Surrogate Active Subspaces for Jump-Discontinuous Functions

Nathan Wycoff
The McCourt School's Massive Data Institute, Georgetown University, Washington, D.C.

Abstract

Surrogate modeling and active subspaces have emerged as powerful paradigms in computational science and engineering. Porting such techniques to computational models in the social sciences brings into sharp relief their limitations in dealing with discontinuous simulators, such as Agent-Based Models, which have discrete outputs. Nevertheless, prior applied work has shown that surrogate estimates of active subspaces for such estimators can yield interesting results. But given that active subspaces are defined by way of gradients, it is not clear what quantity is being estimated when this methodology is applied to a discontinuous simulator. We begin this article by showing some pathologies that can arise when conducting such an analysis. This motivates an extension of active subspaces to discontinuous functions, clarifying what is actually being estimated in such analyses. We also conduct numerical experiments on synthetic test functions to compare Gaussian process estimates of active subspaces on continuous and discontinuous functions. Finally, we deploy our methodology on Flee, an agent-based model of refugee movement, yielding novel insights into which parameters of the simulation are most important across 8 displacement crises in Africa and the Middle East.

1 INTRODUCTION

In most fields of science and engineering, cutting edge mathematical models of phenomena are not susceptible to closed-form mathematical analysis, and must instead be studied via numerical simulation. One approach to conducting such a study is via a *sensitivity analysis*, which aims to determine what input parameters, or combinations thereof, the output of a simulator is most influenced by. In this article, we will be concerned with the Active Subspace Method [Constantine, 2015] (ASM, see Section 3.3), a form of global sensitivity analysis based on analyzing the gradient of the target function. When the computer simulation is computationally expensive, a common approach to studying it is to fit a *surrogate model*, that is, to estimate a flexible statistical model to sampled input-output pairs. Furthermore, some early work has explored porting these tools to simulators of social scientific phenomena, such as Agent-Based Models (ABMs). But ABMs represent the sum of discrete choices made by individuals, and so are inherently discontinuous. Nevertheless, nothing prevents an analyst from fitting a smooth surrogate model to a discontinuous simulator and calculating the surrogate's active subspace. For instance, [Notestine, 2022] computes various surrogate estimates of the active subspace of an ABM of social unrest, and shows that useful conclusions can be drawn from such an analysis. We also find promising results in our case study on an ABM of forced displacement, where a surrogate active subspace analysis offers novel conclusions and improves predictive accuracy. But it's not clear what is actually being estimated, since an ABM is not differentiable and is almost everywhere constant, so the "true" active subspace is undefined or $\mathbf{0}$. We might hope that another sensitivity analytic framework, like Sufficient Dimension Reduction (see Section 3.4), can tell us what's going on. But our Corollary 1 shows that this is not the case. In this article, we develop an extension of active subspaces to certain functions with jump-discontinuities. We find that asymptotically, a surrogate active subspace analysis will favor discontinuous directions of variation over continuous ones, and numerically find that in finite samples, the sample size implicitly parameterizes a trade-off between continuous and discontinuous directions of variation.

Though there seems to be significant demand for surrogate modeling of discontinuous simulators as evidenced

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

by the plethora of applied articles expounding their usefulness (see Section 3.2), the surrogate methodologist's conception of a "black-box" is overwhelmingly a continuous one. This article aims to play some small part in filling this methodological gap by making the following contributions:

1. In Section 2, we show that the use of surrogate active subspaces on simulators with jump discontinuities can lead to unexpected pathologies.
2. We develop an extension of active subspaces to discontinuous simulators to explain the observed pathologies and provide a theoretical basis for surrogate active subspace analysis of discontinuous simulators in Section 4.
3. Section 5 studies the fitness of various Gaussian process kernels for estimating the active subspace of discontinuous functions, finding rougher ones to be best.
4. In our case study of Section 6, we show that surrogate active subspace analysis can lead to quantitative results superior to dimension reduction which avoids gradients altogether and to meaningful qualitative insights.

Additionally to our main contributions, the pathologies we reveal in surrogate discontinuous active subspace analysis open the door for significant future work which we briefly overview in Section 7.

2 MOTIVATION

We begin this section with an overview of the *Flee* ABM, the application that motivated this research. Subsequently, we present two distressing observations about the empirical behavior of surrogate sensitivity analysis on discontinuous functions.

2.1 The *Flee* Simulator

The 21st century may well experience unprecedented migration due to the changing climate, both environmental ([Wrathall et al., 2019]) and political (including the ongoing mass-displacement events in Ukraine and Gaza), which motivates the study of human migration. The Flee simulator[1] [Suleimenova et al., 2017] is an ABM of the journeys of forcibly displaced persons. Agents are displaced over time and move from populated areas to refugee camps and neighboring countries according to simulation parameters, of which we study 7 in this article (see Table 1). Beyond these parameters, *Flee* also requires a spatial context in which to simulate movement. We study six crises resulting in forced displacement provided by *Flee*, namely the South Sudanese civil war in 2014, the civil war in the Central African Republic in 2013, the 2012 Malian *coup d'état*, the 2013 escalation of Syria's civil war, the Ethiopian civil war of 2020 and the 2015 civil unrest in Burundi. In each case, the simulator compares its estimates of displacement with ground truth data and provides a scalar error estimate. We wish to study the sensitivity of this error with respect to the model parameters for each context individually. Figure 1 shows the prediction error of Flee in the Syria case study with fixed seed[2]; the discontinuities are prominent.

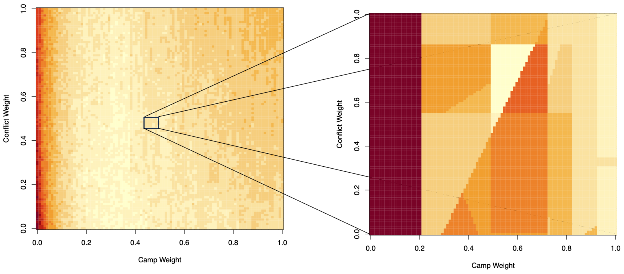

Figure 1: The cost surface of the Flee simulator for fixed seed as two parameters are changed and the others remain fixed.

Parameter	Min	Max	Default
max_move_speed	0.0	40000	200
max_walk_speed	0.0	40000	35
camp_move_chance	0.0	1.0	0.0
conflict_move_chance	0.0	1.0	1.0
default_move_chance	0.0	1.0	0.3
camp_weight	1.0	10.0	2.0
conflict_weight	0.1	1.0	0.2

Table 1: *Flee* model parameters.

2.2 Divergence of the Classical Active Subspace Estimate

Though the notion of an active subspace is not well defined for functions which are not differentiable such as ABMs, we might hope that fitting an almost-everywhere-continuous surrogate would lead to a reasonable estimate. In this section, we consider a 1 dimensional test function given by the heaviside step function centered at 0.5. We interpolate this function at an evenly spaced grid of an $n_g = 2k$ points by simply drawing a line between subsequent observations (Fig-

[1] Flee is licensed under BSD-3.

[2] As of writing, the seed cannot be set in Flee without significantly changing the nature of the simulation (i.e. all agents behaving identically); this figure is for illustrative purposes only.

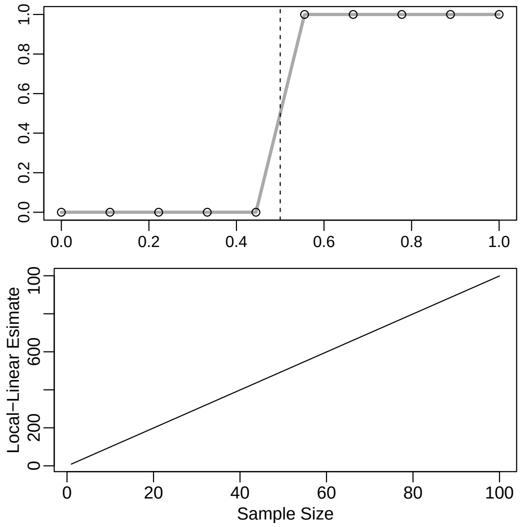

Figure 2: *Top:* Piecewise linear interpolant of the heaviside step function. *Bottom:* The piecewise linear estimate of active subspace, which diverges.

ure 2, top). All line segments are of slope zero except for the center-most segment, which has slope $n-1$, and extent $\frac{1}{n-1}$. Hence, if s_i gives the slope of the ith line segment, the active subspace of the surrogate is given by (see Section 3.3) $\frac{1}{n_g-1}\sum_{i=1}^{n_g-1} s_i^2 = \frac{1}{n_g-1}(n_g-1)^2 = n_g - 1$. We thus see that the active subspace estimate diverges as $n_g \to \infty$ (Figure 2, bottom). It is simple in this case to normalize by n to avoid divergence, and in any case, the scaling of the sensitivity analysis is not important. However, this divergence is indicative of a deeper issue which can lead to unexpected outcomes, as we discuss next.

2.3 Contradictory Sensitivity Analyses in a Mixed Simulator

We now consider the two dimensional function $f(\mathbf{x}) = \mathbb{1}_{[x_1 \geq 0.5]} + 6(x_2 - 0.5)^2$, which varies smoothly along x_2 but has a jump along x_1 (see Figure 3, top). We draw N random points in the unit square and use these to compute a Gaussian process surrogate estimate of the diagonal elements of the active subspace matrix normalized to have norm 1, which are indicators of variable importance (see Section 3.2). For $N \leq 30$, we see that the analysis consistently reports that x_2, the smooth variable is more important (Figure 3, bottom) than x_1. However, for $N \geq 40$, this is reversed. As the design points are placed closer together, the sensitivity estimate in the smooth direction stabilizes, while that in the discontinuous direction diverges. Section 4 develops theory explaining this phenomenon, but we first catch up on the needed methodological background.

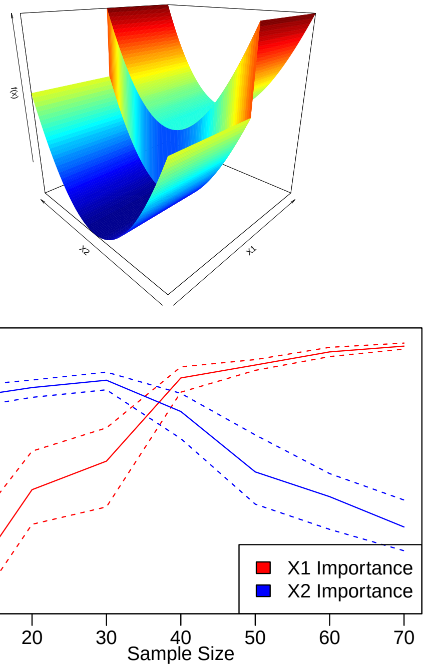

Figure 3: *Top:* A function with a discontinuity along the x_1 direction and a smooth quadratic form along the x_2 direction. *Bottom:* As the sample size increases, the expected importance of the discontinuous direction overtakes the continuous one.

3 BACKGROUND

We review surrogate modeling of discontinuous simulators and some concepts from linear sensitivity analysis.

3.1 Surrogate Modeling of Computer Experiments

The practice of *Surrogate Modeling* [Conti et al., 2009, Gramacy, 2020] corresponds to the use of flexible statistical models to approximate parameterized computer simulations, conceptualized as input-output maps. In this article, we will be interested in studying surrogates of a black-box function f mapping $\mathcal{X} \subseteq \mathbb{R}^P \to \mathbb{R}$. Some of our technical results rely on $\mathcal{X}$ being compact, and in the numerical studies it will be the unit hypercube $[0,1]^P$. Conceptually, a statistical surrogate can be any regression model. In practice, commonly used surrogates include polynomials, Gaussian processes [Rasmussen and Williams, 2005] and other nonparametric models.

3.2 Surrogacy for Discontinuous Simulators

Sometimes, there's an important and well understood discontinuity that we'd like our surrogate model to preserve. For instance, in aerodynamics, the transsonic barrier leads to completely different dynamics on one side from the other. [Dupuis et al., 2018] use different surrogates for subsonic and various supersonic operating conditions to accurately capture the jump as well as faithfully approximate the truth on either side. In the case of ABMs, this is due to their being a model of a sum of discrete choices. In other circumstances, the discontinuity can be a nuisance caused not by an underlying natural phenomenon of interest but rather due to numerical noise or nonconvergence of the simulation. In such cases, the hope is rather that the surrogate will paper over the inadequacies of the model. Take [Huang et al., 2020], whose simulator exhibited discontinuous jumps related to tolerance parameters. Between these two extremes lie a number of other possible situations. In the domain of Structural Optimization for Crashworthiness, the simulator studied by [Niutta et al., 2018] has important discontinuities, but the nature and number of them is not known *a priori*. The authors estimate the number and location of discontinuities using a combined surrogate approach. [Gorodetsky and Marzouk, 2014] propose methodology for estimating the location at which a jump occurs by examining a polynomial interpolant of the function. [Audet et al., 2022] propose an approach for Black-Box optimization under the constraint that the optimum cannot lie near the unknown locations of discontinuity.

We see that some authors choose to model piecewise-discontinuous functions with similarly piecewise-discontinuous surrogates, while others use a global smooth surrogate. In this article, we will study the latter approach, finding both positive and negative results. Some pathologies of fitting continuous interpolants to discontinuous functions have been long known, such as the tendency for Fourier approximations to oscillate when approximating discontinuous functions, which is known as the Gibbs Phenomenon [Arfken and Weber, 1972, Chapter 14.5]. Indeed, this is true of any global smooth approximant to a discontinuous function [Butzer et al., 1987].

3.3 Gradient-Based Global Sensitivity Analysis

For a smooth function f, the gradient $\nabla f(\mathbf{x})$ is a natural way of quantifying the sensitivity of an output to an input. One strategy for turning this local estimate of sensitivity into a global one is to integrate it over the parameter space: $\mathbf{C}_{f,\mu} = \int_{\mathcal{X}} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top d\mu(\mathbf{x})$. Here, μ is a probability measure on the input space. In the computational engineering literature, this often goes by the name Active Subspace Method (ASM) [Constantine, 2015], and such expected gradient outer products have also been called Average Derivative Functionals in the observational context [Samarov, 1993]. Examination of the eigendecomposition of $\mathbf{C}_{f,\mu}$ gives important linear combinations of inputs, which may be viewed as running a principal components analysis on randomly sampled gradients of the target function. When the gradient of the target function is available, a Monte-Carlo estimate is straightforward to form [Constantine and Gleich, 2014]. Otherwise, the strategy of computing the active subspace of a surrogate model can be deployed [Palar and Shimoyama, 2018]. In the observational context, [Fukumizu and Leng, 2014] produce a kernel estimate of the active subspace with respect to the empirical measure of a given sample. [Wycoff et al., 2021] showed that if the surrogate model is a Gaussian process with certain kernel functions, the active subspace is available in closed form.

3.4 Sufficient Dimension Reduction

Another perspective on linear dimension reduction is that of Sufficient Dimension Reduction (SDR). Given a measure μ on $\mathbf{x}$, we say that $\mathcal{U}$ is a sufficient reduction if $P(y|\mathbf{x}) = P(y|\mathbf{U}\mathbf{x})$ [Cook, 1994], where $\mathbf{U}$ is a matrix with range $\mathcal{U}$ (though there are also slightly different definitions and related concepts [Adragni and Cook, 2009]). See [Ma and Zhu, 2013] for a review. One approach to estimating sufficient reductions is Sliced Inverse Regression [Li, 1991], which splits the response y into bins before taking the mean $\mathbf{x}$ value in each bin and performing PCA on the resulting matrix.

3.5 Ridge Functions

Related to both active subspaces and SDR is the concept of a *ridge function* [Logan and Shepp, 1975], that is, a function $f : \mathbb{R}^P \to \mathbb{R}$ which takes $\mathbf{x} \to g(\mathbf{A}\mathbf{x})$ with $\mathbf{A} \in \mathbb{R}^{R \times P}$ for $R < P$ and $g : \mathbb{R}^R \to \mathbb{R}$. Like SDR, it encapsulates the idea of relationships which depend *solely* on certain input dimensions. By comparison, the concept of ASM is fuzzier, allowing *some* variation in all directions, but focusing it along certain ones.

4 AN EXTENSION OF THE ACTIVE SUBSPACE

In this section, we'll develop an extension of Active Subspaces to discontinuous functions in order to explain what is being estimated by the active subspace of a surrogate model fit to a discontinuous simulator. The proofs of all results are given in the Supplementary Materials. Throughout this section, $\|.\|$ will refer to the Euclidean norm, C^1 represents the function space of functions once differentiable on $\mathcal{X}$, $\mathcal{B}_r$ is the ℓ_2 ball of radius r, $\Gamma(x)$ refers to the special function, and μ is a probability measure on $\mathcal{X}$ and for some results it will be assumed to have a continuously differentiable Lebesgue density δ [3]. We will consider simulators abstracted mathematically as functions given by the sum of characteristic functions for sets parameterized by a differentiable function together with a smooth term, that is, $f(\mathbf{x}) = \sum_{j=1}^{J} c_j \mathbb{1}_{[\mathbf{x}\in\mathcal{S}_j]} + g(\mathbf{x})$ where $\mathbb{1}_{[\mathbf{x}\in\mathcal{A}]}$ is the function taking value 1 if $\mathbf{x} \in \mathcal{A}$ and zero otherwise, $\mathcal{S}_j = \{\mathbf{x} \in \mathcal{X} : h_j(\mathbf{x}) \leq 0\}$ [4] where $h_j \in C^1$ for all j, and $g \in C^1$.

Our extension will be built on a continuous analog to a regression coefficient, intuitively given by the limit of the OLS estimate based on sampling points uniformly within a radius r of a given point $\mathbf{x}$ as the sample size tends to infinity.

Definition 1. $\beta_r(\mathbf{x}) = \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}\mathbf{z}^\top]^{-1} \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}f(\mathbf{x}+\mathbf{z})]$.

To work with $\beta^r(\mathbf{x})$, we will need the following elementary results (see Supplementary Material), where the Gamma function arises from the volume of the P-ball:

1. $\int_{\mathbf{z}\in\mathcal{B}_r^P} z_i^2 d\mathbf{z} = \frac{\pi^{\frac{P}{2}} r^{P+2}}{2\Gamma(\frac{P+4}{2})} := \xi_P r^{P+2}$
2. $\mathbb{E}_{\mathbf{z}\in\mathcal{B}_r^P}[z_i^2] = \frac{r^2}{P+2}$

This leads to the following result.

Lemma 1. *If f consists only of a smooth term g we have that $\lim_{r\to 0}\beta_r(\mathbf{x}) = \nabla f(\mathbf{x})$. Otherwise, we have that:*

$$\lim_{r\to 0} r\beta_r(\mathbf{x}) = \begin{cases} A_P \sum_{\{j:\mathbf{x}\in\partial\mathcal{S}_j\}} c_j \frac{\nabla h_j(\mathbf{x})}{\|\nabla h_j(\mathbf{x})\|_2} & \mathbf{x} \in \cup_j \partial\mathcal{S}_j \\ 0 & o.w. \end{cases} \tag{1}$$

Lemma 1 tells us that $\beta^r(\mathbf{x})$ may be viewed as an extension of the gradient to possibly discontinuous functions. We next define an integral of this quantity, analogous to the expected outer product of the gradient in the smooth case:

Definition 2. $\mathbf{B}^r_{f,\mu} = \mathbb{E}_{\mathbf{x}\sim\mu}[\beta_r(\mathbf{x})\beta_r(\mathbf{x})^\top]$.

When the function is indeed smooth, we can recover the classical active subspace by taking $\mathbf{B}^r_{f,\mu}$'s limit.

Theorem 1. *If f is once differentiable (i.e. $f = g$) then $\lim_{r\to 0}\mathbf{B}^r_{f,\mu} = \mathbf{C}_{f,\mu}$.*

However, when the function is not smooth, that limit does not exist. We define our extension of active subspaces to nonsmooth functions as follows:

Definition 3. $\mathbf{B}_{f,\mu} = \lim_{r\to 0} r\mathbf{B}^r_{f,\mu}$.

Next we investigate some properties of $\mathbf{B}_{f,\mu}$ which apply in the general case where f is discontinuous.

Lemma 2. *If $f(\mathbf{x})$ is constant along dimension $\mathbf{u}$, and μ is translation-invariant along $\mathbf{u}$, then $\mathbf{u}^\top\beta^r(\mathbf{x}) = 0$.*

This lemma shows us that even for finite r, the gradient analogue $\beta^r(\mathbf{x})$ will always point in directions which the target function vary in. It leads to the below theorem.

Theorem 2. *If $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ with $\mathbf{A} \in \mathbb{R}^{R\times P}$ and $g : \mathbb{R}^R \to \mathbb{R}$, Range$(\mathbf{B}^r_{f,\mu}) \subseteq$ Range$(\mathbf{A})$.*

This theorem represents our first positive result, showing that ridge functions, including discontinuous ones, have their ridge structure respected by the extended active subspace. But the active subspace on continuous functions can tell us more than simply whether or not a given function has ridge structure or not; it also gives us the relative importance of different directions defined in a sum of squares sense. We now turn to investigating analogous properties of our proposed extension, starting in one dimension to build intuition.

Lemma 3. *If $f : \mathbb{R} \to \mathbb{R}$ is a linear combination of translated heaviside functions, that is $f(x) = \sum_{j=1}^{J} c_j \mathbb{1}_{[x\leq\tau_j]}$, and μ is Lebesgue-continuous with differentiable density δ, then $\lim_{r\to 0} r\mathbf{B}^r_{f,\mu} = \frac{3}{5}\sum_{j=1}^{J} c_j^2 \delta(\tau_j)$.*

This lemma shows us that, when properly normalized, the active subspace extension gives the sum of squared jumps of a discontinuous function, weighted by the density of the measure with respect to which it is defined. The following theorem extends this understanding to P dimensions.

Theorem 3. *Denoting by $B_P = \frac{\Gamma(\frac{P}{2}+1)^2\Gamma(P+2)}{\sqrt{\pi}\Gamma(P+\frac{5}{2})\Gamma(\frac{P+3}{2})^2}$ if the sets $\mathcal{S}_j$ are disjoint [5], then we hve $\frac{1}{B_P}\mathbf{B}_{f,\mu} = \sum_{j=1}^{J} c_j^2 \big[\int_{h_j^{-1}(0)} \frac{1}{\|\nabla h_j(\mathbf{x})\|_2^2} \nabla h_j(\mathbf{x})\nabla h_j(\mathbf{x})^\top \delta(\mathbf{x}) dS_j(\mathbf{x})\big]$ where S_j is the surface measure on $\mathcal{S}_j$.*

Intuitively, the active subspace extension is given by a weighted sum of an active subspace analogue of the functions parameterizing the jump points, weighted by the squared size of the jump and with a degenerate measure confined to the null-set of h_j. However, unlike the standard active subspace definition, note that the expression $\frac{1}{\|\nabla h_j(\mathbf{x})\|_2^2}\nabla h_j(\mathbf{x})\nabla h_j(\mathbf{x})^\top$ is invariant

[3] not to be confused with the Dirac delta function, which does not appear in this article

[4] Our results hold for $\mathcal{S}_j$ defined either by strict or nonstrict inequality, leading to either open or closed sets. For notational simplicity, we use closed sets throughout.

[5] The expression without the disjointness assumption is also given in the proof of this theorem in the Supplementary Material.

to smooth monotonic transformation to any h_j, which is necessary given that this kind of transformation will have no effect on f. The following is an immediate consequence of the fact that the expression in the preceding theorem does not depend on g, and is our main negative result.

Corollary 1. *For f with both smooth and discontinuous components, the range of $\mathbf{B}_{f,\mu}$ does not necessarily contain the SDR space.*

This tells us that the extended active subspace ignores smooth directions of hybrid smooth-discontinuous functions. It helps to explain the contradictory behavior we observed when estimating a surrogate's active subspace fit to a discontinuous function in Section 2.3.

We now study the limiting behavior of the active subspace of a Nadaraya–Watson kernel regression [Simonoff, 2012] with vanishing kernel bandwidth and a large sample.

Theorem 4. *Let $m(\mathbf{x})$ denote the Nadaraya–Watson kernel estimate using kernel $k(\mathbf{x})1, \mathbf{x}_2) = d(\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2}{\sqrt{l}})$ for decreasing scalar function d. Under regularity conditions enumerated in the supplementary material*

$$\lim_{l \to 0} \lim_{n \to \infty} \sqrt{l} \int_{\mathbf{x} \in [0,1]^P} \nabla m(\mathbf{x}) \nabla m(\mathbf{x})^\top d\mu = C\mathbf{B}_{f,\mu} \quad (2)$$

where C is a constant depending only on k and P.

We suspect that a Gaussian process would exhibit similar behavior.

Conjecture 1. *For suitable kernel functions, the limiting posterior active subspace*

$$\lim_{l \to 0} \lim_{n \to \infty} \sqrt{l} \int_{\mathbf{x} \in [0,1]^P} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top P(f|\mathbf{X}_n, \mathbf{y}_n) d\mu \quad (3)$$

is given by $C\mathbf{B}_{f,\mu}$, where $P(f|\mathbf{X}_n, \mathbf{y}_n)$ is a Gaussian process posterior and C is a constant depending only on the kernel and P.

5 NUMERICAL STUDY OF KERNEL ESTIMATES

We now study the capability of Gaussian Process surrogates to estimate the active subspace of ridge functions with direction $\mathbf{u}$ on two smooth functions $f_1(\mathbf{x}) = (\mathbf{u}^\top \mathbf{x})^2$ and $f_2(\mathbf{x}) = e^{-(\mathbf{u}^\top \mathbf{x})^2}$ and two discontinuous functions $f_3(\mathbf{x}) = \mathbb{1}_{[\mathbf{u}^\top (\mathbf{x} - 0.5\mathbf{1}) \geq 0]}$ and $f_4(\mathbf{x}) = \mathbb{1}_{[\sin\left(\frac{10\pi}{P} \mathbf{u}^\top (\mathbf{x} - \frac{1}{2})\right) \geq 0]}$ (visualized in 2D in Figure 4, top). For each function, with samples sizes $N \in \{50, 100, 150, 200\}$ and in dimensions $P \in \{3, 5, 7\}$, we fit a Gaussian process with Gaussian, Matérn $\frac{5}{2}$, or Matérn $\frac{3}{2}$ kernels, which lead to infinitely differentiable, twice differentiable, or once differentiable surrogates, respectively [Williams and Rasmussen, 2006, Chapter 4] using the defaults of `hetGP` package's `mleHomGP`. We compute their active subspace using the `R` package `activegp`. Then, we measure the cosine of the angle between $\mathbf{u}$ and the leading eigenvector of the estimated active subspace matrix, which serves as our error measure. We repeat the experiment 30 times, sampling $\mathbf{u}$ uniformly at random on the unit P-sphere.

Figure 4 shows the results in dimension 7 (the others are qualitatively similar and in the Supplementary Material). We see that on the smooth functions, the Gaussian and Matérn $\frac{5}{2}$ kernels are better able to exploit smoothness which leads to better subspace estimates. Conversely, when the function is nonsmooth, the rougher Matérn $\frac{3}{2}$ kernel dominates in terms of error. Though, strictly speaking, the Matérn covariance is still "wrong" insofar as the true simulator is not continuous whereas the surrogate is continuously differentiable, it seems that its discontinuous higher order derivatives still allow it to do a better job matching the active subspace than smoother kernels.

6 *Flee* CASE STUDY

In this section we deploy active subspaces to the *Flee* ABM (see section 2.1). We generated a sample of 500 randomly distributed points within the parameter ranges for each of the six case studies and evaluated *Flee* at each of the design points. We calculated active subspace estimates using Matérn $\frac{3}{2}$ kernels fit to the entire dataset, and found that the South Sudan study had an active subspace of dimension 2, the Mali study one of dimension 3, and all others one of dimension 1 (Figure 5, bottom).

6.1 Quantitative Prediction Comparison

To quantitatively evaluate the active subspace sensitivity, we compare the performance of predictive models fit to simulator data, both with and without "prewarping" [Wycoff et al., 2022] the points with the active subspace. As predictive models, we consider K nearest neighbors regression using the `caret` [Kuhn and Max, 2008] R package, local approximate Gaussian processes [Gramacy and Apley, 2015] using the `laGP` [Gramacy, 2016] R package, kernel regression using the `npregbw` command of the `np` [Hayfield and Racine, 2008] R package, and a random forest using the `randomForest` [Liaw and Wiener, 2002] library, all using default parameters.

We fit these models on the raw data as a baseline, and

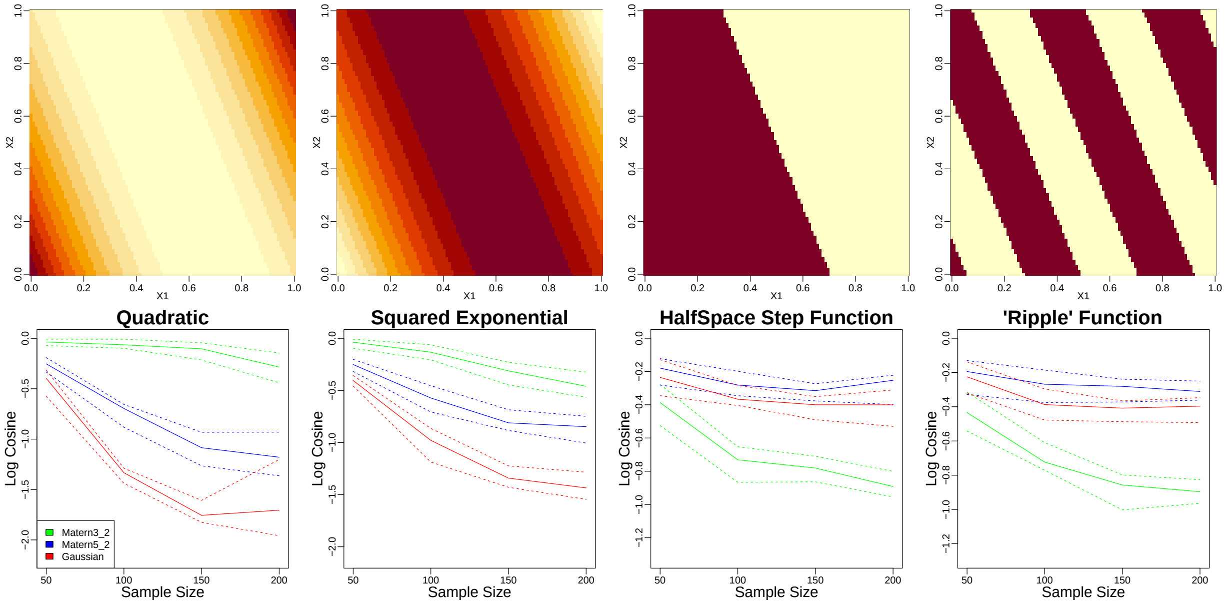

Figure 4: *Top:* Visualizations of test functions in 2D. *Bottom:* Mean subspace error for Gaussian Process estimates of the active subspace for various kernel functions. Solid line gives median and dotted lines give 25th and 75th percentiles.

then compare to several transformations of the input data. First, by transforming $\tilde{\mathbf{x}}_i = \mathbf{L}\mathbf{x}$, where $\mathbf{L} = \Lambda^{\frac{1}{2}}\mathbf{U}$ and $\mathbf{B}_{f,\mu} = \mathbf{U}\Lambda\mathbf{U}^\top$. We also consider truncating the warped data to a lower dimensional space by taking only the leading columns of $\mathbf{L}$. We compare this approach to a projection based on Sliced Inverse Regression, an SDR method. For truncated active subspace and SIR, we set the dimension of the reduced space to that determined by the eigenanalysis of the active subspace estimated on the full data. We perform 10-fold CV to estimate predictive accuracy. On KNN and laGP, the truncated active subspace method tends to perform best, sometimes tying with the SIR model, and losing to it on the CAR case study. Furthermore, it improves over the original KNN by an order of magnitude on South Sudan, Syria, Ethiopia and Burundi. For the kernel regression on the other hand, methods that do not truncate seem to do better. The best performing method is more variable with random forests, though on Burundi and South Sudan case studies the truncated active subspace provides an advantage.

6.2 Qualitative Findings

Interestingly, we find that the majority of the loadings of the first and second eigenvectors tend to map onto a single variable, with the exception of Burundi's second eigenvector (Table 2). We find that the `camp_weight` variable is most important for the Mali, Syria, Ethiopia and Burundi case studies, while `conflict_weight` is most important for the CAR and South Sudan case studies (Table 3).

We compute a projection of the 500 design points using the first two eigenvectors for each case study, shown in Figure 5, middle. For South Sudan, Syria, Ethiopia and Burundi, the surrogate active subspace seems to capture the majority of the variation in the response. This is not the case for Mali, which is unsurprising given the fact that the spectrum of the active subspace matrix indicated a three dimensional subspace. The Central African Republic, on the other hand, has significant outliers not explained by a higher dimensional subspace being present. Furthermore, recall that the quantitative study showed little variation across different methods, indicating that linear dimension reduction may not be suitable for this problem.

Flee has previously been studied via simulation study [Suleimenova et al., 2021], though of a very different kind. Whereas that study focused in on parameter settings very near the actually used parameters, we have here globally explored the simulator across an entire range of hyperparameter settings. Though this gives us a bird's eye view, it may also make it hard to see exactly what's happening the in regions of the space near the commonly used parameter settings.

7 DISCUSSION

Summary: In this article, we discussed some pathologies associated with surrogate estimation of active subspaces for functions with smooth and discontinu-

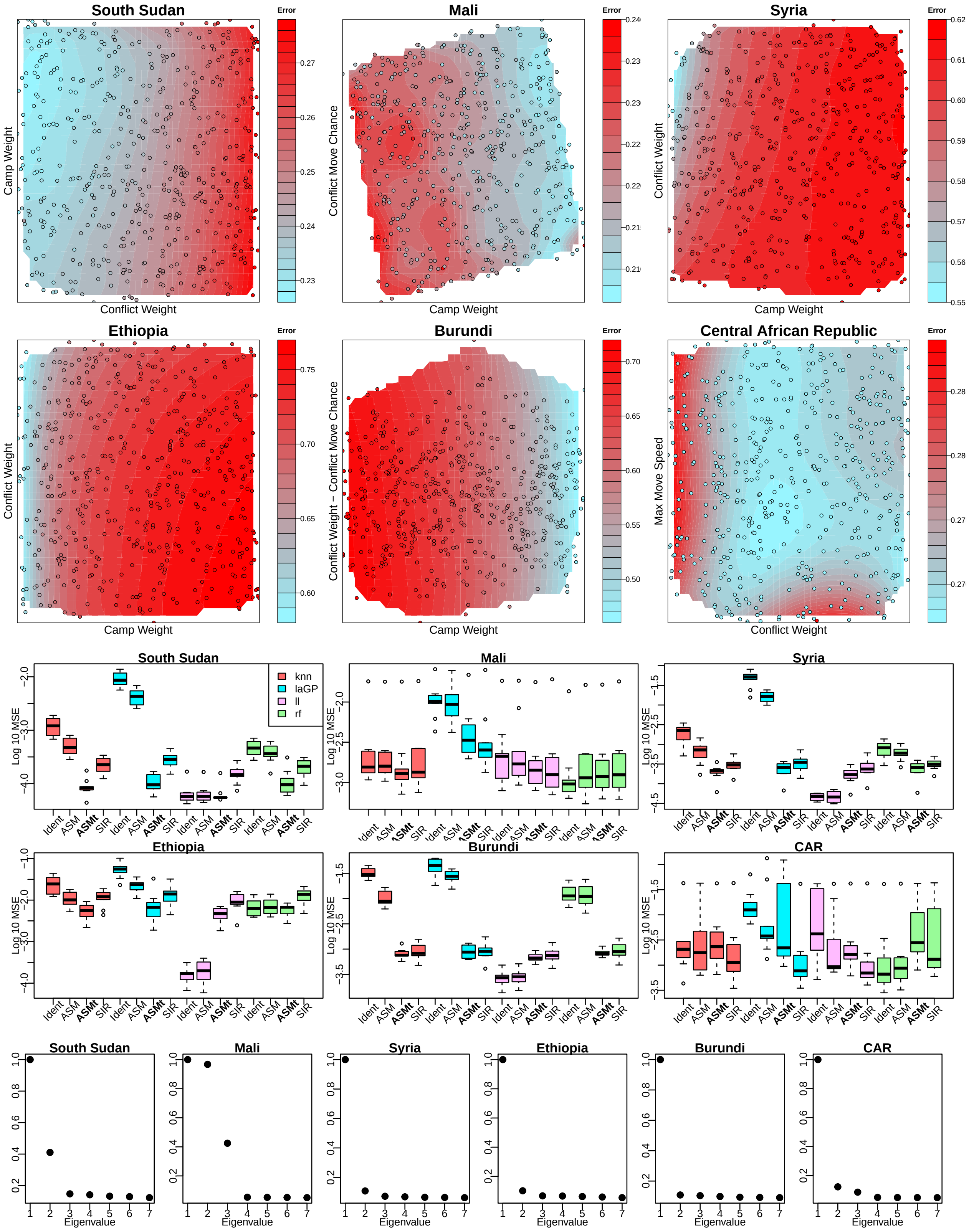

Figure 5: Flee simulator case study. *Top:* Active subspace projections of entire data sets. *Middle:* 10-fold CV predictive MSE with KNN. Competitors from left to right are KNN with no warping (Ident), with active subspace rotation only (ASM), with active subspace rotation and truncation (ASMt), and with SIR projection (SIR); lower is better. *Bottom:* Eigenvalues of surrogate active subspace matrices for each case study; gaps between subsequent eigenvalues indicate presence of active subspace.

1st Eigenvectors:

Param	S. Sudan	Mali	Syria	Ethiopia	Burundi	CAR
MMS	0.01	-0.00	-0.00	-0.00	-0.00	-0.00
MWS	-0.01	0.00	0.00	0.00	-0.00	-0.00
CMC	0.00	-0.00	0.00	-0.01	0.00	-0.00
CoMC	0.02	-0.20	-0.00	0.00	-0.00	0.00
DMC	-0.01	0.00	-0.00	-0.00	-0.01	-0.00
CW	0.01	0.98	1.00	1.00	1.00	0.00
CoW	1.00	0.00	-0.01	0.03	-0.04	1.00

2nd Eigenvectors:

Param	S. Sudan	Mali	Syria	Ethiopia	Burundi	CAR
MMS	-0.01	0.00	-0.04	0.06	-0.09	1.00
MWS	0.01	0.00	0.11	0.05	0.09	0.00
CMC	-0.01	0.00	0.05	0.03	-0.36	0.00
CoMC	0.02	0.98	-0.00	0.01	-0.61	-0.00
DMC	0.01	0.00	-0.06	-0.03	0.33	-0.01
CW	1.00	0.20	0.01	-0.03	0.03	0.02
CoW	-0.01	0.00	0.99	1.00	0.61	0.00

Table 2: Top two eigenvectors of estimated active subspace. MMS is Max Move Speed, MWS is Max Walk Speed, CMC is Camp Move Chance, CoMC is Conflict Move Chance, DMC is Default Move Chance, CW is Camp Weight and CoW is Conflict Weight.

	S. Sudan	Mali	Syria	Ethiopia	Burundi	CAR
1	CoW	CW	CW	CW	CW	CoW
2	CW	CoMC	CoW	CoW	CoW - CMC	MMS

Table 3: Qualitative Representation of Eigenvectors; see Table 2 caption.

ities components, and developed an extension of active subspaces to explain them. In our case study, we found that the surrogate active subspace estimates were for the most part axis-aligned. This was a surprising result; on most case studies to which active subspaces are deployed, the discovered dimensions are combinations of input parameters (e.g. [Lukaczyk et al., 2014, Constantine et al., 2016, Grey and Constantine, 2018]), however, visualization of the projected design points showed that the active subspaces did indeed accurately capture variation in the function, with the exception of the Mali and CAR case studies, which had too high a dimensional subspace or no clear linear subspace, respectively. Furthermore, it was interesting that both the dimension of the active subspaces and the type of active subspace varied from case study to case study, even for the same simulator and set of parameters.

Conclusions: Our numerical and analytic results provide us with several important conclusions. In studying a simulator with important smooth and discontinuous structure, we should keep in mind that by choosing a sample size, we are implicitly choosing a tradeoff between them, and that for sufficiently large sample sizes, the smooth directions will be lost. Furthermore, our analysis, via reasoning by limit arguments, puts into sharp relief a choice that is made when we do active subspaces: by squaring the gradient, we prioritize sharp jumps over gradual ones, even on fully differentiable simulators. This study also proved the viability of estimating the sensitive directions of a piecewise constant discontinuous function using continuous surrogates, namely Gaussian processes, and our numerical experiments suggest that best accuracy may be achieved by using minimally differentiable kernels, namely the Matérn $\frac{3}{2}$. In this article we tried to show what analysts are actually estimating when doing surrogate ASM on discrete simulators, in effect cautiously endorsing such analyses. Another reaction might have been condemnation: why use active subspaces when there are perfectly good dimension reduction tools not reliant on gradients? Our case study shows that on some applications, the ASM does better than SIR, a tool which does not use gradient structure, lending an empirical argument for discrete ASM deployment.

Future Work: This work left open some interesting questions, notably Conjecture 1. Additionally, it would be interesting to determine if it is possible to evaluate $\mathbf{B}_{f,\mu}$ analytically for Gaussian process surrogates that are not differentiable in mean square, such as a Matern $\frac{1}{2}$ process. Also, in revealing some pathologies of the ASM on mixed smooth-discontinuous simulators, we believe we have opened the door to future work which allows for explicit setting of a tradeoff between them. One approach would be a hyperparameter governing the relative strength of the two, by decomposing sensitivity into smooth and nonsmooth parts, or by using a different definition which directly avoids the delineated pathologies. In clarifying the behavior of surrogate active subspaces on fully discontinuous simulators, we hope to lend further theoretical understanding of future applied case studies. Finally, an implication of our work is that surrogate active subspaces may be useful in the context of mostly smooth simulators with unknown discontinuities. Whereas [Gorodetsky and Marzouk, 2014] develop an algorithm for determining *where* discontinuities occur, this proposed future work would determine along which *directions* discontinuities occur simply by conducting surrogate active subspace analysis on the simulator with a sufficiently large sample size. Some work would be required to determine how large is large enough, and how to best benefit from knowledge of these directions.

Acknowledgements

I gratefully acknowledge support from the McCourt Institute and the Massive Data Institute. I would like to thank Diana Suleimenova for technical assistance with the Flee simulator and for high level feedback, as well as Benjamin Harrop-Griffiths for discussions about technical aspects of the proof strategy. This paper was much improved by suggestions made by anonymous reviewers. I take full responsibility for any remaining errors.

References

[Adragni and Cook, 2009] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

[Arfken and Weber, 1972] Arfken, G. B. and Weber, H.-J. (1972). *Mathematical methods for physicists*. Academic Press Orlando, FL.

[Audet et al., 2022] Audet, C., Batailly, A., and Kojtych, S. (2022). Escaping unknown discontinuous regions in blackbox optimization. *SIAM Journal on Optimization*, 32(3):1843–1870.

[Butzer et al., 1987] Butzer, P., Ries, S., and Stens, R. (1987). Approximation of continuous and discontinuous functions by generalized sampling series. *Journal of approximation theory*, 50(1):25–39.

[Constantine and Gleich, 2014] Constantine, P. and Gleich, D. (2014). Computing active subspaces with monte carlo. *arXiv preprint arXiv:1408.0545*.

[Constantine, 2015] Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM.

[Constantine et al., 2016] Constantine, P. G., Kent, C., and Bui-Thanh, T. (2016). Accelerating markov chain monte carlo with active subspaces. *SIAM Journal on Scientific Computing*, 38(5):A2779–A2805.

[Conti et al., 2009] Conti, S., Gosling, J. P., Oakley, J. E., and O'Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3):663–676.

[Cook, 1994] Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

[Dupuis et al., 2018] Dupuis, R., Jouhaud, J.-C., and Sagaut, P. (2018). Surrogate modeling of aerodynamic simulations for multiple operating conditions using machine learning. *Aiaa Journal*, 56(9):3622–3635.

[Fukumizu and Leng, 2014] Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370.

[Gorodetsky and Marzouk, 2014] Gorodetsky, A. and Marzouk, Y. (2014). Efficient localization of discontinuities in complex computational simulations. *SIAM Journal on Scientific Computing*, 36(6):A2584–A2610.

[Gramacy, 2016] Gramacy, R. B. (2016). laGP: Large-scale spatial modeling via local approximate gaussian processes in R. *Journal of Statistical Software*, 72(1):1–46.

[Gramacy, 2020] Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman Hall/CRC, Boca Raton, Florida. `http://bobby.gramacy.com/surrogates/`.

[Gramacy and Apley, 2015] Gramacy, R. B. and Apley, D. W. (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

[Grey and Constantine, 2018] Grey, Z. J. and Constantine, P. G. (2018). Active subspaces of airfoil shape parameterizations. *AIAA Journal*, 56(5):2003–2017.

[Hayfield and Racine, 2008] Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32.

[Huang et al., 2020] Huang, J., Gramacy, R. B., Binois, M., and Libraschi, M. (2020). On-site surrogates for large-scale calibration. *Applied Stochastic Models in Business and Industry*, 36(2):283–304.

[Kuhn and Max, 2008] Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

[Li, 1991] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

[Logan and Shepp, 1975] Logan, B. F. and Shepp, L. A. (1975). Optimal reconstruction of a function from its projections.

[Lukaczyk et al., 2014] Lukaczyk, T. W., Constantine, P., Palacios, F., and Alonso, J. J. (2014). Active subspaces for shape optimization. In *10th AIAA multidisciplinary design optimization conference*, page 1171.

[Ma and Zhu, 2013] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

[Niutta et al., 2018] Niutta, C. B., Wehrle, E. J., Duddeck, F., and Belingardi, G. (2018). Surrogate modeling in design optimization of structures with discontinuous responses. *Structural and Multidisciplinary Optimization*, 57(5):1857–1869.

[Notestine, 2022] Notestine, J. G. (2022). *Sensitivity and Active Subspace Analysis for Agent-Based Models.* North Carolina State University.

[Palar and Shimoyama, 2018] Palar, P. S. and Shimoyama, K. (2018). On the accuracy of kriging model in active subspaces. In *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, page 0913.

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

[Samarov, 1993] Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847.

[Simonoff, 2012] Simonoff, J. S. (2012). *Smoothing methods in statistics.* Springer Science & Business Media.

[Suleimenova et al., 2021] Suleimenova, D., Arabnejad, H., Edeling, W. N., and Groen, D. (2021). Sensitivity-driven simulation development: a case study in forced migration. *Philosophical Transactions of the Royal Society A*, 379(2197):20200077.

[Suleimenova et al., 2017] Suleimenova, D., Bell, D., and Groen, D. (2017). A generalized simulation development approach for predicting refugee destinations. *Scientific reports*, 7(1):13377.

[Williams and Rasmussen, 2006] Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

[Wrathall et al., 2019] Wrathall, D., Mueller, V., Clark, P. U., Bell, A., Oppenheimer, M., Hauer, M., Kulp, S., Gilmore, E., Adams, H., Kopp, R., et al. (2019). Meeting the looming policy challenge of sea-level change and human migration. *Nature Climate Change*, 9(12):898–901.

[Wycoff et al., 2022] Wycoff, N., Binois, M., and Gramacy, R. B. (2022). Sensitivity prewarping for local surrogate modeling. *Technometrics*, 64(4):535–547.

[Wycoff et al., 2021] Wycoff, N., Binois, M., and Wild, S. M. (2021). Sequential learning of active subspaces. *Journal of Computational and Graphical Statistics*, 30(4):1224–1237.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes.
   (b) The license information of the assets, if applicable. Yes.
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes.
   (d) Information about consent from data providers/curators. Yes.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Surrogate Active Subspaces for Jump-Discontinuous Functions: Supplementary Materials

This document contains full proofs for the results given in the main paper's Section 4 in its first Section (Elementary Result 1, Elementary Result 2, Lemma 1, Lemma 2, Theorem 1, Theorem 2, Theorem 3, Theorem 4). It also contains a two-dimensional numerical illustration of Theorem 3. Finally, it provides an extended version of Section 5 of the main paper.

1 Proofs

We begin with a proof of each result. We then provide a numerical illustration of Theorem 3.

Elementary Result 1: $\int_{\mathbf{z}\in\mathcal{B}_r^P} z_i^2 d\mathbf{z} = \frac{\pi^{\frac{P}{2}} r^{P+2}}{2\Gamma(\frac{P+4}{2})} := \xi_P r^{P+2}$

Proof. $\int_{\mathbf{x}\in\mathcal{B}_r^P} x_1^2 = \int_{x_i\in[-r,r]} x_1^2 \int_{\mathbf{x}_c\in\mathcal{B}^{P-1}_{\sqrt{r^2-x_i^2}}} d\mathbf{x}_c dx_i$. Since the volume of the ball with radius ρ in dimension D is $\frac{\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2}+1)}\rho^P$, the integral may be re-expressed as: $\frac{\pi^{\frac{P-1}{2}}}{\Gamma(\frac{P-1}{2}+1)} \int_{x\in[-r,r]} x_i^2(r^2-x_i^2)^{\frac{P-1}{2}} dx_i$, and noting that $\int_{x\in[-r,r]} x_i^2(r^2-x_i^2)^{\frac{P-1}{2}} dx_i = \frac{\sqrt{\pi}r^{P+2}\Gamma(\frac{P+1}{2})}{2\Gamma(\frac{P+4}{2})}$ and that $\frac{\Gamma(\frac{P+1}{2})}{\Gamma(\frac{P-1}{2}+1)\Gamma(\frac{P}{2}+2)} = \frac{1}{\Gamma(\frac{P+4}{2})}$ yields the answer. □

Elementary Result 2:

$$\mathbb{E}_{\mathbf{z}\in\mathcal{B}_r^P}[z_i^2] = \frac{r^2}{P+2} \tag{1}$$

Proof. Follows from dividing the result in the preceding Elementary Result by the volume of the P-ball. □

Lemma 1. *1. If f consists only of a smooth term g we have that $\lim_{r\to 0}\beta_r(\mathbf{x}) = \nabla f(\mathbf{x})$.*

2. Otherwise, we have that:

$$\lim_{r\to 0} r\beta_r(\mathbf{x}) = \begin{cases} A_P \sum_{\{j:\mathbf{x}\in\partial\mathcal{S}_j\}} c_j \frac{\nabla h_j(\mathbf{x})}{\|\nabla h_j(\mathbf{x})\|_2} & \mathbf{x}\in\cup_j\partial\mathcal{S}_j \\ 0 & o.w. \end{cases} \tag{2}$$

Proof. Part 1 Expanding $f(\mathbf{z})$ about $\mathbf{x}$ and plugging in to our expression gives

$$\underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}f(\mathbf{z})] = \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}\Big(f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{z}-\mathbf{x}) + o(r)\Big)] = \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}(\nabla f(\mathbf{x})^\top\mathbf{z})] + o(r) = \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}\mathbf{z}^\top]\nabla f(\mathbf{x}) + o(r). \tag{3}$$

Then $\beta_r(\mathbf{x}) = \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}\mathbf{z}^\top]^{-1} \underset{\mathbf{z}\in\mathcal{B}_r}{\mathbb{E}}[\mathbf{z}f(x+\mathbf{z})] = \nabla f(\mathbf{x}) + o(r)$.

Part 2

Let us first assume that the target function is a single jump function, namely $f(\mathbf{x}) = \mathbb{1}_{\{\mathbf{x}\in\mathcal{S}\}}$. Fix some point $\mathbf{x}\in\cup_j\partial\mathcal{S}_j$ and some other point $\tilde{\mathbf{x}}$ such that $\|\mathbf{x}-\tilde{\mathbf{x}}\|_2 \le r$.

By Elementary Result 1, the first term evaluates to $\frac{1}{\xi_P r^{P+2}}\mathbf{I}$, so that $\beta_r(\tilde{\mathbf{x}}) = \frac{1}{\xi_P r^{P+2}}\int_{\mathbf{z}\in\mathcal{B}_r}\mathbf{z}f(\tilde{\mathbf{x}}+\mathbf{z})d\mathbf{z}$.

Then:

$$\beta_r(\tilde{\mathbf{x}}) = \frac{1}{\xi_P r^{P+2}}\int_{\mathbf{z}\in\mathcal{B}_r}\mathbf{z}f(\tilde{\mathbf{x}}+\mathbf{z})d\mathbf{z} = \frac{1}{\xi_P r^{P+2}}\int_{\{\mathbf{z}\in\mathcal{B}_r:h(\tilde{\mathbf{x}}+\mathbf{z})\le 0\}}\mathbf{z}d\mathbf{z} \tag{4}$$

$$= \frac{1}{\xi_P r^{P+2}}\int_{\{\mathbf{z}\in\mathcal{B}_r:h(\mathbf{x})+\nabla h(\mathbf{x})^\top(\mathbf{z}+\tilde{\mathbf{x}}-\mathbf{x})+o(r)\le 0\}}\mathbf{z}d\mathbf{z} \tag{5}$$

$$= \frac{1}{\xi_P r^{P+2}}\left[\int_{\{\mathbf{z}\in\mathcal{B}_r:\nabla h(\mathbf{x})^\top(\mathbf{z}+\tilde{\mathbf{x}}-\mathbf{x})\le 0\}}\mathbf{z}d\mathbf{z} + \int_{\{\mathbf{z}\in\mathcal{B}_r:|\nabla h(\mathbf{x})^\top(\mathbf{z}+\tilde{\mathbf{x}}-\mathbf{x})|\le|o(r)|\}}\mathbf{z}d\mathbf{z}\right]. \tag{6}$$

We will denote $\delta = \tilde{\mathbf{x}}-\mathbf{x}$.

For the first term in 6, we will take a change of variables, writing $\mathbf{z} = s_h\frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + \mathbf{U}\mathbf{z}_b$ and similarly $\delta = t_h\frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + \mathbf{U}\delta_b$, where $\mathbf{U}$ is an orthonormal basis for the set of vectors orthogonal to $\nabla h(\mathbf{x})$. Then we have:

$$\int_{\{\mathbf{z}\in\mathcal{B}_r:\nabla h(\mathbf{x})^\top(\mathbf{z}+\delta)\le 0\}}\mathbf{z}d\mathbf{z} = \int_{\{s_h\in[-r,r]:\|\nabla h(\mathbf{x})\|(s_h+l_h)\le 0\}}\int_{\{\mathbf{z}_b\in\mathcal{B}_{\sqrt{r^2-z_h}^2}\}} s_h\frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + \mathbf{U}\mathbf{z}_b d\mathbf{z}_b ds_h \tag{7}$$

$$= \frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|}\int_{s_h=-r}^{-l_h} s_h\left[\int_{\{\mathbf{z}_b\in\mathcal{B}_{\sqrt{r^2-z_h^2}}\}} d\mathbf{z}_b + \mathbf{U}s_h\left(\int_{\{\mathbf{z}_b\in\mathcal{B}_{\sqrt{r^2-z_h^2}}\}}\mathbf{z}_b\right)d\mathbf{z}_b\right]ds_h \tag{8}$$

The first inner integral is just the volume of the ball and the second is symmetric and vanishes, leaving us with

$$\int_{\{\mathbf{z}\in\mathcal{B}_r:\nabla h(\mathbf{x})^\top\mathbf{z}\leq 0\}} \mathbf{z}d\mathbf{z} = \frac{\pi^{\frac{P-1}{2}}}{\Gamma(\frac{P-1}{2}+1)} \frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} \int_{s_h=-r}^{-l_h} s_h \Big(r^2 - s_h^2\Big)^{\frac{P-1}{2}} ds_h \tag{9}$$

$$= \frac{\pi^{\frac{P-1}{2}}}{\Gamma(\frac{P+1}{2})} \Big(- \frac{(r^2-l_h^2)^{\frac{P+1}{2}}}{P+1} \Big) \frac{\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} \,. \tag{10}$$

For the second term of 6, we know that

$$\Big\| \int_{\{\mathbf{z}\in\mathcal{B}_r:|\nabla h(\mathbf{x})^\top\mathbf{z}|\leq|g(\mathbf{z})|\}} \mathbf{z}d\mathbf{z} \Big\|_{\ell_\infty} \leq r\mu(\{\mathbf{z}\in\mathcal{B}_r : |\nabla h(\mathbf{x})^\top\mathbf{z}| \leq |g(\mathbf{z})|\}) \leq r\big(\max_{\mathbf{z}\in\mathcal{B}_r} |g(\mathbf{z})| r^{P-1}\big) = \max_{\mathbf{z}\in\mathcal{B}_r} |g(\mathbf{z})| r^P \tag{11}$$

Combining 10 and 11 and defining $C_1 = \frac{\pi^{\frac{P-1}{2}}}{\Gamma(\frac{P+1}{2})(P+1)}$

$$\beta_r(\tilde{\mathbf{x}}) = \frac{1}{\xi_P r^{P+2}} \Bigg[C_1 (r^2 - l_h^2)^{\frac{P+1}{2}} \frac{-\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + o(r) r^P \Bigg] \tag{12}$$

$$\iff r\beta_r(\tilde{\mathbf{x}}) = A_P \frac{(r^2-l_h^2)^{\frac{P+1}{2}}}{r^{P+1}} \frac{-\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + o(1) \tag{13}$$

with $A_P = \frac{C_1}{\xi_P} = \frac{2\Gamma(\frac{P+4}{2})}{\sqrt{\pi}(P+1)\Gamma(\frac{P+1}{2})}$.

In particular, if $\tilde{\mathbf{x}} = \mathbf{x}$ then $t = 0$ and we have:

$$r\beta_r(\tilde{\mathbf{x}}) = A_P \frac{-\nabla h(\mathbf{x})}{\|\nabla h(\mathbf{x})\|} + o(1)\,. \tag{14}$$

This establishes the desired result for the case where f is the characteristic function associated with a single set parameterized by a smooth curve. Now let us consider a general $f = g + \sum_{j=1}^J c_j \mathbb{1}_{\mathcal{S}_j}$. Since $\beta_r(\mathbf{x})$ is linear in f, we have that:

$$\lim_{r\to 0} r\beta_r(\mathbf{x}) = A_P \sum_{\{j:\mathbf{x}\in\partial\mathcal{S}_j\}} c_j \frac{\nabla h_j(\mathbf{x})}{\|\nabla h_j(\mathbf{x})\|} + o(1) \tag{15}$$

since the term involving g goes to zero.

□

Theorem 1. *If f is once differentiable (i.e. $f = g$) and bounded on $\mathcal{X}$, compact, then $\lim_{r\to 0} \mathbf{B}^r_{f,\mu} = \mathbf{C}_{f,\mu}$.*

Proof. We need only show that the limit and integral may be interchanged. The assumptions have been made such that Lebesgue's Dominated Convergence Theorem applies, taking the dominating function to be the constant function with constant given by the max of $|f|$ on $[0,1]^P$. □

Lemma 2. *If $f(\mathbf{x})$ is constant along dimension $\mathbf{u}$, and μ is translation-invariant along $\mathbf{u}$, then $\mathbf{u}^\top\beta^r(\mathbf{x}) = 0$.*

Proof. $\mathop{\mathbb{E}}_{\mathbf{z}\in\mathcal{B}_r}[\mathbf{z}f(\mathbf{z})]^\top\mathbf{u} = \mathop{\mathbb{E}}_{\mathbf{z}\in\mathcal{B}_r}[f(\mathbf{z})\mathbf{z}^\top\mathbf{u}]$ which may be written as $\int_{\mathbf{z}\in\mathcal{B}_r} f(\mathbf{z})\mathbf{z}^\top\mathbf{u}d\mu$. Define $\mathbf{z}^\top\mathbf{u} = x_u$ and the orthogonal component as $\mathbf{z}_c$ and denote by $[\mathbf{z}]$ the equivalence class of $\mathbf{z}$ modulo span($\mathbf{u}$). Then break up the integral as $\int_{\left[\begin{smallmatrix}0\\ \mathbf{z}_c\end{smallmatrix}\right]\in\mathcal{B}_r} \int_{\left[\begin{smallmatrix}z_u\\ \mathbf{z}_c\end{smallmatrix}\right]\in\mathcal{B}_r} f(\mathbf{z})z_u$ or $\int_{\left[\begin{smallmatrix}0\\ \mathbf{z}_c\end{smallmatrix}\right]\in\mathcal{B}_r} f([\mathbf{z}]) \int_{\left[\begin{smallmatrix}z_u\\ \mathbf{z}_c\end{smallmatrix}\right]\in\mathcal{B}_r} z_u$. The proof is concluded by noting that the inner integral vanishes. □

Theorem 2. *If $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ with $\mathbf{A} \in \mathbb{R}^{R\times P}$ and $g : \mathbb{R}^R \to \mathbb{R}$, Range$(\mathbf{B}^r_{f,\mu}) \subseteq$ Range$(\mathbf{A})$.*

Proof. This follows from Lemma 2 together with the fact that integrating over matrices with a common nullspace preserves it. □

Lemma 3. *If $f : \mathbb{R} \to \mathbb{R}$ is a linear combination of translated heaviside functions, that is $f(x) = \sum_{j=1}^J c_j \mathbb{1}_{[x\le\tau_j]}$, and μ is Lebesgue-continuous with differentiable density δ, then $\lim\limits_{r\to 0} \frac{5r}{3}\mathbf{B}^r_{f,\mu} = \sum_{j=1}^J c_j^2\delta(\tau_j)$.*

Proof. For a given x, if $[x - r, x + r]$ does not contain any τ_j, then:

$$\mathop{\mathbb{E}}_{\mathbf{z}\in\mathcal{B}_r}[zf(x + z)] = f(x - r) \mathop{\mathbb{E}}_{\mathbf{z}\in\mathcal{B}_r}[z] = 0\,. \tag{16}$$

For $r < \min_{j_1,j_2} \frac{|\tau_{j_1}-\tau_{j_2}|}{2}$, each region $[x - r, x + r]$ can contain at most one τ_j, and only such regions will contribute to the integral, such that:

$$\int_{x\in[0,1]} \beta^r(x)^2 d\mu = \sum_{j=1}^J \int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2\delta(x)dx\,, \tag{17}$$

where we have assumed for simplicity that there is no τ_j on the boundary (which would correspond simply to shifting the function by a constant).

For an individual $\beta^r(x)$, we get that

$$\beta^r(x) = \frac{\int_{-r}^r zf(x + z)dz}{\int_{-r}^r z^2dr} = \frac{\int_{-r}^r zf(x + z)dz}{\frac{2}{3}r^3}\,. \tag{18}$$

Near x, we can write $f(x + z) = f_0 + \mathbb{1}_{z>\tau_j-x}$, and simplify the integral in the numerator as

$$\int_{-r}^r zf(x + z)dz = f_0 \int_{-r}^r zdz + \int_{-r}^r z\mathbb{1}_{z>\tau_j-x}dz = \int_{\tau_j-x}^r zdz = \frac{r^2 - (\tau_j - x)^2}{2}\,. \tag{19}$$

Thus

$$\beta^r(x) = c_j^2\frac{3}{4}\big(\frac{1}{r} - \frac{(\tau_j - x)^2}{r^3}\big)\,. \tag{20}$$

Now we turn our attention to $\int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2\delta(x)dx$. Expanding the density, $\delta(x) = \delta(\tau_j) + \delta'(\tau_j)(x - r) + o(r)$, and plugging into $\int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2\delta(x)dx$ yields:

$$\delta(\tau_j)\int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2dx + \delta'(\tau_j)\int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2(x - \tau_j)dx + o(r)\int_{x=\tau_j-r}^{x=\tau_j+r} \beta^r(x)^2dx\,. \tag{21}$$

Since $\int_{\tau_j-r}^{\tau_j+r} \big(\frac{1}{r} - \frac{(\tau_j-x)^2}{r^3}\big)^2 = \frac{16}{15r}$, and $\int_{\tau_j-r}^{\tau_j+r} \big(\frac{1}{r} - \frac{(\tau_j-x)^2}{r^3}\big)^2(x-\tau_j) = 0$, the expression resolves to $c_j^2\delta(\tau_j)\frac{3}{5r} + o(r)$. Thus $\lim_{r\to 0} \frac{5r}{3}\int_{x\in[0,1]} \beta^r(x)^2d\mu = \sum_{j=1}^J c_j^2\delta(\tau_j)$.

□

Theorem 3. *Denoting by $B_P = \frac{\Gamma(\frac{P}{2}+1)^2\Gamma(P+2)}{\sqrt{\pi}\Gamma(P+\frac{5}{2})\Gamma(\frac{P+3}{2})^2}$ and by S_j the surface measure of $\partial\mathcal{S}_j$, $\frac{1}{B_P}\mathbf{B}_{f,\mu} = \sum_{j=1}^J c_j^2\big[\int_{h_j^{-1}(0)} \frac{1}{\|\nabla h_j(\mathbf{x})\|_2^2}\nabla h_j(\mathbf{x})\nabla h_j(\mathbf{x})^\top\delta(\mathbf{x})dS_j(\mathbf{x})\big]$.*

Proof. Let's begin again by considering only a single characteristic function $f(\mathbf{x}) = \mathbb{1}_{\{x\in\mathcal{S}\}}$.

We denote $F = r\beta_r\beta_r^\top(\mathbf{x})\delta(\mathbf{x})$. Since $F = 0$ at any $\mathbf{x}$ sufficiently far from $\mathcal{S}$,

$$\int_{[0,1]^P} F(\mathbf{x})d\mathbf{x} = \int_{\{\mathbf{x}:dist(\mathbf{x},\mathcal{S})\leq r\}} F(\mathbf{x})d\mathbf{x} \tag{22}$$

Let $\{\mathcal{U}_i, \psi_i\}_{i=1}^I$ be an atlas for $\mathcal{S}$. I is finite because $\mathcal{S}$ is compact (since it resides in $[0,1]^P$).

We define functions F_i, each compactly supported on $\mathcal{U}_i$, such that $F = \sum_{i=1}^I F_i$ using a partition of unity, and write:

$$\sum_{i=1}^I \int_{\{\mathbf{x}:\exists\tilde{\mathbf{x}}\in\mathcal{U},t\in[-r,r]\; s.t.\; \mathbf{x}=\tilde{\mathbf{x}}+t\frac{\nabla h(\tilde{\mathbf{x}})}{\|\nabla h(\tilde{\mathbf{x}})\|}\}} F_i(\mathbf{x})d\mathbf{x} \tag{23}$$

$$= \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} \int_{t\in[-r,r]} F_i\left(\psi_i(\tilde{\mathbf{x}}) + t\frac{\nabla h(\tilde{\mathbf{x}})}{\|\nabla h(\tilde{\mathbf{x}})\|}\right)|\det \nabla\Psi(\tilde{\mathbf{x}},t)|dtd\tilde{\mathbf{x}} \tag{24}$$

where $\Psi(\tilde{\mathbf{x}},t) = \psi_i(\tilde{\mathbf{x}}) + t\frac{\nabla h(\tilde{\mathbf{x}})}{\|\nabla h(\tilde{\mathbf{x}})\|}$.

We note that $|\det \Psi_i(\tilde{\mathbf{x}},t)| = |\det \Psi_i(\tilde{\mathbf{x}},t)\Psi_i(\tilde{\mathbf{x}},t)^\top|^{\frac{1}{2}} = |\det \psi_i(\tilde{\mathbf{x}})\psi_i(\tilde{\mathbf{x}})^\top|$ because $\nabla_{\tilde{\mathbf{x}}}\psi_i(\tilde{\mathbf{x}})^\top\nabla_{\tilde{\mathbf{x}}}\frac{\nabla h(\tilde{\mathbf{x}})}{\|\nabla h(\tilde{\mathbf{x}})\|} = 0$.

Plugging this in and using 13 a few lines later yields:

$$\sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}}\left[\int_{t\in[-r,r]} F_i\left(\psi_i(\tilde{\mathbf{x}}) + t\frac{\nabla h(\tilde{\mathbf{x}})}{\|\nabla h(\tilde{\mathbf{x}})\|}\right)dt\right]d\tilde{\mathbf{x}} \tag{25}$$

$$= \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}}\left[r\int_{t\in[-r,r]} \beta_r(\tilde{\mathbf{x}})\beta_r(\tilde{\mathbf{x}})^\top\delta(\tilde{\mathbf{x}})dt\right]d\tilde{\mathbf{x}} \tag{26}$$

$$= \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}}\left[A_P^2\frac{\nabla h(\tilde{\mathbf{x}})\nabla h(\tilde{\mathbf{x}})^\top}{\|\nabla h(\tilde{\mathbf{x}})\|^2}\delta(\tilde{\mathbf{x}})r\int_{t\in[-r,r]}\frac{(r^2-t^2)^{P+1}}{r^{2P+4}}dt + o(1)\right]d\tilde{\mathbf{x}} \tag{27}$$

$$= \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}}\left[A_P^2\frac{\nabla h(\tilde{\mathbf{x}})\nabla h(\tilde{\mathbf{x}})^\top}{\|\nabla h(\tilde{\mathbf{x}})\|^2}\delta(\tilde{\mathbf{x}})r\left[\frac{\sqrt{\pi}\Gamma(P+2)}{\Gamma(P+5/2)}\frac{r^{2P+3}}{r^{2P+4}}\right] + o(1)\right]d\tilde{\mathbf{x}} \tag{28}$$

$$:= \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}}\left[B_P\frac{\nabla h(\tilde{\mathbf{x}})\nabla h(\tilde{\mathbf{x}})^\top}{\|\nabla h(\tilde{\mathbf{x}})\|^2}\delta(\tilde{\mathbf{x}}) + o(1)\right]d\tilde{\mathbf{x}} \tag{29}$$

Taking the limit and using the definition of integration over a manifold yields:

$$\lim_{r\to 0}\int_{\mathbf{x}\in[0,1]^P} r\beta_r\beta_r^\top(\mathbf{x})\delta(\mathbf{x})d\mathbf{x} = \sum_{i=1}^I \int_{\tilde{\mathbf{x}}\in\mathcal{U}_i} |\det \nabla\psi(\tilde{\mathbf{x}})\psi(\tilde{\mathbf{x}})^\top|^{\frac{1}{2}} B_P\frac{\nabla h(\tilde{\mathbf{x}})\nabla h(\tilde{\mathbf{x}})^\top}{\|\nabla h(\tilde{\mathbf{x}})\|^2}d\tilde{\mathbf{x}} \tag{30}$$

$$= B_P\int_{\mathbf{x}\in\mathcal{S}} \frac{\nabla h(\mathbf{x})\nabla h(\mathbf{x})^\top}{\|\nabla h(\mathbf{x})\|^2}dS(\mathbf{x}) \tag{31}$$

where S is the surface measure of $\mathcal{S}$, and the constant is given by:

$$B_P = \frac{\Gamma(\frac{P}{2}+1)^2\Gamma(P+2)}{\sqrt{\pi}\Gamma(P+\frac{5}{2})\Gamma(\frac{P+3}{2})^2}, \tag{32}$$

which we can verify resolves to $\frac{3}{5}$ when $P = 1$, as Lemma 3 demands.

This establishes the behavior for f given by a single characteristic function. We'd now like to consider f of the form $f = g + \sum_{j=1}^J c_j\mathbb{1}_{\mathcal{S}_j}$. We will split this into two cases. First, assume that there is a positive r^* that separates the $\mathcal{S}_j$. In such a case, taking $r < \frac{r^*}{2}$, we can split the integral into J independent integrals containing each curve, yielding:

$$\lim_{r\to 0}\int_{\mathbf{x}\in[0,1]} r\beta_r(\mathbf{x})\beta_r(\mathbf{x})^\top d\mu = B_P\sum_{j=1}^{J} c_j^2\int_{\mathbf{x}\in\mathcal{S}}\frac{\nabla h_j(\mathbf{x})\nabla h_j(\mathbf{x})^\top}{\|\nabla h_j(\mathbf{x})\|^2}\delta(\mathbf{x})dS_j(\mathbf{x})\,. \tag{33}$$

This will also cover the case where the boundaries only intersect briefly by passing through each other, or in any way such that the surface measure of the intersections are zero.

But in many cases, such as in the Flee example given in this article, we will have edges shared by more than one $\mathcal{S}_j$ (see Figure 1). But if i and j share an edge, we will have that $\frac{\nabla h_i(\mathbf{x})}{\|\nabla h_i(\mathbf{x})\|} = \frac{\nabla h_j(\mathbf{x})}{\|\nabla h_j(\mathbf{x})\|}$ on that edge. This means that we have

$$\lim_{r\to 0} r\beta_r(\mathbf{x}) = A_P\left(\sum_{\{j:\mathbf{x}\in\mathcal{S}_j\}} c_j\right)\frac{\nabla h_i(\mathbf{x})}{\|\nabla h_i(\mathbf{x})\|} \iff \lim_{r\to 0} r\beta_r(\mathbf{x})\beta_r(\mathbf{x})^\top = A_P^2\left(\sum_{\{j:\mathbf{x}\in\mathcal{S}_j\}} c_j\right)^2\frac{\nabla h_i(\mathbf{x})\nabla h_i(\mathbf{x})^\top}{\|\nabla h_i(\mathbf{x})\|^2} \tag{34}$$

where i is any index in $\{j : \mathbf{x}\in\mathcal{S}_j\}$.

Now we partition $\underset{j\in\{1,\dots,J\}}{\cup}\partial\mathcal{S}_j$ into disjoint intervals $\mathcal{C}_i$ such that for all $\mathbf{x}\in\mathcal{C}_i$, there is some fixed index set $\mathcal{I}_i$ such that $\mathbf{x}$ belongs to the boundary of every $\mathcal{S}_j$ for $j\in\mathcal{I}_i$.

This gives us:

$$\lim_{r\to 0}\int_{\mathbf{x}\in[0,1]} r\beta_r(\mathbf{x})\beta_r(\mathbf{x})^\top d\mu = B_P\sum_i\int_{\{\mathbf{x}\in\mathcal{C}_i\}}\left(\sum_{\{j:\mathbf{x}\in\mathcal{S}_j\}} c_j\right)^2\frac{\nabla h_i(\mathbf{x})\nabla h_i(\mathbf{x})^\top}{\|\nabla h_i(\mathbf{x})\|^2}\delta(\mathbf{x})dS_j(\mathbf{x})\,. \tag{35}$$

In the event of functions h_j which are only piecewise continuous, again as in the Flee simulator, we can apply the above results on each of the continuous parts and sum.

□

Theorem 4. *Given a dataset* $\mathbf{U}, y = f(\mathbf{U})$, *the NW kernel estimate is* $m(\mathbf{x}) = \frac{k(\mathbf{x},\mathbf{U})^\top\mathbf{y}}{k(\mathbf{x},\mathbf{U})^\top\mathbf{1}}$. *Let* $k(\mathbf{x}_1-\mathbf{x}_2) = d(\frac{\|\mathbf{x}_1-\mathbf{x}_2\|_2}{\sqrt{l}})$ *be a stationary kernel such that:*

1. *d is diffeomorphic and decreasing.*
2. $\int_{z=R}^{\infty} d(z) \le e^{\frac{-R^2}{C}}$ *for some constant* C.
3. $d(0) = 1$, $d(\infty) = 0$.
4. $\int_0^\infty \frac{d'(\frac{\tau}{\sqrt{l}})}{\tau}d\tau$ *is finite for sufficiently small* l.

The Gaussian kernel is such a kernel. Assume the data $\mathbf{U}$ *are iid uniformly distributed. Then we have*

$$\lim_{l\to 0}\lim_{n\to\infty}\int_{\mathbf{x}\in[0,1]^P}\sqrt{l}\nabla m(\mathbf{x})\nabla m(\mathbf{x})^\top d\mu = C\lim_{r\to 0}\int_{\mathbf{x}\in[0,1]^P} r\beta_r(\mathbf{x})\beta_r(\mathbf{x})^\top d\mu \tag{36}$$

where C is a constant depending only on k and P.

Proof. If $\nabla k(\mathbf{x},\mathbf{U})\in\mathbb{R}^{N\times P}$ is the Jacobian of the kernel, the gradient of the NW predictor is:

$$\frac{\nabla k(\mathbf{x},\mathbf{U})^\top(k(\mathbf{x},\mathbf{U})^\top\mathbf{1})\mathbf{y} - (k(\mathbf{x},\mathbf{U})^\top\mathbf{y})\nabla k(\mathbf{x},\mathbf{U})\mathbf{1}}{(k(\mathbf{x},\mathbf{U})^\top\mathbf{1})^2}\in\mathbb{R}^P\,. \tag{37}$$

Reorganizing gives:

$$\frac{(\mathbf{y}-\frac{v}{w}\mathbf{1})\nabla k(\mathbf{x},\mathbf{U})}{w}\,, \tag{38}$$

where $w = k(\mathbf{x}, \mathbf{U})^\top \mathbf{1} = \sum_{n=1}^N k(\mathbf{x}, \mathbf{u}_n)$ and $v = k(\mathbf{x}, \mathbf{U})^\top \mathbf{y}$.

Thus, $\lim_{n\to\infty} \frac{w}{n} = \underset{\mathbf{u}\sim U[0,1]}{\mathbb{E}}[k(\mathbf{x}, \mathbf{u})] := w_0$ and $\lim_{n\to\infty} \frac{v}{n} = \underset{\mathbf{u}\sim U[0,1]}{\mathbb{E}}[k(\mathbf{x}, \mathbf{u})y(\mathbf{u})] := v_0$.

$$\lim_{n\to\infty} \frac{\sum_{i=n}^N \left(y(\mathbf{u}_i) - \frac{\frac{v}{N}}{\frac{w}{N}}\right)(\mathbf{x} - \mathbf{u}_i)k(\mathbf{x} - \mathbf{u}_i)}{N\frac{w}{N}} \tag{39}$$

$$= \frac{\mathbb{E}_{\mathbf{u}\sim U[0,1]^P}[(f(\mathbf{u}) - \frac{v_0}{u_0})(\mathbf{x} - \mathbf{u})k(\mathbf{x} - \mathbf{u})]}{w_0} = \frac{1}{w_0}\int_{\mathbf{u}\in[0,1]^P} (f(\mathbf{u}) - \frac{v_0}{w_0})(\mathbf{x} - \mathbf{u})k(\mathbf{x} - \mathbf{u})d\mathbf{u}\,. \tag{40}$$

We now take a change of variables $\mathbf{z} = \mathbf{x} - \mathbf{u}$, yielding:

$$\frac{1}{w_0}\int_{\{\mathbf{z}\in\mathbb{R}^P : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} (f(\mathbf{x} + \mathbf{z}) - \frac{v_0}{w_0})\mathbf{z}k(\mathbf{z})d\mathbf{u} \tag{41}$$

$$= \frac{1}{w_0}\int_{\{\mathbf{z}\in\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} (f(\mathbf{x} + \mathbf{z}) - \frac{v_0}{w_0})\mathbf{z}k(\mathbf{z})d\mathbf{u} + \frac{1}{w_0}\int_{\{\mathbf{z}\in\mathbb{R}^P\setminus\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} (f(\mathbf{x} + \mathbf{z}) - \frac{v_0}{w_0})\mathbf{z}k(\mathbf{z})d\mathbf{u}\,. \tag{42}$$

Let us denote $\mathcal{N} = \{\mathbf{z} \in \mathbb{R}^P \setminus \mathcal{B}_r : \mathbf{z} + \mathbf{x} \in [0, 1]^P\}$.

We can bound the second term, denoting the measure with Lebesgue-density $k(\mathbf{z})$ as λ and f^* the value of f maximizing $|f(\tilde{\mathbf{x}}) - \frac{v_0}{w_0}|$ for $\tilde{\mathbf{x}} \in [0, 1]^P$:

$$\left|\int_{\mathcal{N}} (f(\mathbf{x} + \mathbf{z}) - \frac{v_0}{w_0})\mathbf{z}k(\mathbf{z})d\mathbf{u}\right| \leq |f^* - \frac{v_0}{w_0}|\sqrt{P}\lambda(\mathcal{N}) \leq |f^* - \frac{v_0}{w_0}|\sqrt{P}\frac{1}{\sqrt{l}}e^{-\frac{lr^2}{C_s^2}} := C_1\frac{1}{\sqrt{l}}e^{-\frac{r^2}{lC_s^2}} \tag{43}$$

and since the $\mathbf{z}k(\mathbf{z})$ integrates to $\mathbf{0}$ over the ball, we get:

$$\frac{1}{w_0}\int_{\{\mathbf{z}\in\mathbb{R}^P : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} (f(\mathbf{x} + \mathbf{z}) - \frac{v_0}{w_0})\mathbf{z}k(\mathbf{z})d\mathbf{u} \tag{44}$$

$$= \frac{1}{w_0}\int_{\{\mathbf{z}\in\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} f(\mathbf{x} + \mathbf{z})\mathbf{z}k(\mathbf{z})d\mathbf{z} + O(e^{-\frac{r^2}{lC_s^2}}) \tag{45}$$

We'd like to pick r in terms of l such that as $l \to 0$, $r \to 0$ but $\frac{r^2}{l} \to \infty$. One such $r = l^{\frac{1}{4}}$.

By the fundamental theorem of calculus $d(\frac{t}{\sqrt{l}}) = -\int_t^\infty \frac{1}{\sqrt{l}}d'(\frac{\tau}{\sqrt{l}})d\tau = -\frac{1}{\sqrt{l}}\int_0^\infty d'(\frac{\tau}{\sqrt{l}})\mathbb{1}_{[t\leq\tau]}d\tau$, so:

$$\int_{\{\mathbf{z}\in\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} f(\mathbf{x} + \mathbf{z})\mathbf{z}k(\mathbf{z})d\mathbf{z} \tag{46}$$

$$= \frac{1}{\sqrt{l}}\int_{\{\mathbf{z}\in\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} f(\mathbf{x} + \mathbf{z})\mathbf{z}\int_0^\infty d'(\frac{\tau}{\sqrt{l}})\mathbb{1}_{[t\leq\tau]}d\tau d\mathbf{z} \tag{47}$$

$$= \frac{1}{\sqrt{l}}\int_0^\infty d'(\frac{\tau}{\sqrt{l}})\int_{\{\mathbf{z}\in\mathcal{B}_r : \mathbf{z}+\mathbf{x}\in[0,1]^P\}} f(\mathbf{x} + \mathbf{z})\mathbf{z}\mathbb{1}_{[\frac{\|\mathbf{z}\|_2}{\sqrt{l}}\leq\tau]}d\mathbf{z}d\tau \tag{48}$$

$$= \frac{1}{\sqrt{l}}\int_0^\infty d'(\frac{\tau}{\sqrt{l}})\beta_\tau(\mathbf{x})d\tau = \frac{1}{\sqrt{l}}\int_0^\infty \frac{d'(\frac{\tau}{\sqrt{l}})}{\tau}\tau\beta_\tau(\mathbf{x})d\tau\,. \tag{49}$$

Thus

$$\lim_{l\to 0}\sqrt{l}\int_{\{\mathbf{z}\in\mathcal{B}_r:\mathbf{z}+\mathbf{x}\in[0,1]^P\}} f(\mathbf{x}+\mathbf{z})\mathbf{z}k(\mathbf{z})d\mathbf{z} = \lim_{l\to 0} l\beta_l(\mathbf{x})\int_0^\infty \frac{d'(\frac{\tau}{\sqrt{l}})}{\tau}d\tau := C_d \lim_{l\to 0} l\beta_l(\mathbf{x})\,, \tag{50}$$

where $C_d = \lim_{l\to 0}\int_0^\infty \frac{d'(\frac{\tau}{\sqrt{l}})}{\tau}d\tau$ is a constant independent of $\mathbf{x}$. Since the quantities converge pointwise and the integrand is bounded and over a compact set, we have convergence of the integrals as well up to the constant of proportionality C_d.

□

1.1 Numerical Illustration of Theorem 3

In this section, we examine a toy discontinuous function.

Let:

$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{51}$$

$$\Sigma_2 = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 0.5 \end{bmatrix} \tag{52}$$

Then we define three sets:

$$\mathcal{S}_1 = \{\mathbf{x} : (\mathbf{x}-0.1)^\top\Sigma_1(\mathbf{x}-0.1) \le 0.005\} \tag{53}$$

$$\mathcal{S}_2 = \{\mathbf{x} : (\mathbf{x}-0.5)^\top\Sigma_2(\mathbf{x}-0.5) \le 0.1\} \tag{54}$$

$$\mathcal{S}_3 = \{\mathbf{x} : (\mathbf{x}-0.9)^\top\Sigma_1(\mathbf{x}-0.9) \le 0.005\} \tag{55}$$

and subsequently $f_n = \mathbb{1}_{[\mathbf{x}\in\mathcal{S}_1]} + \mathbb{1}_{[\mathbf{x}\in\mathcal{S}_2]} + \mathbb{1}_{[\mathbf{x}\in\mathcal{S}_3]}$.

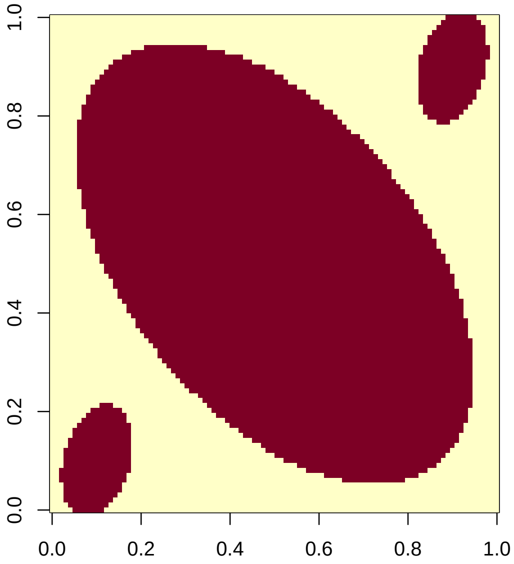

Figure 1: The discontinuous function f_n. The ellipses give the sets $\mathcal{S}_j$.

For this toy example, we will let μ be the Lebesgue measure. Then we can analytically determine the expression:

$$\int_{\mathbf{v}\in h_j^{-1}(0)} \frac{\nabla h(\mathbf{v})}{\|\nabla h(\mathbf{v})\|}\frac{\nabla h(\mathbf{v})}{\|\nabla h(\mathbf{v})\|}^{\top} \delta(\mathbf{v})d\mathbf{v} \tag{56}$$

for each of the three characteristic sets. For sets 1 and 3, it is given by $per(\mathcal{S}_1)\frac{\Sigma_1}{\|\Sigma_1\|}$ where $\|.\|$ is the induced norm/2-norm, and $per(\mathcal{S}_1)$ gives the perimeter of the ellipse. Similarly, $per(\mathcal{S}_2)\frac{\Sigma_2}{\|\Sigma_2\|}$ is the contribution for set 2. Therefore, the analytic solution in this case is proportional to:

$$\mathbb{C}_a = \sum_{j=1}^{3}\int_{\mathbf{v}\in h_j^{-1}(0)} \frac{\nabla h(\mathbf{v})}{\|\nabla h(\mathbf{v})\|}\frac{\nabla h(\mathbf{v})}{\|\nabla h(\mathbf{v})\|}^{\top} d\mathbf{v} = per(\mathcal{S}_2)\frac{\Sigma_2}{\|\Sigma_2\|} + 2per(\mathcal{S}_1)\frac{\Sigma_1}{\|\Sigma_1\|} \tag{57}$$

$$\mathbb{C}_a = \begin{bmatrix} 2.96 & 0.72 \\ 0.72 & 2.42 \end{bmatrix} \tag{58}$$

We next build a Monte Carlo estimator of $\mathbb{E}_{\mathbf{x}\in[0,1]^P}[\beta^r(\mathbf{x})\beta^r(\mathbf{x})^{\top}]$ by randomly sampling $50,000$ $\mathbf{x}_m$ from a uniform distribution and then sampling 100 points within $r = 0.001$ (using antithetic sampling) of $\mathbf{x}_m$. We use OLS to estimate $\beta^r(\mathbf{x}_m)$ and then form $\hat{\mathbf{C}} = \frac{1}{M}\sum_{m=1}^{M}\beta^r(\mathbf{x}_m)\beta^r(\mathbf{x}_m)^{\top}$.

Since we expect them to be only proportional, we compare the relative eigenvalues and the eigenvectors. For $\mathbf{C}_a$, we get normalized eigenvalues of 1 and 0.5546424 and eigenvectors given by :

$$\begin{bmatrix} -0.82 & 0.57 \\ -0.57 & -0.82 \end{bmatrix} \tag{59}$$

For $\hat{\mathbf{C}}_a$, we get normalized eigenvalues of 1 and 0.5282942 and eigenvectors given by:

$$\begin{bmatrix} -0.80 & 0.59 \\ -0.59 & -0.80 \end{bmatrix} \tag{60}$$

These estimates are quite close as predicted.

2 ADDITIONAL EXPERIMENTS

We present an extension of the analysis Section 5 to varying dimensions in Figure 2. We find that the conclusions do not change as the dimension is varied from 7 to 5 and 3. The gap between methods seems greater in the smooth case in smaller dimension compared to in larger dimension. For the discontinuous case, the same patterns are observed across dimensions. We come to the same conclusion as the main text: if we expect to run into a discontinuous function, the rougher Matérn $\frac{3}{2}$ kernel does best on such problems. But it will not be able to take advantage of smooth changes.

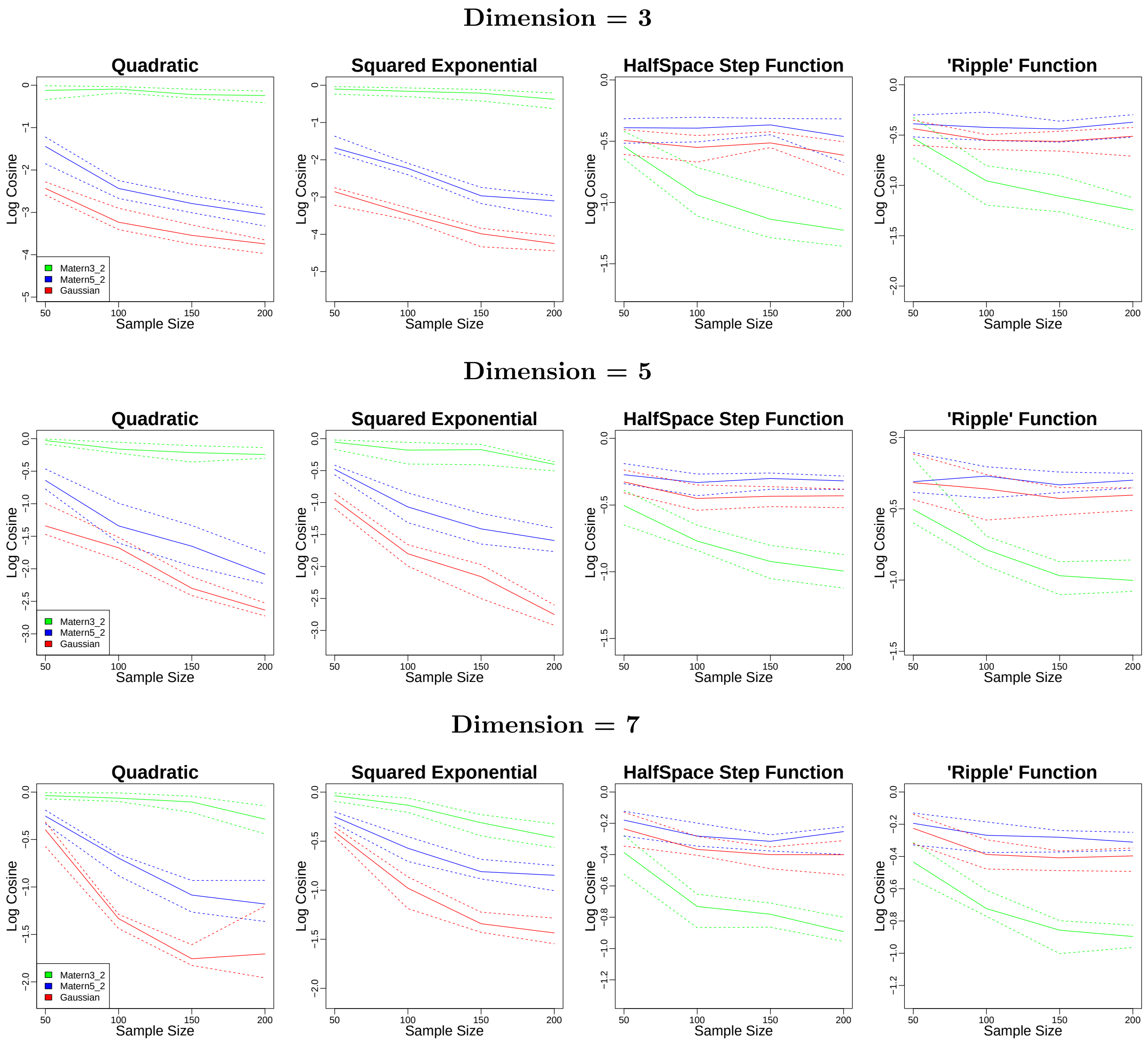

Figure 2: Gaussian Process estimates of the active subspace; lower is better. *Top:* Dimension = 3. *Middle:* Dimension=5. *Bottom:* Dimension=9.