Comparing and generating Latin Hypercube designs in Kriging models

Giovanni Pistone, Grazia Vicario[1]

Politecnico di Torino

Department of Mathematics

Corso Duca degli Abruzzi, 24 – 10129 Torino, Italy

Abstract

The official start of the Computer Experiments (CE) is the paper of McKay et al. (1079) while the contribution by Sachs J. et al. (1989) marked a new step by introducing model based methods. Nowadays, it is acknowledged by any researchers that CE are useful in replacing totally or partially the physical experiments with the computer ones, so that their use has become increasingly a common practice. An actual reason for promoting the use of the CE (in a single or a combined approach with the physical experiments) is that the physical experimentation may be, in a number of circumstances, expensive and unapproachable. On the contrary, the use of the numerical experiments in product/process development phase is quite inexpensive and it has become straightforward. The general availability of comprehensive computing facilities and the recent progresses in software development make numerical simulation of complex systems an attractive alternative option to the execution of the expensive and time consuming physical experiments. Standard modern references are Sasena (2002), Santner et al. (2003), Fang et al. (2006).

In this contest, a careful selection of the design points or training points (see the discussion below) is mandatory if the aim is to study how the observed responses vary when the input variables are set to the values of interest. In physical experimentation, the researcher is asked to comply to a well set protocol in order to achieve correct inferences. Such a protocol is the Design of Experiments (DoE) methodology which is an helpful tool in carrying on the mentioned objectives. Unfortunately, the design of a CE, when it is used as a surrogate of the physical one, differs in several aspects from designing a physical experiment and the applicability of basic principles of D.o.E. is questioned in CE. The selection of an experimental design in CE is a crucial issue to get to an efficient and informative model and cannot be done by merely importing the concepts developed for physical experiments. This means providing efficient strategies for sampling the input space in order to get accurate predictions in untried inputs.

As suggested by the pioneers of the model based CE, the output can be predicted by assuming Gaussian responses with covariance depending parametrically on the distance between the locations, as it was in the Kriging modellization (Krige, 1951; Cressie, 1993). The underlying principle is that the nearer an untried point is to the design points the better is the prediction. Based on this view, a good design strategy is to uniformly spread the points across the experimental region. That prompts to the use of the so called space-filling designs. Random designs, Stratified designs and Latin Hypercube designs are common choices. The first two designs are not really satisfactory because they are not space filling marginally, i.e. in individual directions, and, moreover, are quite not satisfactory for global space filling especially for small number of design points. Whereas, even if the Latin Hypercube (LH) designs are not very satisfactory for space filling, they are satisfactory for space filling in individual directions, and this statement is valid for

1 Corresponding Author: grazia.vicario@polito.it

any number of design points. For a formal treatment of the subject see Butler (2001), Fang (2000), Park (1994), Welch (1992) etc. This is the reason for focussing on the class of LH designs among the different space-filling ones in this paper. We want to investigate which ones are the best in the class of LH designs having the same prediction features.
In our presentation we assume that the Gaussian process has distance-dependent covariance (stationarity) and it is defined on a subset of a regular grid, i.e. the cartesian product of uniform one-dimensional grids. This case correspond to specific applications, see Pistone and Vicario (2009), and it is specially adapted to the methodology we use. The commonly used Euclidean distance does not really fit when considering a regular grid; therefore we switch to the Manhattan distance, as other authors suggest (Santner et al., 2003, pag. 138). The typical design optimization criteria in case of standard DoE is the best prediction of a linear model. Here, we examine the behaviour of different LH design's in minimizing the error of prediction in the single points of the considered regular grid. According the Bayesian methodology for designing and analyzing CE, the comparison is carried out on the variance of the conditional expectation (Santner et al., 2003). Therefore, in this preliminary research we focus on studying the variance of the Gaussian linear prediction (Gaussian conditional expectation) as a function of the design, given that all parameters are known (for sake of simplicity of the discussion, we deliberately ignore the effects of estimating unknown model parameters). For any LH designs, the Mean Square Prediction Error (MSPE) of the linear predictor is the statistical index chosen for comparing different LH designs with the same number of training points. It should be noticed that the issue of identifiability which is of the highest importance in standard DoE is not relevant here.
In performing the mentioned comparisons, there are two tricky computational problems: the computation of variance-covariance matrix of the design points and the computation of the closed form expression of the predictors variance. For the former problem, a solution for computing the variance-covariance matrix for LH designs with any number of levels is presented in a particular case of correlation function, even if is one of the most common one used by the CE practitioners. For the latter, the difficulties are in the computation of the rational functions in the covariances with rational coefficients. We suggest to compute the predictor variances in closed form by a symbolic algebraic software such as CoCoA (Computations in Commutative Algebra), a freely available system for symbolic exact multivariate polynomial computation, see CoCoaTeam. Other computations related with the exponential model for covariances are done with the software R, see http://www.R-project.org/. The final result is a general methodology to analyze the MSPE efficiency of the class of interest of training sets for regular grids and covariances of the form negative exponential of the Manhattan distance.

References

Butler, N.A. (2001). Optimal and Orthogonal Latin Hypercube Designs for Computer Experiments. *Biometrika* **88**: 847-857.

CoCoATeam, CoCoA: a system for doing computations in Commutative Algebra, Version 4.7.4 2008 Available at http://cocoa.dima.unige.it.

Cressie, N. (1993). *Statistics for spatial data*, Wiley, New York

Fang, K.-T., Li R., Sudjianto, A. (2006). Design and Modelling for Computer Experiments Chapman&Hall Boca Raton, FL.

Fang, K.-T., Lin D.K., Winker, P., Zhang, Y. (2000). Uniform Design: Theory and Applications. *Technometrics* **42**: 237–248.

Krige, D.G. (1951). *A statistical approach to some mine valuations and allied problems at the Witwatersrand*, Master's thesis of the University of Witwatersrand.

McKay, M.D. Conover, W.J., and Beckman, R.J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* **21**: 239–245.

Park, J.-S. (1994) Optimal latin-hypercube designs for computer experiments, J. Statist. Plann. Inference **43:** 381-402.

Pistone G., Vicario G. (2009). *Design for Computer Experiments: comparing and generating designs in Kriging models*, In: Erto P., Statistics for Innovation. (pp. 91-102). Springer-Verlag Italia. ISBN: 978-88-470-0815.

Sachs J., Welch W.J., Mitchell T.J., Wynn H.P. (1989). Design and analysis of computer experiments. *Statistical Science*, **4**: 409-423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). The Design and Analysis of Computer Experiments. Springer-Verlag, New York.

Sasena, M. J. (2002). Flexibility and Efficiency Enhacements for Costrained Global Design Optinmization with Kriging Approximations, PhD Thesis University of Michigan.

Welch W. J., Buck, R.J., Sacks J., Wynn H. P., Mitchell T. J. and Morris, M.D. (1992). Sreening, predicting, computer experiments. *Technometrics* **34**: 15–25.