Nonparametric Regression with Common Shocks*

EDUARDO A. SOUZA-RODRIGUES
University of Toronto, Department of Economics

July 2016

Abstract

This paper considers a nonparametric regression model for cross-sectional data in the presence of common shocks. Common shocks are allowed to be very general in nature; they do not need to be finite dimensional with a known (small) number of factors. I investigate the properties of the Nadaraya-Watson kernel estimator and determine how general the common shocks can be while still obtaining meaningful kernel estimates. Restrictions on the common shocks are necessary because kernel estimators typically manipulate conditional densities, and conditional densities do not necessarily exist in the present case. By appealing to disintegration theory, I provide sufficient conditions for the existence of such conditional densities and show that the estimator converges in probability to the Kolmogorov conditional expectation given the sigma-field generated by the common shocks. I also establish the rate of convergence and the asymptotic distribution of the kernel estimator.

JEL Classifications: C13, C14, C21.

KEYWORDS: Nonparametric regression, common shocks, cross-sectional dependence, disintegration theory

*I am grateful to Donald Andrews, Xiaohong Chen, Philip Haile, Steven Berry, Tai Otsu, Yuichi Kitamura, Ed Vytlacil, Peter Phillips, Marfisa Queiroz, and participants of Econometrics Lunch at Yale. Financial support from the Charles V. Hickox Fellowship, Yale University Fellowship and Kernan Brothers Environmental Fellowship at Harvard University are gratefully acknowledged. All errors are mine.

1 Introduction

Cross-sectional dependence has attracted considerable attention among economists recently.[1] It is well-known that ignoring cross-sectional dependence may lead to inconsistent estimators and misleading inference. A popular and successful way to capture cross-sectional dependence is through common factors.[2] Common factor models assume a finite number of unobserved factors that may be the result of economy-wide shocks with impacts on population units that may depend on the characteristics of the unit. Possible common factors include macroeconomic, technological, legal/institutional, political, environmental, health, and sociological shocks, among others. The applied literature has considered, for example, technological shocks (such as new procedures, drugs, and surgical techniques) affecting the relationship between countries' healthcare attainments and their per capita health expenditures and educational levels (Evans, Tandon, Murray, and Lauer (2000)); cross-country cross-industry analysis of returns to R&D, which are affected both by global shocks such as the recent financial crisis, and by local shocks such as spillovers between a limited group of industries or countries (Eberhardt, Helmers, and Strauss (2013)); and the analysis of transnational terrorism, where common factors may arise from common terrorist training camps, common grievances, and demonstration effects (Gaibulloev, Sandler, and Sul (2013)).

Typically, common factor models allow for a small and known number of unobserved factors. Although such an approach is convincing in empirical macro models, in microeconometric models it is often more reasonable to think of a potentially large, possibly unknown (and maybe infinite) number of factors that can influence individuals' behavior. For instance, in studies of individual earnings there are many individual-level observables and unobservables that affect income; as well as several common factors such as region, family, male/female ratio, race composition, education, age composition, and so on (Phillips and Sul (2007)). The number of common factors may increase as we collect more cross-sectional observations, or there may be an infinite number of unobserved factors (see, e.g., Altonji, Conley, Elder, and Taber (2010)).

The purpose of this paper is to study a nonparametric regression model for cross-sectional data in the presence of common shocks that are very general in nature. The common shocks can be of infinite dimension with flexible impact on different population units. For example, common

[1] See, for example, Arbia (2006); Baltagi, et al. (2007); the proceedings of the 2008 Cowles Summer Conference (Andrews, 2008); the special issue of the Journal of Econometrics ("Analysis of Spatially Dependent Data," 2007, 140(1), edited by Baltagi, Kelejian and Prucha); and the special issue of Econometrics ("Spatial Econometrics," 2015, edited by Arbia and Lee). For recent surveys, see Sarafidis and Wansbeek (2012), and Chudik and Pesaran (2014).

[2] See, e.g., Phillips and Sul (2003, 2007), Andrews (2005), Bai and Ng (2006), Pesaran (2006), Bai (2009), Moon and Weidner (2009), Zaffarori (2010), Pesaran and Tosetti (2011), Su and Jin (2012), Huang (2013), Kuersteiner and Prucha (2013), Chudik and Pesaran (2015), Cai, Fang, and Xu (2016), and Forchini and Peng (2016).

shocks could take the form of a nonlinear random function of observable or unobservable individual characteristics with the effect on the i-th observation varying continuously across i depending on the value of the characteristic. We focus on nonparametric models because there may be little guidance (or justification) in practice for selecting a particular functional form for the regression function.

There has been important recent work on nonparametric models with many finite common factors (Su and Jin (2012), Huang (2013), and Cai, Fang, and Xu (2016)). They consider common shocks that enter the regression function additively and with disturbances that are modelled as linear functions of mutually independent unobserved common factors and individual-specific factor loadings. We, in constrast, allow the regression function to be nonseparable for common shocks and we do not require the mutual independence assumption. In other words, we allow for an unknown large, potentially infinite, number of factors that can influence individuals' outcomes and that may interact with observable and unobservable individual characteristics in extremely rich and flexible ways. To the best of our knowledge, this is the first paper that allows for such a flexible framework.

We consider such a flexible setting because we are interested in investigating how general the regression function and the common shocks can be while still allowing for meaningful nonparametric estimates. We focus on the Nadaraya-Watson kernel estimator and study the effects of general common shocks on its asymptotic properties. Asymptotic results for kernel estimators are typically obtained by manipulating conditional densities of random variables. However, if the common shocks are too general, conditional densities do not necessarily exist. Doob (1953, pp. 623-624) and Halmos (1950, Section 48) present some examples of non-existence. If conditional densities do not exist, then what we would expect to be the probability limit of the kernel estimator in the present context is either meaningless or difficult to interpret.[3]

The idea here is to let the common shocks be as general as possible and to work with well-defined conditional densities that adhere as closely as possible to the standard kernel literature. To do so, we appeal to the disintegration theory for conditional distributions that can be found in Pollard (2002), Dellacherie and Meyer (1978), and Hoffmann-Jorgensen (1994). We find that an important sufficient condition to guarantee the existence of conditional densities is that the common shocks must belong to a separable metric space equipped with the Borel σ-field. We conclude that the

[3]Formally, the probability limit of the kernel estimator for a nonstationary process can be obtained using the concept of local time, as in Wang and Phillips (2009). However, the probability limit of the kernel regression estimator may not be measurable with respect to the conditioning variables, including the common shocks. This is a particularly important problem when we extend the results to panel data models, as in Souza-Rodrigues (2014).

sufficient conditions are mild and not very restrictive in practice.[4]

Given the existence of conditional densities, we adjust the standard assumptions of the kernel literature to the present case. We show that the Nadaraya-Watson kernel estimator converges in probability to the Kolmogorov conditional expectation given the sigma-field generated by the common shocks. The optimal rate of convergence is the same as the rate obtained when the observations are i.i.d. The asymptotic distribution is mixed normal with weights depending on the common shocks. It is obtained by exploring a martingale difference sequence central limit theorem. We find that inference depends on how the common shocks affect the regression variables. A dichotomy similar to that of Andrews (2005) is present here: if the dependent variable is mean independent of the common shocks given the explanatory variables, the usual t-test has the correct size; but if the dependent variable is not mean independent, the t statistic diverges to infinity in probability under the null hypothesis.

The closest paper in the literature to ours is that of Andrews (2005), who considers a linear regression model in the presence of general common shocks. He shows that the least-squares estimator converges in probability to Kolmogorov conditional expectations given the σ-field generated by the common shocks. The random probability limit is a well-defined object because the Kolmogorov conditional expectation always exists. Andrews, therefore, does not need to guarantee the existence of conditional densities. Extending his results to a nonparametric model is important because parametric models may be misspecified. We show that the price to be paid is that mild restrictions must then be imposed on the nature of the common shocks.

The nonparametric version of the standard factor model is a special case of the model considered here. For this class of models we show that, even though the kernel regression converges in probability to a random object measurable with respect to the common shocks, it is possible to identify and estimate the slope of the regression function. However, its location (e.g., the intercept in a linear model) is not identified even if we normalize common shocks to have a zero mean. To identify and estimate location, the dependent variable must be mean independent of the common shocks given the regressors.

Common factor models are typically applied to panel data sets (e.g., Pesaran and Tosetti (2011),

[4] Although separability is not a necessary condition, it seems difficult to avoid it if we are to obtain existence of conditional densities – see the discussion about the role of separability in the Appendix. Note that several separable metric spaces satisfying the sufficient conditions are available, but careful interpretation is needed in particular cases. For instance, suppose that an infinite-dimensional common shock can be well-approximated by a finite dimensional object. Because the sigma-field generated by the common shock may be different from the sigma-field generated by the approximating object, the conditional expectations given the common shock and given the finite-dimensional object are different. Ignoring this difference leads to problems such as the Borel paradox.

Kuersteiner and Prucha (2013), Chudik and Pesaran (2015) and Forchini and Peng (2016)). We view the present paper as a first step towards nonparametric panel data models that may incorporate a more general and flexible common factors structure. Indeed, in a companion paper, Souza-Rodrigues (2014) develops a two-step nonparametric estimator that requires a "large-N, large-T" data set for a generalized regression model based on the identification results of Berry and Haile (2009). The estimator applies equally to data sets with a large number of individuals in different groups and a large number of groups. The empirical application in Souza-Rodrigues (2014) considers the impact of hospital volumes of surgical procedures on individual health status (e.g., mortality rate).[5] Group-level observables (i.e., hospital volume of surgeries) may be correlated with group-level unobservables (hospital unobserved quality), which, by its turn, may be indexed by individual characteristics (since an unobserved hospital characteristic that is helpful for patients with some demographic characteristics may not be as helpful for other patients). The strategy proposed by Souza-Rodrigues (2014) is to run a nonparametric regression of individual outcomes on individual observables within each group (hospital) in the first step. It is a nonparametric regression with common shocks where the common shocks are the group-level observables and unobservables. Because the group-level unobservables may be a (random) function of individual characteristics, it is important to allow for this possibility, as we do here.[6] The results of the present paper can be incorporated in other nonlinear panel data settings.

The present paper also relates to the literature of spatial dependence.[7] Typically, in this literature, common shocks are presumed to have predominantly local effects and the dependence is modelled as a function of an exogenously given spatial or economic distance – with some form of stationary mixing condition analogous to the time series data. Recent nonparametric versions of spatial models have been considered by Martins-Filho and Yao (2009) and by Gerolimetto and Magrini (2010), among others. Although the present paper can incorporate common shocks with differential local effects (e.g., assuming that individual factor loadings include geographic location),

[5] The motivation for this application is that numerous studies have documented an inverse relationship between hospital volumes of operations and mortality rates (Finks, Osborne and Birkmeyer (2011)). This suggests that thousands of deaths per year could have been prevented if hospitals with inadequate experience (i.e., with low volume of operations) had performed fewer surgical procedures. The evidence, however, is weak for most operations. Furthermore, existing papers have estimated parametric models that may be misspecified, and have not considered the potential correlation between hospital volume of operations and hospital unobserved quality.

[6] The second step runs a nonparametric instrumental variable regression across groups (hospitals) of the predicted outcome obtained in the first step on the group-level observables. It separates the impacts of group-level observables (hospital volume of surgeries) and unobservables (hospital unobserved quality).

[7] See, e.g., Anselin (1988); Conley (1999); Kelejian and Prucha (1999); Lee (2002); Martins-Filho and Yao (2009); Gerolimetto and Magrini (2010); Lee and Yu (2014); Su and Yang (2015); Bester, Conley, Hansen, and Vogelsang (2016); and the discussion in Arbia (2016). For a recent survey, see Lee and Yu (2010).

we do not allow individual outcomes to depend on the characteristis of other individuals. We therefore view spatial dependence models as complementary to ours.

Robinson (2011) provides an alternative way of modeling cross-sectional dependence. He considers a nonparametric kernel regression in which the disturbances are represented by a (possibly infinite) sum of independent random variables with unknown weights. The structure in the disturbances is sufficiently rich to cover spatial dependence models, but, since it does not require known economic distances, it can accommodate stronger forms of dependence than mixing conditions. Robinson (2011) investigates the properties of kernel estimators and Robinson and Lee (2016) study the properties of sieve estimators within this framework. The present paper can accomodate disturbances of the type represented by Robinson (2011), but with a vector of common shocks in place of the vector of independent random variables. We do not require the vector of common shocks to be independent random variables, and we allow for potentially correlated random weights in the summation term for the disturbances. However, the restrictions we need to impose on the common shocks differ from the assumptions in Robinson (2011). Furthermore, we require i.i.d. sampling schemes that are neither assumed by Robinson (2011) nor by the spatial dependence literature. Our model is therefore neither more general than nor is it a special case of Robinson's model.

The paper is organized as follows: Section 2 presents the regression model and discusses sufficient conditions to guarantee the existence of conditional densities; Section 3 establishes the asymptotic properties of the Nadaraya-Watson kernel regression estimator and discusses its implications; Section 4 concludes. The Appendix presents the disintegration theory and briefly discusses the role of separability of common shocks in the existence of conditional densities. The Supplemental Material (Souza-Rodrigues (2016)) presents results for the kernel density estimator, contains all relevant proofs, and discusses the probabilistic framework adapted from Andrews (2005) that justifies the approach taken here.

2 Regression Model and Conditional Densities

The data set is $\{Y_i, X_i : i = 1, ..., n\}$, where $Y_i \in \mathcal{Y}$ $(\subseteq \mathbb{R})$ and $X_i \in \mathcal{X}$ $(\subseteq \mathbb{R}^k)$. Consider the model:

$$Y_i = m\left(X_i, C\left(S_i\right)\right) + \varepsilon_i, \tag{1}$$

where $S_i \in \mathcal{S}$ $(\subseteq \mathbb{R}^{d_s}$, with $d_s \in \mathbb{N})$ is a vector of individual-specific random variables; $C\left(.\right) \in \mathcal{C}$ is the common shock; and ε_i is the idiosyncratic error. Some components of S_i may be observable (in which case, it may be incorporated in X_i) or it may be completely unobservable. We allow

the common shock $C(.)$ to be either a random vector (possibly infinite-dimensional) or a random function of S_i. In the latter case, the common shocks may affect individuals differently. As usual, we use upper-case letters to denote random quantities and lower-case letters to denote realizations.

The standard parametric factor model is a special case of our model and is typically written as

$$Y_i = \alpha + X_i'\beta + U_i, \quad U_i = \sum_{j=1}^{J} S_{ij} C_j + \varepsilon_i, \tag{2}$$

where $S_i = (S_{i1}, ..., S_{iJ})$ is the vector of individual-specific factor loadings; $C = (C_1, ..., C_J)$ is the vector of unobserved common factors; ε_i is the idiosyncratic error that is independent of (X_i, S_i, C) and has zero mean; and (α, β) is the vector with the parameters of interest. Cross-sectional dependence in the disturbances is generated by the term $S_i'C$. The standard model can also accommodate cross-sectional dependence on regressors X_i. For example, consider the expanded vectors $C = \left(C^1, C^2\right)$ and $S_i = \left(S_i^1, S_i^2\right)$, and take $X_i = S_i^{1\prime} C^1$ and $U_i = S_i^{2\prime} C^2 + \varepsilon_i$. Note that if $C^1 = C^2$, then X_i and U_i are correlated even when S_i^1 and S_i^2 are independent of each other (Andrews (2005), Pesaran (2006), and Bai (2009)). The nonparametric version of (2) takes $Y_i = m_1(X_i) + U_i$, with the same structure for the disturbances U_i.[8]

The standard factor model (2) is a special case of our model (1) with the regression function given by the linear and additively separable $m(X_i, C(S_i)) = \alpha + X_i'\beta + C(S_i)$, and the common shock function given by $C(S_i) = \sum_{j=1}^{J} S_{ij} C_j$. We therefore generalize the standard model in the following ways: (i) we let the regression function $m(.)$ be nonparametric; (ii) we allow the regressors X_i to freely interact with the common shock $C(.)$; and (iii) we let the common shock be a general function of individual-specific factor loadings S_i (subject to the restrictions discussed below). Furthermore, factor models typically impose independence between S_i and (X_i, C) and assume that $C = (C_1, ..., C_J)$ is a mutually independent vector, while we do not need to impose these independence assumptions. We, however, do not consider a fully nonseparable model; we maintain the additive separability assumption in the idiosyncratic error ε_i.

Robinson (2011) also considers a nonparametric version of (2) but with another structure for U_i. He considers the model:

$$Y_i = m_1(X_i) + U_i, \quad U_i = \sigma_i(X_i) \sum_{j=1}^{\infty} b_{ij} e_j, \quad \sum_{j=1}^{\infty} b_{ij}^2 < \infty, \tag{3}$$

where σ_i are scalar unknown functions; $e_j's$ are independent random variables with zero mean and

[8] In a panel data setting, one typically allows for time-varying regressors X_{it}, but restricts S_i so that it does not vary over time, and the common shock C so that it does not vary across individuals. Fixed-effect panel data models let X_{it} and S_i to be correlated.

unit variance; and $b'_{ij}s$ are unknown fixed weights.[9] The present paper compares to Robinson (2011) when the following holds: $m(X_i, C(S_i)) = m_1(X_i) + \sigma_i(X_i) C(S_i)$, with $C(S_i) = \sum_{j=1}^{\infty} S_{ij} C_j$, where $S_{ij} = b_{ij}$ and $C_j = e_j$. Unlike Robinson (2011), we allow for (potentially correlated) random coefficients $b'_{ij}s$, and do not restrict $e'_j s$ to be independent variables with zero mean and unit variances. The restrictions we need to impose on the function $C(S_i)$ are discussed below and are of a different nature than the assumptions used by Robinson (2011).

Data Generation. Denote the vector $W_i = (Y_i, X_i, S_i, C) \in \mathcal{W}$, where $\mathcal{W} \subseteq \mathcal{Y} \times \mathcal{X} \times \mathcal{S} \times \mathcal{C}$. Define the measurable space $(\mathcal{W}, \mathcal{A})$, where $\mathcal{A}$ is the Borel sigma-field. The random elements $\{W_i : i \geq 1\}$ are defined on $\left(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}\right)$, where $\mathcal{W}^{\mathbb{N}}$ is the product space and $\mathcal{A}^{\mathbb{N}}$ is the product Borel sigma-field on $\mathcal{W}^{\mathbb{N}}$. We suppose the common shocks across observations are captured by the σ-field generated by C, denoted by $\sigma(C) \subset \mathcal{A}^{\mathbb{N}}$. We impose the following assumption:

Assumption 1 *The sequence* $\{W_i : i \geq 1\}$ *is i.i.d. conditional on the* σ*-field* $\sigma(C) \subset \mathcal{A}^{\mathbb{N}}$.

As shown by Andrews (2005), this assumption is valid when the units are drawn randomly from the population. One difference between the present paper and Andrews (2005) is that he states the existence of *some* σ-field such that the data is i.i.d. conditional on it without specifying a priori how this σ-field is constructed, while we impose more structure and state explicitly how the σ-field is generated. Andrews' framework is, therefore, more general than ours in this respect. Note that neither the spatial dependence models nor Robinson's (2011) approach require random sampling.[10]

2.1 Existence of Conditional Densities

Because we make use of the Nadaraya-Watson kernel estimator, and because the kernel estimator requires the existence of conditional densities, we now discuss the existence problem.

To guarantee the existence of conditional densities that allow for very general common shocks, we make use of the disintegration theory. Disintegration of a probability measure is a collection of regular conditional probabilities, each satisfying (i) a concentration property (i.e., conditional on an event, the probability of its complement is zero) and (ii) a decomposition property (i.e., the probability of an event is a weighted sum of the conditional probability measures, also known as

[9] Note that this approach does not require known economic distances, but can readily accommodate them by taking $U_i = \sigma_i(X_i) \sum_{j=1}^{\infty} b_{ij} e_i$, $e = (e_1, ..., e_n)$ and by making some assumptions regarding how b_{ij} depends on the distance $|i - j|$.

[10] When Andrews (2005) specializes to factor structure models, he imposes more restrictions on the common shocks, which makes his approach more similar to ours.

the law of total probability).[11] The reader unfamiliar with disintegration theory might want to read the Appendix (or the references cited there) before proceeding.

Define the sub-vector $Z_i = (Y_i, X_i, S_i) \in \mathcal{Z} \subseteq \mathcal{Y} \times \mathcal{X} \times \mathcal{S}$, for $i \geq 1$. We want to guarantee the existence of the conditional density of Z_i given C. By Assumption 1, the probability distribution of $\{W_i : i \geq 1\}$, denoted by $\mathcal{P}^{\mathbb{N}}$, is exchangeable on $\left(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}\right)$. Call P^i the marginal distribution of W_i under $\mathcal{P}^{\mathbb{N}}$. We impose the following:

Assumption 2 *(i) $\mathcal{W}$ is a metric space.*

(ii) λ is a sigma-finite Radon measure on $(\mathcal{W}, \mathcal{A})$.[12]

(iii) C maps $(\mathcal{W}, \mathcal{A})$ into $(\mathcal{C},\mathcal{B})$. $\mathcal{C}$ is a separable metric space and $\mathcal{B}$ is the Borel σ-field.

(iv) μ is a sigma-finite measure on $(\mathcal{C},\mathcal{B})$. Let the measure $\lambda\left(C^{-1}\right)$ induced by C and λ on $(\mathcal{C},\mathcal{B})$ be absolutely continuous with respect to μ.

(v) Let P^i, for any $i \geq 1$, *be absolutely continuous with respect to λ. Denote its Radon-Nikodym density by $f_i(z, c)$.*

Assumption 2(iii) requires $\mathcal{C}$ to be a separable metric space. This is trivially satisfied when $\mathcal{C}$ belongs to a finite-dimensional Euclidean space. However, if C is an infinite dimensional vector of random variables, we need restrictions such as $\mathcal{C} = \ell_p$, for some $1 \leq p < \infty$, where ℓ_p is the space of sequences with finite $\|\cdot\|_p$-norm, and we need to rule out the case $\mathcal{C} = \ell_\infty$, because ℓ_∞ is non-separable. Similarly, if C is a random function of S_i, it must belong to spaces such as the $\mathcal{L}_p(\mathcal{S})$ space for $1 \leq p < \infty$, or the space of bounded and continuos functions defined on a closed bounded subset of $\mathcal{S}$ and equipped with the sup-norm, or the Hölder space, etc. But it cannot belong to the space of bounded functions with the sup-norm, $\mathcal{L}_\infty(\mathcal{S})$, because it is not separable. See the discussion about the role of separability for existence of conditional densities in the Appendix.[13]

The restrictions in Assumption 2 are mild and sufficient to guarantee the existence of conditional densities of Z_i given C for any $i \geq 1$. The reason for sufficiency is the following: first, Assumptions

[11] A regular conditional probability, $\Pr(Y|X = x)$, is a family of probability distribution such that (i) for a fixed x, $\Pr(\cdot|X = x)$ is a probability measure, and (ii) for a fixed measurable set A, $\Pr(A|X = x)$ is a measurable function mapping x to $[0, 1]$.

[12] The measure λ is Radon if (i) $\lambda(K) < \infty$ for each compact K and $\lambda(B) = \sup\{\lambda(K) : B \supseteq K,\ K \text{ compact}\}$.

[13] It is possible to characterize all the objects in Assumption 2 when $\mathcal{W} = \mathcal{Z} \times \mathcal{C}$. First, we have that (i) $\mathcal{W}$ is a separable metric space provided that $\mathcal{C}$ is a separable metric space as well, and (ii) the Borel σ-field $\mathcal{A}$ on $\mathcal{W}$ equals the product Borel σ-field $\mathcal{A}_Z \otimes \mathcal{B}$, where we denote $\mathcal{A}_Z$ the Borel σ-field on $\mathcal{Z}$ (see Folland (1999), Proposition 1.5). Second, let π_c be the projection of $\mathcal{W}$ onto the coordinate space $\mathcal{C}$, i.e., $\pi_c : \mathcal{W} \to \mathcal{C}$. Then, (i) the sub-sigma field $\pi_c^{-1}(\mathcal{B})$ is contained in $\mathcal{A}$ and (ii) because $C(w) = \pi_c(w)$, for all $w \in \mathcal{W}$, the sigma-field generated by C is $\sigma(C) = \pi_c^{-1}(\mathcal{B}) \subset \mathcal{A}$. Furthermore, if we define the sigma-finite Radon λ on $(\mathcal{W}, \mathcal{A})$ to be the product measure $\lambda = \nu \otimes \mu$, where ν is defined on $(\mathcal{Z}, \mathcal{A}_Z)$ and μ on $(\mathcal{C}, \mathcal{B})$, then the measure $\lambda\left(C^{-1}\right)$ induced by C and λ on $(\mathcal{C},\mathcal{B})$ equals μ, and, so, $\lambda\left(C^{-1}\right)$ is (trivially) absolutely continuous with respect to μ. Finally, we have to assume both ν and μ are sigma-finite Radon, so that λ is sigma-finite Radon on $\mathcal{A}$ as well.

2(i)-(iv) are sufficient for the sigma-finite Radon measure λ to have a (C, μ)-disintegration; i.e., they guarantee the existence of a collection of measures, denoted by $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$, that satisfy the aforementioned concentration and decomposition properties (but note that $\lambda_c's$ do not have to be probability measures; see Definition 3 and Theorem 1 in the Appendix).

Second, if the disintegration $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$ exists and the probability measure P^i on $(\mathcal{W}, \mathcal{A})$ is absolutely continuous with respect to λ with density $f_i(z, c)$ (Assumption 2(v)), then two implications follow (see Theorem 2 in the Appendix): (i) the probability distribution of C induced by P^i (i.e., the image measure $Q^i = P^i\left(C^{-1}\right)$) is absolutely continuous with respect to μ with density

$$q_i(c) \equiv \int f_i(\widetilde{z}, c)\, d\lambda_c(\widetilde{z}) \tag{4}$$

and (ii) the probability measure P^i has a conditional distribution given C, denoted by the collection $\mathcal{P}^i = \left\{P_c^i : c \in \mathcal{C}\right\}$, where P_c^i is defined by having density

$$f_i(z|c) \equiv \frac{f_i(z, c)}{q_i(c)} 1\{0 < q_i(c) < \infty\} \tag{5}$$

with respect to λ_c for Q^i-almost all $c \in \mathcal{C}$. The conditional density $f_i(z|c)$ is therefore similar to elementary conditional densities: it is the ratio of the joint density $f_i(z, c)$ and the marginal $q_i(c)$. But it does not require C to belong to a finite-dimensional Euclidean space.

Because C is common to all i, the equality $Q = Q^i$ follows for all $i \geq 1$. In addition, $f_i(z|c) = f_j(z|c)$ for all $i \neq j$ and for Q-almost all $c \in \mathcal{C}$ by Assumption 1. We state this result as a Lemma:

Lemma 1 *Let Assumptions 1 and 2 hold. Then there exist conditional densities of Z_i given C, for all $i \geq 1$, defined by*

$$f_i(z|c) = \frac{f_1(z, c)}{q(c)} 1\{0 < q(c) < \infty\}, \tag{6}$$

for Q-almost all $c \in \mathcal{C}$, where $q(c) \equiv \int f_1(\widetilde{z}, c)\, d\lambda_c(\widetilde{z})$.[14]

Example 1 *Suppose S_i is scalar and $\mathcal{C}$ is the separable Hilbert space $\mathcal{L}_2(\mathcal{S})$. Take a basis $\left\{\phi_j\right\}_{j=1}^{\infty}$ for $\mathcal{L}_2(\mathcal{S})$ and represent the common shock by $C(S_i) = \sum_{j=1}^{\infty} C_j \phi_j(S_i)$, where $C_j \in \mathbb{R}$ for $j \geq 1$. Note that one can define $S_{ij} = \phi_j(S_i)$, in which case the random coefficients are not independent of each other. More important for us is to note that selecting a function in $\mathcal{C}$ is equivalent to selecting the*

[14] Note that we can manipulate the conditional density (6) on $\mathcal{Z} \otimes \mathcal{C}$ as is usually done. Fix $C = c$ and think of $\mathcal{Z} \otimes \{c\}$ as a copy of $\mathcal{Z}$ embedded into the product space. For a fixed $c \in \mathcal{C}$, take the measure λ_c living on $\mathcal{Z} \otimes \{c\}$ to coincide with the Lebesgue measure on $\mathcal{Z}$. If $r(\cdot)$ is a vector-valued function with $E\|r(Z)\| < \infty$, then

$$E[r(Z)|C = c] = \int r(\widetilde{z}) dP_c(\widetilde{z}) = \int r(\widetilde{z}) f(\widetilde{z}|c)\, d\lambda_c(\widetilde{z}) = \int r(\widetilde{z}) f(\widetilde{z}|c)\, d\widetilde{z}.$$

infinite dimensional vector $\{C_j\}_{j=1}^{\infty}$ *in* ℓ_2. *Let* $\mathcal{B}(\mathcal{L}_2)$ *be the Borel* σ*-field on* $\mathcal{L}_2$ *and* $\mathcal{B}(\ell_2)$ *be the Borel* σ*-field on* ℓ_2. *Because the spaces* $\mathcal{L}_2$ *and* ℓ_2 *are homeomorphic, their topologies are equivalent and, so,* $\mathcal{B}(\mathcal{L}_2)$ *and* $\mathcal{B}(\ell_2)$ *are equivalent.*[15] *As a result, the event* $\{C(\cdot) = c\}$ *on* $\mathcal{L}_2$ *is equivalent to the (potentially more intuitive) event* $\{(C_1, C_2, ...) = (c_1, c_2, ...)\}$ *on* ℓ_2. *In addition,* ***conditioning*** *on* $\{C(\cdot) = c\}$ *is equivalent to* ***conditioning*** *on* $\{(C_1, C_2, ...) = (c_1, c_2, ...)\}$. *We have, therefore,* $f(z|c(\cdot)) = f(z|c_1, c_2, ...)$ *and*

$$\begin{aligned} \Pr(Z_i \in A | C(\cdot) = c) &= \Pr(Z_i \in A | C_1 = c_1, C_2 = c_2, ...) \\ &= \int_A f(\widetilde{z}|c_1, c_2, ...)\, d\widetilde{z}, \end{aligned} \tag{7}$$

for any measurable set A.[16]

Example 1 intends to translate properties of conditional probabilities given an element in some abstract space of functions into properties in (hopefully) more concrete spaces defined by random vectors. Example 1, however, does not apply when $\mathcal{C}$ is not a Hilbert space. Although we may approximate any of the separable metric spaces by other simpler spaces, the conditioning argument does not hold without running into problems such as the Borel paradox (see, e.g., Rao (1988)). For instance, take $\mathcal{C}$ to be the set of bounded and continuous functions, $(BC(\mathcal{S}), \|\cdot\|_\infty)$. It is separable and any $c \in \mathcal{C}$ can be well approximated by a polynomial of order $J < \infty$, say $p^J(\cdot)$ with some coefficients $(b_j)_{j=1}^J$. Because we can take J such that $\left\|c - p^J\right\|_\infty < \varepsilon$, for some $\varepsilon > 0$, the probability of the event $\{C(\cdot) = c\}$ is close to the probability of the event $\left\{(B_j)_{j=1}^J = (b_j)_{j=1}^J\right\}$. However, the topology of $(BC(\mathcal{S}), \|\cdot\|_\infty)$ is not the same as the topology of the Euclidean $\mathbb{R}^J$ for any finite J. So the Borel σ-field on $BC(\mathcal{S})$ is different from the Borel σ-field on any $\mathbb{R}^J$. Conditioning on different σ-fields delivers different conditional probability distributions, and, so, we are not guaranteed to have $\Pr(Z \in A | C(\cdot) = c)$ close to $\Pr\left(Z \in A | \cap_{j=1}^J \{B_j = b_j\}\right)$ for all measurable sets A. We can still obtain the existence of conditional densities, but we cannot derive conclusions based on some approximation $\sum_{j=1}^{J} b_j p_j(S_i)$ for $C(S_i)$, no matter how large J is.

[15] Any infinite-dimensional separable Hilbert space, say $\mathcal{H}$, is isometrically isomorphic to a suitable $\ell_2(I)$, where the cardinality of the set I is the cardinality of an arbitrary Hilbertian basis for $\mathcal{H}$, i.e., there exists a linear operator $L : \mathcal{H} \to \ell_2(I)$, such that $\|Lh\|_2 = \|h\|_{\mathcal{H}}$, where $h \in \mathcal{H}$, $\|\cdot\|_{\mathcal{H}}$ is the norm on $\mathcal{H}$ and $\|\cdot\|_2$ is the ℓ_2-norm.

[16] Conditioning on the event $\{(C_1, C_2, ...) = (c_1, c_2, ...)\}$ is only one possibility. For some $a \in \mathbb{R}$, we could condition either on the event $\{C(S_i) = a\} = \left\{\sum_{j=1}^{\infty} C_j \phi_j(S_i) = a\right\}$, or on the event $\{c(S_i) = a\} = \left\{\sum_{j=1}^{\infty} c_j \phi_j(S_i) = a\right\}$, where the randomness of the event comes from S_i, or on $\{C(s) = a\} = \left\{\sum_{j=1}^{\infty} C_j \phi_j(s) = a\right\}$, where the randomness comes from $(C_1, C_2, ...)$.

3 Regression Estimator

Next, we consider the properties of the Nadaraya-Watson kernel regression estimator:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)}, \tag{8}$$

where $K(\cdot)$ is the kernel function and h_n is the bandwidth. As previously mentioned, the objective here is to work as closely as possible to the standard kernel literature. The assumptions we impose are therefore similar to the standard assumptions (see Pagan and Ullah (1999)), but with the population density and regression function substituted for the corresponding conditional functions and with the extra "Q-almost all c" qualifiers added. For brevity, we relegate the properties of the kernel density estimator to the Supplemental Material.

We maintain Assumptions 1 and 2 from now on. In addition, we impose the following conditions:

Condition 1 *Let K be the class of all Borel measurable nonnegative bounded real-valued functions $K(u)$ such that (i) $\int K(u)du = 1$; (ii) $\int |K(u)|\, du < \infty$; (iii) $|K(u)|\, \|u\|^k \to 0$ as $\|u\| \to \infty$; (iv) $\kappa = \int K^2(u)du < \infty$; (v) $\sup_u |K(u)| < \infty$; and (vi) $\mu_2 = \int u^2 K(u)du < \infty$.*

Condition 2 *For Q-almost all $c \in \mathcal{C}$, the conditional density $f(x|c)$ is continuous at any point x_0.*

Condition 3 *(i) $h_n \to 0$ as $n \to \infty$ and (ii) $nh_n \to \infty$ as $n \to \infty$.*

Condition 4 *For Q-almost all c, (i) $f(x|c)$ is twice continuously differentiable with respect to x in some neighborhood of x_0, (ii) the second-order derivatives of $f(x|c)$ with respect to x are bounded in this neighborhood.*

Condition 5 *For Q-almost all c, the point x_0 is in the interior of the support of X conditional on $\{C = c\}$,and $f(x_0|c) \geq \xi > 0$, for some finite ξ.*

Condition 6 *The kernel K is a symmetric function satisfying $\int uK(u)du = 0$.*

Condition 7 *(i) $E[\varepsilon_i | X_i, \sigma(C)] = 0$ a.s., and (ii) let $\sigma^2(x,c) = E\left(\varepsilon_i^2 | X_i = x, C = c\right)$ and assume $\sigma^2(X,C) < \infty$ a.s..*

Condition 8 *For Q-almost all c, the function $m(x,c)$ is twice continuously differentiable with respect to x in some neighborhood of x_0.*

Conditions 1-5 suffice to obtain the asymptotic properties of the kernel density estimator (consistency, rate of convergence, and asymptotic distribution; see the Supplemental Material). Condition 6 is standard in the literature.

Condition 7(i) implies $m(x,C) = E[Y|X=x,\sigma(C)]$. In the standard factor model, this translates into

$$m(x,C) = E[Y|X=x,C] = \alpha + x'\beta + \sum_{j=1}^{J} E[S_{ij}|X=x,C]\,C_j. \tag{9}$$

Note that $m(x,C)$ is a random object because C has not been fixed. Typically in the literature, S_i is assumed to be independent of (X_i,C), in which case $E[S_{ij}|X=x,C] = E[S_{ij}] \equiv b_{ij}$,where b_{ij} is an unknown constant. Unlike the standard model, here we allow J to be infinite (as long as C belongs to an appropriate separable metric space), we do not require S_i to be independent of (X_i,C), and we allow for more complicated interactions between X and C.

Condition 7(ii) allows for conditional heteroskedasticity; and Condition 8 is used to apply Q-almost sure Taylor expansions similar to what is usually done in the kernel literature.

Remark 1 *Condition 8 requires $m(x,c)$ to be twice continuously differentiable in x for almost all c. To fix ideas, consider the following case: let $S_i = X_i$, $C(\cdot) \in \mathcal{L}_2(\mathcal{X})$ and $m(X,C) = m_1(X_i) + C(X_i)$. Conditioned on the event $\{X_i = x\} \cap \{C(\cdot) = c\}$ we have that*

$$\begin{aligned} E[Y_i|X_i = x, C(\cdot) = c] &= m_1(x) + \left(\sum_{j=1}^{\infty} c_j \phi_j(x)\right) \\ &= m(x,c), \end{aligned}$$

while conditioning only on the event $\{X_i = x\}$, we obtain the random object

$$\begin{aligned} E[Y_i|X_i = x, C(x)] &= m_1(x) + \left(\sum_{j=1}^{\infty} C_j \phi_j(x)\right) \\ &= m(x,C). \end{aligned}$$

So, to satisfy Condition 8, we need $E[Y_i|X_i, C(X_i)]$ to be twice continuously differentiable with respect to both the first and second arguments, and we need $C(\cdot)$ to be twice continuously differentiable with respect to x as well.[17]

To obtain consistency of $\widehat{m}(x)$, we first show that the kernel density converges in probability to the conditional density $f(x|C)$. Then we prove that the mean-squared error of $\widehat{m}(x)$ conditional on $\sigma(C)$ converges to zero in probability. Finally, consistency follows by the dominated convergence

[17] It should be clear that it is not possible to separately identify $m_1(X)$ from $C(X)$ in this example.

theorem. We then show that the rate of convergence is the same as the rate of convergence without common shocks. Pointwise asymptotic distribution is obtained using the martingale difference sequence central limit theorem.

Proposition 1 *Let $E[\cdot|X=\{x_i\}_{i=1}^n]$ denote the conditional expectation given x_i, $i=1,...,n$. Let Assumptions 1 and 2 and Conditions 1-8 hold. Then*

1. $\widehat{m}(x) \xrightarrow{p} m(x,C)$ *as* $n \to \infty$.
2. $\widehat{m}(x) - m(x,C) = O_p\left(n^{-\frac{2}{4+k}}\right)$.
3. *Suppose also that* $\int |K(u)|^{2+\delta} du < \infty$ *and* $E\left[|\varepsilon_i|^{2+\delta}\right] < \infty$, *for some* $\delta > 0$. *Define* $\sigma^2(x,C) = E\left(\varepsilon_i^2 | X_i = x, C\right)$. *Then, (i) as* $n \to \infty$

$$\sqrt{nh_n^k}\left(\widehat{m}(x) - E\left[\widehat{m}(x)|X=\{x_i\}_{i=1}^n, C\right]\right) \xrightarrow{d} \left(\frac{\sigma^2(x,C)}{f(x|C)}\int K^2(u)\,du\right) N(0,1)$$

and (ii) if, in addition, $\sqrt{nh_n^k}h_n^2 \to 0$ *as* $n \to \infty$, *then*

$$\sqrt{nh_n^k}\left(\widehat{m}(x) - m(x,C)\right) \xrightarrow{d} \left(\frac{\sigma^2(x,C)}{f(x|C)}\int K^2(u)\,du\right) N(0,1)$$

as $n \to \infty$.

Proposition 1.1 shows that the kernel regression estimator converges in probability to the random object $m(x,C) = E[Y|X=x, \sigma(C)]$. In general, $m(x,C)$ is different from the conditional expectation $m(x) = E[Y|X=x]$: the equality $m(x,C) = m(x)$ only holds when Y is mean independent of C given X. To see how this difference may affect the interpretation of potential estimands, take the standard factor model as an example.[18] In this case, $m(x,C)$ is given by (9), while $m(x)$ is given by

$$m(x) = E[Y|X=x] = \alpha + x'\beta + \sum_{j=1}^{J} E[S_{ij}C_j|X=x]. \tag{10}$$

If we assume, as is usually done, that S_i is independent of (X_i, C), we have that $E[S_{ij}C_j|X=x] = b_{ij}E[C_j|X=x]$. If there is no cross-sectional dependence on regressors resulting from the common shocks, then $E[C_j|X=x] = E[C_j]$. In addition, if we normalize $E[C_j] = 0$ for all j, then $m(x) = \alpha + x'\beta$, while $m(x,C) = \alpha + x'\beta + \sum_{j=1}^{J} b_{ij}C_j$. Because Y is not mean independent of C given X, $\widehat{m}(x) \xrightarrow{p} m(x,C) \neq m(x)$. Although we cannot estimate $m(x)$ consistently, it is possible to

[18] Recall that the nonparametric version of the factor model takes $Y_i = m_1(X_i) + U_i$, with $U_i = \sum_{j=1}^{J} S_{ij}C_j + \varepsilon_i$. The parametric model imposes $m_1(x) = \alpha + x'\beta$.

identify and estimate β by noting that $m(x_1, C) - m(x_2, C) = (x_1 - x_2)'\beta$, for $x_1 \neq x_2$. Similarly, for nonparametric factor models, $Y_i = m_1(X_i) + U_i$, one can identify and estimate the slope of $m_1(x)$. However, the presence of the common shocks $\sum_{j=1}^{J} b_{ij}C_j$ prevents the identification of the intercept α in the linear model (and the identification of the location of $m_1(x)$ in the nonparametric model) even if we normalize $E[C_j] = 0$ for all j.

Remark 2 *The nonparametric factor model with $J = \infty$, $E[S_{ij}|X = x, C] = b_{ij}$, and $E[C_j] = 0$ for all j, has a structure similar to the one proposed by Robinson (2011). Yet, while Robinson shows that the kernel regression estimator converges in probability to $m(x)$, we obtain convergence to $m(x, C)$. An important distinction comes from the assumption on the sampling process. Because we have exchangeable data given the common shocks (Assumption 1), the conditions we impose are not sufficient to "get rid of" C in the limit. Robinson (2011), in contrast, does not impose the conditional i.i.d. sampling process.*

Returning to the standard factor model, if we assume now the presence of cross-sectional dependence on regressors captured by, say, $X_i = S_i^{1\prime}C$ with $S_i = \left(S_i^1, S_i^2\right)$, then $E[C_j|X = x] \neq E[C_j]$ and

$$\begin{aligned} m(x, C) &= \alpha + x'\beta + \sum_{j=1}^{J} b_{ij}C_j, \\ m(x) &= \alpha + x'\beta + \sum_{j=1}^{J} b_{ij}E[C_j|X = x]. \end{aligned}$$

Again, Y is not mean independent of C given X, so that $\widehat{m}(x) \xrightarrow{p} m(x, C) \neq m(x)$, and it is still possible to identify β in the parametric model (and the slope of $m_1(x)$ in the nonparametric version).[19]

In the standard factor model, Y is mean independent of C given X only if the common shocks have no direct effect on Y. This is the case when $E[S_{ij}] = b_{ij} = 0$. When this is true, $m(x, C) = m(x)$ and the kernel regression converge in probability to $m(x)$, even when there is cross-sectional dependence on X. In this case, we identify both parameters α and β in the linear model, and $m_1(x)$ in the nonparametric model. Note that assuming $E[S_{ij}] = 0$ for all j is not an innocuous normalization but a substantive assumption.

Remark 3 *The last case is similar to Andrews (2005). Let $X_i = S_i^{1\prime}C^1$ and $U_i = S_i^{2\prime}C^2 + \varepsilon_i$,*

[19] Note that if we were able to estimate the conditional expectation $m(x)$ instead of $m(x, C)$, it would be impossible to separate $x'\beta$ from $\sum_{j=1}^{J} b_{ij}E[C_j|X = x]$, and, so, we would not be able to identify β.

where $C = \left(C^1, C^2\right)$ *is mutually independent and* $S_i = \left(S_i^1, S_i^2\right)$ *(see Andrews' Assumption SF1).*[20] *Imposing* $E\left[S_{ij}|X = x, C\right] = E\left[S_{ij}\right] = 0$*, is similar to imposing Andrews' condition SF3. Assuming condition 7(i) together with mutually independence* $\left(S_i^1, S_i^2, \varepsilon_i\right)$ *is similar to Andrews' condition SF2.*

Proposition 1.2 shows that the rate of convergence here is the same as the rate of convergence without common shocks. Proposition 1.3 presents the asymptotic distribution of the kernel regression estimator. It shows that even when $\widehat{m}(x) \xrightarrow{p} m(x, C) = m(x)$, the common shocks affect the asymptotic distribution of the kernel regression because they may impact both the conditional variance of Y and the conditional density of X. This result is similar to that of Andrews (2005), Robinson (2011), and others.

Remark 4 *A consequence of Proposition 1.3 is that inference results depend on whether* Y *is mean independent of* C *given* X*. To test a null hypothesis, say,* $H_0 : m(x) = m_0(x)$ *against* $H_1 : m(x) \neq m_0(x)$*, the corresponding t statistics is*

$$T_n = \sqrt{nh_n^k} \frac{\left(\widehat{m}(x) - m_0(x)\right)}{\left(\frac{\widehat{\sigma}^2(x)}{\widehat{f}(x)} \int K^2(u)\, du\right)^{1/2}}.$$

The usual two-sided t test with significance level α *rejects the null if* $|T_n| > z_{1-\alpha/2}$*, where* z_α *is the* α *quantile of the standard normal distribution. If* Y *is mean independent of* C *given* X *then* $\Pr\left(|T_n| > z_{1-\alpha/2}\right) \to \alpha$ *as* $n \to \infty$*. Otherwise, we have* $\Pr\left(|T_n| > z_{1-\alpha/2}\right) \to 1$ *as* $n \to \infty$*.*[21]

Remark 5 *The bandwidth can be chosen by minimizing the approximated integrated mean squared error (AMISE) conditional on* $\sigma(C)$*. The bandwidth must be a* $\sigma(C)$*-measurable random variable,*

[20] Adding the idiosyncratic error is without loss as one can specify $U_i = S_i^{2\prime} C^2$, with the last elements of C and S to be $C_{J+1}^2 = 1$ and $S_{iJ+1}^2 = \varepsilon_i$.

[21] In the Supplemental Material we provide conditions under which the kernel density estimator is consistent: $\widehat{f}(x) \xrightarrow{p} f(x|C)$. For the variance $\sigma^2(x, c) = E\left(Y_i^2 | X_i = x, C = c\right) - [m(x, c)]^2$, we can take $\widehat{\sigma}^2(x)$ to be

$$\widehat{\sigma}^2(x) = \left[\frac{\Sigma_{i=1}^n Y_i^2 K\left(\frac{X_i - x}{h}\right)}{\Sigma_{i=1}^n K\left(\frac{X_i - x}{h}\right)}\right] - [\widehat{m}(x)]^2.$$

The first term on the right hand side converges in probability to $E\left(Y_i^2 | X_i = x, C\right)$ using the same arguments as in Proposition 1. The second term on the right hand side converges in probability to $[m(x, C)]^2$ by the Slutsky theorem. Therefore, $\widehat{\sigma}^2(x) \xrightarrow{p} \sigma^2(x, C)$. Next, note that

$$T_n = \sqrt{nh_n^k} \frac{\left(\widehat{m}(x) - m(x, C)\right)}{\left(\frac{\widehat{\sigma}^2(x)}{\widehat{f}(x)} \int K^2(u)\, du\right)^{1/2}} + \sqrt{nh_n^k} \frac{\left(m(x, C) - m_0(x)\right)}{\left(\frac{\widehat{\sigma}^2(x)}{\widehat{f}(x)} \int K^2(u)\, du\right)^{1/2}}.$$

The first term on the RHS converges in distribution to $N(0, 1)$ by Proposition 1.3(ii). The second term on the RHS is such that: (a) $\widehat{f}(x) \geq \xi > 0$, for some finite ξ, with probability approaching one because $f(x|c) \geq \xi > 0$ for Q-almost all c (see the Supplemental Material). If (b) $\sigma^2(x, C)$ is finite Q-almost surely (implying $\widehat{\sigma}^2(x)$ is finite with probability approaching one); and if (c) $m(x, C) \neq m_0(x)$ with positive probability; then the second term on the RHS diverges in probability to $\pm\infty$. As a result, $|T_n| \to \infty$ as $n \to \infty$ under the null.

$h_n(C)$. In the Supplemental Material, we show that $h_n(C) = O_p\left(n^{-\frac{1}{4+k}}\right)$, and one might expect both plug-in and cross-validation estimators to be consistent. The usual concerns in the literature about how to select the bandwidth are present here, but for brevity we do not investigate the topic further. We only emphasize that the bandwidth choice based on the unconditional $AMISE$ is infeasible because it is impossible to estimate the distribution of C (and integrate that out) using a single cross-sectional data set.

4 Conclusion

In this paper, we investigate a nonparametric regression estimator for cross-sectional data in the presence of very general, potentially infinite-dimension, common shocks. In a companion paper (Souza-Rodrigues (2014)), we extend the results to a "large-N, large-T" panel data framework for a nonlinear generalized regression model. We plan to investigate extensions to finite-T panel data models in the future.

A Appendix: Disintegration Theory

We follow the discussion in Pollard (2002, Chapter 5 and Appendix F).[22] Throughout this section, let the measurable space be $(\Omega,\mathcal{F})$ and let C be a measurable map from $(\Omega,\mathcal{F})$ into $(\mathcal{C},\mathcal{B})$. Let λ be a sigma-finite measure on $\mathcal{F}$ and μ be a sigma-finite measure on $\mathcal{B}$. The definition of conditional distributions given in Pollard (2002, p. 113) is:

Definition 1 *Let P be a probability measure on $(\Omega,\mathcal{F})$ and let Q be the probability distribution of C induced by P. A family $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$ of probability measures on $\mathcal{F}$ is called the* ***conditional probability distribution*** *of P given C if*

1. *$P_c\{C \neq c\} = 0$, for Q-almost all $c \in \mathcal{C}$;*

   and for each nonnegative measurable function f on Ω:

2. *the map $c \mapsto \int f(\omega)\, dP_c(\omega)$ is $\mathcal{B}$-measurable; and*

3. *the equality $\int f(\omega)\, dP = \int \left[\int f(\omega)\, dP_c(\omega)\right] dQ(c)$ holds.*

The conditional probability distribution $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$ is a family of probability measures satisfying (i) a concentration property ($P_c\{C \neq c\} = 0$), (ii) a measurability condition (property

[22] Dellacherie and Meyer (1978), Pachl (1978), Hoffmann-Jorgensen (1994, chap. 6 and section 10.11), and Chang and Pollard (1997) are also important references.

2) and (iii) a decomposition property (property 3). Unfortunately, the conditional probability distribution may not exist. The Kolmogorov conditional expectation, on the other hand, always exists. For completeness, we state the definition below (Pollard (2002, p. 126)):

Definition 2 *Let f be a random variable on Ω and P be a probability measure on $(\Omega,\mathcal{F})$. For each sub-sigma-field $\mathcal{G} \subset \mathcal{F}$, the conditional expectation $E(f|\mathcal{G})$ is the random variable defined on $(\Omega,\mathcal{G})$ such that, for all sets $A \in \mathcal{G}$, with indicator functions I_A,*

$$\int \{I_A(\omega) f(\omega)\} dP(\omega) = \int \{I_A(\omega) [E(f|\mathcal{G})(\omega)]\} dP(\omega). \tag{11}$$

The random variable $E(f|\mathcal{G})$ is called the ***conditional expectation of f given the sub-sigma-field $\mathcal{G}$*** *and it is unique up to P-equivalence.*

If the conditional probability distribution of P given C exists, $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$, the $\mathcal{B}$-measurable function defined by

$$E(f|C = c) = \int f(\omega) dP_c(\omega)$$

satisfies the equality (11) for Q-almost all $c \in \mathcal{C}$.

The problem with the Kolmogorov conditional expectation is that each of its usual properties (mainly, being a linear increasing functional of f satisfying the monotone convergence property) holds Q-almost everywhere, but with possible uncountably many negligible sets in which these properties do not hold. The accumulation of these null sets may lead to paradoxes when one is trying to compute the conditional expectation (see, e.g. Rao (1988)). To avoid these difficulties, topological assumptions are invoked to guarantee the existence of conditional probability distributions $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$ such that all the properties of the Kolmogorov conditional expectation are satisfied except in countably many Q-negligible sets. By collecting all these countably many Q-negligible sets into a single Q-null set, we avoid the problems and paradoxes coming from an accumulation of uncountably many null sets. Under the topological assumptions, the family $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$ is a version of the Kolmogorov conditional expectation that does not run into such difficulties and, as a by-product, guarantees the existence of conditional densities. The conditional densities may then be (carefully) manipulated preserving the intuition we have for the cases where the conditioning event has positive probability.

Existence of conditional probability distribution follows from a general decomposition called disintegration. The definition of disintegration given in Chang and Pollard (1997, p. 292) is:

Definition 3 *The measure λ has a disintegration $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$ with respect to C and μ (or a (C, μ)-disintegration) if*

1. *λ_c is a sigma-finite measure on $\mathcal{F}$ concentrated on $\{C = c\}$, that is, $\lambda_c\{C \neq c\} = 0$ for μ-almost all c;*

   and for each nonnegative measurable function f on Ω:

2. *the map $c \mapsto \int f(\omega)\, d\lambda_c(\omega)$ is $\mathcal{B}$-measurable; and*

3. *the equality $\int f(\omega)\, d\lambda = \int \left[\int f(\omega)\, d\lambda_c(\omega)\right] d\mu(c)$ holds.*

From the definitions, it is clear that a (C, μ)-disintegration $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$ can be a conditional probability distribution $\mathcal{P} = \{P_c : c \in \mathcal{C}\}$. Yet, it is useful to let Λ be a collection of sigma-finite measures, so that we can define conditional densities with respect to dominating measures.

Based on this disintegration, we can define a new measure $\mu \otimes \Lambda$ on the product $(\Omega \times \mathcal{C}, \mathcal{F} \otimes \mathcal{B})$, by the iterated integral

$$(\mu \otimes \Lambda)(A) = \int \left[\int I_A d\lambda_c(\omega)\right] d\mu(c)$$

for all sets $A \in \mathcal{F} \otimes \mathcal{B}$. The measure $\mu \otimes \Lambda$ has to be well-defined to satisfy the condition 3 of the Definition 3.

The existence of disintegration is guaranteed by the following theorem (Theorem 6 in Pollard (2002, Appendix F)).

Theorem 1 *(Existence of Disintegration) Let λ be a sigma-finite Radon measure on the Borel sigma-field $\mathcal{F}$ of a metric space Ω. Let μ be a sigma-finite measure on $\mathcal{B}$ that dominates the image measure $\lambda\left(C^{-1}\right)$ (i.e., the measure on $\mathcal{B}$ induced by the map C and the measure λ). If the set*

$$graph(C) \equiv \{(\omega, c) \in \Omega \times \mathcal{C} : C(\omega) = c\}$$

is $\mathcal{F} \otimes \mathcal{B}$ measurable, then λ has a (C, μ)-disintegration, $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$, uniquely determined up to μ-equivalence (i.e., if $\{\lambda_c^* : c \in \mathcal{C}\}$ is another (C, μ)-disintegration then $\mu\{c \in \mathcal{C} : \lambda_c \neq \lambda_c^*\} = 0$).

To guarantee the existence of the (C, μ)-disintegration we need, therefore, to restrict (i) Ω to be a metric space with the Borel sigma-field $\mathcal{F}$, (ii) λ to be a sigma-finite Radon measure, and (iii) the set $graph(C) \equiv \{(\omega, c) \in \Omega \times \mathcal{F} : C(\omega) = c\}$ to be $\mathcal{F} \otimes \mathcal{B}$ measurable. Depending on the problem at hand, it may be reasonable to assume (i) directly. To see the importance of the requirements (ii) and (iii), we briefly describe how the proof works. We then finally discuss the existence of conditional densities.

First, assume that Ω is a compact metric space and let $\mathcal{K}_0$ be a compact paving. A compact paving is a class of compact sets in Ω that is closed under finite unions and intersections. One can show that $\mathcal{K}_0$ is countable when Ω is compact. The proof carefully constructs a finitely additive measure $\lambda_c : \mathcal{K}_0 \to \mathbb{R}^+$, for some $c \in \mathcal{C}$, so that the desired "measure-like" properties of the disintegration (Definition 3) hold for μ-almost all c. Because $\mathcal{K}_0$ is countable, all the desired properties of λ_c hold except on countably many negligible sets, which can be collected into a single negligible set. It is shown, then, that there exists a unique extension of λ_c to a countably additive measure defined on a sigma-field containing $\mathcal{K}_0$ (Pollard (2002, Appendix A)). The extension is (inner) approximated by compact sets. By construction all the desirable properties hold for the extension of λ_c and for all $c \notin N$, where N is a single set with $\mu(N) = 0$. The proof then shows that $c \longmapsto \lambda_c(A)$ is $\mathcal{B}$-measurable for all Borel sets $A \in \mathcal{F}$. Finally, the argument is extended for Ω that is not compact but the measure λ concentrates all the mass on a disjoint union of countably many compact Borel sets. I.e., the measure λ is a sigma-finite Radon measure. Intuitively, the proof explores compact approximations as a way to obtain countable additivity from finite additivity and to collect the negligible sets into a single null set N.

Pachl (1978) shows that a sigma-finite Radon λ (requirement (ii)) is a necessary condition for existence of disintegration. So, even when Ω is not compact (or not separable), λ must have separable support.[23]

The third requirement – the $\mathcal{F}\otimes\mathcal{B}$-measurability of the set $graph(C)$ – is also necessary because the measure

$$(\mu \otimes \Lambda)(A) = \int \left[\int I_A d\lambda_c(\omega) \right] d\mu(c) = \lambda\{\omega \in \Omega : (\omega, C(\omega)) \in A\}$$

is well-defined only if $A \in \mathcal{F} \otimes \mathcal{B}$. The condition is not innocuous: it is well-known that the $graph(C)$ may not be $\mathcal{F} \otimes \mathcal{B}$-measurable even when C is measurable. The $\mathcal{F} \otimes \mathcal{B}$-measurability can be obtained if the σ-field $\mathcal{B}$ is countably generated and contains all the singleton sets $\{c\}$ (see Pollard 2002, p. 344). In particular, if $\mathcal{B}$ is the Borel σ-field on the separable metric space $\mathcal{C}$, these conditions are satisfied (see Pollard 2002, p. 103).

A separable $\mathcal{C}$ with the Borel σ-field $\mathcal{B}$ is sufficient, but not necessary, for the $\mathcal{F}\otimes\mathcal{B}$-measurability of the $graph(C)$. It is possible but not trivial to obtain such a result for non-separable spaces. Hansell (1988) provides very abstract (and somewhat difficult to interpret) sufficient conditions for the $\mathcal{F} \otimes \mathcal{B}$-measurability when $\mathcal{C}$ is not separable. Yet, even if the $\mathcal{F} \otimes \mathcal{B}$-measurability holds for

[23] Formally, the necessary condition is that λ must be approximated by a compact paving that is closed under countable unions.

a non-separable $\mathcal{C}$, the Radon measure λ puts all mass on a separable subset of $\mathcal{C}$. To see why, let G be a countable union of compact sets on Ω such that $\lambda(G^c) = 0$, where G^c is the complement of G. The map $g : \Omega \to \Omega \times \mathcal{C}$ defined by $g(\omega) = (\omega, C(\omega))$ is such that λ concentrates all mass in the set $g(G)$. If C is Borel measurable, the set $g(G) \subset \Omega \times \mathcal{C}$ is separable, and so is $C(G)$ (Bogachev (2007, Corollary 6.10.17)). The image measure of C under λ therefore puts all mass on a separable subset of $\mathcal{C}$ when $\mathcal{C}$ is non-separable. So, although $\mathcal{C}$ does not have to be separable to obtain existence of disintegration, it seems difficult to get away from separability in this context.

The next theorem provides the conditions under which conditional densities exist (Theorem 12 in Pollard 2002, Chapter 5).

Theorem 2 *(Conditional Densities) Let P be a probability measure on $(\Omega,\mathcal{F})$ with density $f(\omega)$ with respect to the sigma-finite measure λ. Let λ have a (C,μ)-disintegration $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$. Then*

1. *The image measure $Q = P\left(C^{-1}\right)$ (i.e., the probability distribution of C induced by P) is absolutely continuous with respect to μ, with density $q(c) \equiv \int f(\omega)\, d\lambda_c(\omega)$.*

2. *The set $\{(\omega, c) \in \Omega \times \mathcal{C} : q(c) = \infty \text{ or } q(c) = 0\}$ has zero $\mu \otimes \Lambda$ measure.*

3. *The probability measure P has conditional distribution $\{P_c : c \in \mathcal{C}\}$ given C, where P_c is defined by having density*

$$f(\omega|c) \equiv \frac{f(\omega)}{q(c)} \{0 < q(c) < \infty\} \tag{12}$$

with respect to λ_c, for Q-almost all $c \in \mathcal{C}$.

The formula in (12) is the general version of the conditional density as the ratio of the joint density to the marginal density, but not requiring C to belong to a Euclidean space. To guarantee existence of the conditional density, we therefore need the existence of the (C,μ)-disintegration $\Lambda = \{\lambda_c : c \in \mathcal{C}\}$. For a more detailed discussion, see Dellacherie and Meyer (1978), Hoffmann-Jorgensen (1994), Chang and Pollard (1997), and Pollard (2002).

References

[1] Altonji, J., T. Conley, T. E. Elder, and C. R. Taber (2010). "Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables." Yale University.

[2] Anselin, L. (1988). *Spatial Econometric Methods and Models*. Boston: Kluwer.

[3] Andrews, D.W.K. (2005). "Cross-section regression with common shocks." *Econometrica* 73, No. 5, 1551-1585.

[4] Andrews, D.W.K., organizer. (2008). "Handling Dependence: Temporal, Cross-Sectional and Spatial." *Cowles Summer Conference*, June 22-23, 2009, Yale University. http://cowles.yale.edu/handling-dependence-temporal-cross-sectional-and-spatial

[5] Arbia, G. (2006). *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence.* Springer-Verlag, Berlin.

[6] Arbia, G. (2016). "Spatial Econometrics: A Rapidly Evolving Discipline." *Econometrics*, 4(1), 18.

[7] Baltagi, B.H, H.H. Kelejian, and I.R. Prucha, eds. (2007). "Analysis of Spatially Dependent Data" *Journal of Econometrics*(special edition), 140(1), 1-332.

[8] Berry, S. T. and P. A. Haile (2009). "Identification of a nonparametric generalized regression model with group effects." Discussion paper, Yale University.

[9] Bai, J. (2009). "Panel data models with interactive fixed effects." *Econometrica*, 77, 1229-1279.

[10] Bai, J. and S. Ng. (2006). "Evaluating latent and observed factors in macroeconomics and finance." *Journal of Econometrics* 131, 507-537.

[11] Bester, C. A., T. G. Conley, C. B. Hansen, and T. J. Vogelsang (2016). "Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators." *Econometric Theory*, 32: 154-186.

[12] Bogachev, V. I. (2007). *Measure Theory.* Springer-Verlag.

[13] Chang, J. T. and D. Pollard (1997). "Conditioning as disintegration." *Statistica Neerlandica* 51, No. 3, 287-317.

[14] Chudik, A. and M. H. Pesaran (2014). "Large panel data models with cross-sectional dependence: a survey." Forthcoming in B. Baltagi (Ed.), *The Oxford Handbook on Panel Data.* Oxford University Press.

[15] Chudik, A. and M. H. Pesaran (2015). "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors." *Journal of Econometrics*, 188(2), 393-420.

[16] Conley, T. G. (1999). "GMM estimation with cross-sectional dependence." *Journal of Econometrics* 92, 1-45.

[17] Dellacherie, C. and P. A. Meyer (1978). *Probabilities and Potential*. North-Holland, Amsterdam.

[18] Doob, J. L. (1953). *Stochastic Processes*. Wiley. New York.

[19] Eberhardt, M., C. Helmers, and H. Strauss (2013) "Do spillovers matter when estimating private returns to R&D?" *Review of Economics and Statistics*, 95, 436-448.

[20] Evans, D., A. Tandon, C. Murray, and J. Lauer (2000). "The comparative efficiency of national health systems in producing health: An analysis of 191 countries." GPE Discussion Paper No. 29; World Health Organization: Geneva, Switzerland.

[21] Finks, J. F., N. H. Osborne, and J. D. Birkmeyer (2011). "Trends in hospital volume and operative mortality for high-risk surgery." *The New England Journal of Medicine*, 364(22):2128-37. doi: 10.1056/NEJMsa1010705.

[22] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications.* Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts.

[23] Forchini, G. and B. Peng (2016). "A conditional approach to panel data models with common shocks," *Econometrics*, 4(1), 4.

[24] Gerolimetto, M. and S. Magrini (2009). "Nonparametric Regression with Spatially Dependent Data." Dipartimento di Scienze Economiche, Università Ca' Foscari Venezia.

[25] Gaibulloev, K., T. Sandler, and D. Sul (2013). "Common drivers of transnational terrorism: Principal component analysis." *Economic Inquiry*, 51:707-21.

[26] Halmos, P.R. (1950). *Measure Theory*. Van Nostrand, New York. (July 1969 reprinting).

[27] Hansell, R. W. (1988). "Sums, products and continuity of Borel maps in nonseparable metric spaces." *Proceedings of the American Mathematical Society* 104, No. 2, 465-471.

[28] Hoffmann-Jorgensen, J. (1994). *Probability with a View Towards Statistics*. Vol 2, Chapman and Hall, New York.

[29] Kelejian, H. H. and I. R. Prucha (1999). "A generalized moments estimator for the autoregressive parameter in a spatial model." *International Economic Review* 40, 509-533.

[30] Kuersteiner, G.M. and I. R. Prucha (2013). "Limit theory for panel data models with cross sectional dependence and sequential exogeneity." *Journal of Econometrics*, 174, 107-126.

[31] Martins-Filho, C. and F. Yao (2009). "Nonparametric regression estimation with general parametric error covariance." *Journal of Multivariate Analysis*, 100, 309-333.

[32] Moon, H. and Weidner, M. (2009). "Likelihood expansion for panel regression models with factors." Manuscript.

[33] Lee, L.F. (2002). "Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models." *Econometric Theory* 18, 252-277.

[34] Lee, L. and J. Yu. (2010). "Some recent developments in spatial panel data models." *Regional Science and Urban Economics*, 40: 255-271.

[35] Lee, L. and J. Yu (2014). "Efficient GMM estimation of spatial dynamic panel data models with fixed effects." *Journal of Econometrics*, 180: 174-197.

[36] Pachl, J. (1978). "Disintegration and compact measures." *Mathematica Scandinavica* 43, 157-168.

[37] Pagan, A.R. and A. Ullah (1999). *Nonparametric Econometrics*, Cambridge University Press.

[38] Pesaran, M. H. (2006). "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica* 74(4), 967-1012.

[39] Pesaran, M. H. and E. Tosetti (2011). "Large panels with common factors and spatial correlation." *Journal of Econometrics* 161 (2), 182-202.

[40] Phillips, P. C. B. and D. Sul (2003). "Dynamic panel estimation and homogeneity testing under cross section dependence." *Econometrics Journal* 6(1), 217-259.

[41] Phillips, P. C. B. and D. Sul (2007). "Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence." *Journal of Econometrics* 137(1), 162-188.

[42] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[43] Rao, M. M. (1988). "Paradoxes in conditional expectation." *Journal of Multivariate Analysis* 27, Issue 2, 434-446.

[44] Robinson, P. M. (2011). "Asymptotic theory for nonparametric regression with spatial data." *Journal of Econometrics* 165(1), 5-19.

[45] Robinson, P. M. and J. Lee (2016). "Series estimation under cross-sectional dependence." *Journal of Econometrics* 190(1), 1-17.

[46] Sarafidis, V. and T. Wansbeek. (2012). "Cross-sectional dependence in panel data analysis." *Econometric Review*, 31, 483-531.

[47] Souza-Rodrigues, E. A. (2014). "Nonparametric estimation of generalized regression model with group effects," Mimeo.

[48] Souza-Rodrigues, E. A. (2016). "Supplement to Nonparametric Regression with Common Shocks."

[49] Su, L. and Z. Yang (2015). "QML estimation of dynamic panel data models with spatial errors." *Journal of Econometrics*, 185: 230-258.

[50] Wang, Q. and P. C. B. Phillips (2009). "Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegration Regression," *Econometric Theory*, 25, 710-738.

[51] Zaffaroni, P. (2010). "Generalized least squares estimation of panel with common shocks," Manuscript.