# My amazing title

*Your R. Name*
April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisors:
*Advisor F. Name*
*Your Other Advisor*

# Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

# Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1  Introduction

The field of preventive epidemiology involves the identification of potentially modifiable risk factors that contribute to the burden of disease within human populations. Environmental epidemiology, in particular, considers the effect of environmental exposures — chemical or otherwise — which have been increasingly recognized as crucial determinants of human health (Vineis, 2018). Understanding the health effects of exposure to chemical pollutants is especially timely. It has been estimated that human activity releases chemicals at a rate of 220 billion tons per annum (Cribb, 2016). As a result, exposure to low levels of pollutants has become an inevitable peril of daily life. Scholars warn that such conditions of cumulative chronic toxicity pose an acute risk to the wellbeing of humans and our living environment (Naidu et al., 2021). To this, the quantification of such risks through environmental epidemiology studies can prompt critical regulatory action.

Studies concerning chemical pollutants in environmental epidemiology have historically focused on elucidating the effect and mechanisms of single exposures. However, humans are invariably exposed to numerous complex chemical mixtures which together contribute to the progression of adverse health outcomes — risk assessments of single pollutants likely fail to capture the true consequences of these complex exposures (Heys et al., 2016). Assessing mixtures of chemicals can also have more direct implications for public health interventions. The United States Environmental Protection Agency (U.S. EPA) currently passes regulations for individual pollutants.

In practice, though, regulation occurs by controlling the source of pollution, which is responsible for the production of a whole mixture of chemicals with specific joint effects on human health. As a result, the National Academies of Science has advocated for a multipollutant regulatory approach, which is likely to be more protective of human health (Committee on Incorporating 21st Century Science into Risk-Based Evaluations et al., 2017).

Hence, there are clear practical motivations for the development of studies and methodologies that examine the health effects of exposure to co-occurring chemical mixtures, hereafter referred to as exposure mixtures. However, expanding the focus of analysis from one exposure to multiple exposures introduces unique statistical challenges. In addition to the issue of small effect sizes and small sample sizes present in all exposure analyses, multiple exposure analyses must also contend with high-dimensionality, collinearity, non-linear effects, and non-additive interactions (Yu et al., 2022). The classic multiple linear regression framework fails to capture the true effects in this setting. In the past few years, a wide variety of statistical methods have been developed to overcome these challenges (Gibson et al., 2019; Yu et al., 2022), which have been accompanied by a host of comparative simulation studies for general mixture scenarios (e.g., Hoskovec et al., 2021; Lazarevic et al., 2020; Pesenti et al., 2023). However, there is not yet conclusive guidance about the ability of these methods to conduct inference on non-additive interactions between exposures.

The goal of this thesis is to explore the theory of emerging Bayesian regression techniques for quantifying complex interactions between environmental exposures. [**clarify goals**]

In an age where anthropogenic actions have radically reshaped the earth, humanistic inquiry can offer critical insights into our place within a rapidly evolving environment. I begin in Chapter 2 by contextualizing this thesis with a brief overview

of cultural and social understandings of the topic of environmental exposures. Chapter 3 provides background on X Bayesian methods for analyzing exposure mixtures. Chapter 4 assesses the performance of these methods for conducting inference on non-additive interactions using a simulation study based on **X**. Chapter 5 explores an application on **X** data [**TBD**]. I conclude with a discussion of the implications of this work for the future study of complex interactions in exposure mixture studies.

# Chapter 2    R Markdown Basics

# Chapter 3    Bayesian regression methods

## 3.1    Motivation

- mechanistic

- public health

- interactions are difficult to pick up

- consider the number of interactions that would have to be encoded explicitly in

  MLR

- different types of interactions

## 3.2    Bayesian kernel machine regression

Defining notation:

- observations from $i = 1, \ldots, n$

- $\mathbf{x}_m$ is a predictor variable in the predictor matrix $\mathbf{X}$ with $m = 1, \ldots, M$, measuring exposure variables in this case

- $\mathbf{x}_i$ is a vector of values for a single observation in $\mathbf{X}$ with $i = 1, \ldots, n$, measuring the health outcome in this case

- $x_{im}$ is an observation of $\mathbf{x}_m$

- $Y_i$ is an observation of $\mathbf{Y}$

- $\rho$ is the tuning parameter inside the exponential expression of the covariance,

equals $2l$

- $\tau$ is the tuning parameter outside the expontential expression of the covariance, equals $\sigma_f^2$

- $k$ is the smoothing function, the Gaussian in this case

- $\mathbf{K}$ is the $n \times n$ kernel matrix, with (i, j)th element $k(\mathbf{x}_i, \mathbf{x}_j)$

- $h(\mathbf{x}_i)$ the flexible function relating X to Y

- $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, the residuals of the response

### 3.2.1   Kernel machine regression

In this section, we introduce kernel machine regression, with attention to its specific implementation in BKMR. Kernel machine regression is a nonparametric regression technique that can be used to capture nonlinear effects and nonadditive interactions. In the typical linear regression setting,

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$$

where $Y_i$ measures a health outcome at a given point, $\mathbf{x}_i = (x_1, \ldots, x_M)^\top$ is a vector of M exposures, $\boldsymbol{\beta}$ is a vector of weights, and $\epsilon_i$ is a random variable from $\boldsymbol{\epsilon} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We can see that this function assumes that there is a linear relationship between the exposure and the response, and that the combined effects of multiple exposures are additive.

Kernel machine regression defines this relationship using a function $h : \mathbb{R}^M \to \mathbb{R}$, where

$$Y_i = h(\mathbf{x}_i) + \epsilon_i$$

Here, $h(\cdot)$ is represented by the function $k(\cdot, \cdot)$, a kernel. The kernel controls the

covariance, or the similarity, between values of $h(\mathbf{x})$ and as such ensures that points near each other on the prediction surface will have similar values — or, in other words, that the prediction surface will be smooth. In the case of kernel machine regression, we define a positive definite kernel where $k : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$. There are many choices of functions to represent $k$. BKMR uses the Gaussian kernel, also known as the radial basis function or, sometimes, the squared exponential kernel. The Gaussian kernel is defined

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\{-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2l}\}$$

where $||\mathbf{x} - \mathbf{x}'||^2 = \sum_{m=1}^{M} (x_m - x'_m)^2$ for $\mathbf{x}$, a set of exposure values, and $\mathbf{x}$, the exposure profile of a second subject, $\sigma_f^2$ is a tuning parameter that controls how much the function is allowed to vary, and $l$ is a tuning parameter that controls the relationship between the correlation between two points and their distance. Greater values of $l$ will enforce more dependence between points and make the resulting function smoother. Note that BKMR uses $\rho = 2l$ and $\tau = \sigma_f^2$, so, henceforth, we will be referring to these parameters using BKMR's notation.

Now that we have defined $h$ and $k$, we can think about how to characterize the relationship between our response and predictors. Kernel machine regression is a nonparametric technique because it does not specify a functional form for this relationship. Hence, we will think about estimating the response at a particular query point. Operationally, kernel machine regression uses a weighted average of all the observations in the dataset to estimate the response, as follows

$$\bar{Y} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

with some set of weights $\{w_i\}_{i=1}^{n}$. Intuitively, we want to weight the observations that

are closer to the query point more heavily. Using the Gaussian kernel as a weight allows us to achieve this. Replacing the weight with the Gaussian kernel, we get

$$\bar{Y} = \frac{\sum_{i=1}^{n} K(\mathbf{x}, \mathbf{x}_i) Y_i}{\sum_{i=1}^{n} K(\mathbf{x}, \mathbf{x}_i)}$$

As we move through the predictor space, we can think of the prediction as a continuous moving average of local points in the dataset. This gives us the following relationship

$$\text{cor}(h_i, h_j) = \exp\{-(\frac{1}{\rho}) \sum_{m=1}^{M} (x_{im} - x_{jm})^2\}$$

which allows us to see that values $h$ near each other will have a higher correlation and thus similar values. This is also why the resulting function is smooth.

### 3.2.2 Connection with mixed models

It is useful to make connections between this definition of kernel machine regression and mixed models. To do this, we can represent $h(\mathbf{x})$ as following a Gaussian process probability distribution, defined

$$h(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$$

with mean function $m$ and covariance function $k$, where $\mathbf{x}$ is a vector of exposure values, and $\mathbf{x}'$ is the exposure profile of another subject. A Gaussian process is a collection of random variables, of which any finite number follow a multivariate normal distribution. Here, we assume that the expected value of the function with input $\mathbf{x}$ is $\mathbf{0}$. We use $k$ for the covariance function, whic represents the dependence between the function values with inputs $\mathbf{x}$ and $\mathbf{x}'$: $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(h(\mathbf{x}) - m(\mathbf{x}))(h(\mathbf{x}') - m(\mathbf{x}'))]$.

Now, we can represent $h$ as a collection of variables from a Gaussian process. $h$

follows a multivariate normal distribution

$$h(\mathbf{x}) \sim N(\mathbf{0}, \mathbf{K})$$

where $h(\mathbf{x}) = h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_n)$, and $\mathbf{K}$ is the kernel matrix. The kernel matrix is an $n \times n$ matrix with $(i, j)$th element $k(\mathbf{x}_i, \mathbf{x}_j)$. Now, returning back to the regression view, we can think of each $Y_i$ as following the distribution
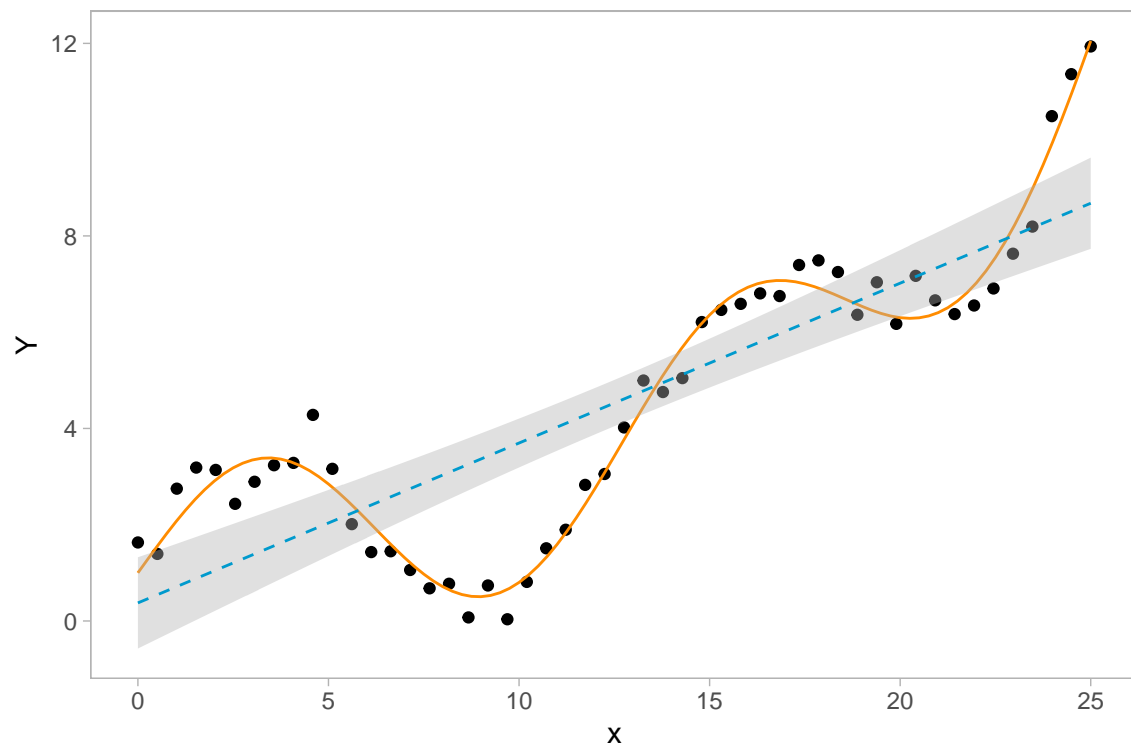
$$Y_i \overset{\text{ind}}{\sim} N(h(\mathbf{x}_i), \sigma^2) \text{ for } i = 1, \ldots, n$$

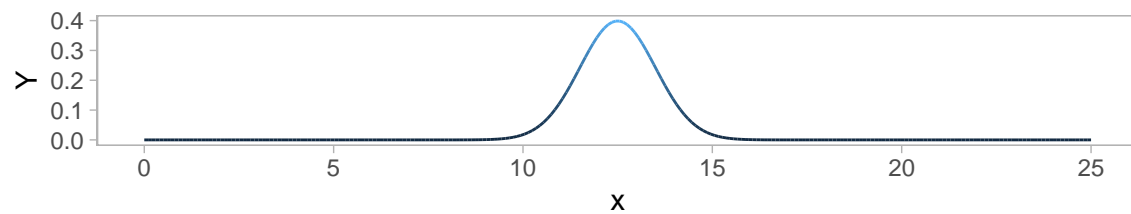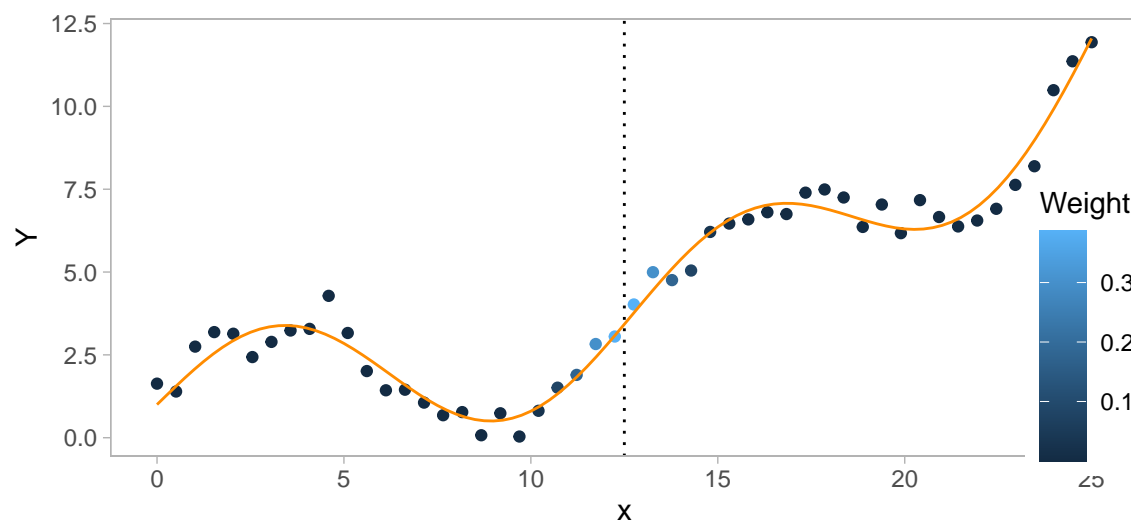where $\sigma^2$ comes from the variance of the residuals.

### 3.2.3 Toy example

In the following section, we introduce the technique with a toy example.

Consider the following case where we want to model the relationship between a single predictor and a response variable. Suppose the true relationship between x and Y is defined $Y = e^{\frac{x}{10}} + 2\sin(\frac{x}{2})$

Define a query point

Note to self, bandwidth represent four times the quartiles of the probability density. In the case of a Gaussian kernel, the bandwidth represents $\frac{8}{3}\sigma$, so using an sd of 1 translates to bandwidth of 8/3
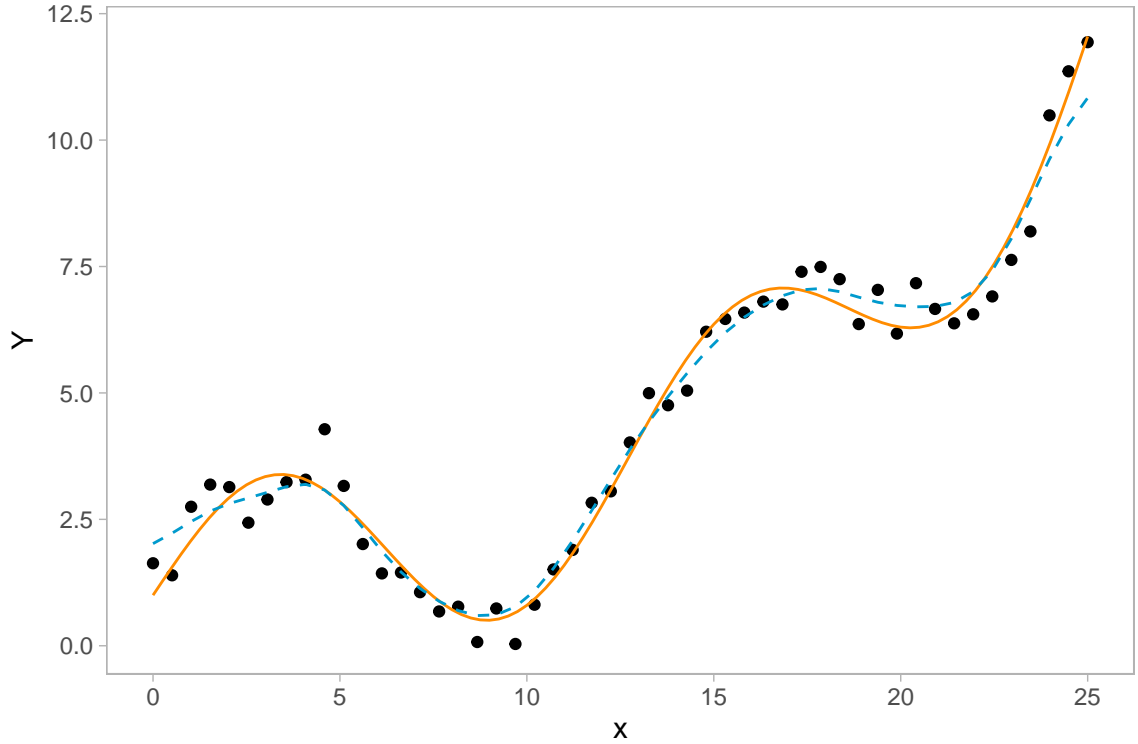


We can also use this example to understand how our choice of the scaling parameter influences model fit.

### 3.2.4 Variable selection

In this section, we discuss the two methods for variable selection in BKMR: hierarchical variable selectiona and component-wise variable selection.

### 3.2.5   Priors in a Bayesian framework

### 3.2.6   The MCMC algorithm

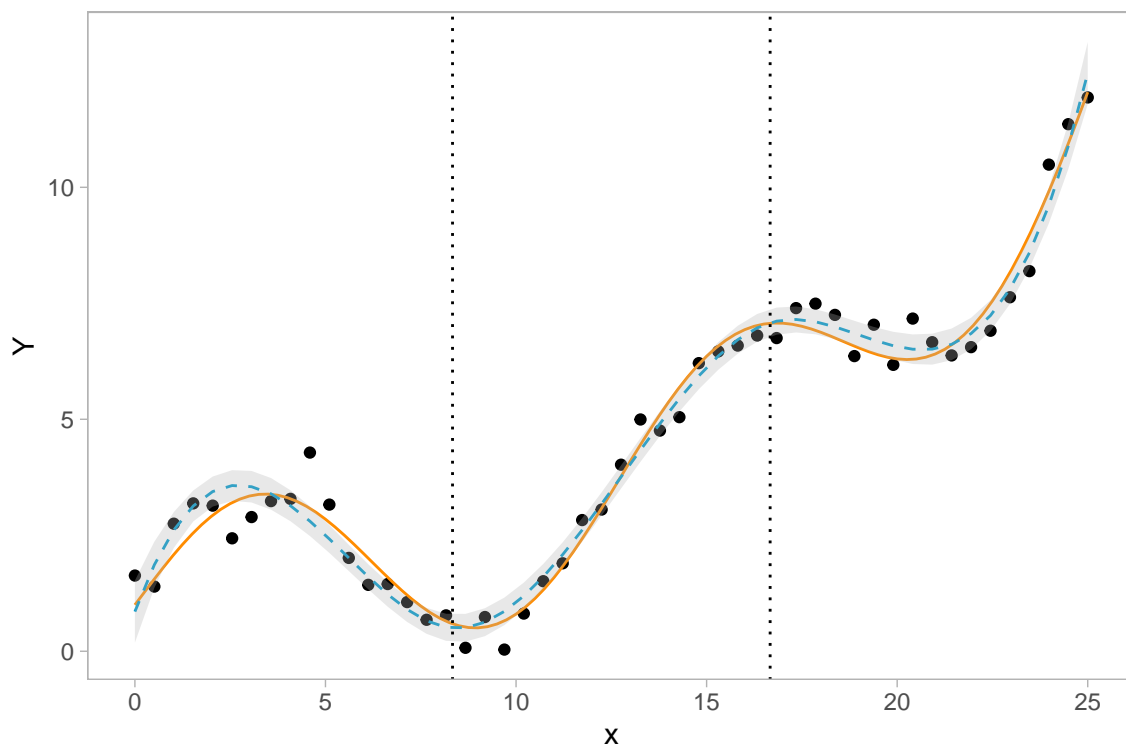### 3.2.7   Visualizing interactions

## 3.3   Bayesian semiparametric regression

differences w/ bkmr

- makes distributional assumptions about the dataset

- kernel regression is computationally intensive w/ large datasets

### 3.3.1   Connections to linear mixed model

### 3.3.2   Toy example



$->$

# Chapter 4   Tables, Graphics, References, and Labels

# Conclusion

If we don't want the conclusion to have a chapter number next to it, we can add the {-} attribute.

**More info**

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

# Appendix A   The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

## A.1   In the main file 4:

## A.2   In Chapter 4:

# Appendix B    The Second Appendix

R code

# Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

# References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with OpenGL.* Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with Quick-Time.* Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel.* Boston, MA: Wesley Addison Longman.