Maximum likelihood estimation for stochastic differential equations using sequential Gaussian-process-based optimization

Peter F. Craigmile

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

Department of Statistical Sciences,

University of Toronto,

Toronto, Canada

29 September 2016

G. Schneider, P. F. Craigmile and R. Herbei (2016), Maximum likelihood estimation for stochastic differential equations using sequential kriging-based optimization. Technometrics, DOI: 10.1080/00401706.2016.1153522.


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

Stochastic differential equations

Many phenomena that arise in finance, biology, ecology, and other areas are modeled in continuous time using a real-valued diffusion process, $\{X_t\}$.

Consider diffusions that are the solution to the **stochastic differential equation (SDE)**

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

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

Stochastic differential equations, cont.

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

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

We 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 $\mathbb{R}^p$.

We further assume the drift and diffusion functions are locally Lipschitz with linear growth bounds so that a weakly unique solution to the diffusion.

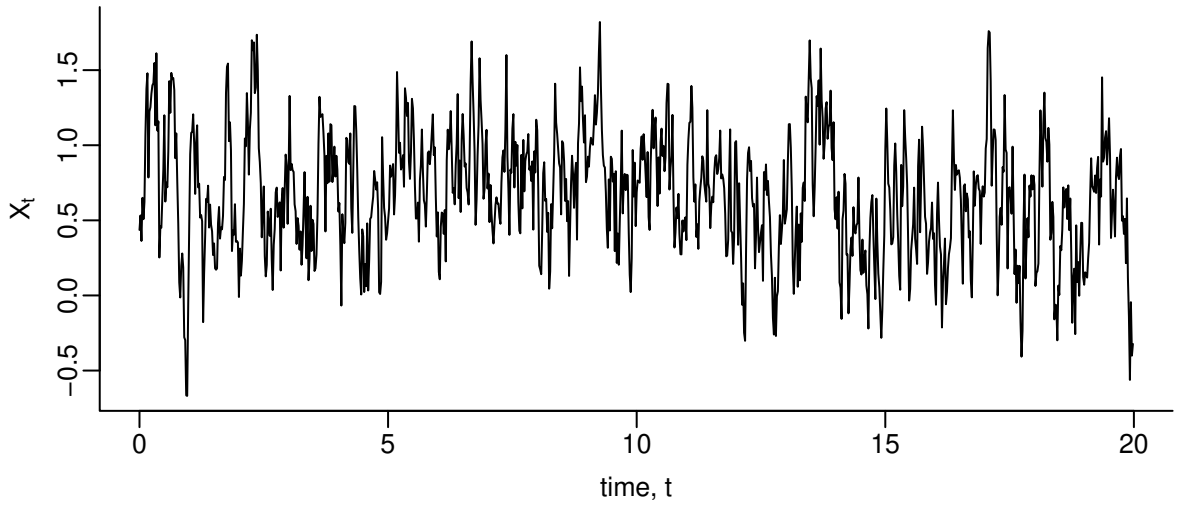
A simple example: The OU process

The **Ornstein-Uhlenbeck (OU) process** [Uhlenbeck and Ornstein, 1930] is

$$dX_t = (\theta_0 + \theta_1 X_t)\, dt + \theta_2\, dW_t, \quad 0 \le t \le T,$$

where $X_0 = x_0$ is the initial value, $\theta_0 \in \mathbb{R}, \theta_1 < 0$, $\theta_2 > 0$, and $\{W_t\}$ is a standard BM.

(Discrete) realization for $\theta_0 = 2$, $\theta_1 = -3$, $\theta_2 = 1$, with sampling interval 0.1:



Our inference problem

Suppose that we observe the process $\{X_t\}$ at time points t_i $(i = 1, \ldots, N)$ where $0 = t_0 < t_1 < \cdots < t_N$.

Based on the data vector $\mathbf{X} = (X_{t_1}, \ldots, X_{t_N})^T$ we want:

1. The **maximum likelihood estimator** of $\boldsymbol{\theta}$;

2. Associated **confidence bounds** for $\boldsymbol{\theta}$.

Defining the likelihood function

Let $p(x|x_{t_{i-1}}, \boldsymbol{\theta})$ represent the conditional probability density of X_{t_i} given $X_{t_{i-1}} = x_{t_{i-1}}$ evaluated at x for a given set of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

- This is the so-called **transition density**.

Treating $X_0 = x_0$ as fixed, we use the **Markov property** to write the **discretized likelihood** of the data as the product of these 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.

Approximating the transition density

Four choices [see Hurn et al., 2007, for a review]:

1. Closed-form Hermite expansions of the transition density [Aït-Sahalia, 2002, Aït-Sahalia, 2008]
2. Importance sampling [Pedersen, 1995, Santa-Clara, 1997, Brandt and Santa-Clara, 2002, Durham and Gallant, 2002]
3. Methods based on the exact simulation of diffusions [Beskos and Roberts, 2005, Beskos et al., 2006, 2008]
4. Approximations derived by numerically solving the Kolmogorov forward equation [Lo, 1988].

Approximating the transition density: Euler approximation

Remember our SDE is

$$dX_t = \mu(X_t, \boldsymbol{\theta})dt + \sigma(X_t, \boldsymbol{\theta})\, dW_t,$$

Using the fact that for a BM $\{W_t\}$,

$$W_\Delta - W_0 \sim N(0, \Delta),$$

the Euler approximation is

$$X_\Delta \approx X_0 + \mu(X_0, \boldsymbol{\theta})\Delta + \sigma(X_0, \boldsymbol{\theta})\sqrt{\Delta}Z,$$

where $Z \sim N(0, 1)$.

Question: How well does this approximation do in practice?

Approximating the transition density: Importance sampling

1. Euler approximation is better for smaller distances.

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

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

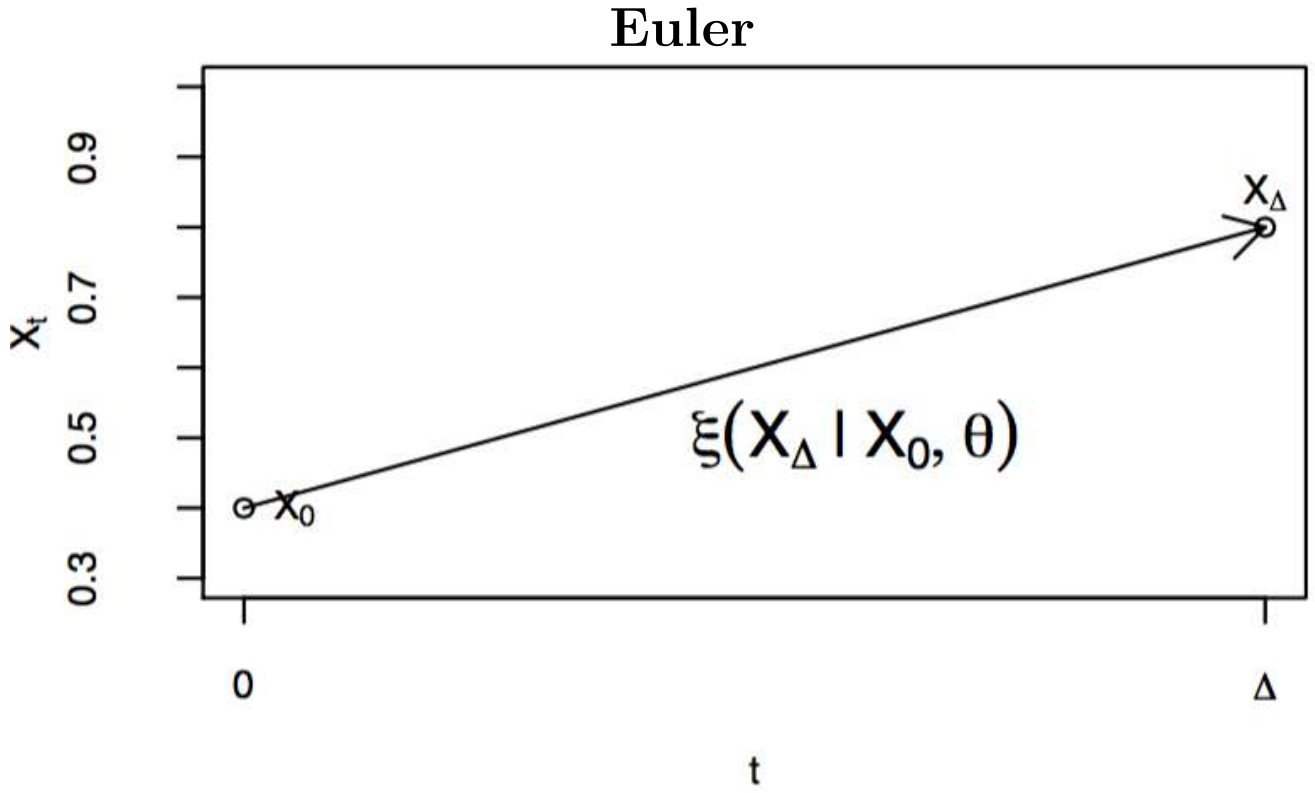
3. The **discretized transition density** [Kloeden and Platen, 1992] is

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

   where λ denotes the Lebesgue measure.

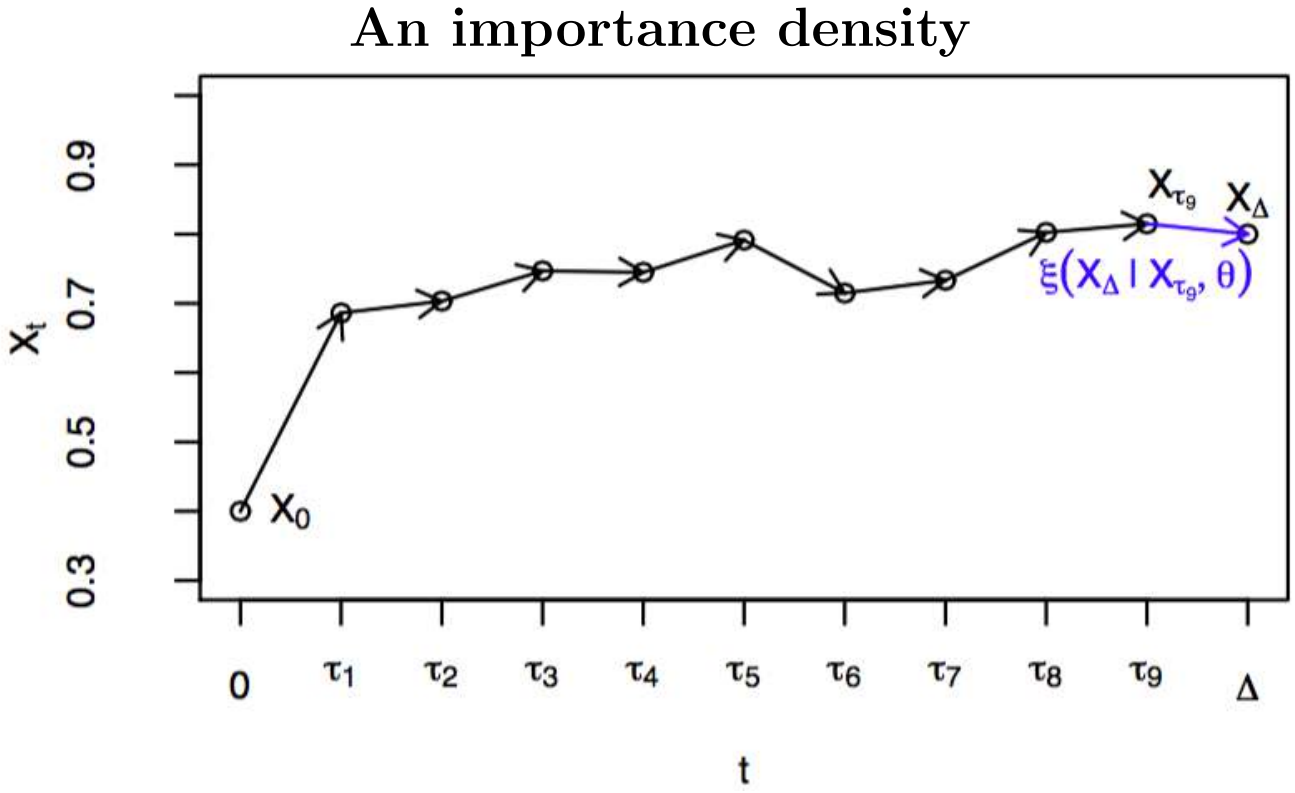
4. We approximate this density by **importance sampling**, using M random samples from an importance density $q(\cdot)$.

Approximating the transition density: Importance sampling



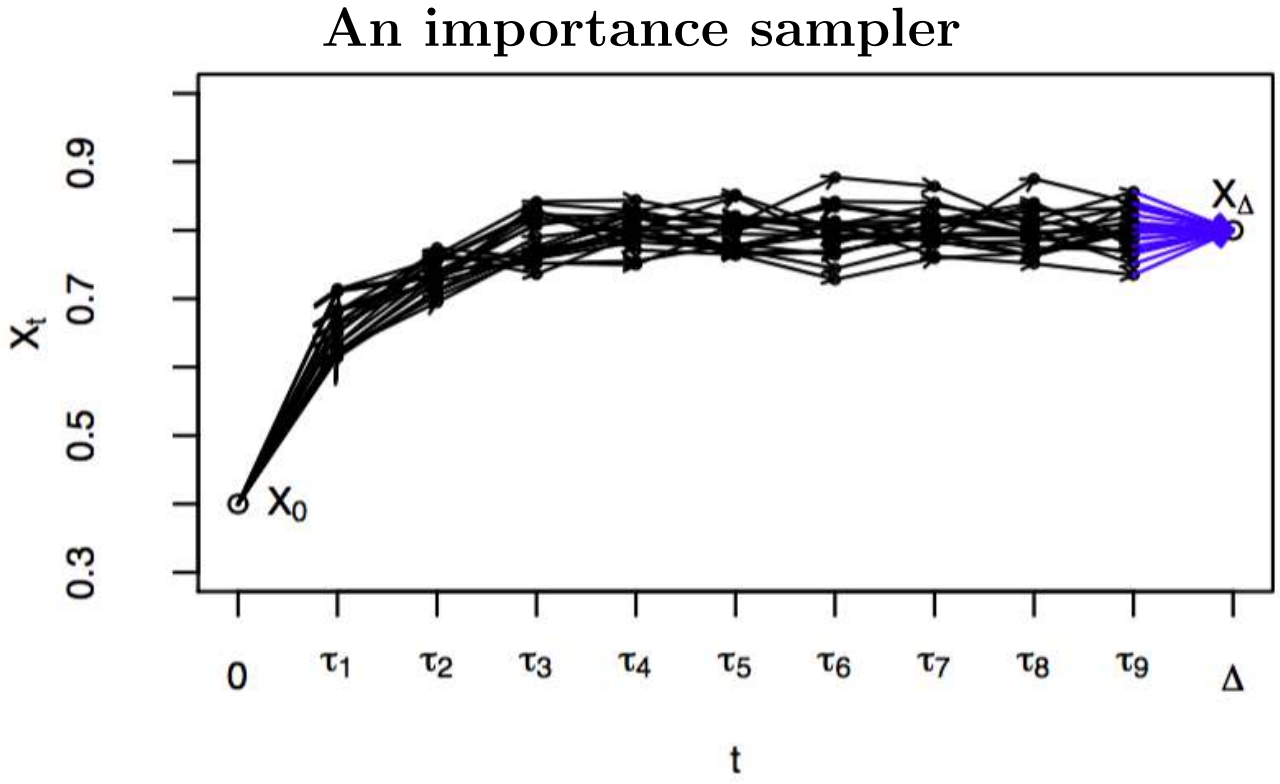
$$p^{(K)}\left(X_\Delta | X_0, \boldsymbol{\theta}\right) = \int \prod_{k=1}^{K} \xi\left(X_{\tau_k} | X_{\tau_{k-1}}, \boldsymbol{\theta}\right) \lambda(d\mathbf{X}_\tau)$$

Approximating the transition density: Importance sampling



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

Approximating the transition density: Importance sampling



with $q(\mathbf{X}_\tau) = \prod_{k=1}^{K-1} \xi\left(X_{\tau_k}|X_{\tau_{k-1}}, \boldsymbol{\theta}\right)$, $p^{(K,M)}(X_\Delta|X_0, \boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \xi\left(X_\Delta|X_{\tau_{K-1},m}, \boldsymbol{\theta}\right)$

[Pedersen, 1995, Brandt and Santa-Clara, 2002]

Approximating the transition density: Importance sampling

Elerian et al. [2001] criticized Pedersen [1995] for its inefficiency and proposed a computationally intensive method of sampling $\mathbf{X}_\tau$ from a multivariate normal or t distribution based on a second-order Taylor expansion.

Exact simulation [Beskos and Roberts, 2005, Beskos et al., 2006, 2008]: sample from $q(\cdot)$ that is exactly $p\left(\mathbf{X}_\tau | X_0, \boldsymbol{\theta}\right)$, but adds another layer of computational complexity [e.g., Bladt and Sørensen, 2014].

As a compromise between accuracy and computational efficiency, we use the **modified Brownian bridge sampler** [Durham and Gallant, 2002] for $q(\cdot)$.

For more discussion see Papaspiliopoulos and Roberts [2012].

Obtaining the MLE for the SDE process parameters

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

Work out the gradient and use a steepest-ascent approach?

1. Too much Monte Carlo variability in estimates of the log-likelihood.
2. Requires prohibitively large sample sizes, M. (We gave up considering this as a competitor after increasing computation time by a factor of around 1,000.)

Approximate the parameter space by some fine grid of points and get a log-likelihood estimate at each parameter value?

1. Takes a long time when $\boldsymbol{\theta}$ is low-dimensional.
2. Takes a prohibitively long time when the dimension is moderate or high.

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!

A Gaussian process (GP) model

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

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 $\mathrm{N}(0, \sigma^2)$ 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.

Remarks on the choice of GP model

Extensive literature on the choice of mean and covariance function for the GP [see, e.g., Cressie and Wikle, 2011].

- Richer choices can more accurately emulate the discretized log-likelihood, at a cost of necessitating larger sample sizes, n, and more computational resources to faithfully model the features of the GP.
- Properties of GP-based estimates of smooth functions are well known [e.g, Van Der Vaart and Van Zanten, 2011].

In computer experiments [e.g., Sacks et al., 1989, Santner et al., 2003], smooth functions are usually modeled using the stationary Gaussian covariance.

Usually simple polynomial models are assumed for the mean function [e.g., Santner et al., 2003].

Predicting the log-likelihood function

By Gaussianity of the GP and data model, the best linear unbiased predictor of $l^{(K)}(\boldsymbol{\theta}^*)$ given the data $\boldsymbol{Y}_n = (Y(\boldsymbol{\theta}_i) : i = 1, \ldots, n)^T$ has a Gaussian distribution 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 $C_{L,n}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j; \boldsymbol{\zeta})$.

Finding the maximum by evaluating the improvement

Now we are in a position to try to maximize this likelihood function wrt $\boldsymbol{\theta}$.

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

Then, the **improvement** [Jones et al., 1998] at $\boldsymbol{\theta}^*$ is

$$I\left(\boldsymbol{\theta}^*\right) = \max\left\{0, l^{(K)}\left(\boldsymbol{\theta}^*\right) - \widetilde{\eta}_{L,n}\right\}.$$

But $l^{(K)}(\boldsymbol{\theta}^*)$ is unknown ...

So we use the expected improvement

Replace the improvement by the **expected improvement** at parameter value $\boldsymbol{\theta}^*$, which can be shown to be equal to [Jones et al., 1998]

$$E(I(\boldsymbol{\theta}^*)\,|\boldsymbol{Y}_n) = [\eta_{L,n}(\boldsymbol{\theta}^*) - \widetilde{\eta}_{L,n}]\,\Phi\left(\frac{\eta_{L,n}(\boldsymbol{\theta}^*) - \widetilde{\eta}_{L,n}}{v_{L,n}(\boldsymbol{\theta}^*)}\right) + v_{L,n}(\boldsymbol{\theta}^*)\;\phi\left(\frac{\eta_{L,n}(\boldsymbol{\theta}^*) - \widetilde{\eta}_{L,n}}{v_{L,n}(\boldsymbol{\theta}^*)}\right),$$

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 Kriging-Based Optimization (SKBO)

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);

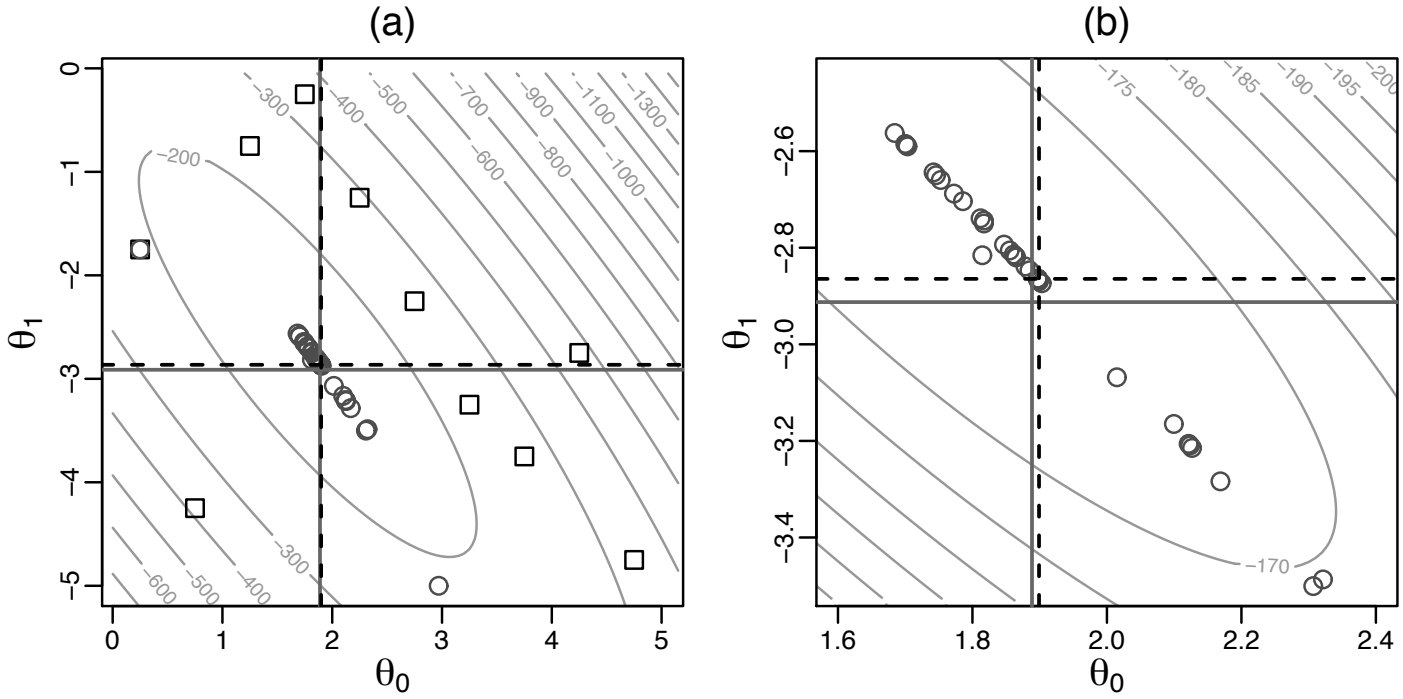
2. The conditional mean and variance for the BLUP.

Now repeat, until some stopping criteria is met (little change in the estimate of the MLE).

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 SKBO 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).

Confidence regions for $\boldsymbol{\theta}$

We obtain an approximate $(1-\alpha)\%$ joint confidence region for $\boldsymbol{\theta}$ directly from the conditional mean based on the likelihood ratio test:

$$\left\{\boldsymbol{\theta} : 2\left(\eta_{L,n}(\widehat{\boldsymbol{\theta}}) - \eta_{L,n}\left(\boldsymbol{\theta}\right)\right) \leq \chi^2_{p;1-\alpha}\right\},$$

where $\chi^2_{p;1-\alpha}$ is the 1-α quantile of a chi-square distribution with p degrees of freedom.

In practice we calculate this confidence interval by calculating the conditional mean on a dense grid of a parameter values in the neighborhood of the estimated MLE.

Simulations – OU model

Simulate 2000 datasets of $N = 1000$ observations from **OU model**:

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

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

For each dataset, compute

- True MLE, est. MLE via naive methods, and est. MLE via SKBO.

Vary K, M, and number of $\boldsymbol{\theta}$ values. Summarize:

1. Bias, SD and RMSE
2. Confidence region coverage %
3. Computation time

		MLE	$K = 5$, $M = 25$				$K = 10$, $M = 100$			
			Naive		SGBO		Naive		SGBO	
Initial pts		–	50	2500	10	20	50	2500	10	20
Avg added		–	–	–	12.0	7.7	–	–	9.7	6.1
	Bias	0.03	0.06	0.08	0.05	0.07	0.05	0.05	0.04	0.05
θ_0	SD	0.20	0.40	0.23	0.31	0.20	0.39	0.22	0.27	0.20
	RMSE	0.20	0.40	0.25	0.31	0.21	0.39	0.22	0.28	0.21
	Bias	0.05	0.09	0.11	0.07	0.10	0.08	0.08	0.06	0.07
θ_1	SD	0.25	0.51	0.31	0.36	0.26	0.50	0.28	0.35	0.25
	RMSE	0.25	0.52	0.33	0.36	0.28	0.50	0.29	0.35	0.26
Coverage %		97.0	97.9	84.9	89.3	94.2	98.7	93.4	91.1	95.6
Avg Times		–	0.6	27.6	0.6	0.7	4.9	231.0	2.1	2.8

Results at lower sample sizes

Repeating with $N = 100$, $N = 250$, and $N = 500$ we found that, as expected, all methods for estimating $\boldsymbol{\theta}$ (including the exact MLE) performed poorer than the $N = 1000$ case.

However, again, the SGBO method with $K = 10$, $M = 100$ and 40 initial points clearly outperformed all the other IS-based methods and was competitive with the exact MLE approach.

Simulations – GCIR model

Generalized Cox-Ingersoll-Ross (GCIR) model, introduced in Chan et al. [1992], and analyzed in Roberts and Stramer [2001]:

$$dX_t = (\theta_0 - \theta_1 X_t)\, dt + \gamma X_t^{\psi}\, dW_t, \quad 0 \le t \le T,$$

where $X_0 = x_0$ is the initial value of the process, $\theta_0, \theta_1 \in \mathbb{R}, \gamma > 0, \psi \in [0, 1]$, and W_t is a standard BM.

Does not have a closed-form likelihood, except for when $\psi = 0$ (the OU process) or $\psi = 0.5$ (the Cox-Ingersoll-Ross model).

We let $\theta_2 = \log(\gamma)$ and $\theta_3 = \log(\psi/(1-\psi))$, and optimize over the real-valued parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)$.

		$K = 5,\ M = 25$			$K = 10,\ M = 100$		
		Naive	SGBO		Naive	SGBO	
Initial pts		100	20	40	100	20	40
Avg added		–	60.3	47.0	-	56.2	46.3
	Bias	0.22	0.05	0.02	0.25	0.05	0.02
θ_0	SD	0.71	0.57	0.39	0.70	0.45	0.29
	RMSE	0.74	0.57	0.39	0.75	0.46	0.29
	Bias	-0.13	-0.07	-0.04	-0.16	-0.08	-0.03
θ_1	SD	0.92	0.65	0.48	0.91	0.54	0.29
	RMSE	0.93	0.65	0.48	0.92	0.55	0.30
	Bias	0.08	0.04	0.03	0.10	0.03	0.01
θ_2	SD	0.17	0.22	0.23	0.18	0.16	0.15
	RMSE	0.19	0.23	0.23	0.20	0.16	0.15
	Bias	-0.14	-0.01	0.01	-0.14	0.04	0.03
θ_3	SD	0.62	0.54	0.42	0.60	0.44	0.33
	RMSE	0.64	0.54	0.42	0.62	0.44	0.33
Avg Time		6.2	13.6	18.8	15.4	24.1	27.1

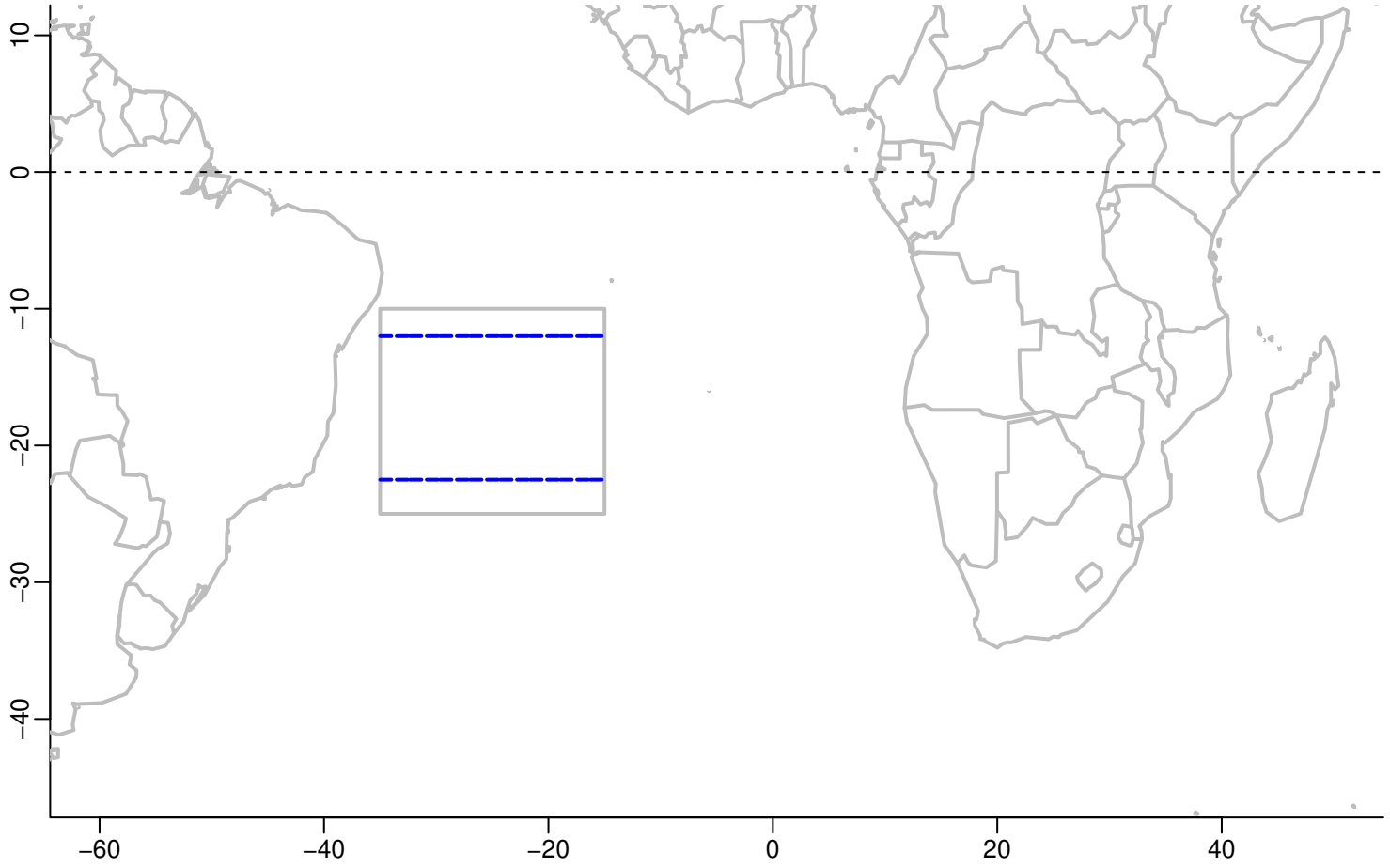
Estimating Deep Flow in the South Atlantic Ocean

Estimating the **state** of the world's oceans is a fundamental problem in modern day science.

Ocean circulation cannot be measured directly, but rather 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].

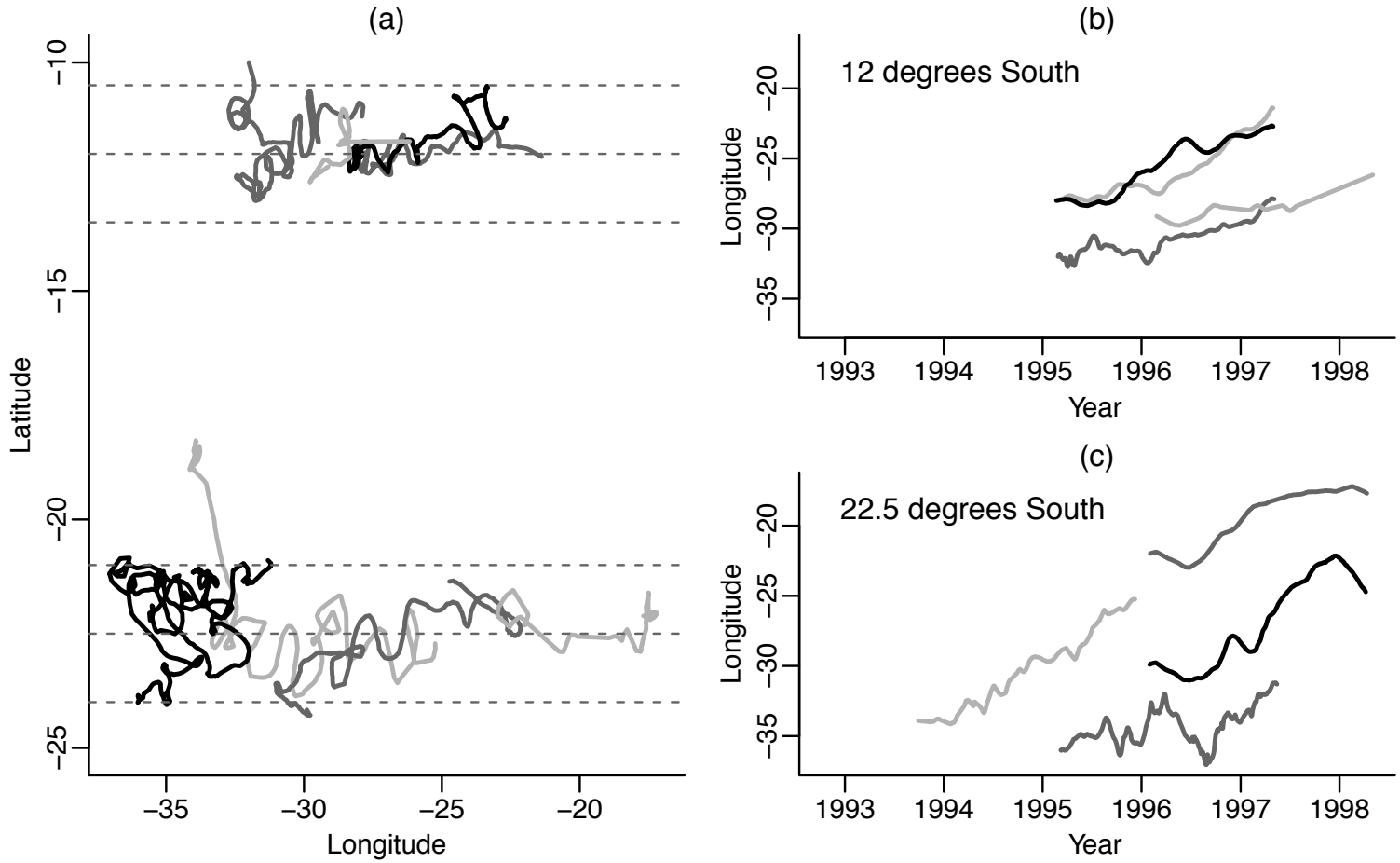
We focus on estimating **water velocities**.

Estimating Deep Flow in the South Atlantic Ocean, cont.



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.

Our data are available every two days.

Then assume $\{X_t^{(i)}\}$ satisfies the SDE

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

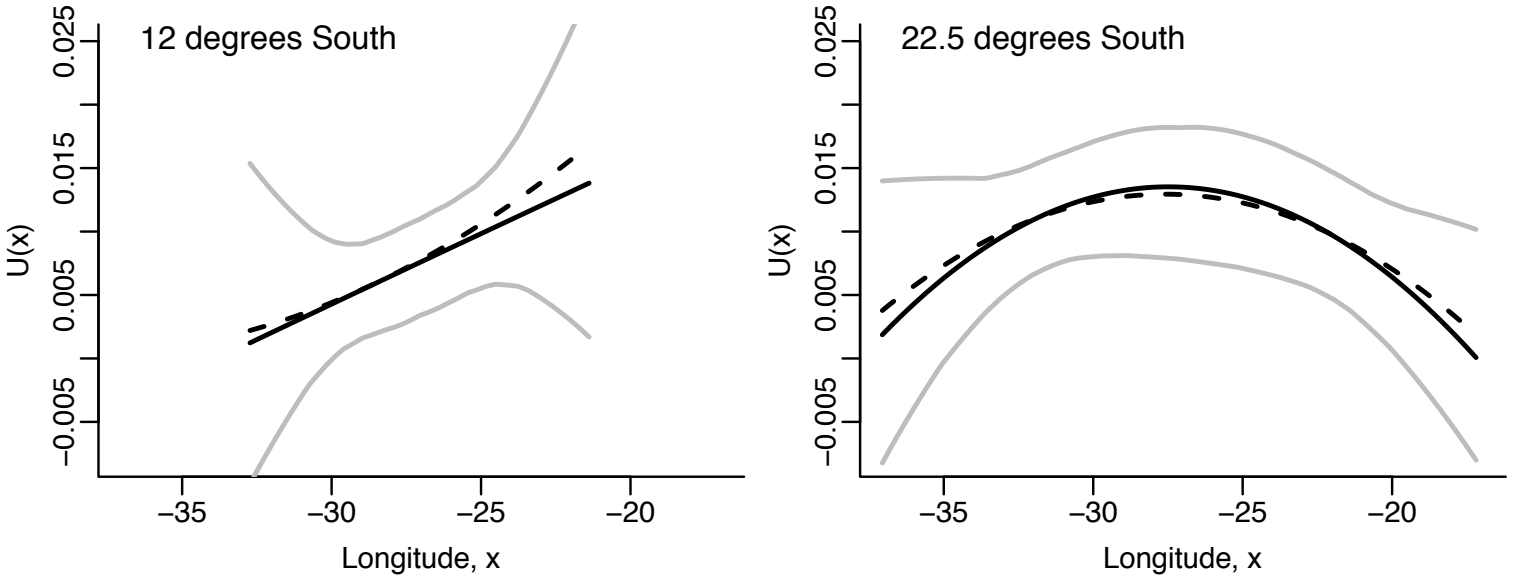
where $U(\cdot)$ is the **zonal velocity of interest** assumed common for each series in a given latitude band, σ is the diffusion coefficient, and $\{W_t\}$ is a standard BM.

Assuming conditional independence over i and σ is known, we estimate the drift function $U(\cdot)$.

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)

Estimating Deep Flow in the South Atlantic Ocean, cont.

This preliminary analysis provides motivation that the zonal velocity function varies by latitude.

Also demonstrates the utility of picking up more accurate velocity structures using the SGBO method, compared to naive Euler-based approximations.

Further research will investigate methods to estimate the parameters of a spatially varying SDE model that can estimate the **zonal and meridonal velocities jointly**1, using more floats deployed in the Atlantic Ocean.

Discussion and further work

We assume data are complete and exact discrete-time observations of an SDE; however, our method can be assumed to other settings.

We still want to consider calibration methods from the computer experiments literature [e.g., Kennedy and O'Hagan, 2001, Higdon et al., 2008] to try to minimize the bias introduced from the IS-estimate of the log-likelihood through an adaptive selection of the values of K and M in the SGBO algorithm.

- Another solution: lognormal kriging [e.g. Cressie and Pavlicová, 2005].

Also considering other ways to ameliorate the under-coverage of confidence regions.

Currently working on hierarchical and spatio-temporal SDEs.

References

Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70:223–262, 2002.

Y. Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36:906–937, 2008.

A. Beskos and G. O. Roberts. Exact simulation of diffusions. *The Annals of Applied Probability*, 15:2422–2444, 2005.

A. Beskos, O. Papaspiliopoulos, and G. O. Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12: 1077–1098, 2006.

A. Beskos, O. Papaspiliopoulos, and G. O. Roberts. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10:85–104, 2008.

M. Bladt and M. Sørensen. Simple simulation of diffusion bridges with application to likelihood inference for diffusions. *Bernoulli*, 20:645–675, 2014.

M. W. Brandt and P. Santa-Clara. Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *Journal of Financial Economics*, 63:161–210, 2002.

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.

N. Cressie and M. Pavlicová. Lognormal kriging: bias adjustment and kriging variances. In *Geostatistics Banff 2004*, pages 1027–1036. Springer, New York, NY, 2005.

N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, New York, 2011.

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.

O. Elerian, S. Chib, and N. Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69:959–993, 2001.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103:570–583, 2008.

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.

A. S. Hurn, J. I. Jeisman, and K. A. Lindsay. Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics*, 5:390–455, 2007.

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.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63:425–464, 2001.

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

A. W. Lo. Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory*, 4:231–247, 1988.

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.

O. Papaspiliopoulos and G. Roberts. Importance sampling techniques for estimation of diffusion models. *Statistical methods for stochastic differential equations*, 124:311–340, 2012.

A. R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22:55–71, 1995.

G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621, 2001.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–423, 1989.

P. Santa-Clara. Simulated likelihood estimation of diffusions with an application to the short term interest rate. Technical report, Anderson Graduate School of Management, UCLA, 1997.

T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer, New York, NY, 2003.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Physical review*, 36:823–841, 1930.

A. Van Der Vaart and H. Van Zanten. Information rates of nonparametric Gaussian process methods. *The Journal of Machine Learning Research*, 12:2095–2119, 2011.

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

The exact OU likelihood

The **OU process** is

$$dX_t = (\theta_0 + \theta_1 X_t)\, dt + \theta_2\, dW_t, \quad 0 \le t \le T,$$

where $X_0 = x_0$ is the initial value, $\theta_0 \in \mathbb{R}, \theta_1 < 0$, $\theta_2 > 0$, and $\{W_t\}$ is a standard BM.

The stationary distribution is

$$\phi\left(X_0;\ \theta_0/\theta_1,\ 1/(2\theta_1)\right).$$

The conditional distribution is

$$p(X_\Delta | X_0, \boldsymbol{\theta}) = \phi\left(X_\Delta; X_0 e^{-\theta_1(1-\Delta)} + \frac{\theta_0}{\theta_1}\left[1 - e^{-\theta_1(1-\Delta)}\right], \frac{1}{2\theta_1}\left[1 - e^{-2\theta_1(1-\Delta)}\right]\right).$$