Theory and Application of Dynamic Spatial Time Series Models

Andree, B.P.J.

2020

document version
Publisher's PDF, also known as Version of record

Chapter 5

Non-parametric Cross-sectional Nonlinearities

Chapter Summary

The UN's Sustainable Development Goals for 2030 aim on one hand at inclusive growth and eradicating poverty, and on the other at preserving environments. The relation between development and the environment has been studied extensively since the 1990s, documenting inverted U-shaped relations between per capita income and indicators of environmental degradation. This paper revisits the issue with machine learning techniques and novel disaggregate data to model these relationships heterogeneously across economic indicators. Results suggest that development gradually improves the efficiency of consuming the earth's nonrenewable resources, but increased efficiency alone is not sufficient to offset growth in scale. Development shifts reliance on one nonrenewable source to another, and on average we find successive inverted U-shapes in deforestation, air pollution and carbon intensities, followed by a J-shape in per capita carbon output. Local economic circumstances further determine the shape, amplitude, and location of tipping points in environmental output. The general implications of the estimated dynamics are explored by extrapolating environmental output to 2030 under simplistic scenario's. The results are a reminder that immediate, and sustained global efforts are required to preserve our environment.[1]

[1]This chapter is based on a compilation of work. It draws from *"Environment and Development"* published by the *World Bank*, the full reference is Andree et al. (2019). An adapted version *"Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission"* of the same authors together with Dr. Eric Koomen is published in the *Journal of Renewable and Sustainable Energy Reviews*. The reference is Andrée et al. (2019). The supplementary appendix is based on the technical background note associated with this publication, available here https://doi.org/10.1016/j.rser.2019.06.028. The material is reproduced here with kind permission from *Elsevier* and the *World Bank*.

5.1 Introduction

Will continuation of economic development increase pressure on the earth's finite resources, or does the increase in income provide the basis for environmental improvement? This question was central to several empirical studies in the early 1990s (Grossman and Krueger, 1991; Shafik and Bandyopadhyay, 1992; Panayotou, 1993). Their initial work suggested the existence of an inverted U-shaped relation between per capita income and environmental degradation indicators related to pollution (e.g. SO2, NOx), deforestation, and carbon emission. Their hypothesized environmental Kuznets curves gained massive following in research and policy as they held the attractive promise that economic development could actually benefit the environment. Early empirical examples of these curves, their possible explanations and policy implications are reviewed by Soumyananda (2004); Bo (2011). From 2000 onwards, however, substantial criticism was formulated in relation to the poor statistical foundation of these curves (Stern, 2004) and the obsession with replicating the exact inverted U-shape (Levinson, 2001). Stern (2004) points out that increases in wealth and income occur simultaneously with a structural transformation process in which the composition of inputs and methods of production gradually shift in favor of less destructive production. So it is not necessarily the increase in income that makes lower emission levels possible, but the gradual adoption of cleaner technology that can occur irrespective of development status (as documented by, for example, Stern and Common (2001); Dasgupta et al. (2002)). Environmental impact is thus determined both by efficiency of production, which may improve nonlinearly across GDP, and by total production size, which varies across panels of countries (Stern et al., 1996). If the scale of the economy is large, minute changes in the efficiency of production can result in large differences in output levels. Therefore, if a panel is constructed that includes economies of widely different scales, the variance in environmen-

tal output levels can be expected to vary with the GDP levels of the countries. To cope with this, one should acknowledge in the model design that environmental output is de facto a result of both a *scale* component, and a *technology* component.

Many empirical approaches have tried to model degradation levels directly, not distinguishing between the role of *scale* and a *technology* separately within the model, and therefore assume a degree of homogeneity in the emission-income relationship that is unrealistic for a panel of widely differing countries. A possible theoretical foundation for the environmental Kuznets curve, found in technological progress and diminishing returns to capital, is discussed by Brock and Taylor (2010). They highlight that modeling a panel relationship between emission levels and per capita income directly is not supported by their theory. Instead, they focus on combined panel data on emission intensities and abatement costs. Others have highlighted that, even when the regression deals with per capita emissions instead of levels, restricting cross-sections to undergo identical experiences over time biases results (List and Gallet, 1999). This suggests flexible approaches that allow for heterogeneous relationships may be more suitable as they allow for locally varying patterns to exist. Vollebergh et al. (2005) pay specific attention to homogeneity assumptions in their environmental impact regressions, and conclude that correctly modeling heterogeneity is essential to prevent spurious correlation in reduced-form panel estimations. Other econometric issues with environmental Kuznets curves relate to inappropriately dealing with the serial dependence and omitted variables bias Stern and Common (2001); Stern (2004). This has partly been addressed by adding control variables as in studies reviewed by Stern (1998), or by deploying fixed-effect approaches Stern (2004). Time series approaches that claim that the error correction approach provides appropriate diagnostic statistics and specification tests for the environment-economic relationship are also widespread (see, for example, Perman and Stern (2003); Stern (2004)). However, even over

time, linearity and constant variance assumptions break for moderate time dimensions because nonlinearities result from the income dependence of the derivatives of the function mapping changes in income to changes in degradation. From that perspective, the non-parametric error correction approach (Shahbaz et al., 2017) improves on previous work. More discussion on nonlinear cointegration in the context of the environmental Kuznets curve can be found in (Wagner, 2015). His general conclusion is that the diagnostics available in the standard framework are not appropriate in the nonlinear case because powers of integrated processes are themselves not integrated. In the non-parametric context on the other hand, causality and other correct-specification arguments are tightly related to the penalization technique, or bandwidth setting, that may take the limit criterion away from the *true* parameter.

In this paper, we revisit the empirical relation between economic growth and the environment using a panel data set on environmental indicators and economic development for a large set of countries applying a flexible kernel model that allows dependencies to vary smoothly throughout the data. We focus on a cross-comparable *technology* component represented by degradation intensities of average per capita wealth production to cope with the heteroskedasticity related to economic scales. The empirical strategy taken, pays tribute to the earlier literature that argued in favor of modeling outcome variables that are cross-comparable, and for using flexible models that allow relationships to vary throughout the data. To allow for a wide variety of potential nonlinearities with minimal parametric assumptions, we deploy a machine learning method that learns from similarities in the data using kernels. The method is known as Kernel Regularized Least Squares (Hainmueller and Hazlett, 2014). The key reason behind this choice is that apart from flexibility and taking full advantage of the kernel learning framework, it is still straightforward enough to back out marginal effects.

The framework is used together with out-of-sample selection of fixed effects to model remotely sensed and reported environmental data for 95 countries that include 85% of the world's population, 83% of global carbon output and 72% of all forest cover. The large sample of countries and simultaneous assessment of three environmental pressures along identical economic data using the same modeling strategy, differentiates our work from recent studies that use various methods to approach the relationship between economic development and specific environmental pressures in individual countries (e.g. (Managi and Jena, 2008; Keene and Deller, 2015; Apergis and Ozturk, 2015)), or the recent wave of research on carbon emissions in more limited samples of countries (e.g. (Apergis, 2016; Özokcu and Özdemir, 2017; Awaworyi Churchill et al., 2018)).

The remainder of this paper is as follows. We discuss estimation methods in Section 5.2. We provide only a non-technical discussion here, more technical discussion is provided in the supplementary background notes made available together with this paper. Section 5.3 details the data used for our empirical analysis in section 5.4. Finally in section 5.4.5, we use our empirical descriptions to explore the implications of continuation of growth on environmental output. Section 5.5 concludes.

5.2 Methods

In most explanatory analysis, the estimated model is assumed to consist of a finite set of parameters. While this makes the interpretation straightforward, it imposes strong assumptions about the behavior of the process being modeled. Specifically, linear models assume that the relationship between two variables Y and X described by a parameter β is constant across levels of Y and X. Such strong assumptions about the data generation process (DGP) are rarely - if ever - justified by economic

theory and can lead to seriously erroneous conclusions. The single parameter elasticities in linear models can at most approximate the average of the nonlinear elasticities locally on a function. Puu (1991) provides an excellent discussion on linear versus nonlinear dynamics in economics, and the key issue that linear approximations can be reasonable within bounds but should not be used to infer change so large that the bounds of the approximation interval are violated. Naturally, these bounds depend on the strength of the nonlinearity (Lorenz, 1993). Linear approaches can thus yield useful evidence, but only within a relatively narrow range of the overall state space, particularly if large parts of the population are likely to pass through that state. However, in general, local approximations fall short when building global arguments. Costanza et al. (1993) provide an excellent discussion on the severe limits of taking simple relationships from a local level and aggregating them up to describe the large-scale behavior of a complex system. The other way around, inferential errors induced by fixing the relationships in a complex system at the average, increase with the divergence between the average observation and the values of the observations of interest. This is undoubtedly the case in the analysis of economic development and environmental output, in which the structural behavior of the outliers, such as the poorest or most polluted, are often of foremost concern to policy makers.

Finite dimensional nonlinear parametric models may address several of these issues, but require strong predictions from underlying economic theory on the implied form of the structural relationship for the parameters to be economically meaningful. Such functions may be difficult to parameterize. Finite series approximators may also provide flexibility, but the resulting conclusions are often different from models in which the order of approximation is allowed to vary along the sample size, see Horowitz (2011) for further discussion.

Non-parametric models make fewer assumptions about the DGP, and

can produce approximations with varying flexibility (Härdle et al., 2004). The key question in this case is how flexible the empirical function should be given the data that has been observed. A regularized non-parametric model exposed to growing data is in a sense an approximator that adjusts its belief of what an appropriate description of the DGP is according to the number of observations that is seen. We apply a particular type of this model in this paper that is known as the Kernel Regularized Least Squares estimator developed by Hainmueller and Hazlett (2014). Key in this approach is that the model adjusts its understanding of the DGP as the data grows. A non-parametric model in which the size of the model is appropriately regulated results in a small size when samples are few, but may increase in dimensionality as the data grows. As a result, the approximation error declines with growing data. Regulation of the order of approximation in non-parametric models occurs through tuning parameters. Correct inference is therefore strongly dependent on values that are not estimated by the criterion, but instead set by the researcher. While the relationship between standard loss minimization and correct parameter inference, as in the linear Least Squares literature, is a basic concept well-known to many researchers, inference based on the estimators of a non-parametric model has to consider the effect of the external parameters for which results do not follow under the same consistency and normality theorems. For example, while Hainmueller and Hazlett (2014) provide consistency and normality results for their model, they state explicitly that these results are different for every level of penalization. We refer the reader to Andree et al. (2019) and the supplementary notes made available together with this paper for an in-depth discussion on this topic and a technical exposition of the model. We also provide more discussion there regarding the assumptions we make about the type of nonlinearities in the data generating process. For

the general reader, it suffices to say that our regression is of the form:

$$\mathbf{y}_t = h(X_t) + g(\epsilon_t) + \boldsymbol{\varepsilon}_t, \tag{5.1}$$

where $\mathbf{y}_t$ is a vector of environmental degradation variables at time t which will be introduced in our next section, X_t is a matrix of economic variables at time t that are similarly introduced in the next section, h is a flexible function that is approximated using Gaussian kernels, ϵ_t are time specific constants with g being some function that determines whether those time-specific error effects should be included, and $\boldsymbol{\varepsilon}_t$ are vectors of residuals at time t.[2] The regression is estimated by minimizing Least Squares using a penalty that discourages overly complex results. The penalty is chosen using cross-validation. This also ensures that, as data grows, the estimated marginal effects can be interpreted as usual, as is explained further in the supplementary background notes.

5.3 Data

We combined measures of tree loss, air pollution concentrations and carbon emissions on one side, and GDP indicators of economic structure on the other. Our data is from a variety of sources and includes 95 countries measured over $1999 - 2014$ containing approximately 85% of the world's population, 83% of the world's carbon output, and 72% of the world's forest cover. We have removed areas below 1500 square kilometers − essentially all small island states − from the analysis. A summary of the data as it enters our regressions is given below.

[2]Note that ϵ_t could be part of X_t, which is how we treat it in the appendix. We have written it here separately as an error component, which may be more recognizable to those that are familiar with the panel regression setting.

5.3.1 Forest cover

We use data from Hansen et al. (2013), which contains estimates of global tree cover extent (2000) and annual tree cover loss (2001-2014) at a spatial resolution of 30 meters.[3] They analyzed satellite images from Landsat 5, 7, and 8 to identify tree cover extent, defined as vegetation taller than 5 meters in height, and loss, defined as complete removal of tree cover canopy. The authors reported the tree cover loss data to have a false positive rate of 13%, a false negative rate of 12% and a ratio of total forest gain to loss over 2001-2012 of 0.34. The derived data differs from statistics reported by the UN-FAO's Forest Resource Assessment, but due to the consistent methodology and definition of forests across countries, we believe this data is better suited for a global analysis. We define forests as pixels with a minimum canopy closure density of 30%. Finally, we convert the data to area measures and sum the data by country to calculate tree cover loss as a percentage of tree cover extent in 2000. Our intention is to examine "natural dense forests", but note that the data also captures forest plantations. Our loss measure is thus only a proxy for deforestation, as there may be many other natural and anthropogenic processes (storm damage, fires, mechanical harvesting) that are reflected by the data. We refrain from including the tree cover gain data as there are significant differences in methodology that limit additivity or comparison with loss.

Forest cover loss included two outliers of respectively 10.8% and 5.4% loss in Namibia (2001, and 2005). For comparison, the median observation across time in this country was 1.7%. We have capped these numbers at 3% which, seemed an appropriate maximum for the range of forest loss after inspecting a kernel density. We applied a three-period simple moving average to further smoothen outliers. Our final data set includes

[3]The data can be found at https://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html.

the losses for countries that held 72% of forest cover in 2000. The largest missing forest patch is that of the Russian Federation.

5.3.2 Air pollution

We use concentrations of fine particulate matter ($PM_{2.5}$), coarse dust particles of 2.5 micrometers in diameter, as a proxy for broader air pollution. The ($0.01^\circ \times 0.01^\circ$ resolution) data is developed by van Donkelaar et al. (2016) and includes global annual ground-level $PM_{2.5}$ (1999-2014) derived from a combination of satellite-, simulation- and monitor-based sources. The data set has been developed from satellite-derived Aerosol Optical Depth reflectance values calibrated to ground-based $PM_{2.5}$ observations using a Geographically Weighted Regression.

Remote sensing methods aim to observe particulate matter but are prone to capturing fine dust released from barren lands that have similar reflectance properties in the high frequency spectral wavelengths. This poses a difficulty in our analysis, as countries in desert regions have high country wide average pollution levels, while large countries, or those with substantial forest cover where ambient pollution is low, have lower average concentration to what the larger population is exposed to on a regular basis. We used gridded population data that is produced using a combination of light at night data and census data, to identify patches of urban areas.[4] We averaged gridded pollution data that falls within urban boundaries to the country-level, defining urban areas as places where population density was higher than 300 people per square kilometer. The results in fig. 5.1 show that this procedure results in higher pollution levels in large countries with known pollution problems in cities (notably China, Nepal, Pakistan) or those with forests (notably Lao PDR, Indonesia, Senegal), and in lower concentrations in areas with

[4]The population grids are from http://sedac.ciesin.columbia.edu/data/collection/gpw-v4, we use the 2000 grids.

known deserts (notably Chad, Tunisia, Morocco).

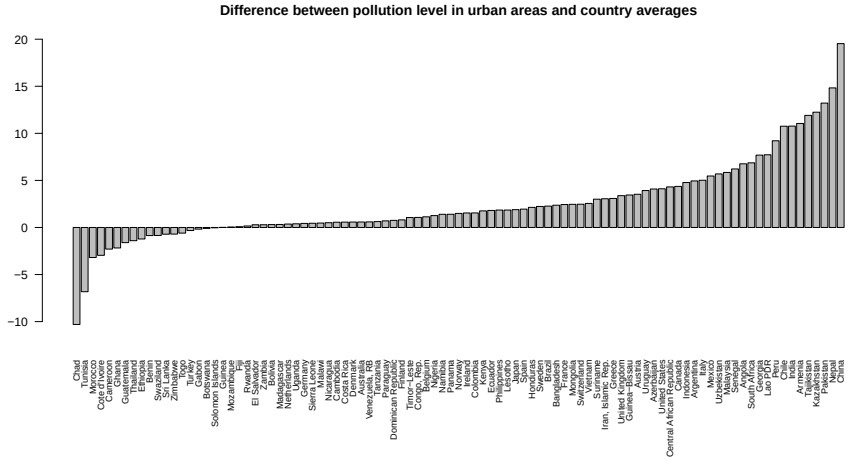

Figure 5.1: Difference between average air pollution ($PM_{2.5}$) in urban areas and country-wide average.

5.3.3 Carbon emission and economic development

For data on carbon dioxide emissions and economic variables, we rely on the World Bank's World Development Indicators (WDI). To ensure cross-country comparability, we use GDP per capita in constant 2011 international dollars adjusted for purchasing power parity (PPP). The PPP adjusted GDP accounts for highly variant costs of living between countries. Such adjustments to GDP have been criticized by some to possibly overstate the wealth of poor countries (Coyle, 2014), while others highlight that the form of PPP adjustments may matter for appropriate international comparison (Davies et al., 2011). Nevertheless, PPP adjustments remain the standard approach for international comparison of GDP figures across countries adopted by international institutions (The World Bank, 2011, 2013; Statistical Office of the European Communities and Organisation for Economic Co-operation and Development, 2012).

Over- or under-stating wealth does not pose a problem to the current analysis if the bias effects countries with similar GDP in a similar way, as we are purely interested in trends across wealth and not necessarily performing an unbiased wealth assessment itself. In the remainder of this paper, when we mention GDP we refer to its adjusted version and with an international dollar we refer to a single unit of GDP.

The CO_2 emissions estimates retrieved from WDI were produced by the U.S. Department of Energy's Carbon Dioxide Information Analysis Center (CDIAC) and include anthropogenic emissions from fossil fuel consumption and world cement manufacturing. The data set includes approximately 83% of global carbon emissions and should be quite representative of missing countries as it is close to the 85% of world population included in our samples.

5.3.4 Treatment of missing data

Almost every ambitious analysis that aims to pull together various sources of data to produce insights supported by a wide range of observations, is eventually plagued by missing data of some form. The WDI data set contains a wealth of information, but some important observations are missing. Immediately, this poses a trade-off between using less, but complete, data, or using more data but having to cope with missingness by deploying an imputation strategy.

The predictive modeling community has generally found that using more information tends to result in better predictions (Kuhn and Johnson, 2013). The view is that the usefulness of the imputation can be inferred from looking at out-of-sample performance of the final model. Hence, this has led to a more relaxed opinion about using sparsely observed variables to build predictive models, and various approaches are widely available (Kuhn, 2008; Kowarik and Templ, 2016).

When interested in inference, however, it is important to understand the reason behind missingness. A common presumption is that imputations introduce additional uncertainty and possibly bias. However, complete-case studies can in fact lead to much more severe problems if the observations are missing for a reason (Westreich, 2012). A trivial example in the current context is the following one. If countries with extreme carbon emissions simply choose not to report them, then surely, we would underestimate carbon output if we drop those cases. While many remain cautious to use imputed data, complete-case analysis is in fact only unbiased under restrictive assumptions (Rubin, 1976), and our default view for better inference is to favor imputations.

Strategies to deal with (extreme) missingness are treated for example in (Little and Rubin, 2012; Graham, 2012; Salgado et al., 2016). A standard approach is the use of fully-conditional regression specifications to fill in missing data, e.g. using regression specifications based on all other available covariates (Audigier et al., 2018), and the most favorable method (but often computationally challenging in the nonlinear case) is possibly that of using multiple impuatations and pooling regression results (Rubin, 1996). This is of interest when one is also concerned about correcting the conditional variance function rather than only the conditional mean function. In either case, flexible modeling strategies tend to produce both improved prediction performance as well as better inference than linear imputation approaches (Murray, 2018). However, tractability and simplicity are not factored in when merely pitting different imputation approaches against one another based on simple diagnostics. For that reason, we adopt different approaches depending on the imputation case.

GDP has only .12% missing, manufacturing and services GDP shares have 5.57% missing. We interpolate these values linearly over the time dimension. The WDI does not report income shares in each country but sometimes reports a Gini index. We used this to back out income

shares. In particular, we make use of the fact that income shares held by a certain share of the population can be read off the Lorenz-curve and that the Gini coefficient is a measure of dispersion of the Lorenz-curve calculated from the summed surface under the Lorenz-curve and the surface under the 45° line. In total we are able to collect 937 observations of both Gini coefficients and income shares held by the first two quintiles. We estimate the nonlinear inverse map with high precision (R^2 of .99) using the penalized non-parametric estimator. We then used the Gini observations to predict the income shares. After this first imputation step, 61.25% of the observations remain missing, but only over the time dimension. In the countries that have more observations in the time dimension, we observe that the income shares are relatively stable over time. We therefore simply interpolate the remaining missing values linearly over time.

Poverty rates has 69.54% missing values. A large part of the missing values are due to statistics that are not produced in high income countries. We fill all missing poverty and undernourishment rates above $23,000$ GDP per capita with zero. The choice of this threshold is because undernourishment rates were not reported above this income value, and only Malaysia, with $24,500$ GDP ppp per capita, had a positive reported poverty rate (1.3%). All other countries already attained 0% poverty rates in the published data. After this first imputation step, 49.54% of poverty remains missing because most countries are not complete in the time dimension. first interpolate these variables by taking a weighted average over time. This works well for most variables, but may yield poor results for poverty, as we have seen a tremendous improvement in most countries in past years. We improve the time dynamics in the interpolated poverty data by using information about time dynamics contained in our other variables. We vectorize the interpolated values, and fit the kernel model using the full set of undernourishment, logarithmic GDP per capita, the share of manufacturing, services, urban population shares,

and bottom 40 income shares. The model reaches an R^2's of .91. We use this model to smoothen the interpolated poverty values by taking an average of the interpolated values and the values predicted by this nonlinear model.

A final caveat is in order. Ultimately, the data on poverty and income shares remains patchy at best. We are well aware that the sparsity and the heavily imputed nature of the variable used in the analysis may remain a controversy to some. To put the missingness of 61.25% and 49.54% of the bottom income shares and poverty rates in perspective, the World Bank and UN-FAO methodology for the calculation of the official undernourishment statistics is based on three-year linear moving averages of model results produced using Household Consumption and Expenditure Surveys that are taken every 3-5 years (Moltedo et al., 2014). At this stage of knowledge, the objective of the research here is to further open up the empirical debate on the poverty-environment relationship. Recent initiatives such as the United Nation's Poverty-Environment Initiative highlight the importance of the relationship for policy, and recent research highlights the importance that poverty plays in the quality of the environment (Dogo et al., 2019). Since the hypothesis behind the environmental Kuznets curve is that poor countries may be more polluting, simply dropping incomplete cases would lead to a severe bias of the result.

5.3.5 Other controls and final data

We use the NDVI from the Moderate resolution Imaging Spectroradiometer (MODIS) derived from NASA's Terra satellite imagery to control for effects that relate to a variety of physical characteristics and natural assets of a country that may, for example, have an impact on ambient pollution levels or forest growth and loss dynamics.[5] This data set pro-

[5]Available at https://modis.gsfc.nasa.gov/data/dataprod/mod13.php.

vides spatial and temporal comparisons of global vegetation conditions. The original data has a monthly frequency at a resolution of 1km. We calculated the mean NDVI value for each year in our analysis, using 2000-2015 data, and summarized the data to the country-level using the mean, minimum, and maximum value to get a broad description of the vegetation in a country.

Table 5.1 summarizes the data, all predictors are mapped into the $[0, 1]$ interval to ensure the penalization effect is not driven by differences in variance of the different variables. We scale back the estimation results for easier interpretation.

Table 5.1: Summary of the data used in our empirical application. Statistics are not weighted and not necessarily representative of the world averages.

Statistic	Mean	St. Dev.	Min	Max
Annual % Tree loss	0.437	0.406	0.009	2.924
Urban $PM_{2.5}$ $\mu g/m^3$	18.924	12.609	0.311	63.498
CO_2 kg/$	0.239	0.199	0.014	1.990
CO_2 ton p.c.	3.402	4.162	0.015	20.208
GDP ppp p.c. 2011 international $	13,468.880	15,056.710	555.560	64,979.840
Population density, people / km^2	101.592	138.270	1.524	1,148.514
Undernourishment rate	15.514	13.546	0.000	64.500
Poverty 1.90$ at 2011 international $	21.177	22.040	0.000	84.740
Manufacturing GDP share	14.371	6.721	0.237	38.733
Services GDP share	70.065	12.734	29.279	93.881
Urban population share	53.685	22.754	12.082	97.818
Bottom 40% income share	15.970	4.130	7.510	28.024
NDVI annual mean	0.502	0.163	0.111	0.762
NDVI annual min	0.327	0.164	−0.027	0.657
NDVI annual max	0.655	0.161	0.170	0.862
Forest cover 2000 extent million ha	3.628	2.815	0.0005	9.883
Country area km^2	978,152	1,999,320	15,007	9,904,700

5.3.6 Transformation to degradation intensities

To address homogeneity concerns related to the scales of economies, we model standardized units of deforestation, pollution, and emissions, standardized per unit of GDP per capita in 2011 international dollars.

The choice to standardize degradation units by GDP per capita, and not by GDP, is for cross-comparability between countries of different economic size. In particular, using E to denote environmental pressure, the ratio $E/GDPpc$ is the environmental intensity of the average person's economic wealth, rather than the intensity of an international dollar. We favor $E/GDPpc$ over E/GDP because countries with larger populations will produce more international dollars and thus have lower intensities per dollar if everything else remains constant. The difference in efficiency of dollar production should not be used to suggest that average wealth production is environmentally more efficient. This is important because the hypothesis of the environmental Kuznets curve is that environmental pressure changes with increases of wealth, conventionally modeled as GDP per capita.

The cross-comparability is also statistically favorable because it reduces heteroskedasticity of the dependent variable across economic dimensions which allows to more robustly interpret variances, without the need of additionally modeling the conditional variance across covariates, such as discussed in Brown and Levine (2007). The general problem is that when the residuals vary strongly across the covariates, then apart from approximating a conditional mean function, one would need to approximate also the conditional variance function. The stabilization allows us to view the average standard errors as reasonable proxies, particularly given the additional approximation errors that a new non-parametric model of conditional variance would introduce.

Figure 5.2 at right shows that the variance in the log of the environmental intensities of GDP per capita is stable across GDP per capita, while the left plots with standardized intensities of international dollars contain widely differing variance. Note that the left-side is not log-transformed. While this would stabilize the data better, it does not change the relationship between the variance and GDP per capita.

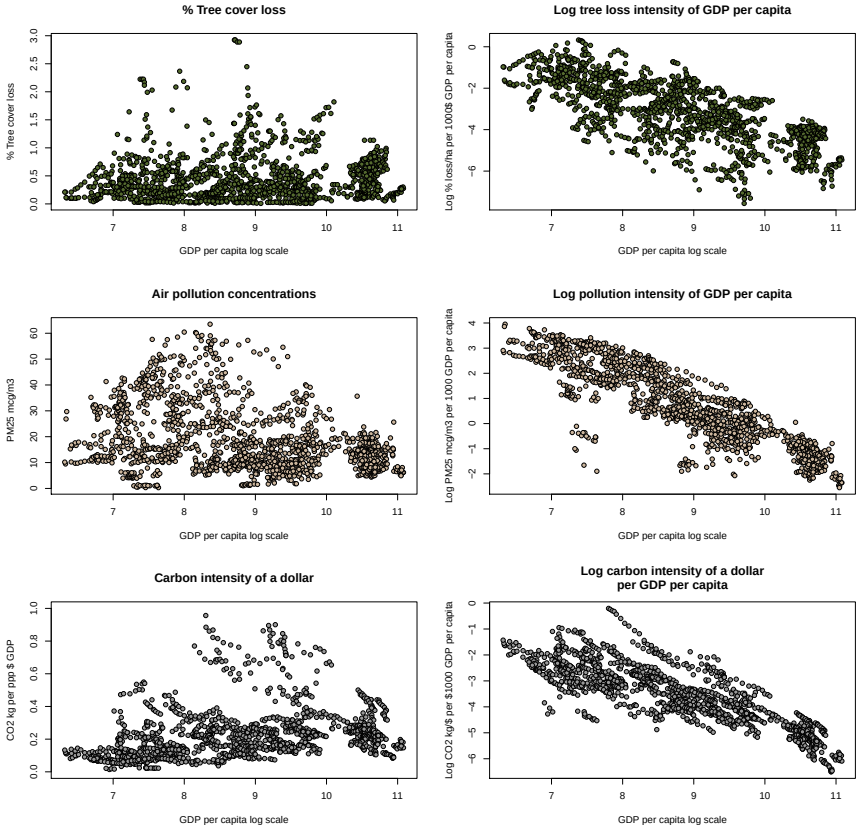

Figure 5.2: Observed degradation intensities and degradation levels across income.

Figure 5.2 immediately reveals a clear relationship between average income and the average environmental pressure for that level of average income. In particular, and unsurprisingly, the average log-linear trend is downward for all environmental pressures indicating that new per capita wealth is generated at lower environmental cost. For environmental pressure to go down on a net basis, we need that the emission intensity of average wealth declines faster than the increase in average wealth. Depending on the acceleration of one versus the other, various environmental output curves can result. Hence, the average trends in fig. 5.2 may

translate into inverted U-shapes in the environmental variables. However, since the left side is in logs, it means that minor deviations from the average may lead to large differences in degradation output explaining why inverted U-shapes are not directly visible in the left side.

5.4 Empirical results

Since, clearly, the environmental pressure of an average person's wealth decreases with GDP per capita, the empirical question is not whether the elasticity is non-zero, which is the question most regression analyses aim to tackle, but whether the elasticity implies that the intensity improves sufficiently fast across increasing income. Hence the empirical analysis of the Kuznets curve can be understood as an analysis of relative speed of change. In the following, we present the results in line with general models of the form

$$\log\left(\frac{E_{it}}{GDPpc_{it}}\right) = \beta_{it}x_{it} + \alpha_i\epsilon_t + \varepsilon_{it}, \tag{5.2}$$

where the β_{it}'s are approximated by our non-parametric regularized kernel estimator. Nevertheless, the interpretation follows as one would usually interpret the above. In particular, table 5.1 lists the variables that enter the regression, including the control variables, and the tables below list how they entered the regression. For example, we use a log transformation of GDP per capita, hence the interpretation is that of a standard elasticity. For a 1 per cent increase in GDP per capita evaluated at it, the quantity $\frac{E_{it}}{GDPpc_{it}}$ is estimated to change by β_{it} per cent. This means that when $\beta_{it} = -1$, a percentage increase in average income is associated with an efficiency improvement of a percentage. For marginal changes, this means that output levels stay approximately unchanged since scale, measured in average income, also increases by a percentage. All the other effects follow log linear interpretation, which is straightforward since the variables already represent rates. For a 1 point

per cent increase, a $\beta_{it} * 100$ per cent change in efficiency is expected, hence all parameters have an interpretation as elasticity.

The results show that the environmental output intensities are well explained by the data and that evidence for nonlinear dependencies is pervasive throughout all three models. Tables 5.2 to 5.4 show the marginal effects for individual models, summarized using the mean, quantiles and medians together with t-statistics.[6] For brevity, we have omitted the control variables in the tables.[7] All models have been checked for time fixed effects, but in all three cases the out-of-sample performance was optimal in the models without fixed components. The appendix contains conditional expectations together with confidence bands for each of the economic variables, holding the effects of all other variables constant at their mean values, which provide guidance throughout our discussion of the results. In particular, while the tables present the range of parameter estimates, it is not immediately possible to understand how effects of variables change along the levels of those variables. The conditional expectation plots plot the expected values for the outcome variables along the levels of individual covariates, and can therefore rovide a sense of the ordering of the local elasticities. Variables for which the marginal effects within the inner 50% of the percentiles range have an identical and significant sign are highlighted in the tables. This reveals that many of the variables contribute both positively as well as negatively to the output intensities depending on the data levels at which effects are evaluated. This shows that nonlinearities are important. We find that income has an unambiguous effect, all three environmental intensities improve with income but not sufficiently to offset scale growth.[8] While increases in

[6] We obtained our results using the R implementation of KRLS. Our out-of-sample shrinkage strategy is not implemented by default, and requires many model fits. We found that an optimized BLAS/LAPACK implementation provided better speed than the $C++$ implementation of bigKRLS.

[7] Annual mean, minimum and maximum NDVI values, forest cover, and country size

[8] The log-log specification allows for a simple interpretation. To offset the scaling effects, the marginal effect of log GDP per capita needs to be smaller than -1, which we do not observe within the 25%-75% range of effects.

GDP provide a basis for the improvement of production efficiency, it appears not to lower the net environmental output. However, as GDP increases, a structural change occurs in which poverty goes down and the share of manufacturing services, and urban population gradually increase. We will highlight several of these structural effects that are best visualized in fig. 5.7 and fig. 5.8. Figure 5.8 shows how poverty, the production composition, urban population shares, and the income distribution trend across GDP.

5.4.1 Individual model results

The results for deforestation show that early increases in population density correlate with a decrease in deforestation intensity while high population densities correlate with an increase. The trend across urban populations is a weak inverted-U. The effects of manufacturing and services are less ambiguous, the move out of an agricultural society and specifically an increasing share in services that occurs with increasing GDP, is a strong correlate of declining deforestation rates. There is some evidence that economies with an unequal income distribution retain a higher deforestation intensity of production. The effect of the poverty variables is, however, mixed as the semi-elasticities contain both negative and positive values. The conditional expectation plot in the appendix also visualizes that there is a very mild U-shape along poverty rates, with deforestation efficiency slightly going down again as countries move below 20 per cent. Reducing the undernourishment rate, in an opposite manner, initially seems to increase deforestation, while the transition out of extreme poverty correlates with a decrease in deforestation intensity. In contrast with the deforestation results, we see that an increase in population density unambiguously drives pollution up. The pollution intensity trend across the urbanization rate is initially flat, but after 50% of the population has urbanized, the trend becomes negative. This

Table 5.2: Deforestation intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
	Dependent: Log deforestation intensity of 1000 GDP p.c.			
Log 1000 GDP per capita**	-0.453***	-0.610***	-0.476***	-0.289***
	(-16.962)	(-22.856)	(-17.83)	(-10.822)
Population density	-0.001***	-0.003***	-0.002***	0.001***
	(-5.351)	(-18.303)	(-9.578)	(8.449)
Undernourishment rate	0.001	-0.008***	0.002	0.011***
	(0.371)	(-4.48)	(0.919)	(6.171)
Poverty 1.90$ rate	-0.005***	-0.013***	-0.004***	0.005***
	(-4.118)	(-10.353)	(-2.878)	(4.297)
Manufacturing GDP share	-0.015***	-0.055***	-0.016***	0.023***
	(-5.459)	(-19.402)	(-5.551)	(8.04)
Services GDP share**	-0.032***	-0.051***	-0.036***	-0.016***
	(-16.57)	(-26.109)	(-18.436)	(-7.961)
Urban population share	0.006***	-0.008***	0.009***	0.019***
	(5.597)	(-6.929)	(7.767)	(17.111)
Bottom 40% income share*	-0.027***	-0.06***	-0.033***	0.003
	(-5.416)	(-12.304)	(-6.671)	(0.55)
N = 1520	R^2= 0.922	λ=0.691.		*p<.1; **p<.05; ***p<.01

Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects.
Model controls for mean, min and max NDVI, forest cover, and country size.
**Approximately 50% of inner marginal effects same sign, but range includes zero.*
***Inner 50% of marginal significantly excludes zero.*

Table 5.3: Pollution intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
	Dependent: Log pollution intensity of 1000 GDP p.c.			
Log 1000 GDP per capita**	-0.691***	-0.842***	-0.692***	-0.567***
	(-47.899)	(-58.388)	(-48.026)	(-39.307)
Population density**	0.002***	0.001***	0.002***	0.003***
	(18.523)	(6.658)	(17.063)	(30.846)
Undernourishment rate	-0.002**	-0.008***	-0.003***	0.003***
	(-2.172)	(-8.675)	(-3.256)	(3.267)
Poverty 1.90$ rate*	0.003***	0.000	0.003	0.008***
	(4.493)	(0.371)	(4.842)	(11.603)
Manufacturing GDP share	-0.009***	-0.023***	-0.01***	0.004***
	(-6.436)	(-15.78)	(-6.756)	(2.492)
Services GDP share*	-0.007***	-0.012***	-0.007***	-0.001
	(-7.00)	(-11.68)	(-6.753)	(-1.286)
Urban population share	-0.006***	-0.015***	-0.008***	0.001**
	(-10.746)	(-24.929)	(-14.101)	(2.547)
Bottom 40% income share	-0.019***	-0.045***	-0.015***	0.012***
	(-7.776)	(-18.221)	(-6.086)	(4.682)
N = 1520	R^2= 0.978	λ=0.691.		*p<.1; **p<.05; ***p<.01

Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects.
Model controls for mean, min and max NDVI, forest cover, and country size.
**Inner 50% of marginal effects same sign, but range includes zero.*
***Inner 50% of marginal significantly excludes zero.*

Table 5.4: Carbon intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
	Dependent: Log carbon intensity of 1000 GDP p.c.			
Log 1000 GDP per capita**	-0.630***	-0.755***	-0.635***	-0.519***
	(-42.341)	(-50.71)	(-42.699)	(-34.903)
Population density	-0.000***	-0.001***	-0.001***	0.000***
	(-4.812)	(-11.919)	(-5.656)	(3.356)
Undernourishment rate	0.006***	-0.001	0.008***	0.014***
	(5.927)	(-0.523)	(8.248)	(14.741)
Poverty 1.90$ rate	0.002**	-0.002***	0.001**	0.008***
	(2.475)	(-3.295)	(2.092)	(11.496)
Manufacturing GDP share*	0.017***	0.001	0.019***	0.036***
	(10.815)	(0.366)	(12.344)	(23.015)
Services GDP share	0.001	-0.007***	0.002	0.01***
	(1.198)	(-6.708)	(1.667)	(9.01)
Urban population share	0.002***	-0.005***	0.002***	0.008***
	(2.826)	(-7.41)	(2.995)	(12.768)
Bottom 40% income share	0.006**	-0.018***	0.002	0.026***
	(2.269)	(-6.492)	(0.825)	(9.297)
N = 1520	R^2= 0.956	λ=0.635.		*p<.1; **p<.05; ***p<.01

Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects. Model controls for mean, min and max NDVI, forest cover, and country size.

**Inner 50% of marginal effects same sign, but range includes zero.*

***Inner 50% of marginal significantly excludes zero.*

indicates that early urbanization is polluting, but that after reaching a tipping point, the city environment becomes cleaner. The trends across manufacturing and services are also primarily downwards. Agricultural societies have a higher pollution intensity of income, while a shift into manufacturing and services reduces the environmental output per unit of production. It remains difficult to say whether the effects reduce pollution on a net basis as this structural transformation occurs jointly with an increase in total productivity. However, for an identical amount of total GDP produced, the data seems to suggest that an agricultural economy produces the highest amount of air pollution. An economy with a high manufacturing share produces less pollution, while an entirely service orientated economy outputs the lowest amount of pollution. This may also relate to a differential in value produced by these sectors which may imply different quality of production processes and differential in the total amount of economic activity for a fixed level of GDP. Across poverty

and undernourishment, we see hyperbolic effects that suggest that the eradication of extreme hunger occurs jointly with an increase in pollution intensity while later poverty eradication eventually occurs jointly with a reduction in pollution intensity. Poverty rates are unambiguously correlated with higher pollution intensities. Again, similar to the deforestation results, it seems that societies with high income inequality are also more polluting.

Carbon intensities also trend with urbanization. We find that the carbon intensities initially increase together with the urbanization process, however after the 50% urban population tipping point, the environment becomes more efficient in carbon consumption. The shift in production composition trends oppositely with those of deforestation and manufacturing. High manufacturing and high services share in the production composition both correlate with higher carbon emission intensities. The initial decline in undernourishment rates occur together with improvements in the carbon emission intensities, poverty reduction however trends with an increase. Finally, we see that equality − a stronger bottom 40% - increases carbon output when everything else is held constant, which is again an opposite trend of what we observed for deforestation and pollution.

5.4.2 Heterogeneity in environmental output

Combined, the results show that income and poverty reduction provide a basis for improvements in the efficiency of economies in their use of finite resources. The economic composition is not unambiguous in its effects. To understand how structural transformation, together with urbanization, poverty reduction and increases in total production, interplay to produce a commonality in environmental output trends, we track the model predictions keeping the control variables at their means. We also keep the income distribution fixed at a mean value as it does not trend clearly with

GDP as seen in fig. 5.8, and keep population densities fixed at means.

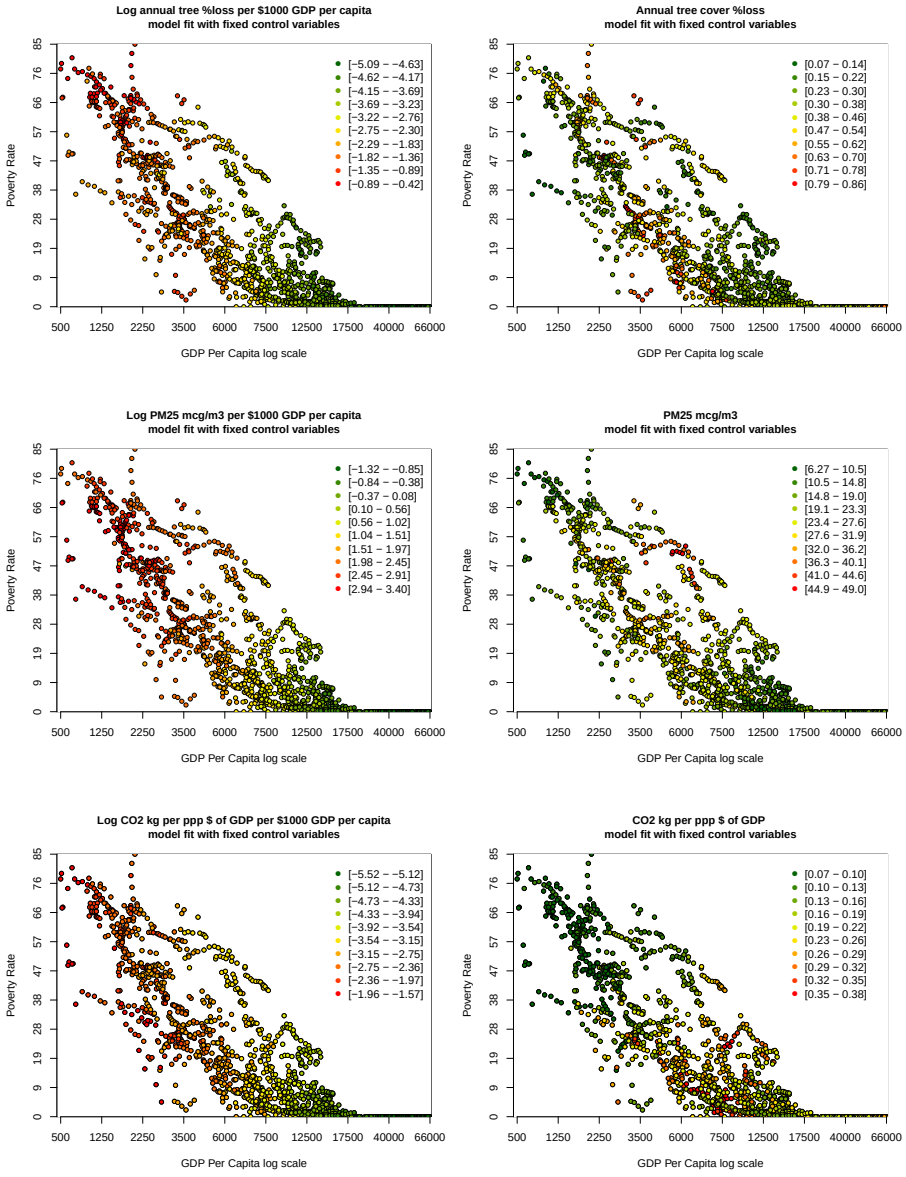

Figure 5.3: Model fits of degradation intensities of log GDP (left) and the rescaled environmental output levels (right) across poverty and income. Population densities and income equality as well as the control variables are held constant at the mean.

Figure 5.3 shows the prediction surfaces using poverty and income as Cartesian coordinates. The model predictions fit the output levels well after scaling the log intensities, see fig. 5.9. While this shows that all countries gradually grow out of poverty and improve their efficiencies following a common pattern, it also reveals that there is significant heterogeneity in the environmental output intensities that relates to differences in poverty and hunger rates, urban population shares and GDP composition. This highlights that the shape of the environmental Kuznets curve strongly depends on the development path of a country across all its dimensions. Furthermore, while the progression in output intensities follows a similar path, slight deviations from the local average may result in large differences in total environmental output. This reveals that while different development paths may relate to relatively small differences in the environmental output intensities, it may produce rather large differences in actual forest loss, air quality and carbon emissions depending on the scale of the economy.

An important takeaway is that heterogeneity in the actual output levels (right), is primarily large around the income levels where output is also highest (around \$4,000 for deforestation, \$6,000 for pollution, and \$8,000 for the carbon weight of a single dollar production value). This indicates that the theorized environmental Kuznets tipping points are also the points at which an averaged result, such as obtained from a linear regression, provides the poorest indication of relationships at the individual country level. While a few general rules could be extracted from the marginal effects, such as inequality, income and population density effects, the larger part of the environmental data seems to relate heterogeneously to economic variables.

5.4.3 Average curvature

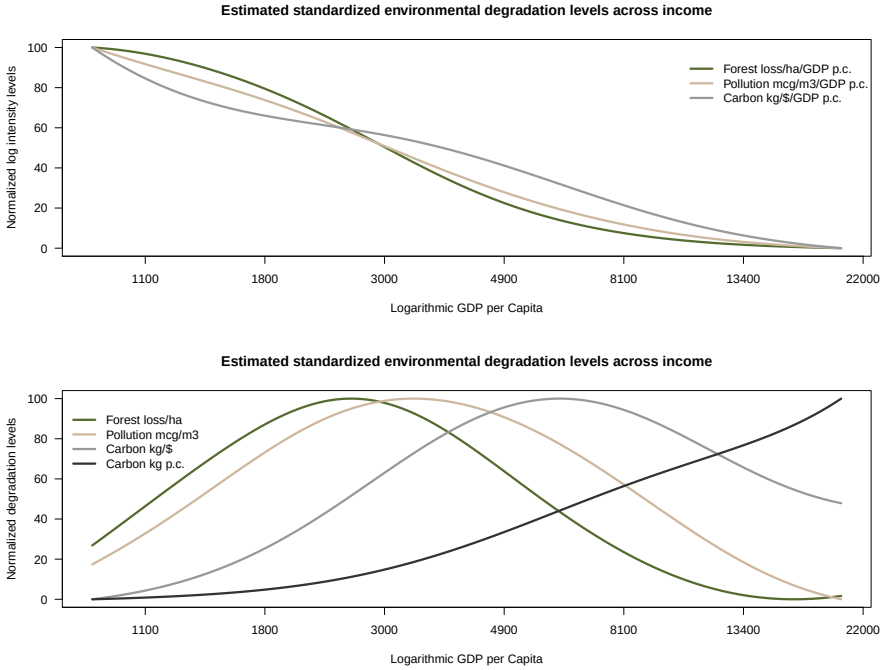

Figure 5.4: Normalized predicted environmental output levels across income. Predictors are held at expectations conditional on GDP. The R^2's from logarithmic GDP per capita to poverty, undernourishment, manufacturing, services and urban population shares are respectively 0.801, 0.633, 0.142, 0.573 and 0.739. The conditional expectations are plotted in fig. 5.8. Population density, income equality, and controls are held at their means.

The heterogeneity in amplitude, and location of tipping points, conditional on the economic variables, implies that a single Kuznets Curve, such as it has often been treated in the literature, is a description that applies only poorly to individual country cases. However, to do some justice to the classical concept we can still construct an average development path and explore how the models fit environmental outputs to that. To do so, we derive conditional expectations for poverty, undernourishment, GDP composition, and urban population shares, using only the logarithmic GDP per capita as an explanatory variable. We then use these conditional

values to build a data set that includes all variables as local averages along with GDP itself. Again, we keep the control variables and the income distribution as well as population densities fixed. We normalize the results to compare the slopes and location tipping points across income.

Figure 5.4 shows the curvatures associated with these development paths. We have dropped the lower 2.5% of GDP observations, and the upper 20%. We focus on this range because of its particular relevance for development policy. We note that the maximum total output associated with the average tipping point is an interesting statistic, but due to heterogeneity this may be a poor approximate to predict whether a country is close to its potential tipping point after observing only environmental output. The deforestation rate associated with the average development path attains a maximum of .66% annually, while that highest pollution concentration maxes at 28.7 $\mu g/m^3$ and the carbon weight of a dollar reaches 0.271 kg.

5.4.4 Heterogeneity in curvature and tipping points

The average pathways accurately describe the transition out of poverty, but they provide less insight into the effects if the economic composition changes. To better understand the importance of deviations in transitional variables, we plot the degradation levels associated with the average development path with additional differences in manufacturing shares, urban population shares and poverty rates.

Figure 5.5 shows that changing these variables, while keeping everything else at the local averages, has important impacts on the location, shape, and height of tipping points. For example, increasing the share of manufacturing by 10 points, shifts the tipping point of deforestation to the left, while economies that retain high agricultural shares reach a tipping point at higher income. This implies that an earlier transition out of agriculture may prevent high deforestation rates at higher income

and lower pollution levels at its peak. This is a slightly counter intuitive result as manufacturing has traditionally been portrayed as the main source of pollution. However, since our data only indicates the share of manufacturing in total GDP and not the quality or quantity of goods produced, higher rates may also correspond to differences in the number of manufacturing sites and the methods of production used.

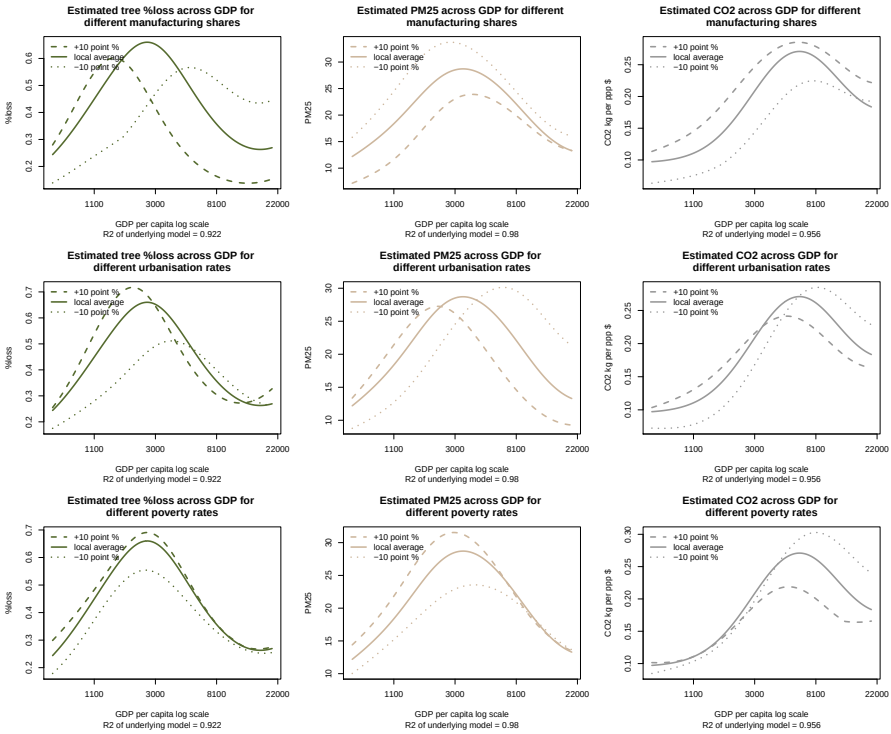

Figure 5.5: predicted environmental output levels across income. Predictors are held at expectations conditional on GDP and one variable has been incremented +/- 10 points in each plot. The local average trend is identical to those in fig. 5.4.

The carbon emissions associated with this structural change are higher, suggesting a trade-off between pollution-heavy and carbon-intense production. In a similar fashion, poor countries that have a high urbanization rate have higher deforestation rates and reach a pollution peak faster. Poor countries that have lower urbanization, on the other hand, even-

tually maintain higher pollution and carbon emissions levels at higher income. This suggests that the draw-down in pollution output is not just a matter of income and productivity, it may relate to attaining critical urban population mass combined with increased income. The effects of poverty, finally, do not impact the location and shape of the environmental output levels. Countries with high poverty rates unambiguously deforest and pollute more, but emit less carbon.

5.4.5 Exploring degradation dynamics under simple 2030 scenario's

To explore whether continuation of current growth can be expected to lower environmental outputs without intervention, we extrapolate GDP into the future and calculate associated model responses under three simplistic scenarios of growth. These explorations are intended to further assess the potential impacts implied by the relationships that are captured by the models. They are by no means an attempt at accurate forecasts of future developments as these will be driven by a wide range of factors and events that are not part of historical data (e.g. unforeseen technological developments, changes in societal preferences, policy agendas). The estimated models can still be applied, however, to sketch how future environmental pressure may advance under current economic and population growth trends in the absence of policy interventions, or new technological successes, based purely on historical relationships. This still provides relevant indications of the magnitude of efforts required to meet environmental objectives.

In a base scenario, this analysis lets each sovereign grow at individual median 1999-2014 compound rates, with the highest growth rate capped at the 90% percentile (5.27% annually). In a pessimistic future, each country continues at one asymmetric deviation unit (-3.67%) below the base rate, and in an opportunistic growth scenario, countries continue

one asymmetric deviation unit (+0.89%) above the base rates.[9] In the opportunistic scenario, rates are capped at the 95% percentile rate (5.87% annually). Finally, we construct Business As Usual (BAU) as the average of the three results to balance between possible asymmetry. Table 5.5 summarizes the growth rates assumed in these simple scenario's.

To limit complexity, we keep population growth slightly below individual country median compound rates, resulting in 8.5 billion people by 2030 (in line with United Nations projections).[10] We use extrapolated GDP levels to derive fits for poverty and undernourishment levels, and the GDP composition using univariate model fits of the penalized kernel model. We let the urban population depend additionally on log population densities.[11] At each point in extrapolated time, we compare the conditional expectations to the predictions of our base year (2014), and compute the percentage change that we then multiply with observed 2014 values. We keep all data points within observed intervals, including a cap on the sum of agriculture and services shares. This means that effectively after reaching a level of $64,980 per capita, our projection halts both the income effect on efficiency improvements and the effect of scale increase on the environmental output for a country (again, highlighting that this analysis does not consider future technological successes beyond what has been achieved by societies so far). In the pessimistic scenario, this does not affect any individual country result, in the base scenario this fixes Norway's output at current levels and caps those of the U.S. and Switzerland in respectively the last 4 and 5 years of the projection (final

[9] Asymmetric deviation units have been calculated as the difference between the median and respectively the 25% and 75% quantiles of growth rates. In the calculations we have dropped the two largest outliers (in absolute value) for each country.

[10] We reduced the population growth rates by 0.05 times the absolute point percentages globally to reduce growth everywhere, then reduced population growth rates by an additional .15 times the percentage rates in the top 40% income countries and additional .25 times in top 20% income countries. This simple scenario is designed to represent relatively higher growth in lower income countries and a slow down in developed countries, in line with UN projections.

[11] The R^2's of the models are, 0.632 for undernourishment, 0.785 for poverty, 0.142 for manufacturing, 0.573 for services, and 0.814 for urban population shares. The uncertainty of the impact of changes in manufacturing remains high in our results.

9 and 10 years in the high growth scenario).

Figure 5.6 presents our results at the global level made by aggregating all country-level results and assuming that the average in-sample trend scales appropriately with missing areas. Results are also available for income segments in table 5.6. In the average scenario, global extreme poverty falls below 7.4% of the global population. The poorest 20% countries in our sample have stronger successes and go from 45% poverty to just over 33%.

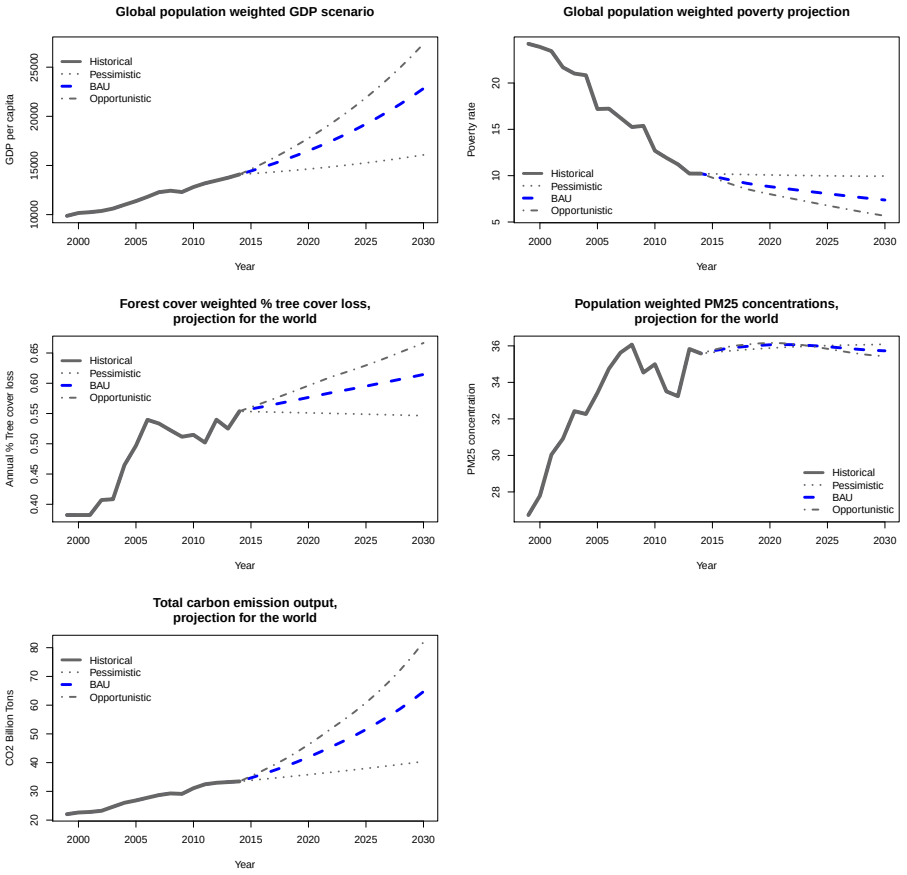

Figure 5.6: Projection for global environmental outcomes. Table 5.6 contains aggregate statistics of the BAU line and highlights the distributional changes across income.

While poverty reductions and GDP increases may improve livelihoods through economic gain, air pollution remains a serious threat to wellbeing, as the average global citizen remains exposed to 36 $\mu g/m^3$, nearly twice the WHO prescribed guidelines. Additionally, development comes at a cost of an annual carbon output that reaches 63GT which is nearly a doubling from the 2014 levels, and a total loss of 242 million hectares of forested land. About 58% of forest loss is in countries with poverty rates above 3% or income in the bottom 60%. Many of those countries have tropical rain forests with slow regrowth rates estimated at 27% (Hansen et al., 2013). Using those statistics, loss in these countries totals 136 million ha between 2014 and 2030, netting over 3.4% of the global 2000 dense forest cover. Other insights include that success in eradicating poverty likely slows as China and India near 0% poverty and populations in poor countries grow faster than those in the developed world. Our modeled data does not signal that development alone will result in successful slowdown in natural capital depletion. At the global level, results suggest ongoing increases in global deforestation rates and carbon emissions. Global pollution exposure stabilizes regardless of the growth scenario, the results instead suggest a distributional shift toward lower income countries with improving and worsening conditions balancing out at the global scale. Air pollution concentrations rise by 28% in the bottom 20% income countries. Table 5.6 shows that the entire bottom 30% income countries of our sample in fact continues to face increasing pollution exposure. Projections of forest cover and carbon emissions on the other hand, are heavily dependent on the economic outlook. Growing wealth in the developing world together with rapid population growth may accelerate future global carbon output. A more extensive discussion on differences in trends in relation to development, and a breakdown of carbon output along income quintiles is provided by Andree et al. (2019).

5.5 Discussion and conclusion

In this paper, we estimated a penalized non-parametric model of environmental output across economic development. This type of modeling works well for nonlinear processes that do not result in overly complex dynamics. We deployed the framework to study environmental data in a panel of 95 countries. We modeled satellite-derived deforestation and air pollution levels and reported carbon emissions. To deal with heteroskedastic variance, we transformed our data to logarithmic degradation intensity of per capita GDP. We used a cross-validation approach to decide which fixed effects should be part of the model and this did not support the inclusion of time fixed effects.

Our results suggest that production gradually favors conserving the earth's finite resources as GDP increases, but that this alone is not sufficient to offset the scale effect of growth. Instead, structural change in the economy shapes environmental output curves. This process shares similarities between sovereigns, but remains largely heterogeneous. These results do not support a single environmental Kuznets rule. Instead, the results emphasize the importance of local economic conditions on environmental results. Across all data levels, some effects hold unambiguously. Poverty and income inequality correlate with higher pollution, higher deforestation, and lower carbon emissions; agricultural GDP shares correlate with deforestation; population densities correlate with pollution; and higher manufacturing shares correlate with increased carbon emissions. We find various tipping points in other variables, notably across urbanization rates. While local conditions may be unique, average development is associated with an inverted U-shape in deforestation, pollution and carbon intensities of production units. Per capita carbon emissions, however, follow a J-curve as the increase in per person productivity is not sufficiently offset by efficiency improvements. Disregarding the level of per capita GDP, we observe that at least one form of natural

capital degradation is high, conflicting with the belief that countries tend to "clean up" as they develop. One could argue that the scope of the impacts of externalities to production increases with development, with the burden falling to increasingly distant households both in time and space. Although local air pollution may be more intrusive on daily life, the consequences of climate change will remain globally impactful for generations to come.

We extrapolated our descriptions forward in time to highlight the daunting implications of development under continuation of current practices without improving policies. Our results are generally in line with emission paths associated to the high radiative forcing scenarios considered in IPCC's 4.9°C world (RCP 8.5). Our projections did not indicate successes on the fronts of reducing deforestation. Air quality improves in some currently severely polluted places, but worsens in poor regions.

In our results, deforestation follows an inverted U-shape across average development in the developing countries. This confirms and extends recent results from Crespo Cuaresma et al. (2017) that provide evidence for a partial environmental Kuznets curve for forest cover at low income. However, we find that economic growth alone is not sufficient to halt forest loss, and we find evidence that within the bottom 60% income countries, deforestation shifts to the bottom 30%, and that countries within the top 40% income do not fully stop deforesting. Others have similarly detailed forest loss in high-income countries, for example in the United States (Sleeter et al., 2012). Future efforts should also aim to understand forest regrowth dynamics across economic development, as we have only used average forest growth rates over the entire study time period as a control variable in our models, rather than investigating how regrowth possibly changes conditional on economic indicators. Generally, the temperate zones have much better regrowth rates. Taking this and projected increases in the bottom 20% into account, the African

forests seem to be at increased risk as economic successes in these areas accelerate, while the Amazon faces only marginal improvements in the immediate future in our modeled projections. Generally, deforestation is related to the economic value of land. Urban land, for instance, can be valued hundred times higher than forest land in the same area (Alig and Plantinga, 2004). Agricultural land usually also yields higher returns and policies focused on protecting forest could address this value gap (e.g. (Hyde et al., 1996)). The payment for ecosystems services schemes may provide an opportunity to conserve essential natural resources while providing an income source to landowners. However, the governance and targeting of these programs must be carefully addressed in order protect both the resources and livelihoods of those dependent upon them (Landell-Mills and Porras, 2002; Grieg-Gran et al., 2005). Extending agricultural subsidies to include renewable perennial crops has the potential to make cleaner alternatives competitive without negatively impacting farmer income or the need to increase aggregate subsidy spending, and could be a way to ensure that environmental damages are at least in part reconciled by positive externalities (Andree et al., 2017b). Other policy interventions that address forest-cover loss can focus on conservation and land-use protection, sustainable forestry, and urban growth boundaries (Alig et al., 2010). The efficacy of these interventions will likely rely upon the local circumstances surrounding forests and nearby populations, yet the potential benefits may be felt globally.

On the pollution side, our model projects rising $PM_{2.5}$ levels in the lowest 30% income countries, with a general decrease in $PM_{2.5}$ in middle income countries. $PM_{2.5}$ remains far above WHO air quality guidelines in many countries, particularly in lower and middle income groups. Given population growth, these levels will expose more people to pollution-related health risks. Currently, about 90% of the global population is exposed to air quality that does not comply with the World Health Organization's Air Quality Guidelines (World Health Organization, 2016).

Tallis et al. (2018) expect that by 2050, business-as-usual development will result in over 4.8 billion people living in countries with worse air quality than in 2010. As a comparison, in our average modeled data 52% of people currently live in places where air quality has worsened by 2030. This totals to approximately 4.4 billion by 2030. Exposure to unsafe levels of particulate matter is estimated to increase the number of premature deaths related to air pollution in coming decades, killing 4.5 million people (or more) by 2030 (International Energy Agency, 2016; Lelieveld et al., 2015). Currently, $PM_{2.5}$ levels peak in middle income countries, and while pollution levels can generally be expected to decline in these countries as their income levels grow, pollution levels will still remain dangerously high in this group. These countries include highly populated areas such as in China, India, and Bangladesh, which have already been identified as hotspots for adverse impacts of air pollution in the coming decades (Organisation for Economic Cooperation and Development, 2016; Pozzer et al., 2012). Eradicating poverty in these places may be one contribution to lowering extreme pollution, but unfortunately there is also evidence that points out that low-income households are also those that are more severely affected by pollution in economic terms Andree et al. (2019).

On the carbon end, our results suggest emission levels that could lead to the high radiative forcing scenario in IPCC's 4.9°C world (RCP 8.5) are largely in line with business-as-usual development in which developing countries follow in the footsteps of wealthier countries. Worse scenarios may in fact be considered as relevant possibilities. Specifically, this could occur if developing countries do not successfully manage to adopt cleaner technologies, or if high income countries revert (part of) their pro-climate policies. Recent studies suggest we are not alone in such a conclusion. See for example Peters et al. (2012) and comments, suggesting - in line with our findings - that reported successes in carbon reduction are short-lived and largely relate to the 2008-2009 crisis and aftermath. Emissions rapidly

increased in many places with the recovery. Furthermore, Peters et al. (2013) and comments thereon reveal that recent emissions continue to track the high end of suggested emission scenarios, making it increasingly unlikely that global warming will stay below 2°C. This is in line with our result that continuing current development puts the world on emissions associated with a 4.9°C pathway. This is further substantiated by the conclusion that developments on the fronts of negative emissions are required to reach a 2°C future Gasser et al. (2015). Combined, the evidence suggests that a worst-case scenario over 4.9°C in 2100 is both not unrealistic and overlooked in both the scientific community and the political arena.

5.6 Appendix

5.6.1 Additional results and figures

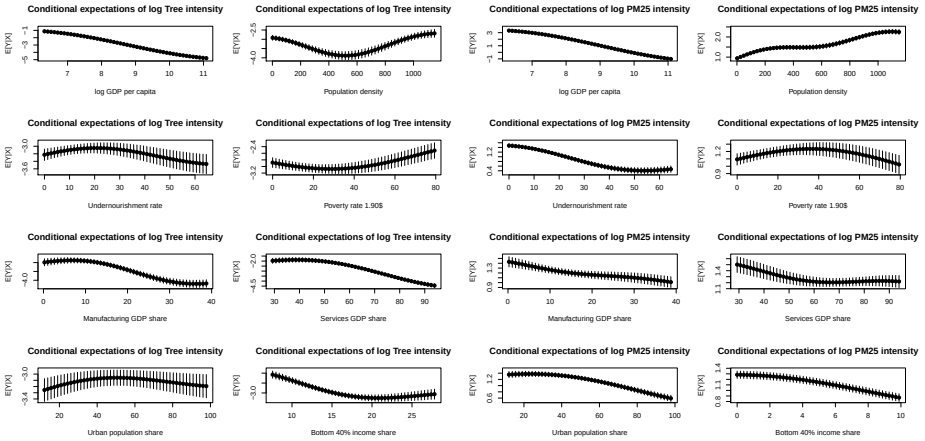

Figure 5.7: Conditional expectations of deforestation (left two columns) and pollution (right two columns) intensity of income for each variable fixing other variables constant at their mean.

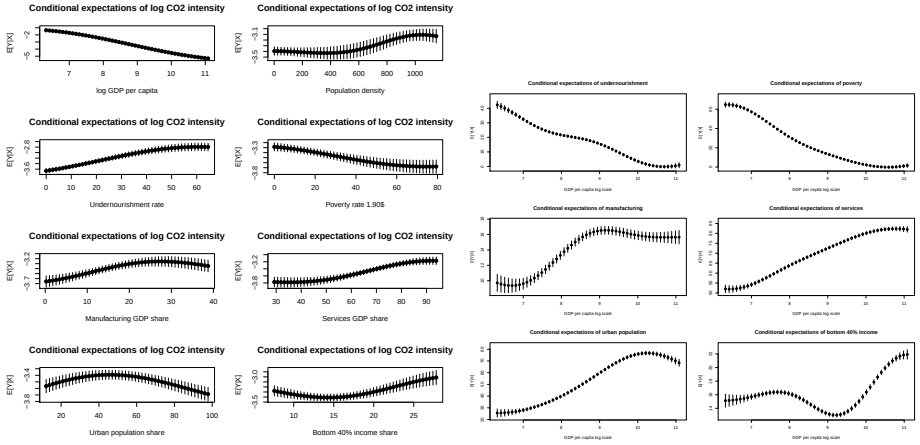

Figure 5.8: Conditional expectations of carbon intensity of income for each variable keeping other variables constant at their mean (left two columns).

Table 5.5: Summary of base rates in percentages by 5-percentiles used in the projection.

	GDP	Population
1	1.33	2.57
2	2.78	2.54
3	2.22	2.28
4	2.93	2.10
5	3.02	1.99
6	2.49	1.98
7	3.69	1.40
8	2.55	1.58
9	2.62	1.51
10	1.67	0.96
11	3.62	0.96
12	2.31	1.31
13	2.97	0.83
14	2.82	1.04
15	2.93	1.15
16	3.85	0.85
17	1.10	0.46
18	1.47	0.30
19	1.47	0.23
20	1.24	0.49

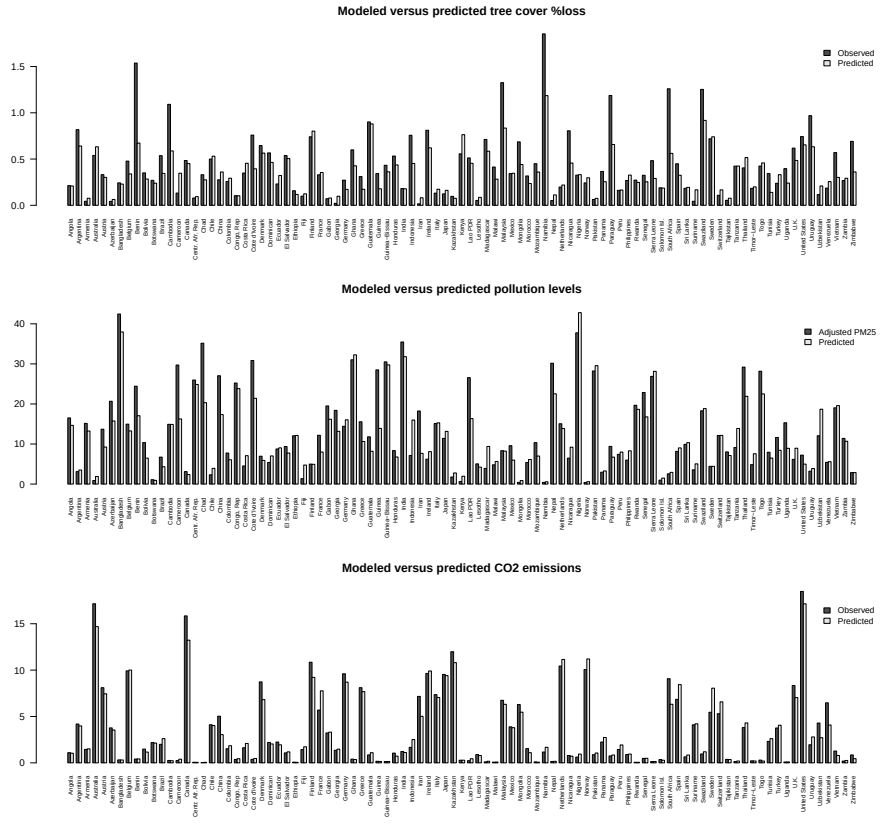

Figure 5.9: Accuracy of predicted degradation levels at high income.

Table 5.6: BAU-2030 and base year data aggregated by 5% percentiles of income. GDP per capita as population weighted averages, population in millions, number of poor people in millions, annual tree loss in square kilometers, $PM_{2.5}$ in population weighted average concentrations, carbon emissions in million tons. World totals are scaled to world totals using multipliers (1.16 for population and carbon based on the share of population in our data, and tree loss 1.42 based on the share of tree cover in the data).

Income Group	GDP p.c. 2014	GDP p.c. 2030	Pop. 2014	Pop. 2030	No. Poor 2014	No. Poor 2030	Treeloss 2014	Treeloss 2030	PM25 2014	PM25 2030	CO2 2014	CO2 2030
1	1,094	1,548	68	101	41	55	3,382	3,640	15	20	10	28
2	1,495	2,611	171	262	74	70	3,750	3,977	17	21	20	109
3	1,899	2,388	46	67	19	24	1,415	1,684	23	25	22	39
4	2,305	3,308	96	139	36	43	1,821	1,931	29	31	27	71
5	2,917	4,720	237	305	57	45	1,295	1,331	47	52	96	285
6	4,249	5,381	265	366	43	48	5,782	5,713	46	48	190	282
7	5,265	9,887	111	133	6	2	4,716	4,031	23	20	170	385
8	5,426	10,301	1,519	1,936	341	288	4,440	3,769	54	51	2,286	5,634
9	6,875	9,344	176	231	21	22	3,297	3,043	12	13	212	399
10	8,219	10,627	15	17	1	1	4,002	3,395	9	9	14	22
11	10,078	16,232	292	352	22	11	17,799	17,676	15	15	540	1,338
12	11,993	16,578	114	141	11	11	3,784	4,037	15	17	646	1,150
13	13,135	24,563	1,691	1,825	49	5	33,540	39,407	44	44	11,212	25,839
14	16,347	20,955	206	244	4	3	2,032	2,132	22	24	1,128	1,864
15	18,386	28,014	162	193	3	2	5,160	5,859	17	17	750	1,402
16	22,607	37,980	53	60	0	0	1,822	2,194	20	24	438	1,112
17	33,000	39,059	204	219	0	0	6,647	8,832	21	22	1,160	1,770
18	38,515	47,345	244	253	0	0	26,962	32,972	12	13	2,324	3,324
19	43,584	53,144	128	132	0	0	4,694	5,794	15	16	1,276	1,788
20	51,694	64,462	354	384	0	0	17,649	19,466	10	11	5,534	7,414
world	14,088	19,899	7,137	8,537	730	630	218,664	242,656	36	36	32,544	62,934

5.7 Supplementary note to the chapter

This supplementary note provides additional methodological discussion around an adaption of Kernel Regularized Least Squares to the dynamic panel context, paying specific attention to automated selection of fixed effects. The model provides an attractive approach to both nonlinearity and interpretability within one integrated framework. Importantly, it allows deriving marginal coefficients at the observational level that are similar to those of parametric models, while leveraging the flexibility provided by the similarity learning framework. Unlike in the standard regression context, the interpretation of these marginal coefficients depends on externally set tuning parameters. The paper discusses the role of the regularization parameter in the interpretation of these basic quantities of interest. The discussion highlights that rigorous hyper-tuning, and out-of-sample prediction performance of models in general, is crucial, even when one is merely interested in inference and not in prediction.

5.7.1 Introduction

The UN's Sustainable Development Goals for 2030 aim on one hand at inclusive growth and eradicating poverty, and on the other at preserving environments. The crucial relation between development and the environment has been studied extensively since the 1990s, and has been revisited recently in the main article associated with this background note. The paper applied a Kernel Regularized Least Squares model on disaggregate data obtained from remote sensing sources to model environmental-economic trends heterogeneously across a number of economic indicators at country level. Results suggested that local economic circumstances played an important role in determining the shape, amplitude, and location of tipping points in environmental output.

This note details how the framework was adapted to the panel context,

paying particular focus to automated selection of fixed effects. The discussion here also goes more deeper into the types of assumptions that are implicitly made about environmental-economic interactions by adopting the kernel approach. The model is attractive as it provides a straightforward approach to nonlinearity and interpretability without having to rely on surrogate approaches, such as the Local Interpretable Model-agnostic Explanations method that locally interrogates a model's output surface (Ribeiro et al., 2016). The interpretation of the marginal coefficients is, however, highly dependent on an externally set hyper-parameter that is not part of the estimated vector of parameters that define the model itself. Instead, consistency results for the Regularized Kernel model are toward a, possibly pseudo-true, parameter for which the limit result is separately defined for each level of penalization. This paper discusses the role of tuning the penalty, or regularization parameter, in ensuring that the marginal coefficients, and associated standard errors, admit to a standard interpretation. The model of interest, and limit theory for it, is provided in Hainmueller and Hazlett (2014). For the more general reader, most of the discussion here is posed in a general way and stretches out to many other relevant cases.

The remainder of this writing is organized as follows. Section 5.7.2 introduces the modeling framework and details the adaption to the panel setting. Section 5.7.3 provides further discussion around the tuning of the penalty and its relationship to the interpretation of the estimation result. Section 5.7.4 concludes.

5.7.2 The modeling framework

Machine learning methods are often developed with different applications in mind than the classical regression models that have been developed primarily for economic inference. Because the limit results in non-parametric models depend on externally set tuning parameters that are not part of

the vector of estimated parameters, it is not immediately clear whether the estimates can be interpreted similarly as those obtained using parametric methods.

As researchers continue to tackle high dimensional problems, such as frequent in the environmental economics domain, we anticipate that machine learning methods will become more popular in the context of inference. For the sake of those less familiar with these approaches we summarize the basic assumptions and key features of the applied non-parametric kernel estimation below, relevant in the setting of the companion paper. We do not introduce new theoretical results, instead we aim to provide an overview that highlights differences with respect to parametric estimation. Particularly, we detail the role that regularization plays in correct inference. Readers that are familiar with penalized kernel models may proceed directly to section 5.7.2, in which we explain how we treat fixed effects in the estimation.

In the following, let $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ denote the sets of the natural, integer, and real numbers. $\mathbb{R}_{>0}$ includes all positive, non-zero, reals. For a set $\mathcal{A}$, we use $\mathfrak{B}(\mathcal{A})$ to denote the Borel σ-algebra over $\mathcal{A}$. We use $t, ..., T \in \mathbb{Z}$ to index time, and $i, ..., N \in \mathbb{N}$ to index cross-sections, $it, ..., NT \in \mathbb{N} \times \mathbb{Z}$ labels all locations in space-time. We use boldfaced letters, e.g., $\mathbf{a} \in \mathcal{A}$ to denote vectors. Furthermore, $\times_{t=1}^{t=T} \mathcal{A} = \mathcal{A}_T$ denotes the Cartesian product of T copies of A, and $\mathcal{A}_\infty = \times_{t=-\infty}^{t=\infty} \mathcal{A}$ is the Cartesian product of infinite copies. For two maps f and g, $f \circ g$ is their composition resulting from a point-wise application, and $\langle \cdot, \cdot \rangle$ denotes the inner product space.[12] Finally, $\| \cdot \|_{\mathcal{A}}$ denotes a norm on $\mathcal{A}$.

[12] As a generalization of the dot product in the Euclidean space, to higher dimensional spaces including infinite dimensional Hilbert spaces.

Assumptions about the data generation process

Suppose, we observe an n_x-variate T-period sequence $\mathbf{x}_T := \{\mathbf{x}\}_{t=1}^{T}$ that describes the state of one economy throughout time. At each point in time, we observe N trajectories of this n_x-variate sequence, i.e., we focus on repeated cross-sectional vectors of length N describing the evolution of N economies. Each vector contains observations of for example income and the composition of the economy, all indexed over a set of locations $i, ..., N$. The matrix $\mathbf{X}_t$ consisting of n_x columns describing different variables $\mathbf{x}$ and N rows describing the different locations, is indexed by time. We consider a second, repeated cross-sectional sequence $\mathbf{y}$, of degradation levels generated by:

$$\mathbf{y} := \{\mathbf{y}_t = h_0(\mathbf{X}_t), t \in \mathbb{Z}\}. \tag{5.3}$$

We can observe $\mathbf{y}_T$, a subset of the results of this process $\mathbf{y}_T := \{\mathbf{y}\}_{t=1}^{T}$. The function $h_0 : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ produces environmental output for every coordinate $\mathbf{X}_t \in \mathcal{X}$.[13] We assume that h_0 is a unique measurable function that for each coordinate $\mathbf{X}_t \in \mathcal{X}$ assigns a *true* value $\mathbf{y}_t \in \mathcal{Y}$ for all $t \in \mathbb{Z}$. In a sense, by assuming this particular form, we assume that the environment does not endogenously degrade itself, i.e. that $\mathbf{y}$ does not endogenously generate itself. Instead, this assumes that the evolution of degradation levels for each economy $\mathbf{y}$ is symptomatic to external, local economic development variables $\mathbf{X}$. This does not exclude the possibility that $\mathbf{y}$ may in part affect elements of $\mathbf{X}$, it requires however that feedback effects are invertible, in turn implied by some form of stability, and follow levels in $\mathbf{X}$ such that h_0 describes the net relationship between $\mathbf{X}$ and $\mathbf{y}$.[14] We also assume that h_0 is smooth, particularly that it maps

[13] Particularly a $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable mapping as kernels with a universal approximating property require at least that the target function is measurable, see for example (Micchelli et al., 2006).

[14] If $\mathbf{y} = f_0(\mathbf{X}) + g_0(\mathbf{y})$ with g_0 describing simultaneous feedback and f_0 describing the contemporaneous exogenous effects, then one can also write $\mathbf{y} = h_0(\mathbf{X})$ if g_0 is invertible, with $h_0(\mathbf{X}) = (I - g_0)^{-1}(f_0(\mathbf{X}))$, hence h_0 arises from the composite $(I - g_0)^{-1} \circ f_0$ and describes the combined effect of exogenous impulses and feedback.

similar coordinates in $\mathcal{X}$ to similar values in $\mathcal{Y}$. This implies that for each state of the economy at each point in time $\mathbf{X}_t$, we observe a level of deforestation, pollution or carbon $\mathbf{y}_t$ that is induced by the state of the economy through the function h_0, and that for small changes in the state of an economy we expect to see small changes in environmental output. Furthermore, it assumes that for two economies that are similar in terms of composition and scale, we expect similar environmental output.

Panel Kernel Regularized Least Squares

In the current case, the environmental Kuznets theory suggests an inverted U-shape between degradation and economic development. The relationships may of course be of a completely other form or differ across environmental variables, while ideally we keep the analysis of both relationships within a similar regression framework. We therefore postulate a very flexible regression of the form

$$\hat{\mathbf{y}} := \{\hat{\mathbf{y}}_t = h(\mathbf{X}_t; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}. \tag{5.4}$$

Our modeled function h is defined as a mapping $h : \mathcal{X} \times \Theta \to \mathcal{Y}$, where Θ is the parameter space. In parametric regressions, Θ is assumed to be compact and finite dimensional. This immediately imposes structure on h, thus translating into assumptions about h_0 if we maintain a belief that $\boldsymbol{\theta}_0 \in \Theta$. By reducing the size of Θ we simplify the possible structure of h, i.e., chances that $\boldsymbol{\theta}_0 \in \Theta$ become increasingly slim. While we minimize assumptions about h_0 by working with Θ as an infinite dimensional space, some assumptions about h_0 are unavoidable as Θ has to be parametrized eventually. In our example, as we shall see, one still has to specify radial basis functions.

In parametric regressions, Θ plays a key role as the Euclidean space containing all the possible coordinates of potential parameter vectors $\boldsymbol{\theta}$.

In the non-parametric case, there is a subtle difference. Suppose that for every $\boldsymbol{\theta} \in \Theta$, there is a function $h(\cdot;\boldsymbol{\theta}) : \mathcal{X} \to \mathcal{Y}$ that is $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable. We can define $\mathcal{H}_\Theta(\mathcal{X})$ as the Hilbert space containing an infinite collection of functions $\{h(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ generated by Θ. We shall use a simplified notation to reduce cluttering and instead write that $\boldsymbol{\theta}$ indexes the functions $h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$. The common notation $\boldsymbol{\theta}_0 \in \Theta$ is thus equivalent to saying $h_0 \in \mathcal{H}_\Theta$, i.e., $\boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t;\boldsymbol{\theta}_0) = h_0(\mathbf{x}_t) \; \forall \; \mathbf{x}_t \in \mathcal{X}$. This clarifies that, while a in a parametric regression problem where we are fore-mostly concerned with searching a compact parameter space Θ for the parameter vector $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$, in the current framework we are explicitly interested in searching across a space of functions produced under some process of generating flexible functions from simple parameter vectors given the sample space, $h_\mathcal{X}$, for infinite $\boldsymbol{\theta} \in \Theta$, for the function that best resembles the target function $h_{\boldsymbol{\theta}} \to h_0$. Specifically, each $\boldsymbol{\theta}$ indexes a member in $\mathcal{H}_\Theta(\mathcal{X})$ according to the map $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ with $h_\mathcal{X}(\boldsymbol{\theta}) := h(\cdot;\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \; \forall \; \boldsymbol{\theta} \in \Theta$. Hence, we can write the estimator also as

$$\hat{h}_T := \arg \min_{h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta} Q_T(\mathbf{y}_T, \mathbf{X}_T; h_{\boldsymbol{\theta}}). \tag{5.5}$$

The criterion function Q_T can also be written as $Q_T(\mathbf{X}_T, h_0(\mathbf{X}_T), h(\mathbf{X}_T;\boldsymbol{\theta}))$, as we started under the notion that $\mathbf{y}_T = \{h_0(\mathbf{X_t})\}_{t=1}^T = h_0(\mathbf{X}_T)$ which reveals the direct connection between the criterion function and target function h_0.

There are many ways to generate $\mathcal{H}_\Theta$. In the current framework, we focus on using a kernel k together with a local parameter θ_i that weights the surface to produce any flexible functional form.

$$h_{\boldsymbol{\theta}} := \sum_i^N \theta_i k(x, x_i) = h(x;\boldsymbol{\theta}). \tag{5.6}$$

The functions $h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$, are allowed to follow any kernel that has the universal approximation property, in this paper we adopt a Gaussian

kernel $k(\mathbf{x}_i, \mathbf{x}_j; n_x) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{n_x}\right)$ with $\|\mathbf{x}_i - \mathbf{x}_j\|$ being the Euclidean distance, and n_x being a fixed bandwidth equal to the dimension of $\mathbf{x}_T$. We count the constant as being part of $\mathbf{x}_T$.

The kernel k can be understood as a measure of similarity, which is seen by applying a Cauchy-Schwarz inequality

$$k(x_i, x_j)^2 \leq k(x_i, x_i)k(x_j, x_j) \ \forall \ (x_i, x_j) \in \mathcal{X},$$

revealing that if x_i and x_j are similar, then $k(x_i, x_j)$ will be close to 1, and close to 0 when x_i and x_j are dissimilar. For a given observed collection $(y, x \in \mathbf{x})$, $h_{\boldsymbol{\theta}}$ is thus a function resulting from placing kernels over x_i and scaling the similarity surface using local coefficients θ_i such that the summated surface approximates the true density of the data. This produces flexible functions that can describe local relationships between $\mathbf{y}$ and an individual covariate $\mathbf{x}$ by assigning similar observations a similar scaling factor that maps onto similar output.

Different parameterizations of the local coefficients θ_i may produce equally well, e.g. perfect fits, such that the problem of estimating the vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)'$ is generally ill-posed without adding further structure to the problem. The specific estimation strategy to learn about the trends in the data is therefore of the form

$$\hat{h}_T := \arg\min_{h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta} Q_T(\mathbf{y}_T, \mathbf{X}_T; h_{\boldsymbol{\theta}}) - \pi(h_{\boldsymbol{\theta}}), \tag{5.7}$$

where $\pi(h_{\boldsymbol{\theta}}) > 0 \ \forall \ h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$ is a strictly positive function that monotonically increases by a measure of complexity defined on $h_{\boldsymbol{\theta}}$. The penalty is critical to ensuring identifiability and consistency of the estimator within simple subset spaces of $\mathcal{H}_\Theta$. At the same time, it allows to fit nonlinearities of varying smoothness while working with a fixed kernel bandwidth that produces a relatively smooth similarity surface, as θ_i is able to scale the nonlinearities locally albeit at a cost $\pi(h_{\boldsymbol{\theta}})$. Hence, it

favors less complex solutions to the criterion function by penalizing the high frequency domain in $\mathcal{H}_\Theta$. Specifically, let K be an $N \times N$ symmetric kernel matrix with entries $k(\mathbf{x}_j, \mathbf{x}_i)$ measuring pair-wise similarities. This yields a model that is a linear combination of basis functions, each measuring similarity of one observation to another observation in the data set, and mapping it to a local output.

$$\mathbf{y}_t = h(\mathbf{X}_t; \boldsymbol{\theta}_t) = K(\mathbf{X}_t)\boldsymbol{\theta}_t$$

$$= \begin{bmatrix} k(\mathbf{x}_{1t}, \mathbf{x}_{1t}) & k(\mathbf{x}_{1t}, \mathbf{x}_{2t}) & \cdots & k(\mathbf{x}_{1t}, \mathbf{x}_{Nt}) \\ k(\mathbf{x}_{2t}, \mathbf{x}_{1t}) & \ddots & & \\ \vdots & & & \\ k(\mathbf{x}_{Nt}, \mathbf{x}_{1t}) & & & k(\mathbf{x}_{Nt}, \mathbf{x}_{Nt}) \end{bmatrix} \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{Nt} \end{bmatrix}. \tag{5.8}$$

The need for a regularization technique is obvious, the parameters $(\theta_{1t}, \theta_{2t}, ..., \theta_{Nt})$ can always rescale the similarity surface to match $\mathbf{y}_t$ perfectly. Instead, the penalized estimator takes into account the complexity of the rescaling by introducing a factor $\lambda\|h_{\boldsymbol{\theta}}\|_K^2$ and chooses the best fitting function by minimizing:

$$\arg\min_{h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta} \sum_i^N \sum_t^T (y_{it} - h(\mathbf{x}_{it}; \boldsymbol{\theta}))^2 + \lambda\|h_{\boldsymbol{\theta}}\|_K^2, \tag{5.9}$$

in which $\sum_i^N \sum_t^T (y_{it} - h(\mathbf{x}_{it}; h_{\boldsymbol{\theta}}))^2$ are the standard sum of squared residuals. $\lambda\|h_{\boldsymbol{\theta}}\|_K^2 = \langle h_{\boldsymbol{\theta}}, h_{\boldsymbol{\theta}} \rangle_{\mathcal{H}_\Theta}$ is a penalty that increases monotonically as a function of the complexity of h under $\boldsymbol{\theta}$. We focus on the L^2 norm. Finally, $\lambda \in \mathbb{R}_{>0}$ is te parameter that determines the strength of the penalty. Using this kernel, we can work with an $NT \times NT$ kernel matrix by defining the dependent variable Y as the NT length vector resulting from stacking the time observations, X as the $NT \times n_x$ matrix resulting from stacking the columns similarly and $\boldsymbol{\theta}$ as an NT length parameter

vector.[15] Using the Gaussian kernel, eq. (5.9) becomes

$$\hat{h}_{NT} = \arg \min_{h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta} (Y - K(X)\boldsymbol{\theta})'(Y - K(X)\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}' K(X)\boldsymbol{\theta}. \tag{5.11}$$

In a panel application, the functions $h_{\boldsymbol{\theta}}$ that result from weighted kernels can produce interesting time-varying dynamics across levels of $\mathbf{X}_t$. This is for example appropriate when a time-varying stationary processes is of interest in which the nonlinearities change throughout the data but are not depending on time itself. Alternatively, one can work with time itself as a covariate, in which case processes that are only locally stationary can be modeled. Intuitively, the kernel approach then results in similar coefficients for similar time. In the case of non-stationary data, the kernel can approximate local conditional means in the data that may vary throughout the sample space.

The role of the penalty

The basic idea of penalizing the criterion function has been explored in many statistical applications, and is for example at the heart of the widely adopted LASSO estimator (Tibshirani, 1996; Zou, 2006). The added structure to the criterion function is a frequentist's analogue to the role that the prior plays within the Bayesian framework. We note

[15] Specifically:

$$Y = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Nt} \\ y_{12} \\ y_{22} \\ \vdots \\ y_{NT} \\ y_{1T} \\ y_{2T} \\ \vdots \\ y_{NT} \end{bmatrix}, X = \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \\ \vdots \\ \mathbf{x}_{Nt} \\ \mathbf{x}_{12} \\ \mathbf{x}_{22} \\ \vdots \\ \mathbf{x}_{NT} \\ \mathbf{x}_{1T} \\ \mathbf{x}_{2T} \\ \vdots \\ \mathbf{x}_{NT} \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \vdots \\ \theta_{N1} \\ \theta_{12} \\ \theta_{22} \\ \vdots \\ \theta_{N2} \\ \theta_{1T} \\ \theta_{2T} \\ \vdots \\ \theta_{NT} \end{bmatrix}. \tag{5.10}$$

that the penalty in the current setting is not primarily a way to improve small sample performance, but that it is in fact the central feature of the learning model that determines what functional forms can be fitted. This differs from kernel approaches in which the bandwidth is the key tuning parameter. In the current approach the bandwidth is fixed to produce smooth functions, but nonlinearities are subsequently locally adjusted using the vector of weights $\boldsymbol{\theta}$ to increase flexibility. The penalization approach is able to shrink the hypothesis space and flexibly establish a subspace in which consistency holds. By balancing between fit and complexity of the locally weighted kernel, the size of the subspace can be regulated by the penalty. In the case of penalized GLM's considered in Blasques and Duplinskiy (2018), nonzero penalties take one away from $\boldsymbol{\theta}_0$ in the limit if the penalty effect does not vanish asymptotically.[16] In that sense, a penalized criterion delivers a pseudo-*true* parameter with a divergence from $\boldsymbol{\theta}_0$ that is controlled by the penalty function. Setting an appropriate penalty therefore determines what one can infer from $\boldsymbol{\theta}_0$. In the current context, positive penalties are a necessity to ensure uniqueness. This might lead to the thought that penalized non-parametric estimators that require positive penalization are biased by definition. The estimate of the weights $\hat{\boldsymbol{\theta}}$ obtained through eq. (5.11) is different for every value of λ. The tuning parameter λ thus represents the researcher's predefined level of tolerance for accepting nonlinear functions. High values of λ force the model to linearize it's dependencies, whereas extreme values for λ will set all coefficients to zero and describe the data using only an average expectation. Hence for every penalty, we find a different functional form $\hat{h}_\lambda$ induced by the estimate $\hat{\boldsymbol{\theta}}_\lambda$ through eq. (5.6) given a specified kernel. Since λ itself is not an estimated parameter, it is generally difficult, if not impossible, to tell whether eq. (5.11) yields an estimate of $\hat{h}$ close

[16] Furthermore, $\boldsymbol{\theta}_0$ in the standard context is the *true* parameter. In the non-parametric context, that *true* parametrization arguably does not exist, however one can think of $\boldsymbol{\theta}_0$ as the parametrization that produces h_0 through the kernel, or alternatively, selects h_0 the *true* (non)linear functional form $\mathcal{H}_\Theta$ that produces the *true* density of the data.

to h_0. Without knowing the magnitude of $\|\hat{h}_\lambda - h_0\|$, the method may difficult to use for economic inference.

Schölkopf and Smola (2001), suggest to set the penalty through an out-of-sample prediction minimization problem to remove the dependence of the results on the external influence of the researcher that determines the level of penalization a priori. Hainmueller and Hazlett (2014) suggest one such strategy for Kernel Regularized Least Squares estimates by minimizing out-of-sample prediction errors over a vector $\lambda \in \Lambda$ based on leave-one-out predictions, noting that it performs well in practice. While practical performance and the removal of external influence on the results provide intuition to set penalties in this way, it does not focus on the question whether $\|\hat{h}_\lambda - h_0\|$ is in fact minimized, which is key to ensure that the marginal coefficients converge to the correct values, e.g. $\|\hat{h}'_\lambda - h'_0\| \to 0$. Here we provide additional discussion on the role of the penalty in ensuring identifiable uniqueness and establishing the consistency and normality results. We discuss that the strategy to set penalties by minimizing an out-of-sample criterion naturally pulls the estimator toward the weight vector that induces the *true* function in the limit, such that inference can be applied as usual. This is because for a given penalty λ, the estimated function conditional on that penalty $\hat{h}|\lambda$ provides the optimal density across all functions $h \in \mathcal{H}_\Theta|\lambda$ induced under that penalty, so choosing the estimate from a set of results found using different penalties $\hat{h}|\lambda \in \Lambda$ that provides the optimal out-of-sample density, also minimizes $\|\hat{h}_\lambda - h_0\|_{\lambda\in\Lambda}$ in the limit since h_0 is the function that by definition provides the best out-of-sample density. In other words, estimating eq. (5.11) while setting λ based on out-of-sample prediction error minimization yields an estimated function that minimizes $\|\hat{h} - h_0\|$ in the limit across the entire family of models generated under all weight vectors and all penalties, which is similar to the standard case in which the criterion converges to the parameter that induces the best conditional

density across the entire parameter space (White et al., 1980; White, 1994).

Fixed effects and out-of-sample shrinkage

Linear effects can be included by using difference estimators as detailed in Hainmueller and Hazlett (2014). Nonlinear effects can be modeled by supplying group-specific trend variables and group identifiers through $\mathbf{X}_t$. In this case all coefficients may depend on time, and similarly across similar groups in the data. Nonlinear fixed effects approaches combined with non-parametric parts around the economic variables may result in models with an enormous size while often the amount of observations locally in the time dimension remains relatively small in environmental economic panels. Model size not only relates to the complexity of functions around the economic variables, but also to the number of fixed effects in the model. In-sample selection strategies to decide on the right number of effects are complicated in the regularized non-parametric context. While in standard regressions additional variables always improve fit, this is not the case in the current context. Adding fixed effects results in different complexity of the local weights vector. Therefore, the effect of the complexity penalty in the criterion may increase such that the penalized estimator adjusts the weighting vector to achieve lower complexity. While this reduces the penalty value, it may possibly lower the in-sample R^2. Comparing models with and without fixed effects is therefore a comparison between functional forms with different complexity and nonlinearities. This is a comparison of non-nested models with an unknown, possibly real valued, difference in degrees of freedom.[17]

To decide on the right number of effects, we start by estimating a model that includes all fixed effects. We then remove the least significant

[17]Degrees of freedom is a parametric concept whose translation to the non-parametric setting is complex. One can approximate the degrees of freedom empirically, which may result in numbers that are real-valued.

dummy, and obtain new results. We repeatedly evaluate the out-of-sample prediction performance while shrinking the effects, and select the model with the optimal out-of-sample density across all fixed effect models. Intuitively, this approach starts with a model similar to a linear fixed effects model, as the penalty heavily discounts the thresholds introduced by the effects resulting in flattened marginal effects, and gradually allows fixed heterogeneity to be explained by nonlinearities across covariates instead. As a result, our final estimates are guaranteed to be preferred over the standard linear fixed effects model, as judged by the out-of-sample criterion.

5.7.3 The role of out-of-sample performance in the interpretation

Non-parametric approaches are capable of producing parametrization mappings that approximate nonlinearities arbitrarily well, but do not necessarily also produce uniquely identifiable solutions to the criterion function if the hypothesis space produces universal approximations that fit the data arbitrarily well for any sample size. Estimation is therefore problematic without additional structure to the estimator, which in our case comes in the form of a penalty to the criterion, but in other settings may relate to bandwidths or other tuning parameters. It is a challenge in its own right to understand how this complexity-penalized estimator is positioned relative to the classical least squares approximation context as considered by White et al. (1980).[18] In the standard context, the best approximation is produced by a unique point in the entire parameter space, while in the penalty context a best approximation exists for every given penalty. Hence, the divergence between the *true* functional form and the *pseudo*-true approximation is not driven by boundaries to the

[18] White et al. (1980) discuss convergence toward the unique least squares approximator that may differ from the *true* parameter in the presence of misspecification bias.

parameter space as in the parametric case, but rather it is driven by the penalty. Ultimately penalization confines the hypothesis space to simple spaces and the use of excessive penalization must reflect a prior belief that the *true* functional form would not result in a large penalty. That prior belief carries over to the limit result if the effect of the penalty does not vanish. This produces a bias, or may even render the result completely arbitrary if the penalty is set without caution.

In the current setting, our penalty arises as a function of an out-of-sample criterion. As a result, the space of functions that are viewed as acceptable solutions to the criterion is generated by the data itself, and the penalized non-parametric method is able to obtain approximations of increasing complexity as the data size tends toward infinity. In the finite sample case, this estimator is appropriate given that the relationship between environmental degradation and indicators of economic development is not dominated by high-frequency components that would result in strong complexity.

Identifiability in nonlinear models

Closely related to regulating the size of non-parametric models is the ill-posedness of unregulated non-parametric models. Before discussing the relationship between penalization and identifiability of the criterion of non-parametric estimators, we provide a simplified discussion on identifiable uniqueness and its relation to inference in the context of finite dimensional nonlinear models.[19]

Hypothesis testing in a framework of finite parameter nonlinear models is often plagued by the problem that verification of the assumptions required

[19]Identifiable uniqueness is a difficult concept, more elaborate general discussion can be found here (Pötscher and Prucha, 1991); formal definitions and discussion at a deeper level regarding strongly unique best approximation in Banach spaces can be found here (Smarzewski, 1986); and discussion on regulated M-estimation can be found here (Kent and Tyler, 2001); and an overview of concepts is written in (Blasques, 2010).

for identifiability, relies itself on the outcome of a hypothesis that may be difficult to test. This is problematic as identifiable uniqueness plays a key role in establishing consistency and normality of test statistics. This is illustrated by a model of the form:

$$y = \delta \exp\left(\frac{-(x-c)^2}{\gamma}\right) + \varepsilon, \tag{5.12}$$

in which the postulated relationship between y and x is assumed to follow a hyperbolic curve across levels of x. In this model the linearity hypothesis $H_0^1 : \gamma = 0$ relating to the non-existence of the curved functional form depends on a second hypothesis $H_0^2 : \delta \neq 0$ being true or false. This follows from the fact that for $\delta = 0$, γ can take any value without changing the predicted density implied by the model. In this case any form of completeness required for identifiable uniqueness of the estimator, holds at most for a subset of the parameter space in which the model would in fact produce an inverted U-shaped form. Distributions corresponding to different values of γ are only sufficiently distinct when δ is sufficiently bounded away from zero. Without establishing existence and uniqueness of a consistent estimator, it is impossible to establish normality, hence the distribution of test statistics remains unknown.[20]

More intuition is found in the following two definitions adapted from Definition 1 and Definition 2 in (Rothenberg, 1971).

DEFINITION. 1. *Two points $\alpha_1 \in \mathcal{A}$ and $\alpha_2 \in \mathcal{A}$ are said to be observationally equivalent with respect to a function h evaluated over x if $h(x;\alpha_1) \equiv h(x;\alpha_2)\ \forall\ x \in \mathbb{R}$.*

[20] Auxiliary test statistics may still be derived, but it is sometimes difficult to ensure that Taylor expansions do not capture nonlinearities of a type not predicted by the economic theory. See for example (Dijk et al., 1999) for a discussion in the threshold framework. Researchers may also choose to rely on information criteria to compare various descriptions of the data and decide between economic theories (Granger et al., 1995). In the limit, Penalized Likelihood Criteria select the model that minimizes Kull-Back Leibler divergence with probability 1 (Sin and White, 1996), but convergence rates depend on the penalty chosen. The acceptance of an economic theory thus relies on information outside the model. In a sense, a researcher has flexibility to corroborate specific theories by designing the information criteria to support them.

DEFINITION. 2. *A point $\alpha^1 \in \mathcal{A}$ is said to be identifiable by a function h evaluated over x if there is no other point $\alpha \in \mathcal{A}$ that is observationally equivalent.*

Let $\boldsymbol{\theta} := (\delta, c, \gamma)'$ denote a vector of parameters, with $\boldsymbol{\theta} \in \Theta$, and $\boldsymbol{\theta}_0 := (\delta_0, c_0, \gamma_0)'$ be the *true* vector of parameters. For consistency toward the *true* parameter, one would not only require $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \to 0$ to be the solution *a.s.* to the criterion $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q(y, x; \boldsymbol{\theta})$ as $N \to \infty$, but it needs to be the *identifiable unique* solution. Following the definitions above, then by the definition $\boldsymbol{\theta}_0$ as the minimizer of $Q(y, x; \boldsymbol{\theta})$, there needs to be assurance of some form that

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q(y, x; \boldsymbol{\theta}_0) < \arg\min_{\boldsymbol{\theta}\in\Theta} Q(y, x; \boldsymbol{\theta}) \ \forall \ \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0, \tag{5.13}$$

excluding

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q(y, x; \boldsymbol{\theta}_0) \leq \arg\min_{\boldsymbol{\theta}\in\Theta} Q(y, x; \boldsymbol{\theta}) \ \forall \ \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0. \tag{5.14}$$

as the alternative. The standard assumption is that Θ is compact. Together with *almost sure continuity* in $\boldsymbol{\theta} \in \Theta$, Weierstrass' theorem implies that $\boldsymbol{\theta}_0$ exists as a non-empty set *a.s.* Equation (5.13) can result directly from the parametrized model $\hat{y} = h(x)$ if

$$h(x; \boldsymbol{\theta}_0) \neq h(x; \boldsymbol{\theta}) \ \forall \ \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0, \tag{5.15}$$

such that there is no point in Θ other than $\boldsymbol{\theta}_0$ that is observationally equivalent to $\boldsymbol{\theta}_0$. Specifically the observational equivalence definition may fail to hold if Θ is high dimensional. If eq. (5.13) is not implied by the nature of h, it can also be provided by additional structure to the criterion $Q(\cdot; \boldsymbol{\theta})$ conditional on regions in Θ, or by limiting the search to remain within a subset $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta_s\subset\Theta} Q(y, x; \boldsymbol{\theta})$, where Θ_s is a compact subset of the parameter space that may possibly grow in complexity along with the sample size.

DEFINITION. 1 and DEFINITION. 2 are intuitive, but provide no testable condition to decide upon the identifiability of an estimator. One insightful definition is the following adapted from (Bates and White, 1985) that ensures that the solution to the criterion is well separated.

DEFINITION. 3. *Suppose* $\boldsymbol{\theta}_0$ *minimizes a real-valued criterion* $Q_\infty(\cdot;\boldsymbol{\theta})$ *on a compact metric space* Θ*, within a circular neighborhood* $\aleph_0(r) \subset \Theta$ *with radius* $r > 0$ *that has a compact complement* $\aleph_0(r)^c : \Theta \setminus \aleph_0(r)$*, then* $\boldsymbol{\theta}_0$ *is uniquely identified on* Θ if and only if *for every* $r > 0$*,*

$$\inf_{\boldsymbol{\theta} \in \aleph_0(r)^c} [Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0.$$

Identifiability in non-parametric models

Non-parametric models aim to learn from the data without assuming that h is up to finitely many parameters, and work under the axiom that the parameter space Θ may in fact be infinitely dimensional. By allowing for that, we minimize the risk that our parametrization assumptions preclude $\boldsymbol{\theta}_0 \in \Theta$, solving for misspecification bias that results from parametric assumptions. However, without imposing further structure to the criterion it is generally not possible to establish consistency of our estimate $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ uniformly over Θ as the compactness assumption on Θ does not hold in infinite dimensions.[21] This poses a problem in verifying DEFINITION. 3, and establishing consistency results such as those of (Domowitz and White, 1982).

One solution is to focus the arguments on establishing a compact subset of the parameter space such that over the complement of the compact subset the criterion function is eventually "large", see (Pötscher and Prucha, 1997). This follows by first constructing a subset $\Theta_s \subset \Theta$ such that $\boldsymbol{\theta}_0 \in \Theta_s$, and such that all the elements outside Θ_s are valued distinguishably different by the criterion than the elements within Θ_s,

[21] By definition a set $\mathcal{A} \in \mathbb{R}^d$ is compact *if and only if* it is closed and bounded.

disregard of the structure outside of Θ_s. The subset is then closed, as its complement is open, and it is bounded as it is contained in a ball of finite radius,which implies that Θ_s is then compact. As a consequence, it is sufficient to show that consistency holds within Θ_s, since any M-estimator must eventually fall within this compact subset. We can summarize identifiable uniqueness of $\boldsymbol{\theta}_0$ in an open space as follows.

DEFINITION. 4 (Identifiability in an open space). *Suppose* $\boldsymbol{\theta}_0$ *minimizes a real-valued criterion* $Q_\infty(\cdot;\boldsymbol{\theta})$ *on an open metric space* Θ. *Suppose furthermore that* $\boldsymbol{\theta}_0$ *minimizes* $Q_\infty(\cdot;\boldsymbol{\theta})$ *within a circular neighborhood* $\Theta_s(d) \subset \Theta$ *that has finite positive radius* $d > 0$ *and that uniformly over* Θ *there exists some positive* ϵ *for which* $[Q_\infty(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - Q_\infty(\boldsymbol{\theta} \in \Theta_s)] > \epsilon$. *If furthermore, there is also a circular neighborhood* $\aleph_0(r) \subset \Theta_s$ *with radius* $r < d$ *that has a compact complement* $\aleph_0(r)^c : \Theta_s \setminus \aleph_0(r)$, *then* $\boldsymbol{\theta}_0$ *is uniquely identified on* Θ if and only if *for every* r, $0 < r < d$,

$$\inf_{\boldsymbol{\theta} \in \aleph_0(r)^c \subset \Theta_s(d) \subset \Theta} [Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0.$$

In a sense, we thus want to exert some control over the structure of $Q_\infty(\cdot;\boldsymbol{\theta})$ on Θ such that $\boldsymbol{\theta}_0$ is uniquely identifiable by the criterion within a neighborhood Θ_s that is distinctly different from elements outside of it, disregard whether the criterion can distinguish differences between the elements outside Θ_s. One such an approach can be found in the well-known kernel estimator. The solution offered by the kernel method depends on selecting an appropriate bandwidth that controls for the size of local neighborhoods in the sample space throughout which nonlinearities smoothly differ. For too small bandwidths, the kernel method creates a subspace $\Theta_k \supset \Theta_s$ that allows overly-flexible fits to the data. This can create an ill-posed problem, in which multiple solutions to the criterion within Θ_k may still deliver equally good fits as judged by the criterion evaluated over Θ_k. It is obvious that DEFINITION. 4 is not applicable in such a context. For too small bandwidths, the kernel method establishes

$\Theta_k \subset \Theta_s$ that is small, and while DEFINITION. 4 may work for Θ_k, we are not sure that in fact $\boldsymbol{\theta}_0 \in \Theta_k$ due to the parametrization assumptions used to construct Θ_k. The role of the bandwidth is therefore extremely important, identifiable uniqueness of the estimator requires the bandwidth to be sufficiently large, while reducing miss-specification bias requires the bandwidth to be sufficiently small. In an ideal framework, both factors are balanced out and Θ_k grows as $N \to \infty$ at an appropriate rate.

The role of the penalty in the estimator

The fitted nonlinearities are allowed to be of any form, but $\lambda > 0$ implies the penalty is never removed completely. Positive penalization is key to ensuring that there exists a finite radius neighborhood $\Theta_s(d) \subset \Theta$ in which any M-estimator must eventually fall uniformly over Θ as $[Q_\infty(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - Q_\infty(\boldsymbol{\theta} \in \Theta_s)] > \epsilon(\boldsymbol{\theta}; \lambda) > 0$, where $\epsilon(\boldsymbol{\theta}; \lambda) > 0$ is ensured for any $\boldsymbol{\theta}$ by $\lambda > 0$. Penalties that vanish completely at a pre-specified rate are interesting when the researcher wishes to impose penalties only when the estimator is confronted with small sample sizes. This requires however that the criterion is uniquely identified at $\lambda = 0$ eventually. Vanishing penalties may improve inference when using estimators that have poor small sample behavior by ensuring that the estimator is relatively inert to weakly nonlinear signals and less likely to overfit the data in local regions of the sample space. Penalties that take values in $\mathbb{R}_{>0}$, can improve small sample behavior, but maintain a bias towards linear solutions that persists in the limit.

Note that eq. (5.11) reveals that convergence of our estimator $\|\hat{h}_{NT} - \hat{h}_\infty\| \to 0$ to a specific target function $\hat{h}_\infty \in \mathcal{H}_\Theta$, where $\hat{h}_\infty$ is possibly the *true* function or the best approximator as judged by the penalized limit criterion, is the same as $\|\hat{\boldsymbol{\theta}}_{NT} - \hat{\boldsymbol{\theta}}_\infty\| \to 0$, which is the more common notation. Hence, we shall use the latter, but really we are interested in ensuring that $\hat{h}_\infty$ is a uniquely identifiable point in $\mathcal{H}_\Theta$ as close to

h_0 as possible. Consistency and normality theorems for eq. (5.9) are provided in Hainmueller and Hazlett (2014). The results ensure a limit convergence toward the best approximation of the conditional expectation function given penalization, hence the limit solution is conditional on the researcher's choice of λ. The theory provided is therefore to be understood in terms of $\hat{\boldsymbol{\theta}}_{NT}$ converging to a *pseudo-true* parameter as $NT \to \infty$, that by construction minimizes the penalized criterion even if the penalty does not vanish. To understand the relationship between the *pseudo-true* parameter and the *true* parameter conditional on the penalty, it is helpful to consider precisely how the penalty influences the criterion and delivers the identifiable uniqueness property.

Let $\hat{\boldsymbol{\theta}}_\pi$ be the point $\hat{\boldsymbol{\theta}}_\pi := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$, that minimizes the penalized criterion, and $\boldsymbol{\theta}_0$ be the point $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{Q}_\infty(\boldsymbol{\theta})$ that minimizes an unpenalized out-sample criterion. $\boldsymbol{\theta}_\pi$ is the best approximator similar to the misspecification case studied in (White et al., 1980), whereas $\boldsymbol{\theta}_0$ is the *true* parameter, that is the weights vector that induces h_0 through the kernel, which is the *true* function that provides the best out-of-sample density by its definition. The function $\hat{h}_\pi := h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi)$ is the best approximator of $h_0 := h(\mathbf{x}_t; \boldsymbol{\theta}_0)$ as judged by the penalized criterion $Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ for a given level of penalization π. The penalty does not imply that $h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi) \neq h(\mathbf{x}_t; \boldsymbol{\theta})\ \forall\ \boldsymbol{\theta} \in \Theta \setminus \hat{\boldsymbol{\theta}}_\pi$ and any $\mathbf{x}_T \in \mathcal{X}$ and all $NT \in \mathbb{N} \times \mathbb{Z}$. However, it ensures that $\hat{\boldsymbol{\theta}}_\pi$ is identifiable unique as the minimizer of the limit criterion even in the case of two observationally equivalent parametrizations $h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi) \equiv h(\mathbf{x}_t; \boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^* \in \Theta \setminus \hat{\boldsymbol{\theta}}_\pi$ and any $\mathbf{x}_T \in \mathcal{X}$ and all $NT \in \mathbb{N} \times \mathbb{Z}$.

PROPOSITION. 4 (Identifiable uniqueness). *The function*

$$\hat{h}_\pi := \arg\min_{h_{\boldsymbol{\theta}}\in\mathcal{H}_\Theta} Q_\infty(h_{\boldsymbol{\theta}}) + \pi(h_{\boldsymbol{\theta}})$$

produced by $h_\mathcal{X}$ *at point* $\hat{\boldsymbol{\theta}}_\pi := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ *is uniquely identified within* $\mathcal{H}_{\Theta_s}$ *a simple subset in the infinite dimensional Hilbert*

space $\mathcal{H}_\Theta$, if π is a strictly positive penalty function continuous on Θ.

Central to the result is that $\pi(\hat{\boldsymbol{\theta}}_\pi) < \pi(\boldsymbol{\theta}^*)$, providing that for two observationally equivalent functions, the identified result is the parameter vector that induces a less complex functional form.

So far we have treated π to be fixed at a pre-specified level. However, for any given level of penalization, the solution to the penalized criterion is different. We can make that more explicit by writing it as the limit estimate conditional on a penalty value $(\hat{\boldsymbol{\theta}}_\infty|\pi)$, and analyzing the role of π in the divergence $\|(\hat{\boldsymbol{\theta}}_\infty|\pi) - \boldsymbol{\theta}_0\|$. This displays the heavy importance on determining an appropriate penalty π as it is crucial to the outcome, see Blasques and Duplinskiy (2018) for some thoughts on how to choose appropriate penalty weights in a general context. Asymptotically, if the impact of π vanishes, for example by using penalties of an $o((NT)^{-\frac{1}{2}})$, consistency toward $\boldsymbol{\theta}_0$ can still be met in the limit, again see Blasques and Duplinskiy (2018) for detail. However, in small samples, similar to the Bayesian case, a researcher can exert influence on the outcome by setting the value of π. In the current framework, $\pi > 0$ prevents a generality claim as it would follow in the parametric case, however we can still focus the argument on finding an optimal penalty that minimizes $\|(\hat{\boldsymbol{\theta}}_\infty|\pi) - \boldsymbol{\theta}_0\|$, or equivalently the function divergence $\|(\hat{h}_\infty|\pi) - h_0\|$.

PROPOSITION. 5 (Best approximation across penalties and weights). *The divergence between the best approximation as judged by the penalized limit criterion given a level of penalization and the* true *function is smaller than the divergence as evaluated at all other limit estimates resulting under other penalty weights*

$$\|(\hat{h}_\infty|\pi_0) - h_0\| < \|(\hat{h}_\infty|\pi) - h_0\| \ \forall \ \pi \ \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

and results under the penalty π_0 that minimizes an out-of-sample criterion

$$\pi_0 : \arg\min_{\pi \in \Pi} Q_\infty(\hat{h}_\infty|\pi), \Pi \subseteq \mathbb{R}_{>\delta}$$

for some small positive δ. Hence, $(\hat{h}_\infty|\pi_0)$ is the best approximation of h_0 over $\mathcal{H}_{\Theta_s \times \Pi} := \{\mathcal{H}_{\Theta_s}|\pi_1 \times \mathcal{H}_{\Theta_s}|\pi_2 \times ... \times \mathcal{H}_{\Theta_s}|\pi\} \; \forall \; \pi \in \Pi$, that is across all penalties and weights.

PROPOSITION. 5 implies that if the penalty is chosen by minimizing a criterion out-of-sample, a weights vector can be estimated that produces the function closest to the target function across all penalties and weights. Effectively, a researcher is able to identify an approximation that is arbitrarily close to the *true* curve, by solving the estimator on very large data iteratively for a sufficiently wide range of penalties and selecting the result that performs optimal as an out-of-sample predictor. This is an intuitive solution as $\boldsymbol{\theta}_0$ carries a natural interpretation as the optimal out-of-sample predictor. The key result, and with that the role of the penalty, is summarized below in fig. 5.10.

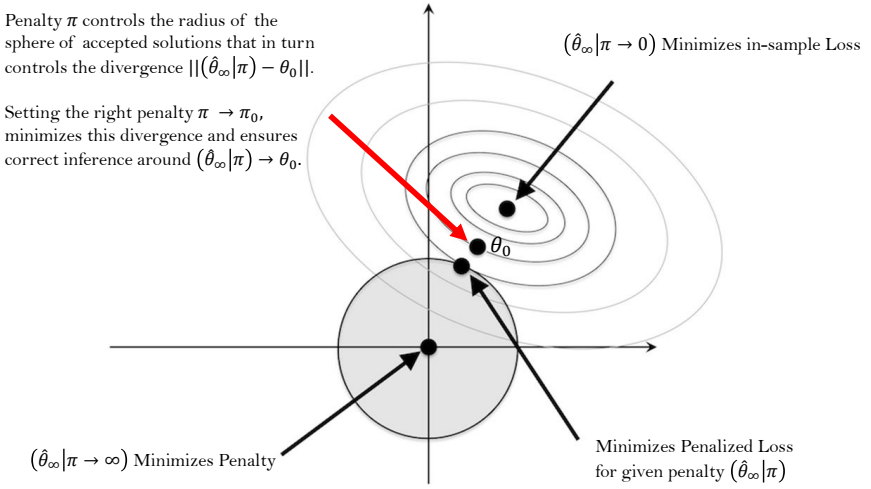

Figure 5.10: For functions h induced under parameter vectors θ, and penalties λ controlled by general penalty functions π, the figure displays graphically the role of the penalty function in managing the closeness of the empirical result to the result that delivers the correct function h_0 associated with the correct marginal coefficients h_0' of interest. The gray shaded area contains the space of accepted solutions that currently includes the result induced under an infinite penalty, but not the correct result, nor the result that would be obtained when the model fully minimizes in-sample loss. Hence the graph corresponds to the mis-specified case in which the data is under-fitted.

5.7.4 Conclusion

This note detailed an adaption of the Kernel Regularized Least Squares model to the panel context, paying particular focus to automated selection of fixed effects. The model was applied in a our main paper to study nonlinear trends between environmental indicators and economic development. The key feature of the model that makes it attractive in this type of applied studies is that it provides a straightforward approach to nonlinearity and interpretability within one integrated framework. The difficulty with the approach is that estimation relies on externally set tuning parameters that are not part of the estimated vector of parameters that define the functional relationships in the data. This makes the interpretation of local marginal coefficients highly dependent on correctly tuning the model. The discussion provided high level arguments for optimizing the estimation criterion on validation samples in order to ensure the coefficients admit to standard interpretation.

The discussion provided some examples that highlighted that penalization in the non-parametric context differs from penalization in the GLM case, such as in the popular LASSO. While penalized GLM's require the penalty to vanish asymptotically for generality claims, positive penalization in the limit may be a necessity to ensure identifiable uniqueness for non-parametric models. Regularization or penalization, while primarily known for dealing with over-fitting, was in fact a way to flexibly establish simple subspaces in which consistency theorems hold. As a result, the consistency and normality limits are uniquely defined for every level of penalization, which makes it less straightforward to interpret the estimator, and its derivatives, with usual confidence. However, penalties may still be found that yield estimates that are conform standard interpretation. Specifically, penalties that result from minimizing an out-of-sample criterion pull the consistency limit toward the result that induces the optimal conditional distribution implied by the weighted kernel across all penalties and

weights, as judged by the out-of-sample criterion. That result is similar to the standard convergence toward the parameter that delivers the best modeled density in terms of divergence with respect to the true density of the data Kullback and Leibler (1951); White (1994). Under that result, the estimator converges to the result that delivers the best approximation to the true distribution of the data.

It is important to stress that the argument is based on out-of-sample optimization of the same criterion function that was used to fit the model in the estimation sample. Particularly in the case of classification problems this may deviate from common practices. For example, classification problems are often tuned by maximizing accuracy measures that involve fractions of correctly predicted or mis-predicted classes. These widely used metrics do not satisfy the smoothness properties imposed on the in-sample criterion function to obtain consistency to a limit result at a given value of the penalty parameter. A straightforward example of one such violated assumption is the assumed continuity of the estimation criterion in all it's arguments, which is needed as part of a standard Consistency proof to ensure limit preserving properties. This continuity breaks because for any level of accuracy (simply the percentage of correctly classified observations), there can exist an infinite number of parameterizations that are judged to be exactly identical by the accuracy criterion. Two simple examples are one model that is correct by predicting .49 and .51 probabilities versus one that predicts 0 versus 1. Moreover, a minute change in a parameter value may change the model's accuracy from 100% to 0%, for example by swapping the margins around probabilities close to .5., so model's with near identical parameters can also be judged as wildly different by an accuracy-based criterion.

Proofs

Proof to PROPOSITION. 4

Proof. $\hat{\boldsymbol{\theta}}_\pi := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$, is by definition the minimizer of $Q_\infty(\cdot;\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ that is by construction of the least squares function and the penalty function $\pi : \Theta \to \mathbb{R}_{>0}$ a real-valued criterion on an open metric space Θ. Furthermore there exists some positive constant ϵ for which

$$[Q_\infty(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) + \pi(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - Q_\infty(\boldsymbol{\theta} \in \Theta_s) + \pi(\boldsymbol{\theta} \in \Theta_s)] > \epsilon,$$

because for $Q_\infty(\boldsymbol{\theta} \in \Theta \setminus \Theta_s) \equiv Q_\infty(\boldsymbol{\theta} \in \Theta_s)$,

$$[\pi(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - \pi(\boldsymbol{\theta} \in \Theta_s)] > \epsilon,$$

by the monotonicity of π on Θ. This implies that $\hat{\boldsymbol{\theta}}_\pi$ minimizes $Q_\infty(\cdot;\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ within a neighborhood $\Theta_s \subset \Theta$. Furthermore $\Theta_s(d)$ has finite radius $d < \infty$ because

$$[\pi(\boldsymbol{\theta} \in \Theta_s(d)) - \pi(\boldsymbol{\theta} \in \Theta \setminus \Theta_s(d))] \leq \epsilon$$

implies $d < \infty$, by finiteness of ϵ in turn implied by continuity of the penalty. Finally, $\Theta_s(d) \subset \Theta$ is compact as it closed because its radius is finite, and its complement $\Theta \setminus \Theta_s$ is open.

We have now established that uniformly over Θ, the estimator must fall eventually inside Θ_s. The rest of the argument follows from standard identifiability arguments in compact parameter spaces as in (Bates and White, 1985; Domowitz and White, 1982) focused on Θ_s. That is, define a circular neighborhood $\aleph_k(r) \subset \Theta_s$ with nonnegative radius $r < d$ that has a compact complement $\aleph_k(r)^c : \Theta_s \setminus \aleph_k(r)$. $\boldsymbol{\theta}_k$ is uniquely identified

on Θ as by $0 < r < d < \infty$, for every (r, d),

$$\inf_{\boldsymbol{\theta} \in \aleph_k(r)^c \subset \Theta_s(d) \subset \Theta} \left[Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}) - Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi) \right] > 0.$$

In our case, this is implied by continuity of the criterion and additionally by the fact that for any observationally equivalent point $\pi(\boldsymbol{\theta}^*)$ such that $Q_\infty(\boldsymbol{\theta}^*) \equiv Q_\infty(\hat{\boldsymbol{\theta}}_\pi)$, by definition of $\hat{\boldsymbol{\theta}}_\pi$ as the minimizer of $\min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ the continuity of π implies

$$\pi(\hat{\boldsymbol{\theta}}_\pi) < \pi(\boldsymbol{\theta}^*).$$

□

Proof to PROPOSITION. 5

Proof. Let $\hat{\boldsymbol{\theta}}_\pi = \hat{\boldsymbol{\theta}}_\infty | \pi := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ be the minimizer of the penalized criterion for a certain level of penalization and $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{Q}_\infty(\boldsymbol{\theta})$ the minimizer of an unpenalized out-of-sample criterion. When plugging the *true* parameter in the penalized criterion, then taking $\|\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_0\| \to 0$ implies that similarly $\left| \left[Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi) \right] - \left[Q_\infty(\boldsymbol{\theta}_0) + \pi(\boldsymbol{\theta}_0) \right] \right| \to 0$. This minimization is solved if the in-sample criterion evaluates $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)| \to 0$ equivalently, as then immediately also $|\pi(\hat{\boldsymbol{\theta}}_\pi) - \pi(\boldsymbol{\theta}_0)| \to 0$. Hence, either $|\pi(\hat{\boldsymbol{\theta}}_\pi) - \pi(\boldsymbol{\theta}_0)| \to 0$ or $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)| \to 0$, is sufficient for $\|\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_0\| \to 0$.

Any such result following from taking both penalties $\pi(\hat{\boldsymbol{\theta}}_\pi)$ and $\pi(\boldsymbol{\theta}_0)$ to zero simultaneously as $N \to \infty$ is prohibited by the fact that $\pi : \Theta \to \mathbb{R}_{>0}$. However $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi)| - |Q_\infty(\boldsymbol{\theta}_0) + \pi(\boldsymbol{\theta}_0)|$ attains a minimum when setting the penalty to minimize the criterion defined on out-of-sample errors. Specifically since $\boldsymbol{\theta}_0$ is by construction the minimum of the out-of-sample criterion in the limit, setting π_0 to minimize

$\arg\min_{\pi\in\Pi\ \forall\ \subseteq\mathbb{R}_{\geq 0}} \mathcal{Q}_\infty(\hat{\boldsymbol{\theta}}_\pi)$ gives

$$\pi_0 : \arg\min_{\pi\in\Pi} |\mathcal{Q}_\infty(\hat{\boldsymbol{\theta}}_\pi) - \mathcal{Q}_\infty(\boldsymbol{\theta}_0)|.$$

and if $|\mathcal{Q}_\infty(\hat{\boldsymbol{\theta}}_\pi) - \mathcal{Q}_\infty(\boldsymbol{\theta}_0)| \to 0$, it must follow that

$$|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)| \to 0,$$

and,

$$|\pi(\hat{\boldsymbol{\theta}}_\pi) - \pi(\boldsymbol{\theta}_0)| \to 0.$$

If $\Pi \subseteq \mathbb{R}_{\geq 0}$ is constructed such that $\pi_0 \in \Pi$ for which the arg min's above reach 0, then $\|(\hat{h}_\infty|\pi_0) - h_0\| = 0$ would follow and we reach the *true* target function. Now $\Pi \subseteq \mathbb{R}_{\geq 0}$ can contain penalties infinitely close to 0, in practice one must work with finite sets for a grid search across Π and construct instead a set $\Pi \subseteq \mathbb{R}_{\geq\delta}$ being the set of all possible parameters bounded away from zero by some arbitrarily small positive constant δ. If $\pi_0 \notin \Pi$ for which $\|(\hat{h}_\infty|\pi_0) - h_0\| = 0$, then still

$$|\mathcal{Q}_\infty(\hat{\boldsymbol{\theta}}_{\pi_0}) - \mathcal{Q}_\infty(\boldsymbol{\theta}_0)| < |\mathcal{Q}_\infty(\hat{\boldsymbol{\theta}}_\pi) - \mathcal{Q}_\infty(\boldsymbol{\theta}_0)|\ \forall\ \pi\ \ \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

thus also

$$|Q_\infty(\hat{\boldsymbol{\theta}}_{\pi_0}) - Q_\infty(\boldsymbol{\theta}_0)| < |Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)|\ \forall\ \pi\ \ \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta}$$

and therefore

$$\|(\hat{\boldsymbol{\theta}}_\infty|\pi_0) - \boldsymbol{\theta}_0\| < \|(\hat{\boldsymbol{\theta}}_\infty|\pi) - \boldsymbol{\theta}_0\|\ \forall\ \pi\ \ \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

which induces through the definition of h as the weighted kernel also

$$\|(\hat{h}_\infty|\pi_0) - h_0\| < \|(\hat{h}_\infty|\pi) - h_0\|\ \forall\ \pi\ \ \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

implying that $(\hat{h}_\infty|\pi_0)$ turns out to be the best approximation of h_0 for all penalties in Π that each result itself as a best approximator within

the subset $\mathcal{H}_{\Theta_s}|\pi$ within the penalized criterion necessarily falls given the level of penalization. In this case π_0 simply plays the role of a pseudo-true penalty that delivers a pseudo-true results, which can be detected when the penalty is at the boundary of the grid Π or is expected when the resolution of the grid is such that Π is not approximately continuous. $\square$