Bayesian Nonparametric Regression Models for Quantifying Complex Interactions in Exposure Mixture Studies

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

Rapid industrial development has created conditions of cumulative chronic toxicity which pose an acute risk to the wellbeing of humans and our living environment. It has been estimated that human activity releases chemicals at a rate of 220 billion tons per annum (Cribb, 2016). Scholars have recently formally declared that, at this global rate of chemical release, humanity has surpassed the safe operating space of the planetary boundary for novel entities (Persson et al., 2022). As a result, exposure to low levels of pollutants has become an inevitable peril of daily life (Naidu et al., 2021; Vineis, 2018). Hence, it is especially timely that we prompt regulatory control of industrial pollution through studies which investigate the health effects of chemical exposures.

For this, we turn to epidemiological studies. The broad 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. However, studies concerning chemical pollutants in environmental epidemiology have historically focused on elucidating the effect and mechanisms of exposures to a single pollutant. In reality, humans are invariably exposed to numerous complex chemical mixtures which together contribute to the progression of adverse health outcomes. Therefore, 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).

There are clear practical motivations for studies 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 a common issue of small effect sizes and small sample sizes present in most exposure analyses, multiple exposure analyses must also contend with high-dimensionality, collinearity, non-linear effects, and non-additive interactions (Yu et al., 2022). In particular, data with numerous pollutants, or predictors, require exponentially greater levels of complexity and time cost in analysis. Collinearity between exposures is common when analyzing pollutants from a single source and can lead to unstable estimates in a generalized linear model if left unaccounted for. Finally, exposures can have both non-linear single effects and non-additive interaction effects, which are difficult to capture unless explicitly specified in the model.

The classic multiple linear regression framework often 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 fill this gap in the literature by exploring the theory and performance of Bayesian regression techniques for quantifying complex interactions between environmental exposures. Specifically, we will compare two recently developed models for estimating the health effects of exposure mixtures: Bayesian Kernel Machine Regression (BKMR, Bobb et al., 2015) and Bayesian Semiparametric Regression (BSR, Antonelli et al., 2020).

In an age where anthropogenic actions have radically reshaped the earth, humanistic inquiry can offer critical insights into how we navigate the hazards of our rapidly changing environment. We begin in Chapter 2 by contextualizing this thesis with a brief overview of cultural and social understandings of toxicity. Chapter 3 explains the motivation for studying interactions and provides background on the theory of Bayesian methods for analyzing exposure mixtures. Chapter 4 assesses the performance of these methods using a simulation study, based on a dataset with information on the relationship between prenatal exposure to heavy metals and gestational weight. Chapter 5 explores an application on X data [TBD]. We conclude with a discussion of the implications of this work for the future study of complex interactions in exposure mixture studies.

Chapter 2 Humanistic perspective

Chapter 3 Bayesian regression methods

3.1 Motivation

We are interested in using Bayesian regression techniques to characterize the nature of complex interactions. In addition to being an interesting statistical challenge, complex interactions are also relevant from both a mechanistic and public health point of view.

- mechanistic
- public health

Drawing from the mechanistic and public health contexts, there are three classes of interactions that are of interest.

- different types of interactions

However, non-additive interactions are difficult to detect.

First, there are different forms that an interaction can take on. For instance, in the linear regression setting, the form of the interaction must be explicitly specified; most commonly, this is done by multiplying two variables together. However,

Moreover, the number of possible interactions to consider quickly becomes intractable in high-dimensional settings. For instance, consider modelling 10 exposures in a linear regression setting. In order to be assessed, each interaction must be explicitly specified as a new term in the model. Including all the possible two-way interactions would involve adding $\binom{10}{2} = 45$ additional terms to the model, and all possible three-way interactions would add $\binom{10}{3} = 120$ additional terms.

- interactions are difficult to pick up
- consider the number of interactions that would have to be encoded explicitly in MLR

It is important also to acknowledge, here, that there is a limit to how many variables can be included in an interaction before it becomes incomprehensible to most humans.

3.2 Bayesian kernel machine regression

Notation for kernel machine regression:

- $\mathbf{x}_m$ is a predictor variable in the predictor matrix $\mathbf{X}$ with $m = 1, \ldots, M$, measuring exposure variables or covariates in this case
- $\mathbf{x}_i$ is a vector of values for a single observation in $\mathbf{X}$ with $i = 1, \ldots, n$
- x_{im} is an observation of $\mathbf{x}_m$
- Y_i is an observation of $\mathbf{Y}$, measuring the health outcome in this case
- $h(\mathbf{x}_i)$ is the flexible function relating X to Y, represented by the kernel
- k is the kernel function, the Gaussian in this case
- $\mathbf{K}$ is the $n \times n$ kernel matrix, with (i, j)th element $k(\mathbf{x}_i, \mathbf{x}_j)$
- ρ is the tuning parameter inside the kernel function which controls smoothness
- τ is the tuning parameter multiplied by the kernel matrix to represent covariance between h values
- $\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are the residuals of the response

BKMR specific notation:

- $r_m = 1/\rho_m$ is an augmented variable in $\mathbf{r}$ in the kernel matrix, for variable selection and controlling smoothness

- δ_m is an indicator variable in $\boldsymbol{\delta}$ which represents inclusion in the model
- $\mathcal{S}_g$ is a group of partitioned predictors with $g = 1, \ldots, G$
- $\{\delta_m | \mathbf{x}_m \in \mathcal{S}_g\}$ is an indicator variable in $\boldsymbol{\delta}_{\mathcal{S}_g}$ which represents inclusion of a parameter in group g in the model
- π is the prior probability of inclusion in the model
- $\lambda = \tau\sigma^{-2}$ is a convenient way to define the prior on τ

3.2.1 Kernel machine regression

In this section, we introduce kernel machine regression, with attention to its specific implementation in BKMR. We follow the presentation of this method provided by Bobb et al. (2015). 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 or covariates (hereafter referred to as predictors), $\boldsymbol{\beta}$ is a vector of weights, and ϵ_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 as

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

where $||\mathbf{x} - \mathbf{x}'||^2 = \sum_{m=1}^{M} (x_m - x'_m)^2$ for a set of predictors values $\mathbf{x}$ and the predictor values of a second subject $\mathbf{x}'$. ρ is a tuning parameter that controls the relationship between the correlation between two points and their distance. Greater values of ρ will enforce more dependence between points and make the resulting function smoother. h is related to k by a multiplicative constant τ, a tuning parameter which controls how much the values of the h function are allowed to vary.

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, [*citation*] demonstrates that kernel machine regression uses a weighted average of all the observations in the dataset to estimate the response, defined as

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

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. The correlation between two values of h is defined as

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

which allows us to see that values of 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. Liu et al. (2007) demonstrated this by representing $h(\mathbf{x})$ as following a Gaussian process probability distribution,

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

with mean function m and covariance function k, where $\mathbf{x}$ is a vector of the predictor values, and $\mathbf{x}'$ contains the predictor values 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 h function with input $\mathbf{x}$ is $\mathbf{0}$. We use k for the covariance function, which represents the dependence between the function values with inputs $\mathbf{x}$ and $\mathbf{x}'$: $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(h(\mathbf{x}) - \mathbf{0})(h(\mathbf{x}') - \mathbf{0})]$.

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

follows a multivariate normal distribution,

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

where $h(\mathbf{x}) = [h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_n)]^\top$ 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

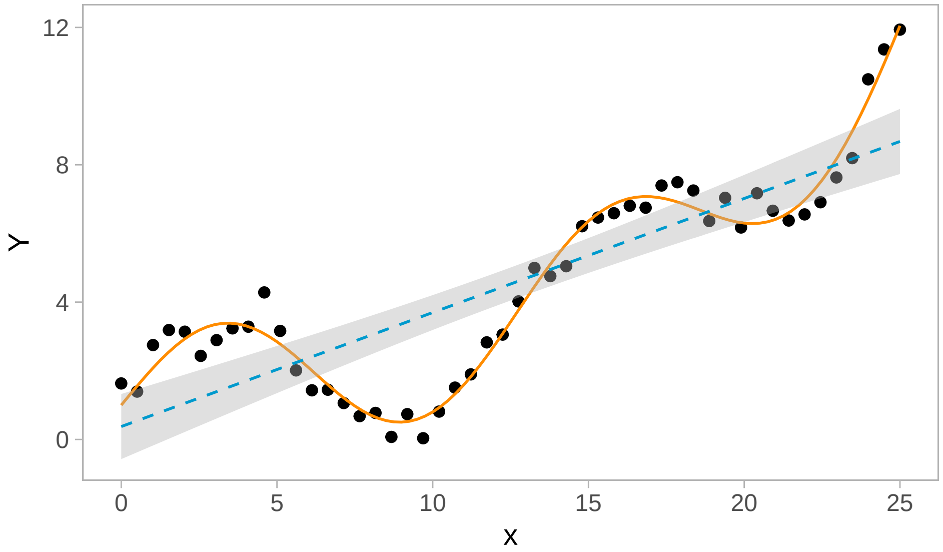
$$Y_i \overset{\text{ind}}{\sim} N(h(\mathbf{x}_i), \sigma^2) \text{ for } i = 1, \ldots, n,$$

where σ^2 comes from the variance of the residuals.

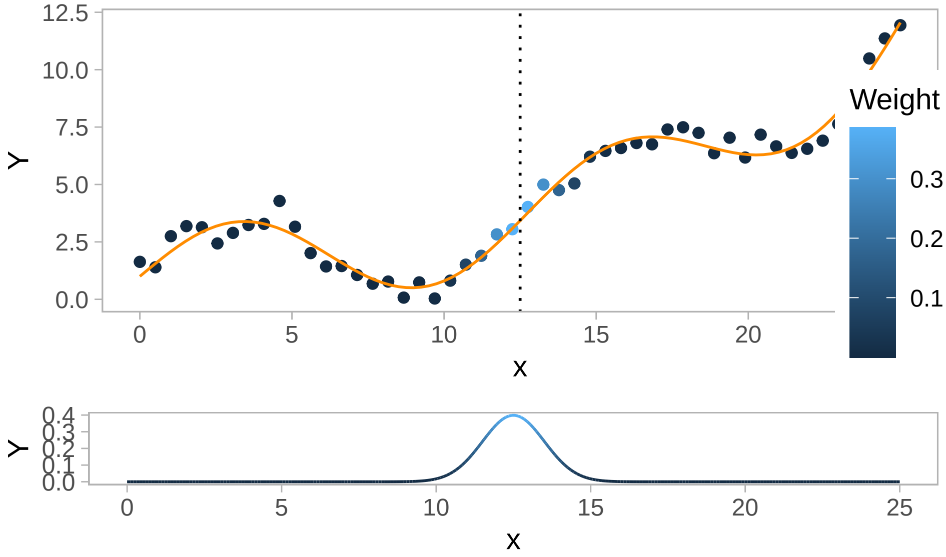
3.2.3 Toy example

In the following section, we introduce kernel machine regression with a toy example.

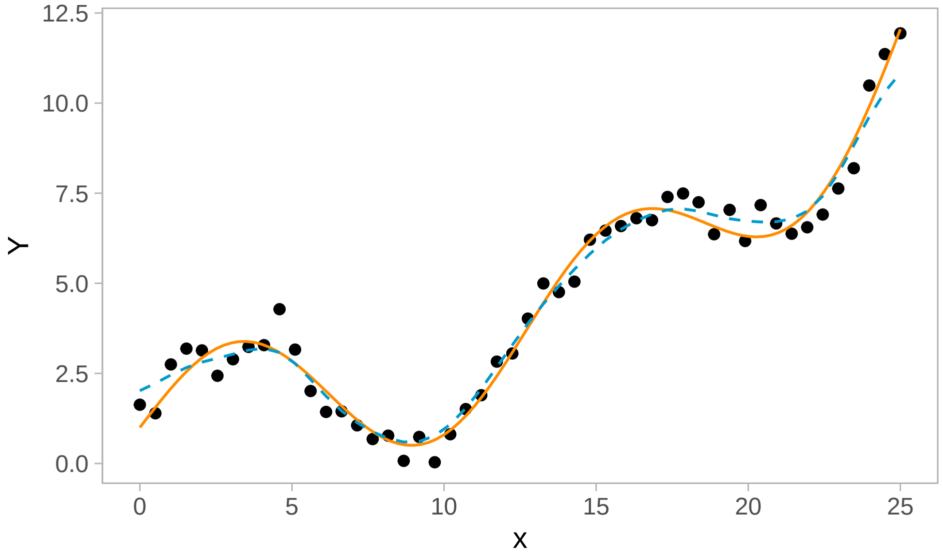
Consider the following case where we want to model the relationship between a single predictor and a response variable. Suppose the true relationship between x and Y is defined $Y = e^{\frac{x}{10}} + 2\sin(\frac{x}{2})$



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

Now that we have defined kernel machine regression, we can extend it to the Bayesian paradigm. In this section, we discuss the two methods for Bayesian variable selection in BKMR: hierarchical variable selection and component-wise variable selection (Bobb et al., 2015).

Component-wise selection follows the same framework as variable selection in a typical Bayesian multiple regression except, instead of augmenting each predictor, we augment the kernel function as

$$k(\mathbf{x}, \mathbf{x}'|\mathbf{r}) = \exp\{-\sum_{m=1}^{M} r_m (x_m - x'_m)^2\},$$

where $\mathbf{r} = [r_1, \ldots, r_M]^\top$. We define $r_m = \frac{1}{\rho_m}$, the inverse of the tuning parameter ρ_m for each $\mathbf{x}_m$, as we want to be able to shrink the influence of unimportant predictors in the model to $r_m = 0$. We now define the kernel matrix $\mathbf{K}_{\mathbf{X},\mathbf{r}}$ as the $n \times n$ matrix with (i, j)th element $k(\mathbf{x}, \mathbf{x}'|\mathbf{r})$. To allow r_m to equal 0 with non-zero probability, we first define an indicator variable determining whether or not a predictor is included in the variable, which is distributed as

$$\delta_m \sim \text{Bernoulli}(\pi),$$

where π is the prior probability of inclusion, and the posterior mean of δ_0 is the posterior probability of inclusion of $\mathbf{x}_m$. Now, we can assume a "slab-and-spike" prior on r_m, distributed as

$$r_m|\delta_m \sim \delta_m f(\cdot) + (1 - \delta_m) P_0,$$

where $f(\cdot)$ is some pdf with support $\mathbb{R}^+$ and P_0 denotes the density with point mass at 0.

While this process of component-wise variable selection works well in a typical multiple regression setting, it can lead to unreliable estimates in situations where the predictors are highly correlated with each other, which is common in exposure mixture studies. In this case, the correlated components contribute similar information to the model, and component-wise variable selection is not able to distinguish which predictor is important. BKMR deals with this problem by introducing hierarchical variable selection.

Hierarchical variable selection involves partitioning the predictors $\mathbf{x}_1, \ldots, \mathbf{x}_M$ a priori into groups $\mathcal{S}_g$ with $g = 1, \ldots, G$. These groups are selected with the aim of keeping within-group correlation high and between-group correlation low. The indicators from $r_m|\delta_m$ are now distributed as

$$\boldsymbol{\delta}_{\mathcal{S}_g}|\omega_g \sim \text{Multinomial}(\omega_g, \boldsymbol{\pi}_{\mathcal{S}_g}), g = 1, \ldots, G, \omega_g \sim \text{Bernoulli}(\pi),$$

where $\boldsymbol{\delta}_{\mathcal{S}_g} = \{\delta_m | \mathbf{x}_m \in \mathcal{S}_g\}$ and $\boldsymbol{\pi}_{\mathcal{S}_g}$ are vectors of indicator variables and prior probabilities, respectively, of a predictor $\mathbf{x}_m$ in group $\mathcal{S}_g$ entering the model. By this approach, at most one predictor in each group is allowed to enter the model.

While hierarchical variable selection resolves the issue of multicollinearity, it requires specifying subgroups of predictors a priori and assumes that one predictor in each group can capture the information of the rest. Hence, care should be taken to justify the partitioning of predictors when taking this approach.

3.2.5 Prior specification

In this section, we specify the prior distributions and parameters used in BKMR.

- $r_m|\delta_m \sim \delta_m \text{Unif}^{-1}(a_r, b_r) + (1 - \delta_m)P_0$

- $\rho_m = 1/r_m \sim \text{Unif}(a_r, b_r)$

- $\delta_m|\pi \sim \text{Bernoulli}(\pi)$

- $\delta_m|\pi \sim \text{Bernoulli}(\pi)$

- $\delta_m|\pi \sim \text{Bernoulli}(\pi)$

- $\pi \sim \text{Beta}(a_\pi, b_\pi)$

- $\sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$

- $\lambda = \tau\sigma^{-2} \sim \text{Gamma}(a_\lambda, b_\lambda)$

Discuss priors, what form do they take? what are the default settings?

3.2.6 The MCMC algorithm

Briefly, we discuss the algorithm used in the BKMR package (Bobb et al., 2015; Bobb, Claus Henn, Valeri, & Coull, 2018), with commentary on its implications for the model fitting process.

BKMR uses a Markov chain Monte Carlo (MCMC) sampler with a hybrid Gibbs and Metropolis-Hastings sampler to estimate the posterior distributions. In particular, a Gibbs step is used to update the distribution of σ^2 while a Metropolis-Hastings step is used to update the distribution of λ. For component-wise and hierarchical variable selection, $(\mathbf{r}, \boldsymbol{\delta}, \boldsymbol{\omega})$ are sampled jointly using a Metropolis-Hastings sampling scheme.

While each distribution generated by the Gibbs step is always accepted, the distributions for λ and r_m generated by the Metroplis-Hastings steps have a probability of acceptance that is dependent on the standard deviation of the proposal distribution (Bobb, 2017). This standard deviation is a tuning parameter; in general, increasing the standard deviation leads to lower acceptance rates. Acceptance rates that are too low lead to slower convergence, but rates that are too high can cause convergence to an non-optimal distribution.

A major computational limitation of BKMR is that at each iteration of the MCMC algorithm, the $n \times n$ augmented kernel matrix $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ must be inverted multiple times. BKMR employs a Gaussian predictive process which involves specifying a set l points, or "knots," that is a subset of the predictor space. The vector of predictors can be approximated by projection on this lower dimensional space, which allows the algorithm to perform inversions on an $l \times l$ matrix.

3.2.7 Detecting interactions

Discuss options for detecting interactions, i.e., can visualize relationship at various quantiles of other predictor, can conduct inference on inter-quantile difference

Potentially provide toy example

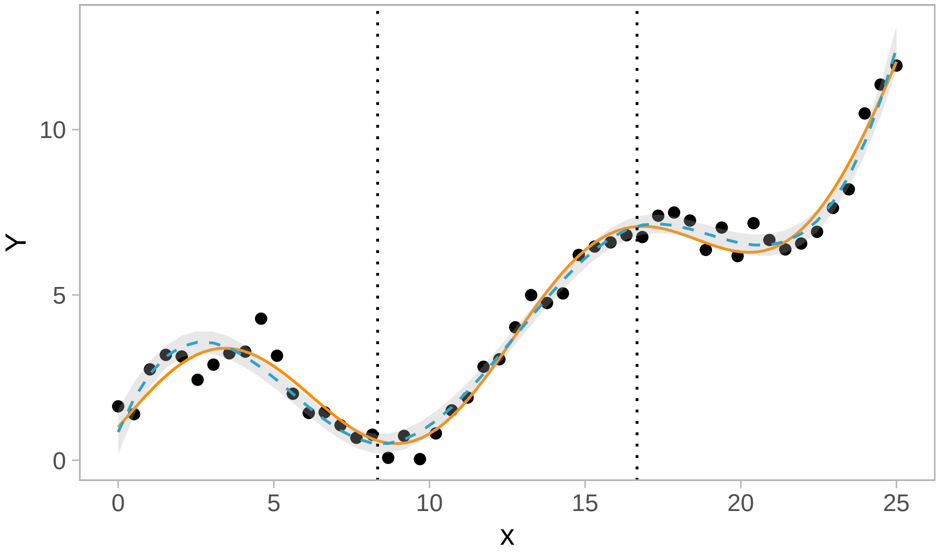
3.3 Bayesian semiparametric regression

differences w/ bkmr

- makes distributional assumptions about the dataset
- kernel regression is computationally intensive w/ large datasets but can handle many predictors
- bsr highly dependent on the choice of function

3.3.1 Connections to linear mixed model

3.3.2 Toy example



–>

Chapter 4 Simulations

4.1 Past simulation studies

4.2 MADRES data

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 ??:

A.2 In Chapter ??:

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

Bobb, J. F. (2017, December). Example using the bkmr R package with simulated data from the NIEHS mixtures workshop. Retrieved from https://jenfb.github.io/bkmr/SimData1.html#1_load_packages_and_download_data

Bobb, J. F., Claus Henn, B., Valeri, L., & Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health*, *17*(1), 67. http://doi.org/10.1186/s12940-018-0413-y

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., … Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, *16*(3), 493–508. http://doi.org/10.1093/biostatistics/kxu058

Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, *63*(4), 1079–1088. http://doi.org/10.1111/j.1541-0420.2007.00799.x