Appendix: Testing Coefficient Variability in Spatial Regression

Ulrich K. Müller and Mark W. Watson
Department of Economics
Princeton University

This Draft: April 2024

This appendix contains three sections. The first section provides instructions for computing the test statistic, p-value and other statistics. The second section discusses modifications of the analysis for IV regressions. The third section describes the data used in the paper.

1 Summary of Calculations for SVP Tests and Inference

This section outlines the calculations necessary to compute:

- The SVP test statistic, ξ_q, and its p-value;
- Confidence intervals and the median unbiased estimate of the magnitude of instability;
- Estimates of the parameter path.

A Matlab program is available on our websites for carrying out these calculations.

Notation:

- The data are $\{y_l, x_l, z_l\}$ for $l = 1, ..., n$. We also use $w_l = (x_l, z_l')'$. y_l and x_l are scalars, z_l is a $p-1$ vector, and w_l is $p \times 1$.
- The locations are denoted by s_l.
- The regression equation is written in two ways:

$$y_l = x_l\beta_l + z_l'\alpha + u_l$$

and

$$y_l = w_l'\delta + e_l, \quad e_l = u_l + (\beta_l - \beta)x_l.$$

We now list a series of calculations needed to compute the various statistics discussed in the paper.

As written, in addition to the data and locations, these steps require two inputs:

- q: The number of eigenvectors/eigenvalues used in the construction of the test statistic. We suggest using $q = 15$.

- $\overline{\rho}$: This is the value of $\overline{\rho}$ used for the kernel used to compute $\hat{V}_0$ in Step 8. (In the examples in the paper we use $\overline{\rho} = 0.015$.)

1. Compute an $n \times n$ matrix $\mathbf{D}$ that contains the (normalized) pairwised distances between the observations. Let $d_{max} = \max_{l,\ell} ||s_l - s_\ell||$ denote the maximum pairwise distance in the sample. Compute the matrix $\mathbf{D}$ as

$$\mathbf{D}_{l,\ell} = \frac{||s_l - s_\ell||}{d_{max}}.$$

   Note that $\mathbf{D}$ imposes the scale normalization that the largest pairwise distance is unity.

2. Compute Σ_L, the Lévy-Brownian motion covariance matrix evaluated at the locations $\{s_l\}_{l=1}^n$. The formula from the text is $cov(s_l, s_\ell) = \frac{1}{2}(||s_l|| + ||s_\ell|| - ||s_l - s_\ell||)$. For the calculations, it is convenient to use s_1 as the origin. And, to set the units for κ estimated in Step14 it is convenient to scale the location by d_{max} from Step 1. This yields

$$\begin{aligned}\Sigma_{L,l,\ell} &= \frac{1}{2}(||s_l - s_1|| + ||s_\ell - s_1|| - ||s_l - s_\ell||)/d_{max} \\ &= \frac{1}{2}(\mathbf{D}_{l,1} + \mathbf{D}_{\ell,1} - \mathbf{D}_{l,\ell})\end{aligned}$$

   where $\mathbf{D}$ was computed in Step 1.

3. Compute $\overline{\Sigma}_L$, the Lévy-Brownian motion covariance matrix after demeaning over the sample locations:

$$\overline{\Sigma}_L = M_1 \Sigma_L M_1 \text{ with } M_1 = I - 1(1'1)^{-1}1'.$$

4. Compute $\hat{\delta}$, the OLS estimate of δ and $\hat{e}_l = y_l - w_l'\hat{\delta}$ the OLS residuals.

5. Compute the q largest eigenvalues of $\overline{\Sigma}_L$; ordered from largest to smallest, these are $\lambda_1, \lambda_2, ..., \lambda_q$. Compute the corresponding eigenvectors, $r_1, r_2, ..., r_q$. Normalize the eigenvectors so that $n^{-1}r_i'r_i = 1$ for all i.

6. Let $r_{j,l}$ denote the lth element of the $n \times 1$ vector r_j. Let $r_{j,l}^e = (r_{j,l} \;\; 0_{1\times(p-1)})'$ (so that $r_{j,l}x_l = r_{j,l}^{e\prime}w_l$). Compute

$$\overline{r}_{j,l}^e = r_{j,l}^e - \left(n^{-1}\sum_{\ell=1}^n w_\ell w_\ell'\right)^{-1} n^{-1}\sum_{\ell=1}^n w_\ell w_\ell' r_{j,\ell}^e \text{ for } j = 1, ..., q \text{ and } l = 1, ..., n.$$

7. Choose the value of c for the kernel k_c used to compute $\hat{V}_0$. In our example we chose c to solve $\overline{\rho} = 0.015 = \frac{1}{n(n-1)} \sum_{l=1}^{n} \sum_{\ell \neq l} k_c(s_l, s_\ell)$ with $k_c(s_l, s_\ell) = \exp(-c \times \mathbf{D}_{l,\ell})$ (where this formula uses the scale normalization of the locations from Step 1). Call this value $c_{\overline{\rho}}$ and the resulting kernel $k_{c_{\overline{\rho}}}$.

8. Compute $\hat{V}_0$: for $i, j = 1, ..., q$ compute

$$\hat{V}_{0,i,j} = n^{-1} \sum_{l,\ell=1}^{n} \hat{\upsilon}_{l,i} k_{c_{\overline{\rho}}}(s_l, s_\ell) \hat{\upsilon}_{\ell,j}, \text{ with } \hat{\upsilon}_{l,j} = \overline{r}^{e\prime}_{j,l} w_l \hat{e}_l.$$

9. Compute $\hat{V}_1$: for $i, j = 1, ..., q$ compute

$$\hat{V}_{1,i,j} = n^{-2} \sum_{l,\ell=1}^{n} b_{l,i} \overline{k}_n(l, \ell) b_{\ell,j}, \text{ with } b_{l,j} = \overline{r}^{e\prime}_{j,l} w_l x_l$$

   where $\overline{k}_n(l, \ell)$ is the l, ℓth element of $\overline{\Sigma}_L$.

10. Compute the test statistic

$$\xi_q = \sum_{j=1}^{q} \lambda_j (n^{-1/2} \sum_{l=1}^{n} r_{j,l} x_l \hat{e}_l)^2$$

11. Generate $Y^*_{0,i} \sim iid\mathcal{N}(0, \hat{V}_0)$ for $i = 1, ..., N$. Generate $Y^*_{1,i} \sim iidN(0, \hat{V}_1)$ for $i = 1, ..., N$, and where $\{Y^*_{0,i}\}$ are independent of $\{Y^*_{1,i}\}$. (Choose N as a large number such as $N = 10,000$).

12. Estimate the p-value from the test as

$$\widehat{p_{value}} = \frac{1}{N} \sum_{i=1}^{N} 1 \left(\left[\sum_{j=1}^{q} \lambda_j (Y^*_{0,i,j})^2 \right] \geq \xi_q \right)$$

   where $Y^*_{0,i,j}$ is the jth element of $Y^*_{0,i}$.

13. Construct a $100 \times (1 - \alpha)\%$ confidence interval from κ as follows:

(a) Using generic notation, for $\upsilon \in (0,1)$ let $\hat{\kappa}_\upsilon$ solve

$$\upsilon = \frac{1}{N}\sum_{i=1}^{N} 1\left(\left[\sum_{j=1}^{q} \lambda_j (Y^*_{0,i,j} + n^{1/2}\hat{\kappa}_\upsilon Y^*_{1,i,j})^2\right] \leq \xi_q\right)$$

(b) The estimate of the $100 \times (1-\alpha)\%$ confidence interval for κ is

$$\kappa \in [\hat{\kappa}_{1-\alpha/2}, \hat{\kappa}_{\alpha/2}].$$

14. Estimate the median unbiased estimate of κ as

$$\hat{\kappa}^{MU} = \hat{\kappa}_{0.5}.$$

- Note: The units for κ. Recall that $\overline{\Sigma}_L$ was computed using the normalized distance matrix $\mathbf{D}$. This implies that the standard deviation of $(\beta_l - \beta_\ell)$ is given by $\kappa||s_l - s_\ell||/d_{max}$. Equivalently, κ is the standard deviation of the change in β_l over the longest pairwise distances in the sample.

15. Estimating the parameter path using an approximation to $\mathbb{E}(\beta_l|Y_n)$:

- Compute the $q \times 1$ vector $\hat{\sigma}_{LY}(s_l)$ with ith element given by

$$\hat{\sigma}_{LY,i}(s_l) = n^{-1}\sum_{\ell=1}^{n} \overline{r}^{e\prime}_{i,\ell} w_\ell x_\ell \overline{k}_n(s_\ell, s_l)$$

- Choose a value of κ, such as $\kappa = \hat{\kappa}^{MU}$, and then compute

$$\hat{\beta}_l = \hat{\beta} + n^{1/2}\kappa^2 \hat{\sigma}_{LY}(s_l)'(\hat{V}_0 + n\kappa^2\hat{V}_1)^{-1} Y_n.$$

2 Generalization for IV regression

Consider an IV regression estimated by 2SLS. Let $\hat{w}_l$ denote the fitted value from the first stage regression, that is, the regression of w_l onto the instruments. Let $\hat{x}_l$ denote the first element of $\hat{w}_l$ and $\hat{e}_l = y_l - w_l'\hat{\delta}_{2SLS}$ denote the IV residual. An IV version of the test statistic

replaces $x_l\hat{e}_l$ with $\hat{x}_l\hat{e}_l$ and uses the test statistic:

$$\xi_{2SLS,q} = \sum_{j=1}^{q} \lambda_i (n^{-1/2} \sum_{l=1}^{n} \hat{x}_l \hat{e}_l)^2.$$

Calculations then proceed as described in the paper (or the previous section in the appendix) with two changes:

1. The term $w_l\hat{e}_l$ is replaced with $\hat{w}_l\hat{e}_l$ throughout.

2. The expression for $\overline{r}^e_{j,l}$ is replaced with

$$\overline{r}^e_{j,l} = r^e_{j,l} - \left(n^{-1} \sum_{\ell=1}^{n} \hat{w}_\ell \hat{w}'_\ell \right)^{-1} n^{-1} \sum_{\ell=1}^{n} \hat{w}_\ell w'_\ell r^e_{j,\ell}$$

3 Data Description

As described in the paper, the data are from the American Community Survey, 5-year estimates from 2018-2022, for the zip codes regions ("zcta") making up the contiguous 48 states and the District of Columbia. The dataset contains sixty-two variables measuring population, educational attainment, income, employment, race, citizenship, health, marital status, mobility, and a handful of other indicators. The underlying dataset is a balanced panel of roughly thirty thousand zip codes. Zip codes containing a small number of observations (generally 250 or fewer) were merged with adjacent zip codes, resulting in a balanced panel of $n = 21,194$ regions. The (approximate) center of each region was used as its location, s_l, and distances between regions are measured by the great circle formula.all of the data. The Excel file **SVP_Data_Description.xlsx** lists each of the 62 variables used in the analysis, the population variable used to normalize the series and the p-values for the spatial unit (LFUR) and stationarity (LFST) tests described in Müller and Watson (2023). Also shown are "Category Indicators" for each series. The bivariate regressions use all possible combinations of variables from different categories, where in each case, one of the variables was randomly assigned to be the regressor and the other to be the regressand.

References

Müller, U. K., and M. W. Watson (2023): "Spatial Unit Roots," *Manuscript, Princeton University.*