Chemometrics and Intelligent Laboratory Systems 135 (2014) 172–182

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Evolutionary Bayesian Network design for high dimensional experiments

Debora Slanzi *, Irene Poli

Department of Environmental Science, Informatics and Statistics, University Ca' Foscari, Cannaregio 873, 30121 Venice, Italy
European Centre for Living Technology, Ca' Minich, S. Marco 2940, 30124 Venice, Italy

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 14 April 2014
Accepted 20 April 2014
Available online 30 April 2014

Keywords:
Design of experiments
Evolutionary Bayesian networks
High dimensional systems
Optimisation
Vesicles self-organisation process

ABSTRACT

Laboratory experimentation is increasingly concerned with systems whose dynamical behaviour can be affected by a very large number of variables. Objectives of experimentation on such systems are generally both the optimisation of some experimental responses and efficiency of experimentation in terms of low investment of resources and low impact on the environment. Design and modelling for high dimensional systems with these objectives present hard and challenging problems, to which much current research is devoted. In this paper, we introduce a novel approach based on the evolutionary principle and Bayesian network models. This approach can discover optimum values while testing just a very limited number of experimental points. The very good performance of the approach is shown both in a simulation analysis and biochemical study concerning the emergence of new functional bio-entities.

1. Introduction

Designing experiments and modelling data in current scientific research increasingly contend with the problem of high-dimensionality. Natural systems in particular are described by a very large number of variables whose dynamical interaction leads to the emergence of a particular behaviour of the system. Understanding the organisation of such systems and identifying the relationships among variables generally involve complex laboratory experimentation. Strategies to plan experiments have to address the difficulty of deriving efficient designs and optimisation procedures testing a small set of experimental points, as each of which may involve great investment of resources and sometimes negative environmental effects. Modelling high-dimensional systems confronts also the difficulties of both building nonlinear dependence relations and estimating a number of parameters that increases rapidly with the dimension of the system. These difficulties suggest the importance of alternatives to standard experimental designs and linear regression models.

As a result of these considerations, many authors have developed strategies to achieve efficient designs by reducing the effective number of parameters through the assumption of sparsity (only a few variables are relevant in determining the dynamics of the system). Georgiou et al. [1,2], Marley et al. [3], and Sun et al. [4], among others, contributed to the construction and evaluation of the large class of supersaturated designs (SSDs), which have been derived as a class of fractional factorial designs for systems with a large number of variables and few experimental points. Building on the pioneering work of Booth and Cox [5], there have been significant developments initially for problems with two-level variables and later for multi-level and mixed-level variables [2]. Based on the classical assumption of the linear main effects model with Gaussian experimental errors, the construction of SSDs is realised through different algorithms and computer search strategies to find optimal designs with respect to a measure built on the information matrix of the model [3].

Another class of designs addressing high dimensionality is Exchange Algorithms and their developments [6–9]. These designs are based on the exchange between the selected design points and points in a candidate list. The exchange is performed with the aim of identifying the design that satisfies a measure of optimality; in these approaches the exchange is realised at any iteration of the algorithm generally for increasing the determinant of the information matrix.

Optimal design procedures based on the information matrix, as SSDs and Exchange Algorithms, are adequate for several problems, but they require a prior knowledge of the form of the response function (linear model) which frequently is not available. Model-robust experimental designs based on Exchange Algorithms have been recently proposed [9] to respond to this issue. These designs do not assume a single model form but allow for a set of user-specified models, and the design is derived by maximising the product of the determinant of the information matrices associated with each of the suggested models.

A different way to address the problem of designing experimentation for high dimensional systems is to use computer experiments [10,11]. In computer experiments, rather than conducting laboratory experimentation or making field observations, simulators based on mathematical models are constructed to study how the model behaves under relevant variables and conditions. Physical processes can be simulated and the simulation code can serve as an efficient mode for

* Corresponding author. Tel.: +39 3469707575.
E-mail addresses: debora.slanzi@unive.it (D. Slanzi), irenpoli@unive.it (I. Poli).

http://dx.doi.org/10.1016/j.chemolab.2014.04.013
0169-7439/

exploring the properties of the process [12–15]. This is becoming a popular approach and it is applied in different research fields; however high dimensionality makes this approach hard to use, since it requires complex models that may be prohibitively expensive to simulate. In designing computer simulations, Latin hypercube sampling has been proposed [16,17]. These designs generally provide uniform samples for the marginal distribution of the variables and have achieved successful results in a set of research studies. Recent contributions, based on these developments, have proposed to design high dimensional experiments by adopting search procedures built on evolutionary approaches, such as genetic algorithms and particle swarm optimisation techniques [18–22].

In this paper, we propose a design strategy that is developed according to the evolutionary approach. Our strategy does not involve an a priori choice of the experimental design but it evolves the design through a number of experimental generations, moving in different areas of the search space. Each generation of planned experiments is achieved by combining evolutionary principles and inference from statistical models. In this procedure we combine the ability of the evolutionary approach to intelligently navigate the search space with the capacity of statistical models to uncover hidden information. The common practice of a priori choice of the experimental points for high dimensional problems may in fact be inappropriate for a premature and possible misleading selection of the experimental points. More specifically, our approach is based on the evolution of probabilistic graphical models, PGMs [23,24]. We focus on a particular class of PGMs, that is the class of Bayesian network models, where nodes in the graph correspond to random variables and arcs between nodes describe the dependence structure that may characterise the set of variables on which we then develop statistical inference.

We study this Evolutionary Bayesian Network design (EBN-design) both in a simulation analysis and laboratory biochemical experimentation concerning the self-organisation of amphiphilic molecules. In both our studies, EBN-design exhibits excellent performance in optimising the response of the systems, in comparison to other common experimental approaches. The paper is organised as follows. Section 2 introduces the design for optimisation and presents the Evolutionary Bayesian Network approach to design high dimensional experiments. Section 3 presents a simulation study to test the efficiency of EBN-design compared with common alternative design of experiment approaches. Section 4 describes a biochemical study concerning the self-organising process of amphiphilic molecules, and Section 5 presents some concluding remarks.

2. Design for optimisation

An optimisation problem can be described in its general structure as follows.

Let S be a subset of the Euclidean space $\mathbb{R}^d$ and f be a real-valued function on S. Let $\mathbf{x} = (x_1,\ldots,x_d)$ be the set of variables defined in S, and y be the response variable, $y = f(x_1,\ldots,x_d)$. The optimisation problem consists in finding $\mathbf{x}^*$ in S such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in S$ (in maximisation problems). The inferential problem concerns the form of f. Experimentation provides the data to construct a function $\hat{f}(x_1,\ldots,x_d)$ that can serve as an approximation to $f(x_1,\ldots,x_d)$ over the domain S of interest. Data from experimentation are collected according to a chosen design ξ_n, consisting of a set of experimental points,

$$\xi_n = (\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_n)$$

where each experimental point x_k, $k = 1,\ldots n$, is a d-dimensional vector, whose values are the levels of the variables. The set of data from experimentation, namely $(\mathbf{X}, \mathbf{y})$ with $\mathbf{X}$ an $(n \times d)$-matrix and $\mathbf{y}$ an n-vector, represents then the evidence for inferring the dependence relations among variables, i.e. the form of f, and for identifying the design points that give the optimum value of the response variable.

In classical design theory, a linear regression model for f is selected to describe the dependence relation of y on $x_1,\ldots,x_d$, that in matrix notation can be represented as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

where β are the unknown parameters of the model and $\mathbf{e}$ is an additive stochastic component with $E(\mathbf{e}) = 0$ and covariance matrix $Var(\mathbf{e}) = \sigma^2\mathbf{I}_n$. The estimates of the parameters, under the Least Squares approach, take the form $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with $Cov\left(\hat{\beta}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and the information matrix associated to the particular n-points design ξ_n is defined as $\mathbf{I}(\xi_n) = (\mathbf{X}'\mathbf{X})/\sigma^2 = \left[Cov\left(\hat{\beta}\right)\right]^{-1}$.

To select an efficient design ξ_n, the theory of optimal design [25–27] prescribes the choice of n experimental points $\mathbf{x}_k$, $k = 1,\ldots,n$, such that some criterion $\phi(\mathbf{I}(\xi_n))$, based on the information matrix, is optimised. The most common criterion is D-optimality, for which $\phi(\mathbf{I}(\xi_n)) = |\mathbf{I}(\xi_n)|$. This approach requires the knowledge of the precise form of f, assuming a linear model. However, the assumption of linearity is difficult to justify when a large number of variables characterise the system and the form f is unknown, or nonlinear, as in complex high dimensional systems.

Our contribution to address this problem is to adopt evolution as a paradigm to build a design strategy and drive the evolution with Bayesian network models.

2.1. Evolving the design with models

The key idea of evolutionary design is to derive a small set of informative experimental points adopting the principle of evolution: the design is evolved through generations of selected experimental points according to a particular function measuring the goodness of the design in reaching the objective of optimisation. The design is then evolved in a sequence of small sub-designs achieved with some probabilistic rules that emulate the laws of genetic evolution [18,28–30,20].

In building the evolutionary design, we create first a small low-dimensional set of experimental points, $\xi_{n_1} = \mathbf{X}_1$ (the design D_1 at time 1), selecting variables and variable levels at random. The random initial design, as a random sample from the population of all possible experimental points, is an intriguing selection criterion since it allows the exploration of the search space in areas not anticipated by prior knowledge but where relevant information may reside. The first population of experimental points is then tested in laboratory (or evaluated in simulation) and generates the first set of data consisting of different combinations of factor levels and their corresponding experimental response $(\mathbf{X}_1, \mathbf{y}_1)$. This initial dataset is small and each experimental point is low dimensional, but the random selection of the choice allows the system to explore through the many dimensions of the whole design space.

In a simple evolutionary design, as in Genetic Algorithm design [31,18,28,20,32], a set of probabilistic transition rules (based on genetic operators) is defined and applied to guide the evolutionary process towards the next generation of candidate experimental points. Probabilistic selection is the most important element in the evolving design. Adopting proportional, truncated or tournament selection approaches, it leads to the construction of a new small population of experimental points selected according to their expected capacity to provide high quality responses. Other operators, such as recombination and mutation, contribute to the evolutionary design, exchanging and changing elements of the experimental points to improve the quality of the compositions. These operators lead the evolution from the initial design to a sequence of successive designs with higher quality and more precise responses. The class of GA-designs has been very successful in addressing search problems in high dimensional spaces and in a variety of research fields.

Another class of evolutionary procedures is the Particle Swarm Optimisation (PSO) [33,22] where each experimental point is regarded as a particle of a swarm. In this procedure particles move in the search space