The Stata Journal (2023)
23, Number 1, pp. 243–264 DOI: 10.1177/1536867X231162035

Uniform nonparametric inference for spatially dependent panel data: The xtnpsreg command

Jia Li
Singapore Management University
Singapore
jiali@smu.edu.sg

Zhipeng Liao
UCLA
Los Angeles, CA
zhipeng.liao@econ.ucla.edu

Wenyu Zhou
Zhejiang University
Hangzhou, China
wenyuzhou@intl.zju.edu.cn

Abstract. In this article, we introduce a command, `xtnpsreg`, that implements a uniform nonparametric inference procedure for possibly unbalanced panel datasets with general forms of spatiotemporal dependence. We demonstrate how to apply this command via several examples, including 1) the nonparametric estimation of the conditional mean function and its marginal response, 2) the construction of uniform confidence bands for these nonparametric functional parameters, 3) specification tests for parametric model restrictions, and 4) the estimation and uniform inference for functional coefficients in semi-nonparametric models.

Keywords: st0707, xtnpsreg, series estimation, spatiotemporal dependence, panel data, nonparametric inference, uniform confidence band

1 Introduction

Nonparametric regression methods provide a flexible way to study the relationship between variables. A popular approach is the series regression, which allows the user to approximate the unknown function with a "large" set of basis functions such as polynomials, splines, wavelets, etc. Conventional econometric theory (see, for example, Andrews [1991] and Newey [1997]) allows one to conduct pointwise inference that is specific to the function's value at a given point. This may be unsatisfactory in practice because applied researchers are often interested in making inferential statements on the conditional mean function as a whole. This more demanding task requires uniform inference methods such as those developed by Belloni et al. (2015) and Li and Liao (2020), respectively, for independent and identically distributed (i.i.d.) and serially dependent time-series data.

Meanwhile, panel datasets are widely used in various areas of empirical research. It is clearly of applied interest to conduct the aforementioned functional inference in the panel-data setting. An immediate benefit is that, by harnessing the richer information from both cross-sectional and time-series dimensions, one may obtain more accurate nonparametric estimates and draw sharper inference. This is a relevant consideration because the practical application of nonparametric methods is often hindered by a small sample size.

Much care is needed for performing reliable inference for panels because these types of data often exhibit spatiotemporal dependence; namely, the observations may be mutually dependent on both cross-sectional and time-series dimensions, which has been emphasized by Bertrand, Duflo, and Mullainathan (2004) and Petersen (2009), among others. Not accounting for such dependence tends to result in an understatement of the sampling variability, leading the empiricist to mistakenly interpret "noise" as "signal".

A popular approach for dealing with spatiotemporal dependence is proposed by Driscoll and Kraay (1998) in the context of generalized method of moments. Driscoll–Kraay standard errors are robust to general forms of weak dependence in the time-series dimension and arbitrarily strong spatial dependence in the cross-sectional dimension. The underlying econometric theory requires "large T" asymptotics but does not restrict the dimensionality of the cross-section. In Stata, `xtscc` implements Driscoll–Kraay standard errors for linear panel regressions (Hoechle 2007). Note that, in the degenerate case where the "panel" contains only a single time series, the Driscoll–Kraay standard error coincides with the classical Newey–West standard error (Newey and West 1987); see [TS] **newey**.

In this article, we propose a new command, `xtnpsreg`, that implements a panel (`xt`) nonparametric (`np`) series regression (`sreg`) and provides valid uniform functional inference that is robust to general forms of spatiotemporal dependence as considered in Driscoll and Kraay (1998). The theoretical validity of the implemented method can be justified by directly invoking the general theory developed by Li and Liao (2020) for growing-dimensional mixingale processes; see the companion article by Li, Liao, and Zhou (2021) for further technical details. The `xtnpsreg` command may be regarded as the nonparametric and functional version of `xtscc`. It is also related to the `tssreg` command developed by Li, Liao, and Gao (2020), which performs a similar task under the time-series setting. Roughly speaking, `xtnpsreg` extends `tssreg` in the same way as `xtscc` extends `newey`.

As we shall demonstrate in detail below, `xtnpsreg` may be conveniently used to perform several types of nonparametric inferential tasks, including 1) the nonparametric estimation of a conditional mean function and its marginal response (that is, the derivative function), 2) the construction of uniform confidence bands (CBs) for these functional parameters, 3) nonparametric specification tests for parametric model restrictions, and 4) the estimation and uniform inference for functional coefficients in semi-nonparametric models.

The remainder of this article is organized as follows. Section 2 provides some heuristics on the econometric and statistical theory underlying the proposed procedure. Section 3 documents the functionalities of the `xtnpsreg` command. Section 4 demonstrates how to use `xtnpsreg` to accomplish various nonparametric inferential tasks in an empirical example using data from the Federal Reserve Bank of Philadelphia Survey of Professional Forecasters.

2 Heuristics for the econometric procedure

In this section, we describe the econometric setting for the nonparametric regression problem and provide some heuristics for the uniform functional inference procedure. For simplicity, we focus on the case with balanced panels in this discussion while noting that unbalanced panels are accommodated by `xtnpsreg` as well.

2.1 Uniform functional inference for the conditional mean function

The baseline setting for the `xtnpsreg` command is the nonparametric panel regression model

$$y_{it} = g(x_{it}) + \epsilon_{it}, \quad \mathbb{E}(\epsilon_{it}|x_{it}) = 0 \tag{1}$$

for $1 \leq i \leq N$ and $1 \leq t \leq T$, where x_{it} is a continuous variable. We assume that $T \to \infty$ but do not impose any restriction on N; that is, N may be fixed or grow to infinity. The object of interest is the conditional mean function $g(\cdot)$, which is implicitly defined as $g(x) \equiv \mathbb{E}(y_{it}|x_{it} = x)$. We focus on a single-equation setting, with the dependent variable (*depvar*) y_{it} being scalar valued. In the current version of `xtnpsreg`, we also require the conditioning variable (*condvar*) x_{it} to be univariate for two reasons. First, although multivariate conditioning is permitted in theory, the resulting nonparametric estimate tends to be imprecise because of the well-known "curse of dimensionality" in nonparametric analysis. Thus, it is advisable to single out a key explanatory variable x_{it} and allow it to enter the model nonparametrically. Second, the univariate conditioning also greatly simplifies the graphical presentation of the functional estimate and its CB, which is desirable for empirical discussions.

The main output of `xtnpsreg` consists of a nonparametric estimate for the conditional mean function $g(\cdot)$ and an associated uniform CB at a user-specified confidence level $1 - \alpha$. To be precise, the uniform CB is given by a pair of functional estimates $[L(\cdot), U(\cdot)]$ such that

$$\mathbb{P}\{L(x) \leq g(x) \leq U(x) \text{ for all } x \in \mathcal{X}\} \to 1 - \alpha, \quad \text{as} \quad T \to \infty$$

In other words, the uniform CB covers the true conditional mean function simultaneously over the entire region $\mathcal{X}$ with approximately $1 - \alpha$ probability in large samples. By default, $\mathcal{X}$ is set to be the observed support of the conditioning variable. In certain applications, the user may want to take $\mathcal{X}$ as a subset of the observed support (so that the uniform nonparametric inference concentrates on a particular subregion of the conditioning space), which is allowed as an option.

The implementation of the statistical procedure proceeds as follows:

- **Step 1 (nonparametric estimation).** The nonparametric estimator for $g(\cdot)$ is constructed by running a series regression. Let $\mathbf{p}(x_{it}) = (p_1(x_{it}), \ldots, p_m(x_{it}))^\top$ denote an m-dimensional vector of approximating functions of x_{it}. Regressing y_{it} on $\mathbf{p}(x_{it})$ yields the regression coefficient

$$\widehat{\mathbf{b}} = \left\{ \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbf{p}(x_{it})\mathbf{p}(x_{it})^\top \right\}^{-1} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbf{p}(x_{it}) y_{it} \right\}$$

and the resulting nonparametric estimator for $g(\cdot)$ is given by

$$\widehat{g}(\cdot) = \mathbf{p}(\cdot)^\top \widehat{\mathbf{b}}$$

The current version of `xtnpsreg` uses Legendre polynomials to form the approximating functions $\mathbf{p}(x_{it})$. Recall that the kth-order Legendre polynomial is given by

$$\frac{1}{2^k k!} \frac{d^k}{dx^k} (x^2 - 1)^k$$

An important property of the Legendre polynomials is that they are orthogonal on the $[-1, 1]$ interval with respect to the uniform distribution. This orthogonality property helps mitigate the multicollinearity among series terms and hence improves the numerical stability of the estimation procedure. Other types of orthogonal series basis may also be adopted to serve the same purpose, and it might be interesting to incorporate them in a future version of `xtnpsreg`.

To better exploit the orthogonality property of Legendre polynomials, one should perform a preliminary transformation on the conditioning variable x_{it} to make it approximately uniformly distributed on the $[-1, 1]$ interval. One way to achieve this is to consider some cumulative distribution function (CDF), say, $F(\cdot)$, and transform x_{it} via $x \mapsto 2F(x) - 1$. If $F(\cdot)$ is the CDF of x_{it}, the transformed variable will be exactly uniformly distributed on $[-1, 1]$. In practice, setting $F(\cdot)$ as any reasonable approximation for the CDF of x_{it} can still achieve this goal to some extent, which will generally improve the numerical stability. By default, `xtnpsreg` uses the CDF of a normal distribution (calibrated to data) to carry out the transformation, which is an adequate choice provided that the distribution of the conditioning variable x_{it} roughly mimics a normal distribution. This default transformation may be disabled via the `method(none)` option, which allows users to customize the transformation of the conditioning variable onto the $[-1, 1]$ interval on their own.

- **Step 2 (critical value).** The second step is to compute a critical value for a "functional t statistic" that is defined as

$$\widehat{\tau} = \sup_{x \in \mathcal{X}} \frac{T^{1/2} |\widehat{g}(x) - g(x)|}{\widehat{\sigma}(x)}$$

where $\widehat{\sigma}(x)$ is the estimated standard error for $\widehat{g}(x)$. Note that $\widehat{\tau}$ is simply the supremum of the absolute value of the pointwise t statistics evaluated at different points over the conditioning space $\mathcal{X}$. The $\widehat{\sigma}(x)$ estimate is computed as

$$\widehat{\sigma}(x) = \sqrt{\mathbf{p}(x)^\top \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{A}} \widehat{\mathbf{Q}}^{-1} \mathbf{p}(x)}$$

where

$$\widehat{\mathbf{Q}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{p}(x_{it})\mathbf{p}(x_{it})^{\top}$$

and $\widehat{\mathbf{A}}$ is a "clustered" Newey–West estimator for the long-run variance–covariance matrix for the score vector $\mathbf{p}(x_{it})\epsilon_{it}$ with the form [denoting $\widehat{\epsilon}_{it} = y_{it} - \widehat{g}(x_{it})$]

$$\begin{cases} \widehat{\mathbf{A}} = \displaystyle\sum_{s=-M_n}^{M_n} \frac{|M_n+1-s|}{M_n+1}\widehat{\mathbf{\Gamma}}_s, \text{ where} \\ \widehat{\mathbf{\Gamma}}_s = \displaystyle\frac{1}{T}\sum_{t=\max\{1,1-s\}}^{\min\{T-s,T\}} \left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{p}(x_{it})\widehat{\epsilon}_{it}\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{p}(x_{it+s})\widehat{\epsilon}_{it+s}\right\}^{\top} \end{cases}$$

The user may specify the bandwidth parameter M_n via the `lag()` option in `xtnpsreg` as in `newey` (see [TS] **newey**). We also note that the $\widehat{\mathbf{A}}$ estimator is constructed in the same spirit as Driscoll and Kraay (1998), and it is robust with respect to general forms of spatiotemporal dependence.

The critical value of interest is an estimate for the $1-\alpha$ quantile of the sup-t statistic $\widehat{\tau}$. Applying the theory of Li and Liao (2020), Li, Liao, and Zhou (2021) show that the distribution of $\widehat{\tau}$ can be approximated in large samples by the conditional (given data) distribution of

$$\widehat{\tau}^* = \sup_{x\in\mathcal{X}} \frac{\left|\mathbf{p}(x)^{\top}\left(\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{A}}\widehat{\mathbf{Q}}^{-1}\right)^{1/2}\mathbf{N}_m^*\right|}{\widehat{\sigma}(x)}$$

where $\mathbf{N}_m^*$ is a generic m-dimensional standard normal random vector [that has the same dimensionality as $\mathbf{p}(x_{it})$]. To compute the critical value, we thus draw $\mathbf{N}_m^*$ from the standard normal distribution many times, and for each draw we compute $\widehat{\tau}^*$ over a discretized mesh of $\mathcal{X}$; we then set the critical value $cv_{1-\alpha}$ as the $1-\alpha$ empirical quantile of the simulated $\widehat{\tau}^*$.

In empirical applications, researchers are often interested in testing whether the conditioning variable may have any effect on the dependent variable, which in the present nonparametric setting amounts to testing the null hypothesis

$$H_0 : g(x) \equiv E(y_{it}|x_{it} = x) = 0, \qquad \text{for all } x \in \mathcal{X}$$

We reject the null hypothesis at significance level α if the sup-t statistic $\widehat{\tau}$ [evaluated at $g(\cdot) = 0$] exceeds the critical value $cv_{1-\alpha}$. The test statistic, critical value, and corresponding p-value are the default output of `xtnpsreg`.

- **Step 3 (uniform CB).** Finally, the $1-\alpha$ level two-sided uniform CB for $g(\cdot)$ is then given by

$$\text{CB}_{1-\alpha}(\cdot) = \left[\widehat{g}(\cdot) - cv_{1-\alpha}T^{-1/2}\widehat{\sigma}(\cdot),\ \widehat{g}(\cdot) + cv_{1-\alpha}T^{-1/2}\widehat{\sigma}(\cdot)\right]$$

The nonparametric functional estimate, together with this CB, can be displayed by activating the `plot` option; also see the `plotu` option for an alternative.

2.2 Marginal response

In linear regression models, the marginal effect of an explanatory variable on the dependent variable is completely summarized by its regression coefficient. For nonparametric regressions, the marginal response is captured by the derivative of the conditional mean function, denoted $\partial g(\cdot)$. By calling the `marginal` option, `xtnpsreg` computes a nonparametric functional estimate for $\partial g(\cdot)$ along with its $1-\alpha$ uniform CB. The implementation is carried out in the background as follows:

- **Step 1**. Compute $\widehat{\mathbf{b}}$, $\widehat{\mathbf{Q}}$, and $\widehat{\mathbf{A}}$ as in section 2.1. Set

$$\partial\widehat{g}(\cdot) = \partial\mathbf{p}(\cdot)^\top\widehat{\mathbf{b}}, \qquad \widetilde{\sigma}(\cdot) = \sqrt{\partial\mathbf{p}(\cdot)^\top\left(\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{A}}\widehat{\mathbf{Q}}^{-1}\right)\partial\mathbf{p}(\cdot)}$$

- **Step 2**. Draw $\mathbf{N}_m^*$ from the m-dimensional standard normal distribution many times, and for each draw, compute

$$\widehat{\tau}'^* = \sup_{x\in\mathcal{X}} \frac{|\partial\mathbf{p}(x)^\top\left(\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{A}}\widehat{\mathbf{Q}}^{-1}\right)^{1/2}\mathbf{N}_m^*|}{\widetilde{\sigma}(x)}$$

  where the supremum can be computed on a discretized mesh of $\mathcal{X}$. Set the critical value $cv'_{1-\alpha}$ as the $1-\alpha$ empirical quantile of the simulated $\widehat{\tau}'^*$.

- **Step 3**. Report the $1-\alpha$ level two-sided uniform CB for $\partial g(\cdot)$ as $[\partial\widehat{g}(\cdot) - cv'_{1-\alpha}T^{-1/2}\widetilde{\sigma}(\cdot),\ \partial\widehat{g}(\cdot) + cv'_{1-\alpha}T^{-1/2}\widetilde{\sigma}(\cdot)]$.

2.3 Functional coefficient model

The aforementioned nonparametric uniform inference method can also be adapted to study linear regression models with functional coefficients. Specifically, consider the specification

$$y_{it} = c + \beta(x_{it})u_{it} + \epsilon_{it}, \quad \mathbb{E}(\epsilon_{it}|x_{it}, u_{it}) = 0 \tag{2}$$

where c is the intercept, u_{it} is a scalar-valued "base" explanatory variable (*basevar*), and $\beta(\cdot)$ is its functional coefficient modeled nonparametrically as a function of the conditioning variable x_{it}. The inferential target is the function $\beta(\cdot)$.

Note that the baseline setting described in section 2.1 may be considered a special case of (2) with $u_{it} = 1$ and $g(x) = c + \beta(x)$. The user may signify this more general functional coefficient setting by specifying the `funccoef` option in `xtnpsreg`. In this situation, the functional inference will concentrate on the functional coefficient $\beta(\cdot)$ without adding the intercept term c.

2.4 Semi-nonparametric model with linear control variables

For the settings discussed above, `xtnpsreg` also allows additional control variables to enter the model linearly. The generalized versions of the baseline nonparametric regression model (1) and the functional coefficient model (2) are given by, respectively,

$$y_{it} = g(x_{it}) + \mathbf{z}_{it}^{\top}\boldsymbol{\gamma} + \epsilon_{it} \tag{3}$$

$$y_{it} = c + \beta(x_{it})u_{it} + \mathbf{z}_{it}^{\top}\boldsymbol{\gamma} + \epsilon_{it} \tag{4}$$

where $\mathbf{z}_{it}$ is a vector of control variables (*controlvar*). Under these settings, the focal point of the functional inference remains $g(\cdot)$ and $\beta(\cdot)$, respectively. As such, `xtnpsreg` complements the existing Stata commands for semiparametric models that mainly focus on pointwise inference; see Verardi (2013) for more details.

2.5 Semi-nonparametric model with individual fixed effects

In many empirical applications, it is often desirable to include individual fixed effects to account for unobserved heterogeneity. A limited form of fixed effect can be implemented by `xtnpsreg`, as we now explain. Consider the specifications

$$y_{it} = g(x_{it}) + \gamma_i + \epsilon_{it}$$
$$y_{it} = c + \beta(x_{it})u_{it} + \gamma_i + \epsilon_{it}$$

where γ_i is the individual fixed effect. If N is "small" (for example, fixed), we may treat these two specifications as special cases of the models (3) and (4), respectively, by setting $\mathbf{z}_{it}$ as a vector of dummy variables for all i's excluding a baseline group. The inference procedure described in section 2.4 can then be used to handle fixed effects. More generally, it is theoretically possible to allow $N \to \infty$ "very slowly" as $T \to \infty$. How to conduct uniform nonparametric inference for the current panel-data setting with many fixed effects remains an open theoretical question. This is beyond the scope of `xtnpsreg` but is an important direction for future development.

3 The xtnpsreg command

This section documents the syntax and functionalities of the `xtnpsreg` command. The command requires the `moremata` package (Jann 2005), which may be installed in command line via `ssc install moremata`. The user must also declare the panel-data structure beforehand via `xtset` *panelvar timevar*.

3.1 Syntax

The syntax of the `xtnpsreg` command is as follows,

`xtnpsreg` *depvar* *condvar* [*basevar*] [*controlvar*] [*if*] [*in*] [, `lag(`#`)` `m(`#`)` `method(`*transtype*`)` `confidencelevel(`#`)` `ngrid(`#`)` `mc(`#`)` `triml(`#`)` `trimr(`#`)` `plot` `plotu` `scatter(`#`)` `funccoef` `marginal` `table` `excel`]

where *depvar* corresponds to the dependent variable y_{it}, *condvar* denotes the nonparametric conditioning variable x_{it}, *basevar* is the univariate "base" explanatory variable u_{it} in the functional coefficient model described in section 2.3, and *controlvar* contains a vector $\mathbf{z}_{it}$ of linear control variables described in section 2.4.

3.2 Options

`lag(`#`)` specifies the number of lags for computing the Newey–West estimator of the long-run variance–covariance matrix. The default is given by the integer part of $0.75T^{1/3}$, where T is the number of time periods.

`m(`#`)` specifies the number of Legendre polynomial terms used in the nonparametric series estimation. The default is `m(6)`.

`method(`*transtype*`)` specifies the transformation implemented on the conditioning variable. The main purpose of doing so is to make the regressors approximately orthogonal, which generally improves the numerical stability of the series regression, especially when many series terms are included. The approximating functions are Legendre polynomials of the transformed variable. The current version supports the following transformation methods, with `method(normal)` set to be the default.

`none`: no transformation.

`normal`: normal transformation $x \mapsto 2\Phi\{(x - \overline{x})/\sigma\} - 1$, where $\overline{x}$ and σ are the sample mean and standard deviation of x, respectively, and Φ is the CDF of the standard normal distribution.

`confidencelevel(`#`)` specifies the confidence level (in percentage) of the uniform CB. The default is `confidencelevel(90)`.

`ngrid(`#`)` specifies the number of grid points used for discretizing the support of the transformed conditioning variable. The default is `ngrid(1000)`.

`mc(`#`)` specifies the number of Monte Carlo simulations used to compute the critical value. The default is `mc(5000)`.

`triml(`#`)` sets the left limit of the conditioning region $\mathcal{X}$ to be the # empirical quantile of *condvar*. The default is `triml(0)`.

`trimr(#)` sets the right limit of the conditioning region $\mathcal{X}$ to be the $1 - \#$ empirical quantile of *condvar*. The default is `trimr(0)`.

`plot` produces a plot of the nonparametric functional estimate and its uniform CB, in which the transformed conditioning variable is plotted on the horizontal axis.

`plotu` produces a plot of the nonparametric functional estimate and its uniform CB, in which the original conditioning variable is plotted on the horizontal axis.

`scatter(#)` adds a scatterplot of the data points, where # is a number between $[0, 100]$ that specifies the fraction of randomly selected data points to be plotted.

`funccoef` signifies that the model of interest is a functional coefficient model with *basevar* as the base explanatory variable.

`marginal` implements the estimation of the marginal response function.

`table` reports the estimated regression coefficients and standard errors in the series estimation.

`excel` generates an Excel file that contains the requisite information for plotting the functional estimate and the associated uniform CB.

3.3 Stored results

The `xtnpsreg` command stores the following results in `e()`:

Scalars	
`e(N)`	number of cross-sectional units
`e(T)`	number of time periods
`e(supt)`	sup-t statistic
`e(cv)`	critical value for the sup-t test
`e(df_r)`	residual degrees of freedom
Macros	
`e(cmd)`	`xtnpsreg`
`e(depvar)`	name of the dependent variable
`e(condvar)`	name of the conditioning variable
`e(method)`	transformation method
Matrices	
`e(b)`	regression coefficients in series estimation
`e(V)`	variance–covariance matrix of the regression coefficients
`e(se)`	standard errors of regression coefficients in series estimation
`e(ygrid)`	functional estimate
`e(xgrid)`	grid points of the conditioning variable
`e(sigma)`	estimate of standard error function

4 An empirical illustration

In this section, we demonstrate `xtnpsreg`'s main usage in an example built on the empirical analysis of Coibion and Gorodnichenko (2015) and Li, Liao, and Zhou (2021). The dataset and implementation code are provided in the online supplement accompanying this article.

4.1 Data description and empirical motivation

spf.dta is constructed from the Survey of Professional Forecasters. It contains quarterly time series of ex post forecast errors (fe) and ex ante forecast revisions (fr) averaged among forecasters from 1969 to 2014 for five macroeconomic variables—gross domestic product price deflator, real gross domestic product, industrial production, housing starts, and unemployment rate—over four forecast horizons. We treat the data for each variable-horizon pair as individual time series. Merging them yields a panel dataset with $N = 20$ and $T = 173$.

Because the time series of forecast errors and forecast revisions have quite different scales across variables, we first normalize them separately so that each time series is averaged at zero with unit standard deviation. The resulting normalized forecast error and forecast revision are stored as fe_norm and fr_norm, respectively. The normalization is implemented as follows:

```
. *** Load data ***
. set seed 12345678
. use spf
. xtset id_N id_date
Panel variable: id_N (strongly balanced)
 Time variable: id_date, 1 to 173
         Delta: 1 unit
. *** Data Normalization ***
. by id_N: egen fe_id_mean = mean(fe)
. by id_N: egen fe_id_sd = sd(fe)
. by id_N: generate fe_norm = (fe-fe_id_mean)/fe_id_sd
. by id_N: egen fr_id_mean = mean(fr)
. by id_N: egen fr_id_sd = sd(fr)
. by id_N: generate fr_norm  = (fr-fr_id_mean)/fr_id_sd
```

The economic motivation for studying the relationship between ex post forecast error and ex ante forecast revision is to examine whether the professional forecasters are collectively rational. Under the rational-expectation hypothesis, the forecast errors are entirely unanticipated, so their conditional expectation given any a priori known information (including the forecast revision) should be zero. Meanwhile, a large literature in macroeconomics argues that the full rationality benchmark may break down because of information stickiness, which in turn implies a positive relationship between forecast error and forecast revision, as shown in Coibion and Gorodnichenko (2015). We may assess the empirical plausibility of these alternative theoretical predictions by nonparametrically regressing the forecast error on the forecast revision. Because macroeconomic time series tend to comove, their associated forecast error and forecast revision series are likely correlated on the cross-section, exhibiting spatial dependence. In addition, because forecasts over multiple horizons are involved, one must account for serial dependence as well. These considerations naturally motivate the adoption of an inference procedure that is robust to spatiotemporal dependence. Below, we demonstrate how to use the xtnpsreg command to implement the nonparametric estimation and the related functional inference in this empirical setting.

4.2 The basic function of xtnpsreg

The main and most basic use of `xtnpsreg` is to nonparametrically estimate the conditional mean function and plot the functional estimate together with its uniform CB. As a first illustration, we nonparametrically regress the (normalized) forecast error `fe_norm` on the forecast revision `fr_norm` as follows:

```
. xtnpsreg fe_norm fr_norm, plot
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.2721	2.5230	0.000

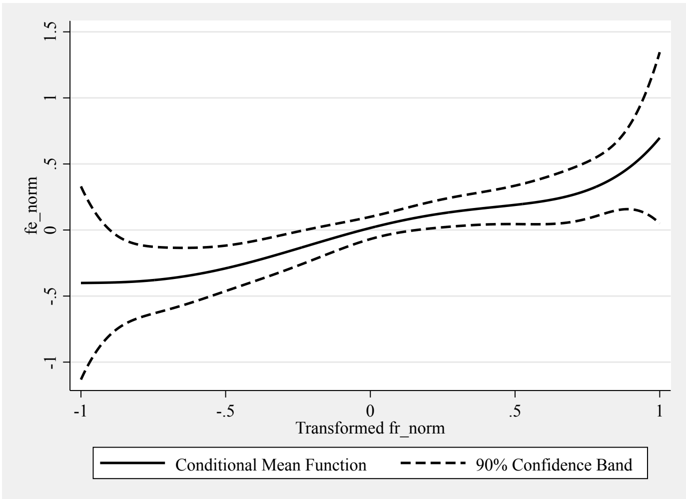


Figure 1. Default output of estimated conditional mean function and the 90% uniform CB

The output table reports the sup-t statistic, critical value, and p-value for testing the null hypothesis that the conditional mean function is identically zero uniformly. The default significance level is $\alpha = 10\%$, which may be changed via the `confidencelevel()` option (for example, `confidencelevel(95)` corresponds to $\alpha = 5\%$). The table above shows that the sup-t statistic is notably greater than the critical value, indicating a strong rejection of the null hypothesis. Indeed, the virtually zero p-value suggests that the null hypothesis is also rejected at, say, the 1% significance level. This finding implies that the forecasts are not fully rational.

Figure 1 plots the estimated conditional mean function and the associated 90% uniform CB. By default, `xtnpsreg` transforms the conditioning variable `fr_norm` onto the $[-1, 1]$ interval via the $x \mapsto 2\Phi(x) - 1$ transformation, where Φ denotes the normal CDF calibrated using the sample mean and variance of the conditioning variable. The fact that the CB does not always cover the zero horizontal line means that the conditional mean function is statistically different from zero as a whole, which, needless to say, is consistent with the aforementioned testing result. The plot also reveals that the conditional mean of the forecast error is an increasing function in the forecast revision and so provides support for theoretical predictions from information-rigidity models. The requisite information for generating figure 1 may be exported to a spreadsheet by calling the `excel` option.

4.3 Robustness checks with respect to tuning parameters

The proposed nonparametric econometric method mainly involves two tuning parameters. One is the number of series terms `m()`. The default specification is `m(6)`, which corresponds to a fifth-order Legendre polynomial. The other is the bandwidth parameter `lag()` stemming from the computation of the Newey–West-type standard error, which is set to be the integer part of $0.75T^{1/3}$ by default. In theory, one should use more series terms for larger samples and use more lags if the data exhibit stronger serial dependence on the t dimension. But it is difficult in practice to pin down these choices "optimally". It is thus useful to check the robustness of empirical findings with respect to these choices.

As a concrete demonstration, we repeat the nonparametric estimation with different numbers of series terms (4, 6, 8, 10) by modifying the `m()` option as follows:

```
. xtnpsreg fe_norm fr_norm, m(4) plot
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.9230	2.3686	0.000

```
. xtnpsreg fe_norm fr_norm, m(6) plot
Notice: fr_norm must be a continous variable.
```

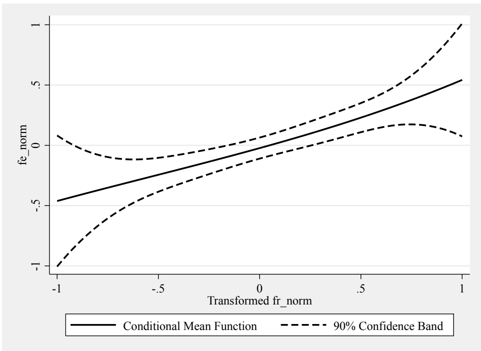
Transformation:	sup-t	10% critical value	P>|t|
Normal	4.2721	2.5508	0.000

```
. xtnpsreg fe_norm fr_norm, m(8) plot
Notice: fr_norm must be a continous variable.
```

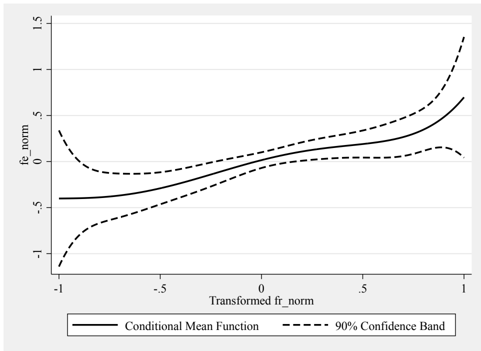
Transformation:	sup-t	10% critical value	P>|t|
Normal	5.2039	2.6139	0.000

```
. xtnpsreg fe_norm fr_norm, m(10) plot
Notice: fr_norm must be a continous variable.
```

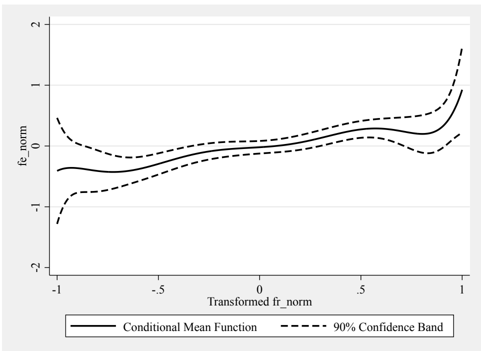
Transformation:	sup-t	10% critical value	P>|t|
Normal	5.4878	2.6763	0.000



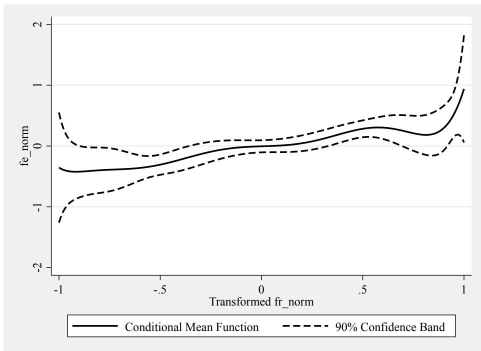
(a) `m(4)`



(b) `m(6)`



(c) `m(8)`



(d) `m(10)`

Figure 2. Nonparametric estimates with different numbers of series terms

We may also check the effect of the Newey–West lag parameter by modifying the `lag()` option as follows:

```
. xtnpsreg fe_norm fr_norm, lag(2) plot
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.5055	2.5081	0.000

```
. xtnpsreg fe_norm fr_norm, lag(4) plot
Notice: fr_norm must be a continous variable.
```

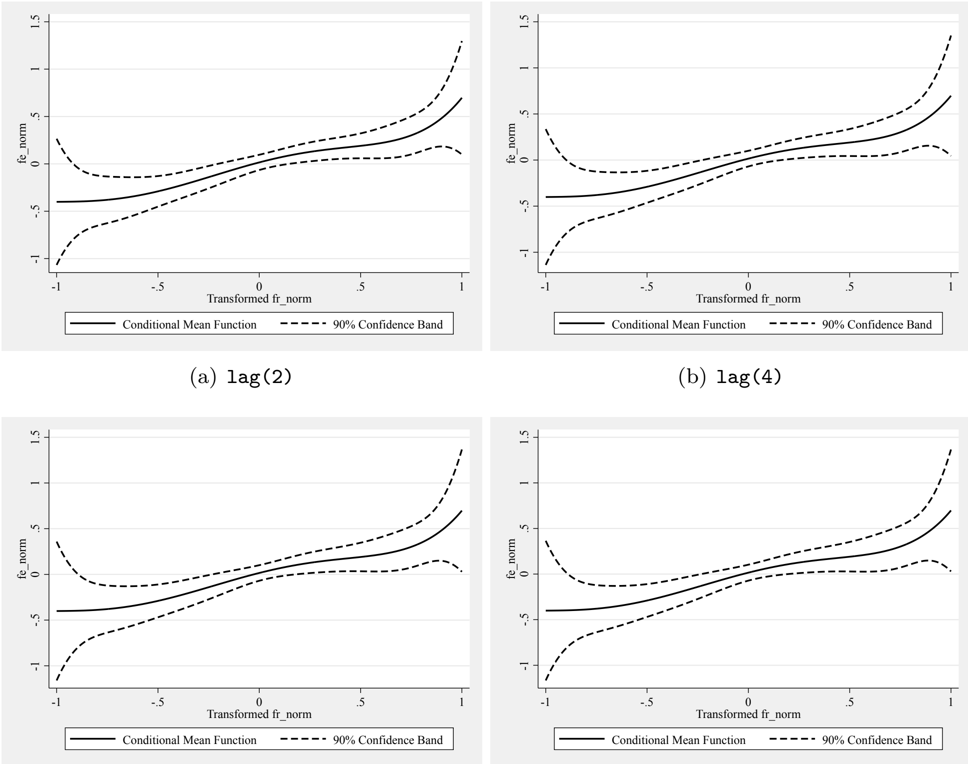
Transformation:	sup-t	10% critical value	P>|t|
Normal	4.2721	2.5428	0.000

```
. xtnpsreg fe_norm fr_norm, lag(6) plot
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.1318	2.5351	0.001

```
. xtnpsreg fe_norm fr_norm, lag(8) plot
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.0872	2.5121	0.001



(a) `lag(2)` (b) `lag(4)`

(c) `lag(6)` (d) `lag(8)`

Figure 3. Nonparametric estimates with different numbers of Newey–West lags

From these tables and the plots in figures 2 and 3, we see that the empirical results are fairly robust with respect to different choices of the number of series terms or the number of Newey–West lags.

4.4 Nonparametric inference for marginal response

Besides the conditional mean function $g(\cdot)$ itself, applied researchers may be interested in estimating the marginal response, defined as the derivative function $\partial g(\cdot)$. The nonparametric estimate and the associated uniform CB can be computed via `xtnpsreg` by calling the `marginal` option as follows:

```
. xtnpsreg fe_norm fr_norm, plot marginal
Notice: fr_norm must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.3819	2.5652	0.000

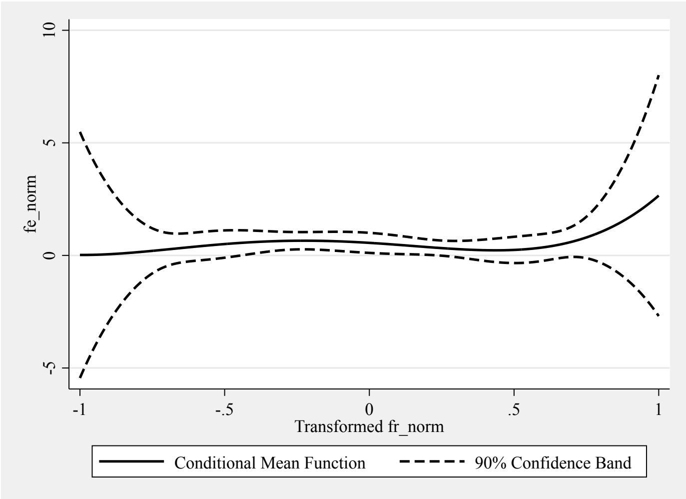


Figure 4. Nonparametric estimate and uniform CB of the derivative function

In this table, the sup-t statistic, critical value, and p-value pertain to testing the null hypothesis that the derivative function $\partial g(\cdot)$ is identically zero. The results suggest that the marginal response function, as a whole, is statistically different from zero. Figure 4 plots the nonparametric estimate of the marginal response and its 90% uniform CB. From the figure, we see that the estimated marginal response is nonnegative over the conditioning space, which is consistent with the previous observation that the conditional mean function of forecast error is increasing in the amount of forecast revision as predicted by information-rigidity models.

4.5 Specification tests

The uniform nonparametric inference method may also be used to conduct nonparametric specification tests against parametric model restrictions. For instance, we may formally test whether a linear specification is sufficient to describe the relationship between forecast error and forecast revision. To do so, we first run an ordinary least-squares regression of forecast error on forecast revision and obtain the residual as follows:

```
. regress fe_norm fr_norm
      Source |       SS           df       MS      Number of obs   =     3,460
-------------+----------------------------------   F(1, 3458)      =    207.14
       Model |  194.412734         1  194.412734   Prob > F        =    0.0000
    Residual |  3245.58726     3,458  .938573529   R-squared       =    0.0565
-------------+----------------------------------   Adj R-squared   =    0.0562
       Total |        3440     3,459  .994507082   Root MSE        =     .9688

------------------------------------------------------------------------------
     fe_norm | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     fr_norm |   .2377295   .0165179    14.39   0.000     .2053437    .2701154
       _cons |   4.32e-10   .0164701     0.00   1.000    -.0322921    .0322921
------------------------------------------------------------------------------

. predict fresidual, residuals
```

If the linear specification for the conditional mean function is correct, the conditional expectation of the residual given the conditioning variable should be zero. To test this formally, we use `xtnpsreg` to implement the nonparametric regression as follows:

```
. xtnpsreg fresidual fr_norm
Notice: fr_norm must be a continous variable.
-----------------------------------------------------------
Transformation: |  sup-t    10% critical value   P>|t|
----------------+------------------------------------------
Normal          |  2.1216         2.5181          0.244
-----------------------------------------------------------
```

From the table, we see that the null hypothesis of correct specification cannot be rejected at the 10% level, which suggests that the linear specification is in fact compatible with the observed data.

Note that the residuals obtained from the linear regression are "generated variables" in that they are noisy approximations for the unobserved disturbance terms. That noted, it can be shown theoretically (see Li and Liao [2020]) that this approximation error is asymptotically negligible for the nonparametric specification test. The intuition is that, in large samples, estimation errors in the linear regression coefficients shrink to zero at a faster rate than the statistical error in the nonparametric test. This theoretical intuition works better when the number of series terms is relatively large. It is thus advisable to check the robustness of the empirical finding by increasing `m()` as shown in the following implementation:

```
. xtnpsreg fresidual fr_norm, m(10)
Notice: fr_norm must be a continous variable.
-----------------------------------------------------------
Transformation: |  sup-t    10% critical value   P>|t|
----------------+------------------------------------------
Normal          |  2.4977         2.6846          0.165
-----------------------------------------------------------
```

4.6 Semi-nonparametric setting with linear control variables

We next describe how to use `xtnpsreg` in the partial linear model (3) where a vector $\mathbf{z}_{it}$ of control variables enters the specification linearly. For this illustration, we randomly generate two control variables, `Z1` and `Z2`, and feed them to `xtnpsreg` as *controlvar*. We also turn on the `table` option to display the estimated coefficients for all series terms and control variables.

```
. generate Z1 = rnormal()
. generate Z2 = rnormal()
. xtnpsreg fe_norm fr_norm Z1 Z2, table
Notice: fr_norm must be a continous variable.
Number of obs          =       3460
Newey-West maximum lag =          4
```

	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
p_1(fr_norm)	-.0099727	.0330602	-0.30	0.763	-.0752483	.055303
p_2(fr_norm)	.278855	.0642563	4.34	0.000	.1519845	.4057255
p_3(fr_norm)	.0168277	.027064	0.62	0.535	-.0366086	.0702641
p_4(fr_norm)	.0053773	.0354446	0.15	0.880	-.0646061	.0753606
p_5(fr_norm)	.039957	.0233346	1.71	0.089	-.0061158	.0860299
p_6(fr_norm)	.0159266	.0301342	0.53	0.598	-.0435717	.0754249
Z1	-.0185829	.0175373	-1.06	0.291	-.0532095	.0160436
Z1	-.0105368	.0157291	-0.67	0.504	-.041593	.0205195

Transformation:	sup-t	10% critical value	P>|t|
Normal	4.2384	2.4951	0.000

As expected, the coefficients of `Z1` and `Z2` are both close to zero and statistically insignificant because they are simply irrelevant for the data-generating process. Meanwhile, the sup-t statistic and its critical value remain very similar to those seen in section 4.2.

4.7 Uniform inference for functional coefficients

We now demonstrate how to use `xtnpsreg` to conduct inference in the functional coefficient model (2) and its generalized version (4). In this illustration, the dependent variable y_{it} remains the forecast error. But we now set the forecast revision as the base explanatory variable u_{it}, and its marginal effect is given by the function $\beta(\cdot)$ of a new conditioning variable x_{it}. We take x_{it} as the log volatility of the U.S. stock market portfolio computed as the logarithm of the standard deviation of daily returns (in percentage) over the preceding month. In this example, x_{it} happens to be a univariate time series not depending on i; this is permitted, but not required, by the estimation procedure.

The time series of log market volatility is stored as `log_mktvol` in `volatility.dta`, provided in the online supplemental material. As a preliminary preparation, we need to convert the univariate volatility series into a panel by merging the original panel dataset, `spf.dta`, with the new one as follows:

```
. *** Merge data ***
. merge m:1 id_date using volatility
    Result                           Number of obs
    ------------------------------------------
    Not matched                                  0
    Matched                                  3,460  (_merge==3)
    ------------------------------------------
```

The functional coefficient model (2) is fit by turning on the `funccoef` option in `xtnpsreg` as follows:

```
. xtnpsreg fe log_mktvol fr, plot funccoef
Notice: fr must be a continous variable.
```

Transformation:	sup-t	10% critical value	P>|t|
Normal	3.6591	2.5632	0.003

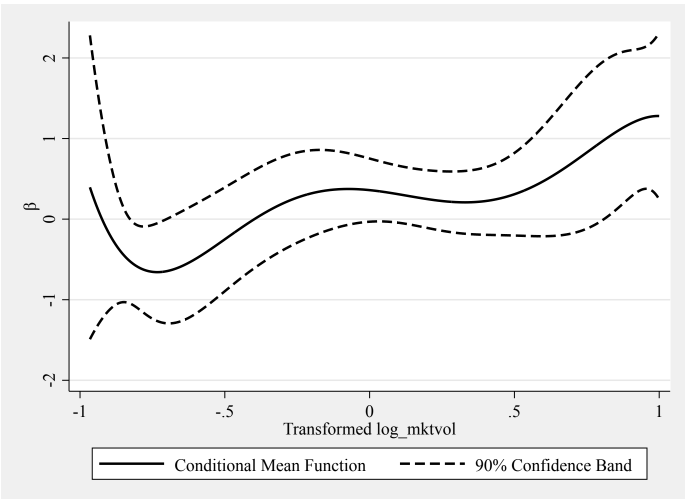


Figure 5. Nonparametric estimate and uniform CB of the functional coefficient $\beta(\cdot)$ plotted on the transformed scale using the `plot` option

It is instructive to clarify the syntax of this command. Here `fe` and `log_mktvol` are parsed as *depvar* and *condvar*, respectively. With the `funccoef` option specified, the variable that immediately follows *condvar* (that is, `fr`) is parsed as *basevar*, and the remaining variables, if there are any, are parsed as *controlvar*. Without turning on `funccoef`, `xtnpsreg` would instead parse all variables following *condvar* as *controlvar*, as described in section 4.6 above. The command above also generates a plot for the nonparametric estimate of the functional coefficient $\beta(\cdot)$ and its uniform CB, as shown in figure 5 above.

As in the baseline setting, `xtnpsreg` transforms the conditioning variable using the normal CDF to the $[-1, 1]$ interval because the `method(normal)` option is active by default. The functional estimates in figure 5 are plotted under the transformed scale, which explains the $[-1, 1]$ domain on the horizontal axis. The user may also obtain plots on the original untransformed scale of *condvar* by replacing `plot` with `plotu`, as shown in figure 6.

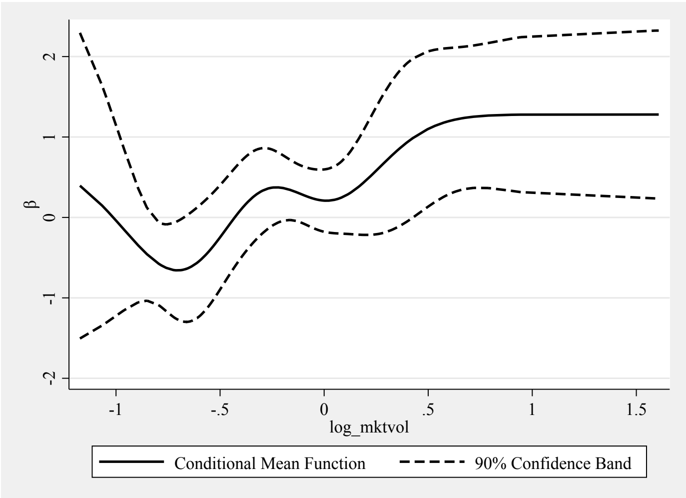


Figure 6. Nonparametric estimate and uniform CB of the functional coefficient $\beta(\cdot)$ plotted on the original scale using the `plotu` option

4.8 Implementation for "large N small T" panels via index swapping

As discussed in section 2, the proposed method relies on a "large T" setting but does not restrict the cross-sectional dimension, which may be fixed or divergent. Correspondingly, the underlying theory also requires that the dependence along the time-series dimension be weak, whereas the cross-sectional dependence is allowed to be arbitrarily strong.

Note that whether one labels i or t as "individual" or "time" is completely inconsequential. For all econometric purposes, what matters is that i indexes the dimension with possibly strong dependence and arbitrary sample size and t indexes the dimension with weak dependence (with independence being a special case) and large sample size.

Therefore, by swapping the roles of i and t (so that i and t become the time and cross-sectional indexes, respectively), we may also apply `xtnpsreg` in short panels with many independent cross-sectional units. Because of the lack of dependence in the new

t dimension, the Newey–West lag should be manually set to `lag(0)`. Arbitrarily strong serial dependence is accommodated because time is now indexed by i.

5 Conclusions

Nonparametric panel regression models have wide applications in empirical research. In this article, we introduced the command `xtnpsreg`, which implements valid uniform inference for semi-nonparametric panel models with general forms of spatiotemporal dependence. We illustrated the usefulness of the command via a detailed example concerning forecast rationality. A potential limitation is that the current version of `xtnpsreg` cannot handle time-fixed effects because of the lack of supporting econometric theory, which we leave as a promising extension for future research.

6 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-1
. net install st0707      (to install program files, if available)
. net get st0707          (to install ancillary files, if available)
```

7 References

Andrews, D. W. K. 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59: 307–345. https://doi.org/10.2307/2938259.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato. 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* 186: 345–366. https://doi.org/10.1016/j.jeconom.2015.02.014.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119: 249–275. https://doi.org/10.1162/003355304772839588.

Coibion, O., and Y. Gorodnichenko. 2015. Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review* 105: 2644–2678. https://doi.org/10.1257/aer.20110306.

Driscoll, J. C., and A. C. Kraay. 1998. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics* 80: 549–560. https://doi.org/10.1162/003465398557825.

Hoechle, D. 2007. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal* 7: 281–312. https://doi.org/10.1177/1536867X0700700301.

Jann, B. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s455001.html.

Li, J., and Z. Liao. 2020. Uniform nonparametric inference for time series. *Journal of Econometrics* 219: 38–51. https://doi.org/10.1016/j.jeconom.2019.09.011.

Li, J., Z. Liao, and M. Gao. 2020. Uniform nonparametric inference for time series using Stata. *Stata Journal* 20: 706–720. https://doi.org/10.1177/1536867X20953576.

Li, J., Z. Liao, and W. Zhou. 2021. Uniform nonparametric inference for spatially dependent panel data. Working Paper. https://www.dropbox.com/s/yw4rmh7y5sewwij/space.pdf?dl=0.

Newey, W. K. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79: 147–168. https://doi.org/10.1016/S0304-4076(97)00011-0.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708. https://doi.org/10.2307/1913610.

Petersen, M. A. 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies* 22: 435–480. https://doi.org/10.1093/rfs/hhn053.

Verardi, V. 2013. Semiparametric regression in Stata. U.K. Stata Users Group meeting proceedings. https://www.stata.com/meeting/uk13/abstracts/materials/uk13_verardi.pdf.

About the authors

Jia Li is the Lee Kong Chian Professor of Economics at the School of Economics, Singapore Management University.

Zhipeng Liao is an associate professor of economics at the Department of Economics, UCLA.

Wenyu Zhou is an assistant professor at the International Business School, Zhejiang University.