Forthcoming in the European Journal for Philsophy of Science

The official version is slightly different!

Simplicity and Model Selection

Guillaume Rochefort-Maranda

October 28, 2015

Contents

1 Introduction

In this paper I compare parametric and nonparametric regression models with the help of a simulated data set. Doing so, I have two main objectives. The first one is to differentiate five concepts of simplicity and assess their respective importance. The second one is to show that the scope of the existing philosophical literature on simplicity and model selection is too narrow because it does not take the nonparametric approach into account (Sober 2002; Forster and Sober 1994; Forster 2001, 2007; Hitchcock and Sober 2004; Mikkelson 2006; Baker 2013).

More precisely, I point out that a measure of simplicity in terms of the number of adjustable parameters is inadequate to characterise nonparametric models and to compare them with parametric models. This allows me to weed out false claims about what makes a model simpler than another.

Furthermore, I show that the importance of simplicity in model selection cannot be captured by the notion of parametric simplicity. 'Simplicity' is an umbrella term. While parametric simplicity can be ignored, there are other notions of simplicity that need to be taken into consideration when we choose a model. Such notions are not discussed in the previously mentioned literature. The latter therefore portrays an incomplete picture of why simplicity matters when we choose a model. Overall I support a pluralist view according to which we cannot give a general and interesting (epistemic or pragmatic) justification for the importance of simplicity in science.

This paper contains two main sections. In the first section, I construct a data set and explain how we can choose a regression model with an additive error term and a linear smoother by using a parametric (polynomial regression) and a nonparametric approach (kernel regression). This allows me to discuss five different concepts of simplicity in the second section. The R codes to recreate the results of the analyses are included in the annex.

2 Selecting a Model

2.1 Constructing the Data Set

In order to construct a data set, I first define the following function $f(x)$:

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-6)^2}{2}} + \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-4)^2}{2}} + \frac{1}{\sqrt{2\pi}}e^{\frac{-(x-8)^2}{2}}$$

Then, I simulate 200 observations of $f(x)$ that are uniformly distributed on the interval $[0, 12]$. For each observation, I add noise that follows a normal distribution with $\mu = 0$ and $\sigma = 0.2$. The resulting couples (x, y) can be visualised in Figure 1. The small blue dots represent the observations and the red line represents $f(x)$.

Those 200 observations now constitute a data set and the scientific problem that I will solve is to estimate $f(x)$ with that data set. I shall consider that $f(x)$ is unknown and make the following three assumptions:

- *The distribution of the error (ϵ_i) follows a normal distribution centred on 0 with an unknown variance.*
- *The error (ϵ_i) is additive.*

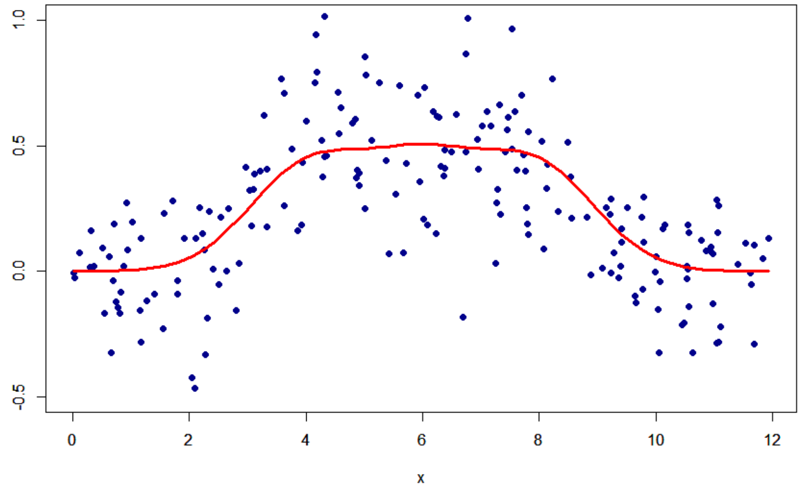

Figure 1: The Data Set

- *The errors are uncorrelated.*

In other words, I will assume the following:

$$y_i = f(x_i) + \epsilon_i,$$
$$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),$$

It is possible to confirm those assumptions but I will take them for granted. Validating them would not serve my purpose which is to compare a parametric and a nonparemetric approach. Given the way in which I constructed the data, the assumptions would be confirmed anyway.

Now, there are many different options to chose from in order to estimate $f(x)$. Here I will attempt to fit a polynomial and a kernel regression model. Both regressions are similar in the sense that their respective estimate $\widehat{f}(x)$ of the function $f(x)$, evaluated on the observed data, can be defined with a linear operator S (a linear smoother) that does not depend on y:

$$\widehat{f}(x) = Sy$$

However, both regressions are different in the sense that a polynomial regression is a parametric model and a kernel regression is a nonparametric model.

This distinction and its implications will become clearer. For the purpose of this paper, it is sufficient to say that a parametric model yields an estimate $\widehat{f}(x)$ such that we only have to know its parameters in order to compute it for any given x. On the other hand, a nonparametric model provides and estimate $\widehat{f}(x)$, such that we always need to know about the observations in our data set in order to compute $\widehat{f}(x)$ for any given x.

Before we move on with the estimation of $f(x)$, it is also worth mentioning that I did not construct that function naively. $f(x)$ has some properties that will highlight an important difference between the parametric and the nonparametric approach. It will help me to illustrate a way in which simplicity can lead to a better approximation of the truth.

2.2 Fitting a Polynomial Regression

To fit a polynomial regression, we must first assume that $f(x)$ has the following form:

$$f(x) = \sum_{k=0}^{p} \beta_k x^k$$

Secondly, we need to estimate the parameters β_k. We can do so by solving the following equation, which determines the parameters that will minimise the square of the difference between the observed y and $f(x)$:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} (y - f(x))^2$$

This will yield an appropriate maximum likelihood estimate of $f(x)$[1]:

$$\widehat{f}(x) = \sum_{k=0}^{p} \widehat{\beta}_k x^k$$

Finally, we need to figure out the number of parameters p that will determine the best estimate for $f(x)$ out of all the possible polynomial regression models that can fit the data set.

To understand the nature of this challenge, let us compare two different models. Figure 2 represents an estimate of $f(x)$ provided by a model with 4 adjustable parameters. Figure 3, on the other hand, represents an estimate provided by a model with 11 adjustable parameters. In the two figures the orange dashed line represents the estimate of the polynomial regression; the red line $f(x)$; and the small blue dots the data.

The question is to determine the best model out of the two. One intuitive criterion would be to compare the mean squared error for each model by using all the observations (x, y) in our data set. This quantity is called the training mean squared error (MSE_{train}).

$$MSE_{train} = \frac{1}{200} \sum_{i=1}^{200} (y_i - \widehat{f}(x_i))^2$$

[1] Under the assumptions made in section 2.1, least squares estimates and maximum likelihood estimates are the same.

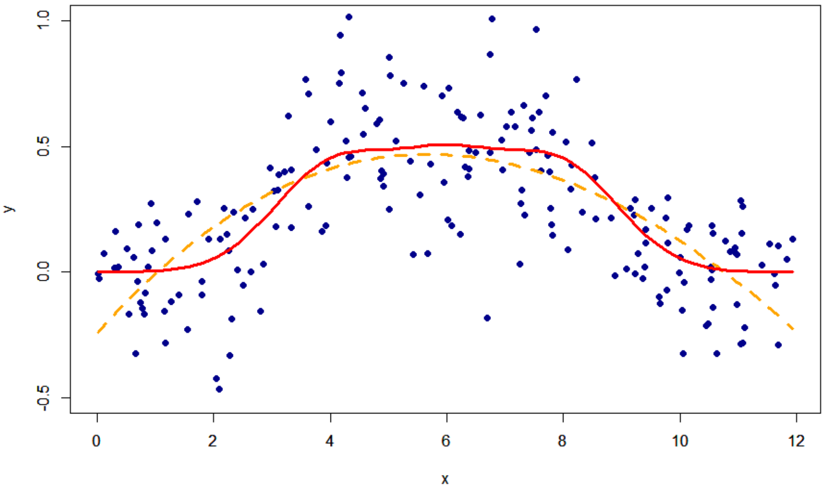

Figure 2: Polynomial Regression, Adjustable Parameters=4

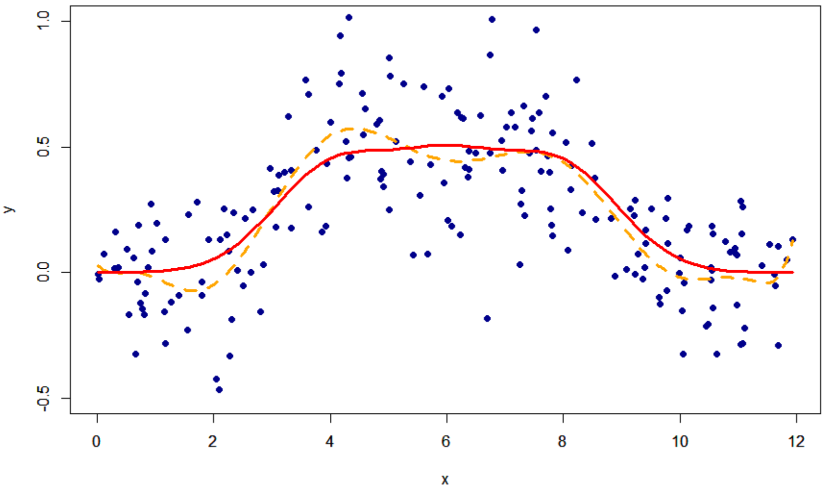

Figure 3: Polynomial Regression, Adjustable Parameters=11

Accordingly, one might conclude that the second model is better than the first because the $MSE_{train}$for the second model is smaller:

$$(0.0406 < 0.0532)$$

But this criterion would be inadequate.

Obviously, we are not interested in a model that can only fit the data at hand. What we really want is a model that fits observations that are not used to construct $\widehat{f}(x)$: $(x_{(new)}, y_{(new)})$. In other words, we would like to choose the model that has the smallest MSE_{test}.

$$MSE_{test} = \frac{1}{n}\sum_{i=1}^{n}(y_{i(new)} - \widehat{f}(x_{i(new)})^2$$

Unfortunately, the model that has the smallest MSE_{train} is not necessarily the one that has the smallest MSE_{test}. For example, when we are trying to fit a polynomial regression model, we can decrease MSE_{train} and increase the MSE_{test} by adding too many adjustable parameters to our model (*i.e.*, parameters whose values are not fixed before we fit the model to the data). When this happens, we say that our model is overfitting the data.

A more judicious choice would be to compute MSE_{test} directly with an independent data set. But in practice, we do not always have the luxury of having such an independent data set that we do not want to use in the construction of our model. A more common approach is to choose the model that minimises MSE_{train} and a penalty for the complexity of the model that is measured with the number of adjustable parameters k. The goal is to choose a model that does not overfit the data.

The Akaike Information Criterion (AIC) is one of the many criteria that implement that idea. For this analysis (under the assumptions made in section 2.1), the AIC can be expressed as follows:

$$AIC = 200 \log(\frac{1}{200} \sum_{i=1}^{200} (y_i - \widehat{f}(x_i))^2) + 2k$$

We will want to choose the model with the smallest AIC.

Another option is to estimate MSE_{test} by cross-validation (CV). One of the many ways to do cross-validation is to remove one observation from our data set; construct a model; and then compute the square of the difference between our prediction of the removed observation and that observation. If we repeat this procedure for every observation in our data set and average the results, we will obtain a value that can guide our choice of model: the smaller the CV the better. Here is how we can express CV, where $\widehat{f}_{(-i)}$ is the estimate of f obtained by omitting the pair x_i, y_i

$$CV = \frac{1}{200} \sum_{i}^{200} (y_i - \widehat{f}_{(-i)}(x_i))^2$$

If we use both criteria to make our choice we will find that the second model as a smaller AIC

$$(-618.6998 < -578.8056)$$

and a smaller CV

$$(0.04445178 < 0.05536721)$$

In fact, further exploration indicates that the second model is the best polynomial model according to both criteria.

2.3 Fitting a Kernel Regression

Now that we have found our best polynomial model (given CV and AIC), let's try to find the best kernel regression model. As we will see, this task will be significantly different. When we constructed the polynomial regression model we used what is called a 'top-down' approach. We determined *a priori* the form of our estimation for $f(x)$ and then tried to find the values of its adjustable parameters that best fit the data set. In other words, our estimation of $f(x)$ was limited to the family of polynomial functions.

On the other hand, when we wish to fit a kernel regression model, we do not make such strong *a priori* restrictions about the form of $f(x)$. In fact, we construct an estimate for $f(x)$ with the assumption that close-by x values must have similar y values. This approach is said to be 'bottom-up' because the estimate will depend more heavily on the observations that we have made.

To be more precise, for any given x_0, a kernel regression will provide a weighted mean value of all the observed y values that are within a certain range h from x_0. Its expression can be written as follows, where K is some unspecified kernel:

$$\widehat{f}(x_0) = \sum_{i=1}^{200} \frac{K(\frac{x_0 - x_i}{h})}{\sum_{i=1}^{200} K(\frac{x_0 - x_i}{h})} y_i$$

A kernel is a function that determines the weight of the nearby observations. In this paper, I will use an Epanechnikov kernel. It is defined as follows:

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2), & \text{if } |u| \leq 1 \\ 0, & \text{if not} \end{cases}$$

The challenge here will be to find the appropriate value for h. If h is too small, our estimate of $f(x)$ will overfit the data. But if h is too large, our estimate will tend to take the form of an horizontal line and the fit with the data set that we have will be awful. Just like in the parametric context, we will not be able to rely on MSE_{train} to choose our model. But, we will be able to rely on the AIC and CV.

If we use CV, we find that the best estimate is obtained with $h = 1.223$. The CV score associate with that h is 0.04400635. We can visualise the resulting estimate in Figure 4. As before, the orange dashed line represents the estimate of the kernel regression; the red line $f(x)$ and the small blue dots the data.

However, the application of the AIC criterion is not as straightforward in this case. As we can see, the only adjustable parameter here is h. It is the only expression in the equation of our model that is not fixed before we attempt to fit a kernel regression model with the data (the kernel has been determined *a priori*). Hence the number of adjustable parameters will be useless as a measure of complexity. To carry on, we will need to use a more general definition of a parameter in order to use the AIC for our kernel regression. We will have to determine what is called ”the effective number of parameters”(Friedman et al. 2001, p.232).

As mentioned in section 2.1, both the polynomial and the kernel regression estimates $\widehat{y}$ on the observed data (x, y) can be defined with a linear

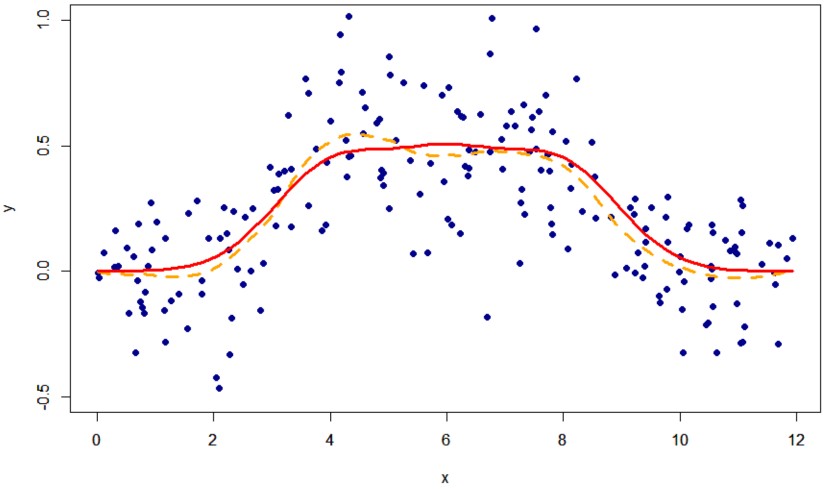

Figure 4: Epanechnikov Kernel Regression (CV)

operator S that does not depend on y:

$$\widehat{f}(x) = Sy$$

S is an interesting matrix because each element of its diagonal tell us how much weight is given to an observed y_i in order to compute the fitted value $\widehat{y_i}$. This means that if we compute the trace of S (the sum of all the elements in the diagonal of S), we will have a useful measure for the complexity of our model. Indeed, the regression line is likely to be more convoluted as we give more weight to each y_i to compute each $\widehat{y_i}$. In fact, the trace of S defines the effective number of parameters (Hurvich et al. 1998). It is used to generalise the AIC since it is also equal to the number of adjustable parameters in the parametric context.

The appropriate definition of the AIC can thus be expressed as follows, where tr(S) is the trace of the matrix S:

$$AIC = 200\log(\frac{1}{200}\sum_{i=1}^{200}(y_i - \widehat{f}(x_i))^2) + 2tr(S)$$

If we apply this criterion to our data set in order to choose a kernel regression model, we find that h=1.222 minimises the AIC with a value of -624.3763 and a number of effective parameters equal to 7.643 (Notice that the number of effective parameters is not necessarily an integer!). Its graph is very similar to the one presented in Figure 4.

If we compare our best parametric model with our best nonparametric models we conclude that we can obtain a better CV and a better AIC with the nonparametric approach. Here are the results

$$(0.04400635_{best_{CVnonpara}} < 0.04445178_{best_{CVpara}})$$

$$(-624.3763_{best_{AICnonpara}} < -618.6998_{best_{AICpara}})$$

3 Five Concepts of Simplicity

Given our data set and the choice-criteria that we have defined, the upshot of the previous analysis is that we will choose a kernel (nonparametric) regression model over a polynomial (parametric) one. The choice of one particular kernel regression estimate however is underdetermined since the two choice-criteria that we used do not converge.

In this section, I rely on that analysis to discuss the importance of simplicity in model selection. I will define 5 different concepts of simplicity. Doing so, I want to bring some important nuances to the existing literature on this topic.

My first objective is to correct a mistake that we often find in the philosophical literature about what makes a model simpler than another. My second objective is to show that the importance that we give for a particular notion of simplicity will depend on the goal that we pursue when we select a model. Therefore, when we wish to explain why simplicity matters in science, we have no choice but to take more than one definition of simplicity into account. In other words, I wish to support a view according to which different goals will justify the importance of different notions of simplicity. This is what I call a pluralist view of simplicity.

This kind of work is different from that of other philosophers, such as Kevin Kelly, who wish to explain why simplicity is important when our goal is to find the truth. See (Kelly 2007b) for example. By looking at other goals, we get a better understanding of the scientific practice of model selection. We will see that the importance of a particular concept of simplicity will depend on whether or not we are interested in a good predictive model; a model that can be constructed under computational or time constraints; an interpretable model; or in the validity of certain kinds of models. The fun fact is that we cannot always achieve all of these goals without making compromises. I will make this clearer in following sections.

3.1 Parametric Simplicity

Looking back at section 2, we see that simplicity plays a crucial role when we used the AIC to select our models. It is one of many criteria, like the Bayesian information criterion (BIC) and the Minimum Description

Length criterion (MDL), that relies on the idea that our model should maximise its fit to the training data and be penalised for its complexity. The justification for these criteria is that we want to avoid models that overfit the data, *i.e.*,we wish to avoid choosing models for which the MSE_{test} is larger than the MSE_{train}. This is essential to obtain a good predictive model.

For this particular reason, philosophers of science have been quick to underscore the importance of parametric simplicity in model selection:

> Model selection involves a trade-off between simplicity and fit for reasons that are now fairly well understood (Forster 2001, p.83).

> **Simplicity matters**. A sufficiently simple hypothesis, formulated on the basis of a given body of data, will not drastically overfit the data. It does not contain too many parameters whose values have been set according to the data. Thus, a simple hypothesis that successfully accommodates a given body of data can be expected to make more accurate predictions about new data than a more complex theory that fits the data equally well (Hitchcock and Sober 2004, p.22).

> Perhaps the most interesting of the standard arguments in favor of simplicity is based upon the concept of ”overfitting”. The idea is that predicting the future by means of an equation with too many free parameters compared to the size of the sample is more likely to produce a prediction far from the true value (Kelly 2007a, p.113).

The scientific relevance of simplicity has long been a matter of debate in philosophical circles. Therefore, it is easy to understand the appeal of a mathematically rigorous justification for the scientific relevance of parametric simplicity in model selection. It is no surprise that parametric simplicity has been the focus of several important articles written by philosophers such as Elliott Sober, Christopher Hitchcock, and Malcolm Forster.

However, the neglect of nonparametric models often results in false claims:

> In statistics, one theory, hypothesis or model is simpler than another if it has fewer adjustable parameters (Mikkelson 2006, p.441).

> there is general agreement among those working in this area that simplicity is to be cashed out in terms of the number of free (or adjustable) parameters of competing hypotheses. (Baker 2013).

> Interestingly, all three methods already mentioned, the MDL criterion, BIC and AIC, *define simplicity* in exactly the same way —as the paucity of adjustable parameters, or more exactly, the dimension of a family of functions [emphasis added on 'define simplicity'] (Forster 2001, p.90).

As we now know, a nonparametric model, like a kernel regression, can be too complex according to a criterion like AIC and have only *1 adjustable parameter*. Therefore, it is false to claim that a model that has fewer adjustable parameter is simpler and that a criterion like AIC defines simplicity in terms of the number of adjustable parameters. I believe that this kind of mistake is symptomatic of a lack of understanding of how parametric complexity can cause overfitting.

In the specific cases discussed in section 2, the number of parameters (effective parameters) is actually a measure of the weight given by a model to the observed y_i in order to compute their corresponding fitted values $\widehat{y_i}$. This is what explains the link between parametric simplicity and overfitting models. The more weight is given to y_i in order to compute its fitted value, the more our model will fit the data and thus model the irreducible error.

But more importantly, there is much more to simplicity than a property that allows us to avoid overfitting models. In fact, a good criterion to avoid overfitting model does not even need to take parametric simplicity into account. As we have seen, we can estimate MSE_{test} with CV and completely eliminate the need to rely on parametric simplicity. See also (Forster 2007; Hitchcock and Sober 2004).

In what follows, I will complete the picture[2]. By comparing parametric with non-parametric models, we can identify at least four other concepts of simplicity: theoretical, computational, epistemic, and dimensional. They are all important facets of simplicity that are not discussed in the literature mentioned in the introduction. They only become apparent

[2] I am not suggesting here that the previously mentioned philosophers are not aware that the picture is incomplete and that more work needs to be done. My intention is to bring the debates forward.

when we compare parametric models with their nonparametric counterparts.

3.2 Theoretical Simplicity VS Theory-ladenness

Going back to section 2.2 we can see that I have made a substantial assumption about the form of $f(x)$ in order to estimate it with a polynomial model. The quality of the estimate depended heavily on this assumption (that is why I defined $f(x)$ the way I did). If we look at figures 2 and 3 and compare the red and the orange dashed lines, we can see that a polynomial estimate will always fail to model the tails of $f(x)$. In other words, a false *a piori* assumption about the form of $f(x)$ can impose a limit on the quality of the estimate. This is why theory-laden approaches can be problematic.

On the other hand, we made no such *a priori* assumptions when we fitted a kernel regression model. We can immediately see how this paid off by looking at Figure 4. We see that the estimate provided by the kernel regression is closer to the true function. Therefore, theoretical simplicity seems to be of the utmost importance in this case.

But let us remember that we are not supposed to know the true function $f(x)$. Thus we are not supposed to see that a polynomial regression will fail to model the tails of $f(x)$ and that the kernel regression estimate is closer to the true function. What we do know however is that we obtained the best CV score with a Kernel regression. This gives us evidence that the MSE_{test} is lower for the Kernel regression than it is for the polynomial regression. Thus, we can now appreciate the importance of theoretical simplicity. Theoretically simpler models can have the best MSE_{test}. In other words they can provide us with better predictive models.

3.3 Computational Simplicity

On the other hand, one of the drawbacks of using nonparametric approaches is that they are computationally intensive. In the case of a kernel regression for example, the computer must find the neighbouring observations for each x, compute a weighted mean and then construct $\widehat{f}(x)$ point by point. In that respect, computational simplicity is a pragmatic virtue that the parametric approach has over the nonparametric.

Generally speaking, if our goal is to provide an estimate of $f(x)$ under time or computational restrictions, then computational simplicity will be an important virtue. Under such restrictions, we might also have to compromise on the idea of finding the best predictive model. But with the increasing power of our computers this is becoming less of an issue. This is why nonparametric approaches are now genuine alternatives to their traditional parametric counterparts. (It is also time for philosophers of science to ′catch the train′.)

3.4 Epistemic Simplicity

Another price to pay for using nonparametric models is that they are much more difficult to interpret. In comparison with parametric regressions, such as linear or polynomial regressions, nonparametric regressions ″can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response″ (James et al.

2013, p.25).

To see this, let us assume that a dependent variable y can be expressed in function of x plus an additive error term. As before, let us assume that the errors are uncorrelated and follow a centred normal distribution. Now consider the following two 2 estimates of the function:

$$\widehat{f}(x) = 4 + 5x \tag{1}$$

$$\widehat{f}(x_0) = \sum_{i=1}^{200} \frac{K(\frac{x_0 - x_i}{1.223})}{\sum_{i=1}^{200} K(\frac{x_0 - x_i}{1.223})} y_i \tag{2}$$

Just by looking at the parametric equation (1), we can easily obtain a wide variety of information about the relation between x and y. With practically no effort, we can interpret the parameters of our model. We can tell that the dependent variable y will increase by 5 units on average when x increases by 1 unit. We can also tell that 4 is the average value of y when $x = 0$. In other words, it is easy to understand how x is related to y.

In contrast, things are not as simple with the nonparametric equation (2). The relation between x and y is much more obscure. For instance, in order to find the roots (if there is any) of equation (2), we would need to find the x_0 (there maybe more than 1 or there maybe be none) such that close by x values in our data set have y values such that their weighted mean is equal to 0. Unless we have a computer, this problem is nowhere as simple as finding the intercept of equation (1).

In other words, the parametric model given by equation (1) is epistemically simpler because it is easier to understand the relationship between x and y. This is how I define epistemic simplicity. This virtue is especially important if we want a model that allows us to understand the relationship between our variables.

The fact of the matter is that there are research contexts where we do not necessarily wish to make predictions with a model, but where we want to know how an independent variable is related with the dependent variables. For instance, a scientist might be interested in knowing if maternal depression is positively related (at to what extend) with a child's learning difficulties in school. In that context, it is important to be able to interpret the resulting estimate of the function between the two variables. We could therefore have to choose a parametric model over a nonparametric one even if the latter makes more accurate predictions and is parametrically simpler.

In fact, there is often a compromise to make if we prefer interpretable over predictive models or *vice versa*. Depending on our goal (understanding or predicting) we might value parametric simplicity and epistemic simplicity differently:

> when inference is the goal, there are clear advantages to using simple [...] statistical learning methods. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest. For instance, if we seek to develop an algorithm to predict the price of a stock, our sole requirement for the algorithm is that it predict accurately -interpretability is not a concern (James et al. 2013, p.25).

Of course, if our nonparametric regression model only has one independent variable x for one dependent variable y, then we can easily plot that model in 2 dimensions in order to visualise and interpret it more easily. This is what I did in section 2. But this is not a solution when the number of dimensions is high. This brings me to one last notion of simplicity that is at play in model selection.

3.5 Dimensional Simplicity

When we fit a regression model of any kind, we must be wary of dimensionality. Dimensional simplicity is important for the same reason as parametric simplicity (they are often the same). For instance, we can severely overfit a regression model by adding independent variables. But again, there is more to dimensional simplicity than a tool to avoid overfitting. Nonparametric models, more specifically, are plagued with what is known as 'the curse of dimensionality' (James et al. 2013, p.108).

Recall that the epistemic foundation for a nonparametric model, like the one I presented in section 2, is the belief that close-by x values will have similar y values. In that example we had 200 observations that have been taken uniformly from 1 independent variable x. The distance between them was small enough to warrant that belief. However, if we were to spread 60 observations uniformly on a space determined by 50 independent variables, the distance between the observations would be so great that this fundamental belief would be very questionable. Generally speaking, the number of observations that we need in order to keep the same quality of estimation grows exponentially with the number of dimensions.

In contexts where further observations are difficult to obtain, this means that it will be useful to implement various techniques, such as principal component analysis[3], in order to reduce the dimension of our data set. In other words, dimensional simplification can be very important when we construct and choose a model. Not only does it allow us to avoid overfitting models but it is essential to maintain the validity a nonparametric model.

This conclusion seems to add weight to Sober's following quotes:

> The legitimacy of parsimony stands or falls, in a particular research context, on subject matter specific (and a posteriori) considerations. (Sober 1994, p.141).

> I have argued in earlier publications that invocations of parsimony in science often should be viewed as expressions of subject-matter-specific background theories; it follows that different invocations in different scientific problems may rest on very different foundations. Thus conceived, the way to understand the use of parsimony in a given scientific domain is to uncover the background theory in play (Sober 2009, p.238).

Against the theoretical background of a kernel regression, dimensional simplicity is particularly relevant. Unless we are in a context where the distance between our observations is too great to adequately fit a model like a kernel regression model (a posteriori consideration), we might not care as much about dimensional simplicity.

[3] Principal component analysis (PCA) is a technique that can "summarise" the variance of a data set into a lower dimension.

4 Conclusion

In sum, I have compared a parametric and a nonparametric approach to regression in order to differentiate five important notions of simplicity at play in model selection. I have therefore given a more complete account of the importance of simplicity in model selection than the one given in the current philosophical literature. The latter neglects the nonparametric approach and therefore has an unjustified and narrow focus on the number of adjustable parameters as a measure of simplicity. Here are four take-away conclusions:

- *The number of adjustable parameters is an inadequate measure of complexity for nonparametric models, like kernel regression models.*
- *The concept of effective parameter is more appropriate to measure simplicity when we are dealing with the family of linear smoother regressions.*
- *Besides parametric simplicity, there are at least four other important concepts of simplicity in model selection: theoretical, computational, epistemic, and dimensional.*
- *This variety of concepts makes it impossible to give a general and interesting (epistemic or pragmatic) justification for the importance of simplicity in model selection. Different goals justify the importance of different notions of simplicity. This is what I call a pluralist view of simplicity.*

Throughout this paper, I chose to discuss model selection within a frequentist framework. This approach is an important part of the current scientific practice. In order to understand the latter, we need to understand the former. By making this choice, I do not mean to imply that other frameworks, such as the Bayesian framework, are less important or justified. In fact, it would be interesting to compare the frequentist and the Bayesian approach to model selection by taking the non-parametric approaches into consideration. This is a topic for future work.

References

Baker, A. (2013). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 ed.).

Forster, M. and E. Sober (1994). How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Forster, M. R. (2001). The New Science of Simplicity. In *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*, pp. 83–119. Cambridge University Press.

Forster, M. R. (2007). A Philosopher's Guide to Empirical Success. *Philosophy of Science 74*(5), 588–600.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*, Volume 1. Springer series in statistics Springer, Berlin.

Friend, M., N. B. Goethe, and V. S. Harizanov (2007). *Induction, Algorithmic Learning Theory, and Philosophy*, Volume 9. Springer Science & Business Media.

Hitchcock, C. and E. Sober (2004). Orediction Versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science 55*(1), 1–34.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(2), 271–293.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Kelly, K. T. (2007a). How Simplicity Helps You Find the Truth Without Pointing at it. In *Induction, algorithmic learning theory, and philosophy*, pp. 111–143. Springer.

Kelly, K. T. (2007b). Ockhams Razor, Empirical Complexity, and Truth-Finding Efficiency. *Theoretical Computer Science 383*(2), 270–289.

Lurz, R. W. (2009). *The Philosophy of Animal Minds*. Cambridge University Press.

Mikkelson, G. M. (2006). Realism Versus Instrumentalism in a New Statistical Framework. *Philosophy of Science 73*(4), 440–447.

Sober, E. (1994). *From a Biological Point of View: Essays in Evolutionary Philosophy*. Cambridge University Press.

Sober, E. (2002). Instrumentalism, Parsimony, and the Akaike Framework. *Philosophy of Science 69*(S3), S112–S123.

Sober, E. (2009). Parsimony and Models of Animal Minds. In *The Philosophy of Animal Minds*, pp. 237–257. Cambridge University Press.

Zellner, A., H. A. Keuzenkamp, and M. McAleer (2001). *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge University Press.

Annex

The real function and the data set

```
library(psych)
f<-function(x) {
  dnorm(x, 6, 1)+ dnorm(x, 4, 1)+ dnorm(x, 8, 1)
}
set.seed(136)
x <- runif(200,0, 12)
y<-f(x) + rnorm(200, 0, 0.2)
d<-data.frame(x, y)
#plot(x, y, ylim=c(-0.5, 1), col="dark blue", lty=1, pch=19, lwd=1)
#par(new=T)
#curve(f(x), from=min(x), to = max(x), ylim=c(-0.5, 1), ylab="", col="red",
lty=1, lwd=3)
```

Polynomial regressions

3rd order model

```
reg3<- lm(y ~ x+I(x^2)+I(x^3))
reg3$coefficients

##  (Intercept)            x       I(x^2)       I(x^3)
## -0.244178747  0.263346283 -0.026423155  0.000377407

fitv<-reg3$fitted.values
datp<-cbind(d, fitv)
datp<-as.data.frame(datp)
datp<-datp[order(datp$x),]
#plot(x,y, ylim=c(-0.5, 1), col="dark blue", pch=19, lwd=1)
#lines(datp$x, datp$fitv, lwd=3, col="orange", lty=2, ylim=c(-0.5, 1))
#par(new=T)
#curve(f, from=min(x), to = max(x), col="red", lwd=3, ylim=c(-0.5, 1), ylab="
")
```

AIC

```
logs<-200*log((sum((reg3$fitted.values-y)^2))/200)
pen=(2*4)

aicrreg3<-logs+pen
aicrreg3
```

```
## [1] -578.8056
```

CV

```
cv3<-rep(NA, 200)
for(i in 1:200){
  reg<- lm(y[-i] ~ x[-i]+I(x[-i]^2)+I(x[-i]^3))
  ypr<-(reg$coefficients[1]+reg$coefficients[2]*(x[i])
        +reg$coefficients[3]*(x[i]^2)+ reg$coefficients[4]*(x[i]^3))
  cv3[i]<-(y[i]-ypr)
  cvr3<-sum(cv3^2)
}
cvr3/200

## [1] 0.05536721
```

10th order model

```
reg10<- lm(y ~
x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10))
reg10$coefficients

##   (Intercept)             x        I(x^2)        I(x^3)        I(x^4)
##  3.329178e-02 -3.291740e-01  1.074385e+00 -1.518393e+00  1.010995e+00
##        I(x^5)        I(x^6)        I(x^7)        I(x^8)        I(x^9)
## -3.576808e-01  7.343625e-02 -9.068850e-03  6.649447e-04 -2.668868e-05
##       I(x^10)
##  4.518782e-07

fitv<-reg10$fitted.values
datp<-cbind(d, fitv)
datp<-as.data.frame(datp)
datp<-datp[order(datp$x),]
#plot(x,y, ylim=c(-0.5, 1), col="dark blue", pch=19, lwd=1)
#lines(datp$x, datp$fitv, lwd=3, col="orange", lty=2, ylim=c(-0.5, 1))
#par(new=T)
#curve(f, from=min(x), to = max(x), col="red", lwd=3, ylim=c(-0.5, 1), ylab="
")
```

AIC

```
logs<-200*log((sum((reg10$fitted.values-y)^2))/200)
pen=(2*11)
aicrreg10<-logs+pen
aicrreg10

## [1] -618.6998
```

CV

```
cv10<-rep(NA, 200)
for(i in 1:200){
  reg<- lm(y[-i] ~ x[-i]+I(x[-i]^2)+I(x[-i]^3)+I(x[-i]^4)
           +I(x[-i]^5)+I(x[-i]^6)+I(x[-i]^7)+I(x[-i]^8)
           +I(x[-i]^9)+I(x[-i]^10))
  ypr<-(reg$coefficients[1]+reg$coefficients[2]*(x[i])
        +reg$coefficients[3]*(x[i]^2)+reg$coefficients[4]*(x[i]^3)
        +reg$coefficients[5]*(x[i]^4)+reg$coefficients[6]*(x[i]^5)
        +reg$coefficients[7]*(x[i]^6) +reg$coefficients[8]*(x[i]^7)
        +reg$coefficients[9]*(x[i]^8)+reg$coefficients[10]*(x[i]^9)
        +reg$coefficients[11]*(x[i]^10))
  cv10[i]<-(y[i]-ypr)
  cvr10<-sum(cv10^2)
}
cvr10/200
```

```
## [1] 0.04445178
```

Kernel regressions

How to find the best h with CV.

```
h = seq(1, 2, 0.001)
cv<-rep(NA, length(h))
for(i in 1:length(h)){
  u<-matrix(NA, nrow = 200, ncol = 200)
  for(j in 1:200){
    u[j,]<-(x[j]-x)/h[i]
  }
  ep<-function(x){
    cond=abs(x)<=1
    ((3/4)*(1-(x^2)))*cond
  }
  M<- ep(u)
  N<-apply(M, 1, sum)
  L = matrix(NA, nrow = 200, ncol = 200)
  for( k in 1:200){
    L[k,] = M[k,]/N[k]
  }
  yhat = L%*%y
  v<-rep(NA, 200)
  for(l in 1:200){
    v[l]<-(y[l]-yhat[l])/(1-L[l,l])
  }
  cv[i]<-(sum(v^2))
}
min(cv)/200
```

```
## [1] 0.04400635
```

```
h[which.min(cv)]
```

```
## [1] 1.223
```

How to find the best h with AIC

```
h = seq(1, 2, 0.001)
aic<-rep(NA, length(h))
for(i in 1:length(h)){
  u<-matrix(NA, nrow = 200, ncol = 200)
  for(j in 1:200){
    u[j,]<-(x[j]-x)/h[i]
  }
  ep<-function(x){
    cond=abs(x)<=1
    ((3/4)*(1-(x^2)))*cond
  }
  M<- ep(u)
  N<-apply(M, 1, sum)
  L = matrix(NA, nrow = 200, ncol = 200)
  for( k in 1:200){
    L[k,] = M[k,]/N[k]
  }
  yhat = L%*%y
  trace=tr(L)
  logsig=200*log(sum((y-yhat)^2)/200)
  pen=(2*tr(L))
  aic[i]<-logsig+pen
}
min(aic)
```

```
## [1] -624.3763
```

```
h[which.min(aic)]
```

```
## [1] 1.222
```

Kernel regression with the best h (CV)

```
hopt=1.223
uopt<-matrix(NA, nrow = 200, ncol = 200)
for(j in 1:200){
  uopt[j,]<-(x[j]-x)/hopt
}

ep<-function(x){
  cond=abs(x)<=1
  ((3/4)*(1-(x^2)))*cond
}
Mopt<- ep(uopt)
Nopt<-apply(Mopt, 1, sum)
```

```
Lopt = matrix(NA, nrow = 200, ncol = 200)
for( k in 1:200){
  Lopt[k,] = Mopt[k,]/Nopt[k]
}
yhatopt = Lopt%*%y
datpred<-cbind(d, yhatopt)
datpred<-as.data.frame(datpred)
datpred<-datpred[order(datpred$x),]
#plot(x, y, ylim=c(-0.5, 1),pch=19, lwd=1, col="dark blue")
#lines(datpred$x,datpred$yhatopt, lwd=3, col="orange", lty=2, ylim=c(-0.5,
1))
#par(new=T)
#curve(f, from=min(x), to = max(x), col="red", lwd=3, ylim=c(-0.5, 1), ylab="
")
```