Relations between spatial design criteria [1]

Werner G. Müller and Helmut Waldl
Department of Applied Statistics, Johannes-Kepler-University Linz,
werner.mueller@jku.at

Abstract: Several papers have recently strengthened the bridge connecting geostatistics and spatial econometrics. For these two fields various criteria have been developed for constructing optimal spatial sampling designs. We will explore relationships between these types of criteria as well as elude to space-filling or not space-filling properties.

Keywords: Empirical kriging, compound D-optimality, Moran's I

1 Introduction

Lindgren et al. (2011) further strengthen the bridge connecting the two somewhat disparate worlds of spatial analysis. One is rooted in the idea of observing continuously varying spatial processes and led to what is largely referred to as geostatistics. The other, which assumes (usually aggregate) observations attached to discrete (mostly irregular) lattices, is commonly known under the name of spatial econometrics. In particular in the latter literature the rift between these two points of view - manifesting itself along various themes - is a constant challenge towards a unified understanding (Griffith and Paelinck, 2007). Also for the more narrow topic of efficient estimation and prediction early contributions can be found there (Griffith and Csillag, 1993) and that the issue is of great current interest is documentable as well (Fernández-Avilés Calderón, 2009). The method of explicitely linking some Gaussian fields to Gaussian Markov random fields on irregular grids given in Lindgren et al. (2011) is certainly a very welcome addition to the equipment connecting the two views as the authors rightfully claim in their discussion section. It remains to be seen whether practitioners will be able to take it up as easily as a perhaps more pragmatic recent suggestion like Nagle et al. (2011).

2 Materials and Methods

But let us draw the attention towards a rather neglected (in the discussion section of Lindgren et al. (2011) as well most of the literature in general) aspect of establishing such a link as above. That is the potential impact of this link on the respective optimal sampling designs and the question of their effective generation.

[1] A considerably shortened and edited version of this paper will be published as a discussion of Lindgren et al. (2011) in JRSS-B.

We will illustrate our points on the same example as used in Section 2.3 of Lindgren et al. (2011), namely the leukaemia survival data, utilizing some of the calculations thankfully provided by the authors.

In geostatistics the optimal sampling design is often based upon the kriging variance over the region of interest $\mathcal{X}$, frequently by minimizing its maximum. It has turned out that this reflects rather not so well the true variation as the uncertainty introduced by estimating covariance parameters γ is thereby neglected. To compensate for that Zhu and Stein (2006) and Zimmerman (2006) have suggested minimizing the modification

$$\max_{x \in \mathcal{X}} \left\{ \mathrm{Var}[\hat{Y}(x)] + \mathrm{tr} \left\{ M_\gamma^{-1} \mathrm{Var}[\partial \hat{Y}(x)/\partial \gamma] \right\} \right\},$$

which the latter has termed the EK(empirical kriging)-criterion. Here M_γ stands for the Fisher information matrix with respect to γ, and we can analogously denote M_β for trend parameters β for later usage.

In spatial econometrics it is common to test for spatial autocorrelation by specifying a spatial linkage or weight matrix W and utilize an overall type measure such as Moran's I. Therefore Gumprecht et al. (2009) have suggested to employ the power of Moran's I under a hypothesized spatial lattice process given by its precision matrix Q as the design criterion; let us call maximization of it the MIP(Moran's I power)-criterion in the following.

3 Results

Now as there is a link established with respect to estimation between the two modelling paradigms, can we expect a similar link with respect to those associated design criteria? Looking at the example a sensible design question we could pose is to which out of the 24 districts in north-west England should we sample if we are limited to a number $k < 24$ for financial reasons. To keep things simple, we will in the following choose $k = 3$, which allows for $\binom{24}{3} = 2024$ different designs. For all those designs we can then calculate the values for the above design criteria and plot them against each other to judge for a potential linkage. As the only covariance parameter, which is not predetermined in the example is ρ, we have $\gamma = \rho$ and EK reduces to scalar operations localized at $\rho = 0.2$. For the MIP we required the precision matrix Q, which was provided by Lindgren et al. (2011). The matrix W was defined by assigning 1 to point pairs with intersite distances less than the range $\rho = 0.2$ and 0 else, which turned out to be an insensitive choice.

At this point we now had to slightly modify the example: since the spatial correlation is so strong in the leukaemia data most of the realized powers were very close to one, thus obscuring all potential patterns. We therefore artifically reduced the number of cases (and thus the powers) by randomly sampling 20 locations from the 3 districts respectively. This resulted in the scatter plot of criteria displayed in the left panel of the figure in the discussion of Lindgren et al. (2011). While

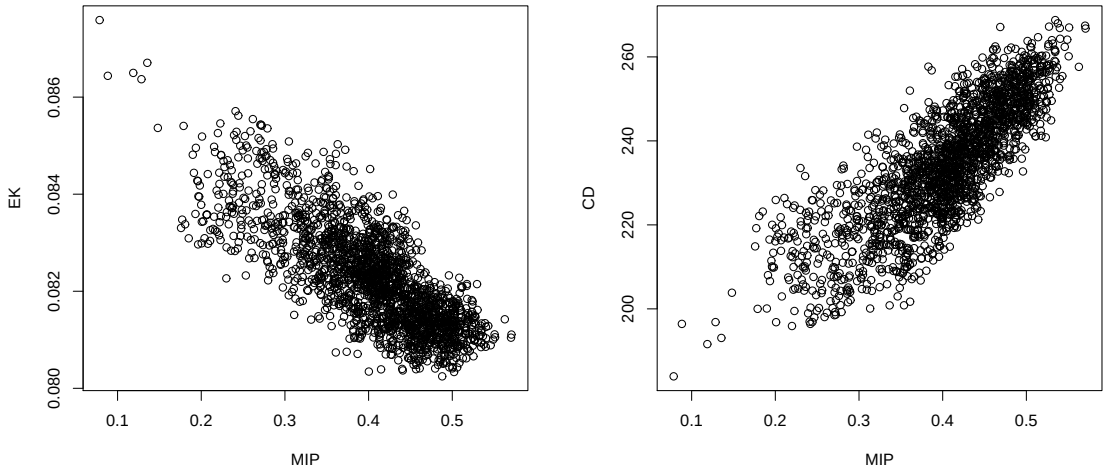

Figure 1: left panel: MIP (horizontal) versus EK (vertical) criterion values; right panel: MIP (horizontal) versus CD (vertical) criterion values.

from this display the link between the criteria already becomes quite evident, we present here in Figure 1 an even stronger one well extending into the corners where the optima lie. This was achieved by simply doubling the diagonal entries of the covariance matrix Q^{-1}, which emulates a stronger nugget effect.

It thus looks that in cases with reasonable localized spatial dependence one could achieve reasonably high design efficiencies by employing one for the other criterion, which offers advantages in both directions. Where MIP requires little prior knowledge its optimization is nonstandard, whereas for EK and related cases well developed theory is available (Müller and Pázman, 2003).

Both criteria, however, are computationally quite intensive and it makes thus sense to look for cheaper alternatives. Motivated by the traditional connection between estimation and prediction based criteria ("equivalence theory"), Müller and Stehlík (2010) have suggested to replace the EK-criterion by a compound criterion for determinants of information matrices, i.e. maximizing

$$|M_\beta)|^\alpha \cdot |M_\gamma|^{(1-\alpha)},$$

with a weighing factor α, which we will call in the following CD_α(compound D)-optimality. The relationship of this criterion (assuming a constant trend β) with an $\alpha = 0.5$ to the MIP is displayed in the right panel of Figure 1. This clearly shows that one could computationally very cheaply find the optimum with respect to CD and still achieve rather high efficiencies on the MIP criterion.

4 Concluding remarks

We must note that our calculations have shown that the dependence between the criteria is related to the specific setup. It turns out that the strength of the relation-

ship between MIP and the other two criteria decreases when the powers approach one, but strongly increases for decreasing ranges and increasing nuggets. Note also the relationships to the ubiquitous space-filling designs as explored in Pronzato and Müller (2011). Summarizing, we believe our discussion showed that the relations between the two linked approaches can go far beyond mere estimation issues.

References

Fernández-Avilés Calderón, G. (2009). Spatial regression analysis vs. kriging methods for spatial estimation. *International Advances in Economic Research 15*(1), 44–58.

Griffith, D. and J. Paelinck (2007). An equation by any other name is still the same: on spatial econometrics and spatial statistics. *The Annals of Regional Science 41*(1), 209–227.

Griffith, D. A. and F. Csillag (1993). Exploring relationships between semivariogram and spatial autoregressive models. *Papers in Regional Science 72*(3), 283–295.

Gumprecht, D., W. G. Müller, and J. M. Rodríguez-Díaz (2009). Designs for detecting spatial dependence. *Geographical Analysis 41*(2), 127–143.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach *Journal of the Royal Statistal Society Series B 73*(4), forthcoming.

Müller, W. G. and A. Pázman (2003). Measures for designs in experiments with correlated errors. *Biometrika 90*(2), 423–434.

Müller, W. G. and M. Stehlík (2010). Compound optimal spatial designs. *Environmetrics 21*(3-4), 354–364.

Nagle, N. N., S. H. Sweeney, and P. C. Kyriakidis (2011). A geostatistical linear regression model for small area data. *Geographical Analysis 43*(1), 38–60.

Pronzato, L. and W. G. Müller (2011). Design of computer experiments: space filling and beyond. *Statistics and Computing*, Online First.

Zhu, Z. and M. L. Stein (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics 11*(1), 24–44.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics 17*(6), 635–652.