the same data using a local GP classifier and local GP. Misclassified 40% of training data and thus overall explanation of data is far less

Outline

- Application
- Gaussian processes
- Deep Gaussian processes
- Toy example
- COMPAS emulation
- Recap

Recap

- Have adapted DGP for emulating computer models with an to include prior information on smoothness of the function
- Developed some theory to explore model performance
- Developed variational Bayes estimation approach
- Applied method emulate COMPAS data
- Next step is to use distribution of observations and emulator to constrain inputs (e.g., a type of calibration problem)

References

- Cressie, N. and Johannesson, G. (2008) "Fixed rank kriging for very large spatial data sets". *JRSS 'B.*
- Dutordoir, V., Knudde, N., van der Herten, J., Couckuyt, I. and Dhaene, T., 2017, December. Deep Gaussian process metamodeling of sequentially sampled non-stationary response surfaces. In *2017 Winter Simulation Conference (WSC)* (pp. 1728-1739). IEEE.
- Dunlop, M.M., Girolami, M.A., Stuart, A.M. and Teckentrup, A.L., 2018. How deep are deep Gaussian processes?. *Journal of Machine Learning Research*, *19*(54), pp.1-46.
- Gramacy, R.B., and Apley, D.W. (2015) Local Gaussian process approximation for large computer experiments". *JCGS*
- Kaufman, C., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. (2011) "Efficient Emulators of Computer Experiments Using Compactly Supported Correlation Functions, With An Application to Cosmology". *Annals of Applied Statistics.*
- Kenndy, M. and O'Hagan, A. (2001) "Bayesian calibration of computer models". *JRSS 'B.*
- Lin, L., Bingham, D., Broekgaarden, F. and Mandel, I., 2021. Uncertainty Quantification of a Computer Model for Binary Black Hole Formation. *arXiv preprint arXiv:2106.01552.*

References

- Matthews, A.G.D.G., Hensman, J., Turner, R. and Ghahramani, Z., 2016, May. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics* (pp. 231-239). PMLR.

- Ming, D., Williamson, D. and Guillas, S., 2021. Deep Gaussian Process Emulation using Stochastic Imputation. *arXiv preprint arXiv:2107.01590*.

- Salimbeni, H. and Deisenroth, M., 2017. Doubly stochastic variational inference for deep Gaussian processes. *arXiv preprint arXiv:1705.08933*.

- Sauer, A., Gramacy, R.B. and Higdon, D., 2020. Active Learning for Deep Gaussian Process Surrogates. *arXiv preprint arXiv:2012.08015*.

- Sampson, P. D. and Guttorp, P. (1992). "Nonparametric estimation of nonstationary spatial co- variance structure." Journal of the American Statistical Association, 87, 417, 108–119.

- Schmidt, A. M. and O'Hagan, A. (2003). "Bayesian inference for non-stationary spatial covariance structure via spatial deformations." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65, 3, 743–758.