Radboud University Nijmegen

Faculty of Social Sciences

A Critical Look at Bayesian Nonparametric Quasi-experimental Design

Thesis BSc Artificial Intelligence

Author:
Rita Matuz
s1010296

Supervisor:
dr. Max Hinne

Second reader:
prof. dr. Tom Heskes

July 2020

Contents

1 Introduction **2**

2 Quasi-experimental designs **2**
2.1 Regression discontinuity design (RDD) . . . 2
2.2 Interrupted time series design (ITS) . . . 4
2.3 Bayesian non-parametric quasi-experimental design (BNQD) . . . 4

3 Assumptions **5**
3.1 Fundamental assumptions . . . 5
3.1.1 Stable unit treatment value assumption (SUTVA) . . . 5
3.1.2 Continuity assumption . . . 5
3.1.3 Functional form assumption . . . 6
3.2 Functional form assumptions of BNQD . . . 8
3.2.1 Same covariance assumption . . . 8
3.2.2 Stationarity assumption . . . 8
3.2.3 Kernel choices . . . 9

4 Simulations **9**
4.1 Simulation data . . . 9
4.1.1 Underlying outcome functions . . . 10
4.1.2 Effect sizes and observation noise levels . . . 11
4.2 Hyperparameter sharing . . . 11
4.2.1 Covariance parameters learned for the different data shapes . . . 11
4.2.2 Effects of hyperparameters sharing between models . . . 13
4.2.3 Discussion . . . 14
4.3 Bandwidths . . . 15
4.3.1 Regression weights in BNQD . . . 15
4.3.2 Effects of the distance of data from the threshold . . . 17
4.3.3 Discarding data . . . 18
4.3.4 Discussion . . . 19
4.4 Kernel properties . . . 20
4.4.1 List of kernels used in simulations . . . 20
4.4.2 Overview comparison between kernels . . . 21
4.4.3 Kernels and statistical power . . . 22
4.4.4 Discussion . . . 25

5 Conclusions **26**

6 References **27**

7 Appendix **28**

1 Introduction

Quasi-experimental designs are a class of research methods that allow for causal inference even though they lack an important component of traditional experimental designs: random assignment to treatment groups. In domains such as social sciences and epidemiology, researchers often face the obstacle that random assignment is infeasible due to ethical concerns or practical limitations. Such circumstances inspired the development of alternative methods, which have been gaining popularity in recent decades (Bärnighausen et al., 2017). Quasi-experimental studies have a strong asset compared to randomized-controlled trials, by avoiding potentially unnatural manipulations they tend to have stronger external validity. Yet, to retain the internal validity, they need to rely on stronger assumptions than randomized-controlled trials.

Hinne, van Gerven & Ambrogioni (2019) have developed a Bayesian non-parametric framework for quasi-experimental designs, named BNQD (Bayesian Non-parametric Quasi-experimental Design). This method provides a flexible framework that can be used in many of the most common quasi-experimental designs and allows for reaping the benefits of Bayesian approaches such as the possibility of providing evidence in favor of the lack of an effect. While the method has great claims, its properties have not yet been studied extensively.

This thesis aims to take a critical look at the BNQD method, assess how strong its assumptions are and evaluate different ways these assumptions might be alleviated. It strives to advise further improvements to the method and to give pointers to researchers on how to use the method without sacrificing the power, internal validity and causal claims of their studies.

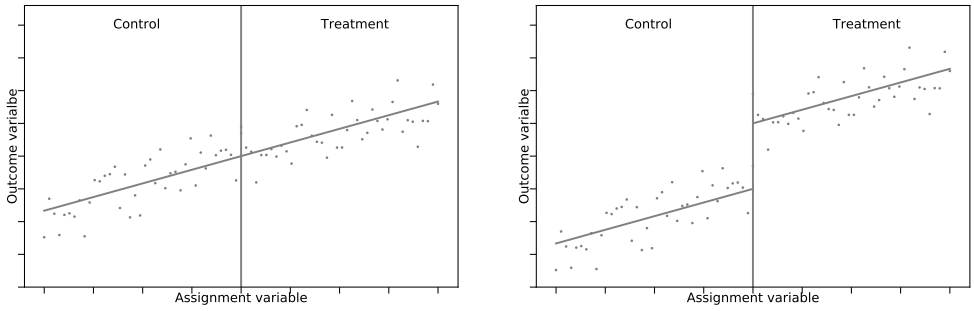
The following section is going to provide an introduction to sharp regression discontinuity and interrupted time series designs, and how these are handled in BNQD. Section 3 will explore the assumptions of these methods, specifically zooming in on the assumptions that are posed by BNQD but are not required other methods for quasi-experimental designs. Section 4 will then through simulations explore the influence of these assumptions and potential ways of alleviating them. Conclusions are drawn in Section 5.

2 Quasi-experimental designs

2.1 Regression discontinuity design (RDD)

Regression discontinuity is one of the most frequently used quasi-experimental designs. It can be used for estimating the causal effect of a treatment on some outcome variable when subjects are assigned to the treated and nontreated groups based on whether their score on some continuous assignment variable reaches a designated threshold. The assignment variable can be any continuous measure such as blood pressure, academic performance or geographic location. Subjects with scores just above or just below the cutoff value are assumed to be similar on all relevant characteristics and therefore all differences in their outcome measures are credited to having received different treatments. Figure 1 shows examples of potential outcomes of regression discontinuity studies.

Regression discontinuity design was first developed by Thistlewaite and Campbell. In 1960 they published the first application of the method, which investigated the effects of winning a national merit scholarship certificate. The method was not immediately widely adopted, however, in recent decades it has been re-discovered and popularized by a number of researchers (Cook, 2008). The revival seems justified since contrary to the initial beliefs of Campbell ("[RDD is] very limited in range of possible applications (...) [and] those limited applications are mainly educational.", Donald T. Campbell, 1963),



(a) No effect **(b)** Effect

Figure 1: Examples of potential outcomes of regression discontinuity studies

RDD has been found to be widely applicable, delivering meaningful results in a broad range of fields, such as economics, political science, psychology and epidemiology. This is partially because decision rules are ubiquitous in everyday life: prescribing medication is often based on whether the patient scores above or below a cutoff score on some physiological measure, governmental measures often apply within clear-cut geographic boundaries, compensatory academic programs are provided to children below a cutoff academic performance score. In many cases, routinely collected administrative data can be used for a regression discontinuity study, providing inexpensive but valuable insights.

Over the history of regression discontinuity design, the methods for analyzing the data from such studies have evolved. To introduce these methods, the following notation will be used: i refers to a particular unit of treatment, most often this is an individual. The score of i on the assignment variable (also referred to as running variable or forcing variable) is denoted by z_i, and the value of the cutoff by c. The binary treatment variable is x_i, having value 1 if unit i was exposed to treatment, and 0 if it was not, this value is fully dependent on z_i. The observed outcome variable is denoted by y_i, which can be either y_{1i}, the outcome if i received the treatment, or y_{0i} otherwise.

The question of interest concerns $y_{1i} - y_{0i}$, that is the difference the treatment makes in the outcome. Given the design there are no units for which we observe both of the potential outcomes, one of these remains a counterfactual and it is hard to reach causal conclusions on the level of the individual unit. In any such between-subjects study designs, one can only aim for finding the average treatment effect (ATE) across the population of interest, that is the mean or expected value of the individual treatment effects for members of the population.

$$ATE = \mathbb{E}[y_{1i} - y_{0i}] = \mathbb{E}[y_{1i}] - \mathbb{E}[y_{0i}] \tag{1}$$

Usually, however, there is no basis to assume that the treatment effect would be uniform across the whole range of z_i. A less ambitious goal of modern regression discontinuity studies is usually to find the average treatment effect at the cutoff point (ATEC).

$$ATEC = \mathbb{E}[y_{1i} - y_{0i} | x_i = c] \tag{2}$$

Under the continuity assumption (discussed in Section 3.1.2), this can be calculated as the differences between the limits towards the cutoff on either side:

$$ATEC = \mathbb{E}[y_{1i} - y_{0i} | x_i = c] = \lim_{z \to c^+} \mathbb{E}[y_i | z_i = c] - \lim_{z \to c^-} \mathbb{E}[y_i | z_i = c] \tag{3}$$

The task is then to fit regression lines to the data on each side of the cutoff value and make an (arbitrarily small) extrapolation towards the limit. A straightforward choice could be to apply linear regression or polynomial regression on both sides. However, the assumption of linearity is unsuitable in many cases, and higher-order polynomial regressions are unable to provide unbiased estimates near the boundaries (Gelman and Imbens, 2018). Since we are most interested in the area around the boundary point, local linear nonparametric regression (LLR) was shown to be a more suitable choice for regression discontinuity design (Hahn et al., 2001, Gelman and Imbens, 2018).

So far we have only looked at the so called sharp regression discontinuity design, where every unit gets treatment according to the decision rule. Not all is lost, however, in the case where this is not true for all units, in the so called fuzzy design, the application of the treatment is probabilistic instead of deterministic given the decision rule. As long as the probability of treatment changes sharply at the cutoff point, we can still find a treatment effect – only this time it is unwarranted to generalize this effect to all of the population. We can instead aim for finding the complier average treatment effect at the cutoff (CATEC).

2.2 Interrupted time series design (ITS)

A special case of regression discontinuity design occurs when the assignment variable is calendar time, this is called interrupted time series design. It is applicable in situations where an intervention happens at a specific point in time, and longitudinal data is collected on the outcome variable from before and after the intervention point. Researchers can then assess the effect of the intervention by looking at the differences of the regression lines at the cutoff timepoint in the same way as treatment effects are evaluated in regression discontinuity design. The nature of the assignment variable makes it somewhat more difficult to meet the assumptions of the available methods, these issues are discussed in Section 3.1.2.

2.3 Bayesian non-parametric quasi-experimental design (BNQD)

The novel framework of BNQD is capable of analyzing more than just regression discontinuity and interrupted time series, but here our discussion is limited to how these designs can be approached using this method as described by Hinne et al. (2019).

In BNQD, the question whether there is a treatment effect, is framed as Bayesian model comparison. Of the two models that are compared, one is the continuous model where a single regression is fit to the entire set of observations, ignoring the existence of the cutoff point. The other model is the discontinuous model, where separate regressions are fit to the treated and non-treated groups. The Bayesian framework allows for finding evidence for either model, making *'the treatment has no effect'* a possible outcome for quasi-experiments. Our belief in the existence of the effect is expressed by the Bayes factor of the comparison between the continuous and discontinuous models. For cases where we would like to incorporate our uncertainty in the existence of the effect, a new effect size measure is provided as well: the Bayesian model averaged effect size.

The models try to capture the latent function that generated the outcomes. While this function is unknown, we have some prior assumptions about the shapes it can take, this is expressed as a Gaussian process (GP) prior. This prior takes two parameters, the mean function $\mu(x)$, which describes the expected mean of the function, and the kernel $k(x, x'; \theta)$ that expresses how different the outcomes are expected to be for two units with given in assignment values.

It is also assumed that the measurement and other noise cause the observations to follow a a Gaussian distribution around the latent function.

$$
\begin{aligned}
f(x) &\sim \mathcal{GP}(\mu(x), k(x, x'; \theta)) \\
y_i &\sim \mathcal{N}(f(x_i), \sigma^2)
\end{aligned}
\tag{4}
$$

The choice for a GP prior is an important modeling decision that can have a large impact on the results of the analysis. The kernel needs to be selected according to the properties of the field of research, expressing the assumptions the modeller has about the underlying processes behind their data. A linear kernel will behave similarly to a linear regression in common RDD practices, but more flexible kernels might make a better choice in situations where the input-output relationships of the data generating process are more complex.

3 Assumptions

3.1 Fundamental assumptions

Like any statistical method, BNQD can only provide correct estimations of causal effects if the analyzed data corresponds to the assumptions the method. Quasi-experiments depend on less strong assumptions than non-experiments, but stronger assumptions than the most commonly preferred methods of true experimentation. The assumptions help exclude potential rival hypotheses that might explain the results differently, obstructing the validity of the causal inference.

This section will provide an overview of the important assumptions of regression discontinuity and interrupted time series designs that have been outlined in previous literature regarding the frequentist methods, and discuss their relations to BNQD.

3.1.1 Stable unit treatment value assumption (SUTVA)

The theory underlying the inference of causal effects in most research is the Rubin causal model or counterfactual model, developed by Donald Rubin and his colleagues (Holland, 1986). It defines the causal effect as a difference between what the outcome is under the treatment and what the outcome would have been, for the same unit under no treatment but otherwise identical circumstances. This definition has a fundamental problem, namely that at least one of these outcomes are not observable, therefore in order to infer the causal effect, we need to have some assumptions in place. Note, that these assumptions are not specific to quasi-experimental designs, all experiments that wish to infer causal effects are subject to them, including randomized-controlled trials.

The assumption of the Rubin causal model is called stable unit treatment value assumption, also known as non-inference assumption. It has two parts: first, the method of assignment to treatment groups (such as random selection or decision thresholds) should have no effect on how a participant responds to the treatment or the lack of thereof. Second, this response should also be independent of what treatments other participants receive. This second part, also known as no spillover effect, is more difficult to ensure as often participants are part of some network of influence.

3.1.2 Continuity assumption

The continuity assumption is the most commonly discussed assumption of quasi-experimental designs. It means that the potential outcome functions are assumed to be continuous at the threshold. Without this assumption, we can never be sure how much discontinuity in the regression lines was caused by the treatment, and how much of it is due to the discontinuity already present in the potential outcome functions.

There are no uniform ways of making sure that the continuity is there, as it depends strongly on the specific nature of the assignment and outcome variables. The goal is to eliminate all potential rival hypotheses by thinking about what else could cause a discontinuity, and reason about why such scenarios are unlikely. For example, in a regression discontinuity study where the decision threshold to administer some medication coincides with a threshold where some physiological change happens in the body, it can be argued that the outcome function would be discontinuous at the decision boundary, even if every patient or none of the patients received the treatment. In the case of interrupted time series designs, the assignment variable is time, which can lead to specific alternative explanations. It could be that some other event co-occurred with the intervention and is causing a discontinuity, there could be changes in the methods or strictness of measurement of the outcome variable to evaluate a policy change, which leads to an unfair comparison of before-and-after. Many interrupted time series studies fail to specify why such alternative hypotheses can be excluded (Ramsay et al., 2003). While excluding all possible rival hypotheses is practically impossible, with thorough knowledge of the field of study and very critical evaluation of the research design we should aim to exclude at least the plausible ones (some pointers for this process can be found in Donald T. Campbell, 1963; Shadish et al., 2002 and Ramsay et al., 2003).

One of the easier ways to make sure assuming continuity is justified, is by proving that there is noise in the measurement of the assignment variable. This noise implies that there is some randomness in treatment assignment around the threshold and thus locally in a very small band around this threshold the conditions of a randomized-controlled trial are met, and estimating a local treatment effect is possible. Noise also makes sure that full manipulation of the assignment variable is not possible, and McCrary (2008) has shown that if there is only partial self-selection, the causal effect can still be inferred. This does not mean that noise in the measurement of the assignment variable is imperative to valid quasi-experimentation, but it can provide a convenient tool for ensuring that the continuity assumption holds.

This assumption is equally indispensable regardless of what method is used for computing regressions in RDD and ITS, and is thus also required for BNQD.

3.1.3 Functional form assumption

The term functional form assumption is used to describe an assortment of different assumptions, the common ground amongst which is that in all interpretations they express some sort of knowledge we claim to have about the shape of the potential outcome functions.

One such potential assumption on the functional form would be to assume that the potential outcome functions are parallel to each other, with their vertical distance being the effect of the treatment. This assumption is met by the outcome functions pictured on Figure 2. In this case, the treatment effect found at the threshold in an RDD study could be generalized to the whole range of the assignment variable and give us the average treatment effect (Equation 1) similarly to a randomized-controlled trial. Most often however, we do not have clear justifications to assume that the treatment would affect all units to the same degree regardless of their score on the assignment variable, and without this assumption we can at most postulate the average treatment effect at the cutoff (Equation 2).

Additional functional form assumptions are also made when regressions are fit to the datapoints, since in order to approximate a function, we need to have at least some prior assumptions about its shape. Our choice of method heavily influences what assumptions we need to make, our discussion here will concern three possible methods: simple regression, local regression and BNQD.

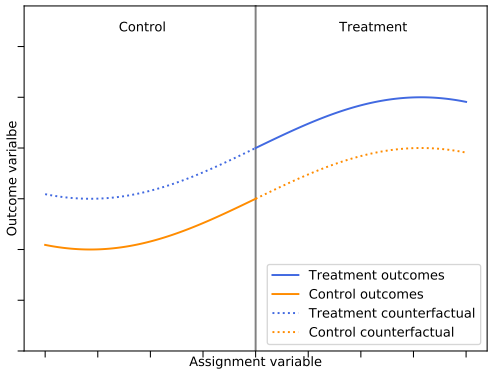


Figure 2: Example of the two potential outcome functions in an RDD study. The blue line indicates the outcome function in case of treatment, the orange line is the outcome function without treatment. Only parts of each function are observed.

If we work with simple (linear or polynomial) regression, we need to specify the form of the outcome functions in the regression equations, therefore this method involves very strong functional form assumptions. It can be argued that real-life relationships are almost never completely linear, and when in order to express more flexible relationships we choose for polynomial regression, we rarely ever have a strong theoretical foundation behind the specific chosen degree of the polynomial (Gelman and Imbens, 2018). The functional form assumptions of the simple regression methods are therefore rather strong and untestable, and in recent decades it is not considered good practice to assume they hold (Bor et al., 2015).

If we accept the idea that anything other than determining the effect at the threshold is too ambitious for regression discontinuity studies, it is possible to also reduce the assumptions on the general shape of the outcome functions. Local (linear or polynomial) regression methods are flexible non-parametric methods that can accommodate non-linear relationships and also allow for assigning more weight to near-threshold points, these methods have been the new standard for RDD in the past decade (Cook, 2008; Imbens and Lemieux, 2008; Lee and Lemieux, 2010). Theoretically, if the only goal is to measure the effect at the threshold and there is enough data close to that threshold, it is possible to calculate the size of the effect without any functional form assumption whatsoever, just by very heavily weighing or only using the data very close to the threshold. In practice, however, data might be sparse and by discarding further away datapoints, statistical power can quickly become an issue. Researchers have to make a choice on the size of the bandwidth for the method, where narrower bandwidths decrease the strength of the necessary functional form assumptions but also decrease the precision of the estimation through not using all of the available data.

There are some differences between the functional form assumptions of the above mentioned methods and those of BNQD. In BNQD, part of the the specification of the functional form comes down to the kernel choice. The properties of the Gaussian process prior we place on our outcome functions are quite explicit representations of what we assume the functions to be like. This generally constitutes a much more relaxed functional form assumption than that of simple regressions, since flexible kernels can fit almost any plausible function adequately well. There are two other assumptions on the functional form of the outcome functions that BNQD requires, namely that the two potential outcome functions are stationary, and share the same covariance parameters.

To examine BNQD in more detail, the following parts will focus on these assumptions that set BNQD apart from existing methods.

3.2 Functional form assumptions of BNQD

3.2.1 Same covariance assumption

The same covariance assumption was first termed and described by Branson et al. (2019), meaning that the two potential outcome functions are assumed to have equal covariance parameters. Note that while Branson et al. also include the stationarity assumption in this term, for notational simplicity purposes here the same covariance assumption is used to refer to only this criteria, separate from the stationarity assumption. The covariance parameters refer to the hyperparameters of the kernel used in the analysis, in the case of the most commonly used squared exponential kernel, these parameters are the *variance* and the *lengthscale*.

Visually inspecting the data from a quasi-experiment, it is rather hard to come to any conclusions over whether this assumption is likely to be violated. This is largely because parts of each outcome function are unobserved counterfactuals (Figure 2 is an example). The observable parts seemingly being characterized by different parameters does not mean that the functions are not similar on the full range, and vica versa.

The same covariance assumption is required mainly for convenience reasons, as this way less parameters need to be fit in the discontinuous model (Branson et al., 2019), and it was proposed that fitting separate parameters in the two sub-models of the discontinuous model would alleviate this assumption. This possibility is explored in Section 4.2.

3.2.2 Stationarity assumption

In mathematics, a stationary process is a stochastic process whose unconditional joint probability distribution does not change when shifted in time (Gagniuc, 2017). Examples of such processes are white noise and colored noise. In machine learning and statistics however, this term tends to be used more leniently and adapted to describe different, but related phenomena. In the Gaussian process literature, the term *stationary* is mostly used in describing properties of kernel: a stationary kernel is one that is invariant to translations in the input space, it is merely a function of the distance between two datapoints (Rasmussen, 2004). Based on that definition, the linear kernel is non-stationary because apart from the relative position of the two datapoints, it also takes into account their absolute location with regards to the origin.

In BNQD, what we assume under the stationarity assumption is that the covariance parameters of the outcome functions do not vary with the assignment variable (Branson et al., 2019). Note that the definition makes this concept of stationarity depend on the covariance function used, as the assumptions are posed on the nature of the parameters of this specific covariance function. Kernel parameters do not always have a clear, intuitive translation into visually discernible features of the process, making the assessment of whether this criterium holds rather intricate. What further complicates this notion is that *a parameter not varying with a variable* is more often than not, a question of scale. Zoom in highly on any two ranges in any arbitrary function, and most likely you will find that they are not best described by the same set of parameters, even if stationarity seems to hold looking at a larger window of data. The only functions where any two equal sized ranges definitely share hyperparameters are functions with no changes in their curvature, essentially straight lines that have a constant first derivative, suggesting that the stationary assumption might in essence equal the assumption of linearity. It is, however, evident that the method is capable of handling highly non-linear, and thus likely non-stationary cases (see the illustrative examples in Hinne et al., 2019).

The extent to which the stationarity assumption is violated with any particular dataset is hard to quantify. While the method is likely robust to small violations, it

would be beneficial if it could be made robust to larger violations as well. If the covariance parameters vary with the assignment variable, in the RDD context our interests lie in their values around the threshold. Following from the idea of stationarity being a question of scale, one approach to make BNQD more robust to this assumption would be to zoom in to the area around the threshold as much as necessary to be left with a sufficiently stationary process, and do the analysis on only this subset of the data. This is a fairly common procedure in the quasi-experimental design literature known as introducing a bandwidth, used to alleviate strong functional form assumptions. This process could only work, however, if there exists a frame with sufficient datapoints around the threshold that can be considered 'stationary enough'. The possibility of introducing a bandwidth in BNQD by discarding datapoints is discussed in section 4.3. Another possible approach would be to place GP priors on the covariance parameters themselves, learn them as a function of the assignment variable and then use the appropriate values from this function that apply to the threshold (Branson et al., 2019). This approach might be overly computationally intense and require more data than we usually have available in a quasi-experiment.

3.2.3 Kernel choices

Our prior beliefs about the shape of the potential outcome functions are expressed in the kernel of the Gaussian process prior, and the choice for this kernel is up to the researcher applying the method. A kernel that is flexible enough to fit just about any potential outcome function is a convenient choice if we do not know anything about its shape, rendering the functional form assumptions to the minimum. If we however, have a good reason to assume certain things about these functions, it is likely beneficial to introduce these assumptions with a fitting prior, to minimize the likelihood that the function is approximated incorrectly. In this sense, BNQD offers great flexibility to make the functional form assumptions as weak or strong as they need to be for each specific purpose.

Making the appropriate kernel choice, however, can be a daunting task for researchers who are not familiar with the workings of Gaussian processes. To assess the magnitude of the impact these choices have on the outcomes of the analysis, this topic is further discussed and addressed with simulations in Section 4.4.

4 Simulations

The goal of this section is to examine the properties of BNQD by exploring some of its aspects that are novel compared to the traditional RDD methods. Namely, Section 4.2 will look at the possibility of hyperparameter sharing and its relation to the same covariance assumption, Section 4.3 will consider the issue of the lack of bandwidths in BNQD and how this relates to the stationarity assumption, and finally Section 4.4 focuses on the kernels, and the differences between them that could inform us in our selections. Before diving into those topics, Section 4.1 discusses how the simulation data used in the remainder of this chapter were created.

4.1 Simulation data

As the first step of generating simulation datasets, the number of samples used in each simulation were equally spaced out to span the range of -10 to 10 of the assignment variable. The outcome value of each sample was generated by passing its assignment value through one of the 8 different underlying functions shown on Figure 3 and discussed in Section 4.1.1, these functions were normalized to have a mean of 0 and standard

deviation (SD) of 10 on the range of −10 to 10. To create a discontinuity with a given underlying effect size, the values on the two sides of the threshold were shifted vertically in the opposite directions. Lastly, a Gaussian noise term was added to each datapoint, with the SD of the Gaussian being a given underlying observation noise.

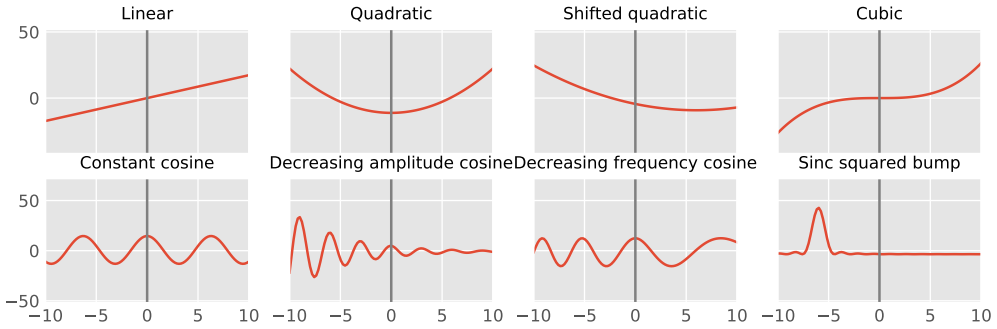


Figure 3: Underlying functions used for generating simulation data

4.1.1 Underlying outcome functions

For none of the resulting datasets can we claim that they have been generated by stationary stochastic processes in the probability theory meaning of the word, due to the underlying trends represented by these 8 functions used to generate outcome values. But it is exactly these latent trend functions that we need to analyze in order to find out whether they are stationary in the way Branson et al. (2019) introduced the term, and thus whether or not the datasets generated through them are appropriate to be used in a BNQD analysis based on the stationarity assumption. Note that even in these simulated examples, we cannot make a claim on whether the shared covariance assumption is violated, since this assumption is connected to the unseen counterfactuals that are also not represented in the simulation data.

The 8 underlying outcome functions used in the simulations are the following:

1. The linear function ($f(x) = x$), a very easily interpretable relationship between two varables. With regards to the covariance parameters of the squared exponential kernel, we can consider this function stationary.

2. The quadratic function ($f(x) = x^2$). Here it might be questionable to claim that the covariance parameters do not vary with x in this function. The two sides of the quadratic function grow fast and no uniform variance or lengthscale is expected to fit the quadratic growth into infinity. Still, if we take only this small range at the middle of this function, it is possible that stationarity is only violated to a small degree.

3. A shifted version of the quadratic function ($f(x) = (x-6)^2$) allows us to examine what happens if we have the same shape as above but the threshold is not laying on the reflection axis and thus if we allow the two sub-models of the discontinuous model to learn separate covariance parameters, they will likely learn distinct ones.

4. The cubic function ($f(x) = x^3$). The same can be said about this function as the quadratic variant, except the threshold is now at the mean of the function instead of its lowest point.

5. A constant cosine function ($f(x) = cos(x)$). While its derivative is not constant, there is not much variation in this function that would suggest a major violation to the stationarity assumption. The entire function can be described well with a single lengthscale and variance, yet being a periodic function it is more complex than linear functions.

6. A cosine with decreasing amplitude ($f(x) = exp(\frac{-x}{5})cos(2\pi\frac{x}{3})$). We can expect the kernel variance to not be constant with respect to x, this dataset is expected to violate the stationarity assumption to a larger extent.

7. A cosine with decreasing amplitude ($f(x) = cos(2\frac{-x}{20}x)$). As the previous function this is a stronger violation to the stationarity assumption, most likely affecting the lengthscale parameter of the squared exponential kernel.

8. A function that is practically constant, except for an unusual 'bump' in the function on one side of the threshold, realized as a shifted, scaled and squared sinc function ($f(x) = (\frac{sin(2x+12)}{2x+12})^2$). While the lengthscale around the threshold would ideally be very high here, the squared exponential kernel is known to fit the lengthscale to the largest spike in the data (Duvenaud, 2014). The spike makes this function nonstationary, although here the area around the threshold remains stationary.

4.1.2 Effect sizes and observation noise levels

Apart from the shape, other essential characteristics of the generated datasets are the sizes of the effect and the observation noise, the simulations in this thesis include multiple combinations of these. Apart from 'no effect', three different effect sizes were used that correspond to the levels defined by Cohen (1988): a small effect that is 20% the size of the standard deviation of the data, a medium effect that is 50%, and a large effect that is 80%. A high noise condition was defined as a condition where the noise is double the effect size, this amount of noise tends to completely mask the discontinuity. Medium noise was created by making the effect size 1.5 times the observation noise, here while the effect is somewhat ambiguous but we would expect our analysis to find it. In the low noise case, where the noise is a third of the effect size, the effects are clearly visible through plotting the data. Figure 4 illustrates these effect size – observation noise combinations on the linear dataset.

4.2 Hyperparameter sharing

This section explores the possibilities of sharing hyperparameters between the submodels of the discontinuous model and its relations to the shared covariance and stationarity assumptions.

4.2.1 Covariance parameters learned for the different data shapes

In Table 1, we can see the learned hyperparameter values that result from running BNQD with the squared exponential kernel on each of the 8 different data shapes described in Section 4.1.1. The values of each parameter are shown for three different cases: the value that the continuous model learns, the value that the discontinuous model learns in the case where the two separate models within the discontinuous model share these parameters, and their two separate values if they do not. This last case represents the the modification that is expected to make the analysis robust to violations of the same covariance assumption.

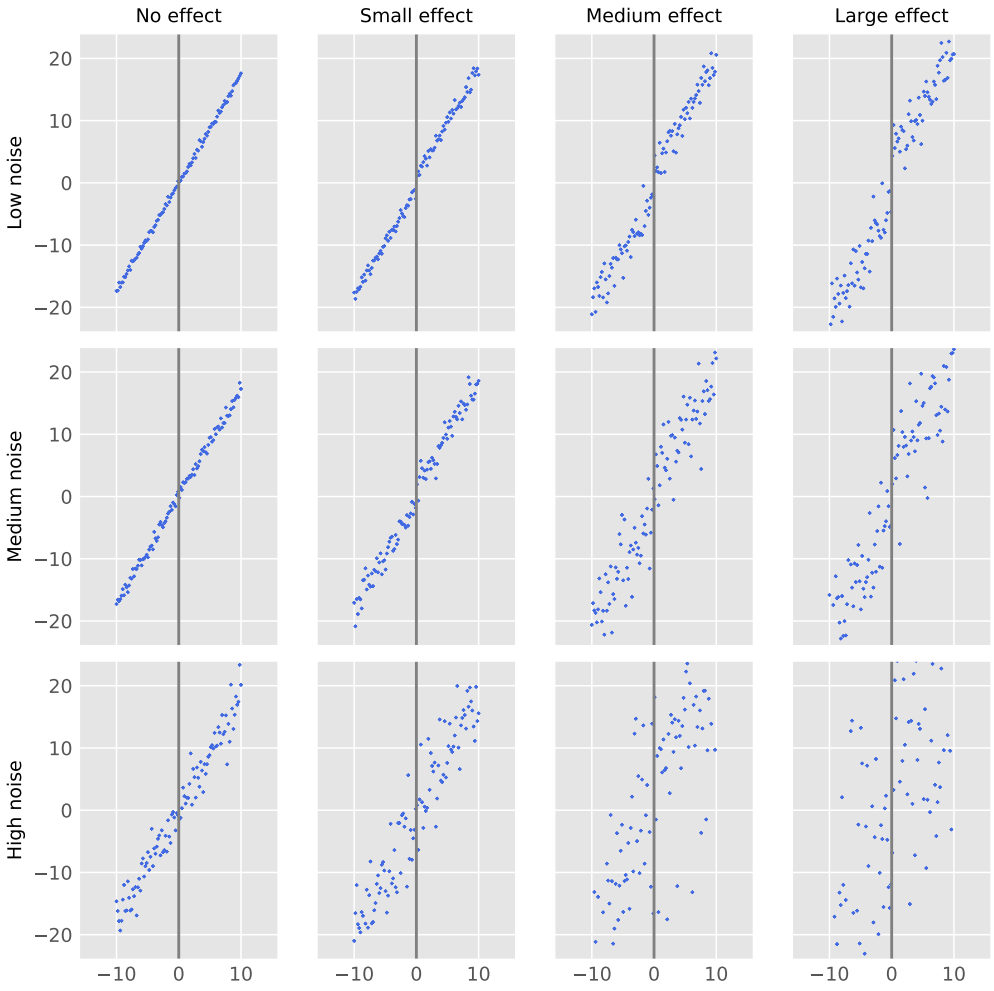


Figure 4: Illustration for the different effect sizes and noise levels used in the simulations

The first hyperparameter, the likelihood variance of the Gaussian process approximates the variance of the noise term, the square of the observation noise level, which in the simulations producing Table 1 is approximately 11.11. The models seem to be able to recover this value fairly accurately, even when only half of the data is available in the case of discontinuous model with separate hyperparameters.

The kernel variance parameter of the squared exponential kernel took a wide range of values between 1.26 and 1492.92. In all simulations, the variance of the data is around 100 units. The learned values of the kernel variance are related to the variance of the data, but not in an definitive manner.

The third hyperparameter is the lengthscale of the squared exponential kernel, varying between values of 0.77 and 629.16. The lower values suggest rapidly changing, fluctuating functions, while higher values imply more constant, slowly changing functions.

In the case of four out of the eight shapes (linear, quadratic, cubic and constant cosine) the two sub-models of the discontinuous model learned virtually the same kernel parameters, this was to be expected as their underlying outcome functions are symmetric with respect to the vertical line at the threshold, or to a point on this line. This also

	Likelihood variance			Kernel variance			Kernel lengthscale		
	Continuous model	Discontinuous model shared parameters	Discontinuous model separate parameters	Continuous model	Discontinuous model shared parameters	Discontinuous model separate parameters	Continuous model	Discontinuous model shared parameters	Discontinuous model separate parameters
Linear	11.62	10.90	10.77 10.91	439.22	324.41	311.61 335.78	14.62	13.17	12.54 13.27
Quadratic	11.39	10.97	10.74 10.98	1469.92	592.16	525.75 643.56	7.41	7.16	6.90 7.18
Shifted quadratic	11.42	10.93	10.75 10.97	243.26	192.81	546.81 22.80	6.73	7.74	9.57 629.16
Cubic	11.28	10.99	10.76 11.01	756.48	585.12	597.96 630.08	5.24	5.97	6.03 6.15
Constant cosine	10.99	10.79	10.55 10.85	172.00	176.88	174.81 181.79	1.71	1.72	1.70 1.73
Decreasing amplitude	10.63	10.54	9.94 10.95	248.06	223.80	592.26 7.95	0.79	0.77	0.84 39.47
Decreasing frequency	10.92	10.82	10.31 10.86	137.19	139.01	147.90 195.91	1.17	1.18	1.02 2.35
Bump	10.68	10.58	10.13 10.76	93.71	90.47	181.98 1.26	0.76	0.75	0.77 404.05

Table 1: Learned hyperparameters of the Gaussian process with squared exponential kernel for the eight different data shapes. Averaged over 100 runs with medium effect size and medium noise.

means that sharing the hyperparameters is not expected to affect the outcomes of the analysis in these cases.

For the other four, non-symmetric shapes the two discontinuous sub-models learned very different kernel parameters. In these cases, and inverse relationship seems to hold between the two covariance parameters: the side of the threshold that gets the higher value for variance tends to get the lower value for lengthscale, and vica versa. This implies that these two hyperparameters are not independent, and that if two sets of data are expected to differ with regards to one of them, they likely differ with regards to the other one too. The results of the BNQD analysis might change if the covariance parameters are allowed to be separate in cases where the separately learned parameters differ greatly from the shared values.

4.2.2 Effects of hyperparameters sharing between models

The results of the BNQD analysis with and without hyperparameter sharing are shown on Figure 5 . The estimated effect sizes averaged over 100 simulations accurately reflect the underlying true effect size in the case of the four datasets whose underlying outcome functions are symmetric, this holds both with and without the sharing of the hyperparameters. In the case of the nonsymmetric data shapes with hyperparameter sharing, the average effect size tends to get underestimated slightly, by about 15%. Without the sharing of the hyperparameters, the effect size estimation becomes highly unpredictable, with the estimate increasing for some and decreasing for other data shapes.

The resulting log Bayes factors are also affected by letting go of the hyperparameter sharing. Positive log Bayes factors suggest evidence for the discontinuous model (there is an effect), while negative ones indicate evidence for the continuous model (no effect). In general, we see an increase in the log Bayes factors for all datasets when we allow for separate hyperparameters, however this increase is much more pronounced for the nonsymmetric shapes, where the effect size estimates also became less accurate.

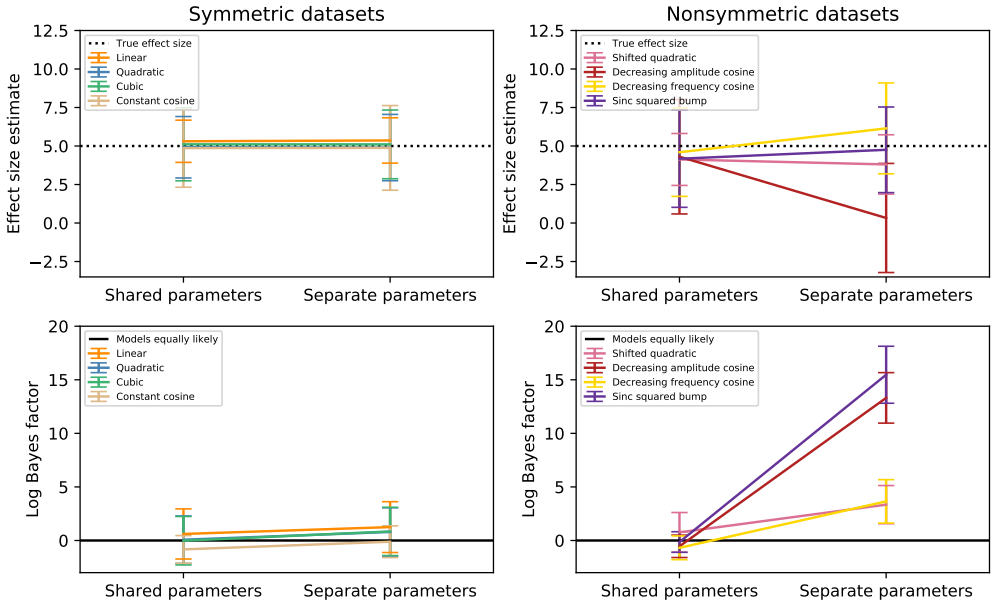


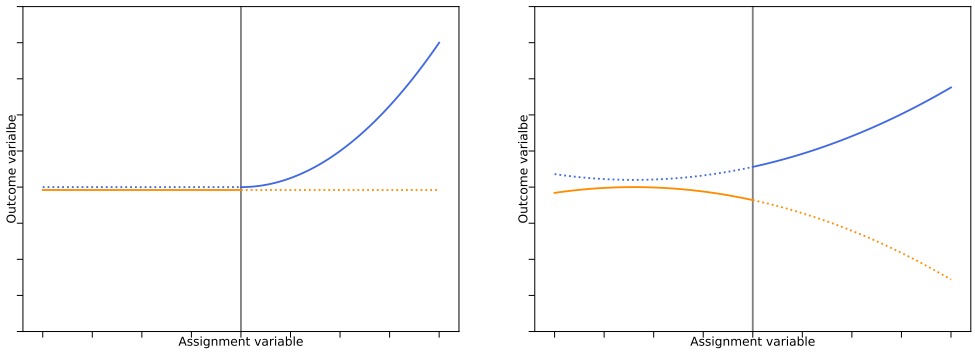
Figure 5: Differences in outcomes between the BNQD analysis run with and without sharing hyperparameters between the two sub-models of the discontinuous model. Averaged over 100 simulations with sample size 100, medium effect size and noise. Error bars indicate one standard deviation from the mean.

4.2.3 Discussion

Separate hyperparameters can help the model become more certain about the presence of an effect, which is likely mediated by better regression fits within the discontinuous model that is then able to dominate over the continuous model in the Bayesian model comparison. The Bayes factor expresses more than just our belief in the existence of a first order discontinuity in the functions, it also incorporates the belief that the functions two sides of the discontinuity have different properties. Since in RDD we are looking for first order discontinuities, we should aim to minimize the effect that other differences make on the Bayes factors, and not allowing the two sides to have different parameters seems to be achieving just that. Coincidentally, the effect size estimates are also more reliable without separating the hyperparameters. The reasons for this latter effect are less clear, but might result from the use of less data leading to less accurate hyperparameter estimations.

The simulations also suggest that BNQD is fairly robust against some violations of the stationarity assumption in datasets where the visible parts of the two underlying functions are nearly symmetric to each other. This is most apparent comparing the quadratic and the shifted quadratic cases, where we can expect stationarity to be violated to a similar degree, yet the non-shifted version allows for more accurate results.

Whether any of the simulated scenarios violate the shared covariance assumption is unclear, because of the unknown nature of the counterfactual parts of the outcome functions. Based on the results, the claim of Branson et al. (2019) that not sharing hyperparameters can lift the shared covariance assumption appears unfounded. Violations to the shared covariance assumption always imply that the two potential outcome functions are not parallel, meaning that there is an interaction between the assignment variable and the treatment. An example is shown on Figure 6a, where the treatment effects are increasingly larger for units scoring higher on the assignment variable. Since



(a) Violating SCA **(b)** Not violating SCA

Figure 6: Examples of potential outcome functions where the treatment effect is not constant with respect to the assignment variable.

this means that the effect size is not constant across the assignment variable, we will not be able to identify more than the average treatment effect at the cutoff (ATEC), this is however the case for RDD designs in general (see Section 3.1.3, and an example on Figure 6b). There are no reasons to believe that the ATEC cannot be accurately estimated with BNQD for datasets where the shared covariance assumption is violated. In these cases however, it is perhaps even more important to share hyperparameters between the two sub-models of the discontinuous model, otherwise the Bayes factors will be more likely to favor the discontinuous model based on the difference in the shapes of the function, even if the ATEC is zero such as on Figure 6a.

Lack of hyperparameter sharing can be an interesting option to explore for possible extensions to BNQD for the detection of discontinuities in the rate of change or in the frequencies a function is composed of. However, in traditional RDD and ITS we are only interested in the first order discontinuities at the threshold, and therefore hyperparameter sharing needs to be kept in the analysis.

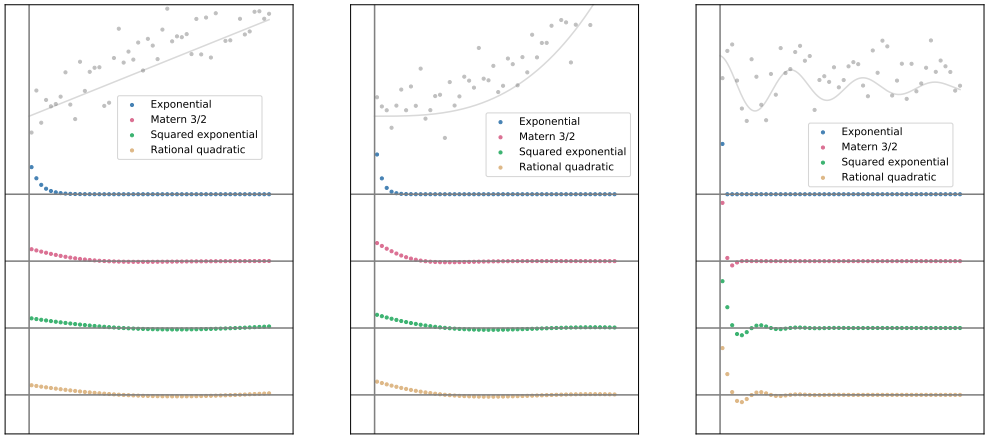
4.3 Bandwidths

To combat functional form assumptions, the most frequently suggested method in the literature of quasi-experimental designs is the use of bandwidths. How a bandwidth is implemented depends on the specifics of the given method used, but generally involves discarding or devaluing data that is far away from the cutoff point.

The goal of this section is to discuss the way and the extent to which this concept is already seamlessly integrated into Gaussian process regression, and then test whether discarding data makes BNQD more robust against the stationarity assumption.

4.3.1 Regression weights in BNQD

When dealing with non-linear data, Gelman and Imbens (2018) argue that fitting higher order polynomials in simple ("global") regression should be avoided. Amongst other reasons, they argue that such regressions tend to place comparatively larger weights on the extreme values furthest away from the threshold, which is the opposite of the ideal behavior expected in RDD. To combat this issue, their recommendation is to use local regression methods (linear or quadratic) with a reasonable bandwidth, as this ensures that more weight is assigned to the values close to the threshold.



(a) Linear data **(b)** Cubic data **(c)** Decreasing amplitude data

Figure 7: The output of the weight function $\mathbf{w}$ at the threshold, corresponding to the inputted datapoints. The inputted datapoints are shown on top in gray. Below, for each kernel the gray horizontal line denotes zero weight, and the vertical distance of the point from this line expresses the relative weight of the datapoint at the given assignment variable value. The hyperparameters used for calculating the weight function are maximum likelihood estimates for the specific dataset at hand.

Gaussian process regression for RDD was presented as an alternative specifically to local regression (Branson et al., 2019), even though the method does not involve explicitly declaring a bandwidth. To predict the value of the underlying function f at the cutoff, it computes a weighted average of the observed values $\mathbf{y}$ as

$$f(c) = k(c, \mathbf{x})^\intercal (k(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I)^{-1} \mathbf{y} \tag{5}$$

where $\mathbf{x}$ denotes the vector of the assignment variable values of all observations on one side of the cutoff point, k implies elementwise application of the kernel function, I is the the identity matrix and σ_n^2 is the likelihood variance hyperparameter of the Gaussian process.

From this, we can derive the weights $\mathbf{w}$ applied to the observations as

$$\mathbf{w}(c) = (k(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I)^{-1} k(c, \mathbf{x}) \tag{6}$$

(Silverman, 1984, Rasmussen, 2004). These weights are thus defined by the covariance function and the hyperparameters of the Gaussian process. Figure 7 shows the weights of each datapoint in simulations, calculated through this weight function for one side of the threshold for three different datasets and four kernels.

Comparing these weights to those of polinomial and local regression demonstrated by Gelman and Imbens (2018), BNQD is in possession of the favorable qualities of local regression. The kernels define the general shape of the weight function, while the hyperparameters of the Gaussian process set how peaked this function is with respect to the estimated point. Parallels can be drawn between the process of estimating discontinuities through BNQD and local regression. First, to estimate the optimal bandwidth for local regression, some method such as the one defined by Imbens and Kalyanaraman (2011) is applied to the full set of data. This is analogous to estimating the set of hyperparameters on the full dataset in BNQD. Then, for both methods, some kernel is selected that defines the shape of the weight function. In local regression for RDD, the

kernel choices do not have a large impact on the analysis (Imbens and Lemieux, 2008), whether the same is true for BNQD is explored in Section 4.4. Lastly, the extrapolated values at the threshold are calculated using the kernel and the learned bandwidth/hyperparameters. BNQD integrates the two-step process of bandwidth estimation and effect size estimation of local regression methods into one method.

While this makes introducing additional bandwidths to BNQD a potentially redundant step, it has been argued that the stationarity assumption of BNQD might be relaxed through discarding datapoints far away from the cutoff point (Branson et al., 2019). While these datapoints are generally not taken into account during the extrapolation step, they influence the selection of hyperparameters, and through that, influence way other datapoints are weighted during the extrapolation. By discarding data and thus zooming in on the near-threshold area, the estimated hyperparameters of a non-stationary underlying function will change to better reflect their values near the threshold. The following section explores whether there is a connection when estimating effect sizes in BNQD, between how stationary a function is and how accurately the near-threshold area is represented in the data. Then, section 4.3.3 explores whether discarding data can help prediction accuracies in BNQD.

4.3.2 Effects of the distance of data from the threshold

Figure 8 shows the accuracy of the effect size predictions (average percentage difference between the predicted and true effect), plotted against how spread out the data is with regards to the assignment variable around the threshold. To create each point on this plot, 100 values of the assignment variable were drawn from a normal distribution with its mean being at the threshold and its standard deviation taking different values between 0 and 10, plotted on the x-axis.

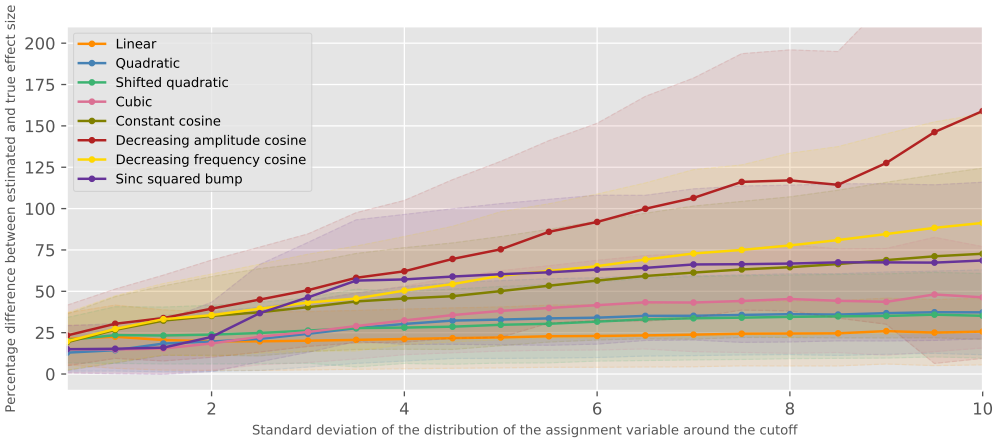


Figure 8: Effect of the spread of the data around the threshold, on the accuracy of the effect size predictions. Run using the squared exponential kernel, medium effect size and medium noise. Averaged over 100 runs, with the shaded areas denoting one standard deviation from the mean.

The results show that the more closely the datapoints center around the threshold, the more accurate the predictions will be. This relationship however, is much more prominent in the case of datasets that are speculated to violate the stationarity assumption the most, especially for the decreasing amplitude and frequency cosines. This implies that the idea of datapoints being close or far away from the cutoff is indeed relevant with regards to stationarity.

In research settings we do not always have the opportunity to collect data in a way that it is less spread out from the threshold. Due to this, a more realistic way of acquiring a dataset with such qualities to increase the prediction accuracy in case of violations to stationarity, might be to simply discard far-away datapoints.

4.3.3 Discarding data

The simulations in Section 4.3.2 have implied that the data focusing less on the threshold can indeed negatively affect the validity of the results when the stationarity assumption is violated, this is the case regardless of the built-in datapoint-weighing properties of Gaussian processes discussed in Section 4.3.1. Branson et al. (2019) recommend further discarding of datapoints to combat this issue.

On Figures 9 and 10 the accuracy of the effect size predictions and the resulting Bayes factors are shown when only specific percentages of the data are retained. The initial arrangement uses 100 datapoints spaced out evenly across the -10 to 10 range of the assignment variable. The discarding of data in the simulation always concerns the given percentage of datapoints furthest away from the threshold, and always results in a decrease in the sample size going into the analysis.

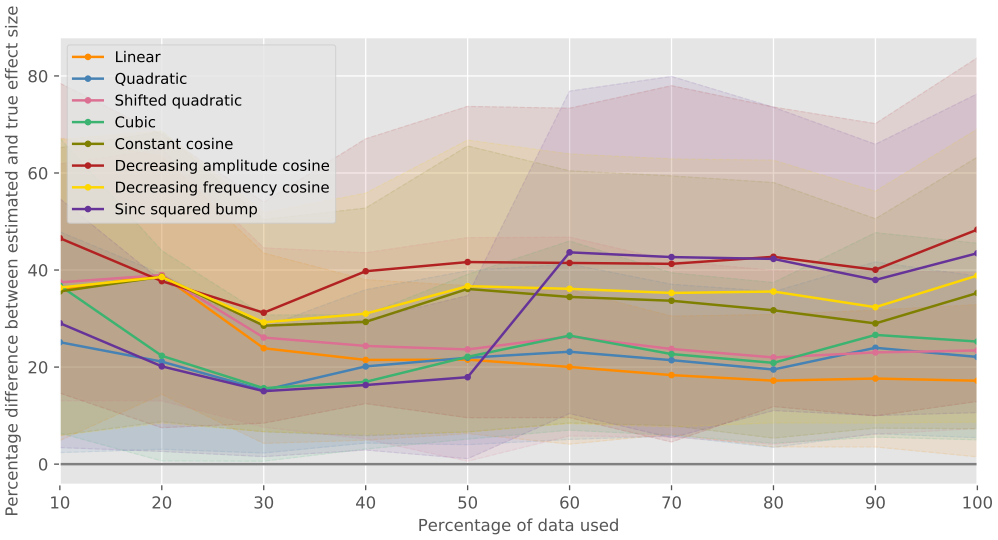


Figure 9: Difference between predicted and true effect size as a percentage of the true effect size, in function of the amount of data left in the analysis. Run using the squared exponential kernel, medium effect size and medium noise. Averaged over 100 runs, with the shaded areas denoting one standard deviation from the mean.

Using less data appears to have almost no effect on how accurately the effect size is estimated and it also does not greatly affect the Bayes factors.

There are two different processes that play into these outcomes. More data allows for more accurate predictions of the underlying function, and thus makes predictions less noisy, yet on the other hand, removing further away datapoints makes the learned hyperparameters more representative of the middle range and can likely allow for better predictions at the threshold. In most cases, these two effects tend to balance each other out completely, but there are two edge cases where we can see an imbalance.

On the one extreme is the most stationary, linear data shape, where the hyperparameters are constant along any window on the whole range of the assignment variable.

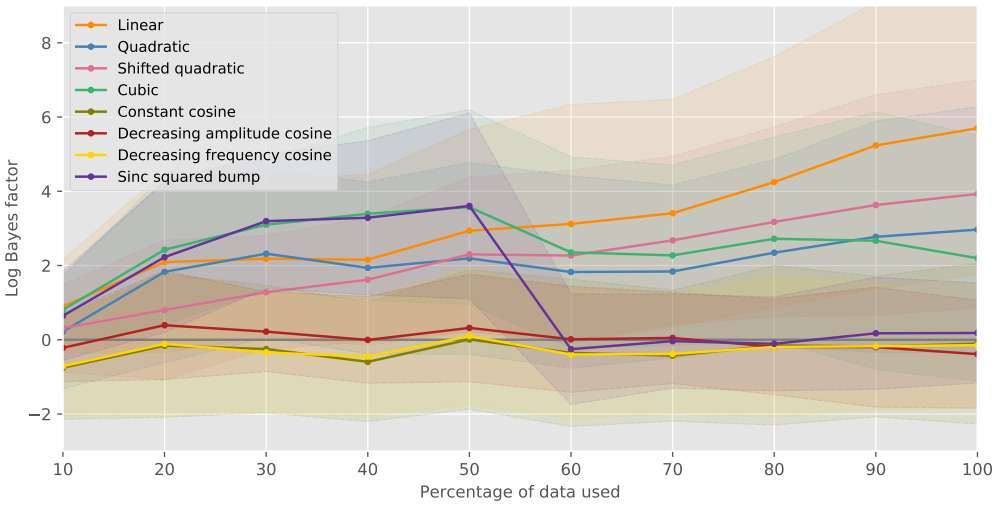


Figure 10: Log Bayes factor, in function of the amount of data left in the analysis. Run using the squared exponential kernel, medium effect size and medium noise. Averaged over 100 runs, with the shaded areas denoting one standard deviation from the mean.

In this case, the discarding of data can never make the hyperparameters any more representative to the near-threshold range. This leads to a subtle but visible decrease in the accuracy of the predictions with less data.

On the other extreme, the 'bump' dataset has a small section of its range which can be described by very different hyperparameters than the range we are interested in, around the threshold. This results in a sudden increase in prediction accuracy when only around 55% of the data is left in the analysis, when the spike in the data gets removed. The Bayes factors also greatly increase at this point.

4.3.4 Discussion

Effect size estimations are the least precise and Bayes factors are the lowest in the cases of the least stationary underlying functions. In the presented simulated example, on average the difference between the estimated and the true effect size is around 40% of the true effect size, but even 70% difference is within one standard deviation over the 100 runs. Discarding far-away datapoints seems to be of help only for one specific dataset, and even there only around a specific bandwidth. This highlights the importance of the type of nonstationarity being dealt with. If there is a small, clearly defined range that causes the nonstationarity, the analysis can benefit from removing the datapoints in this range. If there are good theoretical reasons to assume this nonstationarity does not affect the near-threshold range, this is advised. Including automatic data discarding in BNQD, however, is not deemed necessary, because for most cases it does not improve or even worsens the prediction accuracies, even for highly nonstationary datasets.

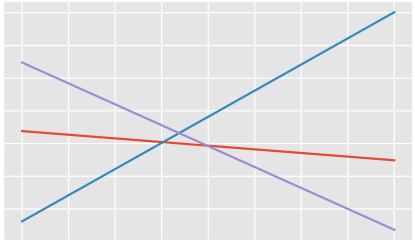
Unfortunately, there does not seem to be an easy fix for the stationarity assumption, however, Section 4.3.1 has also shown that matching the assumptions of BNQD to those of local regression does not yield a fair comparison. This is because BNQD can be a substitute for more than just the regression itself, it also makes another part of the process, bandwidth estimation, unnecessary. Methods for bandwidth estimation place additional assumptions on the underlying functions, see for example Imbens and Kalyanaraman (2011).

4.4 Kernel properties

Selecting an appropriate kernel is the most subjective choice that has to be made when applying BNQD. Being something that is highly dependant on the field and topic of research, there are no clear guidelines on how to make this selection, yet the decision is likely to have an impact on the results. Even in the case of Bayesian model averaging of multiple kernels, these multiple kernels still need to be selected in a sensible manner. The goal of this section is to explore the properties of different kernels, looking for possible indicators and counterindicators to their use.

4.4.1 List of kernels used in simulations

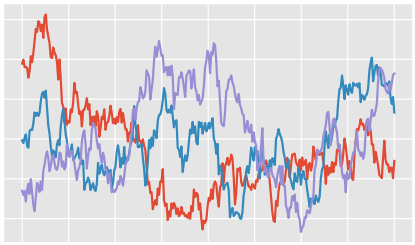
Linear kernel



$$k_L(x, x') = \sigma_0^2 + \sigma^2(x - c)(x' - c)$$

Using this kernel for a Gaussian process equals doing Bayesian linear regression (Duvenaud, 2014). The linear kernel is useful in those, and only those situations where the relationship between assignment and output variables are assumed to be linear, both with and without treatment.

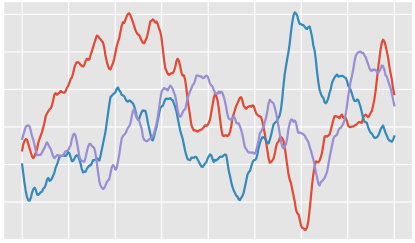
Exponential kernel



$$k_E(x, x') = \sigma^2 exp(\frac{-|x-x'|}{\ell})$$

The exponential kernel is the simplest truly nonparametric kernel option. It produces continuous, but non-differentiable functions.

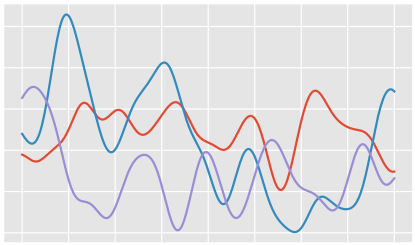
Matérn 3/2 kernel



$$k_M(x, x') = \sigma^2(1 + \frac{\sqrt{3}|x-x'|}{\ell})exp(\frac{-\sqrt{3}|x-x'|}{\ell})$$

The Martén kernel produces functions that are differentiable once, but not infinitely. These functions are therefore smoother than those generated by the exponential kernel, but less smooth than the ones generated by the squared exponential kernel.

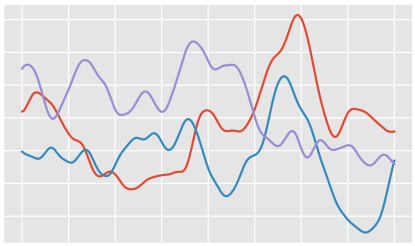
Squared exponential kernel



$$k_{SE}(x, x') = \sigma^2 exp(\frac{-(x-x')^2}{2\ell^2})$$

The squared exponential kernel (also known as Gaussian kernel or radial basis function) allows smooth fits through datapoints, with infinitely differentiable functions. This is a popular choice for cases where the input-output relation is not fully known and is expected to be complex and smooth.

Rational quadratic kernel



$$k_{RQ}(x, x') = \sigma^2(1 + \frac{(x-x')^2}{2\alpha\ell^2})^{-\alpha}$$

The rational quadratic kernel is an infinite sum of squared exponential kernels with different lengthscales. It has an additional parameter compared to the previous kernels (α), which determines what proportions of the function have smaller or larger lengthscales.

4.4.2 Overview comparison between kernels

Figure 11 shows the effect size estimates resulting from running BNQD with the five different kernels on three different data shapes. A more comprehensive report of the simulations, including results for all different noise levels on all data shapes are given in the Appendix.

The results show that most kernels are able to accurately recover the effect size in low and medium noise conditions, albeit they get unpredictable in high noise conditions. On linear data, the less smooth kernels tend to overestimate the effect, while on the shifted quadratic datasets all kernels tend to underestimate it. While the linear kernel is unsuited for most datasets and is therefore unreliable, the differences between the other four kernels are not substantial in most cases. The squared exponential and rational quadratic kernels are particularly similar, yielding the same results in most cases.

The same information is shown about the Bayes factor predictions on Figure 12. For no effect, every kernel manages to conclude negative results in all noise conditions (apart from the linear kernel applied on data that is not linear). The linear kernel yields the highest Bayes factors, followed by the squared exponential kernel. For most datasets,

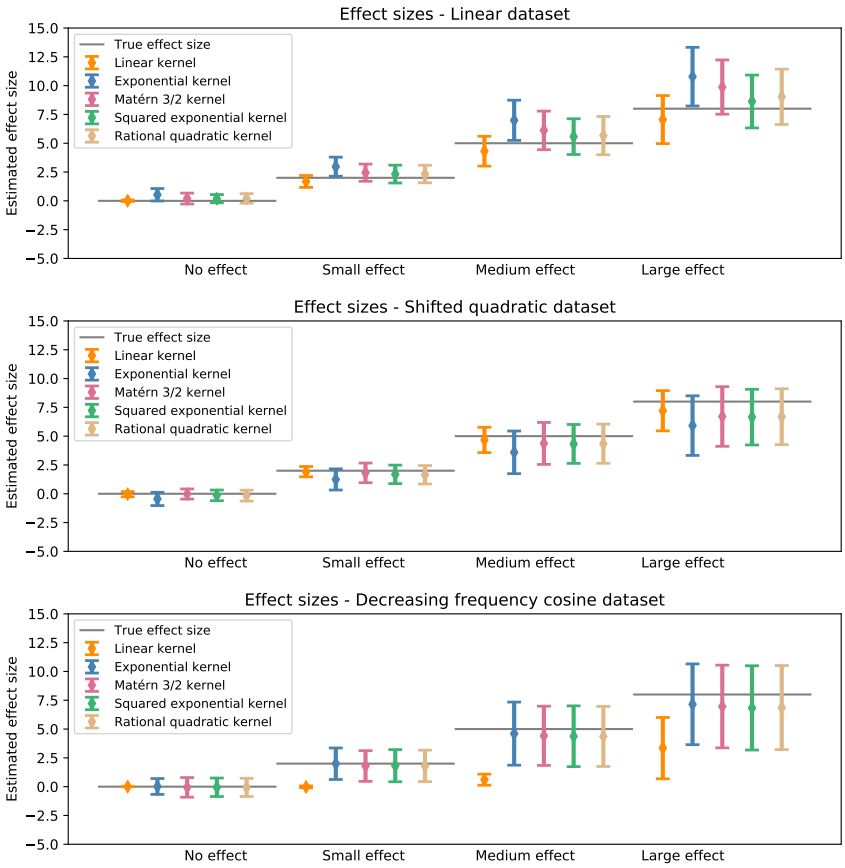


Figure 11: Estimated effect size outcomes per kernel and latent effect size, for three different underlying functions. Averaged over 30 runs with sample size of 100 and medium noise, error bars indicate one standard deviation.

the small effect goes unnoticed, but medium and high effects result in positive log Bayes factors under low or medium noise. The Bayes factors tend to be higher where the latent function is more stationary.

4.4.3 Kernels and statistical power

When the use of different kernels results in different outcomes, it may be possible that the analyses that were less certain about the effect were lacking statistical power. In Section 4.3.1 we have seen that different kernels have different "weight curves" that are related to the concept of bandwidths, which have been shown to have large effects on the statistical power of RDD studies (Deke, 2012). The better extrapolation a Gaussian process allows for, the more certain we can be in our results, and kernel choices might have an impact on this. Additionally, more complex kernels tend to have more parameters to fit to the data, and generally the more parameters that need to be fitted mean larger sample sizes required. This section explores whether there are statistical power differences associated with kernel choices.

The question is complex, as there are complications around measuring statistical power both in quasi-experimental designs and in the Bayesian statistical framework. A highly cited early analysis by Goldberger (1972) showed that compared to a randomized-controlled trial, regression discontinuity design needs 2.75 times more data to be able

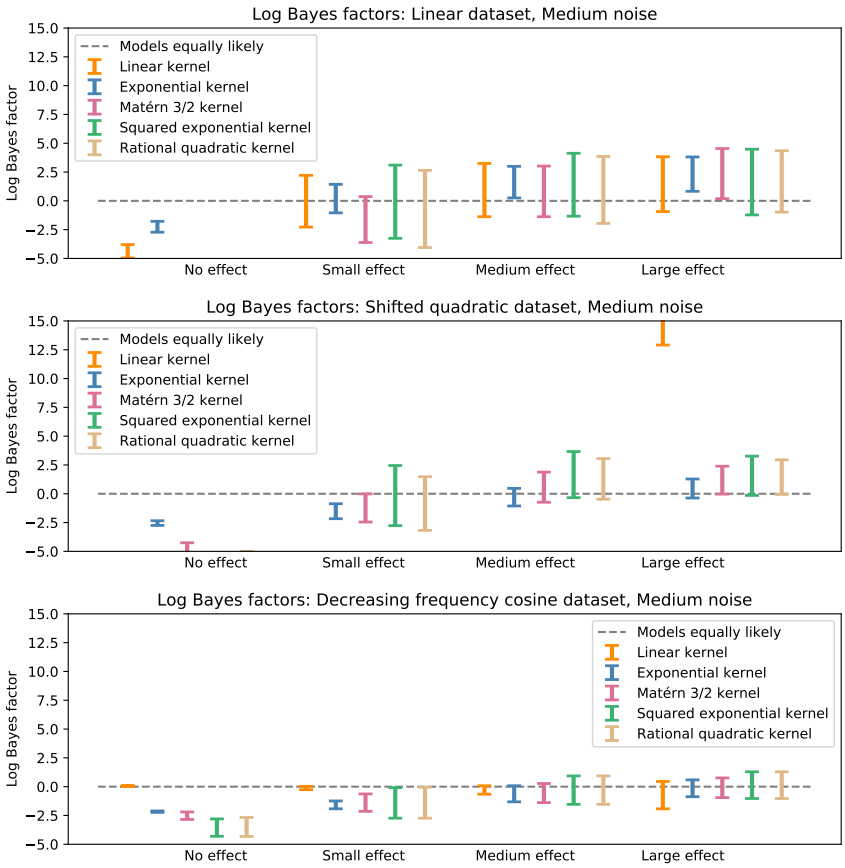


Figure 12: Bayes factor outcomes, per kernel and latent effect size, for three different underlying functions. Averaged over 30 runs with sample size of 100 and medium noise, error bars indicate one standard deviation.

to provide significant results. This number however, heavily depended on the exact properties of the data chosen for the analysis. The fact that in RDD we concentrate on the effect only at a very specific threshold means that statistical power is no longer a simple function of sample size. Instead, adding an extra datapoint might have a different effect based on how far from the cutoff it is, at which side of the cutoff it is added (and whether the two sides are in imbalance), what method and bandwidth is used, whether a clustered design is used, and a number of other factors (Cappelleri et al., 1994; Schochet, 2009; Deke and Dragoset, 2012).

Taking this question to the Bayesian framework further complicates the scene. Generally speaking, statistical power as defined in frequentist statistics does not go well with the Bayesian point of view, because this latter is concerned about degrees of belief rather than binary decisions. Yet, as long as researchers are using BNQD to answer the questions *"is there an effect to this treatment?"* and *"how big is this effect?"*, Bayesian or not, they remain vulnerable to the type II errors that power analyses are supposed to prevent. Due to this, it can still be interesting too look at some measures of statistical power. In randomized-controlled trials, one such Bayesian measure is to look at the width of the 95% highest density interval (HDI) around our outcome measure (Kruschke, 2014). As the width of this interval shrinks, we become more certain about the value of the estimated variable. We can then choose a desired width that we are

hoping to achieve and compare different sample sizes based on how likely they are to yield intervals within this width.

There are two issues around using this methodology to assess BNQD. One of them follows from the issues outlined above stating that in RDD statistical power is not a function of sample size only, and the other being that in BNQD, the effect size measure is the difference of the two Gaussians that are the predicted distributions of the outcome functions at the threshold. The width of this interval is therefore always going to include the uncertainty in the values extrapolated at the cutoff.

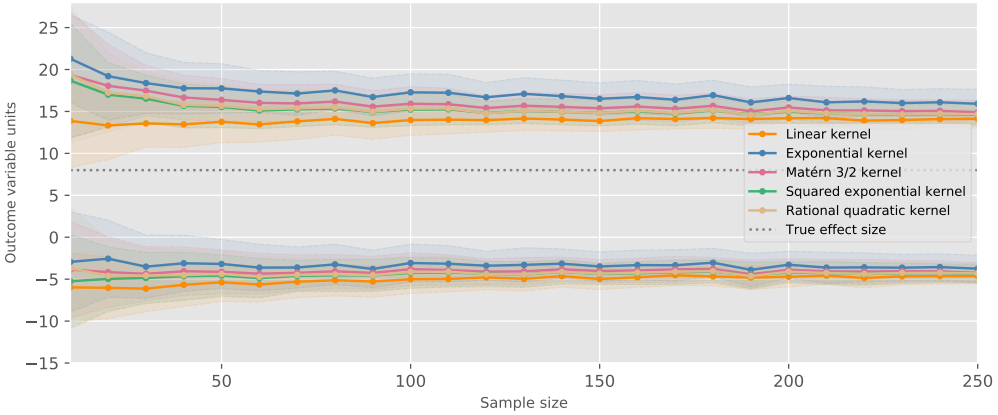


Figure 13: Lower and upper bounds of the 95% HDI around the effect size, as a function of the sample size. Averaged over 100 runs with medium effect size and medium noise on linear data. Shaded areas show one standard deviation.

Figure 13 illustrates the behavior of the 95% HDI around the effect size in function of the sample size. What is shown is that the width of the HDI does not get smaller with more samples, the analysis just becomes more certain in that it is uncertain; that is the variances in the upper and lower bounds become narrower but the HDI itself does not.

Since the width of this interval stems from the Gaussian process, it can be informative to look at how its different hyperparameters play a part in determining this width, this is shown on Figure 14. Each of the hyperparameters were manipulated separately with the other two hyperparameters held constant at the values that the model originally learned for them.

The kernel variance is the hyperparameter that can take the most extreme values in training (see Table 1), but this variable has practically no effect on the width of the HDI.

The kernel lengthscale has a very large effect on the HDI, but only in the cases where its value is small, especially when it is less than 1. Out of the 8 testing functions, this was only the case for highly nonstationary periodic functions (Table 1). At higher values, the lengthscale does not affect the width of the HDI anymore.

The hyperparameter that is affecting the HDI the most is the likelihood variance, which corresponds to the square of the observation noise. This is the only hyperparameter which affects the HDI width at its entire range. In this simulation, the width of the 95% HDI is approximately 6 times the square root of the likelihood variance, that is 3 times the observation noise.

While sample size does not fully account for statistical power, kernels differ in the way they are affected by the sample size. Figure 14 shows the resulting Bayes factors in function of the sample size, these Bayes factors are still visibly increasing even over 200

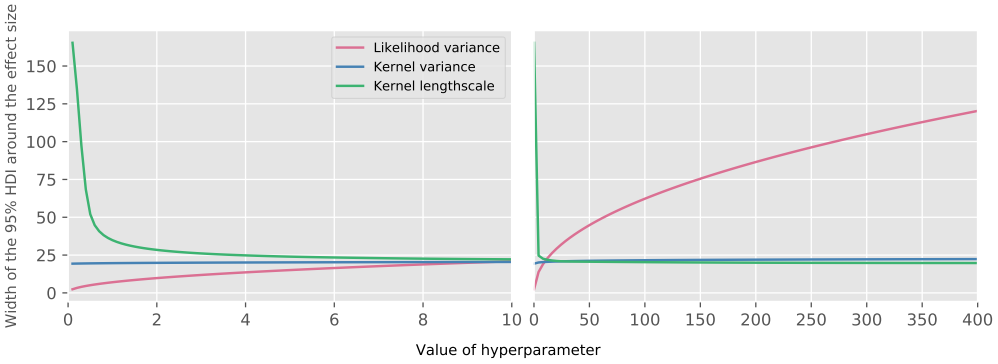


Figure 14: Width of the 95% HDI around the effect size, as a function of the hyperparameters of the Gaussian process with squared exponential kernel. The two plots demonstrate the same functions on two different scales on the x-axis.

datapoints. The rate of increase differs per kernel, the linear kernel is the most affected by sample size while the exponential kernel is the least.

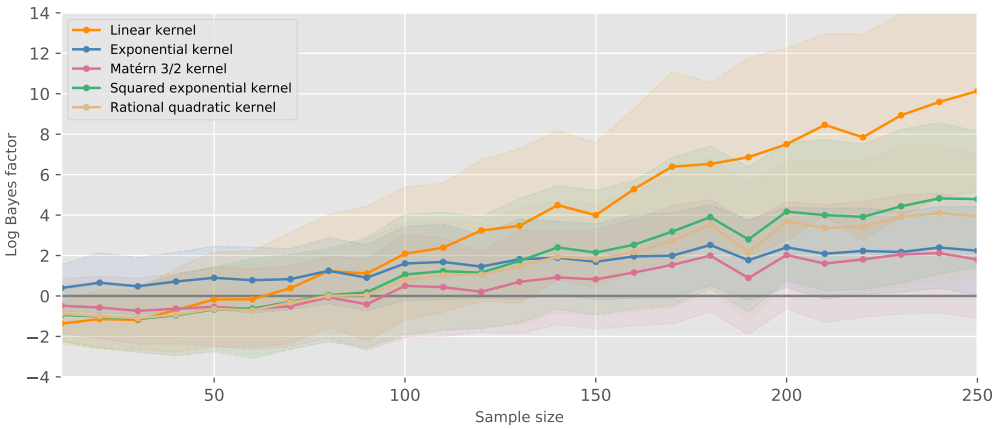


Figure 15: Bayes factor as a function of sample size. Averaged over 100 runs, with the shaded areas showing one standard deviation around this mean. The kernels are contrasted on a linear dataset with medium effect size and medium observation noise.

Figure 16 shows the predicted effect sizes in function of the sample size. The estimates get into the reasonable range around the 30 samples mark, meaning 15 datapoints on each side of the cutoff. The predictions become more accurate even above the 100 samples range, but in that range the improvements are subtle and the error in the prediction is not converging to zero fast, meaning that there is some level of error in the prediction that we cannot get rid of by adding more data. On this linear dataset, the exponential kernel is the least accurate in predicting the effect size, while the other four kernels are similar in their accuracies.

4.4.4 Discussion

Contrasting some of the most commonly known covariance functions, the differences between them seem minor. This is good news for those wishing to apply BNQD, meaning that no strong functional form assumptions are being made when selecting one of the

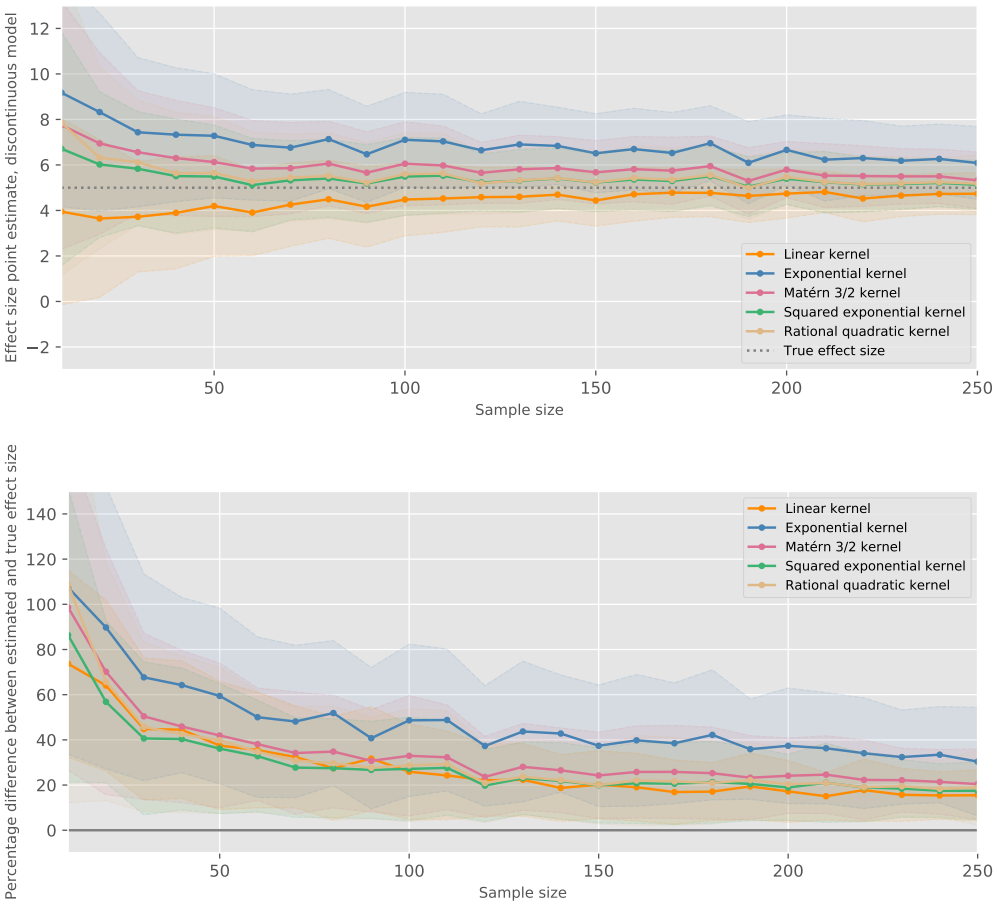


Figure 16: Effects size point estimate as a function of sample size and the percentage error in prediction. The kernels are contrasted on a linear dataset with medium effect size and medium observation noise. Averaged over 100 runs, with the shaded areas showing one standard deviation around this mean.

flexible kernel options. While it is important to make a distinction whether we are expecting a linear or a more curved function, the exact details about the smoothness and differentiability of this function are not crucial.

5 Conclusions

BNQD differs in its assumptions from traditional methods for regression discontinuity in three areas: functional form assumptions through kernel choices, the assumption of stationarity and the assumption of shared covariance parameters between the outcome functions. Based on the simulations presented in this thesis, stationarity is the only assumption out of these three that is somewhat strong. Provided that one of the flexible kernels is selected, BNQD is robust against different kernel choices; and while the shared covariance assumption is a strong assumption, it can be disregarded given that regression discontinuity research is not interested in the treatment effect at any value of the assignment variable other than the threshold.

The method is also robust against small violations to the stationarity assumption. If the outcome functions are near-symmetric with respect to the threshold, even non-stationary signals allow for accurate estimations, and if the nonstationarity is not present in the near-threshold area, discarding datapoints in the nonstationary range can improve predictions.

The stationarity assumption can pose issues, because it is very hard to assess the degree of violation, even on simulated datasets. Whether the covariance parameters vary with the assignment variable depends on which intervals are being looked at. The concept of stationarity is also dependent the kernel choice, meaning that some outcome functions might be stationary with respect to some kernels but less so with respect to other kernels.

Researchers who are investigating whether it is sensible to assume stationarity in their research need to have a high level of intuitive grasp on what the covariance parameters of their selected kernels mean, and how they behave with respect to properties of functions. While the concept of an assumption by nature entails that it is not fully testable, the application of BNQD could be made simpler if future work in this area were able to build metrics that can to detect if such violations are likely.

It is probable that the same datasets that were challenging for BNQD could also only be imprecisely estimated through other methods for regression discontinuity design, investigating this was, however, out of scope for this thesis.

Overall, BNQD provides a valuable new tool for quasi-experimental research, and through the simulations presented here, was found to be robust towards small violations to its assumptions.

6 References

Bärnighausen, T., Røttingen, J.-A., Rockers, P., Shemilt, I., & Tugwell, P. (2017). Quasi-experimental study designs series—paper 1: Introduction: Two historical lineages. *Journal of Clinical Epidemiology*, *89*, 4–11. https://doi.org/10.1016/j.jclinepi.2017.02.020

Bor, J., Moscoe, E., & Bärnighausen, T. (2015). Three approaches to causal inference in regression discontinuity designs. *Epidemiology*, *26*(2), e28–e30. https://doi.org/10.1097/ede.0000000000000256

Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, *202*, 14–30. https://doi.org/10.1016/j.jspi.2019.01.003

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review*, *18*(2), 141–152. https://doi.org/10.1177/0193841x9401800202

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Taylor & Francis. https://books.google.nl/books?id=2v9zDAsLvA0C

Cook, T. D. (2008). waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654. https://doi.org/10.1016/j.jeconom.2007.05.002

Deke, J., & Dragoset, L. (2012). *Statistical power for regression discontinuity designs in education: Empirical estimates of design effects relative to randomized controlled trials* (Mathematica Policy Research Reports). Mathematica Policy Research. https://EconPapers.repec.org/RePEc:mpr:mprres:a4f1d03eb7bf427a8983d4736a4f8f10

Donald T. Campbell, J. C. S. (1963). *Experimental and quasi-experimental designs for research*. Cengage Learning. https://www.xarg.org/ref/a/0395307872/

Duvenaud, D. (2014). Automatic model construction with gaussian processes. https://doi.org/10.17863/CAM.14087

Gagniuc, P. (2017). *Markov chains: From theory to implementation and experimentation.* Wiley. https://books.google.nl/books?id=oNYtDwAAQBAJ

Gelman, A., & Imbens, G. (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations, In *Modelling and evaluating treatment effects in econometrics.* Emerald (MCB UP). https://doi.org/10.1016/s0731-9053(07)00001-1

Hahn, J., Todd, P., & Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209. https://doi.org/10.1111/1468-0262.00183

Hinne, M., van Gerven, M. A. J., & Ambrogioni, L. (2019). Causal inference using bayesian non-parametric quasi-experimental design.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959. https://doi.org/10.1093/restud/rdr043

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan.* Academic Press. https://www.xarg.org/ref/a/0124058884/

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005

Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003). Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care*, *19*(4), 613–623. https://doi.org/10.1017/s0266462303000576

Rasmussen, C. E. (2004). Gaussian processes in machine learning, In *Advanced lectures on machine learning*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*(2), 238–266. https://doi.org/10.3102/1076998609332748

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, *12*(3), 898–916. http://www.jstor.org/stable/2240968

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. https://doi.org/10.1037/h0044319

7 Appendix

Data shape	Kernel	Effect size & Noise levels											
		No effect (0)			Small effect (2)			Medium effect (5)			Large effect (8)		
		Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
Linear	Linear	0.0 ± 0.03	0.01 ± 0.07	0.03 ± 0.2	1.91 ± 0.23	1.68 ± 0.52	0.98 ± 1.21	4.8 ± 0.58	4.31 ± 1.29	2.98 ± 3.35	7.71 ± 0.94	7.05 ± 2.08	6.52 ± 7.3
	Exponential	0.36 ± 0.35	0.53 ± 0.54	1.36 ± 1.15	2.55 ± 0.51	2.97 ± 0.83	4.3 ± 2.01	6.15 ± 0.99	6.99 ± 1.74	9.2 ± 4.06	9.67 ± 1.46	10.78 ± 2.55	9.96 ± 8.22
	Matérn 3/2	0.08 ± 0.23	0.2 ± 0.47	0.55 ± 1.09	2.27 ± 0.43	2.45 ± 0.74	3.39 ± 1.94	5.6 ± 0.95	6.12 ± 1.67	7.66 ± 3.75	8.88 ± 1.39	9.88 ± 2.35	8.37 ± 6.92
	Squared exponential	0.09 ± 0.2	0.19 ± 0.36	0.54 ± 1.02	2.2 ± 0.38	2.32 ± 0.78	2.81 ± 1.82	5.36 ± 0.85	5.57 ± 1.55	6.48 ± 3.85	8.4 ± 1.27	8.63 ± 2.29	8.56 ± 6.6
	Rational quadratic	0.1 ± 0.2	0.21 ± 0.41	0.5 ± 1.09	2.21 ± 0.38	2.33 ± 0.76	2.89 ± 1.93	5.33 ± 0.91	5.66 ± 1.66	10.74 ± 7.43	8.47 ± 1.37	9.03 ± 2.4	10.59 ± 10.77
Quadratic	Linear	-0.01 ± 0.11	-0.02 ± 0.22	-0.07 ± 0.67	1.97 ± 0.22	1.95 ± 0.45	1.85 ± 1.33	4.93 ± 0.56	4.86 ± 1.11	4.5 ± 3.18	7.88 ± 0.89	7.73 ± 1.77	6.81 ± 4.61
	Exponential	0.01 ± 0.35	0.02 ± 0.7	-0.03 ± 1.38	2.01 ± 0.7	1.96 ± 1.07	2.05 ± 2.31	4.94 ± 1.22	4.97 ± 2.01	4.33 ± 4.46	7.92 ± 1.7	7.97 ± 2.87	4.7 ± 6.17
	Matérn 3/2	-0.03 ± 0.27	-0.0 ± 0.52	0.17 ± 1.32	1.98 ± 0.47	2.08 ± 0.84	2.34 ± 2.37	5.12 ± 1.12	5.25 ± 2.04	5.02 ± 4.72	8.18 ± 1.69	8.28 ± 2.97	3.87 ± 6.51
	Squared exponential	0.01 ± 0.28	0.05 ± 0.53	0.2 ± 1.35	2.02 ± 0.54	2.06 ± 0.98	2.32 ± 2.39	5.14 ± 1.16	5.23 ± 2.08	4.88 ± 4.71	8.17 ± 1.74	8.03 ± 3.01	3.04 ± 6.05
	Rational quadratic	0.02 ± 0.27	0.04 ± 0.52	0.14 ± 1.35	2.04 ± 0.52	2.08 ± 0.96	2.25 ± 2.37	5.14 ± 1.16	5.17 ± 2.06	5.1 ± 4.24	8.18 ± 1.73	8.14 ± 2.97	5.43 ± 5.57
Shifted quadratic	Linear	-0.02 ± 0.11	-0.04 ± 0.22	-0.12 ± 0.67	1.96 ± 0.22	1.92 ± 0.45	1.64 ± 1.3	4.88 ± 0.56	4.67 ± 1.1	2.82 ± 2.8	7.74 ± 0.89	7.21 ± 1.75	2.34 ± 3.65
	Exponential	-0.31 ± 0.35	-0.45 ± 0.58	-0.96 ± 1.26	1.55 ± 0.58	1.24 ± 0.92	0.5 ± 2.12	4.1 ± 1.09	3.59 ± 1.85	2.22 ± 3.84	6.72 ± 1.56	5.91 ± 2.59	2.01 ± 4.55
	Matérn 3/2	-0.04 ± 0.25	-0.02 ± 0.44	-0.28 ± 1.21	1.89 ± 0.46	1.81 ± 0.85	1.33 ± 2.08	4.73 ± 1.04	4.37 ± 1.82	2.49 ± 3.59	7.46 ± 1.54	6.71 ± 2.59	2.1 ± 4.37
	Squared exponential	-0.06 ± 0.25	-0.14 ± 0.46	-0.38 ± 1.11	1.85 ± 0.47	1.68 ± 0.8	1.41 ± 1.91	4.59 ± 0.97	4.33 ± 1.69	2.71 ± 3.69	7.36 ± 1.44	6.65 ± 2.42	0.66 ± 2.15
	Rational quadratic	-0.07 ± 0.26	-0.16 ± 0.47	-0.43 ± 1.11	1.83 ± 0.47	1.65 ± 0.8	1.42 ± 1.93	4.54 ± 0.97	4.35 ± 1.7	0.37 ± 5.34	7.29 ± 1.43	6.69 ± 2.43	-0.09 ± 4.17
Cubic	Linear	-10.51 ± 0.12	-10.51 ± 0.24	-10.4 ± 0.71	-8.44 ± 0.24	-8.39 ± 0.48	-7.78 ± 1.51	-5.11 ± 0.64	-4.61 ± 1.34	-2.02 ± 2.8	-1.18 ± 1.07	-0.89 ± 1.3	-0.06 ± 1.55
	Exponential	0.01 ± 0.35	0.01 ± 0.69	-0.01 ± 1.34	2.01 ± 0.68	1.96 ± 1.02	2.15 ± 2.27	4.96 ± 1.18	5.06 ± 1.98	5.5 ± 4.39	8.0 ± 1.67	8.18 ± 2.84	6.61 ± 6.68
	Matérn 3/2	0.03 ± 0.28	-0.05 ± 0.54	0.04 ± 1.33	1.94 ± 0.54	2.0 ± 0.93	2.09 ± 2.42	5.0 ± 1.13	4.91 ± 2.1	4.7 ± 4.75	8.01 ± 1.72	7.96 ± 3.06	4.14 ± 6.08
	Squared exponential	0.0 ± 0.27	-0.08 ± 0.53	0.09 ± 1.58	1.92 ± 0.53	1.93 ± 1.08	2.33 ± 2.73	5.0 ± 1.36	5.41 ± 2.44	5.2 ± 5.27	8.26 ± 2.08	8.53 ± 3.56	2.96 ± 5.77
	Rational quadratic	-0.01 ± 0.28	-0.07 ± 0.52	0.1 ± 1.52	1.98 ± 0.53	1.96 ± 1.04	2.44 ± 2.63	5.04 ± 1.27	5.35 ± 2.33	9.49 ± 6.1	8.22 ± 1.97	8.55 ± 3.42	8.51 ± 9.39
Constant cosine	Linear	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.35 ± 0.46	2.13 ± 1.0	1.89 ± 2.27	5.84 ± 0.73	5.48 ± 1.5	3.43 ± 3.82
	Exponential	0.01 ± 0.35	0.02 ± 0.69	0.05 ± 2.05	2.0 ± 0.69	2.0 ± 1.38	1.79 ± 2.91	4.95 ± 1.72	4.7 ± 2.67	3.56 ± 5.01	7.7 ± 2.42	7.26 ± 3.49	4.05 ± 6.8
	Matérn 3/2	-0.01 ± 0.5	-0.06 ± 0.74	-0.12 ± 1.71	1.93 ± 0.74	1.85 ± 1.25	1.81 ± 2.85	4.78 ± 1.49	4.67 ± 2.52	2.97 ± 5.0	7.61 ± 2.15	7.32 ± 3.54	4.19 ± 6.08
	Squared exponential	-0.06 ± 0.39	-0.09 ± 0.68	-0.04 ± 1.6	1.89 ± 0.68	1.87 ± 1.17	2.01 ± 2.76	4.8 ± 1.39	4.81 ± 2.41	3.92 ± 5.17	7.69 ± 2.03	7.55 ± 3.48	1.91 ± 5.24
	Rational quadratic	-0.06 ± 0.39	-0.09 ± 0.68	-0.04 ± 1.6	1.89 ± 0.68	1.87 ± 1.17	2.01 ± 2.76	4.8 ± 1.39	4.81 ± 2.41	4.26 ± 5.13	7.69 ± 2.03	7.55 ± 3.48	4.18 ± 5.49
Decreasing amplitude cosine	Linear	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.48 ± 0.47	1.41 ± 0.88	1.38 ± 1.53	5.45 ± 0.55	5.26 ± 1.13	3.6 ± 3.28
	Exponential	-0.2 ± 0.33	-0.2 ± 0.66	-0.16 ± 1.95	1.7 ± 0.66	1.7 ± 1.31	1.65 ± 3.68	4.52 ± 1.64	4.39 ± 3.15	2.93 ± 5.63	7.27 ± 2.59	6.7 ± 4.64	2.42 ± 5.86
	Matérn 3/2	0.06 ± 0.72	0.15 ± 1.35	-0.23 ± 2.39	2.13 ± 1.35	2.06 ± 2.17	1.28 ± 3.44	4.8 ± 2.41	4.14 ± 3.15	2.13 ± 4.68	7.18 ± 2.81	6.15 ± 4.04	1.51 ± 4.16
	Squared exponential	0.02 ± 0.56	-0.03 ± 0.97	-0.34 ± 2.29	1.96 ± 0.97	1.8 ± 1.68	1.03 ± 3.41	4.62 ± 2.0	3.99 ± 3.2	2.22 ± 5.11	7.1 ± 2.84	5.92 ± 4.04	3.55 ± 6.44
	Rational quadratic	0.02 ± 0.56	-0.03 ± 0.97	-0.34 ± 2.3	1.96 ± 0.97	1.8 ± 1.68	1.02 ± 3.39	4.62 ± 2.01	3.99 ± 3.2	2.44 ± 5.06	7.09 ± 2.83	5.92 ± 4.04	4.83 ± 5.71
Decreasing frequency cosine	Linear	-0.0 ± 0.0	-0.0 ± 0.0	-0.27 ± 0.68	0.0 ± 0.0	-0.02 ± 0.08	-0.07 ± 0.74	0.73 ± 0.25	0.61 ± 0.48	1.31 ± 1.72	2.58 ± 0.66	3.34 ± 2.66	3.46 ± 4.01
	Exponential	0.01 ± 0.34	0.02 ± 0.69	0.05 ± 2.03	1.99 ± 0.69	1.99 ± 1.37	1.74 ± 2.95	4.92 ± 1.71	4.6 ± 2.74	4.06 ± 4.93	7.64 ± 2.49	7.15 ± 3.51	3.77 ± 6.43
	Matérn 3/2	0.05 ± 0.63	-0.06 ± 0.85	-0.22 ± 1.77	1.92 ± 0.85	1.79 ± 1.33	1.59 ± 2.89	4.66 ± 1.56	4.41 ± 2.57	3.08 ± 4.67	7.37 ± 2.21	6.96 ± 3.59	2.48 ± 5.62
	Squared exponential	-0.01 ± 0.45	-0.05 ± 0.8	-0.17 ± 1.89	1.92 ± 0.8	1.82 ± 1.4	1.64 ± 3.04	4.68 ± 1.66	4.38 ± 2.64	3.92 ± 5.11	7.32 ± 2.29	6.84 ± 3.66	2.12 ± 4.94
	Rational quadratic	-0.02 ± 0.44	-0.06 ± 0.79	-0.21 ± 1.81	1.85 ± 0.78	1.8 ± 1.37	1.56 ± 2.92	4.65 ± 1.61	4.36 ± 2.61	4.11 ± 5.18	7.3 ± 2.25	6.87 ± 3.65	4.95 ± 5.49
Bump	Linear	-1.7 ± 0.22	-1.71 ± 0.44	-2.07 ± 1.42	-0.0 ± 0.0	-0.04 ± 0.12	-0.57 ± 0.93	-0.0 ± 0.0	-0.0 ± 0.0	-0.14 ± 0.33	0.0 ± 0.0	0.0 ± 0.02	0.16 ± 0.65
	Exponential	-0.11 ± 0.34	-0.1 ± 0.68	-0.07 ± 2.0	1.84 ± 0.68	1.84 ± 1.35	1.61 ± 3.17	4.72 ± 1.68	4.45 ± 2.94	3.38 ± 4.81	7.41 ± 2.57	6.59 ± 3.62	2.57 ± 5.45
	Matérn 3/2	0.07 ± 0.68	0.02 ± 1.15	-0.28 ± 1.91	2.01 ± 1.14	1.76 ± 1.51	1.36 ± 2.9	4.56 ± 1.72	4.13 ± 2.6	3.03 ± 4.82	7.05 ± 2.28	6.28 ± 3.45	1.57 ± 4.61
	Squared exponential	-0.01 ± 0.52	-0.09 ± 0.88	-0.38 ± 2.03	1.87 ± 0.88	1.65 ± 1.5	1.17 ± 3.15	4.38 ± 1.76	3.89 ± 2.82	1.85 ± 4.12	6.78 ± 2.45	5.88 ± 3.61	0.99 ± 3.71
	Rational quadratic	-0.01 ± 0.52	-0.09 ± 0.88	-0.44 ± 1.96	1.86 ± 0.88	1.63 ± 1.47	1.1 ± 3.07	4.34 ± 1.72	3.83 ± 2.74	3.15 ± 4.91	6.72 ± 2.38	5.93 ± 3.64	2.88 ± 5.66

Table 2: Estimated effect size outcomes, per data shape, kernel, underlying effect size and observation noise. Averaged over 30 runs with sample size 100.

Data shape	Kernel	Effect size & Noise levels											
		No effect (0)			Small effect (2)			Medium effect (5)			Large effect (8)		
		Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
Linear	Linear	-5.12 ± 0.48	-4.41 ± 0.5	-3.31 ± 0.5	13.61 ± 4.42	0.38 ± 2.54	-2.34 ± 0.78	14.63 ± 4.45	1.37 ± 2.6	-1.44 ± 0.87	15.18 ± 4.48	1.89 ± 2.66	-1.13 ± 0.93
	Exponential	-2.81 ± 0.41	-2.36 ± 0.36	-1.51 ± 0.43	0.93 ± 1.57	-0.25 ± 1.13	-0.4 ± 0.82	3.38 ± 1.96	1.26 ± 1.43	0.43 ± 0.99	4.8 ± 2.18	2.05 ± 1.59	0.56 ± 12.03
	Matérn 3/2	-9.97 ± 0.84	-7.57 ± 0.86	-4.27 ± 0.74	1.97 ± 3.31	-2.27 ± 2.02	-2.16 ± 1.17	4.8 ± 3.25	0.45 ± 2.26	-0.52 ± 1.17	6.21 ± 3.25	1.76 ± 2.23	2.51 ± 11.42
	Squared exponential	-8.97 ± 1.24	-7.22 ± 1.11	-4.35 ± 1.21	8.04 ± 4.41	-0.21 ± 3.08	-2.76 ± 1.11	9.17 ± 4.25	1.15 ± 2.54	-1.1 ± 1.13	9.41 ± 4.02	1.66 ± 2.51	0.96 ± 10.44
	Rational quadratic	-8.96 ± 1.14	-7.18 ± 1.17	-4.44 ± 1.15	7.04 ± 4.03	-0.98 ± 3.02	-2.77 ± 1.19	7.83 ± 3.78	0.88 ± 2.56	-5.82 ± 10.48	8.2 ± 3.51	1.68 ± 2.51	-12.05 ± 30.43
Quadratic	Linear	121.76 ± 1.2	119.64 ± 2.35	101.82 ± 5.52	120.12 ± 2.34	112.52 ± 4.23	71.08 ± 6.63	109.18 ± 4.99	82.26 ± 6.4	23.77 ± 5.36	94.2 ± 6.11	57.38 ± 6.4	9.54 ± 3.71
	Exponential	-3.13 ± 0.08	-2.74 ± 0.16	-1.96 ± 0.17	-1.4 ± 0.95	-1.74 ± 0.65	-1.3 ± 0.4	0.26 ± 1.32	-0.63 ± 0.87	-2.09 ± 5.95	1.38 ± 1.53	0.04 ± 0.98	1.14 ± 5.48
	Matérn 3/2	-5.85 ± 0.86	-4.72 ± 0.74	-2.97 ± 0.6	-0.91 ± 1.92	-2.33 ± 1.2	-1.66 ± 0.65	2.03 ± 2.34	-0.37 ± 1.42	-0.39 ± 4.95	3.35 ± 2.45	0.44 ± 1.48	0.35 ± 8.09
	Squared exponential	-10.23 ± 2.69	-7.8 ± 2.41	-3.98 ± 1.99	4.19 ± 4.64	-1.67 ± 2.53	-1.87 ± 1.46	5.97 ± 3.92	0.44 ± 2.26	0.14 ± 4.15	6.23 ± 3.52	1.02 ± 2.06	1.37 ± 6.4
	Rational quadratic	-10.09 ± 2.09	-7.62 ± 1.95	-4.15 ± 1.28	3.57 ± 3.86	-1.77 ± 2.36	-2.08 ± 1.24	5.01 ± 3.3	0.2 ± 2.15	-1.19 ± 7.77	5.47 ± 3.18	0.83 ± 2.02	-1.25 ± 7.18
Shifted quadratic	Linear	120.76 ± 2.92	109.54 ± 5.02	60.01 ± 6.68	110.14 ± 4.96	82.9 ± 6.5	24.63 ± 5.4	73.79 ± 6.5	34.75 ± 5.73	2.99 ± 2.49	50.81 ± 6.1	17.58 ± 4.57	0.57 ± 1.48
	Exponential	-2.92 ± 0.29	-2.46 ± 0.29	-1.65 ± 0.28	-1.34 ± 0.96	-1.71 ± 0.55	-1.17 ± 0.23	0.56 ± 1.36	-0.52 ± 0.78	-0.55 ± 4.45	1.93 ± 1.57	0.24 ± 0.89	0.67 ± 2.82
	Matérn 3/2	-5.88 ± 0.75	-4.82 ± 0.71	-3.0 ± 0.64	0.63 ± 2.28	-1.68 ± 1.35	-1.58 ± 0.58	3.54 ± 2.41	0.23 ± 1.43	-2.25 ± 7.25	4.65 ± 2.35	0.98 ± 1.36	0.23 ± 3.51
	Squared exponential	-8.52 ± 1.39	-6.59 ± 1.22	-3.77 ± 0.88	5.5 ± 3.95	-0.74 ± 2.21	-1.52 ± 0.95	6.79 ± 3.32	1.13 ± 1.93	-1.46 ± 6.35	6.96 ± 3.05	1.48 ± 1.74	-0.61 ± 1.91
	Rational quadratic	-8.68 ± 1.76	-6.76 ± 1.52	-3.91 ± 1.11	4.48 ± 3.28	-0.89 ± 2.18	-1.58 ± 0.97	5.82 ± 3.0	1.02 ± 1.87	-1.42 ± 7.85	6.11 ± 2.76	1.46 ± 1.69	0.49 ± 1.97
Cubic	Linear	19.56 ± 0.63	18.85 ± 1.21	13.65 ± 2.75	11.81 ± 1.04	10.12 ± 1.84	2.92 ± 2.5	1.52 ± 1.41	-0.33 ± 1.63	-1.61 ± 0.7	-2.73 ± 0.56	-2.39 ± 0.54	-1.5 ± 0.49
	Exponential	-3.13 ± 0.09	-2.7 ± 0.19	-1.91 ± 0.16	-1.32 ± 0.98	-1.69 ± 0.61	-1.26 ± 0.38	0.36 ± 1.31	-0.57 ± 0.87	-0.45 ± 0.48	1.54 ± 1.53	0.14 ± 1.02	2.28 ± 7.52
	Matérn 3/2	-6.01 ± 0.63	-4.73 ± 0.54	-2.92 ± 0.48	-1.4 ± 2.02	-2.45 ± 1.17	-1.68 ± 0.61	1.51 ± 2.28	-0.72 ± 1.4	-0.57 ± 0.68	2.92 ± 2.46	0.2 ± 1.52	0.44 ± 10.44
	Squared exponential	-8.91 ± 1.81	-6.44 ± 1.92	-3.51 ± 1.47	2.24 ± 4.67	-1.45 ± 3.06	-1.59 ± 1.43	4.18 ± 3.98	0.35 ± 2.64	-0.43 ± 1.01	5.22 ± 3.85	1.09 ± 2.34	1.95 ± 9.55
	Rational quadratic	-9.04 ± 1.74	-6.64 ± 1.69	-3.67 ± 1.15	1.28 ± 3.54	-1.97 ± 2.25	-1.53 ± 1.12	3.45 ± 3.44	0.09 ± 2.23	-2.73 ± 7.6	4.44 ± 3.32	0.75 ± 1.93	-7.74 ± 15.79
Constant cosine	Linear	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.01	0.01 ± 0.02	0.02 ± 0.04	0.07 ± 0.15	0.28 ± 0.34	0.35 ± 0.47	-0.04 ± 0.46	0.94 ± 0.36	0.69 ± 0.6	0.28 ± 0.72
	Exponential	-2.56 ± 0.04	-2.45 ± 0.1	-1.91 ± 0.3	-2.02 ± 0.31	-1.89 ± 0.43	-1.38 ± 0.41	-0.7 ± 0.94	-1.08 ± 0.61	-0.66 ± 7.09	-0.04 ± 1.24	-0.6 ± 0.69	-0.69 ± 3.97
	Matérn 3/2	-3.15 ± 0.49	-2.6 ± 0.47	-1.81 ± 0.39	-1.68 ± 2.35	-1.7 ± 0.78	-1.32 ± 0.51	-0.09 ± 1.44	-0.91 ± 0.88	1.9 ± 11.62	0.56 ± 1.56	-0.49 ± 0.93	0.99 ± 6.64
	Squared exponential	-6.47 ± 1.28	-4.29 ± 1.1	-2.19 ± 0.77	-1.17 ± 2.4	-1.98 ± 1.46	-1.33 ± 0.69	0.73 ± 2.3	-0.73 ± 1.29	-0.53 ± 9.92	1.44 ± 2.25	-0.26 ± 1.19	-0.26 ± 6.78
	Rational quadratic	-6.41 ± 1.26	-4.22 ± 1.11	-2.19 ± 0.77	-1.14 ± 2.37	-1.98 ± 1.47	-1.33 ± 0.69	0.73 ± 2.3	-0.73 ± 1.29	-0.33 ± 10.8	1.44 ± 2.25	-0.35 ± 1.2	-2.89 ± 9.23
Decreasing amplitude cosine	Linear	-0.01 ± 0.01	-0.01 ± 0.01	-0.03 ± 0.04	0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	0.15 ± 0.08	0.14 ± 0.12	0.06 ± 0.15	1.3 ± 0.43	1.19 ± 0.55	0.2 ± 0.52
	Exponential	-0.93 ± 0.02	-0.91 ± 0.03	-0.75 ± 0.12	-0.85 ± 0.06	-0.79 ± 0.12	-0.46 ± 0.25	-0.44 ± 0.3	-0.35 ± 0.38	-0.2 ± 0.29	0.05 ± 0.56	-0.12 ± 0.44	-0.17 ± 1.69
	Matérn 3/2	-3.73 ± 0.37	-2.35 ± 0.57	-1.1 ± 0.44	-1.48 ± 1.34	-0.8 ± 1.18	-0.44 ± 0.85	0.1 ± 2.12	-0.48 ± 0.51	-1.12 ± 4.76	0.22 ± 1.2	-0.19 ± 0.51	0.04 ± 2.2
	Squared exponential	-5.37 ± 1.11	-3.83 ± 1.02	-1.68 ± 0.69	-2.3 ± 1.69	-2.09 ± 1.01	-0.69 ± 0.41	-0.46 ± 1.61	-0.58 ± 0.79	1.91 ± 6.81	0.17 ± 1.34	-0.27 ± 0.57	-0.35 ± 3.72
	Rational quadratic	-5.39 ± 1.12	-3.87 ± 1.01	-1.67 ± 0.69	-2.32 ± 1.71	-2.1 ± 1.01	-0.72 ± 0.33	-0.46 ± 1.6	-0.6 ± 0.77	0.3 ± 2.6	0.18 ± 1.31	-0.28 ± 0.58	-0.86 ± 2.02
Decreasing frequency cosine	Linear	0.06 ± 0.02	0.05 ± 0.05	0.02 ± 0.14	-0.12 ± 0.07	-0.13 ± 0.13	-0.17 ± 0.53	-0.27 ± 0.19	-0.28 ± 0.34	-0.53 ± 0.65	-0.12 ± 0.36	-0.53 ± 0.96	-0.3 ± 0.91
	Exponential	-2.25 ± 0.04	-2.16 ± 0.1	-1.64 ± 0.31	-1.81 ± 0.25	-1.66 ± 0.38	-1.14 ± 0.38	-0.6 ± 0.83	-0.94 ± 0.72	-0.49 ± 0.44	0.15 ± 1.07	-0.43 ± 0.65	-0.21 ± 3.52
	Matérn 3/2	-4.67 ± 2.74	-2.55 ± 0.32	-1.67 ± 0.38	-1.48 ± 1.1	-1.66 ± 0.62	-1.15 ± 0.47	-0.34 ± 1.19	-0.86 ± 0.73	0.34 ± 11.28	0.28 ± 1.29	-0.44 ± 0.79	-0.56 ± 2.96
	Squared exponential	-4.7 ± 1.21	-3.4 ± 1.02	-1.75 ± 0.73	-1.25 ± 2.02	-1.73 ± 1.16	-1.11 ± 0.67	0.23 ± 1.85	-0.74 ± 1.02	-0.15 ± 9.0	0.85 ± 1.81	-0.28 ± 0.98	1.02 ± 6.28
	Rational quadratic	-4.63 ± 1.09	-3.39 ± 0.97	-1.78 ± 0.74	-1.44 ± 1.87	-1.78 ± 1.1	-1.11 ± 0.68	0.16 ± 1.81	-0.75 ± 1.02	-0.05 ± 7.44	0.76 ± 1.79	-0.32 ± 0.98	-0.2 ± 6.78
Bump	Linear	-0.94 ± 0.02	-0.94 ± 0.04	-0.91 ± 0.2	-0.74 ± 0.03	-0.73 ± 0.06	-0.59 ± 0.17	-0.15 ± 0.08	-0.12 ± 0.12	-0.11 ± 0.34	-0.0 ± 0.0	-0.0 ± 0.0	-0.03 ± 0.13
	Exponential	-1.26 ± 0.02	-1.21 ± 0.04	-0.87 ± 0.14	-0.98 ± 0.16	-0.86 ± 0.26	-0.48 ± 0.26	-0.05 ± 0.64	-0.23 ± 0.54	0.13 ± 1.68	0.59 ± 0.93	0.05 ± 0.66	-0.69 ± 2.15
	Matérn 3/2	-2.48 ± 0.49	-1.24 ± 1.46	-0.98 ± 0.23	-0.67 ± 1.41	-1.02 ± 0.55	-0.52 ± 0.29	-0.11 ± 1.38	-0.29 ± 0.59	-1.31 ± 4.41	0.59 ± 1.11	-0.01 ± 0.64	0.76 ± 3.41
	Squared exponential	-3.3 ± 0.89	-2.19 ± 0.77	-0.92 ± 0.51	-0.93 ± 1.42	-1.02 ± 0.8	-0.45 ± 0.39	0.26 ± 1.32	-0.24 ± 0.72	-2.09 ± 8.46	0.71 ± 1.31	0.01 ± 0.73	0.35 ± 3.98
	Rational quadratic	-3.3 ± 0.89	-2.2 ± 0.76	-0.91 ± 0.5	-0.98 ± 1.35	-1.08 ± 0.72	-0.47 ± 0.36	0.18 ± 1.24	-0.26 ± 0.67	1.58 ± 5.63	0.67 ± 1.25	0.0 ± 0.68	0.08 ± 3.11

Table 3: Bayes factor outcomes, per data shape, kernel, underlying effect size and observation noise. Averaged over 30 runs with sample size 100.