INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Prediciendo el Presente con *Google Trends*

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA

ANDRÉS POTAPCZYNSKI GUIZA

ASESORA

DRA. CLAUDIA GÓMEZ WULSCHNER

MÉXICO, D.F. 2018

A mis padres.
Quienes dieron
y siguien dando todo su esfuerzo
por verme crecer.

En particular a mi papá
por todo el amor que me dió
y por enseñarme su perspectiva
de vida.

Y en particular a mi mamá
por darme tanto amor y tanto cariño.
También por ser el sustento y la fuerza
detrás de todos mis éxitos

Agradecimientos

A *Lili*, por ser una continua motivación, por sacar lo mejor de mí y por darme tanto cariño. Gracias por todos los momentos que hemos pasado juntos y por los que faltan.

A *Claudia Gómez*, por enseñarme a hacer matemáticas. Los sólidos fundamentos que tengo se los debo a la dedicación, al cariño y al tiempo que me dedicaste en todas las clases que me diste. Más aún, muchas gracias por el apoyo en este trabajo pero, sobre todo, por seguir apoyandome en mis proyectos a futuro.

A *Carlos Bosch*, por enseñarme que hasta las matemáticas más avanzadas (análsis funcional) se pueden entender de una manera intuitiva. Esta intuición la busco constantemente en mi trabajo. También agradezco los consejos que me has dado, aunque debo admitir que te tengo que poner más atención.

A *César Luis García*, por la calidad y el contenido de los cursos que tomé contigo. Esta misma calidad la busco imponer en el trabajo que hago. Más aún, aprecio mucho todas las charlas que hemos tenido. Estas siempre han sido muy enriquecedoras.

A *Guillermo Grabinsky*, por la excelente exposición de tus clases y por motivarme a hacer matemáticas. Hace 6 años tuvimos una conversación sobre análisis matemático que puedo identificar como el comienzo de mi gran gusto por el tema.

A *Beatriz Rumbos*, por estar siempre a la vanguardia y por enseñarme Cálculo Variacional y Teoría de Control. Más importante, por enseñarme que es inútil e imposible modelar la realidad con toda su complejidad. Lo importante de un modelo matemático es que logre extraer las variables más importantes para el fenómeno que busca explicar.

A *Carlos de la Isla*, por enseñarme a pensar críticamente. Actualmente obtenemos conocimientos muy poderosos sin ningún requerimiento moral. Tus clases son el contrapeso que me ha guiado a aplicar estos conocimientos de manera ética.

A *Juan Carlos Mansur*, por engrandecer mi gusto por la filosofía. No puedo negar el papel fundamental que ésta ha tenido en mi perspectiva de vida.

A *Alejando Hernández*, por enseñarme a ver la narrativa que hay detrás de los modelos matemáticos. Puedo decir que tu curso me mostró la relevancia y la claridad que un marco matemático establece para analizar temas sociales.

A *Ignacio Lobato*, por mostrar un gran interés en que me titule. Pero sobre todo por la exigencia de tus cursos de econometría, los cuales han facilitado mucho mi entendimiento en temas más avanzados.

A *Emilio Gutiérrez*, por sugerirme el tema de este trabajo y por abrirme la puerta cada martes. Sin esto, no hubiera podido avanzar tan rápido.

A *Juan Carlos Martínez*, por el tiempo que le has dedicado al presente trabajo y por el apoyo que me has ofrecido para engrandecerlo. También, vale la pena reconocer el buen gusto presente en tus temás de interés.

A *León Berdichevsky*, por hacerme un espacio en tu apretada agenda para revisar mi trabajo y por tus comentarios. Gracias a esto por fin me puedo titular.

A *todos mi amigos y amigas*, por estar siempre ahí cuando los necesito. Y por todas las divertidas e interesantes conversaciones que tenemos.

A *mi familia*, por que puedo contar con ustedes incondicionalmente. Aprovecho para reiterarles que ustedes también cuentan conmigo incondicionalmente.

Summary

Before the *era of Big Data*, macroeconomic, health and other variables had been consistently predicted using standard sources of information and simple linear models. Now, the overcapacity to store data has made new sources and high-frequency information worth capturing. Moreover, the surge of data has been met with a continuous development of readily available technologies. To this extent, the present work explores whether a new particular source of information could potentially help economic, health and other authorities in their decision-making process. Relevant variables usually have a reporting lag. We explore whether real-time internet searches could potentially notify authorities of changes in these variables of interest. Additionally, we explore other non-linear model alternatives to further increase the benefits of these new predictors. We also provide a multivariate proof of the well-known fact that, given sufficient information, a nonparametric approach can estimate any regression function. Finally, we propose an innovative alternative for estimating the nonparametric method. This is our main contribution to the statistic's literature since we find that this strategy is more suitable for small to medium data sets.

Contents

1 Motivation **7**

2 Data **11**
2.1 Google's search queries 11
2.2 Time series data 14

3 Empirical Approach **15**
3.1 Evaluation Strategy 16
3.1.1 Constructing Test Sets for Time Series 16
3.1.2 Choosing an Evaluation Metric 18
3.2 Model Specification 19
3.2.1 Autoregressive Linear Regression 20
3.2.2 Autoregressive Kernel Regression 20
3.3 What approach to choose? 31
3.4 Variable Selection 32
3.4.1 Selecting Relevant Search Terms 32
3.4.2 Deciding what to include in the model 33

4 Results **34**
4.1 ARI Cases Results 34
4.1.1 Data Exploration 34
4.1.2 Test Evaluation Results 37
4.1.3 Test Residuals Examination 38
4.1.4 Prediction versus Actuals 40
4.1.5 Implementation Improvements 40
4.2 Unemployment Rate Results 41
4.2.1 Data Exploration 41
4.2.2 Test Evaluation Results 43
4.2.3 Test Residuals Examination 44
4.2.4 Prediction versus Actuals 46
4.3 Homicide Cases Results 46
4.3.1 Data Exploration 46
4.3.2 Test Evaluation Results 48
4.3.3 Test Residuals Examination 49
4.3.4 Prediction versus Actuals 51

5 Conclusions **52**

A Appendix **54**
A.1 Modes of Convergence 54
A.2 Big-O and Little-o Arithmetic's 55
A.3 Taylor Expansions 56
A.4 Kernel Functions 58
A.4.1 Kernel Definition 58
A.4.2 Kernel Examples 60
A.5 A Note on Integrating Taylor Expansions with Kernels 62
A.6 Useful Propositions 63
A.7 Density Estimation 74

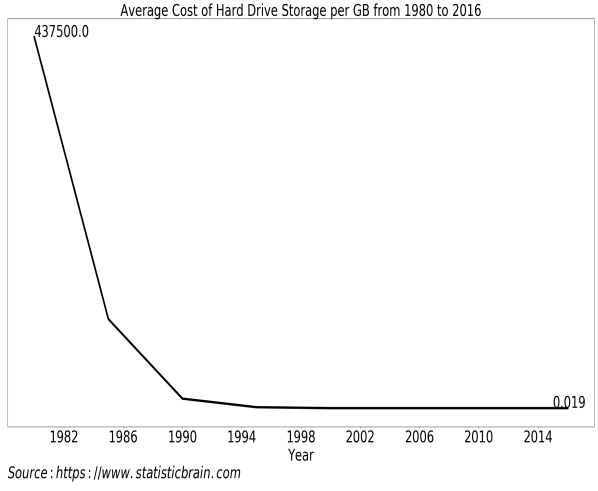
B Bibliography **78**

1 Motivation

The *era of Big Data* is expected to innovate many aspects of our lives. It promises to offer user-specific products (both physical and digital) as well as to automate many of the repetitive tasks that we face on a daily basis. Furthermore, the popularization of Machine Learning techniques is poised to increase the forecasting accuracy in many different industries as well as to uncover hidden patterns and structure in their data bases. Therefore, it is natural to explore how these new set of techniques and models could influence the research and decision making for Economic matters.

To provide some context for this *Big Data* phenomenon it is worth pointing out that this *era* is the result of the conjunction of three important circumstances. First, the cost of storing information has become negligible. Second, we are facing a datafication process where information from all aspects of our life is being systematically captured and stored. Finally, computational power has become inexpensive and therefore highly accessible as well as the technologies to manipulate and extract information from the data.

In terms of the cost of storing information, the graph below shows the evolution of the cost per GB (Gigabyte) through almost the last 20 years.



As it can be seen, the cost has decreased exponentially to the point of almost

mitigating the cost-benefit trade-off of opting for how much information to maintain. Companies no longer need to decide what summaries of their information to backlog, they can simply keep it all. Instead of storing the monthly aggregate sales of their products in different regions, they can now store the daily sales of each product in each of the stores that they have. This circumstance is not exclusive of large enterprises but it is also reflected in our daily activities. For example, every day 100 hours of video are uploaded to YouTube every minute (Murphy, 2012). Moreover, most of our personal use of data comes from the increasing number and quality of the pictures that we take with our cameras. In total, IBM estimates that we are generating 2.5 quintillion bytes of data each day. Of which 90% of this data was created in the last two years (Silver, 2015).

Alongside the low storage costs of information, or perhaps a result from it, we are experiencing a datafication process of every aspect of our live. As Kenneth Neil Cukier and Viktor Mayer-Schoenberger claimed in their article *The Rise of Big Data*, everything we do, online or otherwise, ends up recorded for later examination in someone's data storage units which may even be for sale (Cukier et. al., 2013). Following this line of though, Facebook is datafiying friendships, Twitter stray thoughts and LinkedIn professional networks. We have recently witnessed with the Cambridge Analytica case how this information may even stir the results of democratic elections. This datafiying process will only catalyze with the advent of the Internet of Things (IoT) technology. For our context, this technology can be overly simplified as the inclusion of monitoring software in each of our daily life devices (from phones and watches to water warming kettlebells and refrigerators).

Nonetheless, the abundance of information does not facilitate the insight extraction process. As essayist and statistician Nassim Taleb states, the problem with larger amounts of information is that *the needle comes in an increasingly larger haystack* (Silver, 2015). Moreover, we are even facing a new set of mathematical problems imposed by the colossal set of measurements taken for each observation. The clearest example being the genome encodings databases. For each individual there are $3.8e^9$ genome base pairs that could be included as features for a certain analysis (Murphy, 2012). Now the set of explanatory variables is quite broader than the set of observations; even deciding which variables influence the outcome is a problem on its own (variable selection). Yet, it is also true that the computation power and technology to tackle these problems has become more accessible.

Consider that to run a sophisticated Markov Chain Monte Carlo simulation

we do not need to develop and implement an algorithm or to have access to sophisticated software. Worse yet, we do not even need to understand the method to use it. We could simply write some lines of code on our personal computer and probably get a result faster than the former scientists of the Manhattan Project. This new accessibility paradigm could be considered as one of the propelling forces behind the hype for *Big Data*. Specially since this hype is a result of the fierce demand of the industry to employ the models that statisticians and computer scientists have been working on for the last decades. Most of the machine learning algorithms were not developed recently and have been here for a while. For example, convolutional neural networks, a deep learning technique, was developed around 1980 (Goodfellow et at, 2016). But the new paradigm stems from the off-the-shelve technologies like `Python` and `R` that allow different industries to implement this deep learning models effortlessly.

The term *Big Data* overemphasizes size as the most relevant ingredient of the analysis. It is undeniable that the granularity with which the data is being captured does allow for a more definitive and tailored investigation. However, some *state-of-the-art* applications do not stem from zooming deeper into the data but rather from employing different sources of information (many which did not even exist before). For example, after a speech a politician might get a real-time reaction of the electorate by performing a *sentiment analysis* on Twitter. Or political scientists might get a more precise estimate of the turn-out rate per state by looking at searches in Google related to finding a voting location rather than by running polls (Stephens-Davidowitz, 2013). The present work explores how Google searches could help health, economic and other authorities get a current estimate of relevant variables, such as the number of acute respiratory infection (ARI) cases, the unemployment rate or the homicide count, before they are actually reported and, consequently, generate a *real-time* feedback loop.

This approach promises to increase response times since, depending on the variable of interest, the reports usually have a lag of one to three months or even up to a year before they are announced (like INEGI[1]'s homicide database). In terms of methodology, it is worth pointing out that rather than using previous internet searches to forecast the variable for a future period, it is used to get an estimate on the actual period. To exemplify, we could employ Google searches related to flu symptoms in November to get an estimate of how much we expect the number of cases in that month to increase. In contrast to using this information or previous months to forecast the number of cases for a future month such

[1]Mexico's National Bureau of Statistics and Geography

as December or January. This distinction is vital since what is being hypothesized is that the internet searches used throughout this work are a reflection of the current situation rather than a premonition for the future.

The current work contributes to two growing literatures. The first is the literature related to using search engine data to forecast different macroeconomic variables. This work adds to this literature in mainly two dimensions. First, we expand popular methodologies and proven analyses to the Mexican context (such as health and unemployment rate *nowcasting*). Secondly, we propose beyond linear model specifications and actually explore the use of completely nonparametric approaches. The second contribution is to the statistical learning literature. In that context we prove the convergence (in a multivariate case) of a popular nonparametric approach which is not readily found in classic books of these topics. Moreover, we suggest some implementation changes for the kernel regression optimization strategy that improve the performance when compare to a standard package like `statsmodels` in `Python`. The improvements are the result of a different numerical optimization strategy.

It appears that the first published paper that made the connection between internet searches and macroeconomic forecasts was Ettredge. There they focused on unemployment rates in the U.S. (Ettredge et al., 2005). Also, from 2005 to 2010 there were several publications that related internet searches with epidemiology topics from cancer (Cooper et al., 2005) up to influenza (Polgreen et al., 2008) and others (Ginsberg et al., 2009), (Brownstein et al., 2009), (Valdivia and Monge-Corella, 2010) to mention a few.

The use of search engine data for economic applications was motivated and popularized by Choi and Varian, (2011). There, they employed *Google Trends* data on searches related to automobile sales, unemployment claims, travel destinations and consumer confidence to forecast the movement of the related economic variables. They proposed a simple seasonal linear autoregressive model that, depending on the application, exhibited a 5% to 20% improvement in accuracy when employing relevant *Google Trends* searches. McLaren and Shanbhoge (2011) summarize how search data could benefit central bank's decision making. In that work they introduced and defined the term *nowcasting* for the set of indicators that could help central authorities get a more precise view of the current situation of different economic indicators.

In terms of the variable selection techniques, Scott and Varian (2014) were the first to propose the use of a technique called Bayesian Structural Time Series to determine which variables should be included when *nowcasting* economic time

series. This technique is the combination of three procedures: Kalman Filters to estimate trends and seasonality effects, Spike-and-Slab regression to determine a posteriori distribution on what variables should be included in the model and Bayesian Model Averaging to combine the best performing models for the final forecast. The use of Kalman Filters for time series was first proposed by Harvey (1991), Durbin and Koopman (2001). Spike-and-slab was developed by McCulloch (1997) and Madigan and Raftery (MadRaf, 1994). Finally, averaging over ensemble of models was again motivated by Madigan and Raftery (1994) and Volinsky (2012). This methodology has been successfully employed in a broad set of additional studies.

The non-linear model extensions that the current work proposes are inspired and borrow heavily from Cosma Shalizi (Shalizi, 2017).

The present work is organized as follows. Chapter 2 explains the information used for the analyses. Chapter 3 examines all the time-series model specifications tested as well as the variable selection mechanism used to discern which internet searches had predictive power. Chapter 4 presents the results of employing the model to the data. Chapter 5 offers concluding remarks. Finally, the Appendix contains all the material needed to fully comprehend the proofs presented in this work.

2 Data

There are two different kinds of data used throughout the present work: Google's search queries and the different time series variables. INEGI is the source for both the Mexico's urban unemployment rate and Mexico's city homicide felonies, while Secretaria de Salud is the source for the acute respiratory infections (ARI). Below is a detailed description of all the information.

2.1 Google's search queries

Google provides two services, namely *Google Correlate* and *Google Trends*, that allows us, respectively, to find which search terms might be relevant for a certain time series and to download this information.

Google Correlate is a web-service backed by an algorithm that, for a given time series we input, it returns the search terms that are most closely correlated with it. Intuitively, it takes the input time series, normalizes it by subtracting the mean and standardizing it to 1 and then employs an approximate nearest

neighbor search to find the top 100 queries that are closer to this time series based on the Pearson correlation metric. Moreover, to avoid irrelevant searches it employs different techniques to filter terms that might not have predictive power. Additionally, in order to efficiently explore all the search terms space, it uses a hash function approach (Vanderkam et. al., 2015).

In the context of this work, the *Google Correlate* algorithm was mostly used to hint what type of search terms might be relevant for each of the different analyses. Except for the homicide case, where the exact search term suggested was used. It is worth pointing out that the underlying reason for not using the exact search terms suggested by *Google Correlate* is that those terms suggested were not the most relevant in terms of the context; relevant in the sense of the aggregate number of queries. For example, the *Google Correlate* algorithm suggested that, in the unemployment context, employment finding websites were relevant; however, they were not the most popular sites and hence were substituted for other which were more popular and had a higher predictive power. Also, for the ARI context, the algorithm suggested the use of some flu medications but they were substituted for the more popular medications.

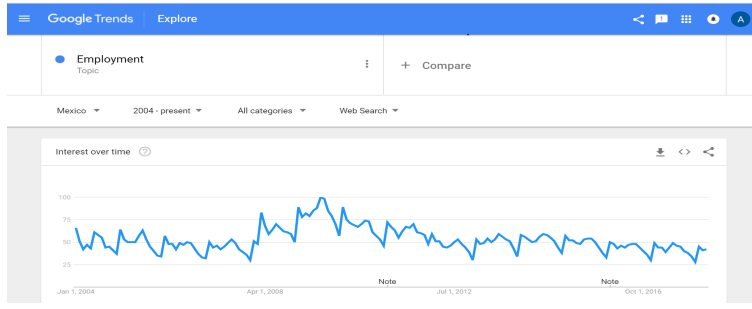
The second service, *Google Trends*, is a web-service that allows us to access the information of a given search term for a specific time range and geographical location. The output data comes as an indexed time series which is constructed in the following manner.

First, it finds all the relevant search terms given the context. To find these search terms it uses word processing and other *Natural Language Processing* techniques. An example is that if someone searches `car` then `automobiles` and `used cars` would be included in the output index. It is also possible to select the exact search them by using double quotation (`"car"`). Moreover, *Google Trends* has 26 categories to group the search terms. The classic example being that if someone searches `apple`, then it is possible to select the results that are only relevant to the `Computer & Electronics` category which would relate to the company Apple. Or if the `Food & Drink` category is selected then the results would relate to the actual fruit. This distinction is achieved by analyzing the search activity. Thus, if the search term in one case was *apple chargers* then it would fall into the first category, whereas for the other case if the subsequent search from *apple* was *orange* then it would fall on the second category.

Then, merging all relevant search terms found in the previous step, a percentage from total is calculated. The percentage is computed by dividing the aggregate activity of the merged search terms by the total number of queries in

that time frame and location. For example, assume that the total activity in a certain period and location is 100. Also, referring back to the previous example, assume that the number of `car` searches is 1, of `automobile` 0 and of `used cars` 1, then the percentage reported would be 2%. However, if for the next year and same location, the total number has quadruple but the search activity has only doubled then the reported percentage would be 1%. This normalization is done to understand how the relevancy of the search terms evolved over time discounting the natural growth of internet activity.

Finally, the reported percentages for each period are normalized so that the highest value in that time frame and geographical location is 100. A reported graph would be similar to



Other considerations are the following. The time frequency outputted changes depending on the length of the time frame. Thus, if the time frame is one day, then the data is reported in an hourly manner. Whereas if the time frame is from the first date available 2004-01 up until 2018-01, then the data is reported in a monthly manner. Also, there is a privacy threshold defined, which if not met then the data is not reported. In other words, there needs to be enough search activity for it to be reported. Finally, the results can vary day by day since the reports are constructed with the given sample of the data used in that day.

The search terms for each of the three analyses are the following. For the ARI analysis, the search terms used fall into two categories. The influenza category where the search term used was `estacional` which related to `Influenza estacional`. The second category relates to flu symptoms and medicines, the search term `"antigripales + rinofaringitis + klaricid + antiflu-des"` was used. The last search term combined flu related medications (`klaricid`

and `antiflu-des`) and the formal medical name of flu: *rhinopharyngitis*. The time frame embarks the beginning of 2010 up to until the beginning of 2015 and it comes at a weekly level. For the unemployment analysis the search queries fell into the `Employment -- Topic`, which contains as its main search terms: `empleo`, `portal empleo`, `occ`, `bolsa de trabajo`, `computrabajo`, `empleo.gob`, and others. The previous search terms are a mixture of employment key words (such as `empleo`) and of employment finding websites (`portal empleo`, `occ`, `computrabajo`, `empleo.gob`). The data goes from 2004 up to 2018 and it comes at a monthly level. Finally, the search term for the homicide rate was `horarios misa` from 2004 up to 2018 which comes at a monthly level. This search term alludes to the religious church mass time schedule.

2.2 Time series data

Moving into the data used for the present work, the acute respiratory infections (ARI) data is provided by the Department of Health's Reported Cases Dataset. The data is collected at a weekly basis and it contains all ARI diagnoses and reported cases per each of the 16,250 clinics in the sample. Only the clinics belonging to the most relevant health subsystems are included (*Seguro Popular, IMSS, IMSS-Oportunidades and ISSSTE*)[2] where the excluded constitute around 1% of public health services in the National Survey of Health and Nutrition 2012 (*ENSANUT* for its Spanish acronym). The data set records the ARI diagnoses based on the J09X up to J22X ICD codes. It is also worth noticing that each public outpatient clinic is legally required to report this information. In terms of reporting the data, while each clinic has immediate knowledge of the number of cases and hence notice a surge on ARIs it is not clear how long it takes to coordinate and aggregate this information at a national level. We assume a one month lag.

The unemployment rate used in this work is provided by INEGI. It is reported in a monthly format, aggregated at a national urban level. It has a reporting lag of one to two months. It is computed by dividing the population over 15 years which is currently unemployed but searching for a job by the population over 15 years which is economically active (which is either working or in the

[2]The Mexican Health Subsystems are: Seguro Popular, Instituto Mexicano del Seguro Social (IMSS), IMSS-Oportunidades which is a joint program between the previous subsystem and the social welfare program Oportunidades. Finally, Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE) which is a health program for government workers.

condition to work). The size of both populations is estimated based on the National Employment and Occupation survey (*ENOE* for its spanish acronym). The survey generally asks in the sense of whether the respondent is currently employed, to describe its job or whether the respondent has been looking for employment (and for how long) between many other questions. Moreover, it has a sampling strategy where the households are stratified and then, selected with a certain probability in a two-phased process based on different criteria. This last procedure guarantees the representativity of the information.

Again, Mexico City's homicide rate is constructed from INEGI's General Deceased data base (*Defunciones Generales*). It is a data base which records the deaths of Mexican citizens and their diverse causes: from health issues, accidents or violence related. INEGI's personnel is in charge of an active recollection of this information which comes in three types of documentation: *acta de defunción*, *certificado de defunción*, *cuaderno de defunciones*. This information is provided by the diverse organizations that manage death related events such as the Civil Registry (*Registro Civil*) or the Public Prosecutor's Office (*Ministerio Público*). Moreover, INEGI's personnel controls the quality and ensures the correct processing of the information into digital form. The variables used for the analysis were the following: (1) the columns used to specify the day, month and year of the event occurrence were: `dia_ocurr`, `mes_ocurr`, `anio_occur`, respectively. (2) The column used to filter information by state was: `ent_ocurr`. Finally, (3) the following three columns allowed to subset the information into homicide events only. The `presunto` column indicates which case stemmed from a homicide event. Additionally, `causa_def` and `lista_mex` provided additional detail about the characteristics of the event.

3 Empirical Approach

This section discusses the different empirical approaches taken throughout this work. The discussion centers around three main topics: how was performance evaluated, what were the model specifications tested and how were the model's predictors selected. The first topic exhibits the design of the *out-of-sample* (or test) sets and the rationale behind the evaluation metrics employed. The second topic analyzes the parametric and nonparametric models used. It exhaustively develops the convergence theory of the nonparametric approach and proposes some implementation adjustments. It concludes by comparing the advantages and disadvantages of the parametric model against the nonparametric one. It

also discusses other possible model alternatives in between. Finally, the third topic deals with the process of selecting what predictors should be included in the model. It illustrates how relevant search terms are selected from the myriad of choices and, additionally, it examines how many predictor lags should be included.

3.1 Evaluation Strategy

3.1.1 Constructing Test Sets for Time Series

It is now standard to evaluate the performance of a model with observations that were not used while it was being estimated. The usual approach is that, given a data set, around 70-80% of the observations are used for model estimation (or training), while the remaining 20-30% are used for model evaluation (or testing). This leads to the so-called, train and test split. This is done since, given enough capacity, some models are able to reduce their train error to zero (essentially interpolating the data). Yet, they will not generalize properly to new observations. The poor generalization occurs due to the high variance introduced by the model, which is really imitating the noise in the training data. The classic manner of splitting the data is to randomly assign 70-80% of the observations to the train set and the remaining to the test set (Hastie et. al., 2009). However, this is only appropriate when the observations are independent but not when they come from a time series.

In the time series context, where most certainly the lags of variables will be used as predictors, it is necessary to maintain the order of the data to take advantage of the time patterns. To argue why, suppose we are analyzing a yearly process that increases, on average, 3% every year. Note in the graph below how, if the data is randomly shuffled, the time pattern is ruined.

Therefore, to preserve time patterns, we are going to follow the next train/test approach. The train set will be formed from all the observations that range from the first period until the period that amounts to roughly ˜70% of the data. The remaining future periods are going to be left to the test set. To represent this, if we have monthly data from 2010-01 to 2018-01 then the split will be

$$\underbrace{2010/01 \quad \cdots \quad 2012/10 \quad \cdots \quad 2015/08}_{train \approx 70\%} \quad \underbrace{2015/09 \quad \cdots \quad 2018/01}_{test \approx 30\%}$$

The previous split deviates from the common time series approach of *forward chaining*. In this approach several train and tests sets are constructed and the prediction result is computed from the mean evaluation of the test sets. This scheme is shown below where ■ denotes a period used for training and ★ for testing

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
■	■	■	■	■	★		
■	■	■	■	■	■	★	
■	■	■	■	■	■	■	★

the result is the mean evaluation of each ★. To compare, our approach is shown below. It is worth noting that we are using the same information for testing

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
■	■	■	■	■	★	★	★

yet we are not iteratively re-training the model each time. This is done truly to have one model at the end and not an ensemble of models (since re-training will

be modifying estimated parameters each time). Additionally, this last approach places greater stress on the model's capacity since it is tested on more periods and it is not allowed to use the new information available.

3.1.2 Choosing an Evaluation Metric

Once the test set is constructed, then the model predictions are compared to the actual values for those periods. Let y_i denote the actual value at testing period i, $\widehat{y}_i$ be the model's prediction in the same period and $e_i \triangleq y_i - \widehat{y}_i$. In the Machine Learning literature, the classic metrics are

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y}_i| = m^{-1}\left\|\mathbf{e}\right\|_1$$

$$RMSE = \left(\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2\right)^{1/2} = m^{-1/2}\left\|\mathbf{e}\right\|_2$$

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \times 100$$

where MAE stands for Mean Absolute Error, $RMSE$ for Root Mean Square Error and $MAPE$ for Mean Absolute Percentage Error. Note that the first two have a $\left\|\cdot\right\|_p$ interpretation and the last one aims at being unitless. Their common goal is to provide a summary statistic of the overall closeness of the predictions, where each data point is considered equally important.

Note that throughout this work, as a pre-processing step, all the predictors are normalized to have mean 0 and standard deviation of 1. This is done primarily to improve the model's performance since the nonparametric approach is receptive to the units of the predictors but also to compare the results between the linear and the nonparametric approach. Due to the preprocessing step, the $MAPE$ loses relevance and the focus is placed on MAE and $RMSE$. Moreover, the most relevant metric for this work is the MAE simply because it is not as sensible to outliers as the $RMSE$.

Even though the quantitative results for this work rely on the outcome of these metrics they are also accompanied by other analyses such as *residuals examination* and *prediction versus actuals plots* to complement details that could be bypassed by these metrics. The motivation for these other analyses will be exposed in the results section.

3.2 Model Specification

The central problem of this section focuses on choosing the best model alternative to estimate a target variable given its past values and possibly some Google search terms. Let y_t denote the target variable at period t and let g_t stand for the Google predictors in that same period. Thus, the problem is to find a function $\widehat{\mu}_n(\cdot)$ that approximates

$$E\left[y_t|\mathbf{y}_{t-1}^{(p)}, g_t\right] \stackrel{\circ}{=} \mu\left(\mathbf{y}_{t-1}^{(p)}, g_t\right)$$

where $\mathbf{y}_{t-1}^{(p)} = y_{t-1}, \ldots, y_{t-p}$ and p indicates the number of lags included in the model. In this work $\stackrel{\circ}{=}$ stands for a new definition. Note that, due to the framing of the discussion, we are not making any assumptions on the distribution of the random variable y_t. This is because we are not trying to uncover the true data generating process but rather to asses which model is better at making *out-of-sample* predictions and, also, if the Google's search terms enhance this accuracy. The first approach analyzed in this section is an autoregressive linear regression of the following form

$$\widehat{\mu}_n\left(\mathbf{y}_{t-1}^{(p)}, g_t|\beta, \mathbf{D}\right) = \beta_0 + \sum_{l=1}^{p} \beta_l y_{t-l} + \beta_{p+1} g_t$$

which is a global linear parametric model on β. The second approach considered is a fully nonparametric model that estimates locally $\mu(\cdot)$. This is the autoregressive kernel regression which is constructed in the following manner.

$$\widehat{\mu}_n\left(\mathbf{y}_{t-1}^{(p)}, g_t|\mathbf{D}\right) = \sum_{i=1}^{n} \frac{K\left(\left(\mathbf{y}_{t-1}^{(p)} - \mathbf{y}_i^{(p)}, g_t - g_i\right) \odot h^{-1}\right) y_{it}}{\sum_{j=1}^{n} K\left(\left(\mathbf{y}_{t-1}^{(p)} - \mathbf{y}_j^{(p)}, g_t - g_j\right) \odot h^{-1}\right)}$$

where $\odot$ denotes element-wise vector multiplication, $h^{-1} = \left(h_1^{-1}, \ldots, h_{p+1}^{-1}\right)^T$, where $\mathbf{h}$ is the kernel's bandwidth which controls the sizes of the neighborhoods in consideration. Additionally, $\mathbf{D}$ denotes the training data set and $K(\cdot)$ a kernel function. The two approaches are fully discussed in the following sections.

3.2.1 Autoregressive Linear Regression

As stated previously, the autoregressive linear regression aims to approximate

$$E\left[y_t|y_{t-1},\ldots,y_{t-p},g_t\right] \approx \beta_0 + \sum_{l=1}^{p}\beta_l y_{t-l} + \beta_{r+1}g_t$$

where the vector parameter β is estimated from the training data. The procedure to estimate β involves solving the least-squares problem

$$\min_{\beta}\sum_{t=1}^{T}\left(y_t - \beta_0 + \sum_{l=1}^{p}\beta_l y_{t-l} + \beta_{p+1}g_t\right)^2$$

This model is commonly known as $AR\left(p\right)$. Finally, it is worth noting that this is the standard model employed in the *nowcasting* literature.

3.2.2 Autoregressive Kernel Regression

Proving Convergence Now, we will analyze the conditions that guarantee the correct functioning of the kernel regression's most attractive feature. Namely, that it recovers or learns any arbitrary $\mu\left(\cdot\right)$ function given enough information. Although this fact is well-known, a proof is not readily found on the classical books of the literature. Li and Racine provide a sketch of the proof for the univariate case. This is used as the general guideline for this work (Li and Racine, 2007). For this, we will drop the specific time series notation for a more general one. Also, since the propositions below use independent sample assumptions, we will proceed with those and, at the end, discuss what changes need to be considered for those assumptions to hold on a time series context.

The main result of this section states that

$$MSE\left(\widehat{\mu}_n\left(\mathbf{x}\right),\mu\left(\mathbf{x}\right)\right) = o\left(\left\|\mathbf{h}\right\|^3 + \left(nh_1\cdots h_p\right)\right)$$

thus, as $n\to+\infty$, $\left\|\mathbf{h}\right\|\to 0$ and $\left(nh_1,\ldots,h_p\right)\to+\infty$ then $\widehat{\mu}_n\left(\mathbf{x}\right)$ converges in

mean-square to $\mu(\mathbf{x})$. For this we are going to work with a particular quotient

$$E\left[\left(\widehat{\mu}_n(\mathbf{x})-\mu(\mathbf{x})\right)^2\right] = E\left[\left(\frac{\left(\widehat{\mu}_n(\mathbf{x})-\mu(\mathbf{x})\right)\widehat{f}_n(\mathbf{x})}{\widehat{f}_n(\mathbf{x})}\right)^2\right]$$
$$\approx E\left[\left(\frac{\left(\widehat{\mu}_n(\mathbf{x})-\mu(\mathbf{x})\right)\widehat{f}_n(\mathbf{x})}{f(\mathbf{x})}\right)^2\right]$$

where $f(\mathbf{x})$ is the density function to which we are taking expectations and $\widehat{f}_n(\mathbf{x})$ is it's kernel density estimator (see Density Estimation in the Appendix). We are going to derive an expression for the past expectation and then argue the overall convergence of $MSE\left(\widehat{\mu}_n(\mathbf{x}),\mu(\mathbf{x})\right)$. Note that by definition

$$E\left[\left(\left(\widehat{\mu}_n(\mathbf{x})-\mu(\mathbf{x})\right)\widehat{f}_n(\mathbf{x})\right)^2\right] = E\left[\left(\sum_{i=1}^{n}\frac{K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(Y_i-\mu(\mathbf{x})\right)}{nh_1\cdots h_p}\right)^2\right]$$

Since $var(\cdot)=E[(\cdot)^2]-E[(\cdot)]^2$, the past expectation becomes

$$= E\left[\sum_{i=1}^{n}\frac{K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(Y_i-\mu(\mathbf{x})\right)}{nh_1\cdots h_p}\right]^2 \qquad (bias^2)$$
$$+\, var\left(\sum_{i=1}^{n}\frac{K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(Y_i-\mu(\mathbf{x})\right)}{nh_1\cdots h_p}\right) \qquad (variance)$$

which brings to mind a square bias and variance decomposition. Let's deal with the bias first

$$E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(Y_i-\mu(\mathbf{x})\right)}{nh_1\cdots h_p}\right] =$$

$$E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(\mu(\mathbf{X}_i)-\mu(\mathbf{x})\right)}{nh_1\cdots h_p}\right]+E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon_i}{nh_1\cdots h_p}\right]$$

where $Y_i=\mu(\mathbf{X}_i)+\epsilon_i$ and $E[\epsilon_i|\mathbf{X}_1,\ldots\mathbf{X}_n]=0$. By proposition 10 of the Appendix, the last summand becomes 0 and since the sample is identically distributed therefore

$$= \frac{E\left[K\left(\left(\mathbf{X}-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(\mu(\mathbf{X})-\mu(\mathbf{x})\right)\right]}{h_1\cdots h_p}$$

by proposition 5 of the Appendix the past expectation becomes

$$= \frac{f(\mathbf{x})\kappa}{2}\sum_{r=1}^{p} h_r^2\left[\frac{2\mu_r(\mathbf{x})f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right] + o\left(\|\mathbf{h}\|^2\right)$$

finally, elevating the previous term to the second power yields the result from proposition 6. Hence the square bias of $MSE\left(\widehat{\mu}_n(\mathbf{x}), \mu(\mathbf{x})\right)$ becomes

$$bias^2 = \left(\sum_{r=1}^{p} \frac{f(\mathbf{x})\kappa h_r^2}{2}\left[\frac{2\mu_r(\mathbf{x})f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right]\right)^2 + o\left(\|\mathbf{h}\|^4\right)$$

In terms of the variance, note that

$$var\left(\sum_{i=1}^{n} \frac{K\left((\mathbf{X}_i - \mathbf{x})\odot\mathbf{h}^{-1}\right)(Y_i - \mu(\mathbf{x}))}{nh_1\cdots h_p}\right) = \frac{var\left(K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)(Y-\mu(\mathbf{x}))\right)}{nh_1^2\cdots h_p^2}$$

since the sample is independent and identically distributed. Also,

$$\begin{aligned} cov(\mu(\mathbf{X}), \epsilon) &= E[\mu(\mathbf{X})E[\epsilon|\mathbf{X}]] - E[E[\epsilon|\mathbf{X}]E[\mu(\mathbf{X})] \\ &= 0 \end{aligned}$$

therefore, the previous variance term becomes

$$= \frac{var\left(K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)(\mu(\mathbf{X})-\mu(\mathbf{x}))\right)}{nh_1^2\cdots h_p^2} + \frac{var\left(K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)\epsilon\right)}{nh_1^2\cdots h_p^2}$$

according to proposition 8 and 11 of the Appendix the expression above reduces to

$$\begin{aligned} variance = &\frac{f(\mathbf{x})\rho\sum_{i=1}^{p} h_i^2\mu_i(\mathbf{x})}{nh_1\cdots h_p} + \frac{\epsilon^2(\mathbf{x})f(\mathbf{x})\omega^p}{nh_1\cdots h_p} \\ &+ o\left((nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right) \end{aligned}$$

Dividing the previous results by $f(\mathbf{x})$ finally leaves us with

$$\begin{aligned} MSE\left(\widehat{\mu}_n(\mathbf{x}), \mu(\mathbf{x})\right) \approx &\left(\sum_{r=1}^{p} \frac{\kappa h_r^2}{2}\left[\frac{2\mu_r(\mathbf{x})f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right]\right)^2 \\ &+ \frac{\rho\sum_{i=1}^{p} h_i^2\mu_i(\mathbf{x}) + \epsilon^2(\mathbf{x})\omega^p}{f(\mathbf{x})(nh_1\cdots h_p)} \\ &+ o\left(\|\mathbf{h}\|^4 + (nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right) \end{aligned}$$

As usual, there is a conflicting bias and variance trade-off. The bias wants to drop $\|\mathbf{h}\|$ to 0 as fast as possible, but in that case the variance would tend to $+\infty$. Withal, if $n \to +\infty$, $\|\mathbf{h}\| \to 0$ but $(nh_1 \cdots h_p) \to +\infty$ then the MSE tends to zero. To accelerate this process, we can choose the kernel's bandwidth (the sizes of the neighborhoods) to minimize the MSE's leading term. In order to get an analytical solution we are going to worsen the approximation by ignoring the first term of the variance. Furthermore, assume that all the predictors are normalized and relevant, then all the bandwidths would be of a similar magnitude (say h) and therefore the leading term becomes

$$\zeta(h, \mathbf{x}) \doteq \frac{h^4}{4}\Omega(\mathbf{x}) + \frac{\varsigma(\mathbf{x})}{nh^p}$$

where

$$\Omega(\mathbf{x}) \doteq \left(\sum_{r=1}^{p} \kappa \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right]\right)^2$$

and

$$\varsigma(\mathbf{x}) \doteq \frac{\epsilon^2(\mathbf{x})\omega^p}{f(\mathbf{x})}$$

If we minimize the above equation with respect to h, then $h_{opt}(n, \mathbf{x})$ would have to solve

$$h_{opt}^3 \Omega(\mathbf{x}) = \frac{p\varsigma(\mathbf{x})}{nh_{opt}^{p+1}}$$

from this equation it follows that

$$h_{opt}(n, \mathbf{x}) = \left(\frac{p\varsigma(\mathbf{x})}{n\Omega(\mathbf{x})}\right)^{\frac{1}{p+4}}$$

which leaves us with

$$h_{opt}(n, \mathbf{x}) \propto n^{\frac{-1}{p+4}}$$

If we substitute this value of h_{opt} the convergence rate of the MSE becomes

$$MSE\left(\widehat{\mu}_n(\mathbf{x}), \mu(\mathbf{x})\right) = O\left(n^{\frac{-4}{p+4}}\right)$$

The above $h_{opt}(n, \mathbf{x})$ depends on $\mathbf{x}$ so if we want to get a general optimal bandwidth we could integrate the above MSE with respect to the probability

measure and hence work with

$$\zeta(h) \triangleq \frac{h^4}{4}\Omega + \frac{\varsigma}{nh^p}$$

where $\Omega = \int \Omega(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ and $\varsigma = \int \varsigma(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$. Following the same analysis, we equally obtain

$$E\left[MSE\left(\widehat{\mu}_n(\mathbf{X}), \mu(\mathbf{X})\right)\right] = O\left(n^{\frac{-4}{p+4}}\right)$$

Notice, however, that the optimal bandwidth is a theoretical result since we do not know $f(\cdot)$, $\mu(\cdot)$ or any of the derivatives of the two functions. Nonetheless, in the next section we discuss how to approximate the optimal bandwidth by cross-validation.

Choosing the Bandwidth It is standard to use cross-validation to approximate the generalization error of a model. In our context, since we only have a single sample of the data generating process, we use cross-validation to estimate $E\left[MSE\left(\widehat{\mu}_n(\mathbf{x}), \mu(\mathbf{x})\right)\right]$. This estimator is defined as

$$CV_k^J(\mathbf{h}) \triangleq \frac{1}{k}\sum_{r=1}^{k}\frac{1}{m}\sum_{j\in I_r^c} J\left(y_j, \widehat{\mu}_{n-m}(\mathbf{x}_j, \mathbf{h}|I_r)\right)$$

where $J(\cdot,\cdot)$ denotes an evaluation metric such as MSE or MAE, I_r denotes the train observations on fold $r = 1, \ldots, k$ while I_r^c denotes the complement, thus the test observations. We assume $|I_r| = n - m$ and $|I_r^c| = m$, hence the total sample size is n. In principle, we choose the bandwidths to minimize

$$\min_{\mathbf{h}} CV_k^J(\mathbf{h})$$

The literature's practice is to use *leave-one-out* cross validation ($k = n$, $m = 1$) with $J(\cdot) = \|\cdot\|_2^2$ as a default. Thus making the above formula become

$$CV_n^2(\mathbf{h}) = \frac{1}{n}\sum_{r=1}^{n}\left(y_r, \widehat{\mu}_{n-1}\left(\mathbf{x}_r, \mathbf{h}|I_r^{loo}\right)\right)^2$$

where I_r^{loo} simply becomes $\{1, \ldots, n\} \setminus \{r\}$ where *loo* stands for *leave-one-out*. In other words, we are estimating our kernel regression with all the data except one

observation and testing the predictions against it. The reason why this estimator is popular is because it has a simplifying formula that avoids estimating a new regression for each of the n folds. To derive this simplification first define the following set of numbers

$$L_r \triangleq \frac{K(\mathbf{0})}{\sum_{j=1}^{n} K\left((\mathbf{x}_r - \mathbf{x}_j) \odot \mathbf{h}^{-1}\right)}, \text{ for } r \in \{1, \ldots, n\}$$

hence the *leave-one-out* cross validation formula as shown in (ANP, 2007) becomes

$$CV_n^2(\mathbf{h}) = \frac{1}{n} \sum_{r=1}^{n} \left(\frac{y_r - \widehat{\mu}_n(\mathbf{x}_r, \mathbf{h})}{1 - L_r} \right)^2$$

where it is only needed to estimate a kernel regression once with all the data and then adjust with L_i. A general proof for the above result can be found in (Hastie et. al., 2009). The computational shortcut does come at a cost. When comparing $CV_n^J(\mathbf{h})$ versus $CV_k^J(\mathbf{h})$, the former has a lower variance but a higher bias. The reason is that all the kernel regression estimates do not vary much; they are computed with almost the same information. Also, note that if the estimates do not change on each fold, then the evaluation used above would highly penalize outliers. This, heuristically, makes the bandwidths smaller.

To address the previous concerns, we propose the following cross-validation strategy

$$CV_k^1(\mathbf{h}) = \frac{1}{k} \sum_{r=1}^{k} \frac{1}{m} \sum_{j \in I_r^c} |y_j, \widehat{\mu}_{n-m}(\mathbf{x}_j, \mathbf{h} | I_r)|$$

where k ranges from 3 to 5 and the evaluation metric used is $J = \|\cdot\|_1$. Moreover, the train/test split is 70/30%. The extra cost of this approach comes from having to estimate k times a kernel regression with $\lceil n(0.7) \rceil$ observations which is $O(k \lceil n(0.7) \rceil)$. In contrast, *leave-one-out* estimates one kernel regression with the full data set, thus $O(n)$. Therefore, the proposed approach increases the number of computations in each iteration by

$$O([k(0.7) - 1] n)$$

which is linear on the number of observations. These are the costs but what are the benefits? Note that, as mentioned previously, *leave-one-out* (or CV_n^2) is a heavily biased estimate. Thus, it becomes an attractive choice when n is really

large because, in this scenario, it both reduces the computational burden and its estimating bias. However, in a time series context, we might not have the sufficient information available for this bias to reduce as needed. Note if we have weekly observations then $n = 250$ amounts to 5 years of data but to get $n = 2000$ we will need more than 38 years. Hence when constrained on n, the bias of CV_n^2 would not decrease as rapidly hence making CV_k^1 a more attractive choice. Also in this case the extra computational cost is not a heavy burden.

Besides the different cross-validation strategy, we propose a different numerical optimization implementation for minimizing the bandwidth. Currently, the `npreg` implementation in `R` and the `statsmodels` implementation in `Python` use *derivative-free* algorithms. The `R` implementation uses *Powell's* algorithm and the `Python` implementation uses *Nelder-Mead's* algorithm. Beyond the obvious, which is avoiding the computational costs of estimating derivatives, we have not found a strong motivation for derivative free algorithms. Therefore, the present work opted to numerically optimize CV_k^1 by a non-linear conjugate gradient algorithm (*Fletcher-Reeves Method*) which indeed uses the derivatives of the objective function (Nocedal and Wright, 2006). For the ARI dataset, there was a significant improvement of employing this approach $\left(CV_k^1, CG \text{ alg}\right)$ versus the implementation in `Python` $\left(CV_n^1, NM \text{ alg}\right)$. A comparison plot is shown in the ARI results section.

Dealing with Time Series The convergence proof assumed that the observations where independent and identically distributed. Since the bias term only relied on taking expectations, the results will still be valid for dependent observations. Nonetheless, the variance term would be now affected by covariances. Referring back to the variance term

$$var\left(\sum_{t=1}^{n} \frac{K\left(\left(\mathbf{X}_t - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(Y_t - \mu\left(\mathbf{x}\right)\right)}{n h_1 \cdots h_p}\right) =$$

$$\frac{var\left(K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(Y - \mu\left(\mathbf{x}\right)\right)\right)}{n h_1 \cdots h_p}$$

$$+ 2\sum_{t=1}^{n}\sum_{s>t}^{n} \frac{cov\left(\xi\left(t, \mathbf{x}, \mathbf{h}\right), \xi\left(s, \mathbf{x}, \mathbf{h}\right)\right)}{n^2 h_1^2 \cdots h_p^2}$$

where

$$\xi\left(t, \mathbf{x}, \mathbf{h}\right) \triangleq K\left(\left(\mathbf{X}_t - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(Y_t - \mu\left(\mathbf{x}\right)\right)$$

Using previous results, the variance first term would equal

$$\frac{var\left(K\left(\left(\mathbf{X}-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\left(Y-\mu\left(\mathbf{x}\right)\right)\right)}{nh_1\cdots h_p}=$$

$$=\frac{\epsilon^2\left(\mathbf{x}\right)f\left(\mathbf{x}\right)\omega^p+f\left(\mathbf{x}\right)\rho\sum_{i=1}^p h_i^2\mu_i\left(\mathbf{x}\right)}{nh_1\cdots h_p}+o\left(\left(nh_1\cdots h_p\right)^{-1}\left\|\mathbf{h}\right\|\right)$$

To avoid worsening the convergence rate we would require that the covariance term not only be bounded but

$$2\sum_{t=1}^n\sum_{s>t}^n\frac{cov\left(\xi\left(t,\mathbf{x},\mathbf{h}\right),\xi\left(s,\mathbf{x},\mathbf{h}\right)\right)}{n^2h_1^2\cdots h_p^2}\leq\frac{\phi\left(\mathbf{x},\mathbf{h}\right)}{nh_1\cdots h_p}$$

This will happen with weakly dependent data such that

$$\left|cov\left(\xi\left(t,\mathbf{x},\mathbf{h}\right),\xi\left(s,\mathbf{x},\mathbf{h}\right)\right)\right|\leq\rho\left(s-t\right)var\left(\xi\left(t,\mathbf{x},\mathbf{h}\right)\right)$$

assuming

$$\begin{aligned}\sum_{t=1}^{n-1}\sum_{s>t}^n\rho\left(s-t\right)&=\sum_{t=1}^{n-1}\sum_{j=1}^n\rho\left(j\right)\\&\leq n\sum_{j=1}^\infty\rho\left(j\right)\end{aligned}$$

where the correlations decrease sufficiently in time so that

$$\sum_{j=1}^\infty\rho\left(j\right)<+\infty$$

Eliminating Design Bias For simplicity, we will now switch to the univariate case. Note how we can derive the Nadaraya-Watson regression estimate by solving the following problem

$$\min_a\sum_{i=1}^n w_i\left(x_i,x,h\right)\left(y_i-a\right)^2$$

where

$$w_i(x_i, x, h) \stackrel{\circ}{=} \frac{K\left(\left(x_i - x\right)h^{-1}\right)}{\sum_{j=1}^{n} K\left(\left(x_j - x\right)h^{-1}\right)}$$

the first order conditions for a imply

$$[a] : 2\sum_{i=1}^{n} w_i(x_i, x, h)\left(y_i - \widehat{a}(x)\right)(-1) = 0$$

solving for a yields

$$\begin{aligned}\widehat{a}(x, h|\mathbf{D}) &= \sum_{i=1}^{n} w_i y_i \\ &= \sum_{i=1}^{n} \frac{K\left(\left(x_i - x\right)h^{-1}\right) y_i}{\sum_{j=1}^{n} K\left(\left(x_j - x\right)h^{-1}\right)} = \widehat{\mu}_n(x, h|\mathbf{D})\end{aligned}$$

where $\mathbf{D}$ denotes conditioning on $X_1 = x_1, \ldots, X_n = x_n$. Therefore, the Nadaraya-Watson estimate is equivalent to finding a constant that minimizes the MSE. Additionally, since the kernel assigns higher weight to values close to the target x, or all of it if the kernel has a compact support, hence, the regression estimate can be interpreted as finding an optimal *local constant* which does vary depending on x. Thus, the central question of this section becomes: is there another local approach that might be better than simply estimating a constant?

Before proposing other local approaches. Let's recover the results from the last section but through a different alley. For the univariate case, the expected bias of $\widehat{\mu}(x)$ is

$$E\left[\widehat{\mu}_n(x) - \mu(x)\right] = \frac{h^2\kappa}{2}\left[\frac{2f_1(x)\mu_1}{f(x)} + \mu_{11}(x)\right] + o\left(h^2\right)$$

This result could be obtained by following the next steps.

$$E\left[\widehat{\mu}_n(x) - \mu(x)\,|X_1 = x_1, \ldots, X_n = x_n\right] \approx E\left[\left(\widehat{\mu}_n(x) - \mu(x)\right)\frac{f_n(x)}{f(x)}|\mathbf{D}\right]$$

this past expectation is

$$E\left[\sum_{i=1}^{n} \frac{K\left(\left(x_i - x\right)h^{-1}\right)\left(y_i - \mu(x)\right)}{nhf(x)}|\mathbf{D}\right]$$

since $y_i = \mu(x_i) + \epsilon_i$ and $E[\epsilon_i | X_1 = x_1, \ldots, X_n = x_n] = 0$. The previous expectation becomes

$$\sum_{i=1}^{n} \frac{K\left(\left(x_i - x\right) h^{-1}\right)\left(\mu(x_i) - \mu(x)\right)}{nhf(x)}$$

Taylor expanding each $\mu(x_i) - \mu(x)$ term around x up to a second order results in

$$\begin{aligned} &= \mu_1(x) \sum_{i=1}^{n} \frac{K\left(\left(x_i - x\right) h^{-1}\right)\left(x_i - x\right)}{nhf(x)} \\ &+ \frac{\mu_{11}(x)}{2} \sum_{i=1}^{n} \frac{K\left(\left(x_i - x\right) h^{-1}\right)\left(x_i - x\right)^2}{nhf(x)} + o\left(h^2\right) \end{aligned}$$

By the law of total expectation

$$E\left[\widehat{\mu}_n(x) - \mu(x)\right] = E\left[E\left[\widehat{\mu}_n(x) - \mu(x) | \mathbf{D}\right]\right].$$

Hence if we take expectations from the Taylor expansion above we have

$$\begin{aligned} =& \frac{\mu_1(x)}{f(x)} E\left[\sum_{i=1}^{n} \frac{K\left(\left(X_i - x\right) h^{-1}\right)\left(X_i - x\right)}{nh}\right] \\ &+ \frac{\mu_{11}(x)}{2f(x)} E\left[\sum_{i=1}^{n} \frac{K\left(\left(X_i - x\right) h^{-1}\right)\left(X_i - x\right)^2}{nh}\right] + o\left(h^2\right). \end{aligned}$$

By the last two propositions from the Appendix, the above expressions is equivalent to

$$E\left[\widehat{\mu}_n(x) - \mu(x)\right] \approx \frac{h^2 \kappa}{2}\left[\frac{2f_1(x)\mu_1(x)}{f(x)} + \mu_{11}(x)\right] + o\left(h^2\right)$$

which is the exact same result we got previously. Note that the only term that depends on the density is

$$\frac{2f_1(x)\mu_1(x)}{f(x)}$$

which is called, by the same reason, the *design bias*. Relating back to the central question of this section, it is quite interesting to note that if we estimate locally a line rather than a constant we eliminate this bias.

To see this, lets pose the following problem

$$\min_{a,b} \sum_{i=1}^{n} w_i \left(x_i, x, h\right) \left(y_i - a - b x_i\right)^2$$

which is the classical weighted least squares problem but where now the weights depend on x. Let's denote the estimated line at each point as

$$\ell_i \left(x, h | \mathbf{D}\right) \stackrel{\circ}{=} \left[(1, x)^T \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} y \right]_i$$

and therefore our new regression estimate becomes

$$\widehat{\mu}_n^{ll} \left(x, h | \mathbf{D}\right) \stackrel{\circ}{=} \sum_{i=1}^{n} \ell_i \left(x, h | \mathbf{D}\right) y_i$$

where *ll* denotes *locally linear*. Performing the same bias analysis as before

$$E\left[\left(\widehat{\mu}_n^{ll} \left(x\right) - \mu \left(x\right) \right) | \mathbf{D} \right] = \sum_{i=1}^{n} \ell \left(x_i, x, h\right) \left(\mu \left(x_i\right) - \mu \left(x\right) \right)$$

since $y_i = \mu \left(x_i\right) + \epsilon_i$ and $E\left[\epsilon_i | X_1 = x_1, \ldots, X_n = x_n\right] = 0$. Again Taylor expanding each $\mu \left(x_i\right) - \mu \left(x\right)$ around x up to a second term we have

$$= \mu_1 \left(x\right) \sum_{i=1}^{n} \ell \left(x_i, x, h\right) \left(x_i - x\right) + \frac{\mu_{11} \left(x\right)}{2} \sum_{i=1}^{n} \ell \left(x_i, x, h\right) \left(x_i - x\right)^2 + o\left(h^2\right)$$

however due to the first order condition $\sum_{i=1}^{n} \ell \left(x_i, x, h\right) \left(x_i - x\right) = 0$ therefore the above term becomes

$$E\left[\left(\widehat{\mu}_n^{ll} \left(x\right) - \mu \left(x\right) \right) | \mathbf{D} \right] \approx \frac{\mu_{11} \left(x\right)}{2} \sum_{i=1}^{n} \ell \left(x_i, x, h\right) \left(x_i - x\right)^2 + o\left(h^2\right)$$

which mitigates the effect of the first order derivatives.

The above results can be expanded to *local polynomials*. However, as always this will imply a wild increase in computational time and variance which could be actually counterproductive.

3.3 What approach to choose?

To set some context for this section, we need to understand the consequences of the *curse of dimensionality*[3] for local methods such as the kernel regression (Hastie et. al., 2009). This phenomenon states that when adding more predictors, a fixed sized neighborhood of a training point becomes increasingly sparser; thus, it needs to exponentially increase the size of its neighborhood to maintain the same amount of members as before. In our context, given a new data point, the kernel regression will weigh more heavily points that are actually not close enough to yield reasonable predictions, therefore leading to a loss in accuracy.

As seen above, the MSE convergence rate

$$E\left[MSE\left(\widehat{\mu}_n\left(\mathbf{X}\right),\mu\left(\mathbf{X}\right)\right)\right]=O\left(n^{\frac{-4}{p+4}}\right)$$

massively increases on the number of predictors. Thus, the considerable drawback of the kernel regression is its inability to deal with several predictors. Therefore, in this case the linear regression approach is preferable since it will at least be able to extract meaningful information from all the predictors. However, if only a few relevant variables are included and there are sufficient observations, then the best modelling alternative is the kernel regression. This is the case since it is guaranteed to converge to whatever function might the $E\left[y_t|y_{t-1},\ldots,y_{t-p},g_t\right]$ follow without requiring any assumptions. Another significant point to consider is computational time. Contrary to the linear regression, the kernel regression needs to numerically optimize the choice of bandwidths which takes more time than solving the closed-form solution to the least squares problem.

As a final comment for this section, it is worth noting that there are model approaches that lie in between the rigidness of the parametric linear model and the full flexibility of the nonparametric approach. These approaches should also be incorporated as standard alternatives as well. One of such approaches would be the *generalized additive linear model (GAM)*. This model approximates the mean response by

$$E\left[y_t|y_{t-1},\ldots,y_{t-p},g_t\right]\approx\beta_0+\sum_{l=1}^{p}\beta_l s_l\left(y_{t-l}\right)+\beta_{p+1}s_{p+1}\left(g_t\right)$$

[3]The expression is coined to Bellman (1961). This curse refers to the phenomena that arise when analyzing high-dimensional spaces that do not appear in low-dimensional spaces.

where $s_i(\cdot)$ denotes a spline function. It is similar to the linear model in the sense that the effect of each variable is treated independently. It differs in that now non-linear relationships with the response are estimated for each predictor. Likewise, the computational estimation for GAMs is closer to the linear model since it requires to iteratively solve least-square problems. This approach was also used for each of the analysis throughout this work. Nonetheless, there was no significant improvement from the linear model and thus they were not included in the results. Our hypothesis for this consequence is twofold. Either the relationships are relatively linear so there was no extra benefit of including more flexibility. Or that to further improve the predictions, it is needed to consider interaction terms. This last point is the strongest hypothesis of why the kernel regression approach would notably improve the results of the ARI analysis while the GAM approach did not.

3.4 Variable Selection

3.4.1 Selecting Relevant Search Terms

Selecting *relevant* search terms from a myriad of possibilities is probably the most important question raised when incorporating *Google Trend*'s data into an analysis. Thankfully, this heavily loading is done by the *Google Correlate* algorithm. As stated in the data section, this algorithm takes as input a given time series and returns the top 100 search terms that most correlate with it.

Once the list of the top 100 terms is provided, the next question is which of those search terms to incorporate. It would not be wise to include all of them since we might not have sufficient observations to sustain that number of predictors or, most certainty, some of them would be highly correlated between them and thus make the parameter estimation unreliable. An initial heuristic manner to trim the list is to first dispose all the spurious search terms that got in and then to cluster the remaining terms in groups. A great example for this procedure is the unemployment case. In that analysis, several job searching websites where suggested by *Google Correlate*. Hence rather than incorporating one by one to the analysis, the `Employment - Topic` aggregates the activity of all of them and this was used as a predictor.

A more automated procedure would be to use variable selection algorithms such as *LASSO* or *Spike-and-Slab*. (Murphy, 2012). However, these off-the-shelf procedures are not guarantee to work with highly correlated data. For example, *LASSO* is guaranteed to converge to the true model when the predictors

are sufficiently uncorrelated (Hastie et. al., 2009). To overcome this, there is a model called *Bayesian Structural Time Series* (*BSTS*) which was actually developed to select predictive Google search terms. This algorithm provides a posterior probability that a search term might belong to the underlying model. In general, this approach combines the *Spike-and-Slab* variable selection algorithm with other time series specific treatments (such as the *Kalman Filter*). For the full details of this procedure read (Varian and Scott, 2013).

3.4.2 Deciding what to include in the model

Once the relevant search terms are found and other predictors are selected, it is now time to actually choose what to include in the models. Both from the predictors available but also from their lags. The standard approach for variable selection in a linear model is to use *LASSO*. However, as declared previously, it has not guaranteed to work with highly correlated data (Hastie et. al., 2009). Furthermore, it cannot be applied to the kernel regression. Consequently, as brute force as it might appear, the variable selection procedure used throughout this work was *Best Subset*.

Best Subset grabs all the subsets that p predictors might generate, tests each of them and then selects the best performing one. This algorithm is not efficient since the number of subsets grows exponentially to the order of 2^p. Nonetheless, the procedure yields the optimal and for the context of this work it was still manageable. Testing $2^{15} = 32,768$ linear regressions in a personal computer takes approximately 1 minute for the ARI weekly data ($N = 218$). Additionally, this algorithm can be employed for the kernel regression. Due to the *curse of dimensionality*, and given the size of the training data, it was only needed to test at most 4 predictors. However, they were selected from a bag of 20 different predictors which added to

$$\binom{20}{1} + \binom{20}{2} + \binom{20}{3} + \binom{20}{4} = 6,192$$

different model propositions (this took roughly 1 hour). There are more sophisticated variable selection algorithms for the kernel regression as seen in (Dudek, 2012). There they use reinforcement learning to create a tournament where the features compete to be included. Nonetheless, the author of this work falls short on those topics.

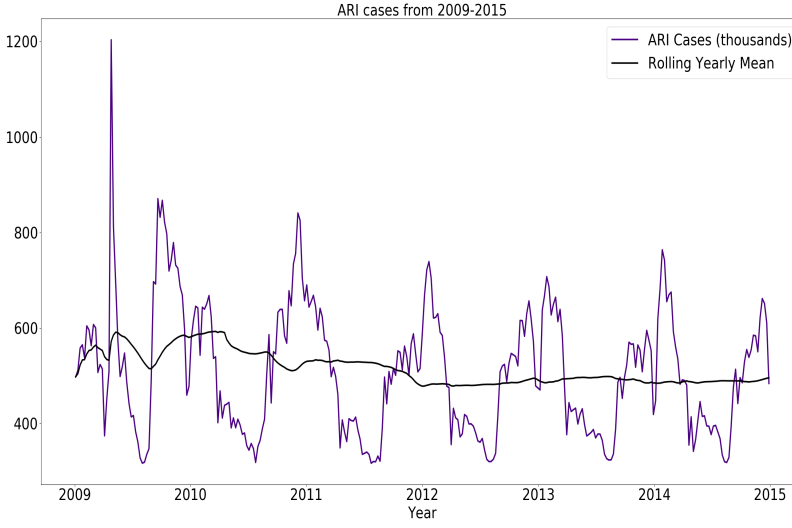
4 Results

In this section it will be shown the quantitative and qualitative results for each of the three analyses (the ARI cases, the unemployment rate and the homicides committed). The overall structure of the exposition is the following. First, a descriptive exploration of the data will be performed. This will set the relevant context for each analysis. The goal is to show trends present in the data, to discuss any unusual events and to exhibit correlation between the search terms and the target times series. Second, the test predictions will be evaluated against the metrics motivated in the past section. This will provide a quantitative improvement of incorporating the Google's search terms as predictors. Third, the *test residuals* of each model will be examined and compared against the baseline model residuals. The ideal behavior for the residuals is to follow a symmetric distribution, with light tails and centered at zero with the lowest variance possible (similar to a normal distribution). Symmetry reveals that the model is not systematically over or under estimating the target variable. The lightness of the tails indicates that the predictions do not make severe mistakes (for example, that the target increased 10% and the prediction was a 10% decrease). Moreover, the low variance concentration at zero is a reflection of the model's accuracy. Finally, a study of the *prediction versus actuals plots* will uncover qualitative details bypassed by the previous tests. For example, these plots will expose if the highest errors are done at the end of the test sample which signals a loss in predictive power of the model.

4.1 ARI Cases Results

4.1.1 Data Exploration

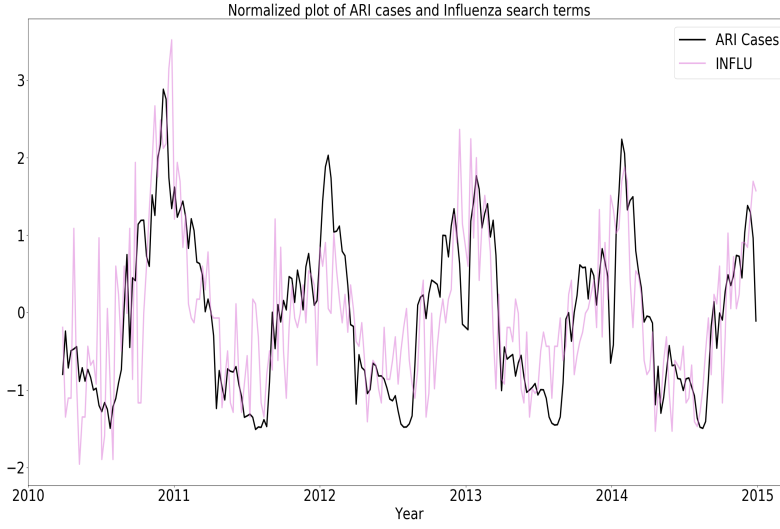
The Acute Respiratory Infections in Mexico follow the next seasonal behavior from 2009 up to 2015



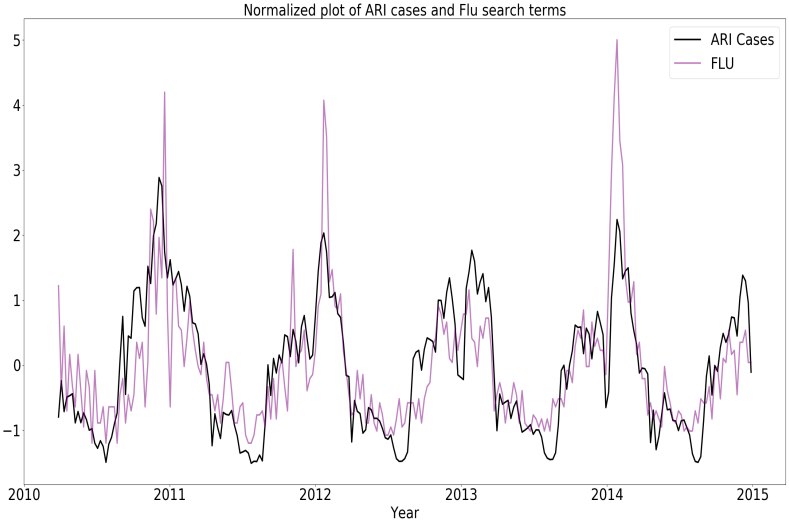
Source : ENSANUT

The ARI time series possesses two distinctive characteristics: it has strong seasonal effects but is nonetheless stationary. As seen from the constant value of the rolling mean, the seasonal highs are compensated by the lows. In terms of seasonality, every year, there is a continuous increase of cases on winter which ultimately peaks around December - February. However, past that time, the number of cases sharply plummets, reaching its lowest values during summer time. The outlier present in 2009 is due to the influenza pandemic in Mexico, which explains the over a 100% increase from the yearly mean. On a good note, since the ARI time series is in absolute numbers. Then since roughly 2011 there has been a continuous decrease in per capita cases.

As mentioned in the data section, the *Google Trends* search terms for this analysis fall into two categories. The ones related with influenza topics ($INFLU$) and the ones related with the common flu and its medications (FLU). Below is a graph that plots together the normalized ARI cases with the normalized values of the Google Indexes; where both series are normalized to have mean zero and standard deviation of one. The idea is to visually see how the movements between the time series follow each other. It is not necessary that the time series have the same values but rather that they share the same valleys and peaks.



Source : ENSANUT and Google Trends



Source : ENSANUT and Google Trends

Regarding both plots, it can be seen that in terms of seasonal movements, both influenza and flu search activity follows closely the ARI time series. However, there is a large amount of noise and volatility present in the *Google Trends* data. A clear example being the tremendous winter surges of flu activity which does not necessarily translate in as many cases. Or the continuous ups and downs from week to week on the influenza time series.

4.1.2 Test Evaluation Results

The following table summarizes the absolute and percentage improvements in evaluation error of incorporating *Google Trends* indexes as predictors. *LM* denotes Linear Model whereas *KM* denotes Kernel Model. Also *wo GT* implies that the model is without the *Google Trends* predictor while *w GT* implies that the model includes it [4].

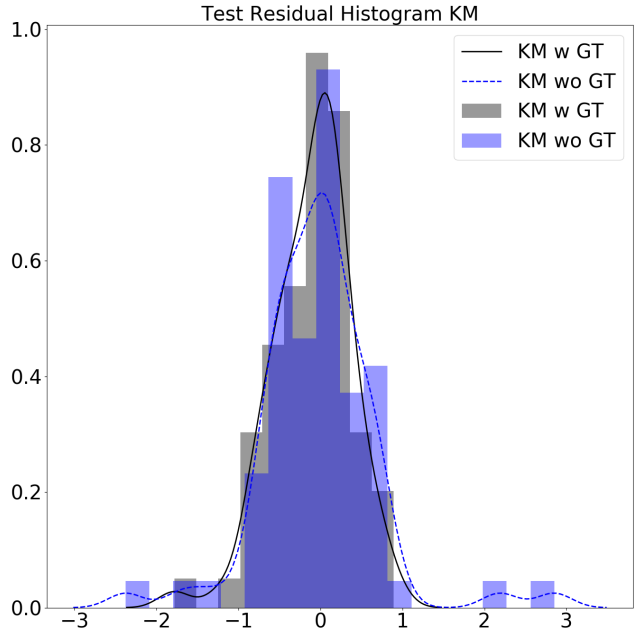
[4]The set of predictors for each model is the following. For LM wo GT the predictors are y_{t-4} and y_{t-12}. For LM w GT: y_{t-4}, y_{t-12}, FLU_t and $INFLU_t$. For KM wo GT: y_{t-4} and y_{t-12} . For KM w GT: y_{t-4}, y_{t-12} and FLU_t. Note that these were the predictors selected by best-subset using MAE as the evaluation metric.

Models	RMSE	%	MAE	%
LM wo GT	0.74		0.57	
KM wo GT	0.70	5	0.47	18
LM w GT	0.48	35	0.39	32
KM w GT	0.47	36	0.35	39

In this scenario, there is an enormous improvement of over 30% for introducing the *Google Trends* indexes into either the linear autoregressive model or the kernel autoregressive model as predictors. Specially in terms of MAE, the best model is able to cut down the baseline by 40%. Regarding the first two models (which do not use Google's search terms) there is a noticeable improvement from moving into the nonparametric approach. The benefit in MAE is higher than in RMSE since the last is driven by outlier mistakes that both the nonparametric model and the baseline model commit. Nonetheless, the model improvements cannot outweigh the benefits of including relevant predictors. With the addition of the Google Trend's data, the linear autoregressive model is able to surpass the initial nonparametric approach. However, the best results are obtained after combining both benefits: introducing the Google's predictors and moving into a nonparametric approach. In terms of MAE, there is a 7% improvement by moving into the nonparametric approach. It is not a large improvement as before since the Google Trend's predictors may not have such a non-linear relationship with the response that the kernel regression can take advantage.

4.1.3 Test Residuals Examination

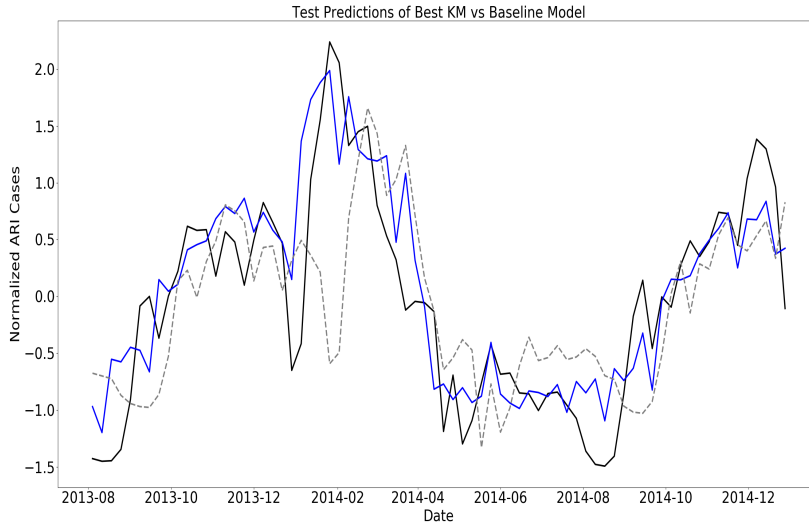
Below is the histogram of the test residuals for the nonparametric models with and without the Google Trend's information. They will be denoted as KM_{Base} and KM_{GT} respectively. We do not compare the nonparametric approach with the baseline model to isolate the benefits of including the search terms as predictors from the effects that result from improving the model specification.



Source : *ENSANUT and Google Trends*

There is a significant improvement from incorporating the Google Trend's predictors. The greatest benefit comes from the KM_{GT} model not committing the large outlier mistakes that the KM_{Base} model does; this translates into the lighter tails from the estimated density of KM_{GT}. Also, related to the previous observation, the residuals are more concentrated around zero. However, there are no benefits in terms of symmetry; even when committing outlier mistakes the KM_{Base} model is not biased to under or over estimate.

4.1.4 Prediction versus Actuals

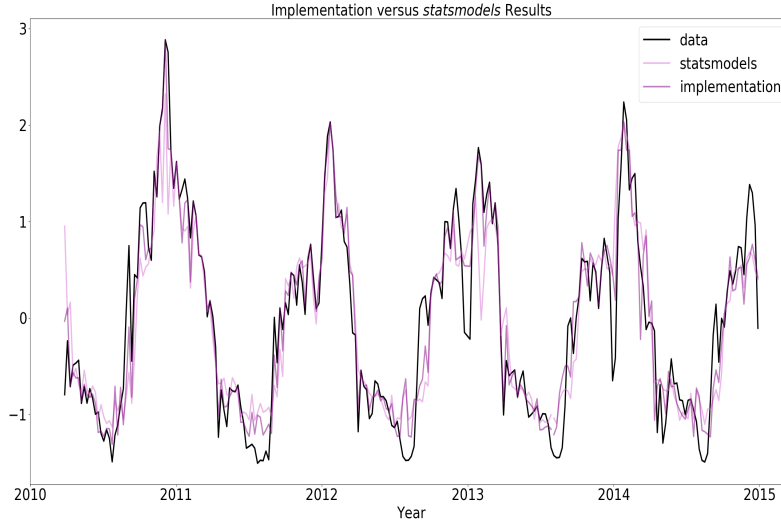


Source : ENSANUT and Google Trends

In general, it is evident that the kernel regression model follows more tightly the actual values than the baseline model. Both models depend on previous lags, however, since the baseline model does not have any other predictor that might indicate a possible trend switch it makes enormous mistakes as seen in 2014-02 and 2014-09 where it essentially repeats previous periods movements. Contrary to this, the Google Index helps the kernel regression correctly anticipate the rise of activity on both 2014-02 and 2014-09. This is essentially the benefits of *nowcasting* with Google's searches: anticipating trend changes when they start to occur.

4.1.5 Implementation Improvements

Below are two graphs that compare different implementations of the kernel regression. The first plot shows our proposed implementation (which was discussed in the past section). The second shows the results from the `statsmodels` implementation in `Python`.

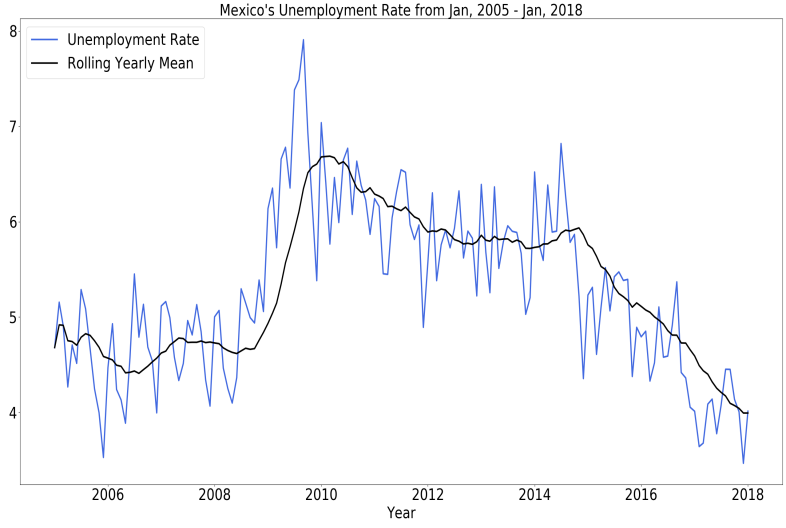


As it can be seen above, our implementation performs significantly better in the train sample. It is able to hit the peaks of 2011 and, notably, 2013. Whereas, for that period, the `statsmodels` implementation moves in the opposite direction. In terms of the test sample, our implementation performs only slightly better. For example, it is able to approximate better the valley at the end of 2014.

4.2 Unemployment Rate Results

4.2.1 Data Exploration

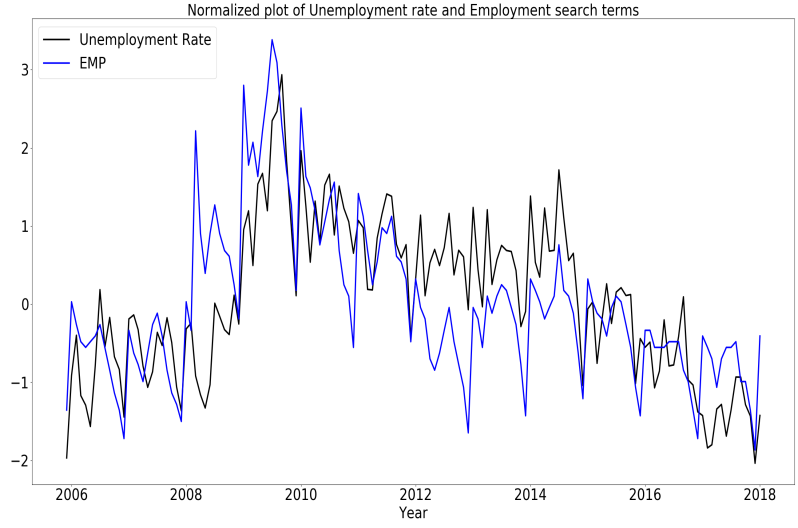
The Mexican urban monthly unemployment rate has the following behavior



Source : INEGI, BIE Desempleo Urbano

Through 2005 and 2009 the unemployment rate appeared to be in a state where its yearly mean value oscillated close to 4.5% (with peaks and valleys that compensated themselves). The sharp increased experience during 2009-2010 is a reflection of the 2008 American financial crisis which took some time to materialize in the Mexican market. This sharp increase set the economy from 2011 up to 2015 in a state whose yearly mean value was roughly over 6%. From 2015 to 2017 the labor economy has experienced a consistent recovery over time reaching historically low levels in the end of 2017.

As stated in the data section the *Google Trends* data for this analysis contains all the search terms included in the `Employment - Topic`. The most relevant search terms for this topic being: `empleo`, `portal empleo`, `occ`, `bolsa de trabajo`, `computrabajo` and `empleo.gob`. Below is a graph that compared the normalized unemployment rate time series versus the normalized *Google Trends* data. Both time series are normalized to have mean zero and standard deviation of one. This is done so that time series could be plotted together.



Source : INEGI, BIE Desempleo Urbano and Google Trends

What is most important from the plot above is that the time series share the same valleys and peaks rather than to necessarily follow themselves closely. For example, throughout 2012-2015 the Google Index is consistently above the unemployment rate; nonetheless, they mostly move in the same directions. As it can be seen above, the *Google Trends* index is not perfect. There are periods where the time series follow themselves closely such as from 2006-2008, 2009-2012, but for some other periods they have large discrepancies like in 2009 where the unemployment rate sharply increases but the search activity sharply decreases. This happens in 2017 but at a lesser degree. Despite this, the favorable cases are more prevalent and hence the Google Trend's index is a valuable predictor as it will be seen below.

4.2.2 Test Evaluation Results

The following table summarizes the improvements (as well as its percentage decrease in evaluation error) of incorporating the *Google Trends* indexes as predictors into the best linear autoregressive model. Again, LM denotes Linear Model whereas KM denotes Kernel Model. Also *wo* GT implies that the model is with-

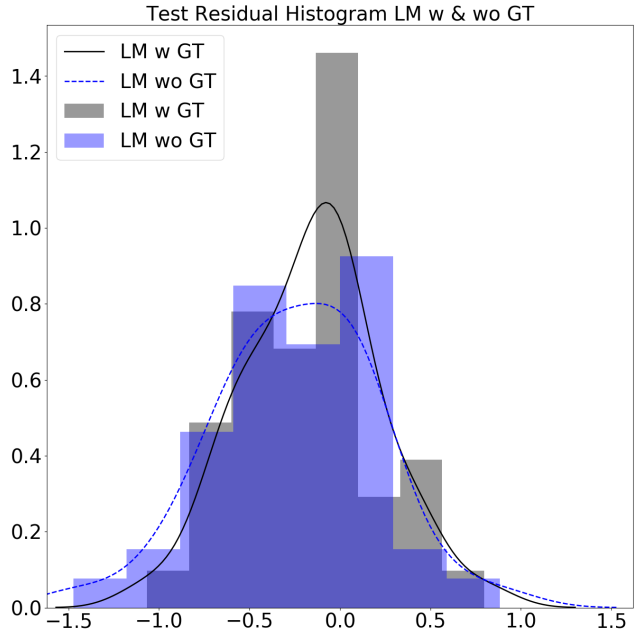
out the *Google Trends* predictor while *w GT* implies that the model includes it [5].

Models	RMSE	%	MAE	%
LM wo GT	0.5056		0.3956	
LM w GT	0.4031	20	0.3095	20

4.2.3 Test Residuals Examination

Below is the histogram of the test residuals for both the linear autoregressive model that incorporates the *Google Trends* data and the linear autoregressive baseline model without this data. For this discussion, let's denote the first model as LM_{GT} and the baseline model as LM_{Base}.

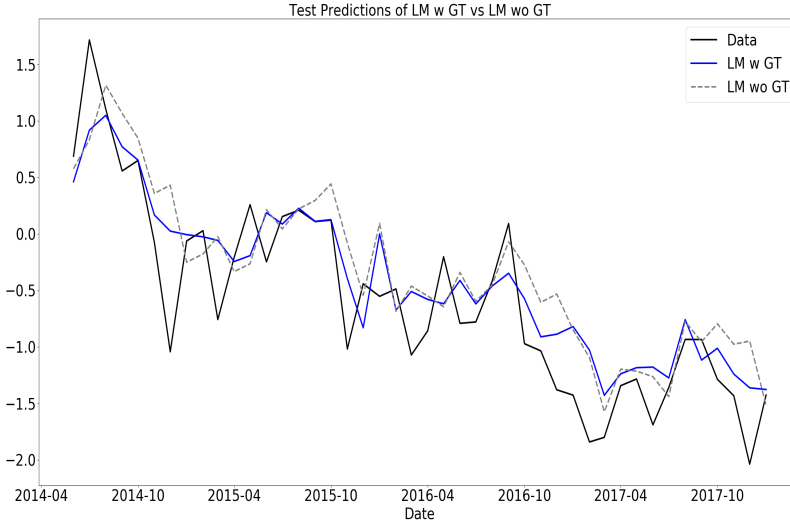
[5]The set of predictors for each model is the following. For LM wo GT: y_{t-1}, y_{t-2}, y_{t-3}, y_{t-9}, y_{t-10} and y_{t-11}. For LM w GT: y_{t-1}, y_{t-2}, y_{t-3}, y_{t-9}, y_{t-10}, y_{t-11} and EMP_t. These were the lags selected by best subset using MAE as the evaluation metric.



Source : INEGI, BIE Desempleo Urbano and Google Trends

There is a strong benefit of incorporating the GI as a predictor. In terms of symmetry, is appears as if the residuals from LM_{Base} are biased to the left, which indicates a systematic underestimation of this model. The residuals from LM_{GT} almost fix this bias however retaining a slight tilt of underestimation. This scenario suggests that there is a predictor missing in LM_{Base} that is correlated with the search terms in LM_{GT} that resolve the underestimation issue. Moreover, including the GI's also makes the tails from the estimated density converge earlier than the tails from the estimated density for LM_{Base}. Finally, the residuals from the LM_{GT} are visibly more concentrated at zero.

4.2.4 Prediction versus Actuals



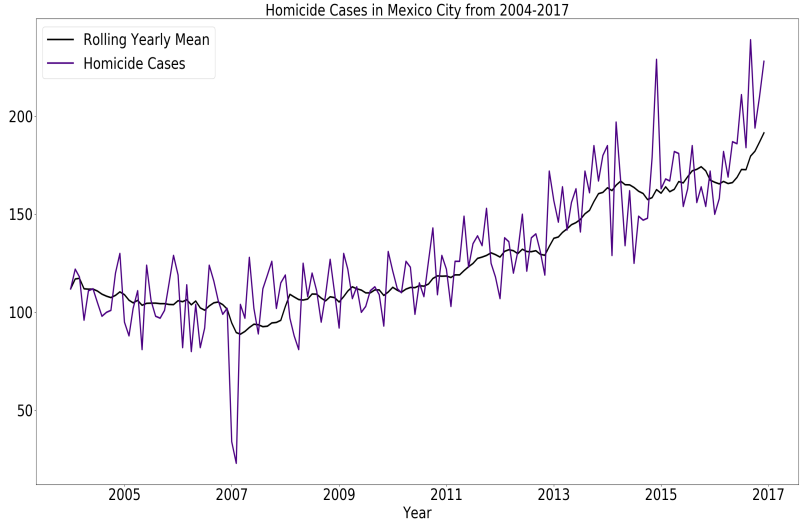
Source : INEGI, BIE Desempleo Urbano and Google Trends

In general, it is evident that the actual unemployment rate has a higher variance that the linear models are not able to predict or follow immediately. However, the linear model that incorporates the *Google Trends* data gains more agility to pivot either upwards or downwards with the data. For example, around 2016-06 the blue line is able to decrease more abruptly than the grey dashed line. Again, after the beginning of 2017, the blue line is able to shift more sharply into an increasing trend and finally, around 2017-06 when the grey line starts to catch up, the blue line again is able to drop faster for the last periods.

4.3 Homicide Cases Results

4.3.1 Data Exploration

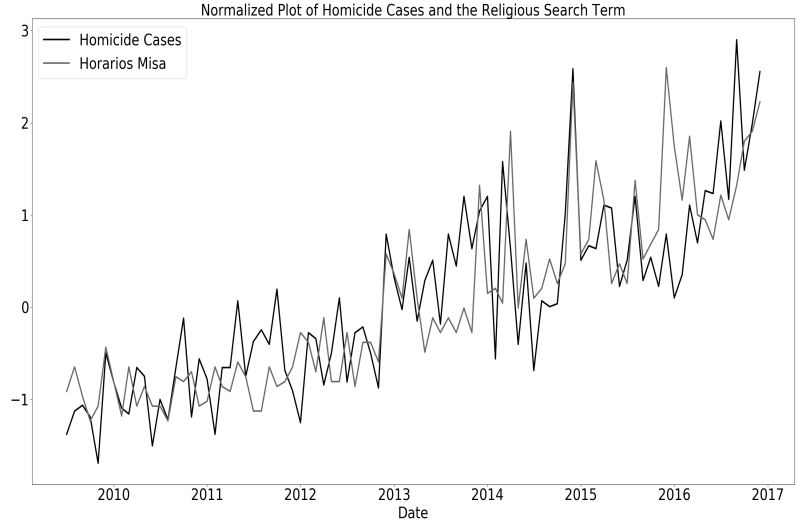
The homicide cases for Mexico City has the subsequent pattern



Source : INEGI Defunciones Generales

From 2004 up until closely 2010 the homicide yearly rolling mean had a value under 110 (definitely, 2007 is a data error). However, right after 2011 an increasing trend starts to gain momentum. It is understandable that a time series such as this would bear an increasing trend due to the population growth and city migration. Nonetheless, there is a close to a 7% yearly growth in the homicide activity. This by far surpasses any natural growth, especially since this time series does not contain any information of Estado de Mexico (a surrounding neighboring state that carries a high criminal activity). We have not read any plausible hypothesis of why this is happening and have intentionally left any discussions about this topic out of the current work. The spikes and valleys do not follow a seasonal pattern. In some years the activity concentrates from July to October, but on some other years from January to April. Overall, we see a sharply increasing trend that has increased variance over time that does not possess any seasonal patterns.

As mentioned previously in the data section here the *Google Trends* data contains a unique search term `horarios misa` which translates to church mass schedules. Below is the well-known graph that compares the normalized time series of both the homicide cases and the search term



Source : INEGI Defunciones Generales and Google Trends

The above plot does not follow the tight pattern perceived in the previous analysis. However, from 2013 to the end of 2015 the time frames chase each other. Yet again, from 2016 until 2017 the search term appears to lose predictive power, with a huge false positive on December 2015. It is worth pointing out that since this data has a 1 year reporting lag, then the effort of finding a predictor search term is really valuable.

4.3.2 Test Evaluation Results

The following table summarizes the absolute and percentage improvements in evaluation error of incorporating *Google Trends* indexes. *LM* denotes Linear Model whereas *KM* denotes Kernel Model. Also *wo GT* implies that the model is without the *Google Trends* predictor while *w GT* implies that the model includes it [6].

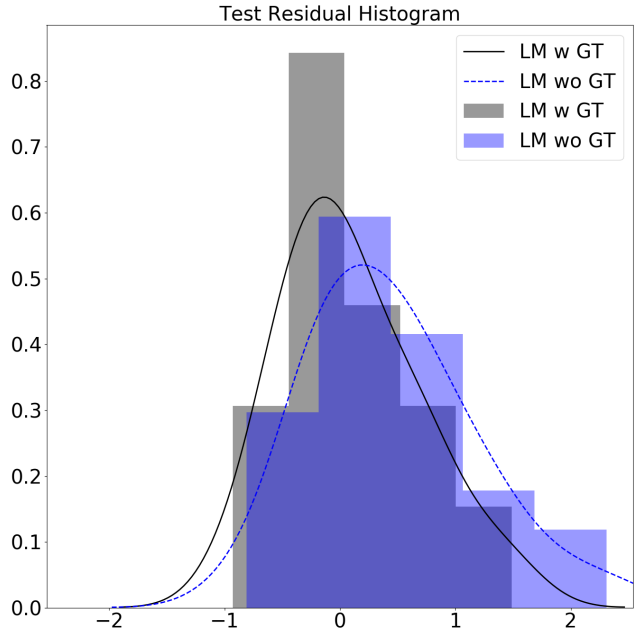
[6] The set of predictors for each model is the following. For LM wo GT: y_{t-1}, y_{t-2} and y_{t-6}. For LM w GT: y_{t-1}, y_{t-2}, y_{t-6} and GI_t. Where GI stands for Google Index. Note that these were the predictors selected by best-subset when using MAE as evaluation metric.

Models	$RMSE$	%	MAE	%
LM wo GT	0.83		0.62	
LM w GT	0.59	29	0.46	26

From the table above, it appears as if there was a large improvement of incorporating the Google Trend Index as predictor. Notwithstanding, the improvement is actually a result of the poor performance of the baseline, not really on the splendid performance of the last model. A MAE of 0.62 is clearly the worst performing baseline when compared with the other models. Thus even adding a rough predictor would improve the model. Yet, the question of whether the search term adds predictive value does find positive evidence from the results above.

4.3.3 Test Residuals Examination

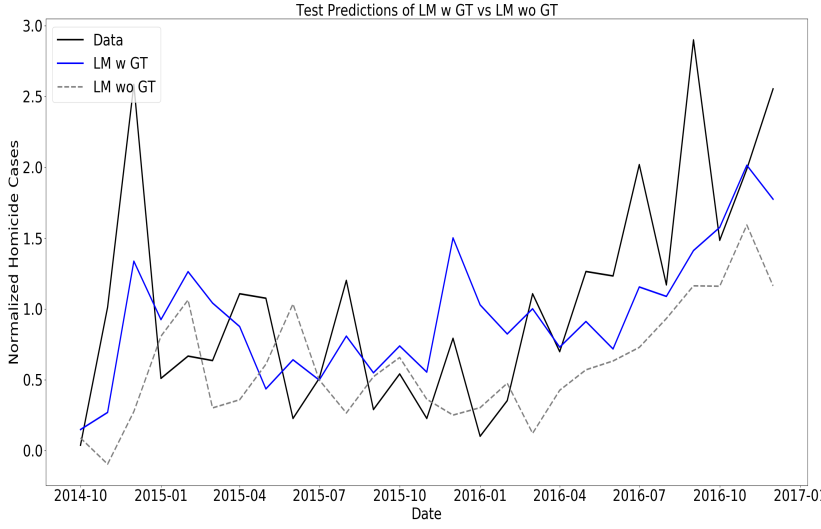
Below is the histogram of the test residuals for both the linear autoregressive model including the search term (LM_{GT}) and the baseline model without (LM_{Base}).



Source : *INEGI Defunciones Generales and Google Trends*

As it can be seen above there is a massive qualitative improvement of adding the search term as predictor. In terms of symmetry, LM_{GT} appears to be unbiased while LM_{Base} is systematically underestimating the actual values. Furthermore, the residuals from LM_{GT} do concentrate more around 0. Nonetheless, the tails of both estimated densities are quite heavy; hinting that the model is not properly is making relevant mistakes and thus has a high variance.

4.3.4 Prediction versus Actuals



Source : INEGI Defunciones Generales and Google Trends

It is undisputable that the blue model loses predictive power at the end of the time frame. However, it does predict a higher activity than the model in grey, thus in this sense, it is more accurate. At least it does predict the overall increase of activity. However, it does not carry the necessary information to predict the huge spikes seen from 2016 and 2017.

5 Conclusions

Relating back to the central question of this work, we find evidence that using internet search queries can help *nowcast* different variables. For all analyzed cases, we find an over 20% improvement in test accuracy when compared to the baseline model[7]. Thus, for these three specific variables, authorities (like *Banxico*[8] or *Secretaria de Salud*) could enrich their predictive models and increase their response time by following the activity of the search terms analyzed. Nonetheless, the predictive power of these search terms needs to be constantly tested. Due to the intrinsic noise and nature of internet activity, it cannot be guaranteed that the search terms will maintain their usefulness overtime. Thus, rather than providing specific search terms for authorities, the present work displays the advantage of incorporating real-time internet search activity into the authorities' toolkit.

Likewise, the present work presents the benefits of employing the kernel regression as a nonparametric model alternative. In the ARI cases, it shows a 7% test improvement over the linear alternative. Also, our multivariate proof exhibits the weaknesses of this nonparametric method. Namely, the dimension constraints and the large information requirements. Finally, we discuss how using a different cross-validation strategy is more suitable for small to medium samples. Regarding implementation, we find better results when optimizing with the conjugate gradient algorithm than with the standard *derivative-free* methods.

Finally, there are two main avenues to further enhance this work. The first avenue relates to implementation improvements. In terms of variable selection, *Best Subset* is not a scalable algorithm. For the linear models there are efficient alternatives such as the *Lasso* regression. However, in the literature there is not a standard variable selection algorithm for the kernel regression and, due to the *curse of dimensionality*, it is of crucial importance. Moreover, neither our numerical strategy nor the standard *derivative-free* algorithms are guaranteed to converge to a global maximum since the objective function is not forced to be convex. Thus, there might be a numerical algorithm that loses generality that leveraging on the structure of the problem guarantees a global optimum. The second avenue relates to the granularity of the analyses performed. For Mexico, the *Google Trends* data comes at a state level and up to an *hour-by-hour* time

[7] It is worth mentioning that this result might be inflated by a weak baseline. Nonetheless, the AR models are the literature's standard.

[8] Mexico's Central Bank

frequency. Therefore, if we trust that are our national predictors do reflect the same reality at a lower geographical level and higher time frequency, then we could gain significant information that is currently not gathered. For example, if we believe that people all around Mexico use the internet when looking for jobs[9], then we could nowcast a weekly state level unemployment rate which is currently not gathered. In principle, even if our *nowcasting* predictors are correct we could never prove this since the actual variable is not reported. However, *INEGI* could execute a *one-off* exercise on a sample of states and at the weekly level to assess how the nowcasting predictors perform.

[9]Let's not forget that states like Oaxaca or Chiapas have a lower internet use and penetration

A Appendix

The Appendix serves two purposes. The first is to remind the reader some useful definitions and classical results (1.1 - 1.3). The second is to present all the essential propositions required to prove that the kernel regression could learn all regression functions (1.4 - 1.7).

A.1 Modes of Convergence

Following are some standard notions of convergence for random variables. Since we are dealing with random samples, we can only guarantee convergence in the terms below.

Definition 1 ***Convergence in Probability***. *Let $(\mathbf{X}_n)_{n=1}^{\infty}$ be a sequence of real random vectors and let $\mathbf{X}$ be a random vector. We say that $\mathbf{X}_n$ converges to $\mathbf{X}$ in probability if, for all $\epsilon > 0$ we have that*

$$P\left(\|\mathbf{X}_n - \mathbf{X}\| < \epsilon\right) \longrightarrow 1, \text{ as } n \to +\infty$$

Definition 2 ***Convergence in r^{th} Mean***. *Let $(\mathbf{X}_n)_{n=1}^{\infty}$ be a sequence of real random vectors and let $\mathbf{X}$ be a random vector with $r > 0$, we say that $\mathbf{X}_n$ converges to $\mathbf{X}$ in r^{th} mean if*

$$E\left[\|\mathbf{X}_n - \mathbf{X}\|^r\right] \longrightarrow 0, \text{ as } n \to +\infty$$

Beneath are some classical results from Measure Theory that will be used in the convergence proof for kernel regression. For a proof of the results below consult (Royden, 2010). These results are contextualized for probability measure spaces.

Definition 3 ***L_p Space***. *For a given $p \in (0, +\infty)$ we say that a Lebesgue measurable function f belongs to L_p if*

$$E\left[|f|^p\right]^{\frac{1}{p}} < +\infty$$

Theorem 4 ***Hölder's Inequality***. *Let W and Z be two real random variables. If $p \in (1, +\infty)$ and $q \in (1, +\infty)$ such that $1/p + 1/q = 1$ then*

$$E\left[|WZ|\right] \leq E\left[|W|^p\right]^{\frac{1}{p}} E\left[|Z|^q\right]^{\frac{1}{q}}$$

Theorem 5 ***Lebesgue's Dominated Convergence Theorem.*** *If $X_n \longrightarrow X$ in probability and if $|X_n| \leq Y$ (almost surely), where $E[Y^r] < +\infty$, then*

$$E[|X_n - X|^r] \to 0 \text{ and } E[X_n^r] \to E[X^r]$$

A.2 Big-O and Little-o Arithmetic's

The Big-O and Little-o notation is quite handy for dealing with upper bounds on approximation theory. Let's define these notions properly.

Definition 1 ***Big-O notation.*** *Let a_n be a sequence. We say that $a_n = O(b_n)$ if there exists a $N \in \mathbb{N}$ and a positive constant C such that*

$$a_n \leq Cb_n, \text{ for all } n \geq N$$

Definition 2 ***Little-o notation.*** *Moreover, we say that $a_n = o(b_n)$ if*

$$\frac{a_n}{b_n} \longrightarrow 0 \text{ as } n \to +\infty$$

To honor the name of the section, below are some useful lemmas that will be used in the following sections. These lemmas are tailored to look as they will appear, nonetheless, the results hold more generally.

Lemma 3 ***Power Lemma.*** *Let $\gamma \in (0, +\infty)$ then*

$$[o(\|\mathbf{h}_n\|^\alpha)]^\gamma = o(\|\mathbf{h}_n\|^{\alpha\gamma})$$

Proof. By continuity

$$\lim_{n+\infty} \frac{a_n^\gamma}{\|\mathbf{h}_n\|^{\alpha\gamma}} = \left(\lim_{n \to +\infty} \frac{a_n}{\|\mathbf{h}_n\|^\alpha} \right)^\gamma = 0$$

■

Lemma 4 ***Multiplication Lemma.*** *Let $\|\mathbf{h}_n\| \to 0$ and α, β be two positive numbers. Then*

$$O(\|\mathbf{h}_n\|^\alpha)\, o\left(\|\mathbf{h}_n\|^\beta\right) = o\left(\|\mathbf{h}\|^{\alpha+\beta}\right)$$

Proof. Let a_n be an arbitrary $O\left(\|\mathbf{h}_n\|^{\alpha}\right)$ sequence and b_n an arbitrary $o\left(\|\mathbf{h}_n\|^{\beta}\right)$ sequence. Then

$$0 \leq \lim_{n\to+\infty} \frac{|a_n b_n|}{\|\mathbf{h}\|^{\alpha+\beta}} \leq \lim_{n\to+\infty} \frac{C_a \|\mathbf{h}\|^{\alpha}}{\|\mathbf{h}\|^{\alpha}} \frac{b_n}{\|\mathbf{h}\|^{\beta}} = 0$$

■

Lemma 5 ***Addition Lemma.*** *Let $\|\mathbf{h}_n\| \to 0$ and α, β be two positive numbers. Then*

$$o\left(\|\mathbf{h}_n\|^{\alpha}\right) \pm o\left(\|\mathbf{h}_n\|^{\beta}\right) = o\left(\|\mathbf{h}_n\|^{\min(\alpha,\beta)}\right)$$

Proof. Let a_n be an arbitrary $o\left(\|\mathbf{h}_n\|^{\alpha}\right)$ sequence and b_n be an arbitrary $o\left(\|\mathbf{h}_n\|^{\beta}\right)$ sequence. Thus,

$$0 \leq \lim_{n\to+\infty} \frac{|a_n + b_n|}{\|\mathbf{h}_n\|^{\min(\alpha.\beta)}} \leq \lim_{n\to+\infty} \frac{|a_n|}{\|\mathbf{h}_n\|^{\alpha}} + \lim_{n\to+\infty} \frac{|b_n|}{\|\mathbf{h}_n\|^{\beta}}$$

■

A.3 Taylor Expansions

One of the most useful results from calculus, both for theoretical proofs and practical applications, is Taylor's Theorem. For us e is a beautiful transcendental number but for the computer it is not but a Taylor expansion of the following sort

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} + R_n, \text{ where } 0 < R_n < \frac{3}{(n+1)!}$$

which can be approximated as precisely as needed. As stated above, it is also a useful theoretical tool that comes handy quite often in statistics.

Theorem 1 ***First-Order Taylor Expansion.*** *Properly, let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be differentiable in an open set U. Then, for $\mathbf{x}_0$ and $\mathbf{x}_0+\mathbf{h} \in U$*

$$\begin{aligned} f(\mathbf{x}_0+\mathbf{h}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T \mathbf{h} + o(\|\mathbf{h}\|) \\ &= f(\mathbf{x}_0) + \sum_{i=1}^{p} h_i f_i(\mathbf{x}_0) + o(\|\mathbf{h}\|) \end{aligned}$$

Theorem 2 ***Second-Order Taylor Expansion.*** *Now let $f(\cdot)$ be two times differentiable in an open set U, then we have*

$$\begin{aligned} f(\mathbf{x}_0+\mathbf{h}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T \mathbf{h} + (1/2)\mathbf{h}^T \nabla^2 f(\mathbf{x}_0)\, \mathbf{h} + o\left(\|\mathbf{h}\|^2\right) \\ &= f(\mathbf{x}_0) + \sum_{i=1}^{p} h_i f_i(\mathbf{x}_0) + \frac{1}{2}\sum_{i,j=1}^{p} h_i h_j f_{ij}(\mathbf{x}_0) + o\left(\|\mathbf{h}\|^2\right) \end{aligned}$$

Although the vector formulation is more elegant, our workhorse will be the element-wise formulation. For a proof consult a classical calculus book such as Marsden's Vector Calculus (Marsden, 2012).

Following are some analytical expressions for the Taylor Expansion of multiplying functions. These cases will constantly appear in the following sections.

Corollary 3 ***First Multiplicative Taylor Expansion.*** *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ and $g : \mathbb{R}^p \mapsto \mathbb{R}$ be two times differentiable at U and both $\mathbf{x}_0$, $\mathbf{x}_0+\mathbf{h} \in U$. Then a first order approximation to $f(\mathbf{x}_0+\mathbf{h})\, g(\mathbf{x}_0+\mathbf{h})$ is*

$$\begin{aligned} & f(\mathbf{x}_0)\, g(\mathbf{x}_0) + f(\mathbf{x}_0)\sum_{r=1}^{p} h_r g_r(\mathbf{x}_0) + g(\mathbf{x}_0)\sum_{i=1}^{p} h_i f_i(\mathbf{x}_0) \\ & + \sum_{i,r=1}^{p} h_i h_r f_i(\mathbf{x}_0)\, g_r(\mathbf{x}_0) + o(\|\mathbf{h}\|) \end{aligned}$$

Corollary 4 ***Second Multiplicative Taylor Expansion.*** *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ and $g : \mathbb{R}^p \mapsto \mathbb{R}$ be three times differentiable at U and both $\mathbf{x}_0$, $\mathbf{x}_0+\mathbf{h} \in U$. A second order approximation to $f(\mathbf{x}_0+\mathbf{h})\, g(\mathbf{x}_0+\mathbf{h})$ is*

$$\begin{aligned} & f(\mathbf{x}_0)\, g(\mathbf{x}_0) + f(\mathbf{x}_0)\sum_{r=1}^{p} h_r g_r(\mathbf{x}_0) + g(\mathbf{x}_0)\sum_{i=1}^{p} h_i f_i(\mathbf{x}_0) \\ & + \sum_{i,r=1}^{p} h_i h_r f_i(\mathbf{x}_0)\, g_r(\mathbf{x}_0) + \frac{1}{2} g(\mathbf{x}_0) \sum_{i,r=1}^{p} h_i h_j f_{ij}(\mathbf{x}_0) \\ & + \frac{1}{2}\sum_{i,j,r=1}^{p} h_i h_j h_r f_{ij}(\mathbf{x}_0)\, g_r(\mathbf{x}_0) + o(\|\mathbf{h}\|) \end{aligned}$$

Corollary 5 ***Quadratic Multiplicative Taylor Expansion.*** *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ and $g : \mathbb{R}^p \mapsto \mathbb{R}$ be differentiable at U and both $\mathbf{x}_0$, $\mathbf{x}_0+\mathbf{h} \in U$. A second order approximation to $\left[f\left(\mathbf{x}_0+\mathbf{h}\right) - f\left(\mathbf{x}_0\right)\right]^2 g\left(\mathbf{x}_0+\mathbf{h}\right)$ is*

$$g\left(\mathbf{x}_0\right)\sum_{i=1}^{p} h_i^2 f_i^2\left(\mathbf{x}_0\right) + \sum_{r,i=1}^{p} h_r h_i^2 g_r\left(\mathbf{x}_0\right) f_i^2\left(\mathbf{x}_0\right) + 2g\left(\mathbf{x}_0\right)\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} h_i h_j f_i\left(\mathbf{x}_0\right) f_j\left(\mathbf{x}_0\right)$$

$$+ 2\sum_{r=1}^{p}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} h_r h_i h_j g_r\left(\mathbf{x}_0\right) f_i\left(\mathbf{x}_0\right) f_j\left(\mathbf{x}_0\right) + o\left(\|\mathbf{h}\|\right)$$

Proof. Use the fact that

$$\left(\sum_{i=1}^{p} x_i\right)^2 = \sum_{i=1}^{p} x_i^2 + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} x_i x_j$$

■

A.4 Kernel Functions

This section serves two purposes. The first is to expose the formal definition of a *kernel function* and to show their properties. The second is to exhibit some examples of these functions.

A.4.1 Kernel Definition

An univariate kernel is a function from $\mathbb{R}$ to $\mathbb{R}$ that

- Preserves constant functions[10]. $\int K\left(v\right) dv = 1$
- Is symmetric. $K\left(v\right) = K\left(-v\right)$
- Has finite second moment. $\kappa \triangleq \int v^2 K\left(v\right) dv \in (0, +\infty)$
- Belongs to L_2. $\omega \triangleq \int K\left(v\right)^2 dv \in (0, +\infty)$
- Has finite second moment in L_2. $\rho \triangleq \int v^2 K\left(v\right)^2 dv \in (0, +\infty)$

[10] If $f\left(x\right) = c$ then $\int f\left(v\right) K\left(v\right) dv = c$

Relevant Integration Property Note that due to the second property (symmetry) there is a relevant integration property

$$\begin{aligned}
\int v^l K(v)^q \, dv &= \lim_{a\to\infty} \left[\int_{-a}^{0} v^l K(v)^q \, dv + \int_0^a v^l K(v)^q \, dv \right] && \text{(definition)} \\
&= \lim_{a\to\infty} \left[-\int_a^0 (-w)^l K(-w)^q \, dw + \int_0^a v^l K(v)^q \, dv \right] && (v = -w) \\
&= \lim_{a\to\infty} \left[-\int_0^a w^l K(w)^q \, dw + \int_0^a v^l K(v)^q \, dv \right] && \text{(symmetry)} \\
&= 0
\end{aligned}$$

for all odd integers l and all q.

To extend the kernel definition for the multivariate case there are two alternatives. To treat $K(\cdot)$ as a product of univariate kernels or to properly define a multivariate kernel with some desirable behavior. For the discussion below, define the following constants for different integrals

$$\begin{aligned}
\sigma_i &\triangleq \int v_i K(\mathbf{v}) \, d\mathbf{v} \\
\sigma_{ij} &\triangleq \int v_i v_j K(\mathbf{v}) \, d\mathbf{v} \\
\sigma_{ijk} &\triangleq \int v_i v_j v_k K(\mathbf{v}) \, d\mathbf{v}
\end{aligned}$$

Product Kernel Define $K : \mathbb{R}^p \to \mathbb{R}$ as a product of independent univariate kernels for each variable

$$K(\mathbf{v}) \triangleq \prod_{i=1}^{p} K_i(v_i)$$

Note that the following properties result as a consequence of the above definition. Each individual kernel integrates independently

$$\int K(\mathbf{v}) \, d\mathbf{v}_{-i} = K_i(v_i)$$

Moreover,

$$\sigma_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ \kappa \text{ if } i = j \end{cases}$$

also

$$\sigma_i = 0 = \sigma_{ijk}$$

and finally, along the same lines,

$$\int K(\mathbf{v})^2 \, d\mathbf{v} = \omega^p$$

since each univariate kernel independently integrates to ω.

Joint Multivariate Kernel If we are to define a joint multivariate kernel then we would require the following

$$\int K(\mathbf{v}) \, d\mathbf{v} = 1$$
$$\int K(\mathbf{v})^2 \, d\mathbf{v} < +\infty$$

also that $\sigma_{ij} \in (0, +\infty)$ (not necessarily 0 for cross terms) and finally that $\sigma_i = 0 = \sigma_{ijr}$.

An example of such a kernel would be a multivariate normal density with nonzero correlation between variables centered at zero. Note that this general setting is only used in the Density Estimation section. For the regression analysis we employ product kernels to simplify the expressions and make the book keeping more manageable.

A.4.2 Kernel Examples

In principle all symmetrical densities are candidates to be kernels (clearly assuming that they also have a finite second moment and belong to L_2). Nonetheless, the only two densities that are commonly found in the literature are the Uniform and the Gaussian.

$$K(v) = \frac{1}{2} I_{[-1,1]}(v) \qquad \text{(Uniform)}$$

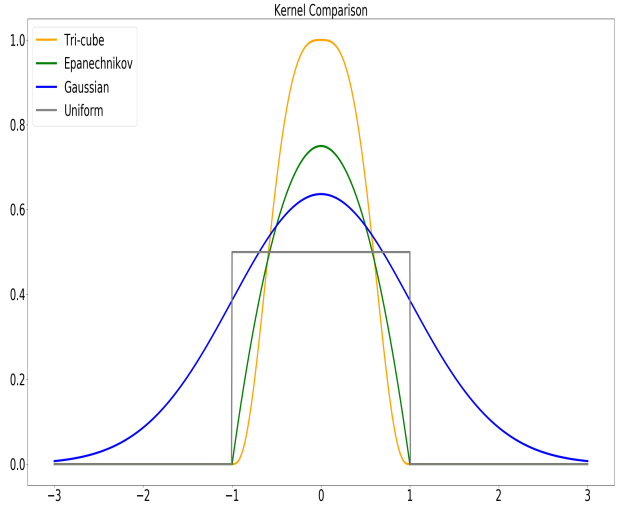
$$K(v) = \frac{1}{2\pi} e^{\frac{-1}{2} v^2} \qquad \text{(Gaussian)}$$

There are also two popular kernels found in the literature.

$$K(v) = \frac{3}{4}\left(1 - v^2\right) I_{[-1,1]}(v) \qquad \text{(Epanechnikov)}$$

$$K(v) = \left(1 - |v|^3\right)^3 I_{[-1,1]}(v) \qquad \text{(Tri-cube)}$$

The differences between these kernels can be seen in the following plot



As seen above, the Tri-cube kernel concentrates the most density around 0 and it is differentiable at the boundaries.

Notice that except for the Gaussian kernel, all others have a compact support. Succinctly, a compact support creates an influence cut-off point and restrains continuous functions. Remember that the goal of the kernel is to assign higher relevance to observations close to the target. A compact support completely eliminates all the influence from observations dimmed distant. Moreover, since continuous functions attain their minimum and maximum value in compact sets, a compact support bounds the influence of the derivatives of $\mu(\mathbf{x}) \triangleq E[Y|\mathbf{X} = \mathbf{x}]$. This last point will be clearly illustrated in the next section.

A.5 A Note on Integrating Taylor Expansions with Kernels

In the next section we will constantly integrate, against a kernel, the residual terms from the Taylor expansion of either the density $f(\cdot)$, the regression function $\mu(\cdot)$ or both multiplied.

To motivate the discussion let's expand a generic function $g(\mathbf{x}+\mathbf{v}\odot\mathbf{h})$ around $\mathbf{x}$, where $\odot$ denotes element-wise multiplication. Using the results from the Taylor Expansion section we have that

$$g(\mathbf{x}+\mathbf{v}\odot\mathbf{h}) = g(\mathbf{x}) + \sum_{i=1}^{p} h_i v_i g_i(\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{p} h_i h_j v_i v_j g_{ij}(\mathbf{x}) + o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right)$$

where the residual has the following shape

$$o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right) = \frac{1}{3!}\sum_{i,j,k=1}^{p} h_i h_j h_k v_i v_j v_k g_{ijk}(\widetilde{\mathbf{x}}_{ijk})$$

and

$$\widetilde{\mathbf{x}}_{ijk} \in B_{\|\mathbf{v}\odot\mathbf{h}\|}(\mathbf{x}) \triangleq \{\mathbf{z} : \|\mathbf{z}-\mathbf{x}\| \leq \|\mathbf{v}\odot\mathbf{h}\|\}$$

We will face the challenge to integrate this residual against the kernel. Mathematically,

$$\int o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right) K(\mathbf{v})\, d\mathbf{v}$$

Thus, the objective of the section is to discuss the conditions that guarantee the equation below

$$\int o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right) K(\mathbf{v})\, d\mathbf{v} = o\left(\|\mathbf{h}\|^2\right)$$

For this, fix $\mathbf{x}_0$ and $\mathbf{h}_0$ and define the following functions implicitly $\phi_{ijk} : \mathbb{R}^p \to \mathbb{R}$ as

$$g(\mathbf{x}_0+\mathbf{v}\odot\mathbf{h}) = g(\mathbf{x}_0) + \sum_{i=1}^{p} h_i v_i g_i(\mathbf{x}_0) + \frac{1}{2}\sum_{i,j=1}^{p} h_i h_j v_i v_j g_{ij}(\mathbf{x}_0)$$
$$+ \frac{1}{3!}\sum_{i,j,k=1}^{p} h_i h_j h_k v_i v_j v_k \phi_{ijk}(\mathbf{v})$$

where for each $\mathbf{v}$

$$\phi_{ijk}(\mathbf{v}) = g_{ijk}(\widetilde{\mathbf{x}}_{ijk}) \text{ , with } \widetilde{\mathbf{x}}_{ijk} \in B_{\|\mathbf{v}\odot\mathbf{h}_0\|}(\mathbf{x}_0)$$

and each $\phi_{ijk}(\cdot)$ inherits continuity. The question then becomes, what guarantees that the following integral makes sense

$$\frac{1}{3!}\sum_{i,j,k=1}^{p} h_i h_j h_k \int v_i v_j v_k \phi_{ijk}(\mathbf{v}) K(\mathbf{v}) d\mathbf{v}$$

The easiest way to escape from this problem would be to require $g(\cdot)$ to have bounded derivatives. Then the integral

$$\frac{1}{3!}\sum_{i,j,k=1}^{p} h_i h_j h_k \int v_i v_j v_k \phi_{ijk}(\mathbf{v}) K(\mathbf{v}) d\mathbf{v} \leq \frac{1}{3!}\sum_{i,j,k=1}^{p} h_i h_j h_k M \int v_i v_j v_k K(\mathbf{v}) d\mathbf{v}$$

However, asking for bounded derivatives totally disregards the annihilating power of the kernel function. From the previous kernel examples, all could deal with derivatives that increase at a polynomial rate. Moreover, the ones with compact supports can deal with more unwanted behavior. Since all $\phi_{ijk}(\mathbf{v})$ are continuous on $[-1,1]^p$ then they will attain their minimum and maximum. Formally,

$$\int v_i v_j v_k \phi_{ijk}(\mathbf{v}) K(\mathbf{v}) d\mathbf{v} \leq \max_{ijk} |\phi_{ijk}(\mathbf{v}^*)| \int_{[-1,1]^p} v_i v_j v_k K(\mathbf{v}) d\mathbf{v}$$

Additionally, in cases where $g(\cdot) = \mu(\cdot)$, we disregard concave functions like $\mu(x_1, x_2) = x_1^{\alpha} x_2^{1-\alpha}$ that usually appear in production economics but have unbounded derivatives.

In conclusion, our implicit requirement for the following propositions is that the kernel is sufficiently strong to bound the behavior of the derivatives of $f(\cdot)$, $\mu(\cdot)$ or both multiplied.

A.6 Useful Propositions

Following is a series of propositions that constitute the step-by-step elements for the different convergence proofs.

Proposition 1 ***Density Expectation****. Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$*. Assume that the density* $f(\cdot) \in C^3(\mathbb{R}^p)$*. Then*

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)\right]}{h_1\cdots h_p} = f(\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{p} h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij} + o\left(\|\mathbf{h}\|^2\right)$$

where $K(\cdot)$ *is a multivariate kernel and* $\sigma_{ij} = \int v_i v_j K(\mathbf{v})\,d\mathbf{v}$*.*

Proof. By definition we have that

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)\right]}{h_1\cdots h_p} = \frac{1}{h_1\cdots h_p}\int K\left((\mathbf{z}-\mathbf{x})\odot\mathbf{h}^{-1}\right) f(\mathbf{z})\,d\mathbf{z}$$

$$\begin{aligned}
&= \frac{1}{h_1\cdots h_p}\int\cdots\int K\left(\frac{z_1-x_1}{h_1},\cdots,\frac{z_p-x_p}{h_p}\right) f(z_1,\ldots z_p)\,dz_1\cdots dz_p \\
&= \int\cdots\int K(v_1,\cdots,v_p)\, f(x_1+v_1h_1,\ldots x_p+v_ph_p)\,dv_1\cdots dv_p \\
&= \int K(\mathbf{v})\, f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\,d\mathbf{v}
\end{aligned}$$

due to the change of variables $z_i = x_i + v_i h_i$. If we Taylor expand $f(\mathbf{x}_0+\mathbf{v}\odot\mathbf{h})$ around $\mathbf{x}$ the above integral becomes

$$\begin{aligned}
&f(\mathbf{x})\int K(\mathbf{v})\,d\mathbf{v} + \sum_{i=1}^{p} f_i(\mathbf{x})\,h_i\int v_i K(\mathbf{v})\,d\mathbf{v} + \\
&\frac{1}{2}\sum_{i,j=1}^{p} f_{ij}(\mathbf{x})\,h_i h_j\int v_i v_j K(\mathbf{v})\,d\mathbf{v} + \int o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right) K(\mathbf{v})\,d\mathbf{v}
\end{aligned}$$

and due to the kernel properties, the above expression reduces to

$$f(\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{p} h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij} + o\left(\|\mathbf{h}\|^2\right)$$

■

Proposition 2 ***Density Square Bias***. *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$ *and* $\|\mathbf{h}\| < 1$. *Assume that the sample is identically distributed from the density* $f(\cdot) \in C^2(\mathbb{R}^p)$. *Then*

$$E\left[\frac{\sum_{i=1}^n K\left(\left(\mathbf{X}_i - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)}{nh_1 \cdots h_p} - f(\mathbf{x})\right]^2 = \left(\frac{1}{2}\sum_{i,j=1}^p h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij}\right)^2 + o\left(\|\mathbf{h}\|^4\right)$$

where $K(\cdot)$ *is a multivariate kernel and* $\sigma_{ij} = \int v_i v_j K(\mathbf{v})\, d\mathbf{v}$.

Proof. Since the sample is identically distributed then

$$\begin{aligned} E\left[\frac{\sum_{i=1}^n K\left(\left(\mathbf{X}_i - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)}{nh_1 \cdots h_p} - f(\mathbf{x})\right] &= \frac{E\left[\sum_{i=1}^n K\left(\left(\mathbf{X}_i - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\right]}{nh_1 \cdots h_p} - f(\mathbf{x}) \\ &= \frac{E\left[K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\right]}{h_1 \cdots h_p} - f(\mathbf{x}) \end{aligned}$$

By the previous proposition we have that the past expression equals

$$= \frac{1}{2}\sum_{i,j=1}^p h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij} + o\left(\|\mathbf{h}\|^2\right)$$

Therefore, squaring the past term yields

$$\frac{1}{2}\sum_{i,j=1}^p h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij} + o\left(\|\mathbf{h}\|^2\right) =$$

$$= \left(\frac{1}{2}\sum_{i,j=1}^p h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij}\right)^2 + \sum_{i,j=1}^p h_i h_j f_{ij}(\mathbf{x})\,\sigma_{ij} o\left(\|\mathbf{h}\|^2\right) + o\left(\|\mathbf{h}\|^4\right)$$

note that the middle term is $O\left(\|\mathbf{h}\|^2\right) o\left(\|\mathbf{h}\|^2\right)$ therefore $o\left(\|\mathbf{h}\|^4\right)$ as in the multiplication lemma. ■

Proposition 3 ***Density Second-Moment***. *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$. *Assume that the density* $f(\cdot) \in C^2(\mathbb{R}^p)$. *Then*

$$\frac{E\left[K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)^2\right]}{h_1 \cdots h_p} = f(\mathbf{x})\,\omega + o(\|\mathbf{h}\|)$$

where $\epsilon^2(\mathbf{x}) = E\left[\epsilon^2|\mathbf{X}=\mathbf{x}\right]$ *,* $K(\cdot)$ *is a product kernel and therefore* $\omega = \int v_i K(\mathbf{v})^2 d\mathbf{v}$.

Proof. Expanding the expectation and making the standard change of variables we have

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\right]}{h_1\cdots h_p} = \int K(\mathbf{v})^2 f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\, d\mathbf{v}$$

Taylor Expanding $f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})$ around $\mathbf{x}$ to a first order yields

$$f(\mathbf{x})\int K(\mathbf{v})^2 d\mathbf{v} + \sum_{i=1}^{p} f_i(\mathbf{x}) h_i \int v_i K(\mathbf{v})^2 dv + \int o\left(\|\mathbf{v}\odot\mathbf{h}\|\right) K(\mathbf{v})^2 d\mathbf{v}$$

■

Proposition 4 ***Density Variance****. Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$ *and* $\|\mathbf{h}\| < 1$*. Assume that the sample is independent and identically distributed from the density* $f(\cdot) \in C^2(\mathbb{R}^p)$*. Then*

$$var\left(\frac{\sum_{i=1}^n K\left((\mathbf{X}_i-\mathbf{x})\odot\mathbf{h}^{-1}\right)}{nh_1\cdots h_p}\right) = \frac{f(\mathbf{x})\omega}{nh_1\cdots h_p} + o\left((nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right)$$

where $\epsilon^2(\mathbf{x}) = E\left[\epsilon^2|\mathbf{X}=\mathbf{x}\right]$*,* $K(\cdot)$ *is a product kernel and therefore* $\omega = \int v_i K(\mathbf{v})^2 d\mathbf{v}$.

Proof. Since the sample is independent and identically distributed then

$$var\left(\frac{\sum_{i=1}^n K\left((\mathbf{X}_i-\mathbf{x})\odot\mathbf{h}^{-1}\right)}{nh_1\cdots h_p}\right) = \frac{var\left(K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)\right)}{nh_1^2\cdots h_p^2}$$

using the fact that $var(\cdot) = E[(\cdot)^2] - E[(\cdot)]^2$ the previous variance becomes

$$= \frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\right]}{nh_1^2\cdots h_p^2} - \frac{1}{n}E\left[\frac{\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)\right]^2}{h_1\cdots h_p}\right]^2$$

$$= \frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\right]}{nh_1^2\cdots h_p^2} - o\left(n^{-1}\|\mathbf{h}\|^3\right)$$

by applying the proposition for the square bias. Now using the previous proposition this last equation becomes

$$= \frac{f(\mathbf{x})\,\omega}{nh_1\cdots h_p} + o\left((nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right)$$

since $o\left((nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right)$ is larger than $o\left(n^{-1}\|\mathbf{h}\|^3\right)$. ■

Proposition 5 ***Regression Bias***. *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$. *Assume that the density* $f(\cdot) \in C^3(\mathbb{R}^p)$. *Then*

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)(\mu(\mathbf{X})-\mu(\mathbf{x}))\right]}{h_1\cdots h_p} =$$

$$= \frac{f(\mathbf{x})\,\kappa}{2}\sum_{r=1}^{p} h_r^2\left[\frac{2\mu_r(\mathbf{x})\,f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right] + o\left(\|\mathbf{h}\|^2\right)$$

where $\mu(\mathbf{x}) \triangleq E[Y|\mathbf{X}=\mathbf{x}]$, $K(\cdot)$ *is a product kernel and therefore* $\kappa = \int v_i^2 K(\mathbf{v})\,d\mathbf{v}$.

Proof. Again, applying the usual change of variables $v_i = x_i + v_i h_i$ and writing the integral of the expectation we have that

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)(\mu(\mathbf{X})-\mu(\mathbf{x}))\right]}{h_1\cdots h_p} =$$

$$= \int K(\mathbf{v})\left(\mu(\mathbf{x}+\mathbf{v}\odot\mathbf{h}) - \mu(\mathbf{x})\right) f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\,d\mathbf{v}$$

now applying the second Taylor expansion for multiplying functions we have

$$= f(\mathbf{x})\sum_{i=1}^{p} h_i\mu_i(\mathbf{x})\int v_i K(\mathbf{v})\,d\mathbf{v} + \mu(\mathbf{x})\sum_{r=1}^{p} h_r f_r(\mathbf{x})\int v_r K(\mathbf{v})\,d\mathbf{v}$$

$$\sum_{i=1}^{p}\sum_{r=1}^{p} h_i h_r \mu_i(\mathbf{x})\,f_r(\mathbf{x})\int v_i v_r K(\mathbf{v})\,d\mathbf{v} + \frac{f(\mathbf{x})}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} h_i h_j \mu_{ij}(\mathbf{x})\int v_i v_j K(\mathbf{v})\,d\mathbf{v}$$

$$+\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{r=1}^{p} h_i h_j h_r \mu_{ij}(\mathbf{x})\,f_r(\mathbf{x})\int v_i v_j v_r K(\mathbf{v})\,d\mathbf{v} + \int o\left(\|\mathbf{v}\odot\mathbf{h}\|^2\right)K(\mathbf{v})\,d\mathbf{v}$$

where the only kernel integrals that survive are the ones of the kind $\kappa = \int v_i^2 K(\mathbf{v})\, d\mathbf{v}$. Therefore, the past expression reduces to

$$= \sum_{r=1}^{p} \kappa h_r^2 \mu_r(\mathbf{x}) f_r(\mathbf{x}) + \frac{f(\mathbf{x})}{2} \sum_{r=1}^{p} h_r^2 \mu_{rr}(\mathbf{x}) \kappa + o\left(\|\mathbf{h}\|^2\right)$$

$$= \frac{\kappa f(\mathbf{x})}{2} \sum_{r=1}^{p} h_r^2 \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right] + o\left(\|\mathbf{h}\|^2\right)$$

■

Proposition 6 ***Regression Square Bias.*** *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$. *Assume that the density* $f(\cdot) \in C^3(\mathbb{R}^p)$. *Then*

$$\left(\frac{E\left[K\left((\mathbf{X}-\mathbf{x}) \odot \mathbf{h}^{-1}\right)(\mu(\mathbf{X}_i) - \mu(\mathbf{x}))\right]}{h_1 \cdots h_p}\right)^2 = o\left(\|\mathbf{h}\|^3\right)$$

$$= \left(\sum_{r=1}^{p} \frac{f(\mathbf{x}) \kappa h_r^2}{2} \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right]\right)^2 + o\left(\|\mathbf{h}\|^4\right)$$

where $\mu(\mathbf{x}) \triangleq E[Y|\mathbf{X} = \mathbf{x}]$, $K(\cdot)$ *is a product kernel and therefore* $\kappa = \int v_i K(\mathbf{v})^2\, d\mathbf{v}$.

Proof. Applying the past proposition we have that

$$\left(\sum_{r=1}^{p} \frac{f(\mathbf{x}) \kappa h_r^2}{2} \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right] + o\left(\|\mathbf{h}\|^2\right)\right)^2 =$$

$$\left(\sum_{r=1}^{p} \frac{f(\mathbf{x}) \kappa h_r^2}{2} \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right]\right)^2$$
$$+ \sum_{r=1}^{p} f(\mathbf{x}) \kappa h_r^2 \left[\frac{2\mu_r(\mathbf{x}) f_r(\mathbf{x})}{f(\mathbf{x})} + \mu_{rr}(\mathbf{x})\right] o\left(\|\mathbf{h}\|^2\right) + o\left(\|\mathbf{h}\|^4\right)$$

note that since the middle term is $O\left(\|\mathbf{h}\|^2\right) o\left(\|\mathbf{h}\|^2\right)$ therefore $o\left(\|\mathbf{h}\|^4\right)$ as in the multiplication lemma. ■

Proposition 7 ***Regression Bias Second Moment.*** *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$*. Assume that the density* $f(\cdot) \in C^3(\mathbb{R}^p)$*. Then*

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\left(\mu(\mathbf{X}_i)-\mu(\mathbf{x})\right)^2\right]}{h_1\cdots h_p} = f(\mathbf{x})\,\rho\sum_{i=1}^{p}h_i^2\mu_i^2(\mathbf{x})+o(\|\mathbf{h}\|)$$

where $\mu(\mathbf{x}) \triangleq E[Y|\mathbf{X}=\mathbf{x}]$*,* $K(\cdot)$ *is a product kernel and therefore* $\rho = \int v_i^2 K(\mathbf{v})^2\,d\mathbf{v}$*.*

Proof. Again, integrating the expectation and doing the usual change of variables we have

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\left(\mu(\mathbf{X}_i)-\mu(\mathbf{x})\right)^2\right]}{h_1\cdots h_p} =$$

$$= \int K(\mathbf{v})^2\left(\mu(\mathbf{x}+\mathbf{v}\odot\mathbf{h})-\mu(\mathbf{x})\right)^2 f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\,d\mathbf{v}$$

by inputting the last Taylor Expansion for the multiplication of two functions we have that

$$\begin{aligned}
&= f(\mathbf{x})\sum_{i=1}^{p}h_i^2\mu_i^2(\mathbf{x})\int v_i^2K(\mathbf{v})^2\,d\mathbf{v}\\
&+2f(\mathbf{x})\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}h_ih_j\mu_i(\mathbf{x}_0)\,\mu_j(\mathbf{x}_0)\int v_iv_jK(\mathbf{v})^2\,d\mathbf{v}\\
&+\sum_{r=1}^{p}\sum_{i=1}^{p}h_rh_i^2f_r(\mathbf{x})\,\mu_i(\mathbf{x})\int v_rv_i^2K(\mathbf{v})^2\,d\mathbf{v}\\
&+2\sum_{r=1}^{p}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}h_rh_ih_jf_r(\mathbf{x})\,\mu_i(\mathbf{x})\,\mu_j(\mathbf{x})\int v_rv_iv_jK(\mathbf{v})^2\,d\mathbf{v}\\
&+\int o(\|\mathbf{v}\odot\mathbf{h}\|)\,K(\mathbf{v})^2\,d\mathbf{v}
\end{aligned}$$

as in the previous propositions the only terms that remain are of the kind $\rho = \int v_i^2K(\mathbf{v})^2\,d\mathbf{v}$. Therefore the past equation reduces to

$$f(\mathbf{x})\,\rho\sum_{i=1}^{p}h_i^2\mu_i^2(\mathbf{x})+o(\|\mathbf{h}\|)$$

■

Proposition 8 ***Regression Estimator Variance***. *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$ *and* $\|\mathbf{h}\| < 1$. *Assume that the sample is independent and identically distributed from the density* $f(\cdot) \in C^2(\mathbb{R}^p)$. *Then*

$$var\left(\frac{\sum_{i=1}^{n} K\left(\left(\mathbf{X}_i - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(\mu\left(\mathbf{X}_i\right) - \mu\left(\mathbf{x}\right)\right)}{nh_1 \cdots h_p}\right) =$$

$$= \frac{f(\mathbf{x})\,\rho \sum_{i=1}^{p} h_i^2 \mu_i^2(\mathbf{x})}{nh_1 \cdots h_p} + o\left(\left(nh_1 \cdots h_p\right)^{-1} \|\mathbf{h}\|\right)$$

where $\mu(\mathbf{x}) \triangleq E[Y|\mathbf{X} = \mathbf{x}]$, $K(\cdot)$ *is a product kernel and therefore* $\rho = \int v_i^2 K(\mathbf{v})^2 \, d\mathbf{v}$.

Proof. Since the sample is independent and identically distributed

$$var\left(\frac{\sum_{i=1}^{n} K\left(\left(\mathbf{X}_i - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(\mu\left(\mathbf{X}_i\right) - \mu\left(\mathbf{x}\right)\right)}{nh_1 \cdots h_p}\right) =$$

$$= \frac{var\left(K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(\mu\left(\mathbf{X}\right) - \mu\left(\mathbf{x}\right)\right)\right)}{nh_1^2 \cdots h_p^2}$$

writing the above variance as a sum of expectation we have that

$$= \frac{E\left[K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)^2 \left(\mu\left(\mathbf{X}_i\right) - \mu\left(\mathbf{x}\right)\right)\right]}{nh_1^2 \cdots h_p^2}$$

$$- \frac{1}{n}\left(\frac{E\left[K\left(\left(\mathbf{X} - \mathbf{x}\right) \odot \mathbf{h}^{-1}\right)\left(\mu\left(\mathbf{X}_i\right) - \mu\left(\mathbf{x}\right)\right)\right]}{nh_1 \cdots h_p}\right)^2$$

Applying the relevant propositions the above summands become

$$= \frac{f(\mathbf{x})\,\rho \sum_{i=1}^{p} h_i^2 \mu_i^2(\mathbf{x})}{nh_1 \cdots h_p} + o\left(\left(nh_1 \cdots h_p\right)^{-1} \|\mathbf{h}\|\right)$$

$$- o\left(n^{-1} \|\mathbf{h}\|^3\right)$$

Note that since $\|\mathbf{h}\| < 1$

$$\left(n^{-1} h_1 \cdots h_p\right)^{-1} \|\mathbf{h}\| > n^{-1} \|\mathbf{h}\|^3$$

therefore the term $o\left(\left(nh_1 \cdots h_p\right)^{-1} \|\mathbf{h}\|\right)$ dominates. ■

Proposition 9 ***Regression Second Error Moment.*** *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$ *and. Assume that the density* $f(\cdot) \in C^2(\mathbb{R}^p)$*. Then*

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\epsilon^2\right]}{h_1\cdots h_p} = \epsilon^2(\mathbf{x})\,f(\mathbf{x})\,\omega^p + o(\|\mathbf{h}\|)$$

where $\epsilon^2(\mathbf{x}) \triangleq E\left[\epsilon^2|\mathbf{X}=\mathbf{x}\right]$*,* $K(\cdot)$ *is a product kernel and therefore* $\omega^p = \int K(\mathbf{v})^2\,d\mathbf{v}$*.*

Proof. By using the law of total expectation we have that

$$\frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\epsilon^2\right]}{h_1\cdots h_p} = \frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2 E\left[\epsilon^2|\mathbf{X}\right]\right]}{h_1\cdots h_p}$$

then integrating

$$= \frac{1}{h_1\cdots h_p}\int K\left((\mathbf{z}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\epsilon^2(\mathbf{z})\,f(\mathbf{z})\,d\mathbf{z}$$

by applying the same usual change of variables the past integral transforms to

$$= \int \epsilon^2(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\,f(\mathbf{x}+\mathbf{v}\odot\mathbf{h})\,K(\mathbf{v})^2\,d\mathbf{v}$$

now by utilizing the first Taylor expansion for multiplying functions we have

$$\epsilon^2(\mathbf{x})\,f(\mathbf{x})\int K(\mathbf{v})^2\,d\mathbf{v} + \epsilon^2(\mathbf{x})\sum_{i=1}^{p} h_i f_i(\mathbf{x})\int v_i K(\mathbf{v})^2\,d\mathbf{v}$$
$$+ f(\mathbf{x})\sum_{r=1}^{p} h_r\epsilon_r^2(\mathbf{x})\int v_r K(\mathbf{v})^2\,d\mathbf{v} + \int o(\|\mathbf{v}\odot\mathbf{h}\|)\,K(\mathbf{v})^2\,d\mathbf{v}$$

due to the kernel properties the second and third element are 0 and hence the expression becomes

$$= \epsilon^2(\mathbf{x})\,f(\mathbf{x})\,\omega^p + o(\|\mathbf{h}\|)$$

■

Proposition 10 ***Regression Error Expectation.*** *Assume that*

$$E\left[\epsilon_i|\mathbf{X}_1=\mathbf{x}_1,\ldots,\mathbf{X}_n=\mathbf{x}_n\right]=0$$

then

$$E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon_i}{nh_1\cdots h_p}\right]=0$$

where $K\left(\cdot\right)$ *is a multivariate kernel.*

Proof. By applying the linear properties of the expectation and due to the law of total expectation we have

$$E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon_i}{nh_1\cdots h_p}\right]=$$

$$=E\left[\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)E\left[\epsilon_i|\mathbf{X}_1,\ldots\mathbf{X}_n\right]}{nh_1\cdots h_p}\right]$$
$$=0$$

■

Proposition 11 ***Regression Error Variance.*** *Let* $\mathbf{x}$ *be an interior point of the support of* $\mathbf{X}$*. Assume that the sample is independent and identically distributed from the density* $f\left(\cdot\right)\in C^2\left(\mathbb{R}^p\right)$*. Then*

$$var\left(\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon_i}{nh_1\cdots h_p}\right)=\frac{\epsilon^2\left(\mathbf{x}\right)f\left(\mathbf{x}\right)\omega^p}{nh_1\cdots h_p}+o\left(\left(nh_1\cdots h_p\right)^{-1}\left\|\mathbf{h}\right\|\right)$$

where $\epsilon^2\left(\mathbf{x}\right)\triangleq E\left[\epsilon^2|\mathbf{X}=\mathbf{x}\right]$*,* $K\left(\cdot\right)$ *is a product kernel and therefore* $\omega^p=\int K\left(\mathbf{v}\right)^2d\mathbf{v}$*.*

Proof. Since the sample is independent then the variance distributes sums therefore

$$var\left(\frac{\sum_{i=1}^{n}K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon_i}{nh_1\cdots h_p}\right)=\frac{var\left(K\left(\left(\mathbf{X}-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)\epsilon\right)}{nh_1^2\cdots h_p^2}$$

using the variance decomposition of $E[(\cdot)^2] - E[(\cdot)]^2$ the above expression becomes

$$= \frac{E\left[K\left((\mathbf{X}-\mathbf{x})\odot\mathbf{h}^{-1}\right)^2\epsilon^2\right]}{nh_1^2\cdots h_p^2} - \frac{0^2}{nh_1^2\cdots h_p^2}$$

where the last term equals zero due to the previous proposition. Applying the corresponding proposition for the first term results in

$$= \frac{\epsilon^2(\mathbf{x})f(\mathbf{x})\omega^p}{nh_1\cdots h_p} + o\left((nh_1\cdots h_p)^{-1}\|\mathbf{h}\|\right)$$

■

Proposition 12 ***First Moment Design Correction****. Let x be an interior point of the support of X. Assume that the sample is identically distributed from the density $f(\cdot)\in C^2(\mathbb{R})$. Then*

$$E\left[(nh)^{-1}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)(X_i-x)\right] = f_1(x)h^2\kappa + o\left(h^2\right)$$

where $K(\cdot)$ is an univariate kernel and $\kappa = \int v^2K(v)\,dv$.

Proof. By the linearity properties of the expectation and since the sample is identically distributed we have

$$\begin{aligned}
E\left[(nh)^{-1}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)(X_i-x)\right] &= \left(h^{-1}\right)E\left[K\left(\frac{X-x}{h}\right)(X-x)\right] \\
&= \left(h^{-1}\right)\int K\left(\frac{z-x}{h}\right)(z-x)f(z)\,dz \\
&= h\int vK(v)f(x+vh)\,dv
\end{aligned}$$

applying the change of variables $z = x+vh$. Taylor expanding $f(x+vh)$ around x up to a first order makes the past integral equal

$$\begin{aligned}
&= f(x)h\int vK(v)\,dv + f_1(x)h^2\int v^2K(v)\,dv + h\int o(|vh|)K(v)\,dv \\
&= 0 + f_1(x)h^2\kappa + ho(|h|)
\end{aligned}$$

due to the kernel properties. ■

Proposition 13 ***Second Moment Design Correction****. Let x be an interior point of the support of X. Assume that the sample is identically distributed from the density $f(\cdot) \in C^1(\mathbb{R})$. Then*

$$E\left[(nh)^{-1}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^2\right] = f(x)\,h^2\kappa + o\left(h^3\right)$$

where $K(\cdot)$ is an univariate kernel.

Proof. By the linearity properties of the expectation and since the sample is identically distributed we have

$$\begin{aligned} E\left[(nh)^{-1}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^2\right] &= \frac{1}{h}E\left[K\left(\frac{X - x}{h}\right)(X - x)^2\right] \\ &= \frac{1}{h}\int K\left(\frac{z - x}{h}\right)(z - x)^2 f(z)\,dz \\ &= h^2\int v^2 K(v)\, f(x + vh)\,dv \end{aligned}$$

applying the change of variables $z = x + vh$. Taylor expanding $f(x + vh)$ around x up to a first order makes the past integral equal

$$\begin{aligned} &= h^2 f(x)\int v^2 K(v)\,dv + h^3 f_1(x)\int v^3 K(v)\,dv + h^2\int o(|vh|)\,K(v)\,dv \\ &= f(x)\,h^2\kappa + h^2 o(h) \end{aligned}$$

■

A.7 Density Estimation

The idea of kernel regression was pioneered simultaneously by Nadaraya and Watson in 1964. Both were inspired by kernel density estimation (Hastie et. al., 2009). Thus, we will prove convergence results for density estimation and then conclude by motivating how this translates to the idea of regression.

First, let's motivate the idea of using a kernel function to estimate a density function. As usual, let $F(\mathbf{x}) = P[X_1 \leq x_1, \ldots, X_p \leq x_p]$ denote the distribution

function. By the frequentist interpretation of probability, if we take a sample, then the cumulative distribution can be approximated by

$$\widehat{F}_n(\mathbf{x}) = \frac{\#\{\mathbf{X}_i \text{ such that } X_{ij} \leq x_j\}}{n}$$

Moreover, the density function would then be approximated by (assuming that the predictors are independent)

$$f(\mathbf{x}) = \lim_{h \to 0} \frac{P[x_1 - h < X_1 \leq x_1 + h]}{2h} \cdots \frac{P[x_p - h \leq X_p \leq x_p + h]}{2h}$$

Restating the previous equation, we have

$$\widehat{f}_n(\mathbf{x}) = \frac{\#\{\mathbf{X}_i \text{ such that } x_j - h_j \leq X_{ij} \leq x_j + h_j\}}{n(2h)^p}$$

using the multivariate Uniform kernel, this last expression becomes

$$\begin{aligned} \widehat{f}_n(\mathbf{x}) &= \frac{1}{n(2h)^p} \sum_{i=1}^{m} I_{[\mathbf{x}-\mathbf{h}, \mathbf{x}+\mathbf{h}]}(\mathbf{X}_i) \\ &= \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right)}{nh^p} \end{aligned}$$

Let $(\mathbf{X}_i)_{i=1}^n$ be a random sample from the density defined by $f(\cdot)$ and $\mathbf{x}, \mathbf{h} \in \mathbb{R}^p$, such that $\mathbf{h} \gg 0$. Our kernel density, as motivated previously, is defined as

$$\widehat{f}_n(\mathbf{x}) \triangleq \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right)}{nh_1 \cdots h_p}$$

where $\mathbf{h}^{-1} \triangleq \left(h_1^{-1}, \ldots, h_p^{-1}\right)^T$ and $\odot$ denotes element-wise multiplication of vectors. The central result from this section states that the random number $\widehat{f}_n(\mathbf{x})$ converges in L_2 to $f(\mathbf{x})$ as $n \to +\infty$, $\|\mathbf{h}\| \to 0$, and $nh_1 \cdots h_p \to +\infty$. For this, let $\mathbf{x}$ be an arbitrary observation, we are now going to split the discussion in terms of bias and variance

$$\begin{aligned} MSE\left(\widehat{f}_n(\mathbf{x}), f(\mathbf{x})\right) &= E\left[\left(\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \\ &= bias\left(\widehat{f}_n(\mathbf{x}), f(\mathbf{x})\right)^2 + var\left(\widehat{f}_n(\mathbf{x})\right) \end{aligned}$$

By the proposition 2 the square bias equals

$$bias\left(\widehat{f}_n\left(\mathbf{x}\right),f\left(\mathbf{x}\right)\right)^2=E\left[\frac{\sum_{i=1}^n K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)}{nh_1\cdots h_p}-f\left(\mathbf{x}\right)\right]^2$$
$$=\left(\frac{1}{2}\sum_{i,j=1}^p h_ih_jf_{ij}\left(\mathbf{x}\right)\sigma_{ij}\right)^2+o\left(\left\|\mathbf{h}\right\|^4\right)$$

whereas by proposition 4 the variance equals

$$var\left(\widehat{f}_n\left(\mathbf{x}\right)\right)=var\left(\frac{\sum_{i=1}^n K\left(\left(\mathbf{X}_i-\mathbf{x}\right)\odot\mathbf{h}^{-1}\right)}{nh_1\cdots h_p}\right)$$
$$=\frac{f\left(\mathbf{x}\right)\omega}{nh_1\cdots h_p}+o\left(\left(nh_1\cdots h_p\right)^{-1}\left\|\mathbf{h}\right\|\right)$$

therefore, the MSE becomes

$$MSE\left(\widehat{f}_n\left(\mathbf{x}\right),f\left(\mathbf{x}\right)\right)=\left(\frac{1}{2}\sum_{i,j=1}^p h_ih_jf_{ij}\left(\mathbf{x}\right)\sigma_{ij}\right)^2+\frac{f\left(\mathbf{x}\right)\omega}{nh_1\cdots h_p}$$
$$+o\left(\left\|\mathbf{h}\right\|^4+\left(nh_1\cdots h_p\right)^{-1}\left\|\mathbf{h}\right\|\right)$$

This last term converges to 0 as, $n\to+\infty$, $\|\mathbf{h}\|\to 0$ and $(nh_1,\dots h_p)\to+\infty$. However, we can minimize the first two leading terms to accelerate this process. Assume that all h_i follow a similar order of magnitude, that is $h_i=h$. Therefore, the first two summands become

$$\zeta\left(h\right)\triangleq\frac{h^4}{4}\left(\sum_{i,j=1}^p f_{ij}\left(\mathbf{x}\right)\sigma_{ij}\right)^2+\frac{f\left(\mathbf{x}\right)\omega}{nh^p}$$

solving the first order conditions for this function we have

$$h_{opt}^3\left(\sum_{i,j=1}^p f_{ij}\left(\mathbf{x}\right)\sigma_{ij}\right)^2=\frac{pf\left(\mathbf{x}\right)\omega}{nh_{opt}^{p+1}}$$

and therefore

$$h_{opt} = \left(\frac{pf(\mathbf{x})\,\omega}{n\left[\sum_{i,j=1}^{p} f_{ij}(\mathbf{x})\,\sigma_{ij}\right]^2} \right)^{\frac{1}{p+4}}, \text{ thus } h_{opt} \propto n^{-\frac{1}{p+4}}$$

substituting this choice of h_{opt} yields

$$MSE\left(\widehat{f}_n(\mathbf{x}), f(\mathbf{x})\right) = O\left(n^{\frac{-4}{p+4}}\right)$$

which does show that the convergence of this method does require large amounts of data.

To motivate how density estimation translates into regression note that naturally

$$E[Y|\mathbf{X}=\mathbf{x}] \approx \int y \frac{\widehat{f}_n(\mathbf{x}, y)}{\widehat{f}_n(\mathbf{x})} dy$$

$$= \int y \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right)}{(nh_1 \cdots h_p)\,\widehat{f}_n(\mathbf{x})} K\left(\frac{Y_i - y}{h_y}\right) dy \qquad \text{(product kernel)}$$

$$= \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right)}{(nh_1 \cdots h_p)\,\widehat{f}_n(\mathbf{x})} \int y K\left(\frac{y - Y_i}{h_y}\right) dy \qquad \text{(symmetry)}$$

making the a variable change of $v = Y_i + h_y y$ yields

$$= \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right) Y_i}{(nh_1 \cdots h_p)\,\widehat{f}_n(\mathbf{x})}$$

therefore the conditional expectation of $E[Y|\mathbf{X}=\mathbf{x}]$ is estimated by

$$E[Y|\mathbf{X}=\mathbf{x}] \approx \frac{\sum_{i=1}^{n} K\left((\mathbf{X}_i - \mathbf{x}) \odot \mathbf{h}^{-1}\right) Y_i}{\sum_{j=1}^{n} K\left((\mathbf{X}_j - \mathbf{x}) \odot \mathbf{h}^{-1}\right)}$$

B Bibliography

[1] Brownstein J. S., Freifeld C. C. and Madoff L. C. (2009) *Digital disease detection-harnessing the web for public health surveillance*. New England Journal of Medicine.

[2] Buja A., Hastie T., Tibshirani R. (1989) *Linear Smoothers and Additive Models*. Annals of Statistics 17 453-555

[3] Cooper C., Mallon S., Pollack L. and Peipins L. (2005) *Cancer internet search activity on a major search engine*. United States 2001-2003. J Med Internet Res, 7.

[4] Cukier K., Mayer-Schoenberger V. (2013) *The Rise of Big Data: How It's Changing the Way We Think About the World*. Foreign Affairs Vol. 92, No. 3 (May/June 2013), pp. 28-40

[5] Dudek, G. (2012) *Variable Selection in the Kernel Regression Based on Short-Term Load Forecasting Methods.* Springer-Verlag Berlin Heidelberg. ICAISC 2012, Part II, LNCS 7268, pp. 557-563.

[6] Durbin J., Koopman S. J. (2001) *Time Series Analysis by State Space Methods*. Oxford University Press.

[7] Ettredge M., Gerdes J. and Karuga G. (2005) *Using web-based search data to predict macroeconomic statistics*. Communications of the ACM, 48(11): 87-92, 2005.

[8] Fan J., Gijbels I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

[9] Fan J., Gijbels I. (2008) *Variable Bandwidth and Local Linear Regression Smoothers*. The Annals of Statistics, Vol. 20, No. 4, 2008.

[10] Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L. Smolinski M. S. and Brilliant L. (2009) *Detecting influenza epidemics using search engine query data*. Nature, pp. 1012-1014.

[11] Goodfellow I., Bengio Y., Courville A. (2016) *Deep Learning* MIT.

[12] Harvey, A. C. (1989) *Forecasting, structural time series models and the Kalman Filter*. Cambridge University Press.

[13] Hastie T.,Tibshirani R., Friedman J. (2009) *The Elements of Statistical Learning*. Springer.

[14] Li Q., Racine J. S. (2004) *Cross-Validated Local linear Nonparametric Regression* Statistica Sinica.

[15] Li Q., Racine J. S. (2007) *Nonparametric Econometrics: theory and practice* Princeton University Press

[16] Madigan D., Raftery A. E. (1994) *Model selection and accounting for model uncertainty in graphical models using Occam's window*. Journal of the American Statistical Association 89, pp. 1535-1546.

[17] Marsden J. E., Tromba, A. (2012) *Vector Calculus* Sixth Edition, W.H. Freeman and Company.

[18] McCulloch R. E., George, E. I. (1997) *Approaches for Bayesian variable selection*. Statistica Sinica 7, 339-374.

[19] Murphy, K. (2012) *Machine Learning A Probabilistic Perspective* MIT Press.

[20] Nocedal J., Wright S. J. (2006) *Numerical Optimization* Second Edition, Springer.

[21] Polgreen P. M., Chen Y., Pennock D. M. and Nelson F. D (2008) *Using internet searches for influenza surveillance*. Clinical Infectious Diseases, 47:1443-1448.

[22] Royden H.L., Fitzpatrick P.M. (2010) *Real Analysis* Fourth Edition, Pearson Education.

[23] Shalizi C. R., (2017)*Advanced Data Analysis from an Elementary Point of View*. Link: http://www.stat.cmu.edu/ cshalizi/ADAfaEPoV/

[24] Silver N (2015) *The Signal and the Noise: Why So Many Predictions Fail–but Some Don't*. Penguin Books.

[25] Stephens-Davidowitz S. (2013) *The Cost of Racial Animus On A Black Presidential Candidate: Using Google Search Data To Find What Surveys Miss*.

[26] Valdivia A., Monge-Corella S. (2010) *Diseases tracked by using Google Trends*. Emerg Infect Dis.

[27] Vanderkam D., Schonberger R., Rowley H., Kumar S. *Nearest Neighbor Search in Google Correlate*.

[28] Varian H. R. (2014) *Big Data: New Tricks for Econometrics* The Journal of Economic Perspectives, Vol. 28, No. 2 (Spring 2014), pp. 3-27.

[29] Varian H., Choi H. (2011) *Predicting the Present with Google Trends* Google Inc.

[30] Varian H., Scott S. L. (2013) *Predicting the Present with Bayesian Structural Time Series* Google Inc.