Title
Sensitivity Analysis of Stochastic Simulators with Information Theory

Author
Huoh, Yu-Jay

Publication Date
2013

Peer reviewed|Thesis/dissertation

Sensitivity Analysis of Stochastic Simulators with Information Theory

By

Yu-Jay Huoh

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Cari G. Kaufman, Chair
Professor Deborah Nolan
Professor Armen Der Kiureghian

Fall 2013

Sensitivity Analysis of Stochastic Simulators with Information Theory

Abstract

Sensitivity Analysis of Stochastic Simulators with Information Theory

by

Yu-Jay Huoh

Doctor of Philosophy in Statistics

University of California, Berkeley

Assistant Professor Cari G. Kaufman, Chair

The increased computational power available today has made the use of "computer models" or "simulators" common in many fields. While there is a widely adopted set of tools for the analysis of simulators, there are still many unsolved problems when dealing with these models. Specifically, the traditional methods for sensitivity analysis of deterministic computer models, based on functional ANOVA decompositions, do not generalize well to simulators with stochastic or nondeterministic output. This paper presents a methodological solution for conducting sensitivity analysis on computer models with stochastic output through the use of information theory and Bayesian density regression. The presented method is applied to the inputs of a near-fault ground motion stochastic simulator of Dabaghi et al. (2011).

I dedicate this dissertation to my family and to true statisticians everywhere.

Contents

Acknowledgments

I would like to express my deepest appreciation to my advisor, Professor Cari Kaufman, for all her insight and guidance during my years working with her. Without her efforts, this dissertation would not have been possible.

I would also like to thank all the members of her research group: Ben, Wayne, Linda, and Regina, for all their help, support, and friendship.

In addition, I am truly grateful for all the efforts of Mayssa Dabaghi. Without her last minute efforts, this dissertation would not have come together in time.

Lastly, I want to thank the rest of the faculty at UC Berkeley, particularly my two other committee members, Professor Deborah Nolan and Professor Armen Der Kiureghian for their understanding and accommodations.

Some portion of this work was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Chapter 1: Introduction

The goal of this dissertation is to present a method for performing global sensitivity analysis on stochastic simulators.

1.1 Simulators

A *simulator* refers to any computer program or model that, when given input parameters, produces some output value. The term simulator is used because, these programs typically represent numerical approximations to any number of real life phenomena. There are simulators for natural processes such as earthquakes or climate. Simulators also exist for smaller scale processes such as automobile factories or motorcycle crashes.

In the past, most simulators have been deterministic: for a given set of inputs, a deterministic simulator always returns the same output value. Because one set of inputs corresponds to a singular output value, we can represent simulators mathematically as functions. Formally, for a fixed input X, a deterministic simulator, F, is a function mapping X to the same output value $Y = F(X)$, regardless of the number of times the function is evaluated or the computer program is run. Here, X does not have to be a scalar; it may be a vector of inputs $X = (x_1, x_2, \ldots, x_p)$. However, for the purposes of this dissertation, the simulators (deterministic or otherwise) being studied will only produce output that is scalar or univariate. That is, for a given set of inputs, the simulator only produces a one-dimensional output. In contrast, a multivariate simulator would produce multiple values on a single simulation run for a given set of parameters. Although our focus will be on the analysis of univariate simulators, the methods described here may still be applicable to multivariate simulators. If the output from a multivariate simulator can be reduced to a single variable summary, then all of our methods can be applied.

1.1.1 Stochastic Simulators

A *stochastic simulator* is also a computer program or model, with the distinction being the output value produced is *random* or *nondeterministic*. That is, for a fixed set of inputs, a stochastic simulator will output a different value each time it is run. Since stochastic simulators do not map a set of inputs to a singular output value, we cannot think of them as functions in the traditional mathematical sense. Instead, we think of a stochastic simulator

as a process mapping input values to output distributions. Formally, a stochastic simulator, F, is a process mapping a set of inputs, X, to a distribution for the output values. We let Y denote the output from a single run of the simulator and we write $Y|X \sim F(X)$ to emphasize both the randomness in Y and the roles of F and X on the *distribution* of Y, not just the output value. When the simulator F is not in question, we write $Y|X$ for the output distribution of the simulator given X.

1.1.2 Input Uncertainty

There are many situations where one would want to attach a distribution to the inputs of a simulator. For example, the simulator under study may represent a real-life process and one of the input parameters may correspond to a physical constant whose true value is unknown because it is difficult or costly to measure. In this case, the distribution on the inputs may account for uncertainty in the true value of this parameter. Alternatively, suppose a factory simulator is being used to identify the effects of varying worker efficiency or the chances of a machine breaking on factory output. In this case, the uncertainty in the inputs represents potential values for employee efficiency or equipment reliability.

This notion of input uncertainty is important in the context of sensitivity analysis for simulators, which will be discussed further in Section 1.2. However, it is for this reason that we take care to emphasize the distinctions in both the outputs of deterministic and stochastic simulators as well as the different kinds of output distributions that arise when talking about them. For a deterministic simulator, any distribution on the outputs is entirely a result of uncertainty in the inputs. If the inputs were known exactly, the output would not be random. Formally, let $\mathcal{G}_X$ be some uncertainty distribution on the inputs X, that is, $X \sim \mathcal{G}_X$. Then this input uncertainty induces a distribution for the output and $Y = F(X)$ would be a random variable following this distribution. If the form of F is known, then the distribution of Y can be derived analytically through a change of variables formula.

In contrast, a stochastic simulator is inherently random by definition, so certainty in the inputs would still result in a distribution for the output Y. Uncertainty in the inputs induces a distribution on distributions for Y. Formally, if $X \sim \mathcal{G}_X$, then $\mathcal{G}_X$ induces a distribution over distributions for the output Y, and $Y|X$ denotes one potential realization from the possible distributions.

1.2 Sensitivity Analysis

Of course, computer models are never run in isolation; researchers often use them to make decisions or inform them about the represented process. Seismologists run earthquake simulations to help them make decisions about building materials or location. Climatologists run climate models to learn how the global climate will reacts to different settings for greenhouse gases or deforestation. In these situations, sensitivity analyses allow researchers to quantify how much their decisions or conclusions are affected by the value of the inputs that are put

into the simulator. In the case of the climate model, the climatologist might be interested in the effect of these human activities on the global mean temperature.

Sensitivity analysis can be done on two scales – local and global. Sensitivity analysis on the local scale looks to determine how small (local) changes in the inputs affect the output at a given location. For example, if a factory manager is using a simulator to identify the optimal parameters steps in his process to maximize efficiency, a local sensitivity study quantifies how small changes near the optimal settings affect the efficiency of the factory. While definitely an interesting problem and applicable in many situations in addition to the one described here, the focus of this dissertation will instead be on sensitivity analysis at the global scale.

Global sensitivity analysis quantifies the impact or relationship of parameters on the outputs across the entire input space. Since global sensitivity analysis is independent of the input values being evaluated, it is a tool for understanding the simulator or underlying process as a whole. Through a global sensitivity study, a climatologist can understand the effect of deforestation on global mean temperature and their conclusions about the climate based on this output.

This global framework for sensitivity analysis is particularly useful when there is uncertainty in the inputs. Specifically, a global sensitivity analysis quantifies the effect of uncertainty in the inputs on the outputs of the simulator. If uncertainty in the input parameters is due to difficulty in measurement, a sensitivity study can identify which parameters have a large influence on the output distribution and would thus be a good utilisation of resources for more accurate measurement. Inputs that have little or no effect on the output distribution could be fixed at the most likely or average value.

The primary resource for global sensitivity analyses are Saltelli et al. (2004) and Saltelli et al. (2008). These books are primarily written to guide non-experts in sensitivity analysies through the process of choosing and implementing the appropriate methods for their problems. However, they contain complete derivations for all the methods described as well as full descriptions to build the necessary intuition required to conduct meaningful analyses.

1.2.1 Sensitivity Analysis for Deterministic Simulators

Global sensitivity analysis for deterministic simulators has been well studied within the last two decades. In the statistics community, one of the preferred methods for conducting these types of analyses is based on the work of Sobol (2001). This method is centered around the principle of variance decomposition for functions. If $F(X) = F(x_1, \ldots, x_p)$ is an integrable function, then it can always be expressed as a sum of orthogonal mean-zero functions.

$$F(X_1, \ldots, x_p) = f_0 + \sum_i F_i(x_i) + \sum_{i<j} F_{ij}(x_i, x_j) + \ldots + F_{1\ldots p}(x_1, \ldots, x_p), \tag{1.1}$$

If the inputs, X, are distributed according to some uncertainty distribution, $\mathcal{G}_X$, then $F(X)$ is now a random variable. If F is also square integrable, then Eq. 1.1 can be used to

decompose the variance of $F(X)$ into

$$Var[F(X)] = \sum_i Var[F_i(X_i)] + \sum_{i<j} Var[F_{ij}(X_i, X_j)] + \ldots + Var[F_{1\ldots p}(X_1, \ldots, X_p)] \quad (1.2)$$

where the variances are integrals over the input distribution $\mathcal{G}_X$.

Written in this from, the $Var[F_i(x_i)]$ quantities could be considered first order sensitivity measures: they represent the proportion of the variability in $F(X)$ attributable to uncertainty in X_i. Similar interpretations exist for the higher order terms, $Var[F_{ij}(X_i, X_j)]$, $Var[F_{i\ldots p}(X_1, \ldots, X_p)]$, etc.

For the most part, these quantities can be estimated through Monte Carlo methods. However, if F is a slow-to-evaluate function or simulator, then it is no longer feasible to take the large samples necessary for Monte Carlo estimates. The most common approach in these situations, first proposed by Oakley and O'Hagan (2004), is to use a statistical surrogate for F that is fast-to-evaluate.

1.2.2 Sensitivity Analysis for Stochastic Simulators

The goal of sensitivity analysis on stochastic simulators is to characterize the relationship between the input parameters and the output distribution. Specifically, we want to quantify how much changing one or more of the inputs changes the distribution of the output – i.e. how *sensitive* the output distribution is to changes in the inputs.

Quantifying the effect of input parameters on the output distribution requires a good way to characterize distributions. A naïve approach would be to summarize the distribution with a mean and a variance. That is, for a given output distribution $Y|X$, we would calculate

$$m(x) = E[Y|X] = \int_{\mathbf{Y}} y f(dy|x)$$

and

$$V(x) = Var[Y|X] = \int_{\mathbf{Y}} (y - m(x))^2 f(dy|x)$$

where f is the conditional density of the output distribution.

The effect of the inputs X on the distribution of the output Y can then be measured by calculating some sensitivity measures on these two functions. In fact, since m and V are deterministic functions, we can apply the previously mentioned methods for deterministic simulators to these functions.

Unfortunately, there are a few problems to this approach. First, sensitivity measures calculated in this manner only quantify the effects of X on the means and variances of the output distribution for Y. Whether or not conclusions drawn from these effects translate back to the output distribution depends entirely on how adequately Y's distribution is characterised by first and second order summaries. For output distributions resembling a normal distribution, this may not be much of a problem. If the form of the output distribution is

unknown or has more complicated behaviors (e.g. asymmetry, multi-modality) or if m and V vary strangely with respect to x, then conclusions about the impact of the inputs on m and V may not necessarily apply to the output distribution.

Additionally, it is not always clear how to simultaneously interpret the results of a sensitivity analyses on these two functions. For example, suppose the simulator has two inputs X_1 and X_2. After you calculate $m(x_1, x_2)$ and $V(x_1, x_2)$ and perform a sensitivity study, you find that X_1 has a large impact on the mean function, but a low impact on the variance. Conversely, X_2 has a low impact on the mean function and a high impact on the variance function. Which of these two factors is more important? Are they equally important? In most cases, the goal of the study may provide some insight on how to answer these questions. Even then, it is often difficult to figure out the best way to simultaneously interpret two different summaries. While having both the mean and variance provides more information about the process than a single number possibly could, situations often arise where just a single summary is desired.

Lastly, computing the two quantities ($V(x)$ in particular) with any certainty requires either replicated evaluations of the simulator at certain inputs or strong assumptions about the form of the output distribution. For fast functions, this required inefficiency may not be a problem. However, if the simulator under study is slow-to-evaluate, then this method suddenly becomes infeasible.

Instead of relying on the mean and variance functions of a stochastic simulator to quantify the relationship between the input parameters and the output distribution, we propose the use of tools from information theory, specifically entropy and mutual information, to quantify said relationship. The entropy of a distribution is a measure of the variability in the possible outcomes. Entropy can also be thought of as a quantification of the amount of uncertainty there is in the value of a random variable. The mutual information between two random variables X and Y quantifies the dependence between the two variables by measuring how much learning the value of one variable reduces the entropy or uncertainty in the other. In the context of stochastic simulators (which may have multiple inputs), the mutual information between an input X_i and Y quantifies how much learning the value of X_i reduces the uncertainty in the value of Y. If an input has a large effect on the output distribution, be it through an effect on the mean or the variance, the mutual information between Y and the input will be high.

1.3 Outline

The material in this dissertation is meant to be self-contained from an implemenation standpoint; all the relevant information needed to implement the methods presented here is included. The rest of the dissertation proceeds as follows.

Chapter 2 introduces the relevant information theoretic tools and definitions that will be used in the proceeding chapters. One new contribution in this chapter is the simulation study comparing different estimators for mutual information.

Chapter 3 introduces a method for Bayesian density regression, utilizing Kernel Stick Breaking Processes, first developed by Dunson and Park (2008). This method will be our main tool for modelling dependence between X and Y. In this chapter, there are novel derivations of the posterior predictive distribution for the model and a new method for illiciting prior parameters based on prior information on the correlation strength.

Chapter 4 presents a method for using the model in Chapter 3 to conduct mutual information-based sensitivity analysis for a stochastic simulator. The majority of this dissertation's novel contributions to the field are contained in this chapter.

Chapter 5 details the application of the method described in Chapter 4 to a stochastic simulator of near fault ground motions.

Finally, in Chapter 6, there will be a brief discussion on the topics covered and potential directions for future work.

Chapter 2: Information Theory

This chapter contains the necessary background on topics in information theory that will be used throughout the rest of the dissertation. It is written to contain all relevant information to understand the applications in the later sections, but with only a cursory overview of less-related topics. For a deeper treatment of the field with more details, a good resource would be Cover and Thomas (2012).

The first section defines and develops the notion of entropy for both discrete and continuous random variables. The next section gives some estimators for entropy as well as some of their properties. The chapter concludes with a section that extends the idea of entropy to mutual information and frames it as a tool for conducting sensitivity analysis. That last section also contains methods for estimating the mutual information as well as a simulation study comparing the different estimators.

2.1 Definitions

Although the methods in this dissertation are primarily based on differential entropy, the discussion on entropy and information theory in this chapter begins with Shannon entropy in hopes of developing improved understanding and intuition of entropy as a whole.

2.1.1 Shannon Entropy

Let X be a random variable with probability mass function p on some discrete space $\mathbf{X}$. The *entropy* of X, as first presented in Shannon and Weaver (1948), is defined as

$$H(X) = -\sum_{x \in \mathbf{X}} log\left\{p(x)\right\} p(x). \tag{2.1}$$

The unit of entropy depends on the base of the log being used. The two most common are base 2, which would give bits, and base e, which gives nats. Unless otherwise denoted, all the logs used in this dissertation will be base e logs. Note that the entropy is not affected by the values that X could take on, it depends solely on the distribution p. For this reason, in certain fields the entropy may sometimes be written as $H(p)$. For the most part, this type of notation will be avoided throughout this dissertation.

Notice that the Shannon entropy is always nonnegative. This is an immediate result due to the fact that since X is discrete, $0 \leq p(x) \leq 1$ and hence $-log\{p(x)\} \geq 0$, with the inequality being strict unless X is a constant.

There are many common interpretations of entropy, with the predominant one being dependent on the field of study. However, all interpretations boil down to entropy being a measure of the uncertainty or randomness in X. That is, if the entropy of X is large, then the value of realizations are more unpredictable. In the special case that X is fixed, then it has zero entropy.

This notion of entropy as a measure of randomness naturally leads to a popular interpretation of entropy amongst computer scientists - the entropy is the expected or average number of bits or nats (depending on the base of the logarithm) required to describe the outcome of the random variable. This interpretation also arises directly when the entropy is written as the following expectation

$$H(X) = -\sum_{x \in \mathbf{X}} log\left\{p(x)\right\} p(x) = -E\left[log\left\{p(X)\right\}\right]. \tag{2.2}$$

To see how this expression corresponds to the bit-average, consider the case where $\mathbf{X}$ is a set with n components. In order to uniquely identify the different elements in $\mathbf{X}$, it would take $log(n)$ bits or nats (depending on the base) to represent them.

Example: Coin tossing

Let X be the outcome of a coin toss, with the chance of a heads being p. That is,

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

From our definition of entropy, we see

$$H(X) = -plog(p) - (1-p)log(1-p) \tag{2.3}$$

If $p = \frac{1}{2}$ and the coin is fair, then $H(X) = 1$ (when the log is base 2). As the coin gets biased (towards heads when p increases or towards tails when p decreases), then the entropy decreases. Biasing the coin towards an outcome decreases the uncertainty because an outcome is more likely. Eventually, at the ends of the spectrum where $p = 0$ or $p = 1$, the coin always has the same outcome and there is no uncertainty and hence the entropy will be 0.

2.1.2 Joint and Conditional Entropy

There is nothing preventing us from generalizing our definition of entropy for a single random variable directly to pairs of random variables (or more). Let (X, Y) be a pair of discrete

random variables on $\mathbf{X}$ and $\mathbf{Y}$ with joint distribution $p(x, y)$. Then the *joint entropy* of (X, Y) is

$$H(X,Y) = -\sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(x,y)\}p(x,y), \tag{2.4}$$

or as an expectation,

$$H(X,Y) = E\left[log\{p(X,Y)\}\right]. \tag{2.5}$$

Because the generalization of entropy to joint entropy is so direct, all the interpretations of entropy also translate straightforwardly to joint entropy as well. Additionally, when working with two random variables, we can also define the *conditional entropy*

$$H(Y|X) = -\sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(y|x)\}p(x,y) \tag{2.6}$$

$$= \sum_{x\in\mathbf{X}} p_x(x)\sum_{y\in\mathbf{Y}} log\{p(y|x)\}p(y|x), \tag{2.7}$$

where $p_x(x) = \sum_{y\in\mathbf{Y}} p(x,y)$ is the marginal distribution for x, and $p(y|x) = \frac{p(x,y)}{p_x(x)}$ is the conditional distribution for y given x. As an expectation, the conditional entropy based on Eq. 2.6 is

$$H(Y|X) = -E\left[log\left\{p(Y|X)\right\}\right]. \tag{2.8}$$

When using Eq 2.7, the conditional entropy as an expectation is

$$H(Y|X) = -E_X\left[E_{Y|X}\left[log\left\{p(Y|X)\right\}\right]\right]. \tag{2.9}$$

One useful result that follows conveniently from these natural definitions for joint and conditional entropy is the following expression, often referred to as the *chain rule*,

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned} \tag{2.10}$$

In words, the joint entropy for a pair of random variables is equal to the sum of the entropy of one of the variables and the conditional entropy for the other. Formally, the chain rule is true because

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(x,y)\}p(x,y) = -\sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p_x(x)p(y|x)\}p(x,y) \\ &= -\sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p_x(x)\}p(x,y) - \sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(y|x)\}p(x,y) \\ &= -\sum_{x\in\mathbf{X}} log\{p_x(x)\}\sum_{y\in\mathbf{Y}} p(x,y) - \sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(y|x)\}p(x,y) \\ &= -\sum_{x\in\mathbf{X}} log\{p_x(x)\}p_x(x) - \sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{Y}} log\{p(y|x)\}p(x,y) \\ &= H(X) + H(Y|X) \end{aligned}$$

Alternatively, from a probabilistic point of view, we know that $p(x, y) = p_x(x)p(y|x)$ and hence $log\{p(x, y)\} = log\{p_x(x)\} + log\{p(y|x)\}$. Since expectation is linear, this means

$$\begin{aligned} H(X, Y) &= -E\left[log\{p(X, Y)\}\right] \\ &= -E\left[log\{p_x(X)\}\right] - E\left[log\{p(Y|X)\}\right] \\ &= H(X) + H(Y|X) \end{aligned}$$

One result immediately clear from either of the chain rule derivations is that if X and Y are independent, then the joint entropy is just the sum of the two marginal entropies:

$$H(X, Y) = H(X) + H(Y). \tag{2.11}$$

This result shows up because X and Y being independent means $p(y|x) = p(y)$.

Example: Two Dependent Events

Let X be the outcome of one roll of a six sided die. Given the value of X, Y is the outcome of a fair coin toss if X is even. If X is odd, then Y is the outcome of a biased coin, with probability $p = .75$ of landing heads. Marginally, the outcome of Y is equivalent to tossing a coin with probability $p = 0.625$ of landing heads, so the entropy, using Eq. 2.3, is

$$H(Y) = -0.625 \cdot log(0.625) - (1 - 0.625) \cdot log(1 - 0.625) \approx 0.662 \text{ nats.}$$

The marginal distribution for the die roll X is

k	1	2	3	4	5	6
$P(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

so the entropy for X is $H(X) = -log\left(\frac{1}{6}\right) \approx 1.792$ nats. Since the relationship between X and Y is defined in terms of $Y|X$, we can calculate $H(Y|X)$ directly using Eq. 2.3 and 2.7

$$\begin{aligned} H(Y|X) &= -\frac{1}{2} \cdot [.5 \cdot log(.5) + (1 - .5) \cdot log(1 - .5)] - \frac{1}{2} \cdot [.75 \cdot log(.75) + (1 - .75)log(1 - .75)] \\ &\approx 0.628 \text{ nats.} \end{aligned}$$

For the joint entropy and the conditional entropy $H(X|Y)$, we'll need the joint distribution of X and Y:

Y \ X	1	2	3	4	5	6
1	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{12}$
0	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$

From the joint distribution table, we can calculate both the the joint entropy and the conditional entropy $H(X|Y)$ using Eq 2.6. First, the conditional entropy

$$H(X|Y) = -\sum_{k=1}^{6} log\{P(X=k|Y=0)\}P(X=k,Y=0) - \sum_{k=1}^{6} log\{P(X=k|Y=1)\}P(X=k,Y=1)$$
$$= -3\cdot log\left(\frac{1}{9}\right)\frac{1}{24} - 3\cdot log\left(\frac{2}{9}\right)\frac{1}{12} - 3\cdot log\left(\frac{1}{5}\right)\frac{1}{8} - 3\cdot log\left(\frac{2}{15}\right)\frac{1}{12} \approx 1.758 \text{ nats.}$$

Now, the joint entropy

$$H(X,Y) = 3\cdot log\left(\frac{1}{8}\right)\frac{1}{8} + 6\cdot log\left(\frac{1}{12}\right)\frac{1}{12} + 3\cdot log\left(\frac{1}{24}\right)\frac{1}{24} \approx 2.42 \text{ nats.}$$

Notice that both $H(Y)+H(X|Y) = 0.662+1.758 = 2.42 = H(X,Y)$ and $H(X)+H(Y|X) = 1.792+0.628 = 2.42 = H(X,Y)$, as suggested by the chain rule. Additionally, $H(Y|X) \neq H(X|Y)$, so conditional entropy is *not* symmetric.

2.1.3 Differential Entropy

In this section, we'll describe how one generalizes the concept of Shannon entropy for discrete distributions and random variables to continuous distributions and random variables. For the most part, summations and probability mass functions are replaced with integrals and probability density functions, respectively. However, there are subtle distinctions between the two that can be important and care will be taken to point them out as these discrepancies arise.

Let X be a random variable on some space $\mathbf{X}$ with probability density function (if it exists) $f(x)$. Then the *differential entropy* for X is:

$$H(X) = -\int_{\mathbf{X}} log\left\{f(x)\right\} f(x)dx, \tag{2.12}$$

which is essentially identical to the definition for discrete entropy with an integral replacing the summation. As in the discrete case, the differential entropy can be written as the expectation

$$H(X) = -E\left[log\{f(x)\}\right]. \tag{2.13}$$

Example: Uniform distribution

If X has uniform distribution on the interval (a,b), then the density of X is $f(x) = \frac{1}{b-a}$. Then the differential entropy is

$$H(X) = -\int_{(a,b)} log\{f(x)\}f(x)dx = -\int_{(a,b)} log\left\{\frac{1}{b-a}\right\}\frac{1}{b-a}dx$$
$$= log\left\{b-a\right\}\int_{(a,b)}\frac{1}{b-a}dx = log\left\{b-a\right\}. \tag{2.14}$$

So the differential entropy for a uniform distribution is equal to the log of the length of the interval. If the interval has length one, then the differential entropy will be zero (regardless of the base of the log, in fact). If the length of the interval is less than one, then the differential entropy will be negative. This is in stark contrast to Shannon entropy, which is always nonnegative and is only zero for constants.

The definitions of joint entropy and conditional entropy for pairs (or more) of discrete random variables also translate straightforwardly to continuous random variables. If X and Y are two random variables with joint density function $f(x, y)$, then the *joint differential entropy* for X and Y is

$$H(X,Y) = -\int_{\mathbf{X}}\int_{\mathbf{Y}} log\{f(x,y)\}f(x,y)dxdy. \tag{2.15}$$

As an expectation, the joint differential entropy is

$$H(X,Y) = -E\left[log\{f(X,Y)\}\right] \tag{2.16}$$

Example: Multivariate normal distribution

Let $(X_1, X_2, \ldots, X_n)$ be a realization from a multivariate normal distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$. The joint density for $\mathbf{X} = (X_1, X_2, \ldots, X_n)$

$$f(x_1, \ldots, x_n) = (2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu\right)^T\mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu\right)\right\}, \tag{2.17}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. Then the joint differential entropy for $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned}
H(X_1, \ldots, X_n) &= -\int_{\mathcal{X}} log\left\{f(\mathbf{x})\right\}f(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{2}log\left\{(2\pi)^n|\mathbf{\Sigma}|\right\} + \frac{1}{2}\int_{\mathcal{X}}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)f(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{2}log\left\{(2\pi)^n|\mathbf{\Sigma}|\right\} + \frac{1}{2}E\left[(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right] \\
&= \frac{1}{2}log\left\{(2\pi)^n|\mathbf{\Sigma}|\right\} + \frac{1}{2}Tr\left\{\mathbf{\Sigma}^{-1}Cov(\mathbf{x}-\mu)\right\} + E[\mathbf{x}-\mu]^T\mathbf{\Sigma}^{-1}E[\mathbf{x}-\mu] \\
&= \frac{1}{2}log\left\{(2\pi)^n|\mathbf{\Sigma}|\right\} + \frac{1}{2}Tr\left\{\mathbf{I}_n\right\} \\
&= \frac{1}{2}log\left\{(2\pi e)^n|\mathbf{\Sigma}|\right\}
\end{aligned} \tag{2.18}$$

Here, $\mathcal{X}$ denotes $\mathbb{R}^n$, which is the support of $\mathbf{X} = (X_1, \ldots, X_n)$.

The last definition to generalize to continuous random variables is the conditional entropy. If we let

$$f_X(x) = \int_{\mathbf{Y}} f(x,y)dy$$

denote the marignal density for X and

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

denote the conditional density of Y given X (if it exists), then the *conditional differential entropy* of Y given X is

$$H(Y|X) = -\int_{\mathbf{X}}\int_{\mathbf{Y}} log\left\{f_{Y|X}(y|x)\right\} f(x,y)dydx. \tag{2.19}$$

As an expectation, the conditional differential entropy is

$$H(Y|X) = -E\left[log\left\{f_{Y|X}(Y|X)\right\}\right] \tag{2.20}$$

Just like in the discrete case, we can write both the integral and expectation forms of conditional differential entropy in terms of the marginal distribution

$$\begin{aligned} H(Y|X) &= -\int_{\mathbf{X}}\left[\int_{\mathbf{Y}} log\left\{f_{Y|X}(y|x)\right\} f(y|x)dy\right] f_X(x)dx \\ &= -E_X\left[E_{Y|X}[log\left\{f_{Y|X}(Y|X)\right\}]\right]. \end{aligned} \tag{2.21}$$

Example: A Bayesian Distribution

Let X be a normal random variable with mean μ and variance σ^2. Given the value of X, Y is also a normal random variable, but with mean X and variance τ^2. Formally:

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \\ Y|X &\sim N(X, \tau^2). \end{aligned}$$

First, the entropy of X

$$\begin{aligned} H(X) &= -\int_{\mathbf{X}} log\left\{\left(2\pi\sigma^2\right)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\}\right\}\left(2\pi\sigma^2\right)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\}dx \\ &= \frac{1}{2}log\left\{2\pi\sigma^2\right\} + \frac{1}{2\sigma^2}\int_{\mathbf{X}}\left(x-\mu\right)^2\left(2\pi\sigma^2\right)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\}dx \\ &= \frac{1}{2}log\left\{2\pi\sigma^2\right\} + \frac{1}{2\sigma^2}\sigma^2 \\ &= \frac{1}{2}log\left\{2\pi e\sigma^2\right\} \text{ nats.} \end{aligned} \tag{2.22}$$

Note that the equality in the last line is due to the preceding integral being $E\left[(X-\mu)^2\right] = \sigma^2$.

To get the entropy of Y, a quick glance at the form of the integral for the marginal distribution

$$f_Y(y) = \int_{\mathbf{X}} f_{Y|X}(y|x)f_X(x)dx \propto \int_{\mathbf{X}} exp\left\{-\frac{1}{2\tau^2}(y-x)^2\right\} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\}dx$$

reveals that Y will have a Gaussian marginal distribution, so we only have to calculate the mean and the variance of Y to know the marginal distribution exactly. The expectation is

$$E_Y[Y] = E_X\left[E_{Y|X}[Y|X]\right] = E_X[X] = \mu$$

and the variance is

$$Var_Y[Y] = Var_X\left[E_{Y|X}[Y|X]\right] + E_X\left[Var_{Y|X}[Y|X]\right] = Var_X[X] + E_X[\tau^2] = \sigma^2 + \tau^2. \tag{2.23}$$

So marginally Y has distribution $N(\mu, \sigma^2 + \tau^2)$, and hence (through the use of Eq. 2.22) $H(Y) = \frac{1}{2}log\left\{2\pi e\left(\sigma^2 + \tau^2\right)\right\}$ nats.

Since the dependence between Y and X is defined conditionally through $Y|X$, the conditional entropy $H(Y|X)$ is straightforward to calculate using Eq. 2.21:

$$H(Y|X) = -E_X\left[E_{Y|X}\left[log\left\{f_{Y|X}(Y|X)\right\}\right]\right] = E_X\left[\frac{1}{2}log\left\{2\pi e\tau^2\right\}\right] = \frac{1}{2}log\left\{2\pi e\tau^2\right\} \text{ nats.} \tag{2.24}$$

To get the joint entropy, instead of computing the integral

$$H(X,Y) = \int_{\mathbf{X}}\int_{\mathbf{Y}} log\left\{f_X(x)f_{Y|X}(y,x)\right\} f_X(x)f_{Y|X}(y,x)dydx,$$

we first identify the joint distribution:

$$f(x,y) = f_X(x)f_{Y|X}(y,x) \propto exp\left\{-\frac{1}{2\tau^2}(y-x)^2\right\} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\},$$

which is some kind of multivariate Normal distribution. From above, we know that the marginal means and variances of X and Y are (μ, σ^2) and $(\mu, \sigma^2 + \tau^2)$, respectively. So the only missing component is the $Cov[X,Y]$, the off-diagonal terms in the covariance matrix for (X,Y). To calculate this quantity, first realize that

$$\begin{aligned} Var[X+Y] &= Var_X\left[E_{Y|X}[X+Y|X]\right] + E_X\left[Var_{Y|X}[X+Y|X]\right] \\ &= Var_X[2X] + E_X[\tau^2] = 4\sigma^2 + \tau^2. \end{aligned}$$

This means

$$4\sigma^2 + \tau^2 = Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y] = \sigma^2 + \sigma^2 + \tau^2 + 2Cov[X,Y]$$

and consequently $Cov[X,Y] = \sigma^2$. This means (X,Y) has bivariate normal distribution with mean vector (μ, μ) and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \tau^2 \end{bmatrix}.$$

From Eq 2.18, the joint entropy of (X, Y) is

$$H(X,Y) = \frac{1}{2} log\left\{(2\pi e)^2|\mathbf{\Sigma}|\right\} = \frac{1}{2} log\left\{(2\pi e)^2(\sigma^4 + \sigma^2\tau^2 - \sigma^4)\right\} = \frac{1}{2} log\left\{(2\pi e)^2\sigma^2\tau^2\right\} \text{ nats,}$$

which is exactly

$$H(X) + H(Y|X) = \frac{1}{2} log\left\{2\pi e\sigma^2\right\} + \frac{1}{2} log\left\{2\pi e\tau^2\right\} = \frac{1}{2} log\left\{(2\pi e)^2\sigma^2\tau^2\right\} = H(X,Y).$$

Framed from a Bayesian point of view, $Y|X \sim N(X, \tau^2)$ would be the sampling distribution of the data with X being a parameter of this distribution. Then $N(\mu, \sigma^2)$ would be the prior distribution for X. Of course, thinking of the distributions in this manner does not alter any of the calculations; it only offers a different interpretation. From this point of view, the conditional entropy, $H(Y|X)$ is the amount of uncertainty or randomness inherent to the process generating Y, i.e. after accounting for our prior uncertainty about the parameter X. In this case, it is entirely based on τ^2, the conditional variance of Y.

As we can see from this example, it turns out that since all of our definitions for entropy, both joint and conditional, were generalized in the same way, without any significant modification, the chain rule still applies to differential entropy. The formal derivation of the chain rule for differential entropy mirrors the discrete case exactly and is consequently omitted.

One recurring property of entropy that has not been explicitly mentioned but has shown up in all three examples in this section is the translation invariance of entropy. In the multivariate normal example, the entropy, $\frac{1}{2} log\left\{(2\pi e)^n|\mathbf{\Sigma}|\right\}$ is independent of the mean vector μ. For the uniform example, shifting the support of the distribution left or right doesn't change the width of the interval and consequently won't affect the entropy of the distribution. In the Bayesian distribution example, the mean μ did not show up in any of the quantities calculated: $H(X), H(Y), H(Y|X)$, or $H(X,Y)$. This property of entropy (both Shannon and differential) is a result of entropy being a functional of the density (or distribution for the discrete case) and not on the specific values of the random variable under study.

2.2 Entropy Estimation

Before moving on to mutual information, which will be our primary information theoretic tool when working with stochastic simulators, this section surveys some popular methods for estimating the entropy of a distribution from a sample of points. For the most part, this section will mirror the approach taken by Beirlant et al. (2001), although our primary goal will be to estimate mutual information, not the entropy, so some of the less relevant details will be omitted.

Let $X_1, \ldots X_n$ be a sample of realizations from some distribution f. For the purposes of this section, we will focus primarily on continuous random variables, so f is a density. The

goal of entropy estimation is to estimate the quantity

$$H(X) = -\int log\{f(x)\}f(x)dx, \tag{2.25}$$

where X is a random variable, also with distribution f. We use $\hat{H}_n(X)$ to denote an estimate of $H(X)$ based on the sample of realizations $X_1, \ldots, X_n$. This section is primarily interested in nonparametric methods for estimating $H(X)$, which are typically used when little is known about the underlying process generating X. A parametric entropy estimator requires the assumption of some parametric form $f(x|\theta)$ for f. An estimate $\hat{\theta}_n$ for θ is computed based on $X_1, \ldots, X_n$ that can be plugged into the integral to estimate the entropy

$$\hat{H}_n(X) = -\int log\left\{f(x|\hat{\theta}_n)\right\} f(x|\hat{\theta}_n)dx \approx -\int log\left\{f(x|\theta)\right\} f(x|\theta)dx = H(X).$$

As is the case when comparing any parametric and nonparametric methods, parametric estimators of entropy typically perform a bit better than nonparametric estimators for a given sample size. Unfortunately, these methods require some prior knoweldge about the form of f in order to use the correct parametrization; after all, it is these assumptions we make about the form of the data that allows parametric estimators to be more accurate with smaller samples. If these assumptions are not true, then the performance is not necessarily better and often lead to worse estimates. Often, in the context of computer models, very little is known about the form of f which limits the usefulness of these methods. Consequently, the emphasis of this section will be on nonparametric entropy estimators, despite the need for larger sample sizes to perform as well.

Before discussing the different types of estimators, here are some desirable convergence properties for $\hat{H}_n(X)$:

Weak consistency: $\lim_{n\to\infty} \hat{H}_n(X) = H(X)$ in probability.

Mean square consistency: $\lim_{n\to\infty} E\left[(\hat{H}_n(X) - H(X))^2\right] = 0.$

Strong consistency: $\lim_{n\to\infty} \hat{H}_n(X) = H(X)$ almost surely.

Root-n Asymptotic normality: $n^{1/2}(\hat{H}_n(X) - H(X)) \Rightarrow N(0, \sigma^2)$.

L_2 rate of convergence: $\lim_{n\to\infty} nE\left[(\hat{H}_n(X) - H(X))^2\right] = \sigma^2.$

For both Root-n asymptotic normality and L_2 rate of convergence, σ^2 is some calculable quantity that depends on both the density f and the estimator used. If an estimator mentioned possesses any of these properties, the property along with an idea of any necessary conditions will be highlighted.

2.2.1 Integral Estimates

Perhaps the most natural estimate of entropy one might think up would be to replace the density in the integrand with an estimate generated from the data. Let $\hat{f}_n(x)$ be an estimate of the density f based on the sample $X_1, \ldots, X_n$. Then an *integral estimate* for $H(X)$, first introduced in Dmitriev and Tarasenko (1974), has the form

$$\hat{H}_n(X) = -\int_{\mathbf{A}} \hat{f}_n(x) log\left\{\hat{f}_n(x)\right\} dx \tag{2.26}$$

where $\mathbf{A}$ is usually some set that excludes tail or small values of $\hat{f}_n$ in order to prevent the *log* in the integrand from producing large, negative values. The original publication showed, in one dimension, if $\mathbf{A} = [-b_n, b_n]$ and $\hat{f}_n$ is a kernel density estimator, then H_n is strong consistent. When f_n is a kernel density estimator, the integration usually must be done numerically, which makes this type of estimation difficult for anything higher than two dimensions.

If $\hat{f}_n$ is the histogram estimator, then the integral is easy to compute. Györfi and Van der Meulen (1987) showed that for general dimension and taking $\mathbf{A} = \{x : f_n(x) \geq a_n\}$ with $0 < a_n \to 0$, the histogram based integral estimate is strong consistent under a mild condition on the tails of X.

The main reason these types of estimators are not used more often is because of the need to come up with reasonable domains of integration, $\mathbf{A}$, while allowing the estimate to still have the desired convergence properties. Additionally, in higher dimensions, histogram estimators necessarily require dense samples in order to resemble the true density.

2.2.2 Resubstitution Estimates

The next estimator introduced builds on the idea behind the integral estimate. A *resubstitution estimate* for entropy has the form

$$\hat{H}_n(X) = -\frac{1}{n}\sum_{k=1}^{n} log\left\{\hat{f}_n(X_k)\right\} \tag{2.27}$$

where $\hat{f}_n$ is some estimate of f based on the sample $X_1, \ldots, X_n$. This type of estimator arises from the layering of two approximations: (1) a Monte Carlo approximation to the entropy integral and (2) an approximation to the density.

Formally, the Monte Carlo integral approximation is

$$-\int_{\mathbf{X}} log\left\{f(x)\right\} f(dx) \approx -\frac{1}{n}\sum_{k=1}^{n} log\left\{f(X_k)\right\} \tag{2.28}$$

since the $X_1, \ldots, X_n$ are distributed according to f. As with any Monte Carlo integral, this approximation can also be viewed as an application of the law of large numbers:

$$H(X) = -E\left[log\left\{f(x)\right\}\right] \approx -\frac{1}{n}\sum_{k=1}^{n} log\left\{f(X_k)\right\}$$

Monte Carlo integral approximations are known to be good in multiple dimensions for even modestly sized samples, with the specific level of accuracy for a given sample size being dependent on the true form of the integrand, $log\left\{f(x)\right\}$ (Press et al. 2007).

The density approximation of f in the estimate is

$$-\frac{1}{n}\sum_{k=1}^{n} log\left\{f(X_k)\right\} \approx -\frac{1}{n}\sum_{k=1}^{n} log\left\{\hat{f}_n(X_k)\right\}. \tag{2.29}$$

Two potential choices for $\hat{f}_n$ are kernel density estimators (KDE) or histogram estimators. The accuracy of this approximation depends primarily on how well $\hat{f}_n$ approximates the true density f. Unless noted otherwise, we will be using KDEs for $\hat{f}_n$ whenever a resubstitution type estimator is mentioned.

Some factors affecting how closely $\hat{f}_n$ resembles f, and subsequently the approximation accuracy of Eq. 2.29, are (i) the true form of f, (ii) the dimension of X, (iii) the type of estimator being used, and (iv) the sample size, n. However, the limiting factor for the approximation (for a given dimensionality of X) will almost always be the sample size, n. For any reasonable choice of density estimators, $\hat{f}_n$ should be fairly close to the true density as long as n is large enough. Typically, the size of sample required to produce a good approximation in Eq. 2.29 will also produce an accurate approximation in Eq. 2.28. Heuristically, if the integrand in Eq. 2.28 behaves poorly and requires an atypically large sample to integrate accurately, then $\hat{f}_n$ will also require a large sample before starting to resemble f.

Resubstitution estimators possess many desirable convergence properties. Ahmad and Lin (1976) showed, under some mild conditions, if $\hat{f}_n$ is a kernel density estimate, then $\hat{H}(X)$ is mean square consistent. Hall and Morton (1993) showed that, under certain tail and smoothness conditions, $\hat{H}(Y)$ is root-n asymptotically normal with a calculable variance for both kernel density and histogram estimators. It is these reasonably general convergence properties, along with the fast evaluation speed (the averaging of logs is a fast operation), that makes the resubstitution estimate with a KDE one of the estimators used throughout the rest of this chapter.

Since we'll be using this type of estimator again in this chapter, we'll take a moment to present a resubstitution estimate for the conditional entropy. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from some joint density $f(x, y)$ and let (X, Y) be random variables also with joint density $f(x, y)$. Then a resubstitution estimate of the conditional entropy $H(Y|X)$ would be

$$\hat{H}_n(Y|X) = -\sum_{i} log\left\{\hat{f}_n(y_i|x_i)\right\}, \tag{2.30}$$

where $\hat{f}_n(y|x)$ is a KDE of $f(y|x)$ based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. One more advantage that resubstitution estimates have over integral estimates, despite being further removed from the desired integral, appears in Eq. 2.30. An equivalent integral estimator for the conditional entropy would be

$$\tilde{H}_n(Y|X) = -\int_{\mathbf{X}} \int_{\mathbf{Y}} log\left\{\hat{f}_n(y|x)\right\} \hat{f}_n(y,x)dydx,$$

where $\hat{f}_n(y,x)$ is a KDE of the joint density $f(y,x)$ based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and, as before, $\hat{f}_n(y|x)$ is a kernel density estimate of the conditional density $f_{Y|X}(y,x)$. The integral estimate requires estimates of two densities: the joint density $f(x,y)$ and the conditional density $f_{Y|X}(y|x)$. Since they are estimated from the same data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, the correlation between the two estimated densities makes estimation even more difficult.

2.2.3 Nearest Neighbor Estimates

Before presenting the main estimator of this type, we'll provide some insight into the intuition behind the derivation of it. For a complete derivation as well as details on some properties of this estimator, see Kozachenko and Leonenko (1987).

Recall that the entropy $H(X)$ is the integral

$$H(X) = -\int log\left\{f(x)\right\} f(x)dx,$$

which depends primarily on the density f. Not having much prior knowledge about f precludes the use of any parametric method, which is why we're looking at these nonparametric methods. The previously introduced resubstitution and integral estimates for entropy rely on the estimation of f through some means - either a kernel density estimate or a histogram estimate. The efficacy of those methods depends largely on f being relatively easy to estimate. However, a nearest neighbor based entropy estimate attempts to circumvent this problem altogether by using the distribution of the k-th nearest neighbor distance as a proxy or surrogate (note the distinction from estimate) to the density.

To get some intuition as to why one could possibly use the k-th nearest neighbor density in this way, consider a sample $X_1, \ldots, X_n$ with distribution f. In an area where the density f is low, the sample will contain fewer points in that area and consequently the distribution of the distance to the k-th nearest neighbor will assign more mass to longer distances in that region (regardless of the value of k). Conversely, if f is high in a region, the sample will be denser in that area and hence the distribution of the distance to the k-th nearest neighbor will tend to shorter distances in that region. A formal derivation of this correspondence between the density f and the k-th nearest neighbor distance distribution is available in **?**. Given that there exists a relationship between the two distributions, the k-th nearest neighbor entropy estimate is

$$\hat{H}_n(X) = -\psi(k) + \psi(n) + \frac{p}{n}\sum_{i=1}^{n} log\left\{\epsilon_{i,k}\right\}, \tag{2.31}$$

where ψ is the digamma function,

$$\psi(x) = \frac{d}{dx} log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)},$$

p is the dimension of the input space $\mathbf{X}$, and $\epsilon_{i,k}$ is the distance from x_i to it's k-th nearest neighbor. Here, the term distance refers to the supremum norm i.e.

$$d(\mathbf{x}, \mathbf{y}) = max_j \{x_1 - y_1, \ldots, x_p - y_p\}$$

Under a general distance, the estimator is

$$\hat{H}_n(X) = -\psi(k) + \psi(n) + log(c_p) + \frac{p}{n} \sum_{i=1}^{n} log \{\epsilon_{i,k}\}$$

where c_p is the volume of the unit ball under the given distance. We choose to use the supremum norm because the analogous mutual information estimator in Section 2.3.2 uses the supremum norm. For the purposes of entropy estimation, no distance should have an advantage for a general density f.

Note that, for a fixed sample, the choice of k will change the estimated value. Typically, though, k should be small since the link between the density and the distribution of k-th nearest neighbor distances only holds for small neighborhoods. However, taking k too small, like $k = 1$, may cause numerical issues when computing the estimate for large samples ; in extreme cases, it's possible for two sample points to be so close that they are numerically distance 0 apart.

One nice computational property of this estimate is that the slowest operation in the estimate is the computation of the k-th nearest neighbor distances. Luckily, finding k-th nearest neighbors is a well studied problem in the computer sciences and, through the use of data structures such as a kd-tree, is a very fast operation. Searching a kd-tree for the k-th nearest neighbor is at worse an $O(n)$ operation and an $O(logn)$ operation on average. This is orders of magnitude faster than the computation of a KDE to use in a resubstitution or integral estimate.

In addition to the low computational complexity, nearest neighbor estimates of entropy also possess some useful convergence properties. The original authors, Kozachenko and Leonenko (1987), showed that Eq 2.31 is a mean square consistent for general distance functions. Bickel and Breiman (1983) showed that specific estimates for general functionals of a density can be root-n asymptotically normal. The class of functionals does not include the entropy, but their results suggest that there is hope for root-n asymptotic normality for this entropy estimate.

2.3 Mutual Information and Sensitivity Analysis

In this section, we introduce the concept of mutual information, a quantity based on entropy, and demonstrate how to use it to conduct sensitivity analysis. We also look at methods of

estimating the mutual information given a sample of data. The section concludes with two simulation studies comparing the different methods in order to choose the estimate used in the proceeding chapters.

2.3.1 Definition

We begin with the definition of mutual information. Unlike the way that we worked up to differential entropy for the continuous case by starting with Shannon entropy for discrete distributions, mutual information will be presented entirely from the continuous case since our focus will be on working with continuous distributions. The development for the discrete case is identical in almost every aspect and has consequently been omitted.

Let X and Y be two random variables with joint density $f(x, y)$. Then the *mutual information* between X and Y is

$$I(X,Y) = \int_{\mathbf{X}} \int_{\mathbf{Y}} log \left\{ \frac{f(x,y)}{f_X(x) f_Y(y)} \right\} f(x,y) dy dx, \tag{2.32}$$

where $f_X(x) = \int_{\mathbf{Y}} f(x,y)dy$ and $f_Y(y) = \int_{\mathbf{X}} f(x,y)dx$ are the marginal densities for X and Y, respectively. While Eq. 2.32 is the formal definition for mutual information, in that form it's not immediately clear what the mutual information represents or why such a quantity might be useful. (*Traditional literature on information theory and mutual information present Eq. 2.32 in the context of KL-divergence - a topic that we've omitted entirely. Under that setting, this form of mutual information is actually quite revealing.*)

Luckily, there are other expressions for mutual information which are more illuminating given the topics covered in the preceding sections. First, by manipulating the *log* in the integrand, we can write mutual information in terms of entropies:

$$\begin{aligned} I(X,Y) &= \int_{\mathbf{X}} \int_{\mathbf{Y}} log \left\{ \frac{f(x,y)}{f_X(x) f_Y(y)} \right\} f(x,y) dx dy \\ &= -\int_{\mathbf{X}} \int_{\mathbf{Y}} log \left\{ f_X(x) f_Y(y) \right\} f(x,y) dx dy + \int_{\mathbf{X}} \int_{\mathbf{Y}} log \left\{ f(x,y) \right\} f(x,y) dx dy \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{2.33}$$

Written in this manner, it should be clear that the mutual information is symmetric. That is, $I(X,Y) = I(Y,X)$.

Recall that if X and Y are two independent random variables, then the joint entropy of X and Y is just $H(X) + H(Y)$, the sum of the two marginal entropies. With this in mind, the mutual information is the difference in entropy between the product of the marginal distributions for X and Y (i.e. under independence) and the joint distribution of X and Y. In a sense, mutual information measures the amount of dependence present between X and Y. (Here, the term *measure* is used loosely and not in the formal mathematical sense.) A low mutual information means that knowing X reveals little about the value of Y while a high mutual information means knowing X reveals a lot about the value of Y.

We can further manipulate the expression for mutual information in Eq. 2.33 by applying the chain rule to the joint entropy

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) = H(X) + H(Y) - (H(Y) + H(X|Y)) \\ &= H(X) - H(X|Y). \end{aligned} \tag{2.34}$$

Applying the chain rule in the other direction gives an equivalent version $H(Y) - H(Y|X)$. In both versions, the mutual information is the reduction in entropy of one variable due to knowing the other variable. That is, $H(Y) - H(Y|X)$ is the reduction in the randomness or uncertainty in Y caused by knowing the value of X. This coincides with the above interpretation of mutual information as a quantification of the dependence between X and Y - if knowing the value of X reduces the uncertainty in Y by a large amount, then there must be strong dependence between X and Y. Conversely, if knowing the value of X does not reduce the randomness in Y by very much, then there the dependence between X and Y must be weak.

Example: Mutual Information in a Bayesian Setting

Returning to the example at the end of Section 2.1.3, recall in that example, X has a normal distribution with mean μ and variance σ^2 and $Y|X$ follows a normal distribution with mean X and variance τ^2. In that section, we calculated $H(Y) = \frac{1}{2} log\left\{2\pi e\left(\sigma^2 + \tau^2\right)\right\}$ nats and $H(Y|X) = \frac{1}{2} log\left\{2\pi e\tau^2\right\}$ nats. Using the Y-centric version of 2.34 gives the mutual information of Y and X to be

$$\begin{aligned} I(X,Y) = H(Y) - H(Y|X) &= \frac{1}{2} log\left\{2\pi e\left(\sigma^2 + \tau^2\right)\right\} - \frac{1}{2} log\left\{2\pi e\tau^2\right\} \\ &= \frac{1}{2} log\left\{\frac{\sigma^2 + \tau^2}{\tau^2}\right\} \text{ nats.} \end{aligned} \tag{2.35}$$

If the variance in X, σ^2, is fixed, increasing the value of τ^2, the conditional variance of Y, decreases the mutual information. As τ^2 approaches infinity, the mutual information approaches 0. These behaviors should not be surprising since increasing the marginal variance should decrease the effect of our uncertainty in the mean, X.

Alternatively, for fixed τ^2, increasing the value of the prior variance for the mean, σ^2 will increase the mutual information. Increasing the variance of the mean increases the marginal randomness in Y, since $H(Y) = \frac{1}{2} log\left\{2\pi e\left(\sigma^2 + \tau^2\right)\right\}$, which consequently increases the importance or impact of learning the value of X. So this behavior in the mutual information is as expected.

This example clearly demonstrates how mutual information can be a tool for conducting sensitivity analyses. Being able to quantify how much learning the value of X reduces the uncertainty in Y is in some sense a measure of how important or how much impact X has on the distribution of Y. It is for this reason that we will be using mutual information as our primary measure of sensitivity when studying the output distributions of stochastic simulators.

2.3.2 Estimation

In this section, we'll present two estimators for the mutual information between two random variables. The end of the section contains a simulation study to determine which of the two estimators performs better given limitations such as sample size. As before, we'll begin with some notational details and a clear definition of the quantity being estimated.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of points distributed according to the joint density f. The desired quantity is the mutual information between X and Y, i.e.,

$$I(X,Y) = -\int_{\mathbf{X}}\int_{\mathbf{Y}} log\left\{\frac{f(x,y)}{f_X(x)f_Y(y)}\right\} f(x,y)dydx, \tag{2.36}$$

where (X, Y) are placeholder random variables also with joint density f. We let $\hat{I}_n(X,Y)$ denote an estimate of $I(X,Y)$ based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. As was the case with entropy, the two estimators presented here are nonparametric estimators of mutual information since usually little is known about the form of the joint density.

Resubstitution Estimates

A rather naïve estimator for the mutual information would be to look at one of the expanded forms in terms of entropies, e.g.,

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X), \tag{2.37}$$

and then replacing each of the entropies with the appropriate estimates. An estimator of this type would be

$$\hat{I}_n(X,Y) = \hat{H}_n(X) + \hat{H}_n(Y) - \hat{H}_n(X,Y). \tag{2.38}$$

Here, $\hat{H}_n(X)$ and $\hat{H}_n(Y)$ are estimates of $H(X)$ and $H(Y)$ based on the marginal samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Also, $\hat{H}_n(X,Y)$ is an estimate of $H(X,Y)$ from the joint sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The estimator used could be any of the estimators mentioned in Section 2.2. Since this expression is just a sum, most of the convergence or limiting properties for the individual entropy estimators should translate to this estimate of mutual information.

However regardless of the type of entropy estimator used, any mutual information estimator of the form of Eq. 2.38 does not perform well because the estimation errors for the individual entropies will be additive and there is no reason to expect any kind of cancellation. One might hope to improve on this estimate by using the second expression in Eq. 2.37 to get the estimator

$$\hat{I}_n(X,Y) = \hat{H}_n(Y) - \hat{H}_n(Y|X), \tag{2.39}$$

which has one less entropy estimate and, ostensibly, should improve on accuracy on this premise alone. For the most part, this estimator performs reasonably well when using the

resubstitution estimators in Eq. 2.27 and 2.30 for $\hat{H}_n(Y)$ and $\hat{H}_n(Y|X)$, respectively; the accuracy is comparable to the other estimate presented in this section.

Note that, by symmetry, we could come up with an X-oriented analogue to Eq. 2.39,

$$\hat{I}_n(X,Y) = \hat{H}_n(X) - \hat{H}_n(X|Y), \tag{2.40}$$

which, from a purely statistical point of view, should not offer any advantages. However, in the context of stochastic simulators or Bayesian problems, the quantity of interest, Y, is often defined in terms of or conditionally given X. In the simulation setting, X is typically used to denote the input values and $Y = f(X)$ is the output of a simulator when it is run at a given X-value. In the Bayesian setting, X would be the parameters (e.g. mean, variance, correlation parameter) and $Y|X$ is the generative model given the parameters.

In these types of settings, the distribution of X is often known. For simulators, the distribution for the inputs is often determined by the scientists conducting the experiment and is typically based on how uncertain they are about the parameters. For Bayesian models, the distribution for X would be the prior distribution which has a known form. In both these cases, $H(X)$ can be calculated exactly (or at least numerically, independently of the data), and so an estimate of the mutual information between X and Y is

$$\hat{I}_n(X,Y) = H(X) - \hat{H}_n(X|Y), \tag{2.41}$$

which should clearly be more accurate than Eq. 2.40 due to estimating one less quantity. In principle, having one less estimated quantity should also make it more accurate than the estimator in Eq. 2.39. For a fixed sample size, the accuracy of $H_n(Y|X)$ and $H_n(X|Y)$ to their target quantities is largely dependent on the forms of the true conditional densities $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$, so there are cases where the Y-centric estimator may be more accurate. However, without any prior knowledge on the forms of the conditional densities, it's not possible to know which would be better for a specific case. In the ensuing simulation study, we'll evaluate the performance of Eq. 2.39, Eq. 2.40, and Eq. 2.41 to highlight possible scenarios where one would outperform the others.

Nearest Neighbor Estimates

In this section, we briefly present the two nearest neighbor distance-based estimates of mutual information derived by Kraskov et al. (2004). The presentation here contains only the formulae and a brief desription about the estimators. Formal developments and the intuition behind the estimators can be found in Kraskov et al. (2004).

These estimators of mutual information start off in the same way as in Eq. 2.38 – by replacing the entropies in the expression for mutual information with their respective nearest

neighbor distance based estimates, e.g.,

$$\hat{I}_n(X,Y) = -\psi(k) + \psi(n) + \frac{p_x}{n}\sum_{i=1}^{n} log\left\{\epsilon_{i,k}^{x}\right\} - \psi(k) + \psi(n) + \frac{p_y}{n}\sum_{i=1}^{n} log\left\{\epsilon_{i,k}^{y}\right\} \tag{2.42}$$
$$+ \psi(k) - \psi(n) - \frac{p_{(x,y)}}{n}\sum_{i=1}^{n} log\left\{\epsilon_{i,k}^{(x,y)}\right\}$$

where $\epsilon_{i,k}^{x}$ is the distance between x_i and it's k-th nearest neighbor, $\epsilon_{i,k}^{y}$ is the distance between y_i and it's k-th nearest neighbor, and $\epsilon_{i,k}^{(x,y)}$ is the distance between (x_i, y_i) and it's k-th nearest neighbor. Note that since they are on different spaces, the neighbors do not necessarily correspond for the different ϵ's.

Through some astute insight and reasoning, some of the terms in the entropy estimators can be cancelled out in order to simplify the expression while simultaneously reducing estimation error. The authors present two separate estimators of mutual information derived in this manner.

The first mutual information estimator is

$$\hat{I}_n^{(1)}(X,Y) = \psi(k) + \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\left[\psi(n_x(i)+1) + \psi(n_y(i)+1)\right]. \tag{2.43}$$

As before, ψ is the digamma function. Here, $n_x(i)$ is the number of sample points whose x values are closer than the distance between (x_i, y_i) and it's k-th nearest neighbor. Formally:

$$n_x(i) = \#\left\{x_j : d(x_i, x_j) \leq \epsilon_{i,k}^{(x,y)}\right\}, \tag{2.44}$$

where, as before, $\epsilon_{i,k}^{(x,y)}$ is the distance of (x_i, y_i) from it's k-th nearest neighbor. For both $\epsilon_{i,k}^{(x,y)}$ and $d(x_i, x_j)$, the distance used is the supremum norm. $n_y(i)$ is defined in a similar fashion.

The second estimator of mutual information is

$$\hat{I}_n^{(2)}(X,Y) = \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}\left[\psi(m_x(i)) + \psi(m_y(i))\right], \tag{2.45}$$

where $m_x(i)$ is the number of sample points whose x values are closer than the distance between the x_i and the x-component of (x_i, y_i)'s k-th nearest neighbor. Formally

$$m_x(i) = \#\left\{x_j : d(x_i, x_j) \leq d(x_i, x_{k(i)})\right\}, \tag{2.46}$$

where $x_{k(i)}$ is the x component of the k-th nearest neighbor of (x_i, y_i). That is, $x_{k(i)}$ is such that

$$k = \#\left\{(x_j, y_j) : d\left((x_i, y_i), (x_j, y_j)\right) \leq d\left((x_i, y_i), (x_{k(i)}, y_{k(i)})\right)\right\}.$$

As before, $m_y(i)$ and $y_{k(i)}$ are defined similarly. The original authors found that Eq. 2.43 and Eq. 2.45 perform similarly in most situations, with computation times also being comparable. In our tests, we've found Eq. 2.43 to be easier to implement, and consequently will only use that estimator when comparing against the other estimates mentioned in Section 2.3.2. However, since the two are nearly interchangeable, most of our findings should hold for Eq. 2.45 as well.

An important parameter choice that must be made for both these estimators is the choice of k, the number of neighbors being considered. We've found that, for large sample sizes, setting k too small, say $k = 1$ or 2, may result in numerical issues when computing the estimators since the distance between a point and its nearest neighbor may be numerically 0. However, taking k too large increases the approximation error that occurs when using the distribution of nearest neighbor distances as an analog for the density. This effect becomes quite pronounced in higher dimensions since the distance between subsequent neighbors is quite large when the dimension increases, so smaller values of k are preferable. Conversely, for large samples and small values of k, minor changes in k have little effect on the estimated mutual information aside from eliminating potential numerical issues. For the most part, setting k between 3-5 yields good accuracy.

The most important feature of these nearest neighbor types of estimators is their fast computational times. The slowest operation for both estimates is identifying the k-th nearest neighbor for each point, which is a well studied topic in the computer sciences. Through the use of k-d trees, finding the k-th nearest neighbors for a sample of size n is, on average, an operation of $O(nlogn)$ complexity, which is orders of magnitude faster than the estimation and evaluation of a KDE for a size n sample, which has $O(n^2)$ complexity. So the estimate in Eq. 2.43 is much faster to calculate than both 2.39 and 2.41. If computation time were the only factor, then this would be the primary estimate used.

2.3.3 Simulation Studies

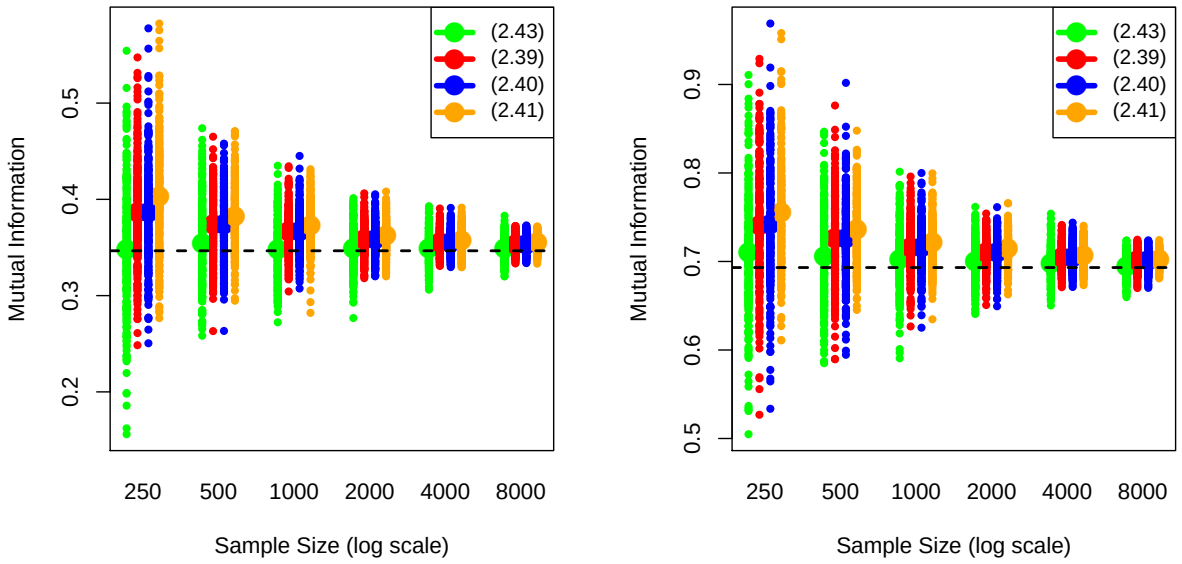
In this section, we'll evaluate the performance of the mutual information estimators presented in Section 2.3.2 on two different example problems. The important factors considered are (i) overall accuracy and (ii) accuracy given a fixed sample size. Differences in computation time will also be mentioned, since we will be estimating the mutual information from each individual MCMC draws from a posterior distribution later on. However, this process is largely parallelizable, so computing time will not have as much influence on the chosen method as accuracy considerations.

The first problem we'll estimate is the Bayesian example from Sections 2.1.3 and 2.3.1. In that example, $X \sim N(\mu, \sigma^2)$ and $Y|X \sim N(X, \tau^2)$. In 2.3.1, we calculated the mutual information between X and Y to be

$$I(X,Y) = \frac{1}{2} log \left\{ \frac{\sigma^2 + \tau^2}{\tau^2} \right\}.$$

In our first set of simulation studies, we will be evaluating the performance of the three resubstitution estimates of mutual information, Eq. 2.39, 2.40, and 2.41, and the first

nearest neighbor distance based mutual information estimator, Eq. 2.43, when estimating the mutual information between Y and X for this model. For the resubstitution estimates, $\hat{H}_n(Y), \hat{H}_n(X), \hat{H}_n(Y|X)$, and $\hat{H}_n(X|Y)$ will be computed with a KDE using a Gaussian kernel. In Eq. 2.40, the true value of $H(X) = \frac{1}{2} log\{2\pi e\sigma^2\}$ is used. The two cases used are when (i) $\sigma^2 = 2$ and $\tau^2 = 2$ and (ii) $\sigma^2 = 3$ and $\tau^2 = 1$. In both cases, the marginal variance for Y is $\sigma^2 + \tau^2 = 4$ (see Eq. 2.23).



(a) $\sigma^2 = 2$, $\tau^2 = 2$ (b) $\sigma^2 = 3$, $\tau^2 = 1$

Figure 2.1: Plots of different estimates of $I(X,Y)$ when $X \sim N(\mu, \sigma^2)$ and $Y|X \sim N(X, \tau^2)$ across varying sample sizes. The parameters were chosen so the marginal variance, $Var(Y)$, remains constant in both Figs. 2.1a and 2.1b. Eq. 2.43, in green, is the nearest-neighbor based estimate of mutual information. Eq. 2.39, in red, is $\hat{I}(X,Y) = \hat{H}(Y) - \hat{H}(Y|X)$. Eq. 2.40, in blue, is $\hat{I}(X,Y) = \hat{H}(X) - \hat{H}(X|Y)$. Eq.2.41, in orange, is $\hat{I}(X,Y) = H(Y) - \hat{H}(Y|X)$.

Figure 2.1 contains the result of the simulation studies for these different values of σ^2 and τ^2. The most noticeable result from these studies is how the three resubstitution type estimators are quite biased for small samples. This is not surprising since, in general, kernel density estimates are biased estimators for the density, so one should not expect estimators using these KDEs to produce an unbiased estimate of mutual information. Like the KDE estimator itself, the bias for these resubstitution estimate decreases with sample size. Eq. 2.43, the nearest neighbor distance based estimator looks unbiased for all sample sizes. However, while the bias is smaller, it appears to have slightly more variability than the resubstitution estimators for each sample sizes.

First, let's consider the case when sample size is not a limiting factor. In this case, it is hard to pick the better estimator based on just performance since the resubstitution estimators are more biased but have less variability while the nearest neighbor estimator

is less biased but has more variability. Luckily, there is a large difference in computation times that makes this decision easy. Estimating the mutual information with the nearest neighbor estimator from a sample with $n = 2000$ is 2 to 3 times faster than calculating the resubstitution estimator for a sample of size $n = 1000$ and the gap widens for even larger sized samples. Since we will be estimating the mutual information from individual MCMC draws, the difference in time between the estimates is on the order of several hours, even after parallelization. Note that the variability for the nearest neighbor estimator at $n = 2000$ is comparable to the variability of the resubstitution estimators when $n = 1000$, so accuracy does not need to be sacrificed if using the faster estimator. Consequently, when sample size is not a restriction, the nearest neighbor estimator should be the preferred choice.

When restricting ourselves to moderately sized samples ($n \leq 500$), the bias in the case of the resubstitution estimators is quite large - on the order of ten percent when $n = 250$. Additionally, the improvement in variance over the nearest neighbor estimate is small, and hardly meaningful when the bias is that large. So when samples are small, the nearest neighbor estimator is also the better choice.

When looking at just the resubstitution estimates, first notice that Eq. 2.40 is less biased than Eq. 2.41 for all sample sizes regardless of the value of σ^2, the variance of X. While there is no immediately obvious reason for any of the estimation error to cancel out when differencing $\hat{H}(X)$ and $\hat{H}(X|Y)$ in Eq. 2.40, it should be clear that error cancellation must be at least partially responsible for Eq. 2.41 having a larger estimation bias. The only difference between the two estimates is the replacement of $\hat{H}(X)$ with the true value, $H(X)$. As for Eq. 2.39 and 2.40, we see that their performance is nearly identical for all sample sizes. This should not be surprising since Y and X are both marginally Normal and both $Y|X$ and $X|Y$ also have Normal distributions. So for this example, changing the point of view should not have an impact on the estimation accuracy since the distributions are of the same family and thus are comparably well-approximated with kernel density estimators.

The second example problem we will look at is the following Normal mixture:

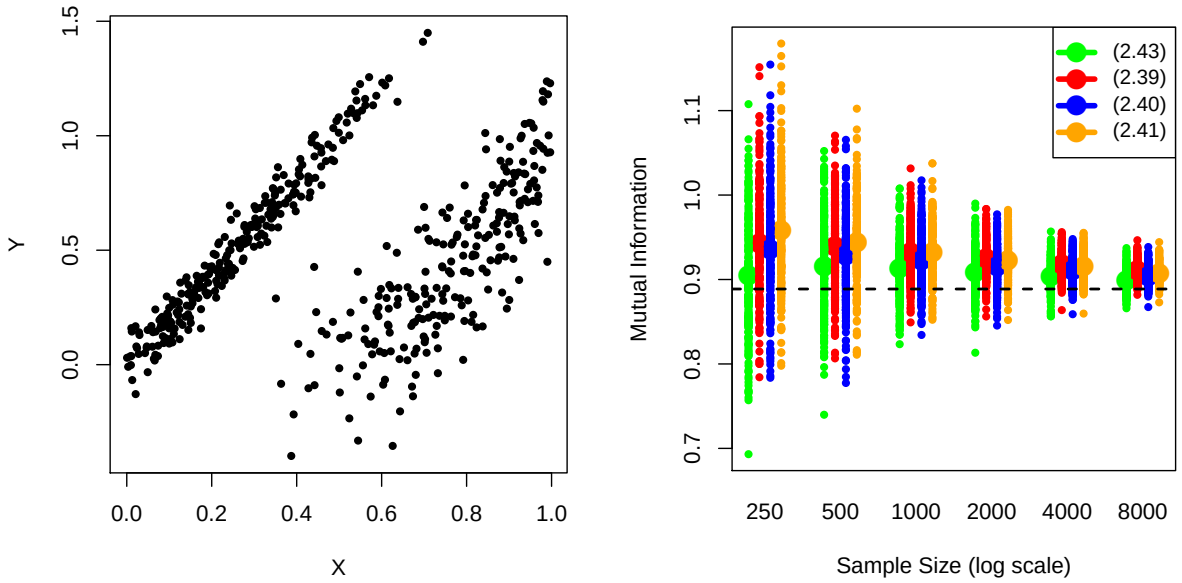
$$\begin{aligned} Y|X &\sim \begin{cases} N(2X, \sqrt{.005}) & \text{with probability } exp\left(-30X^6\right) \\ N(X^4, \sqrt{.04}) & \text{with probability } 1 - exp\left(-30X^6\right) \end{cases} \\ X &\sim Unif(0,1) \end{aligned} \tag{2.47}$$

This example is chosen because the model we use in Chapters 3 and later will be mixtures of this type, so it is important to determine which of the mutual information estimators performs better on these types of models. Since we will be referring to this model later on, we provide the conditional density for $Y|X$ here:

$$f(y|x) = \frac{exp\left(-30x^6\right)}{\sqrt{2\pi(.005)}} exp\left\{-\frac{(y-2x)^2}{2(.005)}\right\} + \frac{1 - exp\left(-30x^6\right)}{\sqrt{2\pi(.04)}} exp\left\{-\frac{(y-x^4)^2}{2(.04)}\right\}. \tag{2.48}$$

A plot of $n = 500$ realizations from Model 2.47 is provided in Fig. 2.2a. The two groups are quite distinct, with both groups only having similar chances of occuring when

$X \in [0.4, 0.6]$. From both the plot of realizations and the formulation of the model, it's clear that both the marginal mean and variance for Y are changing with X.



(a) Realizations from Model 2.47.

(b) $\hat{I}(X, Y)$

Figure 2.2: Fig. 2.2a contains a sample of $n = 500$ draws from the Normal mixture model defined in Model 2.47. Fig. 2.2b contains the result of a simulation study for estimating $I(X, Y)$ from the model. The equations and colouring scheme is the same as in Fig. 2.1. Unlike in Fig. 2.1, the Y-oriented estimator, Eq. 2.39 in red, has larger bias than the X-oriented estimators, Eqs. 2.40 and 2.41 in blue and orange.

Figure 2.2 contains the results of the simulation study on mutual information estimation. As was the case in the previous example, the nearest-neighbor estimator, Eq. 2.43, has a much smaller bias and slightly higher variability than any of the resubstitution estimators. From this, we can safely conclude that whatever property of the estimator is causing the bias advantage and increased variance is at least partially independent of the form of the joint density for X and Y.

Another similar finding from the previous study is the improvement of Eq. 2.40, $\hat{I}(X, Y) = \hat{H}(X) - \hat{H}(X|Y)$, over Eq. 2.41, $\hat{I}(X, Y) = H(X) - \hat{H}(X|Y)$. Again, this implies whatever error cancellation occurs when differencing the two estimates in Eq. 2.40 is at least partially independent of the form of the distributions being studied.

One deviation from our findings in the first study is the improved performance of the X-oriented estimators over Eq. 2.39, the Y-oriented estimator, $\hat{I}(X, Y) = \hat{H}(Y) - \hat{H}(Y|X)$. As alluded to when presenting these estimators Section 2.3.2, the better performing point of view will depend on the distributions of X and Y, both marginally and jointly. In this case, it's clear that the distributions for either Y or $Y|X$ (or both) is harder to approximate with a KDE. Consequently, we see Eq. 2.39 having larger estimation error in Fig. 2.2b.

The difference gets quite pronounced for larger sample sizes, where even Eq. 2.41, which we know to perform poorly, becomes less biased.

From our studies on both the Bayesian model from Sections 2.1.3 and 2.3.1 and the Normal mixture in Model 2.47, the mutual information estimator we will use in the rest of the chapters is the nearest-neighbor estimator, Eq. 2.43. This choice is based primarily on the smaller bias for this estimator and the orders of magnitude faster run times. When we perform mutual information estimation in Chapter 3.5, there will not be a strict limitation on sample size, so given the fast computation speed, the slightly higher variance of this estimator can be accounted for with a larger sample.

2.3.4 Sensitivity Analysis with Mutual Information: A Bayesian Example

In the context of sensitivity analysis for stochastic simulators, it turns out that this interpretation of the mutual information between X and Y as a quantification of how much knowing the value of X reduces the randomness or uncertainty in Y can be quite useful. A full development of how to use mutual information as a tool for sensitivity analysis in that setting is presented in Chapter 4. For now, we merely see how mutual information can be used to conduct sensitivity analysis by taking a thorough look at a canonical Bayesian problem.

Consider the following Bayesian model:

$$\begin{aligned} Y|\mu,\sigma^2 &\sim N(\mu,\sigma^2) \\ \mu|\sigma^2 &\sim N\left(\mu_0,\frac{\sigma^2}{\nu}\right) \\ \sigma^2 &\sim IG(a_0,b_0) \end{aligned} \tag{2.49}$$

Y is normally distributed with mean and variance (μ,σ^2) which have a Normal-Inverse-Gamma (NIG) prior distribution. Utilizing a NIG prior for the parameters introduces dependence between the two. However, the dependent prior was chosen purely for it's conjugacy to the likelihood for Y, not for any illustrative purposes. In general, the inputs to most simulators should be independent unless there is strong justification for dependence amongst the inputs from the researcher.

We set $a_0 = 0.1$ and $b_0 = 0.1$, so that σ^2 has a reasonably non-informative prior. We want to calculate the mutual information between Y and the two parameters, $I(Y,\mu)$ and $I(Y,\sigma^2)$ for different values of ν. The integrals are quite involved, so we will not bother deriving closed form expressions for the desired quantities. Instead, we estimate these quantities from using our estimator of choice - the nearest neighbor estimator in Eq. 2.43. For a given value of ν, we generate a sample of data from Model 2.49 by first sampling (μ,σ^2) according to the NIG prior. For each value of (μ,σ^2) in our sample, we draw a corresponding realization of $Y \sim N(\mu,\sigma^2)$.

Figure 2.3 is a plot of $I(Y,\mu)$ and $I(Y,\sigma^2)$ from Model 2.49 for different values of ν. By

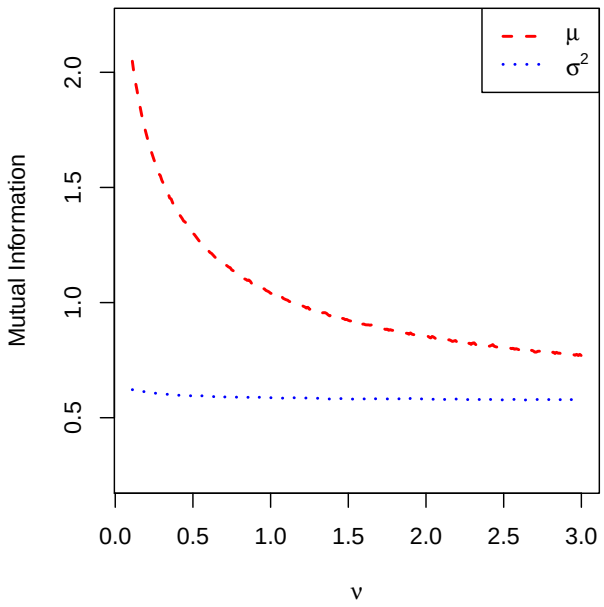


Figure 2.3: Mutual information between (Y, μ) and (Y, σ^2) for different values of ν.

the law of total variance, we know

$$Var(Y) = E\left[Var(Y|(\mu, \sigma^2)\right] + Var\left[E(Y|(\mu, \sigma^2))\right] = E(\sigma^2) + Var(\mu).$$

When ν is low, the marginal variance of μ is large which subsequently increases the marginal variance of Y. When Y has high marginal variability due to uncertainty in μ, learning the value of μ reduces the variance of Y significantly. More formally, when μ has high variance, the difference between $Var(Y)$ and $Var(Y|\mu)$ will be large. Since the mutual information measures the reduction in uncertainty, this means small values of ν result in larger values of $I(Y, \mu)$, as is apparently in the plot.

As ν gets larger, the marginal variability of μ decreases along with the effect of learning it's value on the entropy of Y. This decrease in effect on the entropy is why we see $I(Y, \mu)$ decrease for larger values of ν. For large enough ν, $\frac{\sigma^2}{\nu} = Var(\mu)$ will be small and hence μ will be nearly constant. In this case, μ will have a minimal effect on the variability or uncertainty in Y and hence the mutual information will be close to zero. This turns out to be the case - while not plotted in Figure 2.3, when $log(\nu) \approx 50$, $I(Y, \mu) \leq 10^{-5}$.

For σ^2, recall that in this model, the distribution for σ^2 is independent of ν, so it is not surprising that $I(Y, \sigma^2)$ does not change much based on the value of ν as demonstrated in Fig. 2.3. While the value of σ^2 does have a large effect on Y for modest values of ν, recall that mutual information is determined by the joint density of Y and σ^2, $f(y, \sigma^2)$, which requires marginalization over μ. Since μ, along with any dependence between ν and σ^2, has been integrated out in the joint distribution, $I(Y, \sigma^2)$ should be constant regardless of the value of ν.

In terms of sensitivity, for small values of ν, learning the value of μ tells you more about the value of Y than σ^2. However, as ν increases, the importance of μ decreases, and eventually σ^2 will provide more information about Y.

Chapter 3: Nonparametric Bayesian Density Regression with Kernel Stick Breaking Processes

In this chapter, we present a family of distributions over distributions called *kernel stick breaking processes* (KSBPs). These processes, originally developed in Dunson and Park (2008), are generalizations of Dirichlet Process mixtures (DP mixtures) that allows for predictor dependent mixing probabilities. A more complete treatment of the method and some important properties can be found in the original paper. Here, only the definition and some basic properties of the model and the method for taking samples from the model posterior are presented. Section 3.1.1 presents a method of utilizing KSBPs as a prior in a nonparametric Bayesian density regression model.

A Bayesian method for density estimation is used because sampling from the posterior distribution provides a direct way to calculate the variability of any estimated quantities of interest. In the context of this dissertation, our quantities of interest will primarily be the mutual information between the input and output of a stochastic simulator. Similar frequentist calculations of variability are quite difficult. Additionally, in cases where there is limited data available, frequentist methods often give poor estimates while Bayesian methods can still provide decent estimates - the posterior distribution will just have high uncertainty.

3.1 Kernel Stick Breaking Processes

Kernel stick breaking processes are generalizations of Dirichlet process mixtures that allow for predictor dependence. A typical DP-mixture can be represented with the infinite sum

$$F(\cdot) = \sum_{h=1}^{\infty} W_h G_h(\cdot) \tag{3.1}$$

where the $\{W_h, h = 1, \ldots, \infty\}$ are mixing probabilities that sum to 1, and the $\{G_h(\cdot), h = 1, \ldots, \infty\}$ are distributions. Often, the G_h's have a parametric form, so we can equivalently think of $\{\theta_h, h = 1, \ldots, \infty\}$ with $G_h(\cdot) = G(\cdot|\theta_h)$.

Of course, what makes Eq. 3.1 a DP mixture is if the W_h's are a realization from a Dirichlet process. Under the stick-breaking representation of Dirichlet processes, we can generate the W_h's by taking $W_h = V_h \prod_{l<h}(1 - V_l)$ where the V_h's are i.i.d. $Beta(1, \lambda)$ random variables. The $\{V_h, h = 1, \ldots, \infty\}$ can be thought of proportions of a progressively shrinking probability stick, giving rise to the term "stick-breaking."

KSBP's generalize DP-mixtures by allowing the mixing probabilities, $\{W_h, h = 1, \ldots, \infty\}$, to depend on a predictor x. The intention is for points with similar x values to have similar chances of mixing into each component, giving predictor dependence. Let $\mathcal{X}$ denote the space of input values x. A KSBP is composed of the following components:

- A collection of stick lengths, $\{V_h, h = 1, \ldots, \infty\}$ which are i.i.d. $Beta(1, \lambda)$ random variables. Like in the DP-mixture case, α controls dispersion about the central or modal distribution.

- A collection of knots, $\{\Gamma_h, h = 1, \ldots, \infty\}$ which are elements from some space $\mathcal{D}_\Gamma$. Note that $\mathcal{D}_\Gamma$ does not have to be the same space as $\mathcal{X}$, the predictor space. However, well chosen spaces will have points that have a clear sense of distance relative to points from $\mathcal{X}$.

- A collection of distributions, $\{G_h, h = 1, \ldots, \infty\}$. Typically these are all i.i.d. realizations from some distribution on distributions, $\mathcal{G}$. If the G_h's are parametric distributions, then this is equivalent to a collection of parameters $\{\theta_h, h = 1, \ldots, \infty\}$ with $G_h = G(\cdot|\theta_h)$. If $\mathcal{G}$ is a Dirichlet process, then each G_h would be a mixture of Dirac masses.

- A kernel function, K, that maps $\mathcal{X} \times \mathcal{D}_\Gamma$ to $[0, 1]$. Potential kernel functions could be the supremum norm $K(\mathbf{x}, \Gamma) = max_p \{|x_p - \Gamma_p|\}$ or exponential distance $K(\mathbf{x}, \Gamma) = exp\{-\psi||\mathbf{x} - \Gamma||^2\}$. The kernel function is what governs the level of predictor dependence in the mixture.

For a given predictor value x, the mixing probabilities defined in terms of these components is

$$W_h(\mathbf{x}) = V_h K(\mathbf{x}, \Gamma_h) \prod_{l<h} (1 - V_l K(\mathbf{x}, \Gamma_l)) \tag{3.2}$$

From this expression, it should be clear why the $\{\Gamma_h, h = 1, 2, \ldots\}$ variables are referred to as *knots* - the mixing probabilities for the h-th group for a given $\mathbf{x}$ is dependent on (i) how close $\mathbf{x}$ is to Γ_h and also (ii) far away $\mathbf{x}$ is from Γ_l for $l < h$. Essentially, the inclusion of the $\{\Gamma_h, h = 1, 2, \ldots\}$ variables associates a location to each component in the infinite mixture. How near or far $\mathbf{x}$ is to these locations affects the likelihood of that component being selected.

With the components defined above and the mixing probabilities as defined in Eq. 3.2, a KSBP can be written as

$$F_x(\cdot) = \sum_{h=1}^{\infty} W_h(\mathbf{x}) G_h(\cdot) \tag{3.3}$$

$$= \sum_{h=1}^{\infty} W_h(\mathbf{x}) G(\cdot|\theta_h), \tag{3.4}$$

with Eq. 3.4 being the case when $\mathcal{G}$ is a distribution over a parametric family.

The desired level of predictor dependence is introduced through these mixing probabilities and the components required for their calculations. It turns out that the strength of the dependence is primarily controlled by the kernel function K. The impact of K on the dependence is most apparent when taking K to be the the identity kernel, $K(\mathbf{x}, \Gamma) = 1$, which reduces this model back to the predictor independent DP-mixture in Eq. 3.1.

For nontrivial kernel functions, decreasing the distance between $\mathbf{x}$ and $\mathbf{x}'$ decreases the dissonance between $\{W_h(\mathbf{x}), h = 1, \ldots, \infty\}$ and $\{W_h(\mathbf{x}'), h = 1, \ldots, \infty\}$. That is, as $\mathbf{x}$ and $\mathbf{x}'$ get closer, they are more likely to come from the same mixture component. Throughout the rest of the dissertation, we'll be using the exponential distance kernel $K(\mathbf{x}, \gamma) = exp\{-\psi||\mathbf{x} - \gamma||^2\}$, where ψ is what we'll be referring to as the *distance parameter*.

Increasing the distance parameter ψ makes K decay more quickly, so $K(\mathbf{x}, \Gamma_h)$ and subsequently $W_h(\mathbf{x})$ will be smaller for h's where Γ_h is far from $\mathbf{x}$. This means the selected mixture component (i.e. h value) will be associated with a Γ_h that is close to $\mathbf{x}$. If $\mathbf{x}$ and $\mathbf{x}'$ are close to each other, they will be close to the same Γ_h's and consequently it is more likely that they are assigned to the same mixture component. Thus, increasing the distance parameter ψ increases the predictor dependence and, conversely, lowering the distance parameter lowers the amount of predictor dependence. A more formal calculation involving this relationship is given in 3.1.2.

3.1.1 KSBPs for Bayesian Density Regression

While KSBPs are very flexible models, we don't use them to conduct density regression directly. Instead, we perform nonparametric density regression by using a KSBP prior in

the following Bayesian model of Normal mixtures

$$\begin{aligned}
\beta_i|X_i, \mathbf{V}, \mathbf{\Gamma}, \mathbf{G}, \psi &\sim KSBP(\mathbf{V}, \mathbf{\Gamma}, \mathbf{G}; \psi) \\
Y_i|X_i, \beta, \sigma^2 &\sim N\left(X_i\beta_i, \sigma^2\right) \\
\sigma^2 &\sim IG(a_0, b_0) \\
\psi &\sim log\text{-}N(\mu_\psi, \sigma^2_\psi) \\
V_h &\overset{i.i.d.}{\sim} Beta(1, \lambda) \\
\Gamma_h &\overset{i.i.d.}{\sim} \mathcal{H} \\
G_h &\overset{i.i.d.}{\sim} DP(\alpha G_0),
\end{aligned} \tag{3.5}$$

where $\mathbf{V} = \{V_h, h = 1, \ldots, \infty\}, \mathbf{\Gamma} = \{\Gamma_h, h = 1, \ldots, \infty\}$, and $\mathbf{G} = \{G_h, h = 1, \ldots, \infty\}$. $\mathcal{H}$ is a distribution on the space $\mathcal{D}_\Gamma$. In low dimensions, $\mathcal{H}$ can be a uniform distribution over a fine grid of locations. In higher dimensions, $\mathcal{H}$ is usually some continuous distribution like a multivariate normal. For most stochastic simulators, a sensible choice for $\mathcal{H}$ would be the uncertainty distribution on inputs.

In Model 3.5, the distribution on distributions $\mathcal{G}$ is taken to be a Dirichlet process with dispersion parameter α and base distribution, G_0, being a multivariate normal with mean vector μ_0 and covariance matrix Σ_0. In practice, a prior is also placed on μ_0 and Σ_0, e.g.,

$$\begin{aligned}
G_0|\mu_0, \Sigma_0 &= MVN(\mu_0, \Sigma_0) \\
\mu_0 &\sim MVN(u_0, S_0) \\
\Sigma_0 &\sim Inv\text{-}Wishart(\nu, T_0)
\end{aligned} \tag{3.6}$$

where u_0, S_0, ν, T_0 are the parameters for their respective distributions chosen to make the corresponding priors uninformative.

Figure 3.2 illustrates the effect of placing a $\mathcal{G} = DP(\alpha G_0)$ prior on the $\{G_h, h = 1, 2, \ldots, \infty\}$. Figure 3.2a shows what potential draws from $\mathcal{G}$ will look like. The red line in the backgrounds is G_0, the base distribution governing the locations of the sticks. α, the dispersion parameter, controls the distribution of stick heights.

When working in higher dimensions with only modestly sized samples, the choice of hyperparameters for the log-Normal prior on the distance parameter ψ has an impact on the performance of the model, so they must be chosen sensibly. Section 3.1.2 presents a method for choosing reasonable parameter values for this prior.

Since Model 3.5 is the main generative model for data used throughout the rest of the dissertation, the graphical model representation is provided in Figure 3.1 to help visualize the relationship between the components in the model. We'll also take a moment to introduce some terminology that we'll be using throughout the rest of the dissertation. We refer to the h-th mixture component and the variables that correspond to that component, (V_h, Γ_h, G_h), as the h-th *group*. We say that the data point (X_i, Y_i) *comes from* the h-th group if β_i is from the h-th mixture component, i.e., $\beta_i \sim G_h$. Since G_h is a realization from a Dirichlet

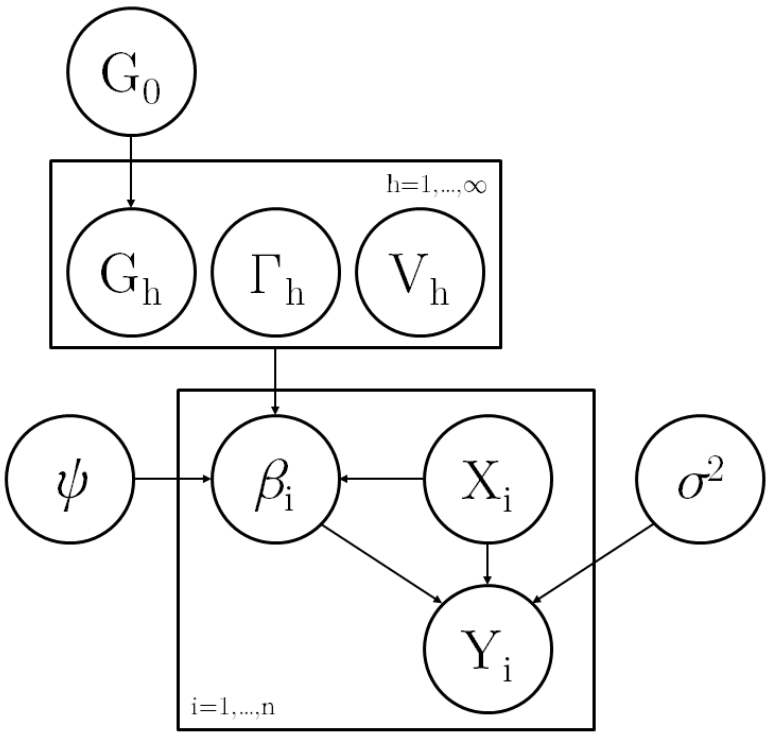


Figure 3.1: Graphical model for density regression using a KSBP.

process, this means the distribution of β_i will be a mixture of Dirac masses. That is,

$$\beta_i \sim \sum_{k=1}^{\infty} p_k^{(h)} \delta_{\theta_k^{(h)}}(\cdot), \tag{3.7}$$

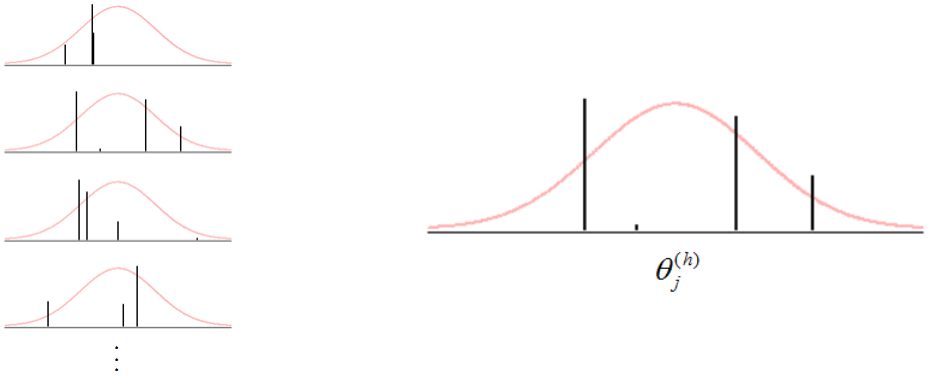
where h is the group that (X_i, Y_i) comes from, $p_k^{(h)}$ is a mixing probability, and $\theta_k^{(h)}$ are the different possible values of β_i sampled according to G_0, the base distribution. Figure 3.2b contains a potential distribution for β_i given that it is from the h-th group.

We can use Eq 3.7 to further separate all the (X_i, Y_i) points from the h-th group into even smaller collections based on the specific values of β_i. We refer to these more finely separated classes as different *clusters*. Formally, the (h, k)-th cluster is the set

$$\left\{ i : (X_i, Y_i) \sim N(X_i \theta_k^{(h)}, \sigma^2) \right\}. \tag{3.8}$$

To clarify, if (X_i, Y_i) and (X_j, Y_j) are from the same cluster, then $\beta_i = \beta_j$. If (X_i, Y_i) and (X_j, Y_j) are from the same group, then $\beta_i \overset{D}{=} \beta_j$.

In Figure 3.3, some example data is provided to highlight the differences between groups and clusters (*Note: the data is not a realization from Model 3.5.*). The points are colored according to which group they are from. Notice how points from the same group are spatially close in the X dimension. The green circles indicate the different clusters within the green



(a) Possible draws from $\mathcal{G} = DP(\alpha G_0)$. (b) Possible values of $\theta_j^{(h)}$ for a given G_h.

Figure 3.2: Figures to help illustrate the effect of having $\mathcal{G} = DP(\alpha G_0)$ on the actual values of β_i. In both figures, the red line indicates G_0.

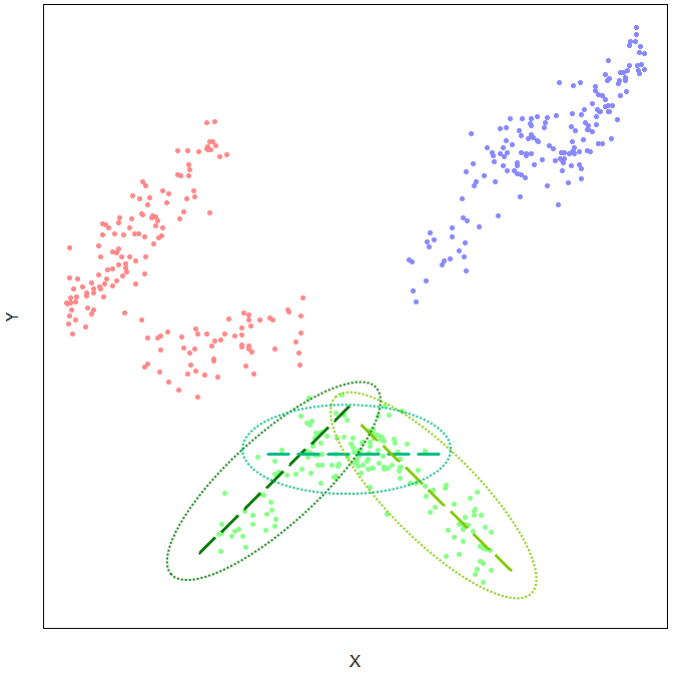


Figure 3.3: Example data illustrating the differences between groups and clusters.

group. The dashed lines in these groups represent the means within each cluster, which are determined by the values of $\theta_k^{(h)}$.

It is important that the definitions for groups and clusters be made very clear here because these terms will be used very frequently in the proceeding sections. Understanding the differences will be particularly important in Section 3.2 since the procedure for sampling from the posterior distribution of the model given data relies on augmenting the model with

cluster assignment indicators for each data point, and group assignment indicators for each cluster.

3.1.2 Prior Specification for ψ

As mentioned at the beginning of Section 3.1, the strength of the spatial relationship between the distributions $\beta_i|X_i = \mathbf{x}$ and $\beta_i|X_i = \mathbf{x}'$ is highly dependent on the value of ψ, the distance parameter of K, the kernel. Consequently, the choice of parameters for the prior distribution of ψ in Model 3.5 has a large effect on the resulting posterior distribution. In this section, we present a way to quantify the strength of the predictor dependence of a KSBP for different values of ψ. This method can be used to derive sensible upper and lower quantiles for ψ which implicitly define parameters to the log-Normal prior distribution on ψ.

Before we begin, recall that in Model 3.5, the KSBP is used as a prior distribution on a collection of X_i-dependent distributions for $\beta_i|X_i$. For concreteness, let $\mathcal{X}$, the space of inputs be a p-dimensional and hence β_i will be a vector in $\mathbf{R}^p$. In this case, a singular "draw" or realization from a KSBP defines a collection of distributions on $\mathbf{R}^p$ of the form defined in Eq. 3.3:

$$F_x(\cdot) = \sum_{h=1}^{\infty} W_h(\mathbf{x}) G_h(\cdot).$$

Here, we must emphasize the role of the subscript $\mathbf{x}$ - different values of $\mathbf{x}$ define different distributions on $\mathbf{R}^p$. Explicitly, for any Borel set $\mathcal{B} \subseteq \mathbf{R}^p$, $F_x(\mathcal{B})$ will, in general, be different from $F_{x'}(\mathcal{B})$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ whenever $\mathbf{x} \neq \mathbf{x}'$.

Since the collection of distributions $\{F_x(\cdot), \mathbf{x} \in \mathcal{X}\}$ is a realization from a KSBP, $F_x(\mathcal{B})$ and $F_{x'}(\mathcal{B})$ are not varying completely at random - the measure depends on $\mathbf{x}$ and $\mathbf{x}'$. If $\mathbf{x}$ and $\mathbf{x}'$ are "close" in some sense, then they might even be similar. In our case, since we are using the ψ-parameterized exponential kernel, $K(\mathbf{x}, \Gamma) = exp\{-\psi||\mathbf{x} - \Gamma||^2\}$, "closeness" corresponds to Euclidean distance and we can actually calculate how similar the two quantities $F_x(\mathcal{B})$ and $F_{x'}(\mathcal{B})$ will be given a value of ψ.

The ensuing calculation relies heavily on the following expression, originally from Dunson and Park (2008):

$$corr\{F_x(\mathcal{B}), F_{x'}(\mathcal{B})\} = \frac{\kappa(\mathbf{x}, \mathbf{x}') \left\{(2+\lambda)\frac{\kappa(\mathbf{x})}{\kappa_2(\mathbf{x})} - 1\right\}^{1/2} \left\{(2+\lambda)\frac{\kappa(\mathbf{x}')}{\kappa_2(\mathbf{x}')} - 1\right\}^{1/2}}{(1+\lambda/2)\{\kappa(\mathbf{x}) + \kappa(\mathbf{x}')\} - \kappa(\mathbf{x}, \mathbf{x}')} \tag{3.9}$$

where $\kappa(\mathbf{x}) = E[K(\mathbf{x}, \Gamma_h)]$, $\kappa_2(\mathbf{x}) = E[K(\mathbf{x}, \Gamma_h)^2]$, and $\kappa(\mathbf{x}, \mathbf{x}') = E[K(\mathbf{x}, \Gamma_h)K(\mathbf{x}', \Gamma_h)]$. Here, we will just accept this statement to be true; the section in the original paper that derives this equation is thorough and acceptably easy to follow. Instead, we focus more on what is being calculated and its implications as well as how to use this expression to specify a prior on ψ.

First, note that many of the components in this expression are expectations but the variables of integration for these expectations have been left out for notational simplicity (in lieu of clarity). While in the end unambiguous, it is not immediately clear which

variables are being integrated out. The correlation on the left, $corr\{F_x(\mathcal{B}), F_{x'}(\mathcal{B})\}$ is actually an integral over the KSBP-defined prior distribution on the collection of distributions $\{F_x(\cdot), \mathbf{x} \in \mathcal{X}\}$. This is equivalent to an integral over all the components that make up F_x: $\{(V_h, \Gamma_h, G_h), h = 1, 2, \ldots\}$. While an integral over multiple infinite sequences of parameters may seem lofty, this integral is possible due to (i) the form of F_x as defined by a KSBP, (ii) the chosen prior distributions for these components, and (iii) the independence across both variables and draws.

In particular, the choice of G_h to be i.i.d. draws from a Dirichlet process prior independently of $\{V_h, h = 1, 2, \ldots\}$ and $\{\Gamma_h, h = 1, 2, \ldots\}$ results in none of the parameters for this prior distribution appearing in the expression on the right. This is primarily a result of a marginalization properties for Dirichlet processes: the marginal distribution of a single draw from F, where F is a realization from a Dirichlet process, is the base distribution. Formally, if $X|F \sim F$ and $F \sim DP(\alpha G_0)$, then marginally, $X \sim G_0$.

Having Γ_h be i.i.d. draws from the distribution $\mathcal{H}$ and the independence between $\{V_h, h = 1, 2, \ldots\}$ allows the expression on the right to contain only integrals of Γ_h ($\kappa(\cdot), \kappa_2(\cdot), \kappa(\cdot, \cdot)$); all terms containing $\{V_h, h = 1, 2, \ldots\}$ have already been integrated out and only λ, the parameter of the $Beta(1, \lambda)$ prior distribution on V_h, remains.

The quantities $\kappa(\mathbf{x}), \kappa_2(\mathbf{x})$, and $\kappa(\mathbf{x}, \mathbf{x}')$ are all expectations with respect to $\mathcal{H}$, the prior distribution on Γ_h. Since K is the exponential distance, the simplest choice, mathematically, for $\mathcal{H}$ is a multivariate Normal with mean $\mathbf{0}$ and covariance matrix I_p. The integrals may be possible with prior distributions, but for now we focus only on the multivariate Normal case. With this choice of $\mathcal{H}$, the desired integrals are

$$\kappa(\mathbf{x}) = \left(\frac{1}{1+2\psi}\right)^{p/2} exp\left\{-\frac{\psi}{1+2\psi}\mathbf{x}^T\mathbf{x}\right\} \tag{3.10}$$

$$\kappa_2(\mathbf{x}) = \left(\frac{1}{1+4\psi}\right)^{p/2} exp\left\{-\frac{2\psi}{1+4\psi}\mathbf{x}^T\mathbf{x}\right\} \tag{3.11}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = \left(\frac{1}{1+4\psi}\right)^{p/2} exp\left\{-\frac{2\psi^2+\psi}{1+4\psi}\left[\mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}'\right] + \frac{2\psi^2}{1+4\psi}\left[\mathbf{x}^T\mathbf{x}' + \mathbf{x}'^T\mathbf{x}\right]\right\} \tag{3.12}$$

$$\kappa(\mathbf{x})/\kappa_2(\mathbf{x}) = \left(\frac{1+4\psi}{1+2\psi}\right)^{p/2} exp\left\{\frac{\psi}{(1+2\psi)(1+4\psi)}\mathbf{x}^T\mathbf{x}\right\} \tag{3.13}$$

The derivations for these expressions are provided in Appendix 7.1. The last expression, Eq. 3.13, is also provided because that ratio shows up twice in the numerator of Eq. 3.9. With these expressions, it is possible to calculate the correlation between the two measures $F_x(\mathcal{B})$ and $F_{x'}(\mathcal{B})$ in terms of ψ for any two values of $\mathbf{x}$ and $\mathbf{x}'$. Since the resulting quantity is independent of the measured set, $\mathcal{B}$, it can be interpreted as a measure of how related F_x and $F_{x'}$ are in terms of the distance parameter ψ. A cursory evaluation of Eq. 3.9 with Eqs. 3.10-3.13 plugged in reveals, for given $\mathbf{x}$ and $\mathbf{x}'$, the correlation decreases as ψ increases and vice versa. Intuitively, when ψ is close to zero, $K(\mathbf{x}, \Gamma) = exp\{\psi||\mathbf{x} - \Gamma||\} \approx 1$ regardless of

the values of $\mathbf{x}$ or Γ. This means $W_h(\mathbf{x}) = V_h K(\mathbf{x}, \Gamma_h) \approx V_h K(\mathbf{x}', \Gamma_h) = W_h(\mathbf{x}')$ for any $\mathbf{x}, \mathbf{x}'$, and h. If the mixing probabilities, $W_h(x)$ and $W_h(\mathbf{x}')$, are close, then the distributions, F_x and $F_{x'}$, must be close as well and hence there will be a strong correlation in their measures.

In order to turn this expression into a measure of spatial dependence, we average over possible values of $\mathbf{x}$ and $\mathbf{x}'$. If taking a fully Bayesian approach then the averaging would be over the input distribution for $\mathbf{x}$ and $\mathbf{x}'$. Alternatively, if a set of data is available, one could resample the input values in the data and then average over the resampled input distribution. In both cases, the target quantity is the expectation

$$E\left[corr\left\{F_x(\mathcal{B}), F_{x'}(\mathcal{B})\right\}\right] = \int_{\mathcal{X}} \int_{\mathcal{X}} corr\left\{F_x(\mathcal{B}), F_{x'}(\mathcal{B})\right\} dP(\mathbf{x}) P(\mathbf{x}'), \tag{3.14}$$

where P is the input distribution on $\mathbf{x}$ and $\mathbf{x}'$. Note, in the integral we've taken $\mathbf{x}$ and $\mathbf{x}'$ to be independent of each other. Taking expectations over $\mathbf{x}$ and $\mathbf{x}'$ means Eq. 3.14 is independent of their values and can be thought of as the average correlation when taking two points at random according to P. In this sense, it is a measure of the spatial strength of the KSBP process for a given value of ψ. In the context of spatial statistics, it is analogous to the range of a spatial process.

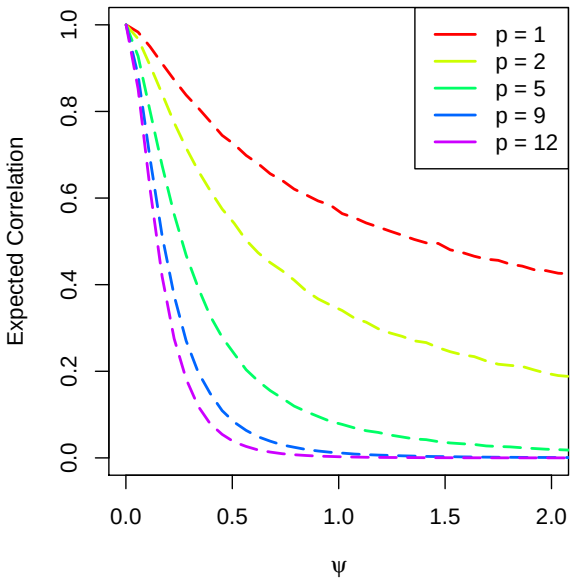


Figure 3.4: Plot of Eq. 3.14, the expected correlation between $G_x(\mathcal{B})$ and $G_{x'}(\mathcal{B})$ for different values of ψ. The differently colored lines indicate different input dimensions, p. The expectation is over $\mathbf{x}$ and $\mathbf{x}'$, which are independent p-variate Normal random variables.

Figure 3.4 contains a plot of the estimated value of Eq. 3.14 for different values of ψ and different dimensions, p. As was the case when $\mathbf{x}$ and $\mathbf{x}'$ were fixed, higher values of ψ means that the expected correlation will be smaller. We see this effect to be more pronounced as

the dimension, p, increases; in 12 dimensions, the expected correlation is almost zero when $\psi = 0.5$. This plot can be used to pick sensible high and low quantiles for ψ.

For example, if $p = 5$, we can take the 2.5th and 97.5th percentile of ψ to be 0.01 and 0.5. These values of ψ correspond to expected correlations of close to 1 when $\psi = 0.01$ and about 0.2 when $\psi = 0.5$. Anything lower than an expected correlation of 0.2 implies almost no smoothness - in which case there would no benefit to using a KSBP prior on $\beta_i|X_i$ in the first place. Since ψ has a log-Normal distribution, given the desired quantiles, then the mean and standard deviation of ψ on the log scale must be -2.649 and 0.998, respectively.

3.2 Posterior Computation

This section presents a method for sampling from the posterior distribution of the KSBP-based density regression model defined in Model 3.5. The method here follows the method presented in the original paper pretty closely, so the emphasis will be less on the derivations and more on the implementation details. One particular deviation from the original method worth noting is the use of slice sampling in Section 3.2.2 for the group assignments for each cluster instead of a retrospective sampler. Many of the descriptions in the original paper are for the more general case. Here, they are specific to the likelihood and priors defined in Model 3.5. The descriptions provided here are intended so that someone should be able to fully implement a sampler from the posterior without needing to reference any other resources.

Given a set of n data points, $\mathbf{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$, sampling from the posterior distribution of Model 3.5 requires sampling from the posterior distribution for $\beta_i|X_i$ given the data. However, due to the form of the $KSBP$ prior placed on $\beta_i|X_i$, sampling from that posterior distribution is equivalent to sampling from the posterior distributions of the variables that make up that distribution - $\mathbf{V} = \{V_h, h = 1, \ldots, \infty\}, \mathbf{\Gamma} = \{\Gamma_h, h = 1, \ldots, \infty\}, \mathbf{G} = \{G_h, h = 1, \ldots, \infty\}$. In addition to these quantities, we also need samples from the posterior distributions for the distance parameter ψ, the process variance σ^2, and the parameters of the base distribution μ_0 and Σ_0.

To facilitate the sampling from this posterior, we augment the data with the following variables:

- The unique values of β_i, which we denote as $\mathbf{\Theta} = \{\theta_j, j = 1, \ldots, J\}$. For a sample of size n, there are at most n unique values of β, so clearly $J \leq n$. Recall from Section 3.1.1, that points in the same cluster have the same values of β, so J is the number of "observed" clusters and there will be one value of θ_j for each cluster.

- Cluster assignments for each data point, $\mathbf{S} = \{S_i, i = 1, \ldots, n\}$. If $S_i = j$, then X_i is from the j-th cluster and hence $E(Y) = X_i\theta_j$ and $Y \sim N(X_i\theta_j, \sigma^2)$. It is these properties that causes us to want to focus on cluster assignments for each data point when sampling (as opposed to group assignments, which would not give this property).

- Group assignments for each cluster, $\mathbf{C} = \{C_j, j = 1, \ldots, J\}$. From Section 3.1.1, clusters are subdivisions of a group resulting from the G_h distributions being realizations

of a $DP(\alpha G_0)$. So if X_i is assigned into the j-th cluster (i.e. $S_i = j$), then it must have been assigned into group C_j at some point.

For notational convenience, we also define $\mathbf{Z} = \{Z_i, i = 1, \ldots, n\}$ to be the group assignments for each data point. We will never sample these variables directly; their values can be inferred using the relationship $Z_i = C_{S_i}$. However, the sampling methods for some of the variables may be easier write out in terms of these variables, so we provide adequate notation here.

When sampling the cluster assignments, S_i, we will sample them one-at-a-time given the values of $S_j, j \neq i$. So it's beneficial to introduce some simplifying notation here. Let $\mathbf{S}^{(i)}$ denote the set of cluster assignments for all the data points except for X_i. Let $\mathbf{\Theta}^{(i)}$ denote all the unique values of β when X_i is excluded (i.e. all the unique values of β_j, $j \neq i$). Similarly, $\mathbf{C}^{(i)}$ denotes the group assignments for all the existing clusters when i is excluded. Lastly, $\mathbf{Z}^{(i)}$ denotes the group assignments for all data points with the exception of X_i. Some care should be taken when implementing this sampler because the lengths of $\mathbf{C}^{(i)}$ and $\mathbf{\Theta}^{(i)}$ will differ based on the value of i. For example, when i is assigned to it's own cluster, $\mathbf{C}^{(i)}$ and $\mathbf{\Theta}^{(i)}$ will have one less entry than $\mathbf{C}$ and $\mathbf{\Theta}$. In these cases, $J^{(i)}$, the number of clusters with X_i removed, is also one less.

With these additional variables, a sample from the posterior for $\mathbf{V}, \mathbf{\Gamma}, \mathbf{G}, \psi, \sigma^2, \mu_0, \Sigma_0$ is equivalent to a sample from the posterior for $\mathbf{V}, \mathbf{\Gamma}, \mathbf{S}, \mathbf{C}, \mathbf{\Theta}, \psi, \sigma^2 \mu_0, \Sigma_0$. In particular, the inclusion of $\mathbf{\Theta}$ and $\mathbf{S}$ allows us to avoid storing any of the distributions in $\mathbf{G}$, since each of the individual G_h's in $\mathbf{G}$ is an infinite mixture.

The rest of this section describes the steps of a Gibbs sampler for drawing from the posterior. Since the sampling method is of the Gibbs type, each step describes how to sample from the full conditional for a variable. In order to help visualize the conditional dependence between the variables, an updated graphical model with the augmented variables is provided in Figure 3.5.

3.2.1 Cluster Assignments for Each Data Point, S_i

The sampling method for the cluster assignment variables is by far the most complicated of the variables being sampled. The derivation for many of the steps for the general case presented in Dunson and Park (2008) is quite lengthy so they will be omitted. In the case of 3.5 and the chosen priors, some simplifications can be made and those are presented here. We first define the following quantities. Let $\mathcal{N}_h^{(i)}$ denote the number of data points assigned to group h with the exception of the i-th point. Since point i is being excluded, this means $\sum_{h=1}^{\infty} \mathcal{N}_h^{(i)} = n - 1$ for all i. In terms of the augmented variables, this quantity can be calculated according to

$$\mathcal{N}_h^{(i)} = \sum_{j \neq i} \mathbf{1}\left\{Z_j^{(i)} = h\right\}.$$

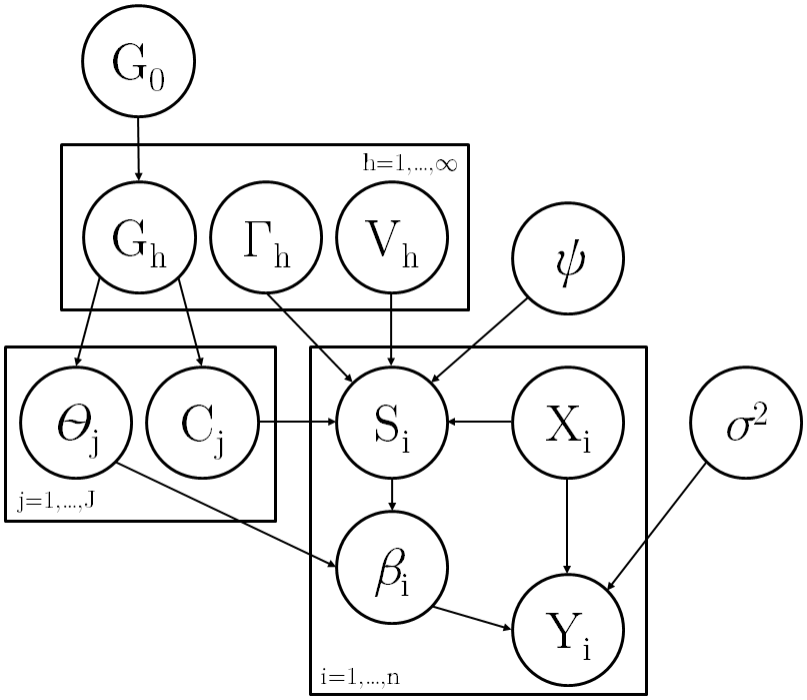


Figure 3.5: Graphical model for density regression using a KSBP with augmented variables included. The inclusion of the augmented variables S_i, C_j, θ_j breaks many of the dependences.

(Note: this definition of $\mathcal{N}_h^{(i)}$ is slightly different from the usage in the original paper.) We use this quantity to define the following two ratios

$$\begin{aligned} l_{ih0} &= \frac{\alpha}{\alpha + \mathcal{N}_h^{(i)}} \\ l_{ihk} &= \frac{1}{\alpha + \mathcal{N}_h^{(i)}}, \quad k = 1, \ldots, n \end{aligned} \tag{3.15}$$

Those familiar with sampling methods for Dirichlet processes will recognize these as Polya urn-style sampling probabilities. In the original paper, the authors use these ratios to define the following weights

$$\begin{aligned} w_{i,0} &= \sum_{h \in I_{oc}^{(i)}} W_h(X_i) l_{ih0} \\ w_{i,j} &= \sum_{k: S_k^{(i)} = j} l_{iC_j^{(i)}k}, \quad j = 1, \ldots, J^{(i)} \\ w_{i,J^{(i)}+1} &= \sum_{h \in I_{uc}^{(i)}} W_h(X_i) l_{ih0} \end{aligned} \tag{3.16}$$

where, as before, $W_h(X_i) = V_h K(X_i, \Gamma_h) \prod_{l<h} (1 - V_l K(X_i, \Gamma_l))$ denotes the probability of assigning point X_i to the h-th group. A brief note about the sets being summed over in Eqs. 3.16. For $w_{i,0}$ and $w_{i,J^{(i)}+1}$, the sets $I_{oc}^{(i)}$ and $I_{uc}^{(i)}$ are used to denote the set of "occupied" and "unoccupied" groups with the effect of point X_i is removed. A group is *occupied* if at least one data point is assigned to a cluster within that group. In terms of the variables being sampling, the set $I_{oc}^{(i)}$ is all the unique values of $\mathbf{C}^{(i)}$. More precisely,

$$I_{oc}^{(i)} = \left\{ h : C_j^{(i)} = h \text{ for some } j \right\}$$

Using this definition, $I_{uc}^{(i)}$ is simply $\{1, 2, \ldots, \infty\} \setminus I_{oc}^{(i)}$. For $w_{i,j}$, the sum is over the set $\{k : S_k^{(i)} = j\}$ which denotes the set of data points that have been assigned to cluster j. Unlike $w_{i,0}$ and $w_{i,J^{(i)}+1}$, which are sums over collections of *groups*, this is a sum over a collection of *data points*. Explicitly, $\{k : S_k^{(i)} = j\} \subseteq \{1, 2, \ldots, n\}$.

The formulae in 3.16 were presented in a very general setting in the original paper. For our purposes, we can use the specific forms of l_{ih0} and l_{ihj} and the definitions of the summation sets to get simplified expressions for these weights. In the case of $w_{i,j}$, notice that from our definition of l_{ihk} in Eq. 3.15 the value of $l_{iC_j^{(i)}k}$ is always

$$l_{iC_j^{(i)}k} = \frac{1}{\alpha + \mathcal{N}_{C_j^{(i)}}^{(i)}}$$

regardless of the value of k. Since the summand is constant, for $j = 1, \ldots, J^{(i)}$, $w_{i,j}$ simplifies to

$$w_{i,j} = \sum_{g:S_g^{(i)}=j} l_{iC_j^{(i)}g} = \frac{\sum_{k=1}^{n} \mathbf{1}\{S_k = j\}}{\alpha + \mathcal{N}_{C_j^{(i)}}^{(i)}}. \tag{3.17}$$

To simplify $w_{i,J^{(i)}+1}$, notice that for any $h \in I_{uc}^{(i)}$, $\mathcal{N}_h^{(i)} = 0$ since, group h is unoccupied. Consequently, for $h \in I_{uc}^{(i)}$, $l_{ih0} = \frac{\alpha}{\alpha + \mathcal{N}_h^{(i)}} = \frac{\alpha}{\alpha + 0} = 1$. With this in mind, $w_{i,J^{(i)}+1}$ simplifies to just

$$w_{i,J^{(i)}+1} = \sum_{h \in I_{uc}^{(i)}} W_h(X_i) l_{ih0} = \sum_{h \in I_{uc}^{(i)}} W_h(X_i) = 1 - \sum_{h \in I_{oc}^{(i)}} W_h(X_i), \tag{3.18}$$

with the last equality being due the fact that $\sum_{h=1}^{\infty} W_h(X_i) = 1$. (Note: this property of KSBP's has not been proven here, but is presented as a theorem in the original paper. Here, we accept it to be true.)

The final necessary definition for sampling is the conditional density

$$f_0(y_i | \mathbf{x}_i) = \int_{\mathbf{\Phi}} f(y_i | \mathbf{x}_i, \phi) dG_0(\phi),$$

which is a marginalization over potential values for the slope of a new cluster. In the case of Model 3.5, since $Y|X_i, \beta_i, \sigma^2, \ldots,$ is a $N(X_i\beta_i, \sigma^2)$ and hence

$$f(y_i|\mathbf{x}_i, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \mathbf{x_i}\phi)^2\right\}. \tag{3.19}$$

Luckily, we do not have to integrate out ϕ in the above expression to figure out the form of f_0 due to our choice for G_0. Since $G_0|\mu_0, \Sigma_0 \sim MVN(\mu_0, \Sigma_0)$ is conditionally conjugate for ϕ with the above likelihood, we know f_0 is the density for a Normal disribution with mean $\mathbf{x}_i^T\mu_0$ and variance $\mathbf{x}_i^T\Sigma_0\mathbf{x}_i + \sigma^2$. Formally,

$$f_0(y_i|\mathbf{x_i}) = \frac{1}{\sqrt{2\pi\left(\mathbf{x}_i^T\Sigma_0\mathbf{x}_i + \sigma^2\right)}} exp\left\{-\frac{1}{2\left(\mathbf{x}_i^T\Sigma_0\mathbf{x}_i + \sigma^2\right)}(y_i - \mathbf{x_i}^T\mu_0)^2\right\}. \tag{3.20}$$

With the quantities defined above, we can sample from the full conditional for S_i according to

$$\begin{aligned} P(S_i = 0|\mathbf{X}, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \mathbf{\Theta}^{(i)}, \mathbf{D}) &\propto w_{i,0} f_0(y|\mathbf{x}_i) \\ P(S_i = j|\mathbf{X}, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \mathbf{\Theta}^{(i)}, \mathbf{D}) &\propto w_{i,j} f(y|\mathbf{x}_i, \theta_j^{(i)}) \\ P(S_i = J^{(i)} + 1|\mathbf{X}, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \mathbf{\Theta}^{(i)}, \mathbf{D}) &\propto w_{i,J^{(i)}+1} f_0(y_i|\mathbf{x}_i). \end{aligned} \tag{3.21}$$

For a finite set of data, $J^{(i)} + 1 \leq n$ is finite and the proportionality constant in this case is just

$$C = \sum_{g=0}^{J^{(i)}+1} P(S_i = g|\mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \mathbf{\Theta}^{(i)}, \mathbf{D})$$

For the most part, $J^{(i)}$ should be much smaller than n, so the calculation of C should not be subject to numerical issues. When $S_i = 1, \ldots, J^{(i)}$ this indicates the assignment of X_i to cluster j and no special steps need to be taken.

If $S_i = 0$ is drawn, then this means that X_i is assigned to a new cluster in an already occupied location. The group assignment for this new cluster is drawn according to

$$P(C_{S_i} = h) = \frac{W_h(X_i)}{\sum_{l \in I_{oc}^{(I)}} W_l(X_i)} \tag{3.22}$$

The use of S_i for the subscript in this expression is intentional since the actual value will be $J^{(i)}+1$ (see below). Essentially, Eq. 3.22 is sampling amongst the occupied groups according to the normalized chance that X_i is assigned to that group since X_i will be the only point assigned to the new cluster.

If $S_i = J^{(i)} + 1$ is drawn, this corresponds to X_i being assigned to a new cluster in an unoccupied location. The group assignment for this new cluster can take on an infinite number of values since $I_{oc}^{(i)}$ is finite and the potential locations are the set $I_{uc}^{(i)} = \{1, 2, \ldots, \infty\} \setminus I_{oc}^{(i)}$. The proposed way in the original paper is to sample the group assignment is through retrospective sampling. It is done as follows

1. Sample $U \sim Unif(0,1)$.
2. Take k to be the first integer such that $\sum_{h=1}^{k-1} W_h(X_i) < U \leq \sum_{h=1}^{k} W_h(X_i)$
3. If $k \in I_{uc}^{(i)}$, then set $C_{J^{(i)}+1} = k$. Otherwise, go back to Step 1 and repeat.

The separation of $S_i = 0$ and $S_i = J^{(i)} + 1$ into two cases was to distinguish between the two possible methods of assiging the new cluster to a group. When actually implementing the sampler in code, S_i should be set to $J^{(i)} + 1$ for both outcomes.

For both the $S_i = 0$ and $S_i = J^{(i)} + 1$ cases, to draw the slope associated with the new cluster, $\theta_{J^{(i)}+1}$, should be drawn according to

$$\theta_{J^{(i)}+1} \propto f(y_i|\mathbf{x}_i, \theta) G_0(\theta) \tag{3.23}$$

In our case, since $f(y_i|\mathbf{x}_i, \theta_{J^{(i)}+1})$, a $N(\mathbf{x}_i^T \theta_{J^{(i)}+1}, \sigma^2)$, and G_0, a $MVN(\mu_0, \Sigma_0)$, are conjugate, we know $\theta_{J^{(i)}+1}$ is distributed according to a multivariate Normal distribution with mean vector and covariance matrix

$$\mu^* = \Sigma^* \left(\Sigma_0^{-1} \mu_0 + \frac{\mathbf{x}_i y_i}{\sigma^2} \right) \tag{3.24}$$

$$\Sigma^* = \left(\Sigma_0^{-1} + \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma^2} \right)^{-1} \tag{3.25}$$

3.2.2 Group Assignments for Each Cluster, C_j

As mentioned at the beginning of this section, the sampling method presented here for the group assignments variables for each cluster deviates from the method described by the original authors; we use a slice sampler instead of a retrospective sampler to update the group assignments for cluster j, C_j. First, the full conditional for C_j is

$$\begin{aligned} P(C_j = h|\mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{\Gamma}, \psi) &\propto \prod_{i:S_i=j} P(X_i \text{ assigned to group } h|\mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{\Gamma}, \psi) \\ &= \prod_{i:S_i=j} \left[V_h K(X_i, \Gamma_h) \prod_{l<h} (1 - V_l K(X_i, \Gamma_l)) \right] \end{aligned} \tag{3.26}$$

While Eq. 3.26 does not look particularly difficult to sample from, remember that h can be any of $\{1, 2, \ldots\}$, which is a countably infinite set of values. To accont for this, and so only finitely many values of V_h and Γ_h need to be stored, we use the following slice sampler. Let p_h denote the quantity in 3.26. Suppose, after t iterations of the sampler, $\theta_j^{(t)} = q$. Then

1. Draw $U \sim Unif\,(0, p_q)$.
2. Let $\mathcal{U}_j^{(t)}$ be the set $\{h : U < p_h\}$.
3. Draw the value for $C_j^{(t+1)}$ uniformly at random from $\mathcal{U}_j^{(t)}$.

Special care needs to be taken in Step 2 when identifying the values of h in $\mathcal{U}_j^{(t)}$. Often, previously unobserved (i.e. new) values of V_h and Γ_h may need to be drawn since p_h must be continuously evaluated until an h^* such that $p_h < U$ for all $h > h^*$ is found. While $|\mathcal{U}_j^{(t)}|$ is almost surely finite and hence h^* is almost surely finite as well, there is no quick way of knowing the value of h^* without knowing the entire sequences $\mathbf{V}$ and $\mathbf{\Gamma}$.

In our implementation, we used the heuristic of stopping at h', where h' is such that $p_{h'} + p_{h'+1} < U$. This heuristic is based on the fact that for large values of h', $p'_h + p_{h'+1} > p_h$ for any $h > h'$ with high probability. This is based on the fact that, while $p_h, h = 1, 2, \ldots$ is not a *strictly* decreasing sequence, it does decrease as along h since $\sum_h p_h < \infty$ and hence the $\lim_h p_h = 0$. Consequently, this implies $h' \geq h^*$ with high probability. Those concerned can stop at h'', where h'' is such that $p_{h''} + p_{h''+1} + p_{h''+2} < U$. Clearly, $h'' \geq h' \geq h^*$ and hence $h'' \geq h^*$ with even higher probability.

3.2.3 Slopes for Each Cluster, θ_j

The slopes for each cluster can be sampled in a way similar to the way of drawing the slope for a new cluster in Eq. 3.23. The value of θ_j should be drawn according to

$$\theta_j \propto \prod_{i:S_i=j} f(y_i|\mathbf{x}_i, \theta) G_0(\theta) \tag{3.27}$$

Note the only difference is the inclusion of more points due to θ_j not necessarily correpsonding to a cluster with only one point assigned to it. By the same conjugacy arguments as before, this means θ_j has a multivariate Normal distribution with mean and covariance matrix

$$\mu_j^* = \Sigma_j^* \left(\Sigma_0^{-1} \mu_0 + \sum_{i:S_i=j} \frac{\mathbf{x}_i y_i}{\sigma^2} \right) \tag{3.28}$$

$$\Sigma_j^* = \left(\Sigma_0^{-1} + \sum_{i:S_i=j} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma^2} \right)^{-1} \tag{3.29}$$

3.2.4 Stick Lengths for Each Group, V_h

The sampling method for the stick lengths, V_h, follows the method presented in the original paper exactly. Adopting their notation, we let $M^{(t)}$ denote the maximum element of I_{oc} across the first t iterations of the sampler. In practical terms, points assigned to newly created groups during one iteration often get reassigned to a larger group during the next iteration of the Gibb's sampler. Despite being empty, both the stick lengths, V_h, and the group centers, Γ_h, for these groups must be sampled during each iteration of the sampler.

The sampling method for V_h relies on a further layer of data augmentation. We introduce two more sequences of Bernoulli random variables - one corresponding to the stick lengths,

V_h, and one corresponding to the effect of the kernel and group centers, Γ_h. Formally, let

$$A_{ih} \sim Bernoulli(V_h) \tag{3.30}$$
$$B_{ih} \sim Bernoulli(K(X_i, \Gamma_h)) \tag{3.31}$$

From the indices, there will be $n \times M^{(t)}$ of these variables, $M^{(t)}$ of them for each data point. These sequences are essentially idicators for the components that make up $W_h(X_i)$. With this in mind, X_i will be assigned to the first group such that both tosses come up heads. Formally, $Z_i = min\{h : A_{ih} = B_{ih} = 1\}$. The graphical representation for this augmented sub-model is provided in Fig. 3.6.

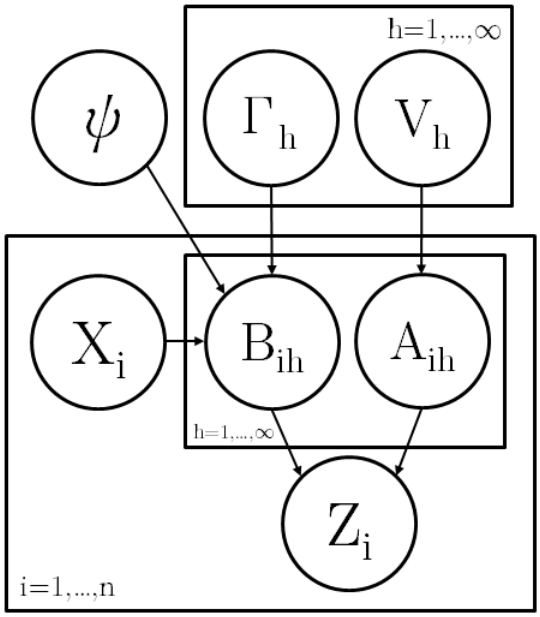


Figure 3.6: Graphical representation of the submodel for $\mathbf{Z}, \mathbf{\Gamma}, \mathbf{V}$ with the augmented variables A_{ih} and B_{ih}. The inclusion of the augmented variables removes the dependence between X_i and $\mathbf{V}$ given the value of Z_i, allowing for simpler sampling of V_h.

To sample V_h, we first draw A_{ih} and B_{ih} from their respective posteriors given X_i and Z_i through the following:

- If $Z_i < h$, then learning the value of Z_i does not provide new information about the values of A_{ih} and B_{ih}, so they are sampled according to their prior distributions in Eqs. 3.30 and 3.31, respectively.
- If $Z_i = h$, then from the definition of Z_i, we know both $A_{ih}, B_{ih} = 1$.
- If $Z_i > h$, then by the way Z_i is defined, this means A_{ih} and B_{ih} cannot both be 1. So jointly draw (A_{ih}, B_{ih}) according to Table 3.1.

(a,b)	$P\left[(A_{ih},B_{ih})=(a,b)\middle\vert Z_i>h\right]$	
$(0,0)$	$\frac{P(A_{ih}=0,B_{ih}=0)}{P((A_{ih},B_{ih})\neq(1,1))}$	$\frac{(1-V_h)(1-K(X_i,\Gamma_h))}{1-V_hK(X_i,\Gamma_h)}$
$(0,1)$	$\frac{P(A_{ih}=0,B_{ih}=1)}{P((A_{ih},B_{ih})\neq(1,1))}$	$\frac{(1-V_h)K(X_i,\Gamma_h)}{1-V_hK(X_i,\Gamma_h)}$
$(1,0)$	$\frac{P(A_{ih}=1,B_{ih}=0)}{P((A_{ih},B_{ih})\neq(1,1))}$	$\frac{V_h(1-K(X_i,\Gamma_h))}{1-V_hK(X_i,\Gamma_h)}$

Table 3.1: Joint distribution table for (A_{ih}, B_{ih}) given $Z_i > h$.

Once A_{ih} and B_{ih} have been sampled from their posterior distributions given the values of Z_i, V_h can be drawn according to

$$V_h|\mathbf{Z}, A_{ih} \sim Beta\left(1+\sum_{i:Z_i\geq h} A_{ih}, \lambda+\sum_{i:Z_i\geq h}(1-A_{ih})\right) \tag{3.32}$$

since the Beta distribution is a conjugate prior for the success chance of a Bernoulli random variable.

3.2.5 Center Locations for Each Group, Γ_h

Here, we use Metropolis-Hastings to take samples from the full conditional for the group centers Γ_h. As with the stick lengths, $\mathbf{V}$, $h = 1, \ldots, M^{(t)} \geq max\{\mathbf{C}\}$, so the group centers for empty groups will be resampled in each iteration. The full conditional for Γ_h is

$$P(\Gamma_h=\gamma|\mathbf{\Gamma}_{-h},\mathbf{V},\mathbf{S},\mathbf{C},\psi) = \prod_{i=1}^{n}\left[V_{Z_i}K(X_i,\Gamma_{Z_i})\prod_{l<Z_i}(1-V_lK(X_i,\Gamma_l))\right]\mathcal{H}(\gamma) \tag{3.33}$$

$$\propto \prod_{i=1}^{n}\left[V_hK(X_i,\gamma)^{\mathbf{1}(Z_i=h)}(1-V_hK(X_i,\gamma))^{\mathbf{1}(Z_i>h)}\right]\mathcal{H}(\gamma) \tag{3.34}$$

where $\mathbf{\Gamma}_{-h}$ denotes the set of all Γ's except for Γ_h and $\mathcal{H}$ indicates the prior distribution on group locations. The proportionality in the second line is due to the variables being conditioned on, $\mathbf{\Gamma}_{-h}$, in particular. Notice that the terms in the product are mutually exclusive; Z_i can only satisfy one or none of the conditions, never both. When $Z_i = h$ and X_i is assigned to group h, it's contribution to the product is $V_hK(X_i, \gamma)$, which encourages values of γ that are close to X_i under K. Intuitively, we would want the center of group h, Γ_h, to be close to the points that are assigned to group h. Conversely, when $Z_i > h$, which means X_i is assigned to a group after h, then the contribution to the product is $1 - V_hK(X_i, \gamma)$, which encourages values of γ that are far from these X_i values. The combination of these two contributions is what should cause different groups to separate themselves. In the case

where $Z_i < h$, that point does not provide any information about the value of Γ_h and it makes no conribution to Eq. 3.34.

Written in the form of Eq. 3.34, it is clear that Γ_h should be sampled one-at-a-time. Luckily, for groups that do not have points assigned into or after them, i.e. when $h \geq max\{\mathbf{C}\}$, then all the terms in the product are 1. In this case, the full conditional reduces to just sampling from the prior, $\mathcal{H}$. Additionally, once the first group with no points assigned into or after it is identified (i.e. the first h such that $h \geq max\{\mathbf{C}\}$, all subsequent groups will also have no points assigned into or after them. This means the group centers for all of those "empty" groups can be sampled at once. In the case where $\mathcal{H}$ is a multivariate Normal, sampling these locations together should provide a significant speed up, especially when there are a lot of such empty groups.

3.2.6 Distance Parameter, ψ

From the graphical model in Fig. 3.5, we know that the full conditional for the distance parameter ψ depends only on $\mathbf{C}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{V}$, and $\mathbf{X}$. The full conditional for ψ is

$$\begin{aligned} p(\psi|\mathbf{V}, \mathbf{\Gamma}, \mathbf{S}, \mathbf{C}, \mathbf{X}) &\propto \left[\prod_{i=1}^{n} P(Z_i = z_i|X_i, \mathbf{V}, \mathbf{\Gamma}, \psi)\right] \pi_\psi(\psi) = \left[\prod_{i=1}^{n} W_{Z_i}(X_i)\right] \pi_\psi(\psi) \\ &= \left[\prod_{i=1}^{n} \left\{ V_{Z_i} K(X_i, \Gamma_{Z_i}) \prod_{l<Z_i} (1 - V_l K(X_i, \Gamma_l)) \right\}\right] \pi_\psi(\psi) \end{aligned} \tag{3.35}$$

where π_ψ denotes the log-Normal prior density specified for ψ in Model 3.5. For clarity, $Z_i = z_i$ in the first line denotes the event the group assignment for X_i is the value of Z_i. A standard Metropolis-Hastings approach should be able to take samples from Eq. 3.35.

3.2.7 Process Variance, σ^2

Given the cluster assignments, $\mathbf{S}$, and the slopes for each cluster, $\mathbf{\Theta}$, we know β_i exactly. From the graphical model in Fig. 3.5, this implies the full conditional for σ^2 depends only on $\mathbf{S}$, $\mathbf{\Theta}$, and the data $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$. The full conditional for σ^2 is

$$P(\sigma^2|\mathbf{S}, \mathbf{\Theta}, \mathbf{X}, \mathbf{Y}) \propto \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, S_i, \mathbf{\Theta}, \sigma^2)\pi_{\sigma^2}(\sigma^2) \tag{3.36}$$

where π_{σ^2} denotes the prior density for σ^2. In our case, $Y_i|X_i, S_i, \mathbf{\Theta}, \sigma^2 \sim N(X_i\theta_{S_i}, \sigma^2)$ and σ^2 has an Inverse-Gamma prior distribution, which is a conjugate prior for the variance of a Normal distribution. This means we know the posterior distribution for σ^2 exactly:

$$\sigma^2|\mathbf{S}, \mathbf{\Theta}, \mathbf{X}, \mathbf{Y} \sim IG\left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^{n} (Y_i - X_i\theta_{S_i})^2\right). \tag{3.37}$$

3.2.8 Mean and Variance of the Base Distribution, μ_0 and Σ_0

Recall from the hyper-Model 3.6, $G_0|\mu_0, \Sigma_0 \sim MVN(\mu_0, \Sigma_0)$. As mentioned earlier, since the G_h's are realizations from a $DP(\alpha G_0)$, this means, marginallly, the θ_j random variables are distributed according to G_0, which is a multivariate Normal. This property gives a simple way to sample from the full conditionals of μ_0 and Σ_0. Since μ_0 and Σ_0 are given conditionally conjugate priors for the multivariate Normal likelihood, there are closed form expressions for the posterior full conditional distributions:

$$\mu_0|\mathbf{\Theta}, \Sigma_0 \sim MVN\left(\left(S_0^{-1} + J\Sigma_0^{-1}\right)^{-1}\left(S_0^{-1}u_0 + J\Sigma^{-1}\overline{\theta}\right), \left(S_0^{-1} + J\Sigma_0^{-1}\right)^{-1}\right) \tag{3.38}$$

$$\Sigma_0|\mathbf{\Theta}, \mu_0 \sim Inv\text{-}Wishart\left(\nu + J, T_0 + \sum_{i=1}^{n}\left(X_i - \mu\right)\left(X_i - \mu\right)^T\right) \tag{3.39}$$

where $J = \#\mathbf{\Theta}$ is the number of clusters and $\overline{\theta} = \frac{1}{J}\sum\theta_j$ is the mean vector of $\mathbf{\Theta}$.

3.3 Posterior Predictive Distribution

Using the Gibb's sampler described in Section 3.2, a draw from the posterior distribution for Model 3.5 given a sample of data $\mathbf{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ would contain the variables $\mathbf{S}, \mathbf{\Theta}, \mathbf{C}, \mathbf{\Gamma}, \mathbf{V}, \psi, \sigma^2, \mu_0, \Sigma_0$. In this section, we will describe how to use these sampled quantities to calculate the posterior distribution for Y given a new set of input values.

For a new point X_{n+1}, the distributions for β_{n+1} and Y_{n+1} given a sample from the posterior of the unaugmented form of Model 3.5 follow

$$Y_{n+1}|X_{n+1}, \beta_{n+1}, \sigma^2 \sim N(X_{n+1}\beta_{n+1}, \sigma^2) \tag{3.40}$$

$$\beta_{n+1}|X_{n+1}, \mathbf{V}, \mathbf{\Gamma}, \mathbf{G}, \psi, \mu_0, \Sigma_0 = \sum_{h=1}^{\infty} W_h(X_{n+1})G_h(\cdot) \tag{3.41}$$

We can separate the sum in Eq. 3.41 into the two cases when $h \in I_{oc}$ and $h \in I_{uc}$

$$\beta_{n+1}|X_{n+1}, \mathbf{V}, \mathbf{\Gamma}, \mathbf{G}, \psi = \sum_{h\in I_{oc}} W_h(X_{n+1})G_h(\cdot) + \sum_{h\in I_{uc}} W_h(X_{n+1})G_h(\cdot). \tag{3.42}$$

Recall, the original purpose of augmenting the model with $\mathbf{S}, \mathbf{C}$ and $\mathbf{\Theta}$ was to avoid the storing or sampling of any of the G_h's, since each one is an infinite mixture. Expressed in terms of our augmented variables, Eq. 3.42 is

$$G_h(\cdot)|\mathbf{S}, \mathbf{C}, \mathbf{\Theta}, \mu_0, \Sigma_0 = \sum_{j:C_j=h} \frac{\sum_i \mathbf{1}\left(S_i = j\right)}{\alpha + \sum_i \mathbf{1}\left(Z_i = h\right)}\delta_{\theta_j}(\cdot) + \frac{\alpha}{\alpha + \sum_i \mathbf{1}\left(Z_i = h\right)}G_0(\cdot). \tag{3.43}$$

The expression in Eq. 3.43 is derived from the Chinese restaurant process. The summation in the left term represents a new customer joining one of the existing tables in group h. The

term on the right is the chance of a new customer sitting down at a new table. Since G_h is a realization from a $DP(\alpha G_0)$, the slope associated with the new table will be a draw from the base distribution $G_0 = MVN(\mu_0, \Sigma_0)$.

From Eq. 3.43, it is clear that for any $h \in I_{uc}$, the set $\mathcal{C}_h = \{j : C_j = h\}$ is empty and hence $\sum_i \mathbf{1}(S_i = j) = 0$ for any $j \in \mathcal{C}_h$ and $\sum_i \mathbf{1}(Z_i = h) = 0$. Consequently, for any $h \in I_{uc}$, $G_h = G_0$.

For convenience, we denote

$$p_{hj} = \frac{\sum_i \mathbf{1}(S_i = j)}{\alpha + \sum_i \mathbf{1}(Z_i = h)}, \tag{3.44}$$

which allows expression of Eq. 3.43 as

$$G_h(\cdot)|\mathbf{S}, \mathbf{C}, \mathbf{\Theta}, \mu_0, \Sigma_0 = \sum_{j \in \mathcal{C}_h} p_{hj}\delta_{\theta_j}(\cdot) + (1 - p_h)\, G_0(\cdot). \tag{3.45}$$

where $p_h = \sum_{j \in \mathcal{C}_\langle} p_{hj}$. Plugging Eq. 3.45 in for $G_h(\cdot)$ in Eq. 3.42 gets

$$\begin{aligned}\beta_{n+1}|X_{n+1}, \mathcal{S} &= \sum_{h \in I_{oc}} W_h(\mathbf{x}_{n+1}) \left[\sum_{j \in \mathcal{C}_h} p_{hj}\delta_{\theta_j}(\cdot) + (1 - p_h)\, G_0(\cdot)\right] + \sum_{h \in I_{uc}} W_h(\mathbf{x}_{n+1}) G_0(\cdot) \\ &= \sum_{h \in I_{oc}} \sum_{j \in \mathcal{C}_h} W_h(\mathbf{x}_{n+1}) p_{hj}\delta_{\theta_j}(\cdot) + \left[\sum_{h \in I_{oc}} W_h(\mathbf{x}_{n+1})(1 - p_h) + \sum_{h \in I_{uc}} W_h(\mathbf{x}_{n+1})\right] G_0(\cdot)\end{aligned} \tag{3.46}$$

where $\mathcal{S}$ is shorthand for the sampled quantities, $\mathbf{V}, \mathbf{\Gamma}, \mathbf{S}, \mathbf{C}, \mathbf{\Theta}, \psi, \mu_0$, and Σ_0.

However, since we avoided sampling of all the V_h and Γ_h values for any $h \geq M$, we cannot draw values of β from the distribution in Eq. 3.46. Instead, notice that by our definitions for p_{hj} and p_h, we must have $\sum_{j \in \mathcal{C}_j} p_{hj} + (1 - p_h) = 1$ and hence

$$\sum_{j \in \mathcal{C}_j} W_h(\mathbf{x}_{n+1}) p_{hj} + W_h(\mathbf{x}_{n+1})(1 - p_h) = W_h(\mathbf{x}_{n+1}).$$

Additionally,

$$\sum_{h=1}^{\infty} W_h(\mathbf{x}_{n+1}) = 1.$$

These two properties allow us to rewrite the distribution in Eq. 3.46 as

$$\beta_{n+1}|X_{n+1}, \mathcal{S} = \sum_{h \in I_{oc}} \sum_{j \in \mathcal{C}_h} W_h(X_{n+1}) p_{hj}\delta_{\theta_j}(\cdot) + \left[1 - \sum_{h \in I_{oc}} \sum_{j \in \mathcal{C}_h} W_h(X_{n+1}) p_{hj}\right] G_0(\cdot), \tag{3.47}$$

which contains only sampled quantities. Combining we can use Eq. 3.47 and the conditional distribution in Eq. 3.40 to get the following posterior predictive distribution for Y_{n+1}

$$Y_{n+1}|X_{n+1},\mathcal{S},\sigma^2 \sim \sum_{h\in I_{oc}}\sum_{j\in\mathcal{C}_h} W_h(X_{n+1})p_{hj}N\left(X_{n+1}\theta_j,\sigma^2\right) + \left[1-\sum_{h\in I_{oc}}\sum_{j\in\mathcal{C}_h} W_h(X_{n+1})p_{hj}\right] N\left(X_{n+1}\mu_0, X_{n+1}^T\Sigma_0 X_{n+1}+\sigma^2\right). \tag{3.48}$$

The normal distribution in the second term is from marginalization of $G_0|\mu_0,\Sigma_0$ (see Eq. 3.20).

3.4 Example: Mixture of Gaussians

In this section, we demonstrate the use of the KSBP-based density regression model to estimate the conditional density for the mixture of Normal densities described by Model 2.47 in Section 2.3.3. Recall that the conditional density (Eq. 2.48) for the model is:

$$f(y|x) = \frac{exp\left(-30x^6\right)}{\sqrt{2\pi(.005)}}exp\left\{-\frac{(y-x)^2}{2(.005)}\right\} + \frac{1-exp\left(-30x^6\right)}{\sqrt{2\pi(.04)}}exp\left\{-\frac{(y-x^4)^2}{2(.04)}\right\}.$$

We generate samples from Model 2.47 by first sampling $x_i, i=1,\ldots,n$ from a $Unif(0,1)$ distribution, and then sampling y_i according to the conditional distributions described in the model. In the context of the density regression model, Model 3.5, we set $Y_i = y_i$ and include an intercept term to the regression component, so $X_i = (1, x_i)$.

The data, $\mathbf{D} = \{(X_i, Y_i), i = 1,\ldots,n\}$, will be a sample of this type of size $n = 75, 150$, and 500. As for the other parameters in Model 3.5:

- a_0 and b_0 are both set to 0.1, which gives a relatively disperse and noninformative prior for the process variance σ^2.
- For the mean and variance for the log-Normal prior on the distance parameter, ψ, we set $\mu_\psi = 2$ and $\sigma^2_\psi = 0.5$.
- Both λ and α, the dispersion parameters for groups and clusters within each group, respectively, are set to 1. This encourages fewer groups and fewer clusters within groups.
- $\mathcal{D}_\Gamma$ is the grid $\{0.00, 0.02, \ldots, 0.98, 1.00\}$ and $\mathcal{H}$ to be the uniform distribution over $\mathcal{D}_\Gamma$.

A brief note on the choice of log-Normal prior parameters for ψ. These are the parameter values suggested in Dunson and Park (2008) for a nearly identical example problem. The method described in Section 3.1.2 was not used to select parameters because the expressions in Eqs. 3.10-3.13 were derived with $\mathcal{H}$ being a normal distribution.

Framing these values in the context of that section, for a $log\text{-}N(2, 0.5)$, the 2.5-th and 97.5-th quantiles are roughly 1.85 and 29.5. While these values seem inordinately high relative to those in Fig. 3.4, that figure is a plot of Eq. 3.14, which averages over a standard Normal distribution for $\mathbf{x}$ and $\mathbf{x}'$. Under that input distribution, the difference between $\mathbf{x}$ and $\mathbf{x}'$ (in one dimension, this is essentially the distance) follows a $N(0, 2)$ distribution. In contrast, when $\mathbf{x}, \mathbf{x}' \sim Unif(0, 1)$, the difference follows a triangular distribution on $-1, 0, 1$. So two points from the uniform input space will be much closer, requiring larger values of ψ for equivalent spatial range.

For the hyper-model on the base distribution of $\mathcal{G} = DP(\alpha G_0)$ (Model 3.6) we set the parameters of μ_0 to be $u0 = \mathbf{0}$ and $S_0 = I_2$. For Σ_0, we take $\nu = 2$ and $T_0 = I_2$. In the Bayesian context, the degrees of freedom for an Inverse-Wishart distribution, ν, corresponds to sample-size used to generate our prior knoweldge. To make make our prior as noninformative as possible, ν should be as low as possible, which is $\nu = 2$. The impact of the choices for S_0 and T_0 will be discussed in Chapter 4.

For the described prior parameter values, the Gibb's sampler described in Section 3.2 was used to produce samples from the posterior distribution for Model 3.5 given a set of observations $\mathbf{D}$. The outcome of these samples for different sized data sets are presented in Figs. 3.7 - 3.10.

In Figs. 3.7 - 3.9, for each posterior sample, Eq. 3.48 is used to calculate the posterior predictive density for $x = 0.1, 0.3, 0.5, 0.7$, and 0.9. In Fig. 3.7, the 75 points that made up the data, $\mathbf{D}$ are plotted in the top left plot. The blue dashed line represents the true conditional density for the generative model, which is given in Eq. 2.48. The red line indicates the pointwise-median posterior predictive density. The dashed black lines indicate pointwise 95% credible intervals for the predictive density. Figures 3.8 and 3.9 contain similar plots, but with datasets of size $n = 150$ and $n = 500$.

In the case of $n = 75$, we see the posterior density manages to get the location of the modes correctly for most values of X and the posterior predictive densities look reasonable (good coverage, correct shape), except when $X = 0.1$ and $X = 0.3$. The posterior predictive densities in these two panels are furthest from the truth due to lacking the height of the true density. This is mostly an effect of the small sample size causing the posterior to lack certainty about the values in that region. For a sample of size $n = 75$ distributed uniformly on the unit interval, there are only 7-15 points near $X = 0.1$ or $X = 0.3$. As n increases to 150 and 500, we see that the height of the posterior predictive densities for those panels increases and become more similar to the true conditional density.

Figure 3.10 contains visualizations of individual realizations for some of the sampled variables of Section 3.2. In the top left plot, points are colored according to their cluster assignments, S_i for a single draw of the posterior. The lines indicate the slopes, θ_j, for each of the clusters. The length of the lines are proportional to the standard deviations of the X

values for points assigned to that cluster. Short lines indicate clusters where the assigned points span a short range. Long lines indicate clusters whose assigned points cover a large area of the input space.

In the top right plot of Fig. 3.10, the points are colored according to the group assignments, Z_i for a single draw of the posterior. The colored squares indicate the location of the centers for each of the groups. The height of the center (recall the centers are only on the X-space) corresponds to the average value of Y_i for X_i's assigned to that group. The shading in the background of the top right plot is indicative of the true mixing probabilities for the two mixtures from the data generating process, Model 2.47.

The plots in the top left and top right represent the same individual draw. However, *the colors do not have any meanings across plots.* That is, the green cluster in the top left does not necessarily have any relationship with the green group in the top right plot. The two plots in the middle are identical to the top two plots, but given a sample of size $n = 150$. The bottom two plots are also identical to the top two plots, but given a sample of size $n = 500$.

There are a few things of note in Fig. 3.10. In every panel, the $KSBP$ model manages to separate the points from each component well. In the group assignment plots, for $n = 75$ and 150, it knows that there are predominantly two main groups - red and green or blue in the top right, and red and yellow for the middle right plot. In the bottom right plot, while there are more occupied groups in this realization, they are pretty well separated - no group spans both mixture components.

In the cluster assignment plots, for points originally from the mixture component with the linear mean (roughly $X \leq 0.5$), when $n = 75$ or 150, it groups nearly all of those points into one cluster with only a single slope. This is sensible because it should only take a singular line to represent a linear process. For points from the mixture component with a quartic mean (roughly $X \geq 0.5$), it assigns those points into multiple smaller clusters with multiple slopes. Again, this is desirable behavior because this indicates that the model realizes that the process generating these points is non-linear and is attempting represent it using multiple locally linear approximations.

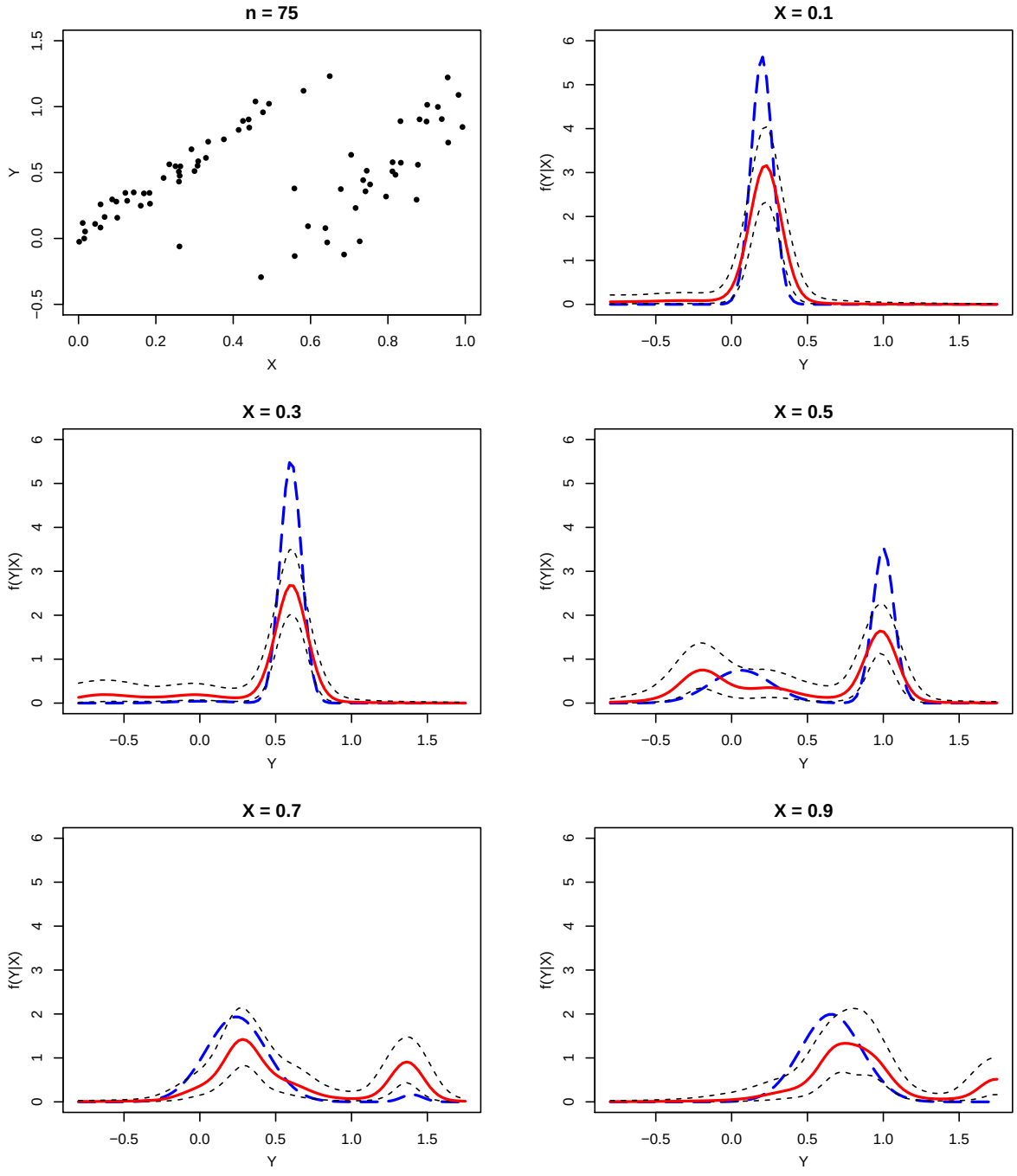


Figure 3.7: Results of fitting the $KSBP$-based density regression model on a sample of size $n = 75$ from Model 2.47. The points in the sample are plotted in the top left panel. The remaining panels contain the predictive density for different values of X. The blue dashed line is the true density. The red is the pointwise posterior median. The dashed black lines are pointwise 95% credible intervals.

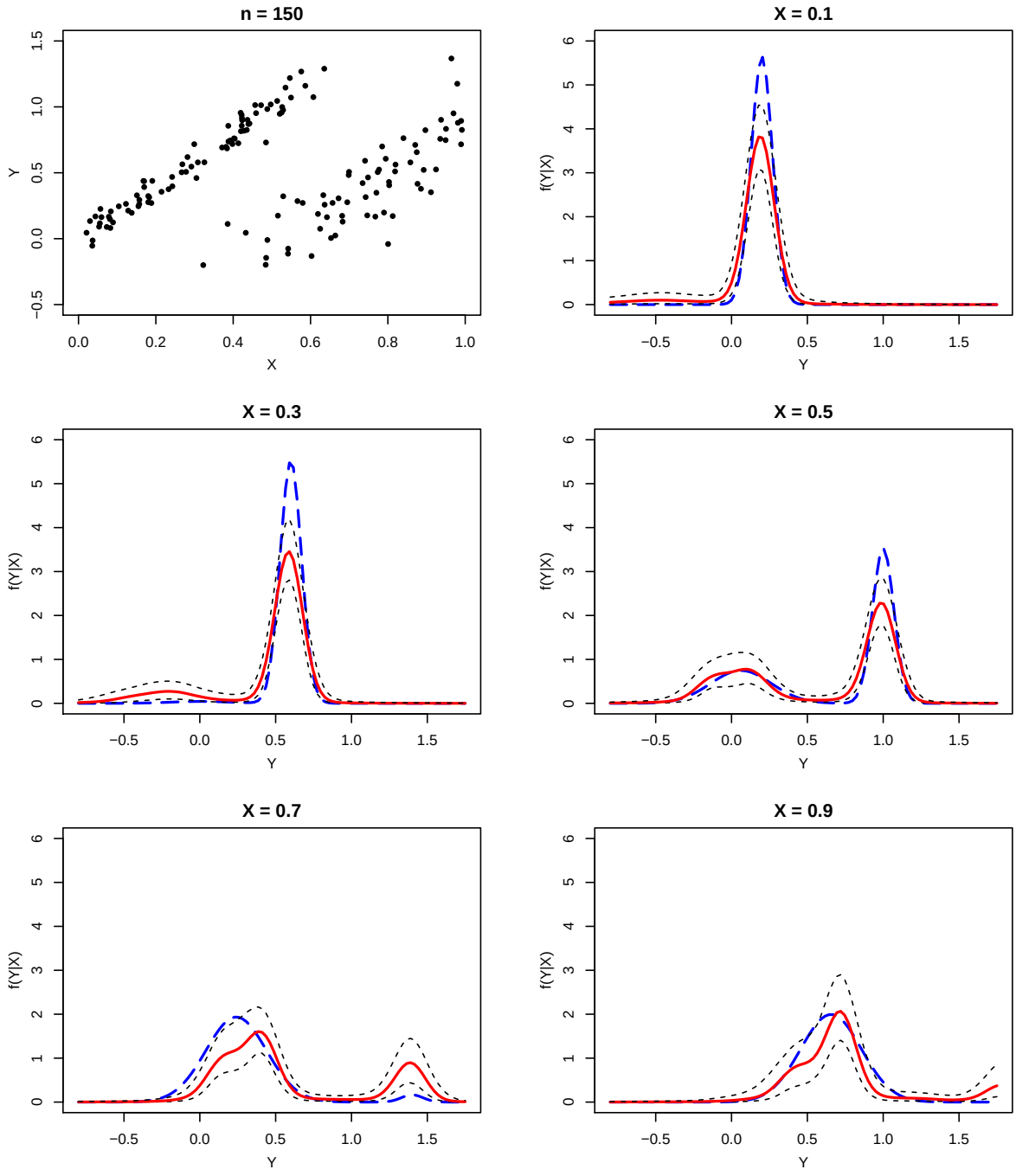


Figure 3.8: Results of fitting the $KSBP$-based density regression model on a sample of size $n = 150$ from Model 2.47. The points in the sample are plotted in the top left panel. Here, the lines have the same meaning as in Fig. 3.7. Blue - true density, red - pointwise median, black - pointwise 95 credible intervals.

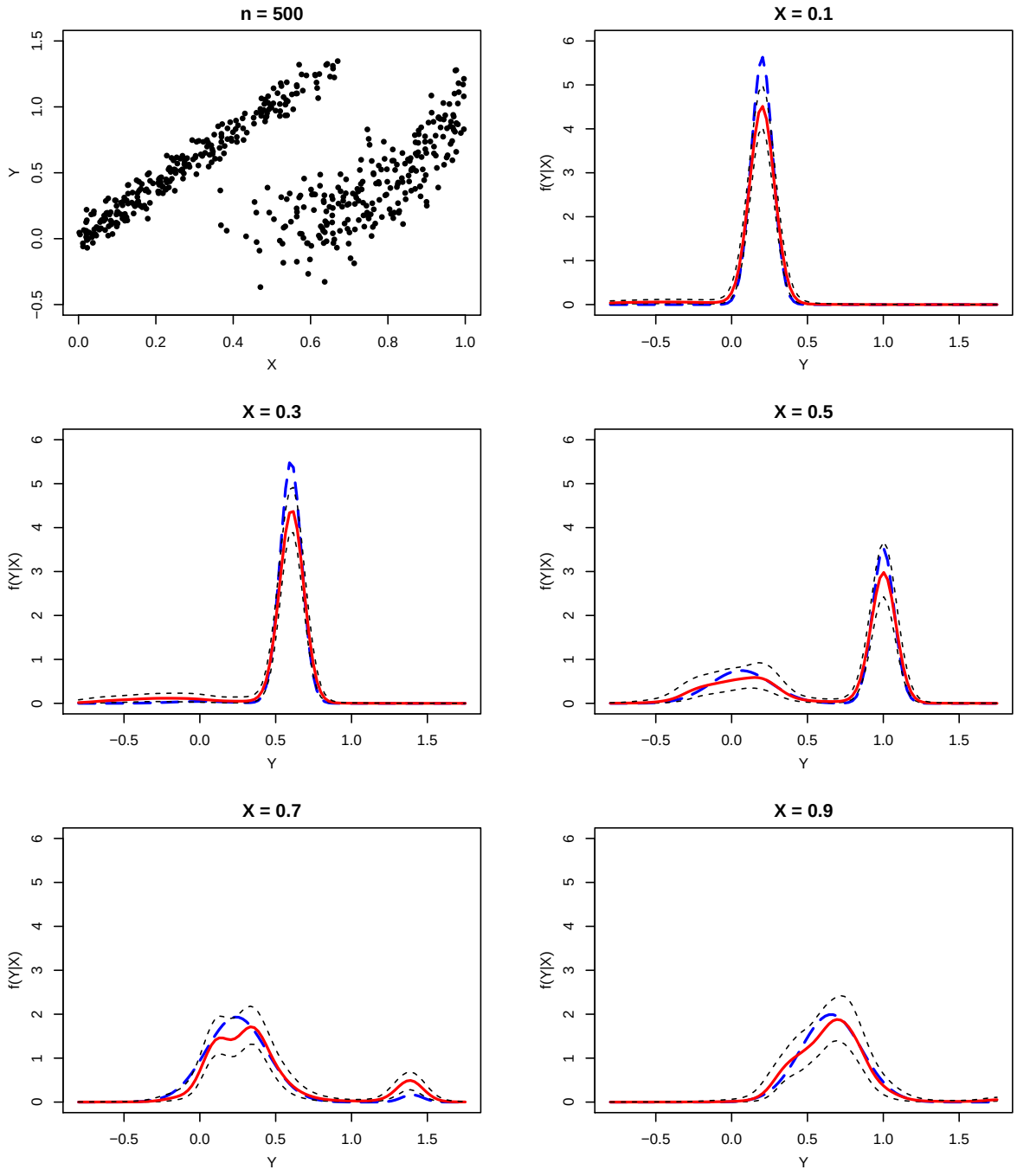


Figure 3.9: Results of fitting the $KSBP$-based density regression model on a sample of size $n = 500$ from Model 2.47. The points in the sample are plotted in the top left panel. Here, the lines have the same meaning as in Fig. 3.7. Blue - true density, red - pointwise median, black - pointwise 95 credible intervals.

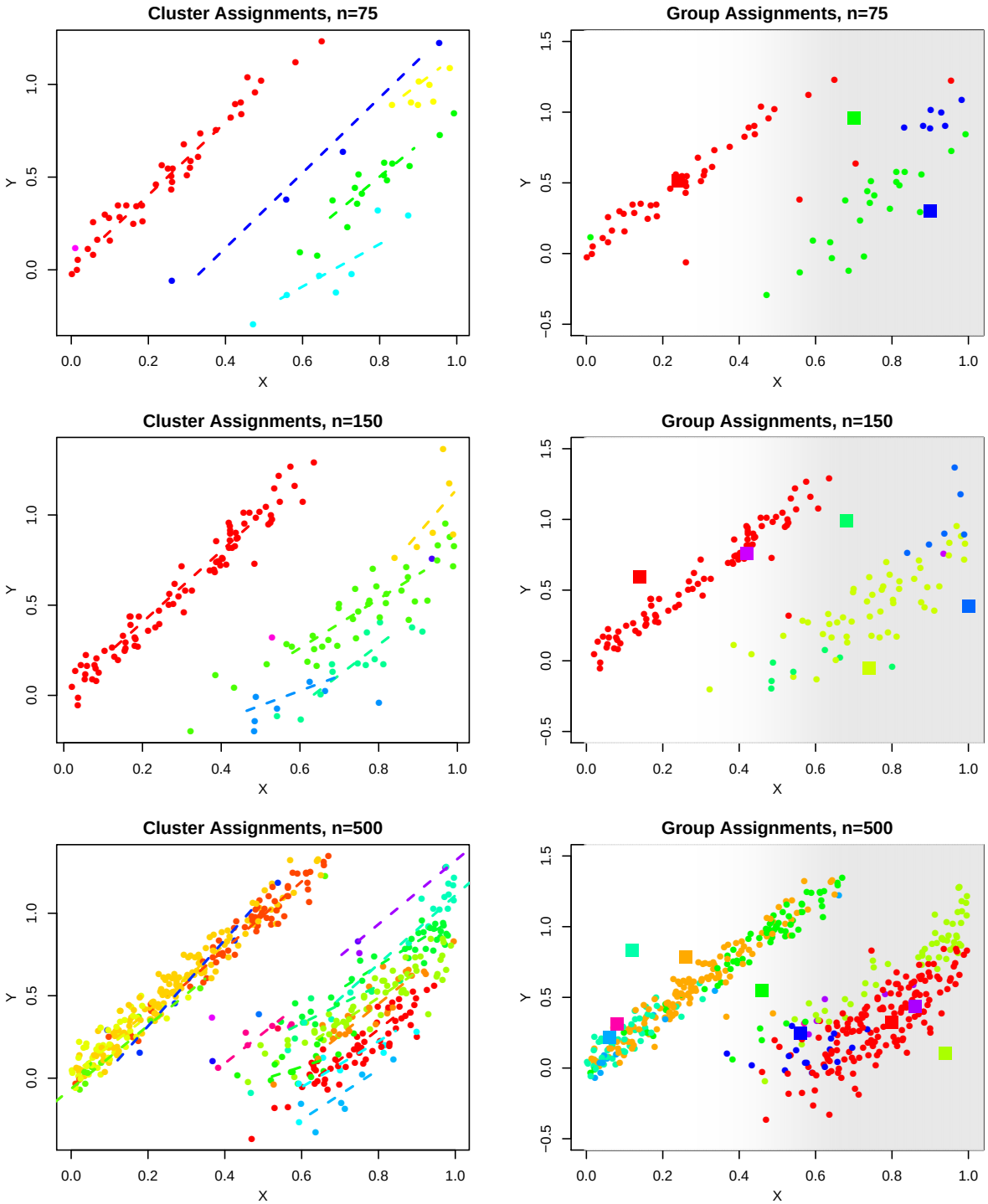


Figure 3.10: Visualizations of individual realizations from posterior distribution given datasets of size 75 (top), 150 (middle), and 500 (bottom). The left plots have points colored according to their cluster assignments (S_i) for the given realization. Lines indicate the slopes (θ_j) for each cluster. Line length is proportional to the SD of the X-values of points in cluster. The right plots have points colored according to their group assignments (Z_i) for the given realization. The colored squares indicate the center of each group (Γ_h). Backround shading corresponds to mixing probabilities in the original model. *Note: The colors do not correspond to similar clusters or groups; they are separate across plots.*

Chapter 4: KSBP for Sensitivity Analysis of Stochastic Functions

In this chapter we present a novel method for conducting sensitivity analysis on a stochastic simulator. We begin by formalizing the problem and then describing the proposed solution. The chapter concludes with an application of the method on a previously seen toy model from Chapters 2 and 3 as well as a short analysis of the impact of the prior choice on the outcome of the study. Application of this technique to a real stochastic simulator is available in Chapter 5.

4.1 Sensitivity Analysis of Stochastic Simulators

In this section we formalize our definition for stochastic simulators and what we mean by sensitivity analysis. This was done briefly in Section 1.2.1, and this section builds upon the ideas introduced there.

Let $\mathbf{X}$ be some space of inputs. For the purposes of this chapter, we let $\mathbf{X} \subseteq \mathbb{R}^p$. However, this may not always be the case when working with stochastic simulators. We will often refer to an element $X \in \mathbf{X}$ as a *set* of inputs because $X = (x_1, x_2, \ldots, x_p)$ is a vector and contains multiple values. In the rare cases when we refer to multiple elements from $\mathbf{X}$, we will use the term *collection* to refer to $\{X_1, X_2, \ldots, X_n\} \subseteq \mathbf{X}$.

A stochastic simulator, F maps elements of $\mathbf{X}$ to output distributions, which we denote $F(X)$. If Y is the output from a single run of F at a set of inputs X, since F is stochastic, Y is a random variable. We write $Y|X \sim F(X)$ to indicate both the randomness in Y and the role of F as the process mapping the inputs, X, to the distribution governing $Y|X$. When the simulator being studied is clear, we omit F and refer to $Y|X$ as the output distribution.

To conduct a first-order global sensitivity analysis on the simulator F, we want to quantify the impact of the one of the input dimentions, say x_j, on the output distribution $Y|X$. In the case where there is an uncertainty distribution for the inputs, $\mathcal{G}_X$, we'd also like to determine how much of the variability or uncertainty in the output distribution $Y|X$ is due to uncertainty in x_i under $\mathcal{G}_X$. Higher order analyses would be centered around the calculation of these quantities for two or more inputs, say $\{x_j, x_k\}$.

Our primary tool for calculating these quantities will be the mutual information. Explicitly, our measure of the sensitivity of the input x_j on the output distribution $Y|X$ will

be

$$I(Y, x_j) = H(Y) + H(x_j) - H(Y, x_j) = H(Y) - H(Y|x_j) \tag{4.1}$$

$$= -\int_{\mathbf{Y}} log\left\{f_Y(y)\right\} f_Y(y)dy + \int_{\mathbf{X}_j}\int_{\mathbf{Y}} log\left\{f_{Y|x_j}(y|x_j)\right\} f_{Y,x_j}(y, x_j)dydx_j \tag{4.2}$$

where $\mathbf{X}_j$ is the domain of input x_j.

Eq. 4.1 provides two equivalent forms for the mutual information. The first form is to emphasize what is actually being calculated - a functional of Eq. 4.5, the joint distribution for (Y, x_j). The second form is to highlight the desired interpretation for the purposes of sensitivity analysis. Recall from Section 2.3, the mutual information between Y and x_j, in the form of the right-most expression of Eq. 4.1, measures how much learning the value of x_j reduces the entropy of Y. Specifically, it is how much of the uncertainty in Y is due to the input uncertainty in x_j, i.e. by $\mathcal{G}_{X_j}$, which is exactly the type of quantity needed for performing global sensitivity analyses for stochastic simulators.

In Eq. 4.2, f_Y, $f_{Y|x_j}$, and f_{Y,x_j} are the the marginal density for Y, the conditional density for $Y|x_j$, and the joint density for (Y, x_j), respectively. Formally, these quantities are

$$f_Y(y) = \int_{\mathbf{X}} f(y|x)g(x)dx \tag{4.3}$$

$$f_{Y|x_j}(y|x_j) = \int_{\mathbf{X}_{-j}} f(y|x)g_{-j|j}(x_{-j}|x_j)dx_{-j} \tag{4.4}$$

$$f_{Y,x_j}(y, x_j) = \int_{\mathbf{X}_{-j}} f(y|x)g(x)dx_{-j} \tag{4.5}$$

where g is the density of the input distribution $\mathcal{G}_X$ and $g_{-j|j}$ is the conditional density for the inputs when conditioning on the value of x_j. Here, x_{-j} is shorthand for all inputs except the j-th one i.e. $x_{-j} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$. Similarly, $\mathbf{X_{-j}}$ is the space of all inputs except for the j-th dimension i.e. $\mathbf{X}_{-j} = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_{j-1} \times \mathbf{X}_{j+1} \times \ldots \times \mathbf{X}_p$.

Of note in Eqs. 4.3-4.5 is the reoccurrence of $\mathcal{G}_X$ and its density g. While $Y|X \sim F(X)$ is random for a deterministic value of X, it is the uncertainty in the inputs X that makes $I(Y, x_j)$ meaningful. If X and hence the x_j's were deterministic constants, then $I(Y, x_j)$ would be identically zero; a deterministic constant does not provide any information about a random variable.

Figure 4.1 illustrates the role of the stochastic simulator F and our desired quantity, $I(Y, x_j)$. The stochastic simulator F maps inputs X to conditional densities. When averaged against an input distribution $\mathbf{G}_X$, these conditional densities define a joint density f_{Y,x_j} (Eq. 4.5) which corresponds to a mutual information.

One point worth addressing is how to evaluate the magnitude of the calculated mutual informations. Depeding on the base of the log being used, the mutual information will be in bits or nats. However, it is not immediately clear how many nats constitutes a high or low level of sensitivity. Critchfield and Willard (1986) and later Lüdtke et al. (2008)

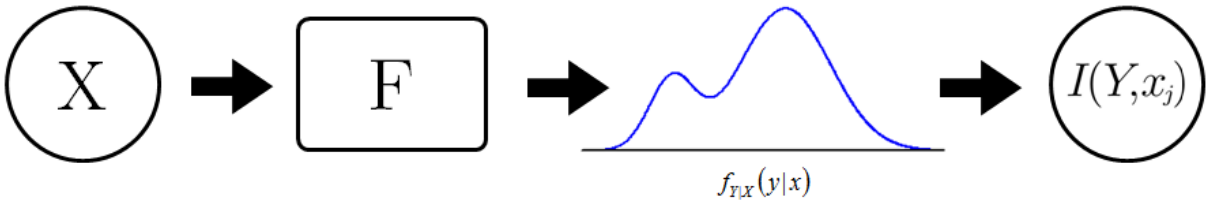


Figure 4.1: Diagram illustrating the relationship between the stochastic simulator F and our desired quantities, $I(Y, x_j)$. F maps inputs X to conditional densities which corresponds to a joint density with mutual informations $I(Y, x_j)$ for $j = 1, \ldots, p$..

have suggested normalizing the mutual information by the marginal entropy of the response, $H(Y)$, to get

$$S_j = \frac{I(Y, x_j)}{H(Y)} \tag{4.6}$$

which they call the *mutual information index*. When dealing with discrete random variables or distributions and using the Shannon entropy, these quantities now reflect percentages - e.g. S_j represents the percentage reduction in entropy caused from learning the true value of x_j. Using the Shannon entropy ensures that $I(Y, X) = H(Y) - H(Y|X)$ will always be less than $H(Y)$ since the Shannon entropy - conditional or otherwise - is always non-negative. However, we are working with continuous random variables and distributions so when we use H to denote the differential entropy which could potentially be negative. This eliminates the interpretation of S_j as a percentage reduction. However, this does suggest a valid way to evaluate the magnitude of the calculated mutual informations - by comparing against the marginal entropy $H(Y)$.

Equations 4.2 - 4.5 may appear cumbersome, but their inclusion is only to clarify what the mutual information represents. In most cases, we never know the form of any of those densities, and so $I(Y, x_j)$ is never calculated directly. For any of our estimators from Section 2.3.2, the only quantities needed to estimate $I(Y, x_j)$ are samples from the joint distribution of (x_j, Y). Given such a sample, any of our estimators from Section 2.3.2 will produce an estimate of $I(Y, x_j)$.

Let $\mathbf{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ denote a collection of n runs from a stochastic simulator F, that is $X_i \sim \mathcal{G}_X$ and $Y_i|X_i \sim F(X_i)$. Since the inputs to F may be a vector, under this notation, $X_i = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})$. Given $\mathbf{D}$, to estimate $I(Y, x_j)$, realize that the set $\mathbf{D}_j = \{(x_{i,j}, Y_i), i = 1, \ldots, n\}$ constitutes a size n sample from the joint distribution of (x_j, Y) which can be used with any of the mutual information estimators. In our case, we will be using the nearest-neighbor estimator in Eq. 2.43. A diagram illustrating this estimation process is provided in Fig. 4.2.

For large enough samples, as demonstrated by the simulation studies in Section 2.3.3, estimates of mutual information calculated in this manner should be close to the true value of $I(Y, x_j)$. In some situations, a point estimate may be all that is desired. However, to conduct

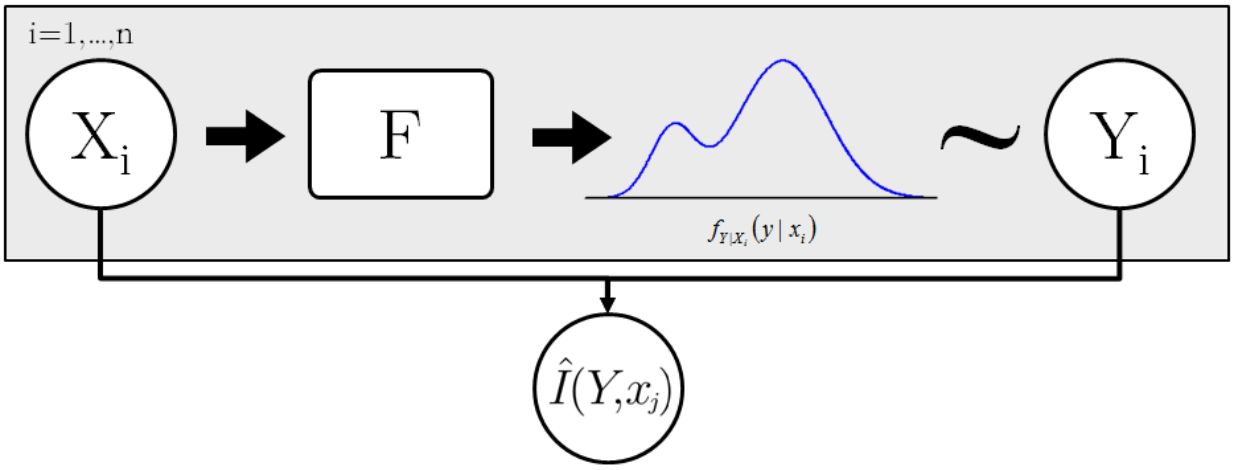


Figure 4.2: Diagram illustrating how a set of observations from F are used to generate the point estimate $\hat{I}(Y, x_j)$.

any meaningful form of inference, a variance or some other measure of uncertainty in the estimated value is required. Since Eq. 2.43 or any other estimate of mutual information, is an estimate of a functional of the densities in Eqs. 4.3 - 4.5, it is difficult to derive expressions for the desired variances.

In situations where the variance is hard to calculate, one common approach is to estimate the variance through resampling of the data. However, since our estimates of mutual information are based on either nearest-neighbor distances or KDE's, it is unclear what effect having replicates in the resampled datasets will have on the desired variance estimates.

4.2 Nonparametric Bayesian Density Regression

To circumvent the inferential difficulties in the above method, we instead propose using the data, $\mathbf{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$, to estimate a model for F, and then calculating the mutual information between Y and the inputs implied by the model. Since stochastic simulators map inputs to a distribution, a model for F will be some kind of density regression model. In our case, we choose to use the KSBP-based Bayesian nonparametric density regression model described in Model 3.5.

Model 3.5 defines a distribution over a collection of X-dependent conditional densities - that is, realizations from this model map values of X to conditional densities for Y. Under the model prior, that is, given no observations, the shape of these conditional densities depends entirely on the prior distributions for the model parameters. Given a set of observations, realizations from the model posterior will produce conditional densities resembling the conditional densities of the generative process, F. As illustrated in Section 3.4, the level of similarity increases as more observations are seen. A set of draws from the model posterior

thus corresponds to a collection of densities resembling the original conditional density of F. The mutual information of Y and x_j calculated from each element of the collection should then be close to the true mutual information, $I(Y, x_j)$, with the dissimilarity decreasing with sample size. A diagram illustrating this process is provided in Fig. 4.3.

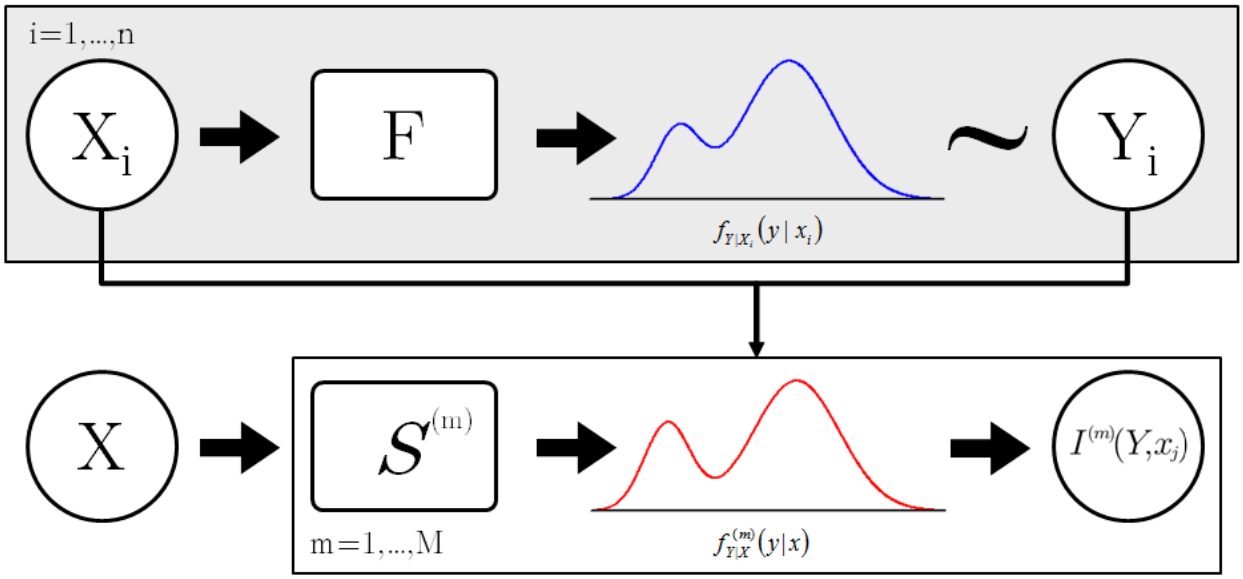


Figure 4.3: Diagram illustrating the benefit of using a density regression model. The sample from F informs our density regression model. Realizations from the model posterior correspond to conditional densities similar to the one generating the original data. Each of these densities has an associated mutual information - the density regression model provides us a distribution for the mutual information.

A nonparametric density regression method is chosen because simulators requiring such in-depth analysis are often complex; for such complicated models, scientists rarely have any ideas about the form of the output distribution. Since we will be estimating the mutual information - a functional of densities - from the fitted model, we must be especially careful about any assumptions made. Density regression models have been studied for a long time and the impact of any assumptions on the resulting densities are well understood. However, the impact of these assumptions on the mutual information of the output density has, to our knowledge, never been investigated and is often unclear. For this reason, we refrain from imposing unnecessary or unvalidated assumptions about the output distribution and employ a nonparametric model.

We choose a Bayesian method for density regression because the posterior distribution from a Bayesian model is a very natural tool for conducting inference - desired variances or intervals are straightforward to calculate from a posterior distribution. As seen in Fig. 4.3, the use of a Bayesian model gives a distribution for the mutual information given the data. Additionally, when dealing with complex simulators, often the amount of data available will

be limited. In these limited data situations, it is clear how the estimated mutual information from a Bayesian model will behave - the posterior distribution will have more uncertainty (given reasonable prior distributions) and the outcome will reflect more of the prior knoweldge. For a frequentist model, it is unclear how limiting the data points available will effect such estimates aside from the it being more likely that the estimated quantities are far from the true value.

Additionally, the Bayesian approach provides a framework for scientists to apply any prior knowledge they have about the simulator and its outputs. While not explicitly used in any of our examples, there are definitely models complicated enough to require prior input and researchers who would benefit from this capability.

One benefit to fitting a model for F given the data is, given a good, well-fitting model, any quantification of dependence or sensitivity can be calculated, not just mutual information. If there is a more appropriate or preferred sensitivity measure for a specific simulator, fitting a density regression model to F allows the calculation of that measure as well.

4.3 Methodology

In this section, we present the following method for conducting Bayesian estimation and inference for $I(Y, x_j)$. As emphasized in Section 4.2, the posterior distribution for Model 3.5 given $\mathbf{D}$ will be a distribution on a set of distributions and draws from the model posterior correspond to realizations of potential distributions.

For each realization from the posterior, Eq. 3.48 gives an analytic form for the posterior predictive of Y given X. While informative, calculating the mutual information between Y and x_j from this posterior predictive requires marginalization of the $\mathbf{X}_{-j}$ dimensions in order to get a conditional or joint density of the forms in Eq. 4.4 or 4.5. The input dependence on the mixing probabilities makes such marginalization too difficult and we instead use the posterior predictive to generate new samples for (Y, x_j). We then estimate the mutual information between Y and x_j for that realization from this new sample. The size of the generated sample is arbitrary and can be large enough to achieve any desired level of accuracy. An illustration outlining the method described in this section is provided in Figure 4.4.

Let $\mathbf{D} = \{(Y_i, X_i), i = 1, 2, \ldots, n\}$ denote a set of observations with $X_i \sim \mathcal{G}_X$ and $Y_i|X_i \sim F(X_i)$. We draw a sample from the posterior distribution for Model 3.5 given $\mathbf{D}$ according to the method described in Section 3.2. Recall that each draw from the posterior distribution consists of the variables $\mathcal{S} = \{\mathbf{V}, \mathbf{\Gamma}, \mathbf{S}, \mathbf{C}, \mathbf{\Theta}, \psi, \mu_0, \Sigma_0, \sigma^2\}$ (Note: this definition of $\mathcal{S}$ differs slightly from the definition Section 3.3 by including σ^2). We denote individual draws from the posterior distribution as $\mathcal{S}^{(m)}, m = 1, \ldots, M$ where M is the total number of posterior samples taken.

For the m-th sample from the posterior, $m = 1, \ldots, M$:

1. Generate a collection of inputs of size N according to the input distribution $\mathcal{G}_X$. We denote this collection of inputs by $\mathcal{X}^{(m)} = \left\{X_1^{(m)}, X_2^{(m)}, \ldots, X_N^{(m)}\right\}$.

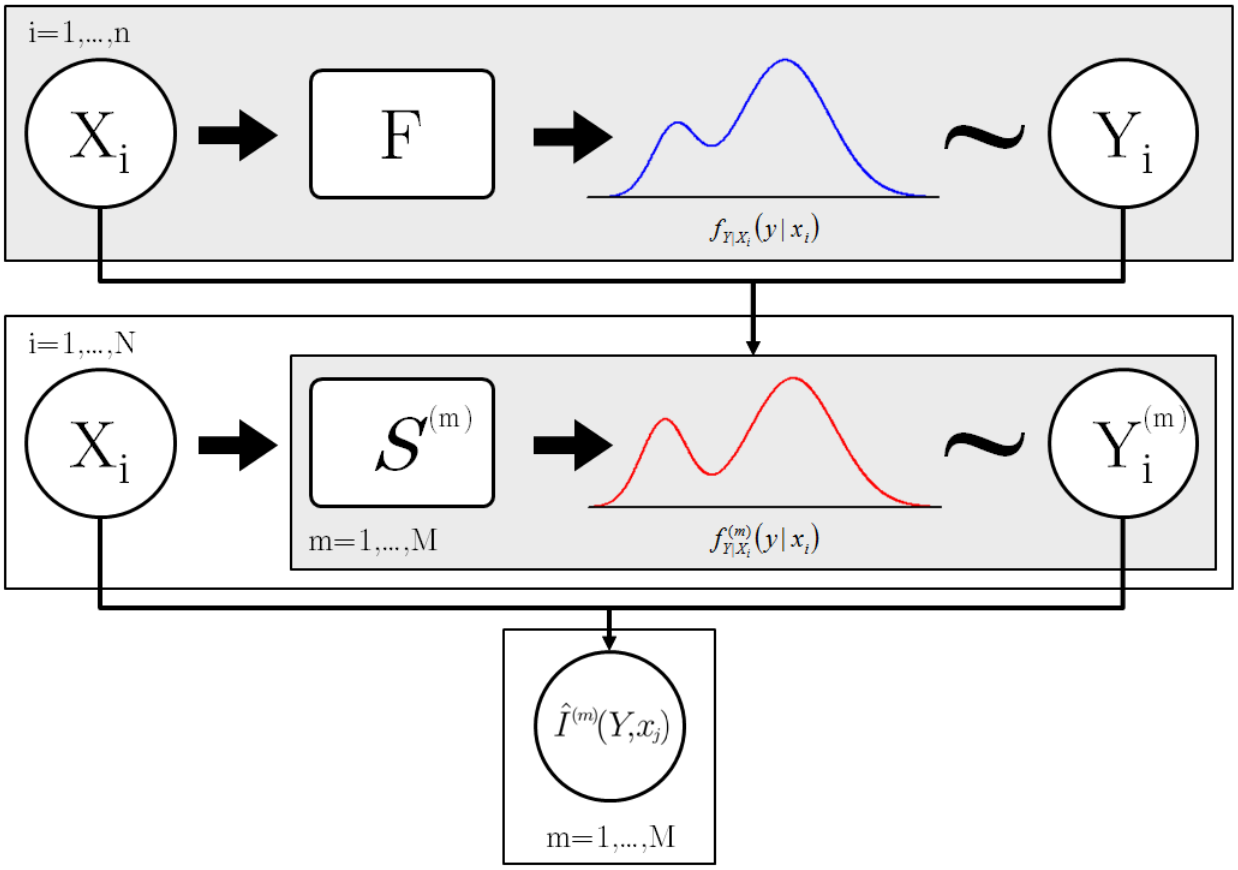


Figure 4.4: Diagram illustrating generation of a sample of mutual informations. Observations from F inform a density regression model. New data samples, $\left(X_i, Y_i^{(m)}\right)$, are generated according to realizations from the model posterior, $S^{(m)}$. The mutual information for each of these new samples is calculated.

2. For each $X_i^{(m)} \in \mathcal{X}^{(m)}$, draw a corresponding $Y_i^{(m)}$ according to the posterior predictive distribution $Y_i^{(m)}|X_i^{(m)}, \mathcal{S}^{(m)}$ as defined by $\mathcal{S}^{(m)}$ and Eq. 3.48.

3. Focusing only on the j-th component and ignoring the remaining dimensions of $X_i^{(m)}$, $(x_{i,1}^{(m)}, \ldots, x_{i,j-1}^{(m)}, x_{i,j+1}^{(m)}, \ldots, x_{i,p}^{(m)})$, the set $\mathcal{D}_j^{(m)} = \left\{\left(x_{i,j}^{(m)}, Y_i^{(m)}\right), i = 1, \ldots, N\right\}$ is a sample of size N from the joint distribution $f_{Y,x_j}^{(m)}(y, x_j)$ (details to follow). So given $\mathcal{D}_j^{(m)}$, we can calculate, using the nearest-neighbor estimator of Eq. 2.43, the estimate

$$\hat{I}^{(m)}(Y, x_j) = \psi(k) + \psi(N) - \frac{1}{N}\sum_{i=1}^{N}\left[\psi(n_{x_j}(i)+1) + \psi(n_y(i)+1)\right], \tag{4.7}$$

In Step 3, the set $\mathcal{D}^{(m)} = \left\{\left(X_i^{(m)}, Y_i^{(m)}\right), i = 1, \ldots, N\right\}$, generated in the manner de-

scribed by Steps 1 and 2, constitutes a sample of size N from the joint distribution

$$f^{(m)}(y, x) = f^{(m)}(y|x)g(x)$$

where g is the density for the input distribuion $\mathcal{G}_X$, and $f^{(m)}(y|x)$ is the density of the posterior predictive distribution $Y|X, \mathcal{S}^{(m)}$. Focusing on the j-th component $x_{i,j}^{(m)}$ and ignoring the other components to get $\mathcal{D}_j^{(m)}$ as described in Step 3 is equivalent to marginalizing over the remaining components. Formally, $\mathcal{D}_j^{(m)}$ is a sample of size N from the joint density

$$f_{Y,x_j}^{(m)}(y, x_j) = \int_{\mathbf{X}_{-j}} f^{(m)}(y, x)dx_{-j} = \int_{\mathbf{X}_{-j}} f^{(m)}(y|x)g(x)dx_{-j}. \tag{4.8}$$

and consequently, the estimates $\hat{I}^{(m)}(Y, x_j)$ calculated from the samples $\mathcal{D}_j^{(m)}$ will be estimates of the mutual information $I^{(m)}(Y, x_j)$ for the joint density of (Y, x_j) given in Eq. 4.8.

The set of densities $\left\{f_{Y,x_j}^{(m)}(y, x_j), m = 1, \ldots, M\right\}$ is a set of realizations from the posterior of our nonparametric density regression model on the process F. This means $I^{(m)}(Y, x_j)$, the mutual information between Y and x_j under the joint density $f_{Y,x_j}^{(m)}(y, x_j)$, is a functional of a realization from the posterior of our model on F. In this sense, the posterior distribution for our model on F implies a distribution on $I^{(m)}(Y, x_j)$ and the set $\mathcal{I}_j = \left\{I^{(m)}(Y, x_j), m = 1, \ldots, M\right\}$ is a sample from this implied distribution.

In Step 1, the value of N, the size of the sample, $\mathcal{D}^{(m)}$, used to estimate $\hat{I}^{(m)}(Y, x_j)$, is an arbitrarily chosen value. From our findings in Section 2.3.3, this means we can make $\hat{I}^{(m)}(x_j)$ as close as we want to $I^{(m)}(Y, x_j)$ by taking N large enough. This means, for a suitably large value of N and enough computing time, the set $\hat{\mathcal{I}}_j = \left\{\hat{I}^{(m)}(Y, x_j), m = 1, \ldots, M\right\}$ should be nearly indistinguishable from the target set, $\mathcal{I}_j$.

Of course, our goal is not to produce a sample resembling $\mathcal{I}_j$, it is to produce a sample that tells us about the true quantity of interest, $I(Y, x_j)$ - the mutual information between Y and x_j under the stochastic simulator F. Whether or not the quantities $\hat{I}^{(m)}(Y, x_j)$ are close to $I(Y, x_j)$ depends entirely on how well our KSBP-based density regression model approximates the true process F. If the posterior distribution for Model 3.5 produces densities close to the true densities of F - that is, if the realizations from the model posterior in $\left\{f_j^{(m)}(Y, x_j), m = 1, \ldots, M\right\}$ are densities close to the joint density $f_{Y,x_j}(y, x_j)$ of Eq. 4.5 - then the set $\hat{\mathcal{I}}_j$ should contain values close to $I(Y, x_j)$. From our example in Section 3.4, we know that one of the factors with the most influence on the goodness-of-fit of the model posterior is n, the number of observed data points used in the computation of the posterior. For enough observations in $\mathbf{D}$, the model posterior for F should produce densities close to the true density and hence the values in $\hat{\mathcal{I}}_j$ should be close to the true value of $I(Y, x_j)$.

4.4 Application to a Toy Model

In this section, we apply the global sensitivity analysis method described in Section 4.3 to the Normal mixture example from Sections 2.3.3 and 3.4, Model 2.47.

In Section 3.4, we took samples from the posterior distribution for Model 3.5 given data sets of size $n = 75, 150$, and 500. Figures 3.7 - 3.9 contain plots of the observed data set (upper left) as well as visualizations of the conditional density associated with each sample from the posterior distribution (remaining panels). These conditional densities are analogous to the red curve in Fig. 4.4. Similarly, the dashed blue line denoting the true density in these panels is analogous to the blue curve in Fig. 4.1 - 4.4.

Using the process described in Section 4.3, a sample of size $N = 2000$ is generated for each draw from the posterior given a set of n observations. From each of these samples, we estimate the mutual information of the conditional density implied by each posterior realization - denoted $\hat{I}^{(m)}$ - with the nearest-neighbor mutual information estimate of Eq. 2.43. This produces a posterior sample for the mutual information between Y and X, which we denote $\hat{\mathcal{I}}$; this is the set $\hat{\mathcal{I}}_j$ from section 4.3 only the j subscript has been omitted since there is only one input parameter to this model.

Since this process was repeated for $n = 75, 150$, and 500, we denote these three different samples of the mutual information $\hat{\mathcal{I}}^{(75)}, \hat{\mathcal{I}}^{(150)}$, and $\hat{\mathcal{I}}^{(500)}$. Figure 4.5 contains plots of the resulting samples. The top left panel contains a plot of all three samples together as well as 95% credible intervals for $I(Y, X)$, indicated by the red shading. The dotted black line indicates the true value of $I(Y, X) = .889$. The solid red line goes through the mean of each sample. The remaining panels are histograms of the estimated mutual informations for the three samples. Again, the red line indicates the sample mean and the dotted black line indicates the true value.

As the number of observations used to fit the model n increases, the mean of the samples gets closer to the true value. This is not surprising considering the conditional densities in Figs. 3.7 - 3.9 also converge toward the true density. As the number of observations increases, the variability in the sampled mutual informations is decreasing. For Bayesian models, as the number of data points increases, the posterior uncertainty should decrease and that is exactly what is happening here - both in the sampled values of $\hat{I}^{(m)}(Y, X)$ and the realizations from the model posterior illustrated in Figs. 3.7 - 3.9.

One troubling behavior illustrated in Fig. 4.5 is the sizeable bias seen at all sample sizes. Even when the model is fit with $n = 500$ data points, the sampled posterior mean is too small by nearly 10%. We conjecture that this underestimation is primarily an effect of the prior distributions of the model. Since these are Bayesian estimates, they will be biased slightly toward the value defined by the prior, with the amount of prior influence depending on the number of observations n. This behavior, while not entirely unexpected, is quite unsettling so Section 4.5 is dedicated to its discussion.

It is worth noting that in other simulation studies (not presented here), the method for sampling the mutual information presented here successfully identifies relevant and irrelevant variables; only the magnitudes are not captured correctly. Additionally, recall that the

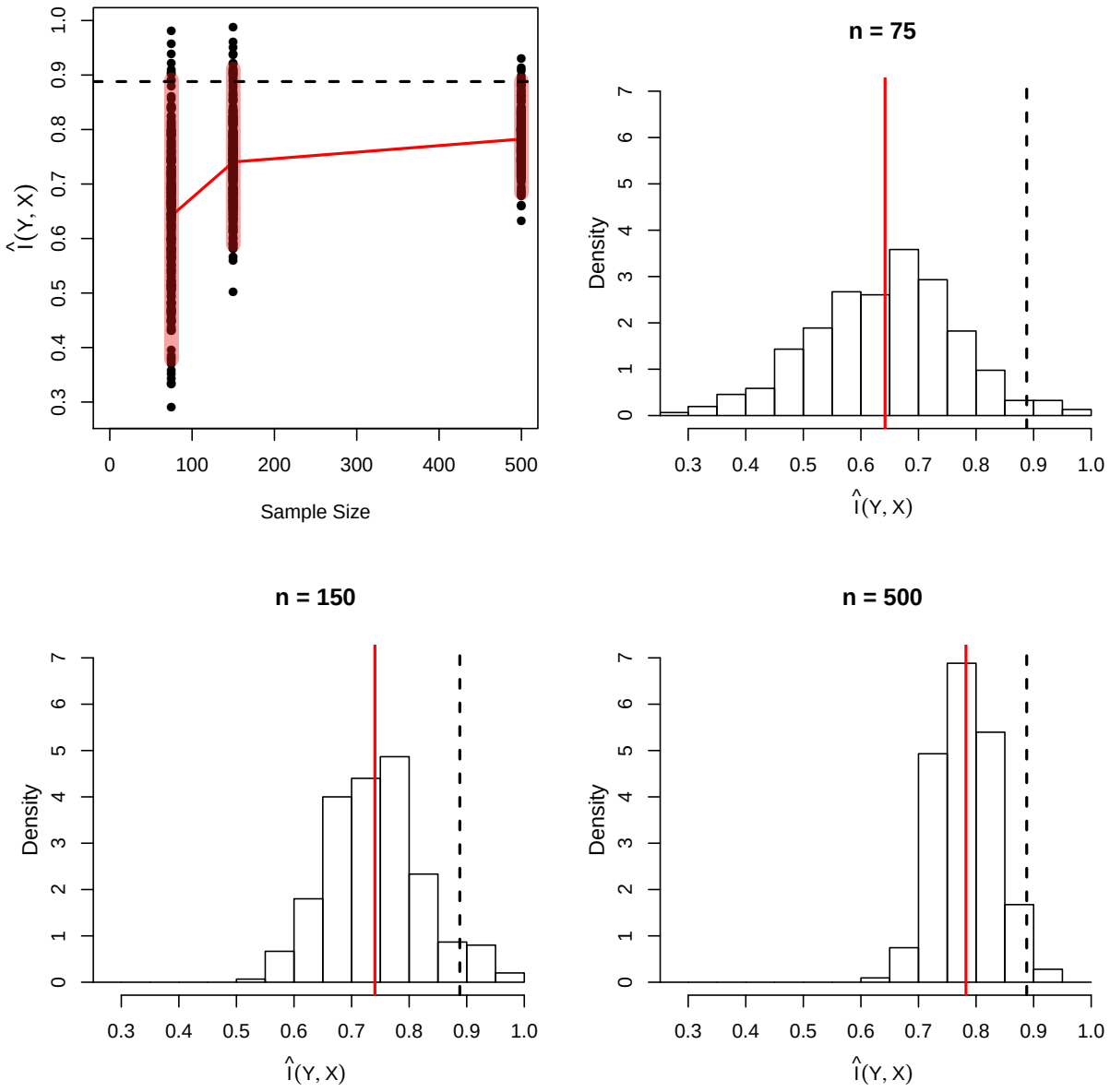


Figure 4.5: Sensitivity analysis using method of Section 4.3 on Model 2.47 for samples of size $n = 75, 150, 500$. The red lines denote the sample means. The dashed black line indicates the true value $I(Y, X) = .8887$. The shaded red areas in the upper-left plot indicate 95% credible intervals.

inputs, X, are distributed uniformly on $[0, 1]$, so the sample size is actually quite limited. To put this in perspective, if instead of estimating a conditional density process, ten separate conditional densities are estimated, corresponding to $[0, 0.1], [0.1, 0.2], \ldots$, then when $n = 75$ or 150, each density would be based on roughly 10-20 points, which is quite limited. That

the posterior credible interval for the mutual information even covers the true value in these cases is impressive.

4.5 KSBP Priors on Mutual Information

In this section, we discuss the impact of parts of the prior distribution on the sampled values in $\hat{\mathcal{I}}_j$ from the model posterior. The ideas covered here are primarily conjectures based on simulation experiments. Consequently, no formal proofs will be provided and the discussion will be centered around providing intuition as to why these conjectures should be true. Where applicable, a potential method of proof will be provided for those interested.

As mentioned in Section 4.4 and illustrated in Fig. 4.5, the sampled values are negatively biased toward the prior mean. In the case of Model 3.5, the prior mean will be close to zero. Given most input distributions $\mathcal{G}_X$, the mutual information should be non-negative, so the prior mean for the mutual information should be positive.

To see why the mutual information under the model prior is small, notice that under the model prior, the conditional distribution for $Y|X$ is an infinite mixture of Normal linear regression models, with the form being similar to the posterior predictive of Eq. 3.48. Because of the Dirichlet process prior on the G_h's, the slope of the regression components in the mixture are i.i.d. realizations from a multivariate normal with mean μ_0, which itself is a random variable with mean $u_0 = \mathbf{0}$. Through the law of iterated expectations, this means the slopes of the regression components have mean zero and, on average, X will not have much effect on Y. Since mutual information is a measure of dependence, on average, the mutual information between X and Y will also be low. This link between the slopes of the regression components and the mutual information is actually quite important, and will be further explored in Section 4.5.2.

Notice that Model 3.5 is overparametrized - in the example of Section 3.4, an infinite mixture of Normal regression models is used to conduct inference on a two component mixture. The prior distribution helps regularize the parameters, but the prior choice affects the types of realizations that are favored in the model posterior - e.g. priors can be chosen to favor realizations with fewer clusters and more process variance. For the purposes of density estimation, the densities from the model posterior are not particularly sensitive to which types of models are preferred - after all, it is an overparametrized model for *densities*. However, it is not necessarily an overparametrized model for mutual information - the mutual information of realizations from the posterior are definitely affected by the types of models, especially when estimated according to our sampling-based method. Consequently, the ensuing sections will be emphasizing the interactions between the different parameters to Model 3.5 and the impact of these interactions on the sampled values of mutual information.

4.5.1 Process Variance Effects (σ^2)

In the Bayesian example of Section 2.3.1, we derived the mutual information of a normal linear model to be (Eq. 2.35):

$$I(X,Y) = \frac{1}{2} log \left\{ \frac{\sigma^2 + \tau^2}{\tau^2} \right\}.$$

Note the notational differences between that example and here - in the above expression, σ^2 is the variance of the mean, which would be equivalent to the variance of θ_j. τ^2 is the process variance, which we now denote σ^2. This expression is decreasing in the process variance, hence overestimation of σ^2, causes underestimation of the mutual information, and vice versa.

The clearest way to visualize the role σ^2 plays in the model posterior is in the cluster assignment plots on the left side of Fig. 3.10. The process variance is determined by the spread of the colored points about the line of that color. If the the lines corresponding to each cluster do not produce accurate approximations of the underlying process, then the process variance will be inflated. As a concrete example, in Model 2.47, when $X \geq 0.5$, Y comes from a Normal distribution whose mean is quartic in X, so the lines for clusters in that region of the input space should be non-parallel in order to approximate the quartic behavior in Y. If they are nearly parallel, then it is not accurately approximating the quartic mean process in that region and the process variance will be inflated to make up for this. In terms of the sampled conditional densities, this effect is barely noticeable. For the sampled mutual informations, the inflated process variance will cause underestimation.

Potential ways to account for this would be to choose a prior that encourages more clusters and consequently more lines in those plots. More clusters would allow for better fitting should reduce the sampled process variance, which would increase the mutual information. A more straightforward way would be to choose a prior that allows more posterior flexibility for the slopes, making it more likely that the slopes associated with each cluster produce accurate approximations.

4.5.2 Slope-Related Effects ($\theta_j, \mu_0, \Sigma_0$)

We've already mentioned the two primary ways the slope of the regression components affects the mutual information: through the prior mean, which the posterior mean for mutual information will be biased towards, and through potential overestimation of the process variance. In this section, we discuss how our prior choices impact the effect of the slope on the mutual information of the model posterior.

In the case of biasing toward the prior mean, the usual way to address this is to make the prior more disperse, since the magnitude of the bias is proportional to the prior mass. However, due to the hierarchical nature of Model 3.5 and the degrees of separation between the slopes and the mutual information, it is not immediately clear how to increase the dispersion of the distribution of mutual information under the model prior.

A naive approach might be to increase the variance of the slopes by encouraging Σ_0 to take on larger values (see Model 3.6). Through marginal properties of Dirichlet processes, this should increase the dispersion in θ_j since $\theta_j \sim MVN(\mu_0, \Sigma_0)$. However, this does not increase the dispersion in the distribution of mutual information under the model prior. Under the model prior, points will be assigned into clusters nearly at random so the mutual information for a realization will be dependent on how similar (i.e. near parallel) the sampled slopes are. Since the prior mean of μ_0 is $u_0 = \mathbf{0}$, θ_j must also have mean zero. Increasing the dispersion in θ_j actually lowers the chance of sampled slopes pointing in similar directions which means the mutual information will be lower.

A better approach to increasing the dispersion of the distribution for mutual information under the model prior would be to increase the dispersion in μ_0 through increasing the diagonals of S_0, the prior variance of μ_0. As mentioned previously, the θ_j will have mean μ_0. Because the prior mean of μ_0 is $u_0 = \mathbf{0}$, increasing the dispersion in μ_0 makes large-magnitude values of μ_0 more likely. For a given value of Σ_0, this makes values θ_j more likely to be pointing in similar directions and makes larger values of mutual information more likely.

One potential benefit of increasing the dispersion in the prior distribution for μ_0 is that it addresses the second way the slopes affect mutual information. As mentioned in Section 4.5.1, poor linear approximations to the true process results will overestimate the process variance causing an artificially lower mutual information. Increasing the dispersion in μ_0 makes the posterior draws of μ_0 less influenced by their prior mean, u_0. This makes it more likely that the Normal linear regression components can accurately approximate the form of the mean for the true generative process.

On this front, encouraging larger entries in Σ_0 encourages potentially different values of θ_j, which also aids in accurate linear approximations and hence should increase the sampled mutual informations for similar reasons. This is contrary to its effect on the distribution for the mutual information under the prior mean, so there is a bit of a trade-off when it comes to the prior for Σ_0 and further investigation could be beneficial.

Worth noting is the effective sample size when considering the effects on the mutual information from the priors for μ_0 and Σ_0. From Eqs. 3.38 and 3.39, the effective sample size for these two parameters is the number of clusters. For the two realizations pictured in the top-left and middle-left panels of Fig. 3.10, the effective sample size for μ_0 and Σ_0 is the same (5) for that realization, depsite the middle-left panel having twice the number of observations. For the most part, the number of clusters will be small, so the influence of the prior distributions for μ_0 and Σ_0 on the sampled values of θ_j will be present in even moderately sized samples. So caution must be exercised when choosing priors based on how they effect the distribution of mutual information from the model posterior.

Chapter 5: Case Study: Sensitivity Analysis for a Stochastic Simulator of Near Fault Ground Motions

In this chapter we demonstrate the application of the sensitivity analysis method from Chaptper 4 on the stochastic simulator for near fault ground motions presented in Dabaghi et al. (2011). Sites in the near-field (< 30 km) region of a fault rupture may experience ground motion with atypically large velocity pulses. These large-amplitude velocity pulses may impose extreme demands on a structure, which warrants investigation into the underlying process. Current simulators do not adequately represent these types of ground motions, although there have been attempts at doing so in recent years, see Shahi and Baker (2011). Of particular interest is their inclusion in probabilistic seismic hazard studies and performance-based earthquake engineering due to their potentially damaging effects on structures, like the study in Taflanidis and Jia (2011). However, our approach differs in that there is less emphasis on risk assessment and more on model understanding.

Note that there are many potential types of ground motions that can occur in the near-field region. For simplicity, the study in this section is focused primarily on the directivity effect of strike-slip near fault ground motions in the strike-normal (SN) direction that have pulse-like behaviors.

This chapter begins with a short description of the conceptual model for ground motion represented in the simulator. Next, the method for determining the distribution of model parameters given a specified set of source and site characteristics is described. Section 5.3 describes how to frame the simulator and generative distribution under the sensitivity analysis methodology developed in Chapter 4. The remaining sections assess the model and present the sensitivity findings.

5.1 Model Description

This section contains a brief description of the conceptual model underlying the ground motion simulator that we use. A detailed description of this model is given in Dabaghi et al. (2011). The conceptual model can be separated into two components - a deterministic

component for modelling the pulse-like behavior in the velocity of a motion and a stochastic component modelling the acceleration of the remaining or residual motion after accounting for velocity pulses. To generate a ground motion from a velocity pulse and a realization of the residual acceration, the derivative of the velocity pulse is added to the residual acceleration to give the total acceleration. Given the total acceleration, it is straightforward to calculate both the total velocity and total displacement.

Although this study only uses the model to generate a very specific type of ground motion - ground motion in the SN direction from strike-slip faults with pulse-like behavior in the velocity - the actual model is quite general; it is a valid model for near-fault ground motions with and without pulse-like behaviors for both the strike-slip and dip-slip faults.

5.1.1 Velocity Pulse Process

The submodel used for the velocity pulse in the model by Dabaghi et al. (2011) is a modified version of the idealized pulse model presented in Mavroeidis and Papageorgiou (2003). The modification made to their model is to ensure that the pulse model achieves zero residual displacement. The form of the modified pulse submodel used in the simulator is

$$v(t) = \begin{cases} \left[\frac{V_p}{2} cos\left(\frac{2\pi(t-t_{max,p})}{T_p} + \nu\right) - \frac{D_r}{\gamma T_p}\right]\left[1 + cos\left(\frac{2\pi(t-t_{max,p})}{\gamma T_p}\right)\right] & \text{if } -\frac{\gamma}{2}T_p < t - t_{max,p} \leq \frac{\gamma}{2}T_p \\ 0 & \text{elsewhere} \end{cases} \tag{5.1}$$

where $(V_p, T_p, \gamma, \nu, t_{max,p})$ are the parameters to this component of the model, which we denote $\mathbf{x}_P$. V_p is the pulse amplitude, T_p is the pulse period, γ is a parameter controlling the number of oscillations in the pulse, ν is the phase angle, and $t_{max,p}$ is the time of the envelope peak.

5.1.2 Residual Acceleration Process

The residual acceleration corresponds to the remaining acceleration after the effect of the velocity pulse has been removed. The submodel used by Dabaghi et al. (2011) for the residual acceleration is a modulated, filtered white-noise process with time-varying filter parameters. The residual acceleration is a realization from the solution to the following stochastic integral

$$a(t) = q(t)\left\{\frac{1}{\sigma_h(t)}\int_{-\infty}^{t} h[t-\tau, \lambda(\tau)]w(\tau)d\tau\right\} \tag{5.2}$$

where w is a white-noise process, $h[t-\tau, \lambda(\tau)]$ is a unit-impulse response function (IRF) for a time-varying filter, $\sigma_h(t)$ is the standard deviation of the integrand, and $q(t)$ is a time modulating function characterizing the root-mean-square of the acceleration. Since we are dividing by the standard deviation, the solution to the stochastic integral has unit variance. $\lambda(\tau) = [\omega_f(\tau), \zeta_f(\tau)]$ denotes the time-varying parameters of the IRF - $\omega_f(\tau)$ is the filter

frequency at time τ and $\zeta_f(\tau)$ is the filter damping at time τ. Since the residual acceleration, $a(t)$, is a realization from a stochastic process, this is what makes both the model and any simulator based on the model, stochastic.

The IRF, h, is chosen according to Rezaeian and Der Kiureghian (2008) and is taken to be

$$h[t-\tau,\lambda(\tau)] = \begin{cases} \frac{\omega_f(\tau)exp\left\{-\zeta_f(\tau)\omega_f(\tau)(t-\tau)\right\}sin\left\{\omega_f(\tau)\sqrt{1-\zeta_f^2(\tau)}(t-\tau)\right\}}{\sqrt{1-\tau_f^2(\tau)}} & \tau \leq t \\ 0 & otherwise \end{cases} \tag{5.3}$$

where $\omega_f(\tau)$ and $\zeta_f(\tau)$ are the time-varying components of $\lambda(\tau)$ described previously. We give ω_f the following linear form

$$\omega_f(\tau) = \omega_{mid} + \omega'(\tau - t_{45}). \tag{5.4}$$

Here, ω_{mid} is the filter frequency at the time to the 45% Arias intensity value of the residual motion. ω' is the rate of change of the frequency over time. $\zeta_f(\tau)$ is taken to be a constant

$$\zeta_f(\tau) = \zeta_f \tag{5.5}$$

The modulating function, q, is proposed by Dabaghi et al. (2011) and has the form

$$q(t) = \begin{cases} 0 & t \leq t_0 \\ c\left(\frac{t-t_0}{t_{max,r}-t_0}\right)^{\alpha} & t_0 < t \leq t_{max,r} \\ c\ exp\left\{-\beta\left(t-t_{max,r}\right)\right\} & t_{max,r} < t \end{cases} \tag{5.6}$$

where t_0 is the time of the peak of the envelope and $t_{max,r}$ is the time of the maximum root-mean-square acceleration. Using the method of Rezaeian and Der Kiureghian (2010), the remaining parameters c, α, and β are determined by a mapping onto the physical parameters I_a, the expected Arias intensity, $D_{5,95}$, the effective duration, and t_{30}, the time to the 30% Arias intensity value. The details of the mapping onto these parameters is omitted for brevity. They can be found both in the original paper as well as in Dabaghi et al. (2011).

From Eqs. 5.2 - 5.6, the parameters to the submodel for residual acceleration are $(I_a, D_{5,95}, t_{30}, t_{max,r}, \omega_{mid}, \omega', \zeta_f)$, which we denote $\mathbf{x}_R$.

5.1.3 Displacement of an Inelastic Single Degree of Freedom Oscillator

The output of the ground motion model is a continuous process over time. In numerical implementations of this model, the simulator outputs a multivariate response, with the actual number of outputs determined by the number of discretization steps. The mutual information based sensitivity analysis method described in Chapter 4 is for simulators with continuous, univariate responses. In order to apply this method to the model described

here, the simulated ground motions must be condensed to a univariate response of some kind. There are a variety of potential univariate responses that can be calculated from simulated ground motion with varying degrees of complexity and usefulness - such as the probability that the peak displacement will be above a value, whether or not a simulated motion is considered pulse-like, etc. In our study, we look at the displacement of an inelastic single degree of freedom oscillator when subjected to the simulated motion. For details on the specific quantity being calcualted, see Dabaghi et al. (2013) and Chopra (2011).

5.2 Parameter Generation

The parameters of the model from Section 5.1 are easy to interpret from a physically, particularly given that some of them are mapped on to physical parameters. However, they are less useful in design or engineering situations because many of these parameters are hard to measure or often unavailable in many cases. In the context of probabilistic seismic studies, these difficulties are worsened since it unlikely for a researcher to know the distribution for parameters whose values are difficult to measure.

Instead, it is preferable to have a method for generating these parameters for a given set of source and site characteristics - magnitude, distance, type of faulting, etc. This is particularly useful because there are limited numbers of recorded ground motions for many potential characteristic combinations, especially when restricting ourselves to near-fault ground motions. This section describes the method developed and used in Dabaghi et al. (2011) to produce a generative distribution for the model parameters given a set of source and site characteristics. Given a collection of recorded ground motions at sites with known site characteristics, the model parameters are estimated for each ground motion in the collection. From these estimated model parameters, an empircal predictive distribution for the model parameters is derived through regression.

The collection of near-fault ground motions used in the derivation of the empirical predictive distribution is from the Pacific Earthquake Engineering Research (PEER)'s Next Generation Attenuation (NGA) database. There are 100 recorded near-fault ground motions in this database with pulse-like behavior in the strike-normal direction.

5.2.1 Pulse Extraction and Estimation

In order to properly estimate the parameters of the two component model described in Section 5.1, it is necessary to separate a measured ground motion into a pulse like component for the velocity and a residual acceleration, corresponding to the two submodels. For a measured ground motion, the measured velocity pulse is extracted from the velocity time history according to a method proposed by Baker (2007). The method iteratively fits wavelet expansions based on 4-th order Daubechies wavelets to the recorded velocity time history to identify the velocity pulse. Complete details on the implementation can be found in the original paper. Figure 5.1 is a visualization of the wavelets used in the pulse extraction

process.

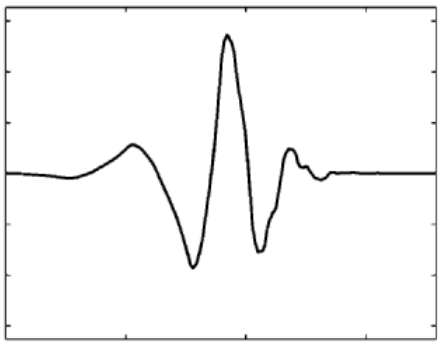


Figure 5.1: A Daubechies wavelet of order 4. Transformed and scaled versions of this wavelet are used in the pulse extraction process. From Baker (2007).

Once the velocity pulse has been extracted from the recorded velocity time history, the parameters to Model 5.1, $(V_p, t_p, \gamma, \nu, t_{max,p})$, are estimated from the extracted pulse through least squares.

For a ground motion record, the residual velocity time history is taken to be the recorded time history minus the modeled (fitted) velocity pulse. This residual velocity time history corresponds to the residual component of the model described by Eqs. 5.2-5.6. The parameters to this model, $(I_a, D_{5,95}, t_{30}, t_{max,r}, \omega_{mid}, \omega', \zeta_f)$ are estimated from this component through a many step process, which is described in Dabaghi et al. (2011) and explained in detail in Rezaeian and Der Kiureghian (2010).

5.2.2 Empirical Predictive Distribution

For each of the 100 ground motions in the database, multiple site and source characteristics are also recorded. The characteristics used in Dabaghi et al. (2011) for the predictive distribution are:

- M_w, the earthquake moment magnitude
- F, an indicator for the type of faulting
- R, the closest distance between the site and rupture, in kilometers
- V_{s30} the shear-wave velocity in the top 30 meters of the soil profile, in meters per second
- s or d, the length or width of the rupture between the hypocenter and the site, in kilometers
- θ or ϕ, the angle between the rupture plane and the direction between the site and hypocenter, in degrees

The last two parameters both depend on the type of fault and occur in pairs - i.e. either (s, θ) or (d, ϕ). For strike-slip earthquakes, θ is the horizontal angle between the rupture plane

and the direction between the site and epicenter. For dip-slip or oblique-slip earthquakes, ϕ is the vertical angle between the rupture plane and the direction between the site and hypocenter. For justifications of the choices of these variables, see Dabaghi et al. (2011), Somerville (2000), Abrahamson et al. (2008), and Rezaeian and Der Kiureghian (2010).

With a set of model parameters, $X_i, i = 1, \ldots, 100$ and source and site characteristics $W_i = (M_{i,w}, F_i, R_i, V_{i,s30}, s_i, \theta_i)$ in the strike-slip case or $W_i = (M_{i,w}, F_i, R_i, V_{i,s30}, d_i, \phi_i)$ in the dip/oblique-slip case, $i = 1, \ldots, 100$, it is possible to derive a predictive distribution for X_{i+1} for a new set of source and site characteristics W_{i+1}.

First, the components of X_i are transformed to the standard normal space through the transformation

$$z_{i,j} = \Phi^{-1}\left(F_j(x_{i,j})\right) \tag{5.7}$$

where $x_{i,j}$ is the j-th component of X_i and F_j is the marginal c.d.f. for the j-th component of X_i. The family for each of the F_j's is chosen through visual inspection, with the parameters for that family being estimated through maximum likelihood. The specific families and estimated parameters as well as more details on the transformation process can be found in Dabaghi et al. (2011).

If F_j accurately represents the true marginal c.d.f. of the components of X_i, then the transformation in Eq. 5.7 should produce $Z_i = (z_{i,1}, \ldots, z_{i,12})$, which has marginally standard normal components. Each component is then regressed one at a time according to one of the following regression models

$$\begin{aligned} z_{i,j} = \beta_{j,0} + \beta_{j,1}F_i + f_{M,j}(M_{w,i}) + f_{R,j}(R_i) + f_{MR,j}(M_{w,i}, R_i) \\ + f_{V,j}(V_{s30,i}) + f_{dir,j}(\theta_i, s_i) + \epsilon_{i,j} \end{aligned} \tag{5.8}$$

$$z_{i,j} = \beta_{j,0} + \beta_{j,1}F_i + f_{M,j}(M_{w,i}) + f_{R,j}(R_i) + f_{MR,j}(M_{w,i}, R_i) + f_{V,j}(V_{s30,i}) + \epsilon_{i,j} \tag{5.9}$$

with Eq. 5.8 being the model used for the components of the pulse parameters $\mathbf{z}_{i,j}$, $j = 1, \ldots, 5$ and Eq. 5.9 the model for components of the residual parameters $\mathbf{z}_{i,j}$, $J = 6, \ldots, 12$. The f functions in Eq. 5.8 and 5.9 are defined as

$$\begin{aligned} f_{M,j}(M_w) &= \beta_{j,2}M_w + \beta_{j,3}M_w^2 \\ f_{R,j}(R) &= \beta_{j,4}R + \beta_{j,5}log(R) + \beta_{j,6}log\left(R^2 + 10\right) \\ f_{MR,j}(M_w, R) &= \beta_{j,7}M_w log(R) \\ f_{v,j}(V_{s30}) &= \beta_{j,8}V_{s30} + \beta_{j,9}log(V_{s30}) \\ f_{dir,j}(\theta, s) &= \beta_{j,10}\theta + \beta_{j,11}s \end{aligned} \tag{5.10}$$

In the last equation in Eqs. 5.10, θ denotes both angles θ and ϕ and s denotes both rupture lengths s and d due to the limited number of data points. There are physical interpretations bethind the expressions in Eqs. 5.10, with a detailed explanation of these interpretations being available in Dabaghi et al. (2011).

If we let $\hat{z}_{i,j}$ denote the fitted value from the regression model, then the regression residual is $e_{i,j} = z_{i,j} - \hat{z}_{i,j}$. To account for any potential correlations in the model parameters, a

correlation matrix can be estimated from these regression residuals. The estimated regression matrix would have entries

$$\Sigma_{h,k} = \frac{\sum_{i=1}^{100} e_{i,h} e_{i,k}}{\sqrt{\left(\sum_{i=1}^{100} e_{i,h}^2\right)\left(\sum_{i=1}^{100} e_{i,k}^2\right)}} \tag{5.11}$$

The estimated values of the β's and correlation matrix from the collection of 100 ground motions are not reproduced here, but can be found in Chapter 3 of Dabaghi et al. (2011). Given a new set of source and site characteristics, W_{i+1}, a generative distribution for Z_{i+1}, the model parameters in the normal space is

$$Z_{i+1}|W_{i+1} \sim MVN\left(\hat{\mathbf{z}}_{i+1}, \Sigma\right) \tag{5.12}$$

where Σ is the correlation matrix with entries defined by Eq. 5.11 and $\hat{\mathbf{z}}_{i+1}$ is the vector of predicted regression values. That is,

$$\begin{aligned}\hat{z}_{i+1,j} = \hat{\beta}_{j,0} + \hat{\beta}_{j,1} F_{i+1} + \hat{f}_{M,j}(M_{w,i+1}) + \hat{f}_{R,j}(R_{i+1}) + \hat{f}_{MR,j}(M_{w,i+1}, R_{i+1}) \\ + \hat{f}_{V,j}(V_{s30,i+1}) + \hat{f}_{dir,j}(\theta_{i+1}, s_{i+1})\end{aligned} \tag{5.13}$$

for $j = 1, \ldots, 5$ with $z_{i+1,j}$ defined similarly when $j = 6, \ldots, 12$. To generate model parameters in the unnormalized space, first generate normalized model parameters according to Eq. 5.12 and then transform the normalized variables using the inverse transformation of Eq. 5.7.

5.3 Experimental Setup

In this section, we demonstrate how the simulator and generative distributions described in Sections 5.1 and 5.2 fit into the framework developed in Chapter 4. We begin with a description of the dataset used in the study and then follow that by clear definitions of both the input distribution and the stochastic simulator being studied. Using the estimated regression parameters and correlation matrix from Section 5.2, a collection of $n = 1000$ model parameters in the transformed space were taken according to the multivariate Normal distribution in Eq. 5.12 with the following source and site characteristics

$$\begin{aligned} F &= 1 & V_{s30} &= 400 \text{ m/s} \\ M_w &= 6.5 & R &= 10 \text{ km} \\ s &= 30 \text{ km} & \theta &= 20^o \end{aligned} \tag{5.14}$$

We denote each element of the collection of model parameters as $X_i, i = 1, \ldots, 1000$. The model parameters will be in the transformed normal space, not in the original parameter space. For a set of model parameters, X_i, a realization from the model described in Section 5.1 is drawn. Using this generated ground motion, the displacement of an inelastic single

degree of freedom oscillator is calculated, which we denote Y_i. The initial period of oscillator being studied is 2 seconds. In the normalized space, this corresponds to a value of roughly -0.243. The yield displacement of the oscillator is 0.15 meters.

Figure 5.2 is an illustration of the different parameters and their roles in this process. The red box encloses what is considered the stochastic simulator (denoted F in Chapter 4) when viewed in this manner.

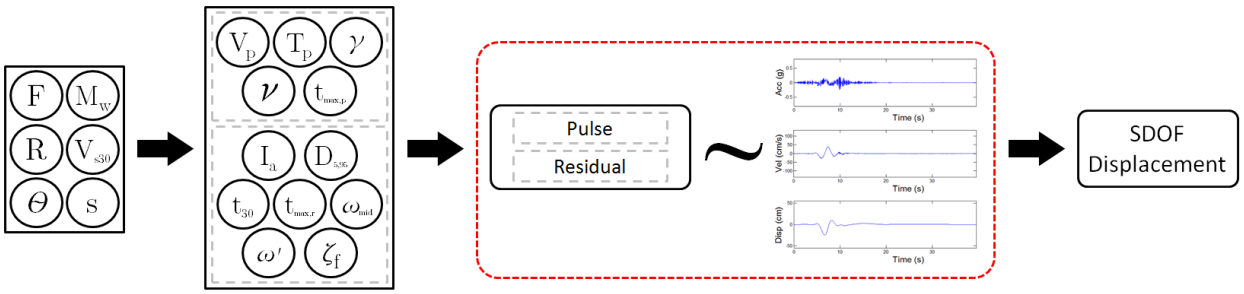


Figure 5.2: Diagram of the full simulation process when considering all 12 model parameters. Site characteristics define a distribution for model parameters. These model parameters define a distribution for ground motions, and one draw is taken from this distribution. Using the simulated ground motion, the displacement of an inelastic single degree of freedom oscillator is calculated.

However, there are 12 model parameters and only 1000 (X_i, Y_i) observations in our data set, which is quite limited. To address this, we restrict the input space to five parameters of interest: $T_p, V_p, I_a, D_{5,95}, \omega_{mid}$ on the transformed space. To emphasize the transformation, we write $\tilde{T}_p, \tilde{V}_p, \tilde{I}_a, \tilde{D}_{5,95}, \tilde{\omega}_{mid}$. For convenience, we write $X_i = (\mathbf{x}_{i,I}, \mathbf{x}_{i,R})$ where $\mathbf{x}_{i,I}$ and $\mathbf{x}_{i,R}$ denote the parameters of interest (I) and the remaining parameters (R), respectively. Similarly, we also denote the mean values of the transformed model parameters as $\mu_x = (\mu_{x,I}, \mu_{x,R})$. The exact value of μ_x can be calculated through Eq. 5.13. However, we will not just ignore or fix the remaining seven parameters of the model. Instead, we incorporate their distributions as another component of the "stochastic simulator" being studied.

From Eq. 5.12, we know that $X_i = (\mathbf{x}_{i,I}, \mathbf{x}_{i,R})$ follows a multivariate normal distribution with mean $\mu_x = (\mu_{x,I}, \mu_{x,R})$ and covariance matrix which we write as

$$\Sigma = \begin{bmatrix} \Sigma_I & \Sigma_{IR} \\ \Sigma_{IR} & \Sigma_R \end{bmatrix}$$

where Σ_I and Σ_R denotes the covariance matrices for the parameters of interest and the remaining parameters, respectively, and Σ_{IR} denotes the matrix of covariances between the two parameter sets. Using the Schur complement, we know the exact distribution for $\mathbf{x}_{i,R}$ given the values of $\mathbf{x}_{i,I}$

$$\mathbf{x}_{i,R}|\mathbf{x}_{i,I} = x_{i,I} \sim MVN\left(\mu_{x,R} + \Sigma_{IR}\Sigma_I^{-1}\left(x_{i,I} - \mu_{x,I}\right), \Sigma_R - \Sigma_{IR}\Sigma_I^{-1}\Sigma_{IR}\right) \tag{5.15}$$

We can think of the stochastic simulator generating Y_i for a given set of parameters of interest, $\mathbf{x}_{i,I}$ as a black box that first generates $\mathbf{x}_{i,R}$ according to the conditional distribution

of Eq. 5.15. The remaining steps are now the same - based on the 12 parameters $(\mathbf{x}_{i,I}, \mathbf{x}_{i,R})$ and the model in Section 5.1, a ground motion is simulated. Using this simulated ground motion, the displacement of an inelastic single degree of freedom oscillator is calculated. Figure 5.3 contains a visualization of this process. The red box denotes what is considered the stochastic simulator, F, under this point of view.

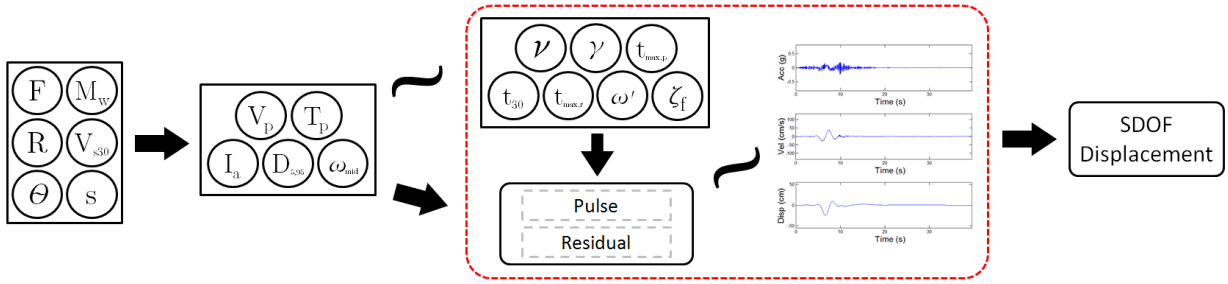


Figure 5.3: Diagram of the simulation process when considering only the five parameters of interest. Site characteristics define a distribution for these five parameters. Conditional on these parameters, there is a known distribution for the remaining seven parameters and a realization of is taken from this distribution. Together, these twelve parameters define a distribution for ground motions and a draw is taken from this distribution. From the simulated ground motion, the displacement of an inleastic single degree of freedom oscillator is calculated.

Note that expanding what is considered the "stochastic simulator" in this manner is preferable over fixing the remaining parameters because the sampled ground motions will keep the same amount of variability. This expansion is similar to marginalizing over the parameters of the submodel for residual acceleration. However, there will be some differences in the interpretation of the calculated mutual information - since they are a quantification of the dependence between the parameters and the distribution of the output, some of the dependence between the parameters of the velocity pulse parameters and the residual acceleration parameters will be reflected in the sampled mutual informations. For the most part, the estimated correlations are small (see Dabaghi et al. (2011) for the exact values) and there are many degrees of separation between the parameters and the final calculated displacement, so the effect of this dependence should be small relative to their effect on the measured displacement and worth the reduction in dimension.

In summary, the inputs to the simulator are the 5 parameters of interest $\mathbf{x}_{i,I} = (\tilde{V}_{p,i}, \tilde{T}_{p,i}, \tilde{I}_{a,i}, \tilde{D}_{5,95,i}, \tilde{\omega}_{mid,i})$. The predictive or generative distribution for these inputs is

$$\mathbf{x}_{i,I} \sim MVN(\mu_{x,I}, \Sigma_I)$$

where $\mu_{x,I}$ and Σ_I are defined as before. The stochastic simulator, F, is the process mapping these five inputs to a distribution for ground motions based on the model of Dabaghi et al. (2011) in Section 5.1. Using the simulated motion, the displacement of an inelastic single degree of freedom oscillator is calculated, which we denote Y_i. We have a set of 1000 $(\mathbf{x}_{i,I}, Y_i)$ realizations from the stochastic simulator. Figure 5.4 contains a histogram of the observed Y_i values and also marginal plots of the displacement Y_i against the five parameters of interest.

We see that V_p and T_p should have large effects compared to the others due to their noticeable effect on the mean. It also seems that especially large values of T_p cause lower variance. I_a has a moderate effect on the mean.

The goal of our sensitivity study is to get posterior distributions for mutual informations $I(Y, V_p), I(Y, T_p), I(Y, I_a), I(Y, D_{5,95})$, and $I(Y, \omega_{mid})$. We will follow the method described in Section 4.3 - draws from the model posterior of Model 3.5 given the 1000 observations are used to get samples of the desired mutual informations.

For our prior parameters, we used

- a_0 and b_0 are both set to 0.1, which gives a relatively disperse and noninformative prior for the process variance σ^2.
- For the mean and variance for the log-Normal prior on the distance parameter, ψ, we set $\mu_\psi = -1.4067$ and $\sigma^2_\psi = 0.515$.
- Both λ and α, the dispersion parameters for groups and clusters within each group, respectively, are set to 1. This encourages fewer groups and fewer clusters within groups.
- $\mathcal{D}_\Gamma$ is $\mathbf{R}^5$ and $\mathcal{H}$ is a mean zero multivariate normal with identity covariance.
- The mean of μ_0 is $u_0 = \mathbf{0}$ and we took the covariance matrix to be the identity (marginal variance of 1).
- For Σ_0, we take $\nu = 6$, the smallest possible value, and the mode T_0 to be the identity matrix.

The prior parameters for the distance parameter, ψ were chosen according to the method described in Section 3.1.2. The chosen parameters correspond to a log-Normal whose 2.5th and 97.5th quantiles are 0.06 and 1, respectively. In terms of the expected correlation, those values of ψ correspond to correlations of nearly 1 and .1. Note that the input distribution is not a standard multivariate normal, so the values in Fig. 3.4 do not apply exactly. Nonetheless, they should still provide a general idea of the expected correlation.

5.4 Posterior Model Assessments

In this section, we look at a few diagnostics from the model posterior to ensure that draws from the model posterior are producing reasonable conditional densities. After all, the accuracy of our mutual information estimates depends largely on how well these conditional densities reflect the conditional densities generated by the true process.

Figures 5.5-5.9 contain the posterior predictive distributions for Y for different values of each of the parameters of interest. The values chosen are two standard deviations below the mean (top right), one standard deviation below the mean (middle left), the mean (middle right), one standard deviation above the mean (bottom left), and two standard deviations

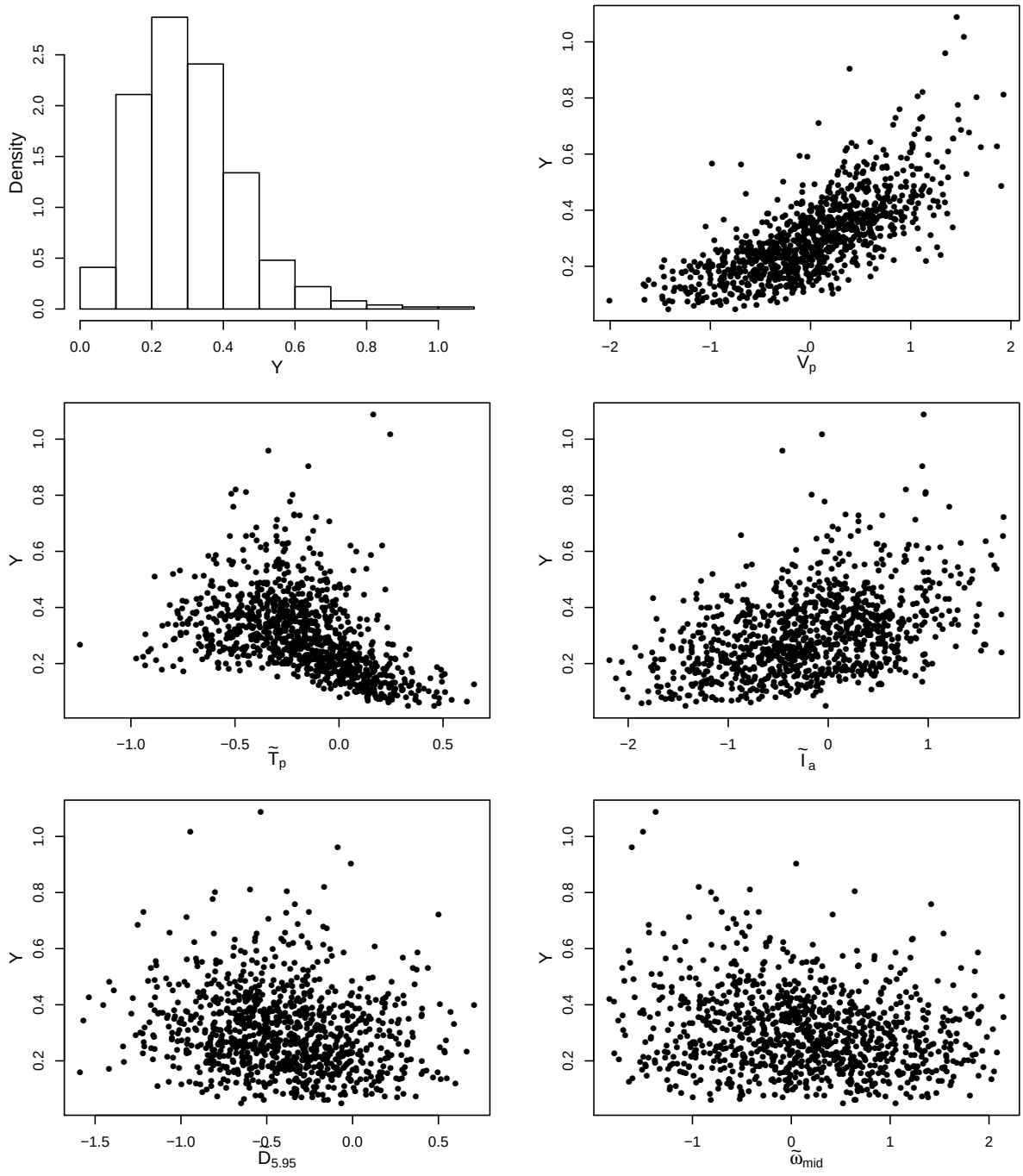


Figure 5.4: The top left figure panel a histogram of the observed displacement, Y. The remaining panels are scatter plots of the displacement against the five parameters of interest $\tilde{V}_p, \tilde{T}_p, \tilde{I}_a, \tilde{D}_{5,95}, \tilde{\omega}_{mid}$ in the normal space.

above the mean (bottom right). When not listed in the titles, the remaining variables are set at their mean values. The middle right panel should be the same across all figures because it is the posterior predictive distribution with all parameters at their mean values.

The behavior for both Fig. 5.5 ($\tilde{V}_p$) and Fig. 5.6 ($\tilde{T}_p$) look reasonable. In the plots for $\tilde{V}_p$, the conditional densities are shifting to the right, corresponding to the increase in the mean we saw in Fig. 5.4. For $\tilde{T}_p$, we see the spread decrease at the highest values, corresponding to the decreased variance we saw in the marginal plots. The peak in response values near $\tilde{T}_p = -.25$ probably corresponds to the period of the oscillator being studied.

In Fig. 5.7, we see the same rightward shift in the conditional densities for $\tilde{I}_a$ corresponding to a mean effect, but the magnitude of the shift is smaller than that of $\tilde{V}_p$. In Figs. 5.8 and 5.9, we see little to no change across different panels, which suggests there is little to low effect from these parameters and consequently the mutual informations $I(Y, \tilde{D}_{5,95})$ and $I(Y, \tilde{\omega}_{mid})$ should both be low.

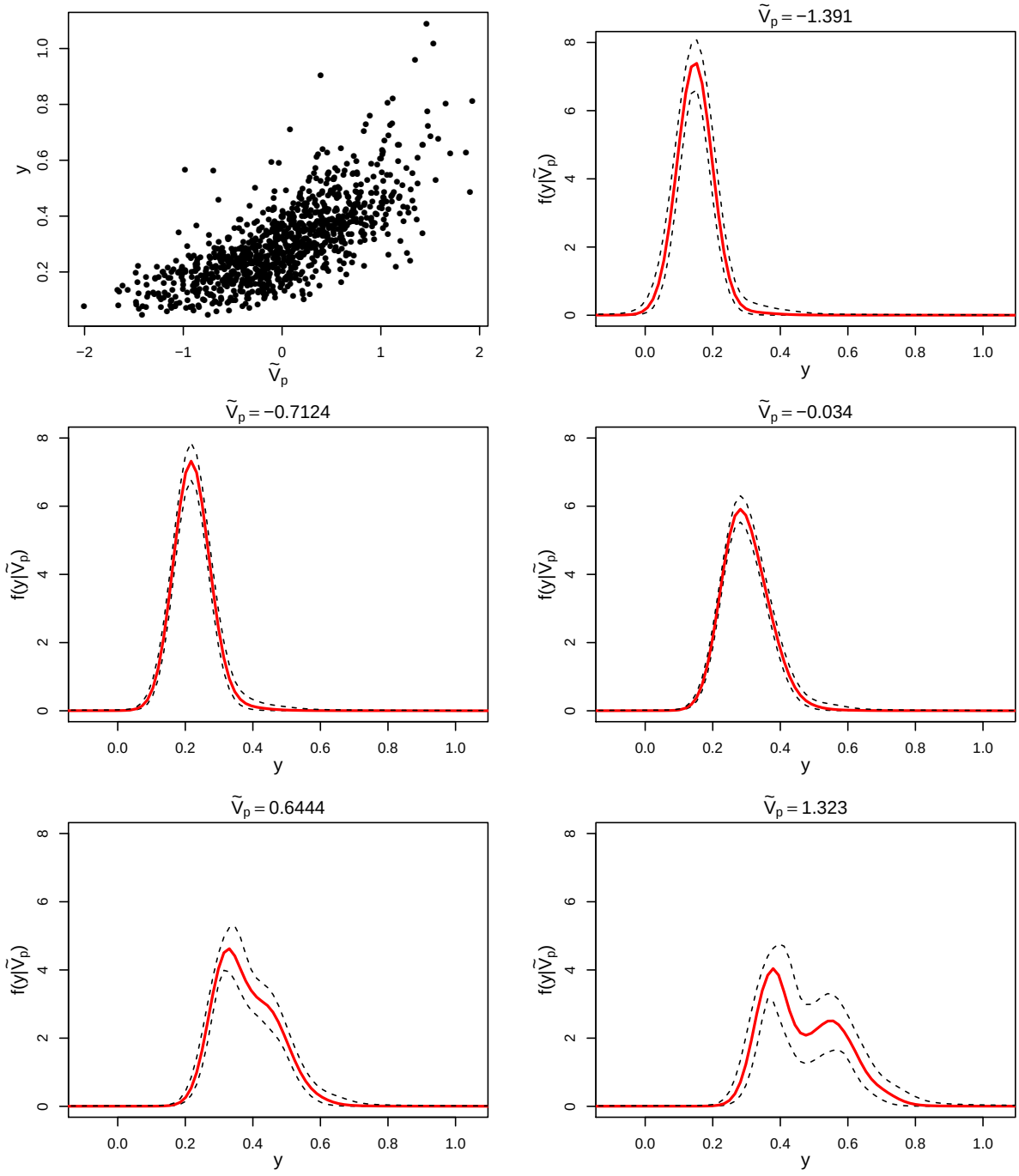


Figure 5.5: Plots of the posterior predictive for the displacement Y given different values of $\tilde{V}_p$, with the unlisted parameters being set at their means. The values of $\tilde{V}_p$ are the mean, the mean $\pm$ 1 SD, and the mean $\pm$ 2 SDs. The red lines indicate the pointwise posterior medians. The black dotted lines indicate pointwise 95% credible intervals. For reference, the marginal scatter plot of the observations is given in the upper left.

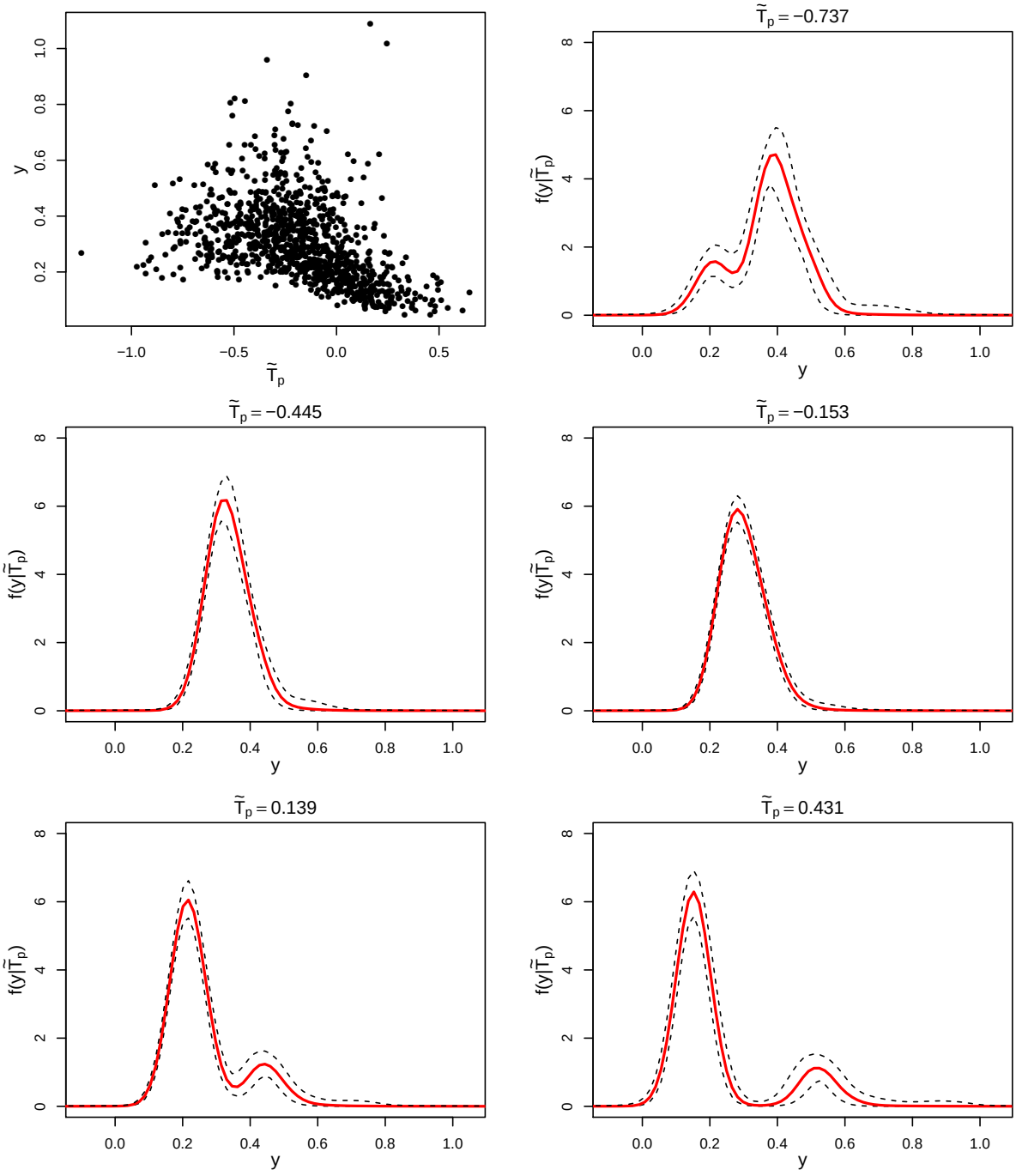


Figure 5.6: Plots of the posterior predictive for the displacement Y given different values of $\tilde{T}_p$, with the unlisted parameters being set at their means. The values of $\tilde{T}_p$ are the mean, the mean $\pm$ 1 SD, and the mean $\pm$ 2 SDs. The red lines indicate the pointwise posterior medians. The black dotted lines indicate pointwise 95% credible intervals. For reference, the marginal scatter plot of the observations is given in the upper left.

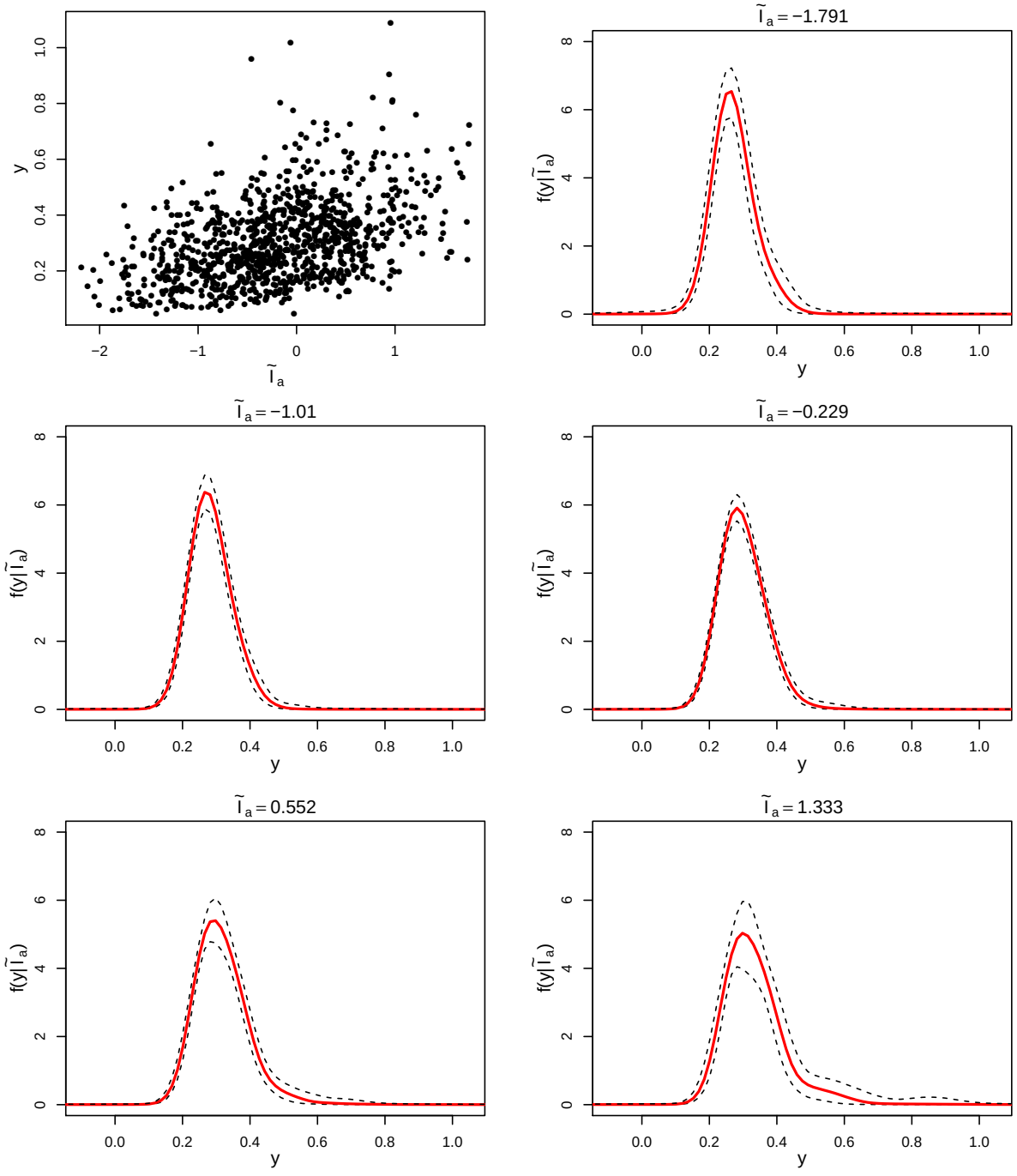


Figure 5.7: Plots of the posterior predictive for the displacement Y given different values of $\tilde{I}_a$, with the unlisted parameters being set at their means. The values of $\tilde{I}_a$ are the mean, the mean $\pm$ 1 SD, and the mean $\pm$ 2 SDs. The red lines indicate the pointwise posterior medians. The black dotted lines indicate pointwise 95% credible intervals. For reference, the marginal scatter plot of the observations is given in the upper left.

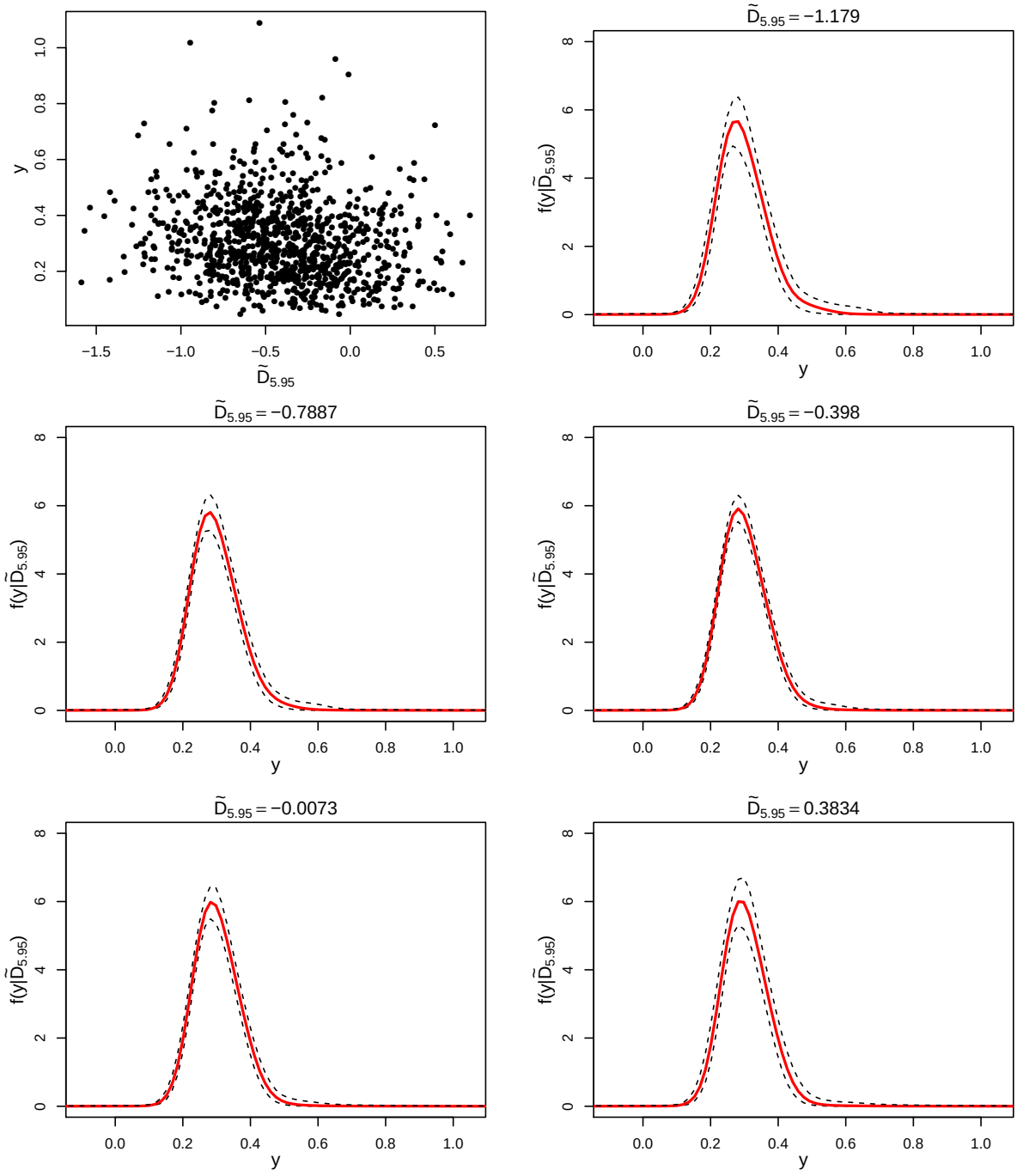


Figure 5.8: Plots of the posterior predictive for the displacement Y given different values of $\tilde{D}_{5,95}$, with the unlisted parameters being set at their means. The values of $\tilde{D}_{5,95}$ are the mean, the mean $\pm$ 1 SD, and the mean $\pm$ 2 SDs. The red lines indicate the pointwise posterior medians. The black dotted lines indicate pointwise 95% credible intervals. For reference, the marginal scatter plot of the observations is given in the upper left.

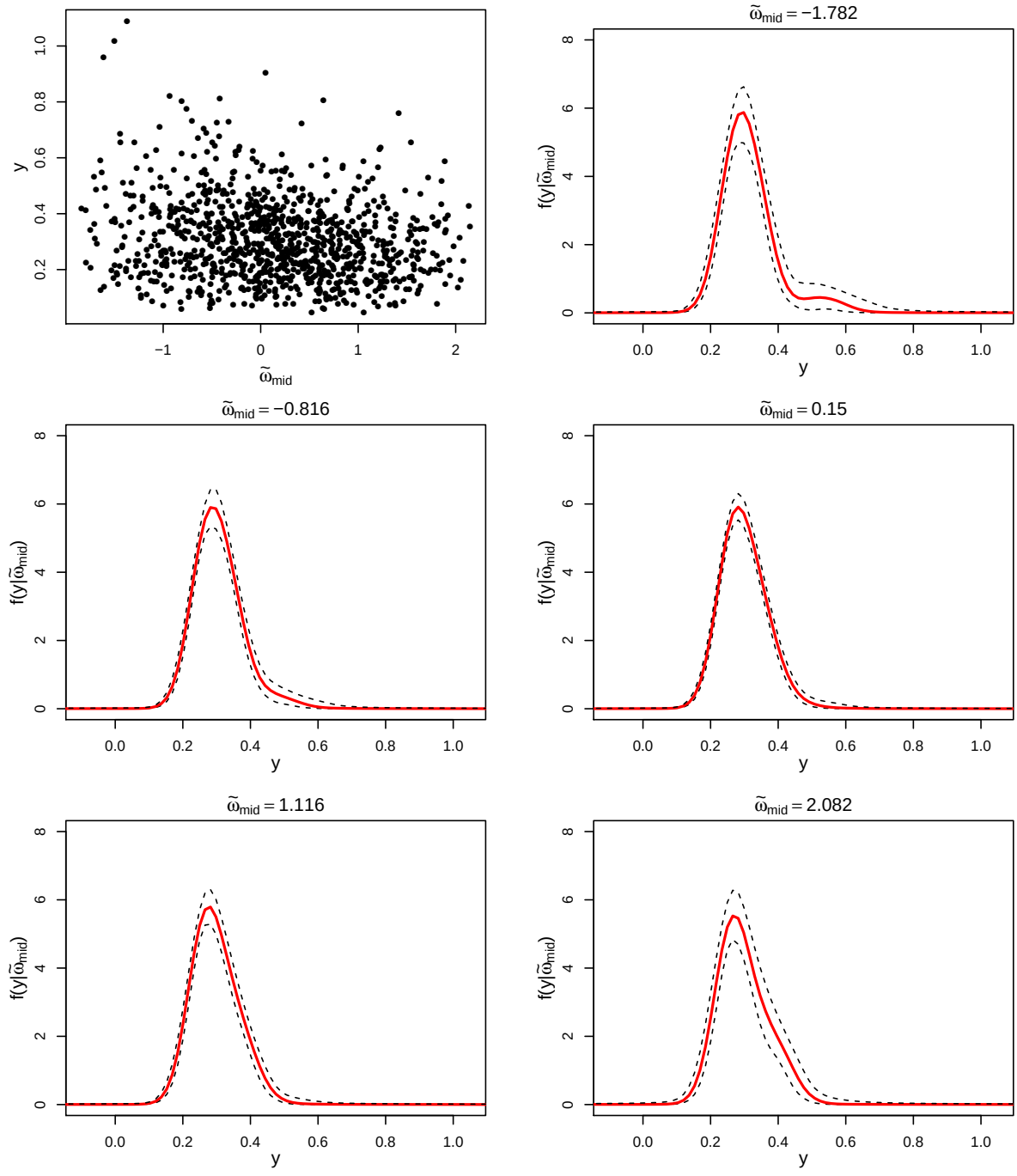


Figure 5.9: Plots of the posterior predictive for the displacement Y given different values of $\tilde{\omega}_{mid}$, with the unlisted parameters being set at their means. The values of $\tilde{\omega}_{mid}$ are the mean, the mean $\pm$ 1 SD, and the mean $\pm$ 2 SDs. The red lines indicate the pointwise posterior medians. The black dotted lines indicate pointwise 95% credible intervals. For reference, the marginal scatter plot of the observations is given in the upper left.

Figures 5.10 and 5.11 contain marginal scatter plots of the data points used to fit the KSBP-based density regression model. The points are colored according to their cluster (Fig. 5.10) and group (Fig. 5.11) assignments for one realization from the posterior. The individual assignments may change slightly between different draws, but overall shapes and trends do not vary much. The panels of Fig. 5.10 also contais lines indicating the value of the slope for that parameter from the given draw. The length of the line is proportional to the product of the log cluster size and the standard deviation of the parameter value within the groups. That is, in the top left panel, the red line is proportional to the product of the log of the number of red coloured points and the standard deviation of $\tilde{V}_p$ for the points that are colored red. Long lines correspond to clusters that either have a lot of points or span a wide range of the parameter. Since the length of the line increases with the number of points in the group, this is an indicator of how strongly its direction will affect the mutual information - it is more likely for a new point to be assigned to a cluster with a long line and the associated slope will show up more often. The length of the line is also an indicator of how stable the slope estimates are - short lines were drawn from distributions based on fewer points so there is more variability in their slope. In all the panels, the red line is the lowest, followed by yellow, orange, and then the two greens. It appears that the clusters are grouped based on the amount of yielding in the response.

For $\tilde{D}_{5,95}$, and $\tilde{\omega}_{mid}$ (the bottom two panels), the three longest lines (orange, yellow, and red) are all nearly horizontal - so the mutual information for these parameters is likely to be small. The two green lines, while clearly not horizontal, are relatively short, so it is unlikely for points to be assigned to these clusters. There is also more variability in the sampled slopes, so their values will change greatly between different realizations from the posterior.

For $\tilde{V}_p$, all the lines have a positive slope, so $I(Y, \tilde{V}_p)$ is probably relatively large. For $\tilde{T}_p$, there are two lines with positive slope (orange and the first green) and two lines with negative slope (red and the other green). The mutual information for $\tilde{T}_p$ will certainly be nonzero, although it's not clear why the clusters are separated in that way. This is potentially an effect of one of the parameters not plotted; the different positive-sloped lines could correspond to different values of one of the other parameters. Since there is not clear separation of clusters in any of the other panels, this is likely due to an interaction effect. An interaction is plausible scientifically; for larger amplitudes $\tilde{V}_p$, the oscillator softens which increases the effective period which changes the effect of $\tilde{T}_p$ on the response.

In Fig. 5.11, we see a possible interaction in the top two panels - the green group consists of points with high values of $\tilde{V}_p$ and low values of $\tilde{T}_p$. For the other parameters, $\tilde{I}_a, \tilde{D}_{5,95}, \tilde{\omega}_{mid}$, there does not seem to be as much separation.

5.5 Sensitivity Results

The results of our sensitivity study are depicted in Fig. 5.12. The black lines indicate 95% credible intervals. The blue dots indicate the sample medians. For reference, the marginal entropy of the displacement is $I(Y) \approx -0.60$. It is clear that $\tilde{V}_p$, $\tilde{T}_p$, and $\tilde{I}_a$ have a much

larger effect on the measured response than $\tilde{D}_{5,95}$, or $\tilde{\omega}_{mid}$. The low effects for $\tilde{D}_{5,95}$ and $\tilde{\omega}_{mid}$ should not be surprising given their lack of mean effects in Fig. 5.4. The magnitudes of the mutual informations for $\tilde{V}_p, \tilde{T}_p$, and $\tilde{I}_a$ coincide with the magnitudes of their effects on the mean.

From an interpretation standpoint, $\tilde{V}_p$ and $\tilde{T}_p$ are the parameters controlling the amplitude and period of the pulse model. Ground motions with large amplitude pulses in their velocity components should definitely cause larger displacements in objects, so this is a reasonable finding. The finding for the period, while not as obvious, is also reasonable given the possibility of resonance for inelastic oscillators due to softening. As the oscillator yields, its natural period elongates and grows closer to the pulse period.

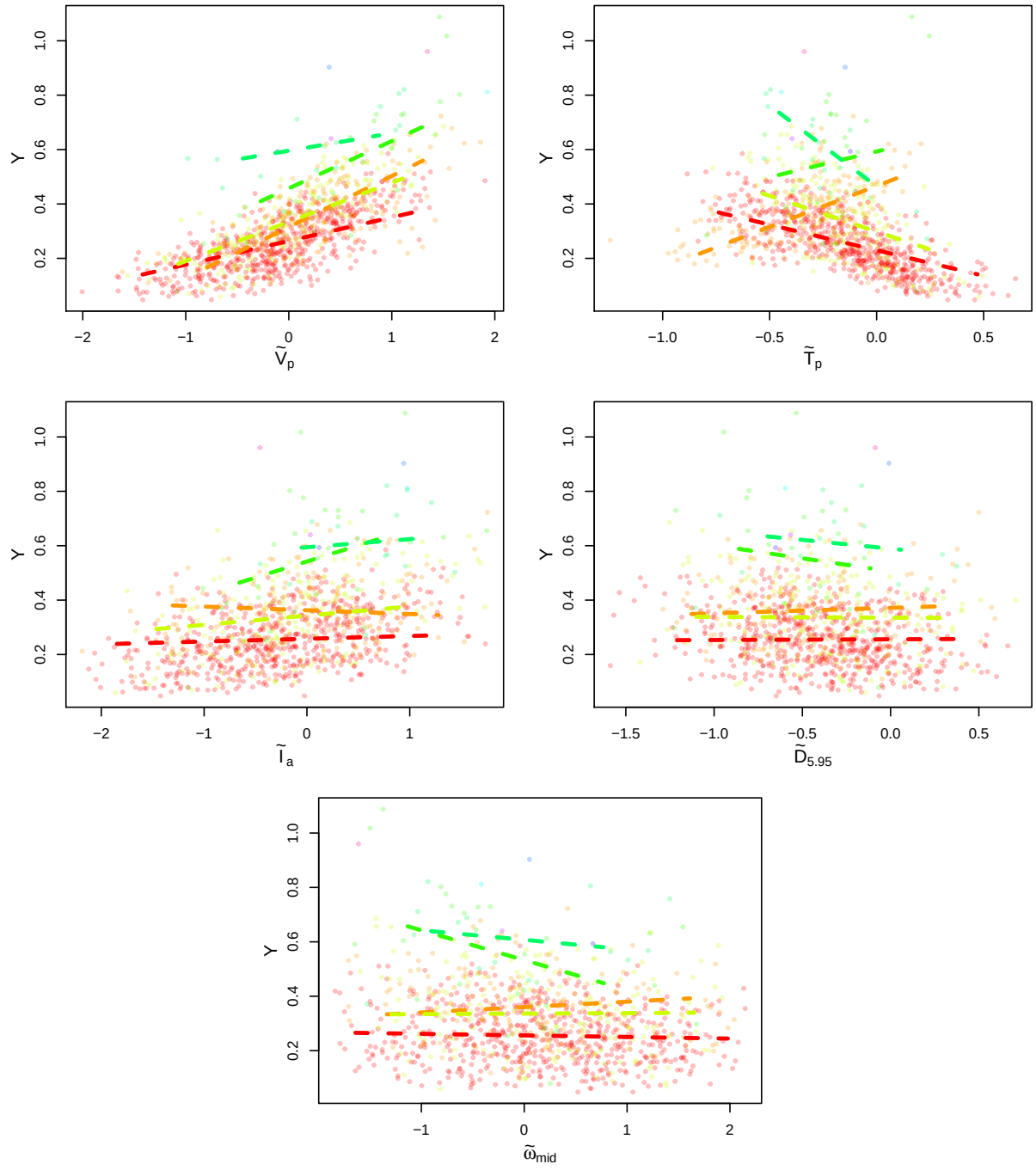


Figure 5.10: Scatter plots of the parameters of interest against the displacement, Y. Points are colored according to their cluster assignments. Coloring is consistent across panels - e.g. red points in each panel correspond to the same cluster. The lines indicate the marginal means for the plotted parameter with lengths proportional to the product of the log cluster size and the SD of the within-group parameter values.

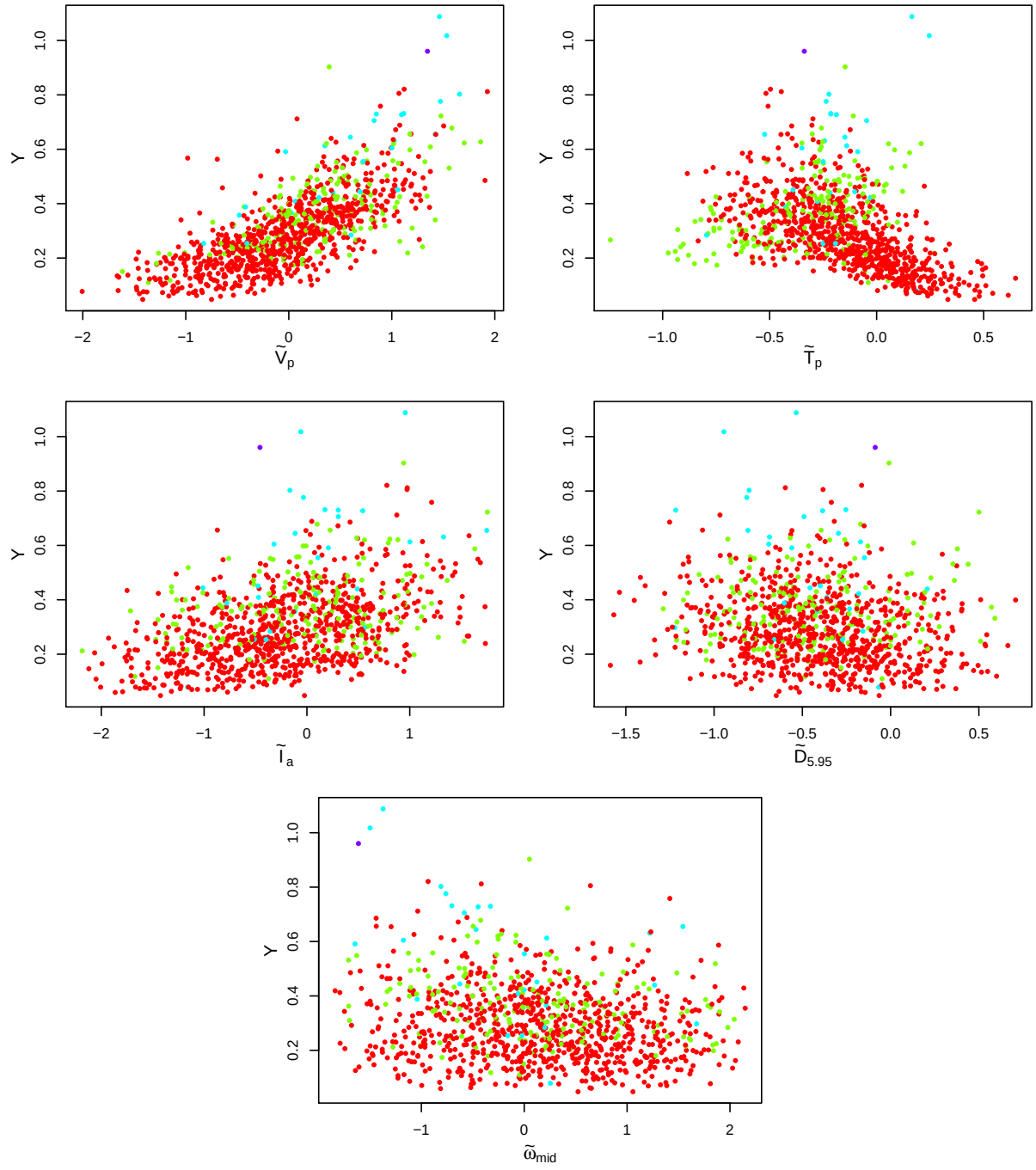


Figure 5.11: Scatter plots of the parameters of interest against the displacement, Y. Points are colored according to their group assignments. Coloring is consistent across panels - e.g. red points in each panels correspond to the same group.

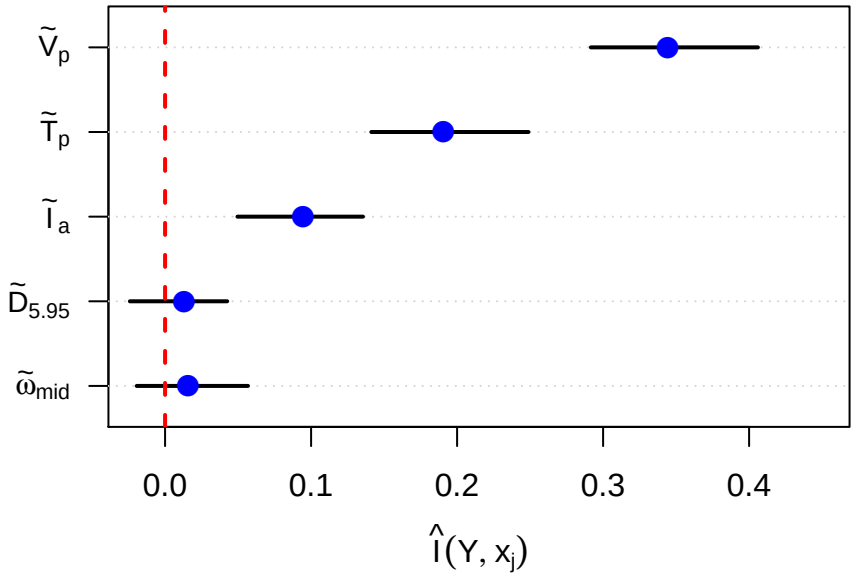


Figure 5.12: Plot of posterior samples for the mutual infromation between the measured response, Y and the parameters of interest. The black lines indicate 95% credible intervals. The blue dots indicate the median of the samples.

Chapter 6: Conclusion

6.1 Summary

The primary contribution of this dissertation is the proposed method for Bayesian probabilistic sensitivity analysis of stochastic simulators through sampling from the posterior distribution for the mutual information presented in Chapter 4. This method is based upon the notion of mathematically characterizing stochastic simulators or computer models as processes mapping input parameters to conditional distributions for the response variable, which was developed in Section 4.1.

Chapter 3 contains ancillary contributions to the field of Bayesian nonparametric density regression, specifically through the use of Kernel Stick Breaking Processes. Section 3.1.2 contains a proposed method for prior specification for the distance parameter (ψ) of a KSBP. Section 3.2 contains an implementation-complete description of a Gibbs sampler for the posterior for Model 3.5. While the sampling method is not entirely novel, one of the steps has been updated to use a slice sampler, which is simpler to implement. To our knowledge, there has not been a complete resource available for anyone interested in the implementation of the KSBP-based Bayesian nonparametric density regression model described in Chapter 3.1.1.

In Chapter 2, the findings of our simulation study comparing estimators of mutual information in Section 2.3.3, while not entirely surprising, are at least novel; to our knowledge there has been no such comparison for such models.

6.2 Future Work

6.2.1 Choice of Design Points

In Chapter 4, the model fitting was performed under the assumption that the input values of the observations were distributed according to the input or uncertainty distribution $\mathcal{G}_X$. Explicitly, the set of observations was defined as $\mathbf{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ where $X_i \sim \mathcal{G}_X$ and $Y_i|X_i \sim F(X_i)$. However, the requirement that $X_i \sim \mathcal{G}_X$ is not actually necessary. In the sensitivity analaysis method, $\mathcal{G}_X$ is the density that the conditional densities are averaged against in Eq. 4.3.

In contrast, the role of the design points, $\mathbf{D}$ is to inform the density regression model about the way that the stochastic simulator F maps inputs to output densities. The relationship between the input parameters and output densities should be independent of the distribution the input parameters in $\mathbf{D}$. In fact, if there is prior knowledge about the behavior of F - e.g. it is more variable in certain regions of the input space - then it is preferable to choose a design that can capitalize on this prior knowledge. In the case of knowing that F is more variable in certain regions, more design points can be allocated to these regions so that there is less posterior uncertainty.

For the most part, modifying the design points in this way does not affect the sensitivity analysis method described in Section 4.3. The main difference will be the interpretation of $\mathcal{G}_X$ not necessarily corresponding to an input distribution, but instead the uncertainty distribution over the input parameters. The sampled values $\hat{I}^{(m)}(Y, x_j)$ will still be from the same posterior distribution. Confusion could arise if caution is not taken to be explicit about what the sampled mutual informations are estimating, which is the primary reason this option was not mentioned when the method was first presented.

However, allowing this flexibility presents a different challenge - how should the design points be chosen? Should they be chosen to optimize the amount of information gained about F? If so, what is the optimal design? Alternatively, the design points could be chosen to minimize the estimation error, although the design achieving this type of optimality is also unknown. These types of questions are still unexplored and would definitely improve the outcomes of sensitivity analyses performed using this method.

6.2.2 Different Types of Response Variables

In this dissertation, we only considered stochastic simulators with a continuous, univariate, response variable on the real line. As we saw in Chapter 5, it is sometimes possible to transform or condense the output of a stochastic simulator to a response of this type. For example, for a simulator that outputs positive numbers, the log transforms the output to the real line. However, there are definitely stochastic simulators where such a transformation is not appropriate. For example, if the simulator output is an event probability or is a multinomial (e.g. different classes), then there is no transformation for this type of simulator.

Luckily, the generalization of both the KSBP-based Model 3.5 to these response types should be reasonably straightforward. Much like the generalization of ordinary regression to generalized linear models, the only change would be in the likelihood of $Y_i|X_i, \beta_i$ (the first equation of Model 3.5) and not in the KSBP prior for $\beta_i|X_i$. This change is minor and should not require major modification of the posterior sampling method in Section 3.2.

Once draws from the modified model posterior are taken, then the sensitivity analysis method described in Section 4.3 should be applicable directly, without any modification. This means, for the most part, all the existing tools for conducting sensitivity analysis for continuous responses on the real line should be applicable, with only minor changes needing to be developed.

6.2.3 Different Bayesian Density Regression Models

The method for generating a sample from the posterior for the mutual information between the response and an input parameter described in Chapter 4 is not limited to the KSBP-based nonparametric density regression model from Chapter 3. One alternative nonparametric density regression model that could potentially be used is the dirichlet process mixture of general linear models proposed by Hannah et al. (2011). Really, any type of Bayesian density regression model can be used in place of Model 3.5 - even parametric models, if one is willing to make the required assumptions on the form of the conditional output density.

As we saw in Section 4.4, the sampled values of mutual information using the method of Section 4.3 are biased downward, which is, at least partly, a result of using the KSBP-based nonparametric Bayesian density regression model. It is possible that replacing with Model 3.5 with a different model may increase performance (i.e. accuracy) or decrease the bias. We chose the nonparametric KSBP-based model due to the lack of required assumptions along with the inherent spatial dependence in the input parameters. These properties may not be as important to other researchers, so there definitely exist alternative density regression models that can be used instead.

6.2.4 Higher Order Indices and Interactions

As mentioned in Section 4.1, $I(Y, x_j)$, the mutual information between Y and x_j, corresponds to a first order sensitivity measure. The definition of mutual information generalizes directly to sets of variables. For example for two inputs x_i and x_j, the mutual information between Y and $\{x_i, x_j\}$ is

$$I\left(Y, \{x_i, x_j\}\right) = -\int_{\mathbf{X}_{ij}} \int_{\mathbf{Y}} log\left\{\frac{f(y, x_i, x_j)}{f_Y(y) f_{ij}(x_i, x_j)}\right\} f(y, x_i, x_j) dy dx_i dx_j.$$

However, this quantity does not represent a second order measure of sensitivity - it is partially confounded with input interactions and first order sensitivity. Lüdtke et al. (2008) advocates the use of *conditional mutual information* for higher order sensitivity measures. The conditional mutual information between x_i and x_j given Y is

$$I(x_i, x_j|Y) = \int_{\mathbf{Y}} \left(\int_{\mathbf{X}_{ij}} log\left\{\frac{f_{ij}(x_i, x_j|y)}{f_i(x_i|y) f_j(x_j|y)}\right\} f_{ij}(x_i, x_j|y) dx_i dx_j \right) f(y) dy$$

With this definition for conditional mutual information, a second order interaction is derived from the following identity

$$I(Y, \{x_i, x_j\}) - I(Y, x_i) - I(Y, x_j) = I(x_i, x_j|Y) - I(x_i, x_j)$$

where $I(x_i, x_j)$ is analogous to the input correlation between x_i and x_j. In the case of input independence, the second order sensitivity measure is the remainder after differencing the first order sensitivities ($I(Y, x_i)$ and $I(Y, x_j)$) from the joint sensitivity ($(I(Y, \{x_i, x_j\}))$).

A similar expression exists for third and higher order interactions as well. While these derivations are sound, the resulting expressions are unsatisfactory from an interpretation point of view. So there are many research possibilities on that front, even if they are just framing his expressions more intuitively from a statistical standpoint.

The notion of a total sensitivity measure - the sum of all effects involving an input - is also pervasive in the mathematical modelling community. The development of a mutual information based measure of total sensitivity would thus be beneficial to the study of stochastic simulators.

While the emphasis on potential research directions in this secton has been on ways to expand the types of analyses done with mutual information, the possibility of a completely different sensitivity measure should not be ruled out. After all, one of the benefits of conducting Bayesian inference of a density regression model on the stochastic simulator is the ability to calculate any desired quantity given realizations from the model posterior.

Chapter 7: Appendix

7.1 Derivation of Expectations in 3.1.2

This section contains full derivations for the following expectations:

$$\begin{aligned}\kappa(\mathbf{x}) &= E\left[K(\mathbf{x},\Gamma_h)\right]\\ \kappa_2(\mathbf{x}) &= E\left[K(\mathbf{x},\Gamma_h)^2\right]\\ \kappa(\mathbf{x},\mathbf{x}') &= E\left[K(\mathbf{x},\Gamma_h)K(\mathbf{x}',\Gamma_h)\right].\end{aligned}$$

with $\mathbf{x},\mathbf{x}'$ being independent draws from a $N(\mathbf{0},I_p)$ distribution and Γ_h also having a $N(\mathbf{0},I_p)$ distribution.

First, $\kappa(\mathbf{x})$:

$$\begin{aligned}E\left[K(\mathbf{x},\Gamma_h)\right] &= \int_{\mathcal{D}_\Gamma} exp\left\{-\psi||\mathbf{x}-\gamma||^2\right\}(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}\gamma^T\gamma\right\}d\gamma\\ &= \int_{\mathcal{D}_\Gamma} exp\left\{-\psi(\mathbf{x}-\gamma)^T(\mathbf{x}-\gamma)\right\}(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}\gamma^T\gamma\right\}d\gamma\\ &= \int_{\mathcal{D}_\Gamma} exp\left\{-\frac{1}{2}(\mathbf{x}-\gamma)^T\left((1/2\psi)I_p\right)^{-1}(\mathbf{x}-\gamma)\right\}(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}\gamma^T\gamma\right\}d\gamma\\ &= (2\pi)^{\frac{p}{2}}(1/2\psi)^{\frac{p}{2}}\int_{\mathcal{D}_\Gamma}(2\pi)^{-\frac{p}{2}}|(1/2\psi)I_p|^{-1}exp\left\{-\frac{1}{2}(\mathbf{x}-\gamma)^T((1/2\psi)I_p)^{-1}(\mathbf{x}-\gamma)\right\}\\ &\qquad\qquad \times(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}\gamma^T\gamma\right\}d\gamma\\ &= (2\pi)^{\frac{p}{2}}(1/2\psi)^{\frac{p}{2}}(2\pi)^{-\frac{p}{2}}(1/2\psi+1)^{-\frac{p}{2}}exp\left\{-\frac{1}{2\left[1/2\psi+1\right]}\mathbf{x}^T\mathbf{x}\right\}\\ &= \left(\frac{1}{1+2\psi}\right)^{\frac{p}{2}}exp\left\{-\frac{\psi}{1+2\psi}\mathbf{x}^T\mathbf{x}\right\}\end{aligned}$$

The second to last equality is due to, under the following hierarchical model,

$$\begin{aligned}X|\Gamma &\sim N(\Gamma,\sigma^2I_p)\\ \Gamma &\sim N(\mathbf{0},I_p),\end{aligned}$$

the marginal distribution for X is a $N(\mathbf{0}, (\sigma^2+1)I_p)$. The integral for $\kappa_2(\mathbf{x})$ is nearly identical:

$$\begin{aligned}
E\left[K(\mathbf{x},\Gamma_h)^2\right] &= \int_{\mathcal{D}_\Gamma} \left(exp\left\{-\psi||\mathbf{x}-\gamma||^2\right\}\right)^2 (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}\gamma^T\gamma\right\} d\gamma \\
&= \int_{\mathcal{D}_\Gamma} exp\left\{-2\psi||\mathbf{x}-\gamma||^2\right\} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}\gamma^T\gamma\right\} d\gamma \\
&= (2\pi)^{\frac{p}{2}}(1/4\psi)^{\frac{p}{2}} \int_{\mathcal{D}_\Gamma} (2\pi)^{-\frac{p}{2}} |(1/4\psi)I_p|^{-1} exp\left\{-\frac{1}{2}(\mathbf{x}-\gamma)^T((1/4\psi)I_p)^{-1}(\mathbf{x}-\gamma)\right\} \\
&\qquad \times (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}\gamma^T\gamma\right\} d\gamma \\
&= (2\pi)^{\frac{p}{2}}(1/4\psi)^{\frac{p}{2}}(2\pi)^{-\frac{p}{2}}(1/4\psi+1)^{-\frac{p}{2}} exp\left\{-\frac{1}{2\left[1/4\psi+1\right]}\mathbf{x}^T\mathbf{x}\right\} \\
&= \left(\frac{1}{1+4\psi}\right)^{\frac{p}{2}} exp\left\{-\frac{2\psi}{1+4\psi}\mathbf{x}^T\mathbf{x}\right\},
\end{aligned}$$

with the integral disappearing in the second to last equality for the same reasons as before. Lastly, $\kappa(\mathbf{x},\mathbf{x}')$:

$$\begin{aligned}
\kappa(\mathbf{x},\mathbf{x}') &= E\left[K_h(\mathbf{x})K_h(\mathbf{x}')\right] = E\left[K(\mathbf{x},\Gamma_h)K(\mathbf{x}',\Gamma_h)\right] \\
&= \int_{\mathcal{D}_\Gamma} exp\left\{-\psi||\mathbf{x}-\gamma||^2\right\} exp\left\{-\psi||\mathbf{x}'-\gamma||^2\right\} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}\gamma^T\gamma\right\} d\gamma \\
&= (2\pi)^{-\frac{p}{2}} \int_{\mathcal{D}_\Gamma} exp\left\{-\frac{1}{2(1/2\psi)}\left(||\mathbf{x}-\gamma||^2 + ||\mathbf{x}'-\gamma||^2 + \frac{\gamma^T\gamma}{2\psi}\right)\right\} d\gamma \\
&= (2\pi)^{-\frac{p}{2}} \int_{\mathcal{D}_\Gamma} exp\left\{-\frac{1}{2(1/2\psi)}\left(a\gamma^T\gamma - \gamma^T(\mathbf{x}+\mathbf{x}') - (\mathbf{x}+\mathbf{x}')^T\gamma + \mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}'\right)\right\} d\gamma
\end{aligned}$$

where $a = \frac{4\psi+1}{2\psi}$. Completing the square by adding and subtracting $\frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a}$ gives:

$$\begin{aligned}
&= (2\pi)^{-\frac{p}{2}} \int_{\mathcal{D}_\Gamma} exp\left\{-\frac{1}{2(1/2\psi)}\left(a\left(\gamma - \frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a}\right)^T\left(\gamma - \frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a}\right)\right.\right. \\
&\qquad \left.\left. - \frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a} + \mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}'\right)\right\} d\gamma \\
&= exp\left\{-\psi\left(\frac{(a-1)\left[\mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}'\right] - \mathbf{x}'^T\mathbf{x} - \mathbf{x}^T\mathbf{x}'}{a}\right)\right\} \times \\
&\qquad \int_{\mathcal{D}_\Gamma} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2(1/2a\psi)}\left(\gamma - \frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a}\right)^T\left(\gamma - \frac{(\mathbf{x}+\mathbf{x}')^T(\mathbf{x}+\mathbf{x}')}{a}\right)\right\} d\gamma \\
&= \left(\frac{1}{2a\psi}\right)^{\frac{p}{2}} exp\left\{-\psi\left(\frac{(a-1)\left[\mathbf{x}^T\mathbf{x} + \mathbf{x}'^T\mathbf{x}'\right] - \mathbf{x}'^T\mathbf{x} - \mathbf{x}^T\mathbf{x}'}{a}\right)\right\}
\end{aligned}$$

The last equality is due to the integrand having the form of a mean zero multivariate Normal with covariance matrix $\frac{1}{2a\psi}I_p$. Plugging $a = \frac{4\psi+1}{2\psi}$ into this expression gives

$$\begin{aligned}
\kappa(\mathbf{x}, \mathbf{x}') &= \left(\frac{1}{2\psi(4\psi+1)/(2\psi)}\right)^{\frac{p}{2}} exp\left\{-\psi\left(\frac{\left(\frac{4\psi+1}{2\psi}-1\right)\left[\mathbf{x}^T\mathbf{x}+\mathbf{x}'^T\mathbf{x}'\right]-\mathbf{x}'^T\mathbf{x}-\mathbf{x}^T\mathbf{x}'}{\frac{4\psi+1}{2\psi}}\right)\right\} \\
&= (4\psi+1)^{-\frac{p}{2}} exp\left\{-\psi\left(\frac{(2\psi+1)\left[\mathbf{x}^T\mathbf{x}+\mathbf{x}'^T\mathbf{x}'\right]-2\psi\left(\mathbf{x}'^T\mathbf{x}-\mathbf{x}^T\mathbf{x}'\right)}{4\psi+1}\right)\right\} \\
&= (4\psi+1)^{-\frac{p}{2}} exp\left\{-\frac{2\psi^2+\psi}{4\psi+1}\left[\mathbf{x}^T\mathbf{x}+\mathbf{x}'^T\mathbf{x}'\right]+\frac{2\psi^2}{4\psi+1}\left[\mathbf{x}^T\mathbf{x}'+\mathbf{x}'^T\mathbf{x}\right]\right\}
\end{aligned}$$

In summary

$$\begin{aligned}
\kappa(\mathbf{x}) &= \left(\frac{1}{1+2\psi}\right)^{\frac{p}{2}} exp\left\{-\frac{\psi}{1+2\psi}\mathbf{x}^T\mathbf{x}\right\} \\
\kappa_2(\mathbf{x}) &= \left(\frac{1}{1+4\psi}\right)^{\frac{p}{2}} exp\left\{-\frac{2\psi}{1+4\psi}\mathbf{x}^T\mathbf{x}\right\} \\
\kappa(\mathbf{x}, \mathbf{x}') &= \left(\frac{1}{1+4\psi}\right)^{\frac{p}{2}} exp\left\{-\frac{2\psi^2+\psi}{1+4\psi}\left[\mathbf{x}^T\mathbf{x}+\mathbf{x}'^T\mathbf{x}'\right]+\frac{2\psi^2}{1+4\psi}\left[\mathbf{x}^T\mathbf{x}'+\mathbf{x}'^T\mathbf{x}\right]\right\} \\
\kappa(\mathbf{x})/\kappa_2(\mathbf{x}) &= \left(\frac{1+4\psi}{1+2\psi}\right)^{\frac{p}{2}} exp\left\{\frac{\psi}{(1+2\psi)(1+4\psi)}\mathbf{x}^T\mathbf{x}\right\}
\end{aligned}$$

Bibliography

Norman Abrahamson, Gail Atkinson, David Boore, Yousef Bozorgnia, Kenneth Campbell, Brian Chiou, IM Idriss, Walter Silva, and Robert Youngs. Comparisons of the nga ground-motion relations. *Earthquake Spectra*, 24(1):45–66, 2008.

I Ahmad and Pi-Erh Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *Information Theory, IEEE Transactions on*, 22(3):372–375, 1976.

Jack W Baker. Quantitative classification of near-fault ground motions using wavelet analysis. *Bulletin of the Seismological Society of America*, 97(5):1486–1501, 2007.

J. Beirlant, E.J.D.L. Györfi, and EC van der Meulen. Nonparametric entropy estimation: an overview. 2001.

Peter J Bickel and Leo Breiman. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, pages 185–214, 1983.

Tamara Broderick, Lester Mackey, John Paisley, and Michael I Jordan. Combinatorial clustering and the beta negative binomial process. *arXiv preprint arXiv:1111.1802*, 2011.

Anil K Chopra. *Dynamics of structures*. Prentice Hall/Pearson Education, 2011.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-interscience, 2012.

G.C. Critchfield and K.E. Willard. Probabilistic analysis of decision trees using monte carlo simulation. *Medical Decision Making*, 6(2):85–92, 1986.

A Dabaghi, M Der Kiureghian, S Rezaeian, and N. Luco. Seismic hazard analysis using simulated ground motions. In *Proceedings of the 11th International Conference on Structural Safety and Reliability (ICOSSAR)*, 2013.

M Dabaghi, S Rezaeian, and A Der Kiureghian. Stochastic simulation of near-fault ground motions for specified earthquake and site chracteristics. In *Proceedings of the International Conference on Appliations of Statistics and Probability in Civil Engineering (ICASP)*, 2011.

Yu G Dmitriev and FP Tarasenko. On the estimation of functionals of the probability density and its derivatives. *Theory of Probability & Its Applications*, 18(3):628–633, 1974.

D.B. Dunson and J.H. Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

László Györfi and Edward C Van der Meulen. Density-free convergence properties of various estimators of entropy. *Computational Statistics & Data Analysis*, 5(4):425–436, 1987.

Peter Hall and Sally C Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1):69–88, 1993.

Lauren Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

Niklas Lüdtke, Stefano Panzeri, Martin Brown, David S Broomhead, Joshua Knowles, Marcelo A Montemurro, and Douglas B Kell. Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks. *Journal of The Royal Society Interface*, 5(19):223–235, 2008.

George P Mavroeidis and Apostolos S Papageorgiou. A mathematical representation of near-fault ground motions. *Bulletin of the Seismological Society of America*, 93(3):1099–1131, 2003.

J.E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models. *Journal of the Royal Statistical Society: Series B*, 66(3):751–769, 2004.

Omiros Papaspiliopoulos and Gareth O Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

Sanaz Rezaeian and Armen Der Kiureghian. A stochastic ground motion model with separable temporal and spectral nonstationarities. *Earthquake Engineering & Structural Dynamics*, 37(13):1565–1584, 2008.

Sanaz Rezaeian and Armen Der Kiureghian. Simulation of synthetic ground motions for specified earthquake and site characteristics. *Earthquake Engineering & Structural Dynamics*, 39(10):1155–1180, 2010.

Fabrice Rossi, Amaury Lendasse, Damien François, Vincent Wertz, and Michel Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and intelligent laboratory systems*, 80(2):215–226, 2006.

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, 2004.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. Wiley. com, 2008.

Shrey K Shahi and Jack W Baker. An empirically calibrated framework for including the effects of near-fault directivity in probabilistic seismic hazard analysis. *Bulletin of the Seismological Society of America*, 101(2):742–755, 2011.

Claude Elwood Shannon and Warren Weaver. A mathematical theory of communication, 1948.

I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

PAUL Somerville. Seismic hazard evaluation. *Bulletin of the New Zealand Society for Earthquake Engineering*, 33(3):371–386, 2000.

Alexandros A Taflanidis and Gaofeng Jia. A simulation-based framework for risk assessment and probabilistic sensitivity analysis of base-isolated structures. *Earthquake Engineering & Structural Dynamics*, 40(14):1629–1651, 2011.

AB Tsybakov and EC Van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, pages 75–83, 1996.

Stephen G Walker. Sampling the dirichlet mixture model with slices. *Communications in StatisticsâĂŤSimulation and Computation®*, 36(1):45–54, 2007.