Bayesian optimal design for prediction of correlated processes

Maria Adamou
Sue Lewis, Sujit Sahu and Dave Woods
University of Southampton
ma5g10@southampton.ac.uk

Southampton Statistical Sciences Research Institute

Introduction

Data collected from correlated processes arise in many diverse application areas including studies in environmental and ecological science and in both real and computer experiments. Often the main aim of the study is to predict the process at unobserved points.

In this work, we illustrate a new approach for the selection of Bayesian optimal designs using applications from both spatial statistics and computer experiments.

A Motivating Example

Networks of monitoring stations are regularly used to collect data to measure pollutant levels in water or air. Figure 1 shows such a network in the Eastern USA for monitoring chemical deposition.

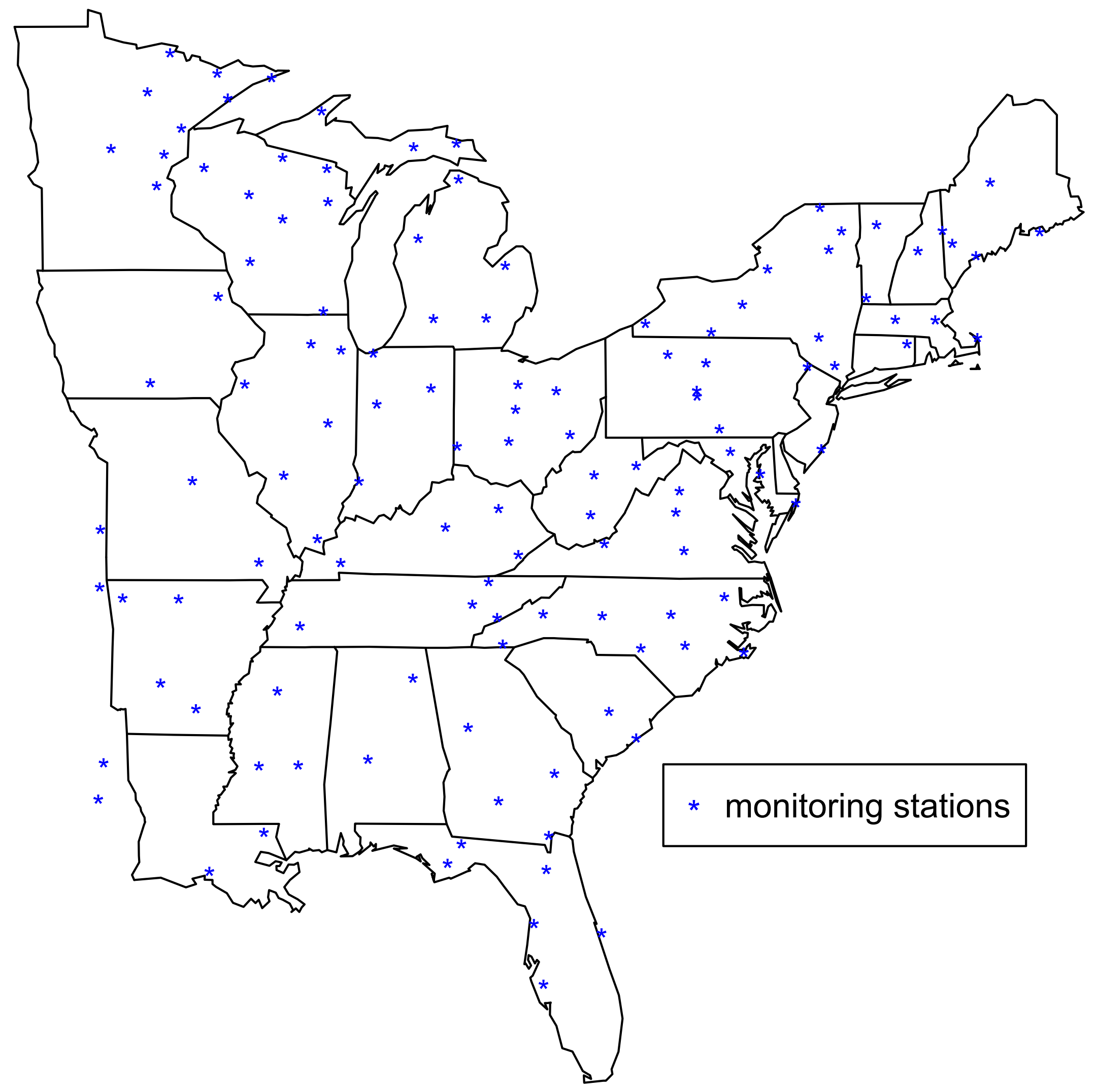

Figure 1: Network of monitoring stations in the Eastern USA.

An important problem is to identify the best locations of stations across the region of interest (Diggle and Lophaven, 2006; Zimmerman, 2006). Usually, observations at geographically close stations are assumed to be highly correlated, and this correlation helps to predict deposition levels at unobserved locations.

Statistical model

We assume a Gaussian regression model

$$Y(x)|\theta, \sigma^2, \phi, \tau^2 \sim N(X\theta, \sigma^2 C(\phi) + \tau^2 I) \qquad (1)$$

where $\boldsymbol{X}$ is the model matrix, θ contains regression coefficients, C is a correlation matrix from a stationary and isotropic correlation function determined by decay parameter ϕ, and σ^2, τ^2 are the spatial and pure error (nugget) terms respectively.

If a Bayesian approach is adopted, the model specification is completed by assignment of prior distributions to the unknown parameters.

Bayesian optimal design

The aim of our experiments is to enable accurate prediction of the response $Y(\boldsymbol{x})$ at unobserved $\boldsymbol{x}$. We adopt a decision theoretic approach (Chaloner and Verdinelli, 1995) to find Bayesian optimal designs that minimise the posterior predictive variance. To avoid the computational burden usually associated with Bayesian designs we have developed a new closed form approximation that allows quick calculation of the variance.

The designs are found using the coordinate exchange algorithm (Meyer and Nachtsheim, 1995), and illustrated with the following examples.

Example 1 - Optimal designs for spatial data

Here observations are assumed to follow model (1) with

- covariance function $\rho(d_{ij};\phi) = \exp(-\phi d_{ij})$, where d_{ij} is the Euclidean distance between points $\boldsymbol{x}_i, \boldsymbol{x}_j \in [-1,1]^2$
- prior distributions $\theta|\sigma^2 \sim N(0, \sigma^2 I)$, $\sigma^2 \sim$ Inverse Gamma$(3,1)$ and $\phi \sim$ Uniform$(0.1,1)$
- $\nu^2 = \tau^2/\sigma^2$ is fixed at one of four values: 0, 0.5, 1, 2.5.

Optimal designs with $n = 10$ points are shown in Figure 2.

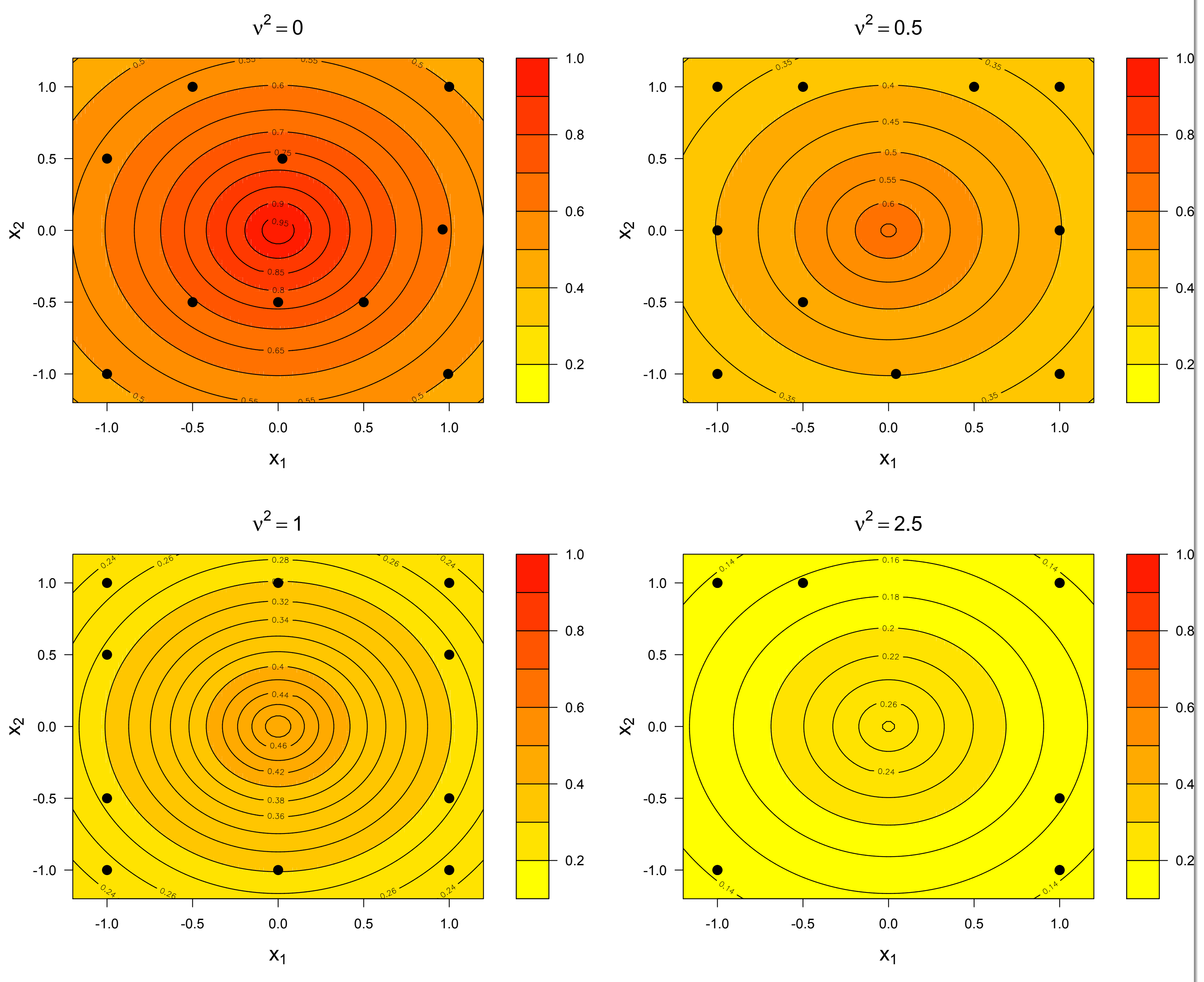

Figure 2: Optimal designs and average correlation surfaces for known ratio $\nu^2 = \tau^2/\sigma^2$.

The filled contours in Figure 2 give the correlation between each point of the design region and the centre of the design region, averaged across the prior for ϕ; the minimum values of these correlations are 0.5, 0.34, 0.25 and 0.13 for $\nu^2 = 0, 0.5, 1, 2.5$ respectively.

Clearly, the optimal designs change with ν^2 and the correlation structure. Low values of ν^2, which result in high correlation, give rise to designs where points are spread across the design region. In contrast, high values of ν^2 (low correlation) produce designs with points at the boundaries and corners of the design region. For $\nu^2 = 2.5$, the design is similar to a standard optimal design for a regression model with uncorrelated errors.

Example 2 - Computer experiments

Model (1) is commonly used to describe data from a deterministic computer experiment. Our example uses data from a simple simulator of a helical compression spring (Tudose and Jucan, 2007; Forrester et al., 2008). The example has three variables and

- $\rho(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\phi}) = \prod_{k=1}^{3} \exp(-\phi_k |x_{ik} - x_{jk}|)$ for $\boldsymbol{x}_i, \boldsymbol{x}_j \in [-1,1]^3$
- priors for θ, σ^2 as in the first example

Figure 3 shows Bayesian optimal designs with $n = 10$ runs for two different prior distributions

prior 1: ν^2=0, $\phi_1 \sim$ Unif$(1,3)$, $\phi_2 \sim$ Unif$(3,5)$, $\phi_3 \simeq 0$
prior 2: ν^2=0.5, $\phi_1 \sim$ Unif$(1,3)$, $\phi_2 \sim$ Unif$(1,3)$, $\phi_3 = 0$

These priors were obtained by analysing data from a maximin Latin hypercube design (LHD) (Morris and Mitchell, 1995), also shown in Figure 3.

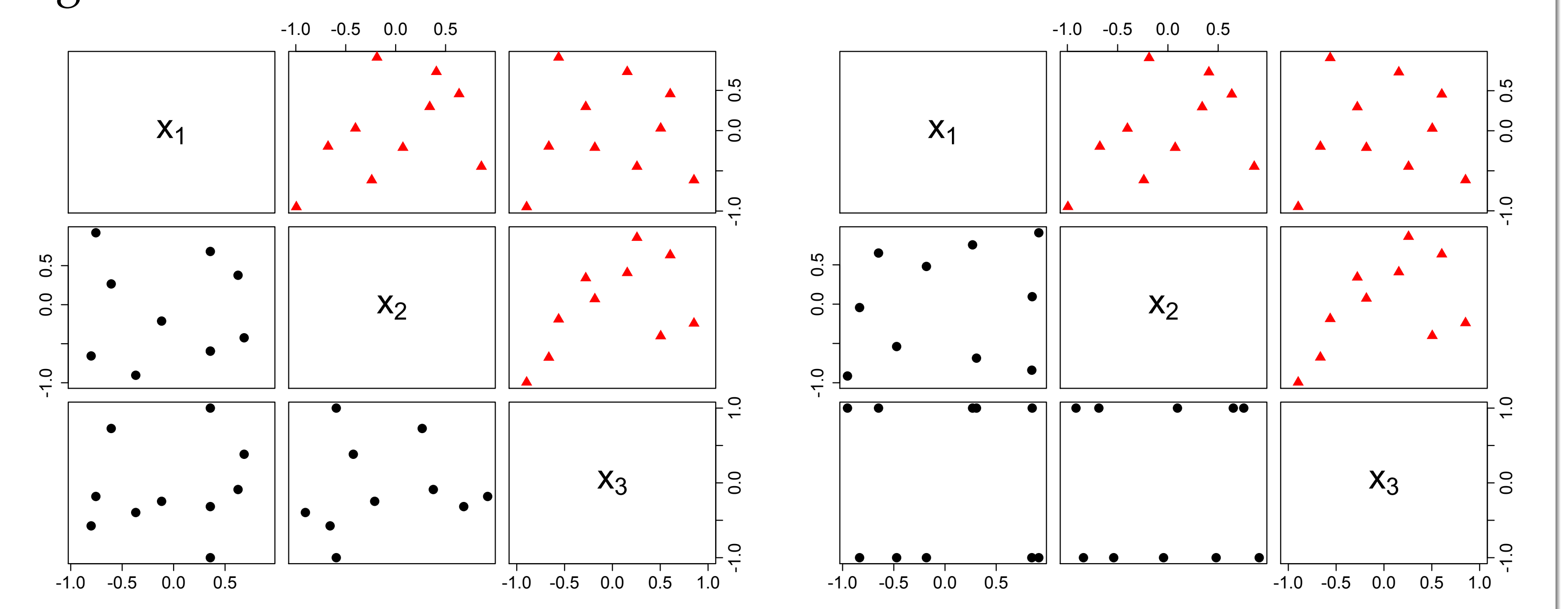

Figure 3: Bayesian optimal design (●) and Latin hypercube designs (▲) for prior 1 (left) and prior 2 (right).

For prior 1, the Bayesian optimal design has similar space-filling properties as the LHD (average intra-point distance of 1.43 vs 1.40) but has 30% smaller average posterior predictive variance. For prior 2, where there is little change in correlation with x_3, the design points in the x_3 dimension collapse onto the extremes (determined by the linear trend). This second design has posterior predictive variance 18% lower than that of the LHD.

Conclusions

We have investigated Bayesian optimal design for collecting correlated data when the aim of the experiment is accurate prediction. The designs we have studied are influenced by the degree of correlation, with higher correlation leading to designs which are close to space-filling, and provide lower prediction variance than LHDs.

References

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.

Diggle, P. and Lophaven, S. (2006) Bayesian geostatistical design. *Scandinavian Journal of Statistics*, **33**, 53–64.

Forrester, A., Sobester, A. and Keane, A. (2008) *Engineering Design via Surrogate Modelling*. Chichester: Willey.

Meyer, R. and Nachtsheim, C. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, **37**, 60–69.

Morris, D. and Mitchell, J. (1995) Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.

Tudose, L. and Jucan, D. (2007) Pareto approach in multi-objective optimal design of helical compression springs. *Annals of the Oradea University, Fascicle of Management and Technological Engineering*, **6(16)**, 991–998.

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

Design and Analysis of Experiments 2012

University of Georgia, Athens, October 2012