Flexible Bayesian 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

Abstract . . . . **i**

Acknowledgments . . . . **iii**

List of Tables . . . . **vi**

List of Figures . . . . **vii**

Chapter 1: Introduction **1**

Chapter 2: Humanistic Perspective **5**

Chapter 3: Bayesian Regression Methods **7**

- 3.1 Motivation 7
  - 3.1.1 Interactions from a statistical perspective 7
  - 3.1.2 Mechanistic and public health relevance 9
- 3.2 Bayesian kernel machine regression 10
  - 3.2.1 Kernel machine regression 11
  - 3.2.2 Connection to mixed models 14
  - 3.2.3 Toy example 15
  - 3.2.4 Variable selection 18
  - 3.2.5 Prior specification 20
  - 3.2.6 The MCMC algorithm 21
- 3.3 Bayesian semiparametric regression 22
  - 3.3.1 Spline regression 22

3.3.2 Toy example 24

3.3.3 Model formulation in BSR 28

3.3.4 Sparsity inducing priors 29

3.3.5 Prior specification 30

3.3.6 The MCMC algorithm 31

3.4 Detecting interactions 31

3.4.1 BKMR 31

3.4.2 BSR 31

3.4.3 Differences between BKMR and BSR 32

Chapter 4: Simulations **33**

4.1 Past simulation studies 33

4.2 Methods 33

4.2.1 Models 33

4.2.2 MADRES data 33

4.2.3 Simulating data 34

4.2.4 Model assessment 34

4.3 Results 34

Conclusion **35**

Appendix A: Code **37**

A.1 Code for Chapter 3: 37

Appendix B: The Second Appendix **43**

Corrections **45**

References **47**

List of Tables

List of Figures

3.1 Non-linear data with a true distribution (orange) and a fitted linear regression (blue). 15

3.2 A query point of 12.5 and the weights of neighboring observations based on a Gaussian kernel 16

3.3 Fitted kernel machine regression (blue) with $\rho = 2$ compared to the true distribution (orange). 17

3.4 Fitted kernel machine regression with $\rho = 0.02$ and $\rho = 50$. 17

3.5 Linear spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange). 25

3.6 Cubic spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange). 26

3.7 Natural spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange). 27

3.8 Natural and cubic spline regression (blue) compared to the true distribution (orange) extrapolated outside the bounds of x (dotted lines). . 28

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. In fact, it has been estimated that, globally, human activity releases chemicals at a rate of 220 billion tons per annum (Cribb, 2016). These staggering levels of pollution have led scholars to formally declare that 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). In this new era of pervasive toxicity, understanding the nature and severity of health effects associated with chemical exposures is especially timely.

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, Shore, Pereira, Jones, & Martin, 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, Board on Environmental Studies and Toxicology, Division on Earth and Life Studies, & National Academies of Sciences, Engineering, and Medicine, 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, Benka-Coker, Severson, Magzamen, & Wilson (2021); lazarevic_performance_2020; pesenti_comparative_2023]. However, to our knowledge, there has yet to be a simulation study which provides conclusive guidance about the ability of these methods to conduct inference on non-additive interactions between exposures, particularly when the nature and effect sizes of these interactions vary.

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 multiple environmental exposures and related covariates. 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. 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

List of loose ideas:

- goal of modernism: define single causes for single effects
  - study chemicals in isolation to obtain purely mechanistic explanation of their toxicity
- contrast with idea of relationality: entities cannot be understood without considering their *relationality* to other surrounding entities
  - relationality disrupts the notion of bounded objects
  - motivation for chemical mixtures: chemicals themselves are not independent from surrounding chemicals
  - motivation for chemical mixtures in the context of social epidemiology: the effects of chemicals are modulated by structural/social conditions
- relationality also disrupts Cartesian split between body and mind
  - racial hierarchy positions certain groups closer to the bounds of the corporal body, while other groups have transcended these bounds and are defined by their intellect (i.e., the mind)
  - result –> some bodies are seen as inherently more susceptible to chemical exposure, more "porous"
  - leads to damage centered research which, while well-intentioned, inadvertently de-humanizes marginalized groups

- remedy: relationality leads into concept of alterlife, modern life is inseparable from alteration due to chemical exposure

Chapter 3 Bayesian Regression Methods

3.1 Motivation

We are interested in using Bayesian regression techniques to characterize the nature of complex interactions in exposure mixture studies. We begin by reviewing definitions for what constitutes a complex interaction and why interactions are relevant from public health and biological relevance.

3.1.1 Interactions from a statistical perspective

First, we define additivity and non-additivity in the traditional statistical paradigm (Siemiatycki & Thomas, 1981). Suppose we have two variables x_1 and x_2, and we want to consider their effect on some outcome of interest. If specifying [effect due to x_1 and x_2] = [effect due to x_1] + [effect due to x_2] can adequately capture this relationship, then we say that x_1 and x_2 each have an **additive effect** on the outcome and that there is no interaction between them. On the other hand, if there is variability in the outcome that can be captured by an additional term equal to some function of x_1 and x_2, we say that there is a **non-additive interaction** between x_1 and x_2. In this case, [effect due to x_1 and x_2] = [effect due to x_1] + [effect due to x_2] + [effect due to $f(x_1, x_2)$], where f is non-zero for some values of x_1 and x_2.

For our sake, we consider any non-additive interaction to be complex, meaning that they are difficult to detect. To see why, let us consider running a linear regression for Y on x_1 and x_2. The regression would be defined as

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_{12} f(x_1, x_2),$$

where the β's represent the effect sizes. We can see that the form of the interaction must be explicitly specified in the formulation of the model. Most commonly, a multiplicative interaction is assessed, where $f(x_1, x_2) = x_1 * x_2$. However, a non-additive interaction can take on many different forms, the true nature of which is difficult to determine analytically.

We used a two-predictor case above, but interactions can also exist between two or more variables (i.e., two-way by $f(x_1, x_2)$, three-way by $f(x_1, x_2, x_3)$, etc.). And if we wanted to assess all possible interactions, the number to consider quickly becomes intractable in high-dimensional settings. For instance, consider modelling 10 predictors in the above linear regression setting. In order to be assessed, each interaction must be explicitly specified as a new term in the model. Even if we only considered one form for the interaction, including all 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.

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. For instance, Halford, Baker, McCredden, & Bain (2005) suggest that there is a steep decline in interpretability from three- to four-way interactions, and that five-way interactions are only interpreted correctly at chance level (Halford et al., 2005). Hence, for practical purposes, we will limit our exploration to two- and three-way interactions.

3.1.2 Mechanistic and public health relevance

Thus far, we have discussed interactions within a statistical paradigm. However, in addition to being an interesting estimation challenge, non-additive interactions are also relevant in exposure mixture studies from both a mechanistic and public health point of view.

From a mechanistic perspective, a non-additive statistical interaction between two chemical exposures suggests that these molecules may be functionally interacting with each other. Theoretical models propose that such interactions can be classified as either synergistic or antagonistic (Heys et al., 2016; Plackett & Hewlett, 1952). In a synergistic interaction, the joint effects of a mixture exceed the independent effects of each component. This usually occurs if a chemical induces an enzyme involved with the activation of a second chemical or if a chemical inhibits an enzyme that would have otherwise degraded a second chemical. For example, various synergistic interactions between heavy metals have been documented in the literature (e.g., Paithankar, Saini, Dwivedi, Sharma, & Chowdhuri, 2021; Wang, Mukherjee, & Park, 2018).

On the other hand, in an antagonistic interaction, the joint effects of a mixture are less than their independent effects. This can occur either through competition at the target site of an enzyme or through direct chemical reactions with each other. In general, synergistic interactions are more concerning in risk assessments, as they lead to underestimation of the true toxicity of a mixture.

It should be noted, though, that while statistical interactions may provide some insight into how exposure mixtures are related to health, they cannot confirm their underlying biology (VanderWeele & Knol, 2014). If the goal is to assess a meaningful biological interaction, then the discovery of a statistical interaction should be followed up by a functional study.

Now, from a public health perspective, we might be interested in how exposure mixtures interact with other covariates, or, in other words, how social and health factors might mediate the relationship between a health outcome and chemical exposures (VanderWeele & Knol, 2014). In our case, we can include these additional covariates in the exposure mixture model, where, statistically, they would contribute to the model in the same manner as another chemical exposure: a predictor. A statistical interaction in our model between a covariate and an exposure would indicate that the *magnitude* of the effect of reducing the level of an exposure might differ across various levels of the covariate. This finding could be relevant to public health policy makers, as the potential benefit of regulating a pollutant might differ across groups. For instance, it has been suggested that nutritional intake may modify susceptibility to chemical exposures (e.g., Kannan, Misra, Dvonch, & Krishnakumar, 2006; Kordas, Lönnerdal, & Stoltzfus, 2007).

In many cases, we might assess a covariate related to health inequity, such as socioeconomic status. We provide a cautionary comment, here, that an interaction term should not be the *sole* measure used to measure a health disparity (Ward et al., 2019). In this case, we should first consider the independent, additive association between the covariate and levels of exposure or rates of a health outcome, in order to contextualize the meaning of a potential interaction term.

3.2 Bayesian kernel machine regression

In this section, we introduce the theory of BKMR. First, we define the notation that we will be using for kernel machine regression:

- X_m is a predictor variable in the predictor matrix $\mathbf{X}$ with $m = 1, \ldots, M$, measuring exposure variables or covariates

- $\mathbf{x}_i$ is a vector of values for a single observation in $\mathbf{X}$ with $i = 1, \ldots, n$
- x_{im} is the ith observation of X_m
- Y_i is an observation of $\mathbf{Y}$, measuring the health outcome in this case
- $h(\cdot)$ is the flexible function relating $\mathbf{x}$ to $\mathbf{Y}$
- 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 parameter inside the kernel function which controls smoothness
- τ is the parameter multiplied by the kernel matrix to relate $\mathbf{K}$ to h, and
- $\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are the residuals of the response.

And, we define the notation that we will be using specific to BKMR:

- $r_m = 1/\rho_m$ is an augmented variable in $\mathbf{r}$ in the kernel matrix, 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 of a predictor in the model, and
- $\lambda \equiv \tau\sigma^{-2}$ is a convenient way to define the prior on τ.

3.2.1 Kernel machine regression

We begin by introducing kernel machine regression, with attention to its specific implementation in BKMR. First proposed by Nadaraya (1964) and Watson (1964), kernel machine regression is a nonparametric regression technique that can be used to capture non-linear effects and non-additive interactions. In this introduction, we follow the presentation of kernel machine regression provided by Bobb et al. (2015).

To contextualize this method, we start at the typical linear regression setting,

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

where Y_i measures a health outcome at a given point, $\mathbf{x}_i = [x_{i1}, \ldots, x_{iM}]$ 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 flexible function $h : \mathbb{R}^M \rightarrow \mathbb{R}$, 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 $k : \mathbb{R}^M \times \mathbb{R}^M \rightarrow \mathbb{R}$.

There are many choices of functions for 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

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

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 another subject $\mathbf{x}'$, and 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 the vertical scale of h.

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, Müller (1987) demonstrated 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

$$\mathrm{cor}(h_i, h_j) = \exp\left\{ - \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\rho} \right\},$$

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 to mixed models

It is useful to make connections between this definition of kernel machine regression and mixed models. Liu, Lin, & Ghosh (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 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 (Schulz, Speekenbrink, & Krause, 2018). 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. Here, h can be interpreted as a random effect.

3.2.3 Toy example

In the following section, we illustrate 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 $Y = e^{\frac{x}{10}} + 2\sin(\frac{x}{2})$. We simulate 51 equally spaced observations of x from 0 to 25, with error $\epsilon_i \overset{\text{iid}}{\sim} N(0, 0.25)$.

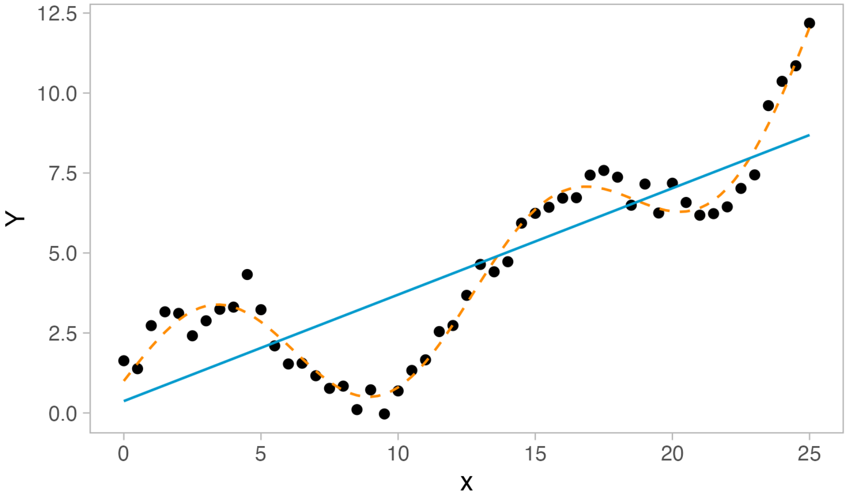

Figure 3.1: Non-linear data with a true distribution (orange) and a fitted linear regression (blue).

Figure 3.1 illustrates the shape of our simulated non-linear data and the fit proposed by a simple linear regression. We can observe that the linear regression fails to capture the true non-linear relationship. In this case, this would lead to an underestimation of the true association between x and Y. Now, we will try to capture this relationship using kernel machine regression.

To visualize how kernel machine regression works as a moving weighted average, we can consider a query point of 12.5.

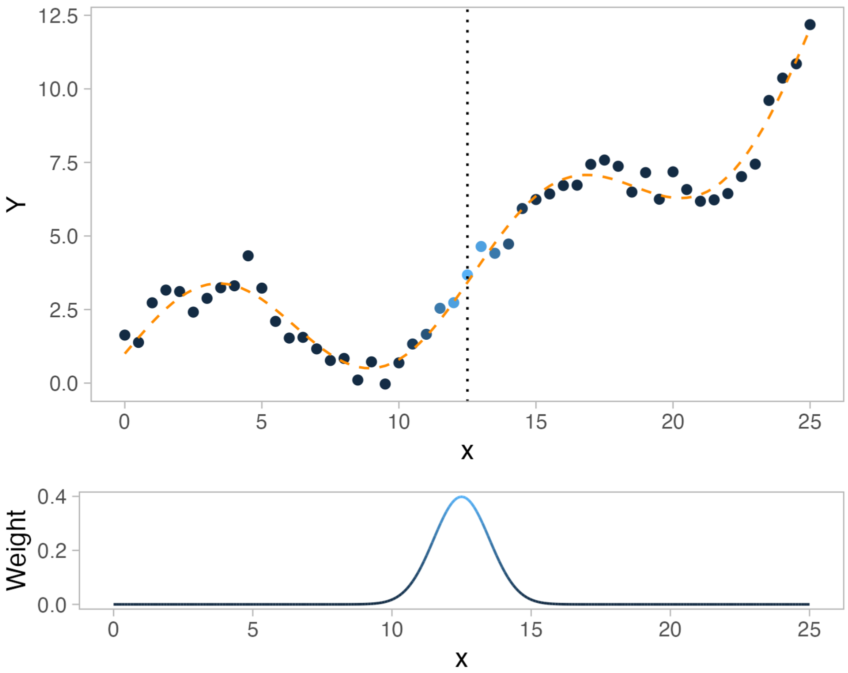

Figure 3.2: A query point of 12.5 and the weights of neighboring observations based on a Gaussian kernel

Figure 3.2 identifies the query point and assigns corresponding weights to the neighboring points based on a normal distribution, which shares the same density as the Gaussian kernel. In this case, we will specify $\rho = 2$, which is synonymous with assigning the weights using a normal distribution with $\sigma^2 = 1$. We can see how an appropriate estimate for $h(12.5)$ can be obtained by taking a weighted average of the Y's, with those observations nearby weighted the most heavily.

Now, we fit a kernel machine regression on this data with $\rho = 2$ using the `stats` package in R (R Core Team, 2013).

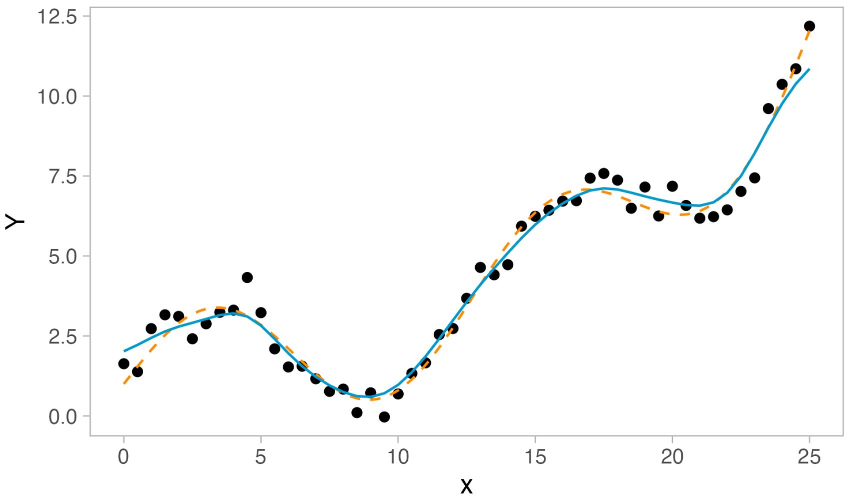

Figure 3.3: Fitted kernel machine regression (blue) with $\rho = 2$ compared to the true distribution (orange).

We can see in Figure 3.3 that kernel machine regression captures the complex non-linear relationship between Y and x and closely follows the true distribution. We do note, though, that the estimation is less precise at the tails, where there is less information provided by local observations. We can also use this example to consider the effect of various values of ρ on the smoothness of the h function.

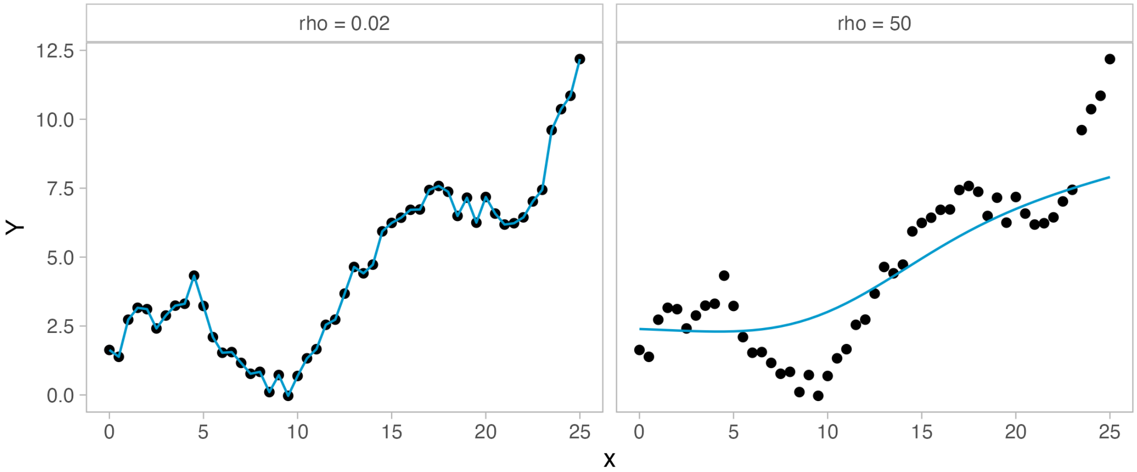

Figure 3.4: Fitted kernel machine regression with $\rho = 0.02$ and $\rho = 50$.

Figure 3.4 demonstrates the effect of relatively smaller and larger values of ρ on h. Decreasing the value of ρ allows kernel machine regression to overfit to the noise in the data by relaxing the dependence of neighboring values of h to each other. On the other hand, increasing the value of ρ enforces more dependence in h and as such results in an underfit estimation. Hence, the choice of ρ has a strong effect on the performance of kernel machine regression.

3.2.4 Variable selection

Now that we have defined kernel machine regression, we can extend it to the Bayesian paradigm. Bobb et al. (2015) showed that the Bayesian approach can outperform frequentist kernel machine regression because simultaneous variable selection and estimation can better capture the exposure-response relationship. 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

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

where $\mathbf{r} = [r_1, \ldots, r_M]^\top$. We define $r_m = 1/\rho_m$, the inverse of the tuning parameter ρ_m for each $\mathbf{x}_m$, as we want r_m to be able to take on the value of 0 when a variable is removed during the selection process. 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. Now, we can assume a "slab-and-spike" prior on r_m, distributed as

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_M$ into groups $\mathcal{S}_g$ with $g = 1, \ldots, G$. These groups should be selected by the user based on a prior knowledge, 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.

Note also that the posterior means of δ_m generated from these variable selection procedures represent the posterior probability of inclusion of $\mathbf{x}_m$. We can interpret these posterior "inclusion probabilities" as measures of the relative importance of each predictor. These measures can be used to understand the contribution of each exposure or covariate to the health outcome of interest in the model.

3.2.5 Prior specification

In this section, we specify the default prior distributions and parameters used by the BKMR algorithm (Bobb et al., 2015).

BKMR, by default, assumes $\rho_m = 1/r_m \sim \text{Unif}(a_r, b_r)$, a flat prior between a_r and b_r for which the default values are 0 and 100, respectively (Bobb, 2017a). This defines the prior probability of ρ as equally distributed across any value from 0 to 100. This inverse of this prior corresponds to the slab component of the "slab-and-spike" prior, where $r_m|\delta_m \sim \delta_m \text{Unif}^{-1}(a_r, b_r) + (1 - \delta_m)P_0$. As a flat prior, this distribution should be chosen when we have no prior knowledge about the smoothness of the exposure-response function, with hyperparameters a_r and b_r selected to represent the range of values we expect ρ to potentially span.

We have seen that the smoothness of a kernel machine regression responds strongly to different values of $r_m = 1/\rho$, and, accordingly, the model fit of BKMR is sensitive to their prior distribution. In general, the posterior inclusion probabilities generated from the variable selection procedure are particularly sensitive to this prior, though their relative importance tends to remain stable (Bobb, Claus Henn, Valeri, & Coull, 2018). As such, the BKMR algorithm also offers the options to define uniform and gamma priors for the $r_m = 1/\rho$.

Moreover, BKMR assumes that the prior probability of including a predictor (δ_m) or group of predictors (ω_g) in the model is distributed $\pi \sim \text{Beta}(a_\pi, b_\pi)$. The default hyperparameters are $a_\pi = b_\pi = 1$, which represent a flat, uninformative prior between 0 and 1. When the hierarchical selection approach is applied, equal values for $\boldsymbol{\pi}_{\mathcal{S}_g}$, the probabilities of inclusion for each component in group $\mathcal{S}_g$, are assumed.

Finally, BKMR assumes that the inverse of the variance of the residuals is distributed $\sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$, with default values of $a_\sigma = b_\sigma = 0.001$, and that the vertical scale of h is parameterized by $\lambda \equiv \tau\sigma^{-2} \sim \text{Gamma}(a_\lambda, b_\lambda)$, with default values of a_λ, b_λ such that the mean and variance of λ are both equal to 10.

3.2.6 The MCMC algorithm

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

BKMR uses a Markov chain Monte Carlo (MCMC) algorithm with a mix of Gibbs and Metropolis-Hastings samplers 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 are accepted based on an acceptance rate (Wagaman & Dobrow, 2021). The standard deviation of the proposal distribution controls the acceptance rate and as such acts as a tuning parameter (Bobb, 2017b). 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 can employ a Gaussian predictive process which involves specifying a set of l points, or "knots," that are 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. A general suggestion is to use this approach to speed up the algorithm when n is large and to specify $l \approx n/10$ (Bellavia, 2021).

3.3 Bayesian semiparametric regression

Defining notation for spline regression:

- x

BSR specific notation:

- x

3.3.1 Spline regression

We begin by introducing spline regression, with attention to its specific implementation in BSR. Spline regression is a semiparametric regression technique that can be used to capture non-linear effects. In this introduction, we follow the presentation of spline regression provided by Antonelli et al. (2020), with additional details and explanation from Hastie, Tibshirani, & Friedman (2009).

Spline regression defines the regression relationship as

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

where f is defined by a set of basis functions and ϵ_i is a random variable from $\boldsymbol{\epsilon} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. BSR uses natural spline bases. In order to understand how these are constructed, we start with a broad definition of basis expansions, before exploring linear, cubic, and then natural spline bases.

We determine the basis expansion by considering a piece-wise function of X_i with some order P and some set of K knots defining $K+1$ disjoint intervals. BSR places knots at uniformly sized quantiles within the boundaries of X_i. The most commonly used orders are $P = 1, 2,$ and 4, the constant, linear, and cubic splines, respectively. To begin, let us consider a continuous piece-wise linear spline basis (i.e., $P = 2$) of a one-dimensional X with two interior knots. In this case, we use the following four basis functions:

$$b_1(X) = 1, \qquad b_2(X) = X, \qquad b_3(X) = (X - \xi_1)_+, \qquad b_4(X) = (X - \xi_2)_+,$$

where ξ_1 and ξ_2 are the two interior knots, and t_+ denotes the positive part. These bases are used to construct the regression model $f(X) = \sum_{j=1}^{4} \beta_j b_j(X)$, which requires estimating $K + P = 4$ parameters. We can check the continuity restrictions at the knots by seeing that $f(\xi_1^-) = \beta_1 + \xi_1 \beta_2$ and $f(\xi_1^+) = \beta_1 + \xi_1 \beta_2 + (\xi_1 - \xi_1)\beta_3$ are equal.

Now, in the case of exposure mixtures, we want smoother functions that can capture the non-linear relationship between the response and the predictors. We can achieve this by increasing the order to $P = 4$ and using a cubic spline, with continuous first and second derivatives at the knots. The cubic spline is the lowest-order spline for which knot-discontinuity cannot be detected by the human eye. For example, for one X with two interior knots, we use the following six basis functions:

$$b_1(X) = 1, \qquad b_2(X) = X, \qquad b_3(X) = X^2,$$
$$b_4(X) = X^3, \qquad b_5(X) = (X - \xi_1)^3_+, \qquad b_6(X) = (X - \xi_2)^3_+.$$

Now, the regression model is defined as $f(X) = \sum_{j=1}^{6} \beta_j b_j(X)$ and requires estimating $K + P = 6$ parameters. It can be shown that $f'(\xi_i^-) = f'(\xi_i^+)$ and $f''(\xi_i^-) = f''(\xi_i^+)$.

However, the behavior of polynomials near the boundaries of X, where there is less information, can be erratic. Natural cubic splines, also referred to as just natural splines, address this by imposing an additional restriction of linearity at the boundaries of X. Paradoxically, this also leads to a simpler model with four fewer parameters to estimate. A general definition of the K basis functions for a natural spline with interior knots ξ_j, $j = 1, \ldots, K$, is given by:

$$N_1(X) = 1, \qquad N_2(X) = X, \qquad N_{k+2}(X) = d_k(X) - d_{K-1}(X),$$
$$d_k(X) = \frac{(X - \xi_k)^3_+ - (X - \xi_K)^3_+}{\xi_K - \xi_k}.$$

Here, the regression model is defined as $f(X) = \sum_{j=1}^{K} \beta_j N_j(X)$, with K parameters. BSR uses natural splines to specify the regression relationship.

3.3.2 Toy example

In the following section, we illustrate spline regression using the same toy example used to introduce kernel machiner regression. See Section 3.2.3 and Figure 3.1 for details on the parameters used to generate simulated data.

As in Section 3.2.3, we consider a case where we want to model the relationship between a single predictor and a response variable, where the true relationship between x and Y is defined $Y = e^{\frac{x}{10}} + 2\sin(\frac{x}{2})$. We fit a series of linear, cubic, and then natural spline regressions to illustrate the general framework of a natural spline regression.

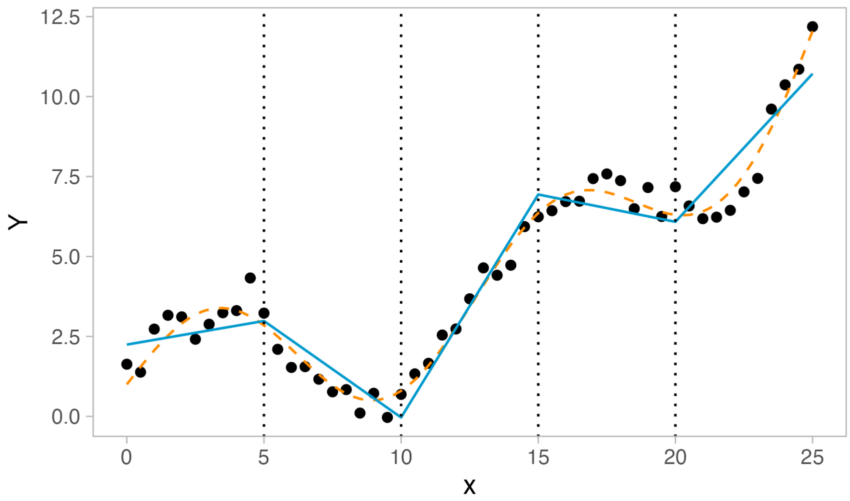

Figure 3.5: Linear spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange).

Figure 3.5 illustrates the fit proposed by a linear spline regression with order $P = 2$. We can see that the implementation of knots allows for even a linear fit to capture more of the nuances in this nonlinear relationship, as compared to a standard linear regression.

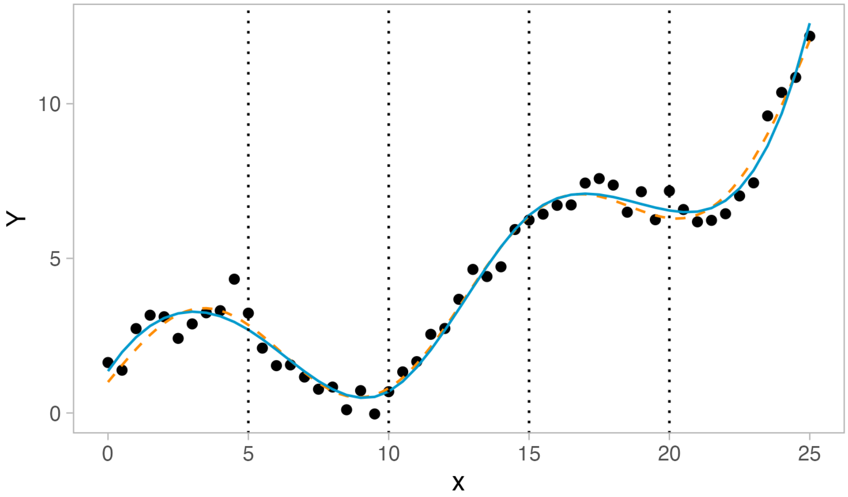

Figure 3.6: Cubic spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange).

However, this linear spline regression is still unable to fully estimate the nonlinearity in our example. Increasing the order to $P = 4$ with a cubic spline regression offers additional flexibility. Figure 3.6 illustrates the fit proposed by this model. Here, we can see the benefits of using a cubic polynomial relationship in a nonlinear setting: the estimated relationship is continuous at the knots, and the nonlinear relationship has been flexibly captured.

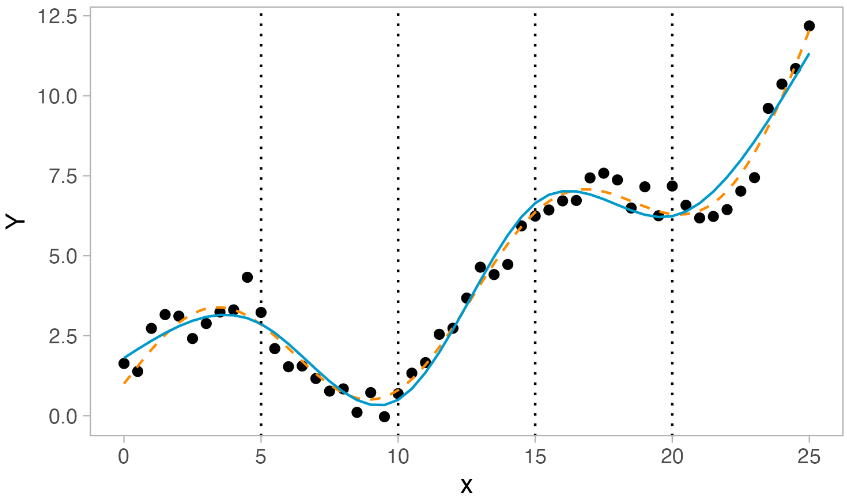

Figure 3.7: Natural spline regression (blue) with four knots (dotted lines) compared to the true distribution (orange).

Our final modification involves imposing linearity constraints on the boundaries of x to implement a natural spline regression. Figure 3.7 shows that the fit estimated using a natural spline regression. The fitted line is only slightly different than that proposed by a cubic spline regression in Figure 3.6. The most noticeable difference is that the slopes of the tails of the natural spline regression are less extreme than for the cubic spline regression.

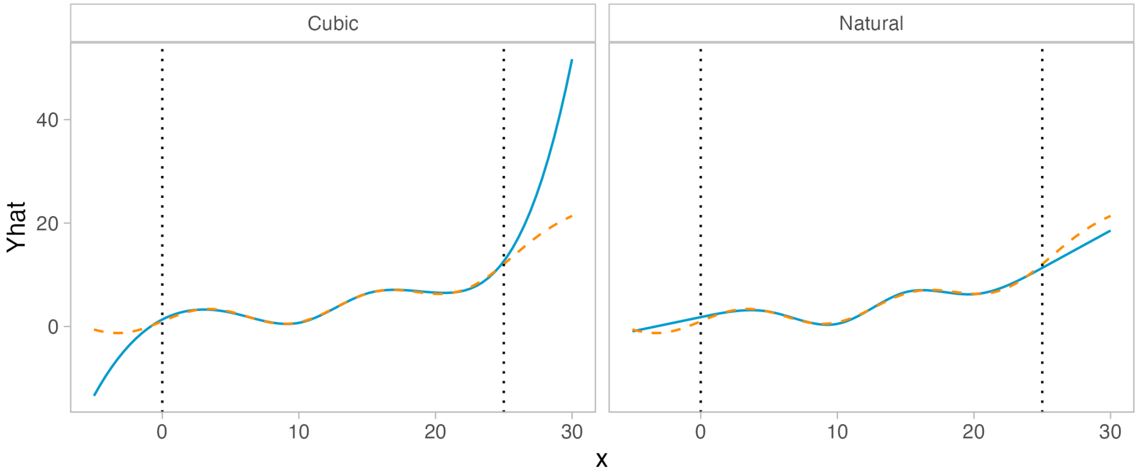

Figure 3.8: Natural and cubic spline regression (blue) compared to the true distribution (orange) extrapolated outside the bounds of x (dotted lines).

Extrapolation outside the scope of x allows us to see the effect of the lineary constraints imposed by natural regression. Figure 3.8 demonstrates how the cubic spline regression behaves erratically outside the bounds of x, as the cubic polynomial lines tend toward $\pm\infty$, while the natural spline regression follows a more appropriate linear trend. As the natural spline regression is more reliable near the boundaries of x and also simpler to estimate, BSR adopts the use of natural splines.

3.3.3 Model formulation in BSR

Now that we have defined natural splines, we introduce BSR, following the presentation in Antonelli et al. (2020). We demonstrate the construction of f in BSR by first assuming a two-dimensional case with predictors X_1 and X_2. We define

$$f(X_1) = \widetilde{X}_1\boldsymbol{\beta}_1, \qquad f(X_2) = \widetilde{X}_2\boldsymbol{\beta}_2,$$

$$f(X_1, X_2) = \widetilde{X}_1\boldsymbol{\beta}_1 + \widetilde{X}_2\boldsymbol{\beta}_2 + \widetilde{X}_{12}\boldsymbol{\beta}_{12},$$

where $\widetilde{X}_m = [b_{m1}(X_m), \ldots, b_{md}(X_m)]$ represents a d-dimensional basis function expansion for $m = 1, 2$, and $\widetilde{X}_{12} = [b_{11}(X_1)b_{21}(X_2), b_{11}(X_1)b_{22}(X_2), \ldots, b_{1d}(X_1)b_{2d}(X_2)]$ represents a d^2-dimensional basis expansion of the interaction between X_1 and X_2. Note that we must explicitly model the effect of the interaction term by assuming a multiplicative interaction between the basis functions of the predictors.

Extending to the multi-dimensional setting, BSR assumes the following general model formulation:

$$f(\mathbf{x}_i) = \sum_{h=1}^{p} f^{(h)}(\mathbf{x}_i),$$
$$f^{(h)}(\mathbf{x}_i) = \sum_{j_1=1}^{M} \widetilde{x}_{ij_1}\boldsymbol{\beta}_{j_1}^{(h)} + \sum_{j_1=2}^{M} \sum_{j_2<j_1} \widetilde{x}_{ij_1j_2}\boldsymbol{\beta}_{j_1j_2}^{(h)} + \ldots,$$

where $f^{(h)}(\mathbf{x}_i)$ includes a summation of all p-way interactions. The inclusion of all p-way interactions makes the model far too overparameterized. Moreover, $f(\mathbf{x}_i)$ is a sum of k different functions $f^{(h)}(\mathbf{x}_i)$ where a value for k is selected in order to capture all exposure effects in the model. Each of the k functions has the same functional form, and so the regression coefficients for a function $f^{(h)}(\mathbf{x}_i)$ are only identifiable up to a constant — this means that there are multiple sets of coefficients that could be estimated from the same data.

3.3.4 Sparsity inducing priors

In order to handle the overparameterization and non-identifiability of the model, BSR implements multivariate sparsity inducing priors. In this section, we follow the presentation provided in Antonelli et al. (2020).

First, we define indicators $\boldsymbol{\zeta} = \{\zeta_{jh}\}$ representing whether the jth exposure is included in the hth function:

$$P(\zeta_{jh}) = \tau_h^{\zeta_{jh}}(1-\tau_h)^{1-\zeta_{jh}} I(A_h \not\subset A_{h'} \forall h' \neq h \text{ or } A_h = \{\}),$$
$$\text{where } A_h = \{j : \zeta_h = 1\}.$$

Here, the indicators follow a Bernoulli distribution with prior probability of inclusion τ_h. The posterior means of $\boldsymbol{\zeta}$ can be interpreted as measures of relative variable importance. We include an indicator function $I()$ that represents whether the function h contains a unique set of predictors. This indicator ensures that no function contains predictors that are a subset of those in another function, in which case this function would be redundant and thus removed from the model entirely.

Now, we assume a multivariate slab-and-spike prior on the regression coefficients:

$$P(\boldsymbol{\beta}_X^{(h)}|\boldsymbol{\zeta}) = \left(1 - \prod_{j \in S} \zeta_{jh}\right) P_{\mathbf{0}} + \left(\prod_{j \in S} \zeta_{jh}\right) \psi_1(\boldsymbol{\beta}_S^{(h)}),$$
$$\text{where } S \text{ is some subset of } 1, 2, \ldots, p.$$

Here, $P_{\mathbf{0}}$ denotes the density with point mass at $\mathbf{0}$, and $\psi_1()$ is a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_\beta$, a diagonal matrix with $\sigma^2\sigma_\beta^2$ on the diagonals.

3.3.5 Prior specification

- hyper prior vs. empirical bayes strategy for sigma beta
- lower bound on slab variance
- properties of priors for higher-order interactions

3.3.6 The MCMC algorithm

- include brief detail on how MCMC algorithm deals with lower-order interactions that are a subset of higher-order interactions

3.4 Detecting interactions

In Section 3.1, we highlighted the challenges of analytically testing for the presence of interactions in exposure mixture studies. These challenges motivated a theoretical exploration of BKMR and BSR in Sections 3.2 and 3.3. Now, we discuss and compare the options that BKMR and BSR provide for inference on the presence of interactions. We also include discussion on theoretical advantages and disadvantages to each.

3.4.1 BKMR

Since the flexible h function in kernel machine regression allows us to forgo any assumptions about the nature of the relationship between the health outome and predictors, BKMR can potentially capture complex interactions between predictors. The challenge with using BKMR to do this, however, is that there is no formal framework for conducting inference on the presence of interactions.

And, if we specify hierarchical variable selection, then

How to detect interactions:

- can visualize relationship at various quantiles of other predictor
- can conduct inference on inter-quantile difference — has yet to be a formal assessment of how this approach performs in various settings

3.4.2 BSR

BSR can conduct inference

3.4.3 Differences between BKMR and BSR

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

Chapter 4 Simulations

4.1 Past simulation studies

Lit review of past sim studies which include bayesian regression approaches + others, note how they haven't focused on interactions

4.2 Methods

What are our aims?

4.2.1 Models

Models compared, specify the parameters for each (justify them!)

- MLR
- MLR with known form of interactions specified
- BKMR with component-wise
- BKMR with hierarchical
- BMIM? extension of BKMR with hierarchical
- BSR

4.2.2 MADRES data

- cite original study design paper

- cite Howe et al. 2021, which applied BKMR and BSR to this data, what did they find?

4.2.3 Simulating data

- use original predictor values in MADRES data. allows us to explore these methods on predictors with realistic correlation structure
- simulate outcome using a formula w/ different types of interaction
- change effect size (three levels?)
- change nature of interaction (mathematical formulation, two- or three-way, b/t just exposures or b/t exposures and covariates)

4.2.4 Model assessment

- how many times is interaction picked up?

4.3 Results

- figures + tables w/ model performance

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 Code

This first appendix includes all of the R chunks of code that were hidden throughout the document.

A.1 Code for Chapter 3:

The code for this chapter generates a toy example, used to demonstrate the kernel machine regression and spline regression techniques.

```
#load packages
library(tidyverse)
library(stats)
library(splines)
```

```
# set theme for plots
theme_set(theme_light())
theme_update(panel.grid.major = element_blank(),
             panel.grid.minor = element_blank())
theme_update(
  strip.background = element_rect(color="gray", fill="white"),
  strip.text = element_text(color = "gray30")
  )
```

```
########
# generate simulated points
########

# generate data from distribution
```

```r
set.seed(0) # reproducibility
x <- seq(0, 25, length.out = 51)
Y <- exp(x/10) + 2*sin(x/2) + rnorm(51, mean = 0, sd = 0.5)
df <- data.frame(x, Y)

# plot data and linear regression line
q1 <- ggplot(df, aes(x, Y)) +
  geom_point() +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_smooth(method = "lm", formula = "y~x",
              color = "deepskyblue3", fill = "gray70",
              linewidth = 0.5, se = F)

# save plot
ggsave("index/figures/ch3_toy1.png", plot = q1, device = "png",
       width = 5, height = 3)
```

```r
########
# kernel regression
########

# get normal distribution of weights around query points
df$Weight <- dnorm(df$x, mean = 12.5, sd = 1)

# plot points colored by their weights
p1 <- ggplot(df, aes(x, Y)) +
  geom_point(aes(color = Weight)) +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_vline(xintercept = 12.5, linetype = "dotted") +
  theme(legend.position = "none")

# plot a curve of weights
normcurv <- data.frame(x = seq(0, 25, length.out = 250))
normcurv$Weight <- dnorm(normcurv$x, mean = 12.5, sd = 1)
p2 <- ggplot(normcurv, aes(x, Weight, color = Weight)) +
  geom_line() +
  scale_y_continuous(breaks = c(0, 0.2, 0.4)) +
  theme(legend.position = "none")
```

```r
# stitch plots together
q2 <- cowplot::plot_grid(p1, p2, ncol = 1, rel_heights = c(0.7, 0.3))
q2

# save plot
ggsave("index/figures/ch3_toy2.png", plot = q2, device = "png",
       width = 5, height = 4)

# fit kernel regression with sigma = 1, bandwidth = 8/3
kmr_toy <- ksmooth(df$x, df$Y, kernel = "normal",
                   bandwidth = 8/3, x.points = df$x)
df <- df |>
  left_join(as.data.frame(kmr_toy), by = "x") |>
  rename(Yhat = y)

# plot kernel regression estimation
q3 <- ggplot(df) +
  geom_point(aes(x, Y)) +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_line(aes(x, Yhat), color = "deepskyblue3")
q3

# save plot
ggsave("index/figures/ch3_toy3.png", plot = q3, device = "png",
       width = 5, height = 3)

# fit kernel regression with sigma = 5, bandwidth = 40/3
kmr_toy_5 <- ksmooth(df$x, df$Y, kernel = "normal",
                     bandwidth = 40/3, x.points = df$x)

# fit kernel regression with sigma = 0.1, bandwith = 8/30
kmr_toy_1 <- ksmooth(df$x, df$Y, kernel = "normal",
                     bandwidth = 8/30, x.points = df$x)

# re-join data
dfrho <- df |>
  left_join(as.data.frame(kmr_toy_5), by = "x") |>
  rename("rho = 50" = y) |>
  left_join(as.data.frame(kmr_toy_1), by = "x") |>
  rename("rho = 0.02" = y) |>
  select(-Yhat) |>
  pivot_longer(cols = c("rho = 50", "rho = 0.02"), values_to = "Yhat")

# plot kernel regression with two values of rho
qrho <- ggplot(dfrho) +
  geom_point(aes(x, Y)) +
  geom_line(aes(x, Yhat), color = "deepskyblue3") +
  facet_wrap(~name)
qrho

# save plot
ggsave("index/figures/ch3_toyrho.png", plot = qrho, device = "png",
       width = 7, height = 3)
```

```
########
# spline regression
########

kn <- c(5, 10, 15, 20) # 4 knots of equal width

# fit linear spline regression
spline_toy_line <- lm(Y ~ bs(x, knots = kn, degree = 1), data = df)
p_line <- predict(spline_toy_line, se = T)
df$Yhats_line <- p_line$fit

q4 <- ggplot(df) +
  geom_point(aes(x, Y)) +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_line(aes(x, Yhats_line), color = "deepskyblue3") +
  geom_vline(xintercept = kn, linetype = "dotted")
q4

# save plot
ggsave("index/figures/ch3_toy4.png", plot = q4, device = "png",
       width = 5, height = 3)

# fit cubic spline regression
spline_toy_cub <- lm(Y ~ bs(x, knots = kn), data = df)
p_cub <- predict(spline_toy_cub, se = T)
df$Yhats_cub <- p_cub$fit
```

```r
# plot spline regression estimation
q5 <- ggplot(df) +
  geom_point(aes(x, Y)) +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_line(aes(x, Yhats_cub), color = "deepskyblue3") +
  geom_vline(xintercept = kn, linetype = "dotted")
q5

# save plot
ggsave("index/figures/ch3_toy5.png", plot = q5, device = "png",
       width = 5, height = 3)

# fit natural spline regression
spline_toy_nat <- lm(Y ~ ns(x, knots = kn), data = df)
p_nat <- predict(spline_toy_nat, se = T)
df$Yhats_nat <- p_nat$fit

# plot spline regression estimation
q6 <- ggplot(df) +
  geom_point(aes(x, Y)) +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_line(aes(x, Yhats_nat), color = "deepskyblue3") +
  geom_vline(xintercept = c(5, 10, 15, 20), linetype = "dotted")
q6

# save plot
ggsave("index/figures/ch3_toy6.png", plot = q6, device = "png",
       width = 5, height = 3)

# see what happens outside of the bounds
x_longer <- seq(-5, 30, length.out = 81)
y_longer_cub <- predict(spline_toy_cub,
                        newdata = data.frame(x = x_longer))
y_longer_nat <- predict(spline_toy_nat,
                        newdata = data.frame(x = x_longer))

df_longer <- data.frame(
  x = c(x_longer, x_longer),
  spline = c(rep("Cubic", 81), rep("Natural", 81)),
  Yhat = c(y_longer_cub, y_longer_nat)
)

# plot outside of bounds
qbounds <- ggplot(df_longer) +
  geom_line(aes(x, Yhat), color = "deepskyblue3") +
  geom_function(fun = function(x) exp(x/10) + 2*sin(x/2),
                linetype = "dashed", color = "darkorange") +
  geom_vline(xintercept = c(0, 25), linetype = "dotted") +
  facet_wrap(~spline)
qbounds

# save plot
ggsave("index/figures/ch3_toybounds.png", plot = qbounds,
       device = "png", width = 7, height = 3)
```

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

Antonelli, J., Mazumdar, M., Bellinger, D., Christiani, D., Wright, R., & Coull, B. (2020). Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *The Annals of Applied Statistics*, *14*(1), 257–275. http://doi.org/10.1214/19-AOAS1307

Bellavia, A. (2021). *Statistical Methods for Environmental Mixtures.* Retrieved from `https://bookdown.org/andreabellavia/mixtures/preface.html`

Bobb, J. F. (2017a, March). Introduction to Bayesian kernel machine regression and the bkmr R package. Retrieved from `https://jenfb.github.io/bkmr/overview.html`

Bobb, J. F. (2017b, 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

Committee on Incorporating 21st Century Science into Risk-Based Evaluations, Board on Environmental Studies and Toxicology, Division on Earth and Life Studies, & National Academies of Sciences, Engineering, and Medicine. (2017). *Using 21st Century Science to Improve Risk-Related Evaluations*. Washington, D.C.: National Academies Press. http://doi.org/10.17226/24635

Cribb, J. (2016). *Surviving the 21st century: Humanity's ten great challenges and how we can overcome them*. Springer.

Gibson, E. A., Nunez, Y., Abuawad, A., Zota, A. R., Renzetti, S., Devick, K. L., … Kioumourtzoglou, M.-A. (2019). An overview of methods to address distinct research questions on environmental mixtures: An application to persistent organic pollutants and leukocyte telomere length. *Environmental Health*, *18*(1), 76. http://doi.org/10.1186/s12940-019-0515-1

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How Many Variables Can Humans Process? *Psychological Science*, *16*(1), 70–76. http://doi.org/10.1111/j.0956-7976.2005.00782.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-84858-7

Heys, K., Shore, R., Pereira, M., Jones, K., & Martin, F. (2016). Risk assessment of environmental mixture effects. *RSC Advances*, *6*(53), 47844–47857. http://doi.org/10.1039/C6RA05406D

Hoskovec, L., Benka-Coker, W., Severson, R., Magzamen, S., & Wilson, A. (2021).

Model choice for estimating the association between exposure to chemical mixtures and health outcomes: A simulation study. *PLOS ONE*, *16*(3), e0249236. http://doi.org/10.1371/journal.pone.0249236

Kannan, S., Misra, D. P., Dvonch, J. T., & Krishnakumar, A. (2006). Exposures to Airborne Particulate Matter and Adverse Perinatal Outcomes: A Biologically Plausible Mechanistic Framework for Exploring Potential Effect Modification by Nutrition. *Environmental Health Perspectives*, *114*(11), 1636–1642. http://doi.org/10.1289/ehp.9081

Kordas, K., Lönnerdal, B., & Stoltzfus, R. J. (2007). Interactions between nutrition and environmental exposures: Effects on health outcomes in women and children. *The Journal of Nutrition*, *137*(12), 2794–2797. http://doi.org/10.1093/jn/137.12.2794

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

Müller, H.-G. (1987). Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting. *Journal of the American Statistical Association*, *82*(397), 231–238. http://doi.org/10.1080/01621459.1987.10478425

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications*, *9*(1), 141–142. http://doi.org/10.1137/1109020

Naidu, R., Biswas, B., Willett, I. R., Cribb, J., Kumar Singh, B., Paul Nathanail, C., … Aitken, R. J. (2021). Chemical pollution: A growing peril and potential catastrophic risk to humanity. *Environment International*, *156*, 106616. http://doi.org/10.1016/j.envint.2021.106616

Paithankar, J. G., Saini, S., Dwivedi, S., Sharma, A., & Chowdhuri, D. K. (2021). Heavy metal associated health hazards: An interplay of oxidative stress and signal transduction. *Chemosphere*, *262*, 128350. http://doi.org/10.1016/j.chemosphere.2020.128350

Persson, L., Carney Almroth, B. M., Collins, C. D., Cornell, S., De Wit, C. A., Diamond, M. L., … Hauschild, M. Z. (2022). Outside the Safe Operating Space of the Planetary Boundary for Novel Entities. *Environmental Science & Technology*, *56*(3), 1510–1521. http://doi.org/10.1021/acs.est.1c04158

Plackett, R. L., & Hewlett, P. S. (1952). Quantal Responses to Mixtures of Poisons. *Journal of the Royal Statistical Society: Series B (Methodological)*, *14*(2), 141–154. http://doi.org/10.1111/j.2517-6161.1952.tb00108.x

R Core Team. (2013). *R: A language and environment for statistical computing: Reference index*. Vienna: R Foundation for Statistical Computing.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16. http://doi.org/10.1016/j.jmp.2018.03.001

Siemiatycki, J., & Thomas, D. C. (1981). Biological Models and Statistical Interactions: An Example from Multistage Carcinogenesis. *International Journal of Epidemiology*, *10*(4), 383–387. http://doi.org/10.1093/ije/10.4.383

VanderWeele, T. J., & Knol, M. J. (2014). A Tutorial on Interaction. *Epidemiologic Methods*, *3*(1), 33–72. http://doi.org/10.1515/em-2013-0005

Vineis, P. (2018). From John Snow to omics: The long journey of environmental epidemiology. *European Journal of Epidemiology*, *33*(4), 355–363. http://doi.org/10.1007/s10654-018-0398-4

Wagaman, A. S., & Dobrow, R. P. (2021). *Probability: With Applications and R* (1st ed.). Wiley. http://doi.org/10.1002/9781119692430

Wang, X., Mukherjee, B., & Park, S. K. (2018). Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among U.S. Adults in NHANES 2003–2014. *Environment International*, *121*, 683–694. http://doi.org/10.1016/j.envint.2018.09.035

Ward, J. B., Gartner, D. R., Keyes, K. M., Fliss, M. D., McClure, E. S., & Robinson, W. R. (2019). How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies. *Annals of Epidemiology*, *29*, 1–7. http://doi.org/10.1016/j.annepidem.2018.09.007

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics*, *26*(4), 359–372.

Yu, L., Liu, W., Wang, X., Ye, Z., Tan, Q., Qiu, W., … Chen, W. (2022). A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environmental Pollution*, *306*, 119356. http://doi.org/10.1016/j.envpol.2022.119356