Enhancing statistical inference for stochastic processes using modern statistical methods

Peter F. Craigmile

The Ohio State University department of STATISTICS

http://www.stat.osu.edu/~pfc/

Department of Applied Mathematics, University of Colorado Boulder

Sep 28, 2018

Supported in part by the US National Science Foundation (DMS-1209142, DMS-1407604, and SES-1424481) and the National Cancer Institute (R21CA212308)

Collaborators

Grant Schneider 

 Radu Herbei

Sophie (Huong) Nguyen 

 Matthew Pratola

 C. Devon Lin

Introduction

There are many situations in which an inference or statistical design problem associated with these processes is **intractable**, and **approximations** are then required.

Traditionally these approximations often come without measures of **quality**.

We can frame such problems from a **statistical perspective** so that we can **probabilistically** quantify uncertainties when making approximations.

We start with an example involving stochastic differential equations (inference).

(We will talk about design at the end.)

Stochastic differential equations (SDEs)

$$dX_t = \mu(X_t, \boldsymbol{\theta})\, dt + \sigma(X_t, \boldsymbol{\theta})\, dW_t, \quad 0 \le t \le T,$$

where $X_0 = x_0$ is the initial value of the process and $\{W_t\}$ is a standard Brownian motion (BM).

Assume the **drift function** $\mu(\cdot, \cdot)$ and **diffusion function** $\sigma(\cdot, \cdot)$ are known up to the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is some compact set in $\mathbb{R}^p$.

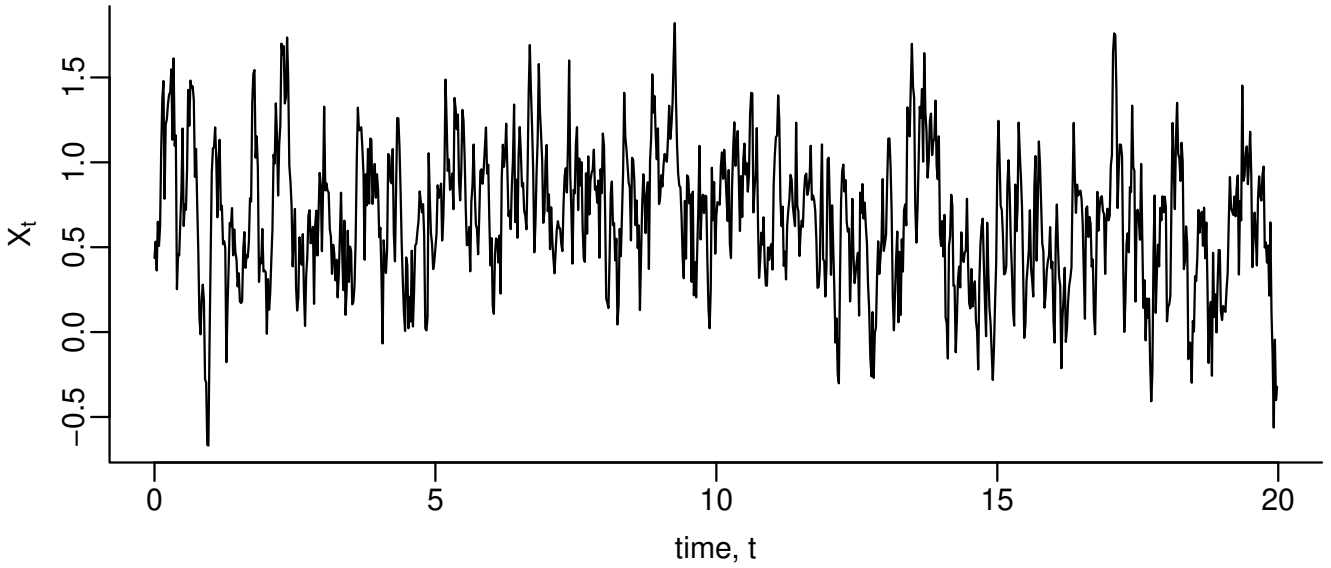
(Also assume locally Lipschitz with linear growth bounds so that a weakly unique solution exists.)

The inference problem: **maximum likelihood estimation (MLE)** of $\boldsymbol{\theta}$ when $\{X_t\}$ is observed at N time points $\{t_i : i = 1, \ldots, N\}$.

Example: Ornstein-Uhlenbeck (OU) process

$$dX_t = (\theta_0 + \theta_1 X_t)\, dt + dW_t, \quad 0 \le t \le T,$$

where $X_0 = x_0$ is the initial value, $\theta_0 \in \mathbb{R}, \theta_1 < 0$, and W_t is a std. BM.



The likelihood function

Treating $X_0 = x_0$ as fixed, we use the **Markov property** to write the **likelihood** as the product of individual **transition densities**:

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{N} p(X_{t_i}|X_{t_{i-1}}, \boldsymbol{\theta}).$$

As the transition density **does not exist in closed-form** except for a handful of cases, **approximations** are typically necessary.

The most commonly used **Euler approximation** is

$$\xi(X_\Delta|X_0, \boldsymbol{\theta}) = n(X_\Delta; X_0 + \mu(X_0, \boldsymbol{\theta})\Delta, \sigma^2(X_0, \boldsymbol{\theta})\Delta).$$

Can do well for small Δ, but there are better approximations ...

Approximating the transition density with importance sampling

1. Partition $[0, \Delta)$ into K subintervals of width Δ/K with endpoints

   $0 = \tau_0 < \tau_1 < \ldots < \tau_K = \Delta.$

2. The **discretized transition density** [Kloeden and Platen, 1992] is

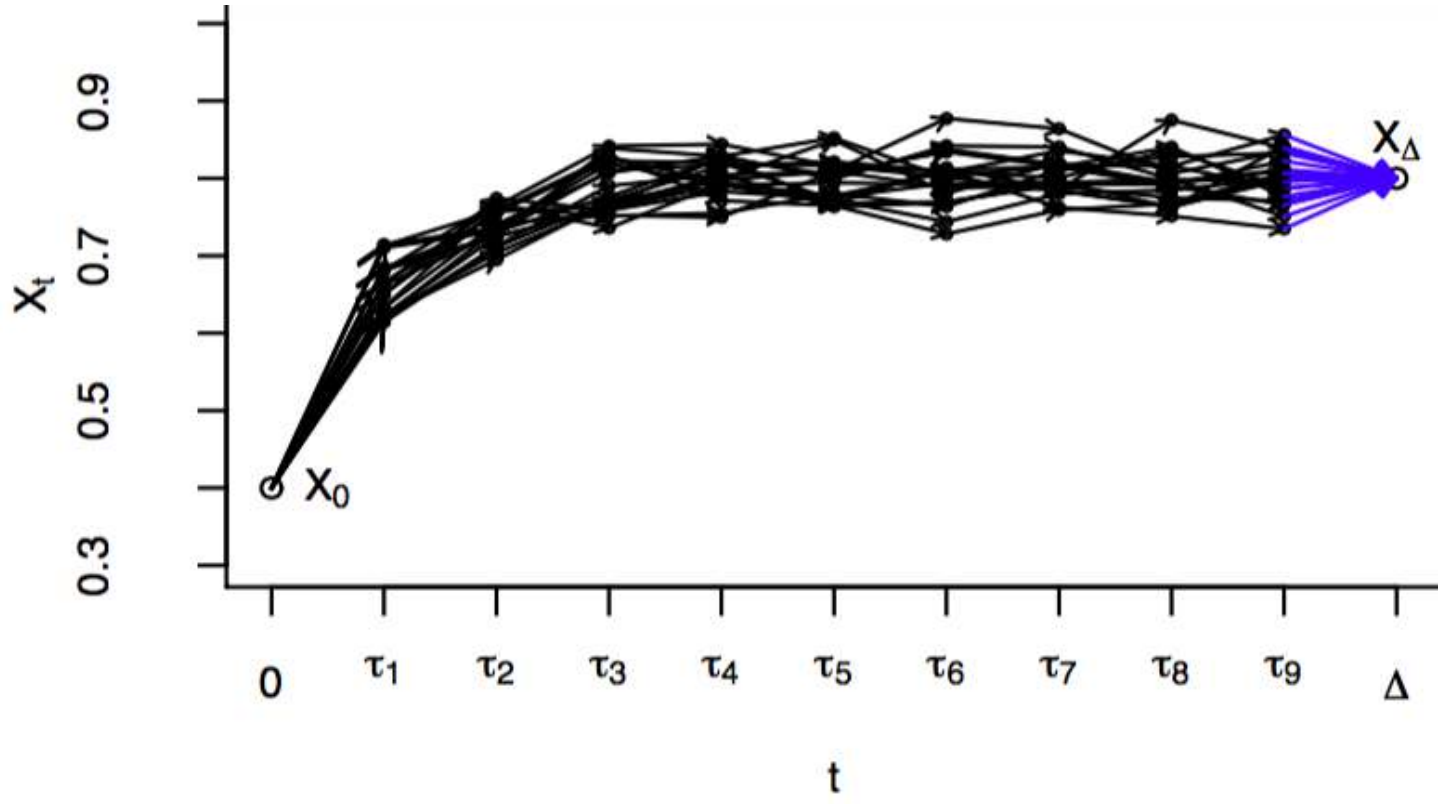
$$p^{(K)}\left(X_\Delta | X_0, \boldsymbol{\theta}\right) = \int \prod_{k=1}^{K} \xi\left(X_{\tau_k} | X_{\tau_{k-1}}, \boldsymbol{\theta}\right) \lambda(d\mathbf{X}_\tau)\,,$$

   where λ denotes the Lebesgue measure.

3. We approximate this density by **importance sampling**, using M random samples from an importance density $q(\cdot)$.

The modified Brownian bridge sampler

[Due to Durham and Gallant, 2002] Provides a compromise between accuracy and computational efficiency of approximating the transition density $p(X_\Delta|X_0)$:



Obtaining the MLE for the SDE process parameters

Limited discussion in the literature about exploring $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Fine-scale grid based methods are very computationally intensive and not efficient.

Gradient based methods suffer from Monte Carlo variability, and are also computationally intensive (require prohibitively large sample sizes, M) – we return to this later.

Using methodology from computer experiments

We believe the underlying discretized log-likelihood function

$$l^{(K)}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p^{(K)}\left(X_{t_i}; X_{t_{i-1}}, \boldsymbol{\theta}\right)$$

is smooth in $\boldsymbol{\theta}$, but our estimates are:

1. Subject to Monte Carlo variability.

2. Expensive to make – $O(KMN)$.

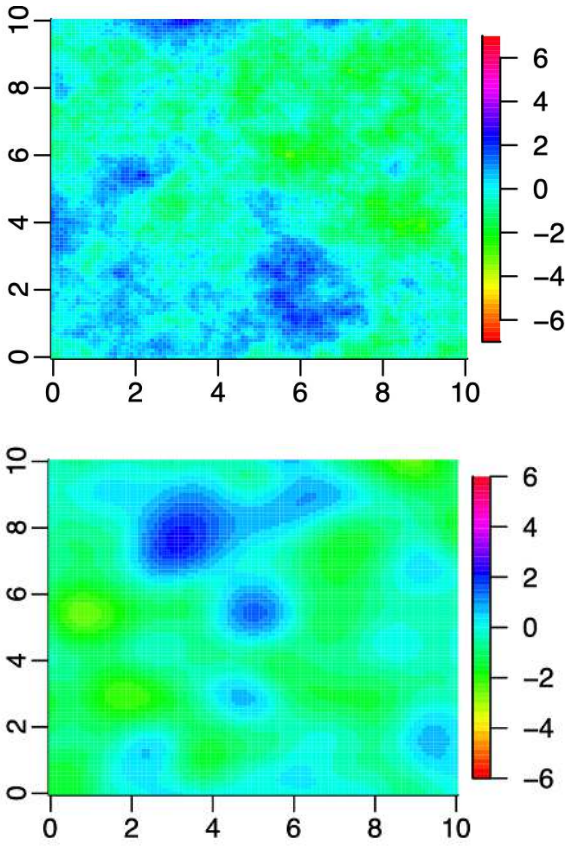
The statistical methodology for **computer experiments** deals with estimation and prediction of expensive-to-evaluate functions (here, measured under uncertainty). A good fit for what we want to do!

What is a Gaussian process?

A **Gaussian process (GP)** is a stochastic process indexed in space over some domain D which is usually a subset of $\mathbb{R}^d$.

Any finite subcollection of random variables of this process will have a joint multivariate normal distribution.

The joint distribution of a GP is characterized by its mean function $\mu(\boldsymbol{s})$ and covariance function $C(\boldsymbol{s}, \boldsymbol{s}')$ – makes it fairly easy to model.



A GP model for the estimated log-likelihood function

Start with estimates at n **initial parameter values** $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)^T$ chosen based on some **space-filling design**. (We will return to this point later.)

Letting $Y(\boldsymbol{\theta}_i)$ denote the estimates of $l^{(K)}(\boldsymbol{\theta}_i)$, assume

$$Y(\boldsymbol{\theta}_i) = l^{(K)}(\boldsymbol{\theta}_i) + \epsilon(\boldsymbol{\theta}_i), \quad i = 1, \ldots, n,$$

where $\{\epsilon(\boldsymbol{\theta}_i) : i = 1, \ldots, n\}$ is a set of independent of N$(0, \sigma^2)$ errors.

Model $\{l^{(K)}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ using a GP with mean function $\mu_L(\boldsymbol{\theta}; \beta)$ and some valid covariance function $C_L(\boldsymbol{\theta}, \boldsymbol{\theta}'; \zeta)$, where β, ζ are unknown parameters.

Predicting the log-likelihood function

By Gaussianity of the GP and data model, the distribution of $l^{(K)}(\boldsymbol{\theta}^*)$ given the data $\boldsymbol{Y}_n = (Y(\boldsymbol{\theta}_i) : i = 1, \ldots, n)^T$ is normal with a conditional mean of

$$\eta_{L,n}(\boldsymbol{\theta}^*) \;=\; \mu_{L,n}(\boldsymbol{\theta}^*) + \boldsymbol{c}_{L,n}^T(\boldsymbol{\Sigma}_{L,n} + \sigma^2 \boldsymbol{I}_n)^{-1}[\boldsymbol{Y}_n - \boldsymbol{\mu}_{L,n}],$$

and conditional variance given by

$$v_{L,n}^2(\boldsymbol{\theta}^*) \;=\; C_{L,n}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \boldsymbol{c}_{L,n}^T(\boldsymbol{\Sigma}_{L,n} + \sigma^2 \boldsymbol{I}_n)^{-1}\boldsymbol{c}_{L,n}.$$

$\boldsymbol{\mu}_{L,n}$ is a mean vector of length n with ith element $\mu_{L,n}(\boldsymbol{\theta}_i; \boldsymbol{\beta})$,

$\boldsymbol{c}_{L,n}$ is a covariance vector of length n with ith element $C_{L,n}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i; \boldsymbol{\zeta})$, and

$\boldsymbol{\Sigma}_{L,n}$ is the $n \times n$ covariance matrix with (i, j) element $C_{L,n}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j; \boldsymbol{\zeta})$.

Maximizing the likelihood using expected improvement

Let $\widetilde{\eta}_{L,n} = \max_{i=1,\ldots,n} \eta_L(\boldsymbol{\theta}_i)$ denote the maximum of the conditional mean over the explored n values of $\boldsymbol{\theta}$.

The **expected improvement** at parameter value $\boldsymbol{\theta}^*$ is [Jones et al., 1998]

$$\left[\eta_{L,n}\left(\boldsymbol{\theta}^*\right) - \widetilde{\eta}_{L,n}\right] \Phi\left(\frac{\eta_{L,n}\left(\boldsymbol{\theta}^*\right) - \widetilde{\eta}_{L,n}}{v_{L,n}\left(\boldsymbol{\theta}^*\right)}\right) + v_{L,n}(\boldsymbol{\theta}^*)\ \phi\left(\frac{\eta_{L,n}\left(\boldsymbol{\theta}^*\right) - \widetilde{\eta}_{L,n}}{v_{L,n}\left(\boldsymbol{\theta}^*\right)}\right),$$

where $\Phi(\cdot)/\phi(\cdot)$ is the standard Gaussian cdf/pdf.

The expected improvement **balances** the need to

maximize the discretized log-likelihood (the first term)

while cognizant of

the uncertainty in estimating the log-likelihood (the second term).

Sequential Gaussian-Process-Based Optimization (SGBO)

We add the parameter value $\boldsymbol{\theta}^*$ that maximizes the expected improvement.

- Estimate the discretized log-likelihood at that value, yielding $Y(\boldsymbol{\theta}^*)$.

From the new data vector and vector of estimated log-likelihoods, we update:

1. GP parameter estimates (using an empirical Bayes technique);

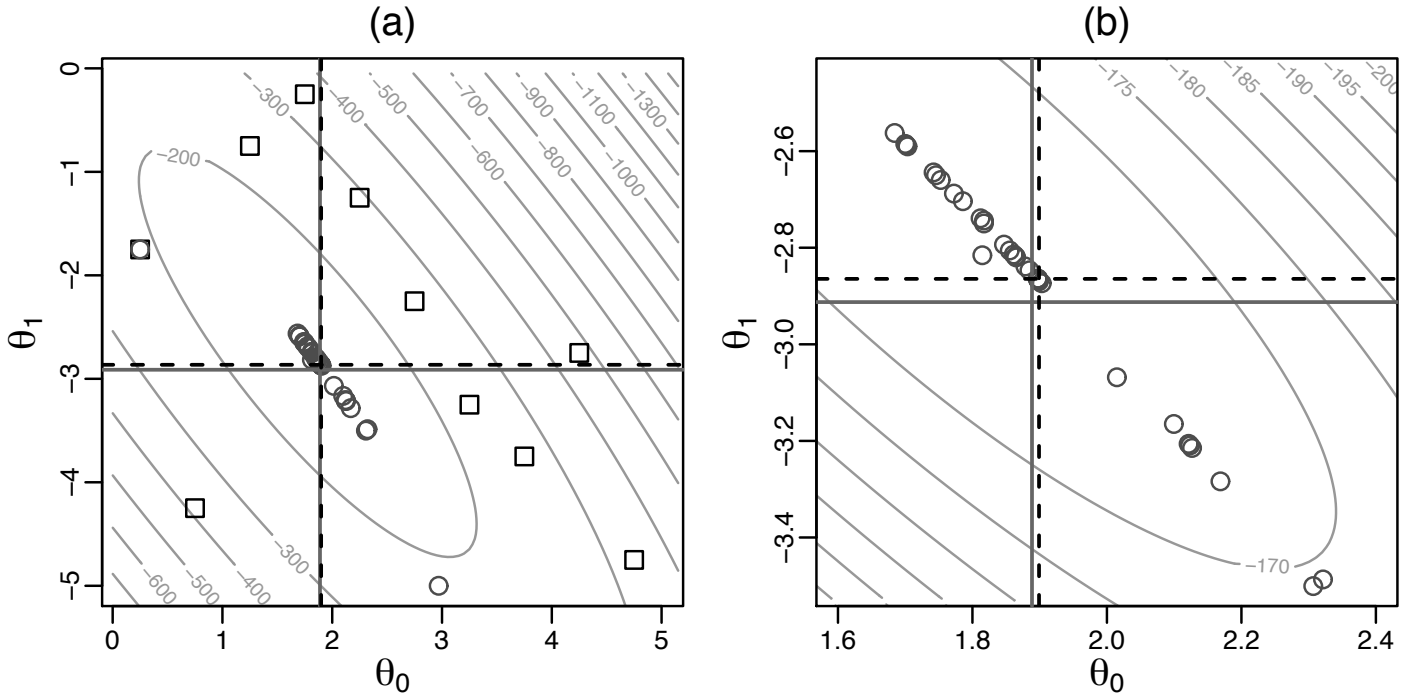
2. The conditional mean and variance for the BLUP.

Repeat until some stopping criteria is met (little change in estimated MLEs).

After n steps, straightforward to obtain the estimated MLE,

$$\widehat{\boldsymbol{\theta}} = \arg \max_{i=1,\ldots,n} \eta_{L,n}(\boldsymbol{\theta}_i).$$

An example SGBO path



(a) A contour plot of the discretized log-likelihood for an OU process with $\theta_0 = 2$ and $\theta_1 = -3$ ($\theta_2 = 1$ is fixed). The solid horizontal and vertical lines denote the exact MLEs of θ_0 and θ_1 respectively, and the dashed lines denote the SGBO-based estimate. For the SGBO method, the squares indicate the initial parameter values, and the circles denote the values added sequentially. (b) is a zoomed in version of (a).

Remarks on the SGBO method

- We can obtain approximate $(1-\alpha)\%$ joint confidence regions for $\boldsymbol{\theta}$ directly from the conditional mean using a **likelihood ratio test**.
- We carried out **simulations** to estimate $\boldsymbol{\theta}$ for OU and Generalized Cox-Ingersoll-Ross (GCIR, Chan et al. [1992]) processes:
  - For OU processes there is no appreciable difference between the SGBO-based estimator and the exact MLE.
  - SGBO clearly outperforms grid methods.
  - SGBO with $K = 10$ outperforms $K = 5$.
  - Methods improve with M with a loss in computational efficiency; it helps to start with more initial points.

Estimating Deep Flow in the South Atlantic Ocean

Estimating the **state** of the world's oceans is a fundamental problem.

Ocean circulation cannot be measured directly.

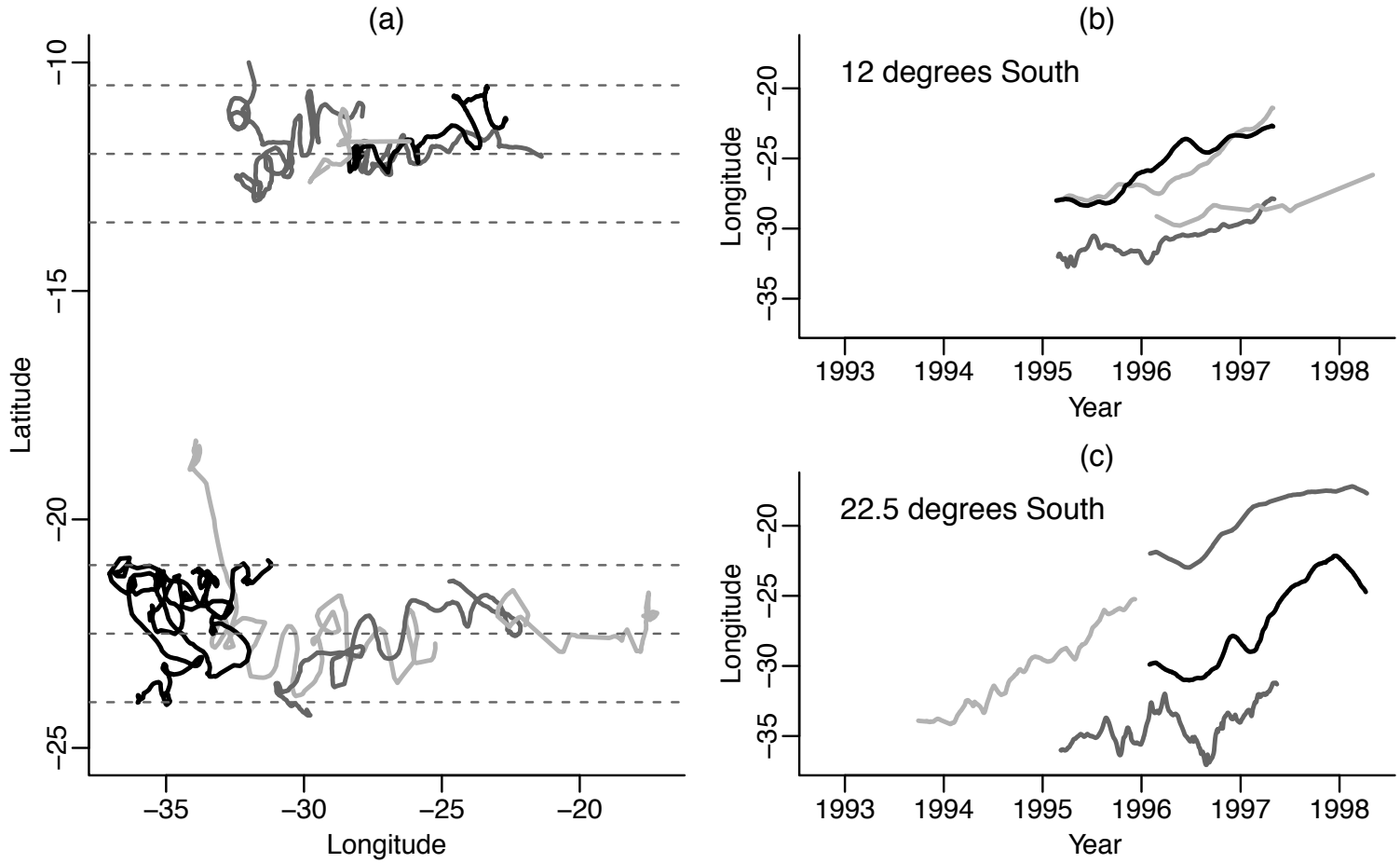
It is inferred from other physical and chemical properties of the ocean, such as measurements of water temperature, salinity, silica, etc.

This is an **inverse problem** [Wunsch, 1996].

We focus on estimating **water velocities**.

Using **float data** [Hogg and Owens, 1999], we estimate the **deep flow** ($2500m$) in two latitude bands (12^oS and $22.5\ ^oS$) in the western South Atlantic Ocean.

Estimating Deep Flow in the South Atlantic Ocean, cont.



In this area the circulation structure is dominated by strong alternating zonal jets [Hogg and Owens, 1999, McKeague et al., 2005].

Estimating Deep Flow in the South Atlantic Ocean, cont.

Let $\{X_t^{(i)} : t \in [0,T]\}$ denote the underlying (continuous) longitude process for float i in a specific latitude band. Data are available every two days.

Assume $\{X_t^{(i)}\}$ satisfies

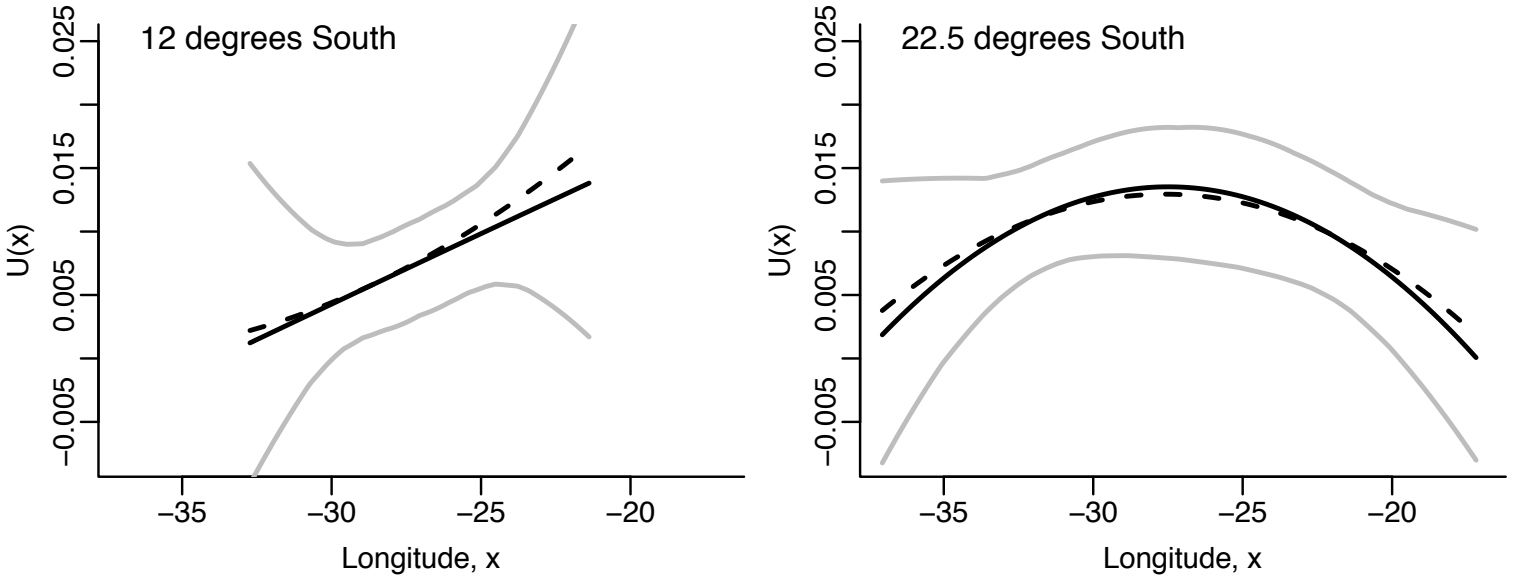
$$dX_t^{(i)} = U(X_t^{(i)})\, dt + \sigma\, dW_t^{(i)} \quad X_0^{(i)} = x_0^{(i)}, \quad t \in [0,T], \tag{1}$$

where $U(\cdot)$ is the **zonal velocity of interest** assumed common for each series in a given latitude band, σ is the diffusion coefficient, and $\{W_t\}$ is a standard BM. We assume conditional independence over i.

Estimating Deep Flow in the South Atlantic Ocean, cont.

We estimated the three parameters of a quadratic model for the common zonal velocity function $U(x)$ using the SGBO method.

$K = 10$, $M = K^2$ and $n = 20 \times 3 = 60$ initial points.



(Dashed line: Euler approximation)

The initial sampling design, revisited

"Start with estimates at n **initial parameter values** $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)^T$ chosen based on some **space-filling design**."

How do we pick a **good** spatial design?

Design of experiments (DOE)

Typically DOE targets a particular aspect of a model which is deemed **important**, such as the prediction error in a spatial model, or the variance of some set of parameters of a regression model.

Given n input settings $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ drawn from some set $\chi \subset \mathbb{R}^d$, a **design criterion** $\mathcal{J}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$ is specified.

The n**-run optimal design** involves finding the input settings $\boldsymbol{\Xi}_n$ that minimize this criterion:

$$\boldsymbol{\Xi}_n = \arg \min_{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n} \mathcal{J}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n).$$

Finding the n-run optimal design is difficult

Assuming χ is a discretized candidate set of size N, the number of possible designs to explore is $\binom{N}{n}$.

The (Federov) **Exchange Algorithm** [Fedorov, 1972], the most popular approach to solving this problem, performs one-at-a-time updates to the design.

Problems: there are a large number of possible designs and the optimization problem is often multi-modal.

Modern optimization alogrithms such as particle-swarm methods and simulated annealing [e.g. Chen et al., 2013] can be difficult to implement reliably in modern settings [Nguyen, 2018].

Two popular design criteria for GPs

1. **Integrated mean squared prediction error (IMSPE)** optimal designs, $\mathcal{J} = \int_{\mathbf{x}} \left(Y(\mathbf{x}) - E[Y(\mathbf{x})|\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n]\right)^2 d\mathbf{x}$.

   (Minimizes the error in out-of-sample predictions.)

2. **Entropy optimal designs**, $\mathcal{J} = E\left[-\log(f_{\mathbf{Y}})\right]$ where $f_{\mathbf{Y}}$ is the a multivariate Gaussian density.

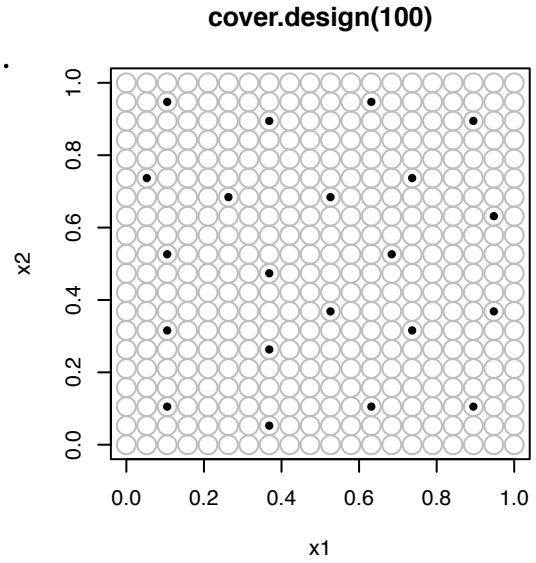
   (Provides improved estimates of the GP correlation parameter – useful for statistical inference.)

Evaluating these criteria are expensive

Requires $\mathcal{O}(n^3)$ operations on the potentially large $n \times n$ correlation matrix.

One way to avoid this is to use **geometrically-motivated** designs:

- Latin hypercube sampling (LHS) designs [McKay et al., 1979];
- Minimax distance designs [Johnson et al., 1990].



Both approaches lead to "space-filling designs", but are now removed from considering $\mathcal{J}$.

Using random measures for design

Using $\mathcal{J}$, there is no guarantee that any optimization algorithm will find the true optimal design.

If we use geometrically-motivated designs, we do not know how good the design is relative to other designs.

Key idea: We use **statistics** to quantify the uncertainty in picking a design.

Two approaches

1. Estimate the distribution of the criterion function $\mathcal{J}$ under **uniform random sampling**, and use this **criterion distribution** to guide the search for **near-optimal designs** [Nguyen et al., 2018];

2. **Emulate** spatial designs for GPs using **spatial point processes** [Pratola et al., 2018].

Near-optimal design

We call the set of **eligible designs** $\mathcal{D}$ the collection of n-run designs that we are able to sample from.

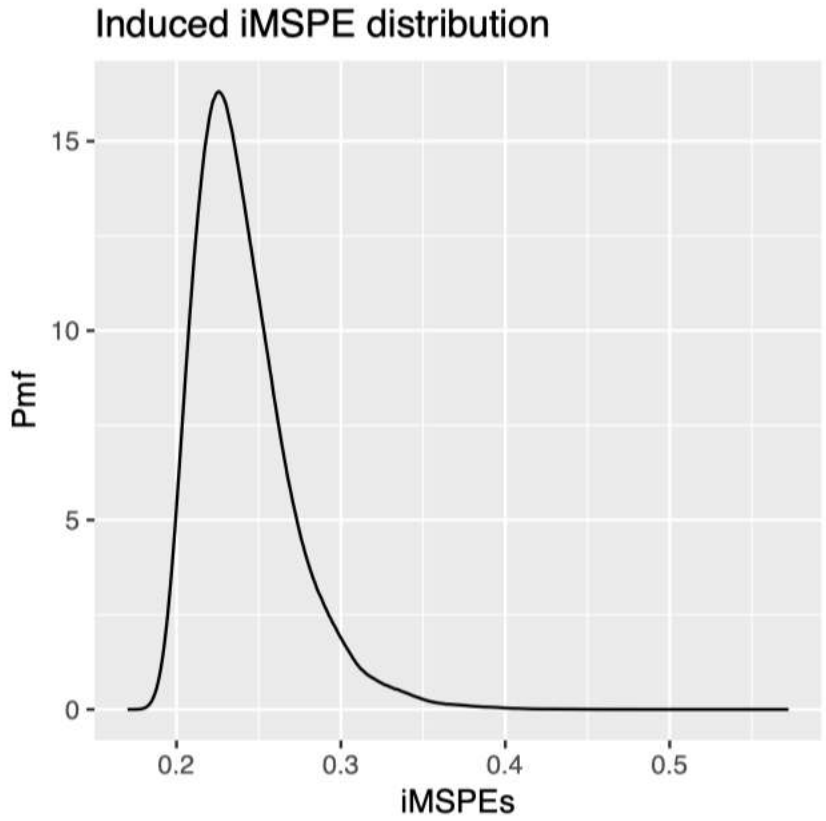
Let $\mathcal{H}$ be a **uniform random measure** over these **eligible designs**.

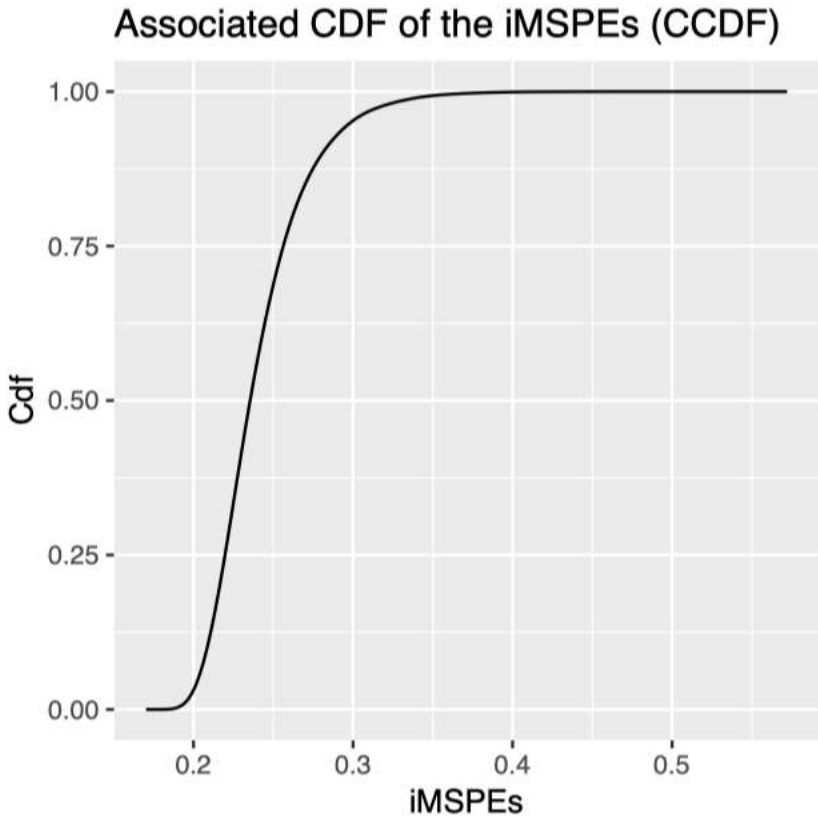
For $\Xi \sim \mathcal{H}$ then $Q = \mathcal{J}(\Xi)$ denotes the random variable induced by $\mathcal{H}$ acting upon the criterion function $\mathcal{J}$.

We say that $F_Q(q) = P(Q \leq q)$ is the **criterion cumulative distribution function (CCDF)**.

An example

Suppose we want to find the near-optimal IMSPE design of size $N = 6$ from the spatial domain $[0, 1]^2$.





Near optimal design

Definitions: The $(100p)$**-percentile near optimal class** is the set of all designs whose criteria are less than $\kappa_Q(p)$, which is the $(100p)$-percentile of the CCDF $F_Q(\cdot)$ induced by uniform measure $\mathcal{H}$ over the design space $\mathcal{D}$.

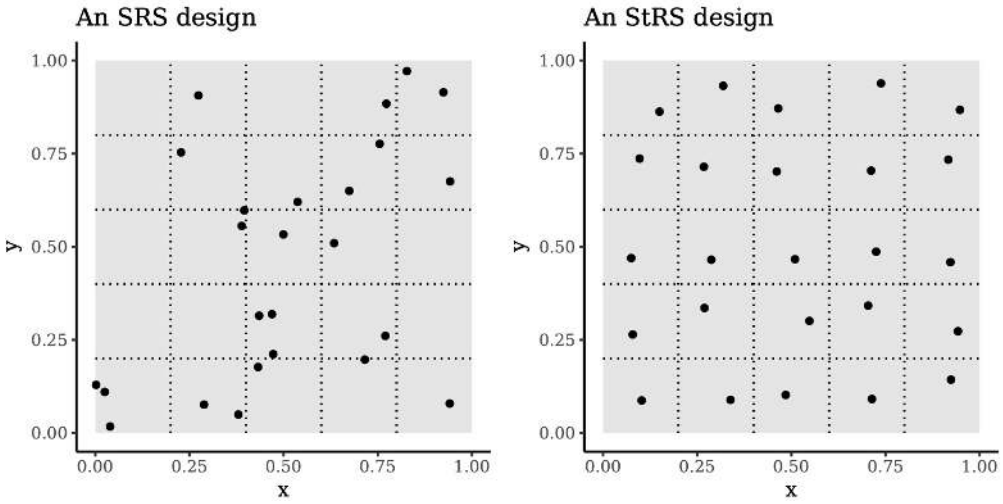
A **near optimal design (NOD)** is a design that lies in the near optimal class.

This motivates a new approach to the design optimization problem:

- We **search** for a $(100p)$-percentile NOD within the design space with respect to a criterion function and uniform random measure.
- In some cases this requires estimating $\kappa_Q(p)$ (a statistical problem!)

Some statistical strategies to find a NOD

1. Simple random sampling
2. Stratified random sampling
3. A cross-design approach via quantile regression



In the second approach, we build a **design emulator** to mimic what makes a good design. We can also use a spatial point process as an emulator [Pratola et al., 2018].

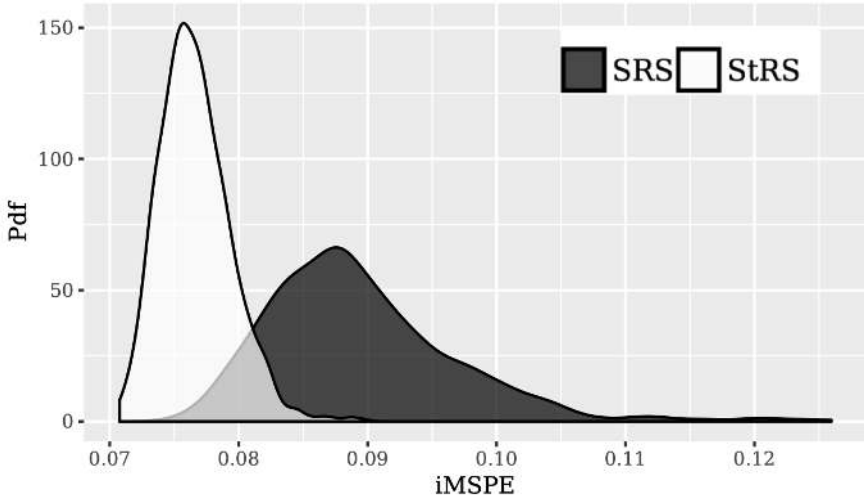


Illustration: Designed Stochastic Gradient Descent (SGD)

SGD used extensively in statistical machine learning to scale model training to big data for a variety of applications including linear models, clustering, GP regression and deep neural networks.

Idea:

1. We take small random subsets of the data, called **batches**, to estimate the gradient.
2. We sacrifice an increase in estimation variance for computational gain so the parameter space of the model can be more efficiently explored when fitting models to big data.

What if we use a **design emulator** to improve batch selection?

Illustrative simulation

Generate observations from a 5-dimensional linear regression model:

$$Y(\mathbf{x}) = \beta_0 + \beta_1 \sin(2\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2 + \beta_3 (x_4 - 0.5)^2 + \beta_4 x_4 + \beta_5 x_5 + \epsilon.$$

$\beta_0, \ldots, \beta_5$ indep Unif$(-10, 10)$, $\epsilon \sim N(0, 1)$.

The model was fit using SGD with batches of size 23, 42, 63 and 83.

SGD iterates over all the batches of data in random order and repeats this entire process a number of times, called **epochs** – we use 200 epochs.

Sample size $N = 50 \times$ `batchsize`.

We repeat 100 times.

Results

Subset size	MSE ratio $\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
23	1.97	1.47	1.27	2.92	1.97
43	1.62	1.04	1.45	2.72	1.67
63	1.56	0.94	2.00	2.59	1.56
83	1.43	1.20	1.32	2.38	1.44

Ratio of MSE of parameter estimates for random subsets versus designed subsets when using SGD to fit the model.

Values greater than 1.0 indicate that designed subsets outperformed random subsets.

Conclusions

Modern statistical methods can be used to evaluate the uncertainty inherent in approximations across different statistical inference and design problems.

This has a wide application to many areas.

In the future:

1. High dimensional spatio-temporal inference.
2. Bayesian inference.

Thank you!

References

K. C. Chan, G. A. Karolyi, F. A. Longstaff, and A. B. Sanders. An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance*, 47:1209–1227, 1992.

R. B. Chen, D. N. Hsieh, Y. Hung, and W. Wang. Optimizing latin hypercube designs by particle swarm. *Statistics and Computing*, 23:663–676, 2013.

G. B. Durham and A. R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20:297–316, 2002.

V. V. Fedorov. *Theory of Optimal Experiments*. Elsevier, New York, New York, 1972.

N. G. Hogg and W. B. Owens. Direct measurement of the deep circulation within the Brazil Basin. *Deep Sea Research Part II: Topical Studies in Oceanography*, 46:335–353, 1999.

M. E. Johnson, L. M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13: 455–492, 1998.

P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer, New York, NY, 1992.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42:55–61, 1979.

I. W. McKeague, G. Nicholls, K. Speer, and R. Herbei. Statistical inversion of South Atlantic circulation in an abyssal neutral density layer. *Journal of Marine Research*, 63:683–704, 2005.

H. Nguyen. *Near-optimal designs for Gaussian Process regression models*. PhD thesis, The Ohio State University, 2018.

H. Nguyen, P. F. Craigmile, and M. T. Pratola. Near-optimal designs for gaussian process regression models. 2018. Under preparation.

M. T. Pratola, C. D. Lin, and P. F. Craigmile. Optimal design emulators: A point process approach. *arXiv preprint arXiv:1804.02089*, 2018.

C. Wunsch. *The Ocean Circulation Inverse Problem*. Cambridge University Press, Cambridge, England, 1996.