# Spatial correlation

Hans Martinez

2023-07-06

## Intro

Open questions to address in Spatial Correlation Principal Component (SCPC) (Müller and Watson 2022b, 2022a):

1. What's the "worst case" covariance matrix? How does that work?
2. Can SCPC accommodate a vector of variables?
3. Does the method work if there exists measurement error in the locations?

## 1 Replication

Here is the first try in replicating Table 1 of Müller and Watson (2022b). In all models, $s_l \sim U(0,1)$. In most models $e_l \sim \mathcal{G}_{exp}(c_{0.03})$, however, to optimize time I used $c = 75.167$ which gives $\bar{\rho} \approx 0.03$. In model 1, $x_l = 1$. In model 2, $x_l \sim \mathcal{G}_{exp}(c_{0.03}/2)$ and $e_l \sim \mathcal{G}_{exp}(c_{0.03}/2)$, however I used $c = 75.167/2$ for both. In model 3, $x_l$ is a step function with $x_l = -0.15$ for the 85% of the locations closest to $s = 0$, and $x_l = 0.85$ for the remaining locations closest to $s = 1$. In model 4, $x_l$ follows a demeaned random walk.

I ran 200 simulations for each model. I used 250 observations for each iteration. To draw from $\mathcal{G}_{exp}(c_{0.03})$, first I estimated the covariance matrix $\Sigma(c)$ using the covariance function $cov(e_l, e_{l'}) = \exp(-c||s_l - s_{l'}||)$ for a value of c. Then, I used the covariance matrix to draw from a multivariate normal distribution with mean zero and covariance matrix $\Sigma(c)$. The rejection frequencies are just the times that the p-value was above 0.05 divided by the number of simulations. The HR method uses the `robust` method in Stata. The SCPC uses the post-estimation command `scpc` provided by the authors.

The rejection frequencies in Table 1 are far from what Müller and Watson (2022b) show. I will correct the calibration of the $c$ parameter to get $\bar{\rho} = 0.03$ consistently, however, this will increase the simulation time significantly because I would have to use an optimizer to select $c_{0.03}$.

Table 1: Rejection frequencies

| Model | Method.HR | Method.SCPC |
|---|---|---|
| 1 | 0.000 | 0.000 |
| 2 | 0.395 | 0.950 |
| 3 | 0.610 | 0.995 |
| 4 | 1.000 | 1.000 |

# 2 Worst case covariance function

Müller and Watson (2022b) assume a functional form for the covariance function of the error term that depends on a single parameter. This single parameter has a unique value for a specific average pairwise correlation. The researcher has to provide before the estimation, a maximum average pairwise correlation. Then, because the average pairwise correlation is a decreasing function of the parameter of the covariance function, this parameter induces the largest value of the error variance for the given average pairwise correlation, according to the authors. In the STATA and MATLAB codes, this parameter is set to 0.03 in both, like in their paper. The authors provide instructions on how to adjust the parameter in the help section of the STATA code. In the MATLAB code, a parameter can be adjusted before estimations.

## 2.1 The setting

$$y_l = x_l\beta + e_l$$

$\mathbb{E}[e_l|x_l] = 0$, and $(y_l, x_l, e_l)$ are associated with observed spatial location $s_l \in \mathbb{R}^d$. Spatial location in time is $d = 1$, two dimensions, like in altitude and latitude, $d = 2$, and so on.

$e_l$ is generated by a Gaussian process with covariance function $cov(e_l, e_{l'}) = \exp(-c||s_l - s_{l'}||)$, where $s_l$ and $s_{l'}$ denote the spatial locations of $e_l$ and $e_{l'}$, and $c > 0$ is a parameter that governs the strength of the spatial correlation. The value of $c$ is calibrated to induce a specific average pairwise correlation $\bar{\rho} = [n(n-1)]^{-1} \sum_{l,l' \neq l} cov(e_l, e_{l'})$. In other words, $c = c_{\bar{\rho}}$ solves $[n(n-1)]^{-1} \sum_{l,l' \neq l} \exp(-c_{\bar{\rho}}||s_l - s_{l'}||) = \bar{\rho}$.

Let $\Sigma(c)$ be the covariance matrix of $e_l$ evaluated at the sample locations, so that $\Sigma(c)_{l,l'} = \exp(-c||s_l - s_{l'}||)$, and let $\bar{\rho}(c)$ denote the resulting average pairwise correlation $\bar{\rho}(c) = [n(n-1)]^{-1} \sum_{l,l' \neq l} \Sigma(c)_{l,l'}$. If the researcher desires a test that controls size for values of $\bar{\rho}$ as large as $\bar{\rho}_{max}$, then he can choose $c_{min}$ such that $\bar{\rho}(c_{min}) = \bar{\rho}_{max}$. Then, $\Sigma(c_{min})$ is the **worst case** covariance matrix in the sense that it induces the largest value of $\sigma^2$ among all $\Sigma(c)$ with $\bar{\rho} \leq \bar{\rho}_{max}$

# 3 Literature Review

## 3.1 Müller and Watson (2022b)

The objective of the paper is to develop a robustified version of their Spatial Correlation Principal Components (SCPC) (Müller and Watson 2022a). In particular, the authors propose modifications to deal with non-stationarities and strong dependence in finite samples (small samples).

According to the authors, the SPCPC method addresses the challenge of spatial correlation robust inference under small samples and *empirically relevant* forms of strong dependence (which ones?).

According to the authors, Conley (1999) doesn't work well in small samples because it relies on the consistency of the estimator of $\sigma^2$, while SCPC (and fixed-b type approaches) rely on the *stationarity* of $u_l$. Stationarity might break in practice, for example, when $x_l$ is a dummy for treatment, and treatment is more likely in one region than another region.

The SCPC method is based on a principal component estimator of $\sigma^2$ based on a pre-specified "worst-case" exponential covariance function conditional on the observed locations.

(On going…)

# References


Conley, T G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92: 1–45.

Müller, Ulrich K, and Mark W Watson. 2022a. "SPATIAL CORRELATION ROBUST INFERENCE." *Econometrica* 90: 2901–35. https://doi.org/10.3982/ECTA19465.

———. 2022b. "Spatial Correlation Robust Inference in Linear Regression and Panel Models." *Journal of Business & Economic Statistics* 00: 1–15. https://doi.org/10.1080/07350015.2022.2127737.