Online Appendix: Testing Coefficient Variability in Spatial Regression

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

This Draft: July 2025

This appendix contains two sections. The first section provides instructions for computing the test statistic, critical values and the test's p-value. The second 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^s, the 'optimal' value of q as described in Section 4.2, critical values and the test's p-value. These are computed in the Matlab function **svp_test.m**. The script **SVP_Example.m** is a template for a main program: it reads in sample data (in SVP_Dataset_1.csv or SVP_Dataset_2.csv) and executes **svp_test.m**. This section of the Appendix serves as documentation for the **svp_test.m** function.

The **svp_test.m** function carries out the following steps:

1. It computes 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. The matrix $\mathbf{D}$ is computed 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. (In **svp_test.m**, the distances are computed as Euclidean distances or using the great circle formula (appropriate for latitude-longitude locations on the globe) as indicated by the indicator variable *latlonflag.* Euclidian distance is used when *latlonflag=0* and the great circle formula is used when *latlonflag=1.* The calculations are done in **get-distmat_normalized.m**.)

2. It then computes $\overline{\Sigma}_L$, the Lévy-Brownian motion covariance matrix after demeaning over the sample locations. (The calculations are done in **get_sigma_lbm_dm.m**.) This done in two steps.

   (a) It computes 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. This normalization 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.

(b) The demeaned value of the covariance matrix is

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

3. Compute the $q_{max} = 50$ largest eigenvalues and eigenvectors of $\overline{\Sigma}_L$. Denote the $n \times q_{max}$ matrix of eigenvectors as R and the eigenvalues as $(\lambda_1, ..., \lambda_{q_{max}})$.

4. Carry out two sub-steps: (both substeps are carried out in **form_omega_matrices.m**.)

   (a) Compute $c_{\overline{\rho}_{max}}$ where $\overline{\rho}_{max} = 0.01$ is the largest average spatial correlation for which the test controls size. (The function **svp_test.m** also computes a grid of value of c, say $\{c_i\}_{i=1}^{n_c}$ where $c_{\overline{\rho}_{max}} \leq c_i \leq c_{\overline{\rho}_{min}}$ where $c_{\overline{\rho}_{min}}$ is a large number. In the program it is computed as $c_{0.00001}$.)

   (b) For each value of c in the grid from Step 4(a) compute $\Omega_c = R'\Sigma_c R$ where $\Sigma_c = \exp(-cD)$ is the $\mathcal{G}_c$ covariance matrix evaluated at the sample locations.

5. Determine q^*, the value of $2 \leq q \leq 50$ that results in the greatest power, as described in Section 4.2. (This is carried out in the progam **get_qstar.m**.). This is implemented the following substeps:

   (a) Suppose $\tilde{Y}_c \sim N(0, \Omega_c)$. Let

$$\tilde{\xi}_{c,q} = \frac{\sum_{i=1}^{q} \lambda_i \tilde{Y}_{i,c}^2}{\sum_{i=1}^{q} \tilde{Y}_{i,c}^2}.$$

   For each value of q, find a critical cv_q where

$$\max_{i=1,...,n_c} \mathbb{P}(\tilde{\xi}_{c_i,q} \geq cv_q) = 0.05$$

These are the (approximate) 5% critical values for each value of q. (In the function, these calculations are carried out in **findcv.m**. The probabilities are computed using Imhof's formula after approximating the required integral using Gaussian quadrature.)

(b) Suppose that $\tilde{Y}(\kappa) \sim N(0, I + \kappa\Lambda)$ where Λ is a diagonal matrix with λ_i on the diagonal. Let

$$\tilde{\xi}_q(\kappa) = \frac{\sum_{i=1}^{q} \lambda_i \tilde{Y}(\kappa)_i^2}{\sum_{i=1}^{q} \tilde{Y}(\kappa)_i^2}$$

and then find the value of κ_q that yields

$$\mathbb{P}(\tilde{\xi}_q(\kappa_q) > cv_q) = 0.50$$

where κ_q is set to a large number if $\lim_{\kappa_q \to \infty} \mathbb{P}(\tilde{\xi}_q(\kappa_q) > cv_q) < 0.5$. (These calcuations are carried out in **find_kappa_50.m**.)

(c) Set $q^* = \text{argmin}\{\kappa_q\}_{q=2}^{q_{max}}$.

6. Using this value of q, compute 1%, 5% and 10% critical values following the procedure in step 5(a).

7. Using the sample data, regress y_i on $\{x_i, z_i\}$. Let $\hat{e}_i$ denote the residual. Compute

$$Y_j = \sum_{i=1}^{n} R_{ij} x_i \hat{e}_i$$

for $j = 1, ..., q^*$. Compute the test statistic

$$\xi_{q^*}^s = \frac{\sum_{i=1}^{q^*} \lambda_i Y_i^2}{\sum_{i=1}^{q^*} Y_i^2}$$

8. Compute the p-value as

$$p\text{-value} = \max_{i=1,...,n_c} \mathbb{P}(\tilde{\xi}_{c_i,q^*} \geq \xi_{q^*}^s)$$

where $\tilde{\xi}_{c_i,q^*}$ is defined in step 5(a) and the probability is computed using $\tilde{Y}_c \sim N(0, \Omega_c)$, also from 5(a).

2 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 (2024). Also shown are "Category Indicators" for each series. The bivariate regressions use all possible combinations of variables from different categories.

References

Müller, U. K., and M. W. Watson (2024): "Spatial Unit Roots and Spurious Regression," *Econometrica*, 92(5), 1661–1695.