SAMSI

2008-09 Program on Algebraic Methods in Systems Biology and Statistics

Algebraic statistical models in kriging

Giovanni Pistone

DIMAT Politecnico di Torino

Transition Workshop June 18-20, 2009

Based on a joint paper with Grazia Vicario, Politecnico di Torino, to be presented at the ENBIS-EMSE2009 conference, St-Etienne FR, July 1-3, 2009.

Overview

- Kriging is an old methodology (1951), named after the inventor D. G. Krige, used to predict unobserved values in a random field.
- The underlying model is a parametric model of Gaussian type. If the space of locations is a grid, the model is an algebraic statistical model. More precisely, it is a parametric sub-model of a graphical model.
- The prediction is based on the concept of Linear Unbiased Predictor LUP and unknown parameters are estimated and plugged into the predictor formula. The final estimator is a rational function of the observed values.
- A problem of interest is the design of experiment, i.e. the choice of a training set with good performance when evaluated with respect to the Mean Squared Prediction Error MSPE criteria, or other criteria, such that minimum entropy.
- We expect the study of algebraic Gaussian models in this setting to produce interesting variations of the standard theory.
- Moreover, it is of interest to show the potential applicability of algebraic software in this area of technological applications. Namely, we show how combinatorial/geometrical properties of the training set are captured by the representation of the optimality criteria as rational functions.

- In kriging, the random variable Y_0 is predicted from observed random variables as $\widehat{Y_0} = \sum_{i=1}^{n} a_i Y_i$. Weights are estimated from a statistical model.
- Daniel G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa*, 52(6):119139, 1951
- Georges Matheron. *Traité de géostatistique appliqué*. Editions Technip, 1962
- In the 80's it was suggested to use kriging in computer experiments. A computer experiment does not have technical errors. Nonetheless a sound statistical methodology is of use. The kriging model can be considered an empirical bayesian approach to computer experiments.
- J. Sacks, W. J. Welch, T. J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments (with discussion). *Statistical Science*, 4: 409–435, 1989a
- Thomas J. Santner, Brian J. Williams, and William I. Notz. *The design and analysis of computer experiments*. Springer Series in Statistics. Springer-Verlag, New York, 2003. ISBN 0-387-95420-1
- Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and modeling for computer experiments*. Computer Science and Data Analysis Series. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-546-7; 1-58488-546-7

Best Linear Unbiased Prediction BLUE

- Let $Y_0, Y_1, \ldots, Y_n$ be Gaussian random variables with mean vector μ and covariances $\Gamma = [\Gamma_{i,j}]_{i,j=0,1,\ldots,n}$. We assume the existence of $\Gamma_{1..n}^{-1}$.
- If μ and Γ are known, the conditional expectation

$$\mathrm{E}\left(Y_0 | Y_1, \ldots, Y_n\right) = \mu_0 + \Gamma_{0,1..n}\Gamma_{1..n}^{-1}(Y_{1..n} - \mu_{1..n})$$

  is unbiased and has minimum Mean Squared Prediction Error MSPE. Such a minimum value is

$$\Gamma_{0,0} - \Gamma_{0,1..n}\Gamma_{1..n}^{-1}\Gamma_{1..n,0}$$

- Assume $\beta = \mu_i$, $i = 0, 1, \ldots, n$, unknown. An Linear Predictor LP $\widehat{Y_0} = a_0 + \sum_{i=1}^n a_n Y_n$ is unbiased if, and only if,

$$\beta = a_0 + \beta \sum_{i=1}^n a_i$$

  for all β, i.e. $a_0 = 0$ and $\sum_{i=1}^n a_i = 1$.

- $\mathcal{D} = \{1 \ldots n\}^s$ is a factorial design (lattice) with s factors and n levels on each factor. $\mathcal{F} \subset \mathcal{D}$ is a fraction of $\mathcal{D}$ (training set).
- The ordinary kriging statistical model is

$$Y_x = \beta + Z_x, \quad x \in \mathcal{D}, \quad \beta \in \mathbb{R},$$
$$Z = (Z_x)_{x \in D} \sim \text{Normal}(0, \Gamma), \quad \Gamma_{x,y} = t^{d_s(x,y)}.$$

Here, d_s is the L^1 (Manhattan) distance and $t \in]0,1[$.

- The BLUP is used to predict the value of an unobserved Y_y, $y \in \mathcal{D} \setminus \mathcal{F}$. The unknown value of the parameter t is estimated from the training set, e.g. by maximum likelihood, and plugged into the BLUE formula.
- The matrix of distances $[d(x,y)]_{x,y \in \mathcal{F}}$ is an iterated Kronecker sum of the one-dimensional distance matrix $[d_1(i,j)]_{i,j=1,\ldots,n} = [|i-j|]_{i,j=1,\ldots,n}$, so that the covariance matrix of Z is the iterated Kronecker product of the one-dimensional covariance matrix $[t^{|i-j|}]_{i,j=1,\ldots,n}$.

Example 4^2

Let us assume a 4×4 grid $\mathcal{D}$. The 1-dimensional distance matrix is

$$\text{Outer}\left(1..4, 1..4, d_1\right) = L_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix} \end{array}$$

The distance matrix on $\mathcal{D}$ is $L_2 = \text{Kronecker}\left(L_1, L_1, +\right)$:

$$\begin{array}{c} \\ 11 \\ 12 \\ 13 \\ 14 \\ 21 \\ 22 \\ 23 \\ 24 \\ 31 \\ 32 \\ 33 \\ 34 \\ 41 \\ 42 \\ 43 \\ 44 \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} 11 & 12 & 13 & 14 & 21 & 22 & 23 & 24 & 31 & 32 & 33 & 34 & 41 & 42 & 43 & 44 \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 & 2 & 3 & 4 & 5 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 3 & 4 & 4 & 3 & 4 & 5 \\ 2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 & 4 & 3 & 2 & 3 & 5 & 4 & 3 & 4 \\ 3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 & 5 & 4 & 3 & 2 & 6 & 5 & 4 & 3 \\ 1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 & 2 & 3 & 4 & 5 \\ 2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 3 & 4 \\ 3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 & 4 & 3 & 2 & 3 \\ 4 & 3 & 2 & 1 & 3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 & 5 & 4 & 3 & 2 \\ 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 & 1 & 2 & 3 & 4 \\ 3 & 2 & 3 & 4 & 2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 3 \\ 4 & 3 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 & 3 & 2 & 1 & 2 \\ 5 & 4 & 3 & 2 & 4 & 3 & 2 & 1 & 3 & 2 & 1 & 0 & 4 & 3 & 2 & 1 \\ 3 & 4 & 5 & 6 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 0 & 1 & 2 & 3 \\ 4 & 3 & 4 & 5 & 3 & 2 & 3 & 4 & 2 & 1 & 2 & 3 & 1 & 0 & 1 & 2 \\ 5 & 4 & 3 & 4 & 4 & 3 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 1 & 0 & 1 \\ 6 & 5 & 4 & 3 & 5 & 4 & 3 & 2 & 4 & 3 & 2 & 1 & 3 & 2 & 1 & 0 \end{bmatrix} \end{array}$$

The 1-dimensional covariance matrix is

$$\Gamma^1 = t^{L_1} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} t^0 & t^1 & t^2 & t^3 \\ t^1 & t^0 & t^1 & t^2 \\ t^2 & t^1 & t^0 & t^1 \\ t^3 & t^2 & t^1 & t^0 \end{bmatrix} \end{array} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 1 & t & t^2 & t^3 \\ t & 1 & t & t^2 \\ t^2 & t & 1 & t \\ t^3 & t^2 & t & 1 \end{bmatrix} \end{array}$$

The covariance matrix on $\mathcal{D}$ is $\Gamma = t^{L_2} = \text{Kronecker}\left(\Gamma^1, \Gamma^1, \times\right)$:

$$\begin{array}{c c} & \begin{array}{cccccccccccccccc} 11 & 12 & 13 & 14 & 21 & 22 & 23 & 24 & 31 & 32 & 33 & 34 & 41 & 42 & 43 & 44 \end{array} \\ \begin{array}{c} 11 \\ 12 \\ 13 \\ 14 \\ 21 \\ 22 \\ 23 \\ 24 \\ 31 \\ 32 \\ 33 \\ 34 \\ 41 \\ 42 \\ 43 \\ 44 \end{array} & \begin{bmatrix} 1 & t & t^2 & t^3 & t & t^2 & t^3 & t^4 & t^2 & t^3 & t^4 & t^5 & t^3 & t^4 & t^5 & t^6 \\ t & 1 & t & t^2 & t^2 & t & t^2 & t^3 & t^3 & t^2 & t^3 & t^4 & t^4 & t^3 & t^4 & t^5 \\ t^2 & t & 1 & t & t^3 & t^2 & t & t^2 & t^4 & t^3 & t^2 & t^3 & t^5 & t^4 & t^3 & t^4 \\ t^3 & t^2 & t & 1 & t^4 & t^3 & t^2 & t & t^5 & t^4 & t^3 & t^2 & t^6 & t^5 & t^4 & t^3 \\ t & t^2 & t^3 & t^4 & 1 & t & t^2 & t^3 & t & t^2 & t^3 & t^4 & t^2 & t^3 & t^4 & t^5 \\ t^2 & t & t^2 & t^3 & t & 1 & t & t^2 & t^2 & t & t^2 & t^3 & t^3 & t^2 & t^3 & t^4 \\ t^3 & t^2 & t & t^2 & t^2 & t & 1 & t & t^3 & t^2 & t & t^2 & t^4 & t^3 & t^2 & t^3 \\ t^4 & t^3 & t^2 & t & t^3 & t^2 & t & 1 & t^4 & t^3 & t^2 & t & t^5 & t^4 & t^3 & t^2 \\ t^2 & t^3 & t^4 & t^5 & t & t^2 & t^3 & t^4 & 1 & t & t^2 & t^3 & t & t^2 & t^3 & t^4 \\ t^3 & t^2 & t^3 & t^4 & t^2 & t & t^2 & t^3 & t & 1 & t & t^2 & t^2 & t & t^2 & t^3 \\ t^4 & t^3 & t^2 & t^3 & t^3 & t^2 & t & t^2 & t^2 & t & 1 & t & t^3 & t^2 & t & t^2 \\ t^5 & t^4 & t^3 & t^2 & t^4 & t^3 & t^2 & t & t^3 & t^2 & t & 1 & t^4 & t^3 & t^2 & t \\ t^3 & t^4 & t^5 & t^6 & t^2 & t^3 & t^4 & t^5 & t & t^2 & t^3 & t^4 & 1 & t & t^2 & t^3 \\ t^4 & t^3 & t^4 & t^5 & t^3 & t^2 & t^3 & t^4 & t^2 & t & t^2 & t^3 & t & 1 & t & t^2 \\ t^5 & t^4 & t^3 & t^4 & t^4 & t^3 & t^2 & t^3 & t^3 & t^2 & t & t^2 & t^2 & t & 1 & t \\ t^6 & t^5 & t^4 & t^3 & t^5 & t^4 & t^3 & t^2 & t^4 & t^3 & t^2 & t & t^3 & t^2 & t & 1 \end{bmatrix} \end{array}$$

Latin Hypercube Design LHD

- A Latin Hypercube Design LHD is a subset of the n^s grid $\mathcal{D}$ with n points that fully projects on each dimension. LHD were introduced in Computer Experiments CE by M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.
- In a n^s grid there are $(n!)^{s-1}$ LHD's and the generation reduces to the generation of all permutations of $1..n$. Other types of fraction are much more difficult to sample from. The computation of a BLUP from a LHD requires the inversion of the sub-matrix $\Gamma_{\mathcal{F}}$. We use CoCoA.
- One expects different LHD's perform differently with respect to the criteria of the Mean Square Prediction Error MSPE. Some LHD's, such as the diagonal, show a bad geometric coverage of the grid. However, this optimality issue requires careful study, as it was observed that in practice the choice of the LHD has little overall influence.
- A popular index is the Total Mean Squared Prediction Error, TMSPE$= \sum_{y \notin \mathcal{F}}$ MSPE(y). Other indexes are based on entropy computations. The algebraic approach should provide a practical tool to classify LHD's with respect to the TMSPE criteria.

The aim of this piece of research is investigate the algebraic features (if any) of the various optimality criteria.

- M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14:165–170, 1987 suggested to use an entropy functional of the LHD.
- Michael Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987. ISSN 0040-1706, Michael Stein. Correction: "Large sample properties of simulations using Latin hypercube sampling" [Technometrics **29** (1987), no. 2, 143–151; MR0887702 (88e:62055)]. *Technometrics*, 32(3):367, 1990. ISSN 0040-1706 has asymptotic results on the estimator variance.
- Jerome Sacks, Susannah B. Schiller, and William J. Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989b. ISSN 0040-1706 use the TMSPE.
- Boxin Tang. Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.*, 88(424):1392–1397, 1993. ISSN 0162-1459 has a special class of LHD derived from Orthogonal Arrays.
- R. A. Bates, Eva Buck, R. J. ad Riccomagno, and Henry P. Wynn. Experimental designs and observations for large systems. *Journal of the Royal Statistical Society B*, 58:77–94, 1996 consider designs generated by a translation mod n, i.e. a subgroup of order n in $\mathbb{Z}_n^s$.

LHDs of the 4^2 factorial design

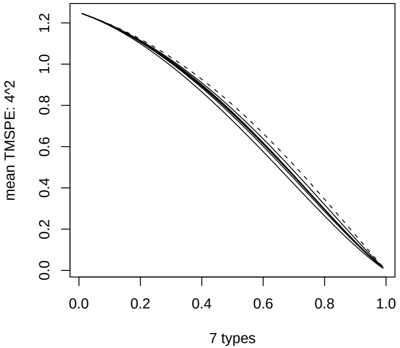
- The $4! = 24$ LHD's of a 4^2 factorial are:

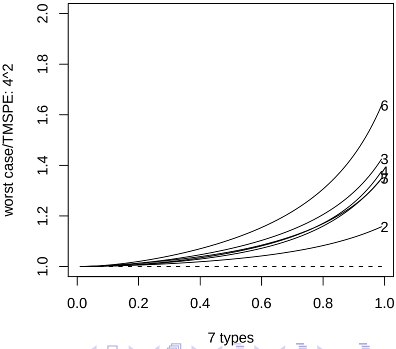
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	11	11	11	14	14	11	11	11	13	13	13	14	14	13	13	13	12	12	12	14	14	12	12	12
2	22	22	24	21	21	24	23	23	21	21	24	23	23	24	22	22	23	23	24	22	22	24	21	21
3	33	34	32	32	33	33	34	32	32	34	31	31	32	32	34	31	31	34	33	33	31	31	34	33
4	44	43	43	43	42	42	42	44	44	42	42	42	41	41	41	44	44	41	41	41	43	43	43	44

- For each of the 24 LHD, the TMSPE is a rational function. There are only 7 different rational functions (edited CoCoA output):

```
(2t^8-4t^7-12t^6+4t^5+8t^4+20t^3+12t-30)/(t^4-t^2-2)
(t^13+t^12+1/2t^11-11/2t^10-31/2t^9-19/2t^8-5/2t^7+71/2t^6+43/2t^5+71/2t^4+13t^3-27t^2-18t-30)/
     (t^7+t^6-2t^4-2t^3-4t^2-2t-2)
(1/2t^15+1/2t^14+t^13-4t^12+3/2t^11-17/2t^10-47/2t^9+9/2t^8-7t^7+48t^6+59/2t^5+49/2t^4+16t^3-35t^2-18t-30)/
     (t^8+t^7-4t^4-2t^3-5t^2-2t-2)
(1/2t^14-1/2t^12-2t^11+3t^10-4t^9-18t^8+12t^7-27/2t^6+40t^5+3/2t^4+38t^3-39t^2+12t-30)/(t^6-t^4-4t^2-2)
(t^15+t^14+t^13-7t^12-5t^11-5t^10-23t^9+13t^8+2t^7+42t^6+20t^5+20t^4+22t^3-34t^2-18t-30)/
     (t^9+t^8+2t^7-2t^6-4t^4-t^3-5t^2-2t-2)
(t^9-3t^8+2t^7+2t^6-22t^5+8t^4+28t^3+8t^2-9t-15)/(t^3-t^2-t-1)
(3t^8-2t^7+6t^6-9t^5-5/2t^4-10t^3+10t^2-3t+15/2)/(t^2+1/2)
```

- Each class has a different performance wrt the TMSPE criteria (first plot). The worst case are LHDs # 1 and 13, i.e. the two diagonals (dashed line).
- The second plot shows the relative performance wrt the worst case. The formal computation allows a precise evaluation of the behaviour near $t = 1$.
- For a given number of factors s, the difference increases with the number of levels n.

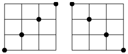




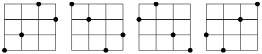
- The following table shows in the bottom line the classification of LHDs based on the TMSPE:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	11	11	11	14	14	11	11	11	13	13	13	14	14	13	13	13	12	12	12	14	14	12	12	12
2	22	22	24	21	21	24	23	23	21	21	24	23	23	24	22	22	23	23	24	22	22	24	21	21
3	33	34	32	32	33	33	34	32	32	34	31	31	32	32	34	31	31	34	33	33	31	31	34	33
4	44	43	43	43	42	42	42	44	44	42	42	42	41	41	41	44	44	41	41	41	43	43	43	44
T	1	2	3	4	3	4	3	5	3	6	7	2	1	2	3	4	3	4	3	5	3	6	7	2

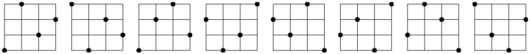
Class 1: 1,13



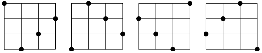
Class 2: 2,12,14,24



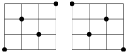
Class 3: 3,5,7,9,15,17,19,21



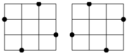
Class 4: 4,6,16,18



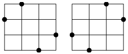
Class 5: 8,20



Class 6: 10,22



Class 7: 11,23



- Class 6 is the best one; classes 3,4,7 are essentially equivalent and worse than class 6; class 2 is second worse.
- Inspection of the pictures shows the relevance of geometric symmetries of the LHD. However, we do not have a theoretical explanation of this specific behaviour.
- Class 6 consists of U-design according B. Tang (1993).

- One among the Entropy criteria requires the computation of the determinant of the covariance matrix of each LHD. There are again types of polynomial function for the determinant (edited CoCoA output):

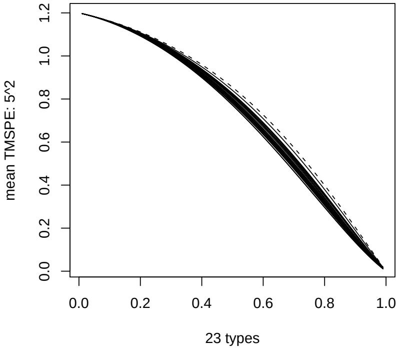
```
-t^12+3t^8-3t^4+1
-2t^12+2t^10+3t^8-2t^6-2t^4+1
t^16-4t^14+2t^12+3t^10+t^8-3t^6-t^4+1
t^16-4t^14+2t^12+4t^10-2t^8-2t^4+1
t^16-4t^14+3t^12+4t^8-4t^6-t^4+1
t^16-4t^14+8t^10-2t^8-4t^6+1
-4t^12+8t^10-3t^8-2t^4+1
```

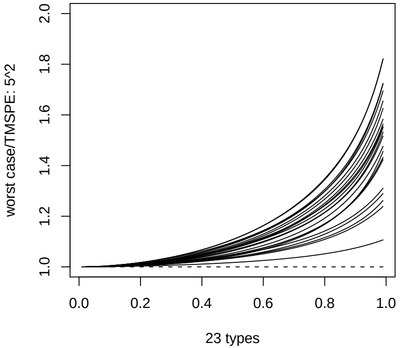
- The classification induced by the determinant is equal to the classification induced by the TMSPE:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	11	11	11	14	14	11	11	11	13	13	13	14	14	13	13	13	12	12	12	14	14	12	12	12
2	22	22	24	21	21	24	23	23	21	21	24	23	23	24	22	22	23	23	24	22	22	24	21	21
3	33	34	32	32	33	33	34	32	32	34	31	31	32	32	34	31	31	34	33	33	31	31	34	33
4	44	43	43	43	42	42	42	44	44	42	42	42	41	41	41	44	44	41	41	41	43	43	43	44
T	1	2	3	4	3	4	3	5	3	6	7	2	1	2	3	4	3	4	3	5	3	6	7	2
D	1	2	3	4	3	4	3	5	3	6	7	2	1	2	3	4	3	4	3	5	3	6	7	2

LHDs of the 5^2 factorial design

- There are $5! = 120$ LHD's of a 5^2 factorial.
- For each of the 120 LHD, the TMSPE is a rational function. There are 23 different rational functions for the TMSPE as shown below. The Determinant produces the same classification.





LHDs of the 3^3 factorial design

- There are $(3!)^2 = 36$ LHD's of a 3^3 factorial.
- For each of the 36 LHD, the TMSPE is a rational function. There are 3 different rational functions for the TMSPE as shown below. The Determinant produces the same classification.
- In this case, the TMSPE for each LHD's do not show relevant variations.