

Application of Robust Statistical Optimal Designs for a Growth Rate Model in Predictive Food Microbiology

Jean-Pierre Gauchi¹ and Jean-Christophe Augustin²

^{1,3}Department of Applied Mathematics and Computing (UR0341MIA)
INRA, Jouy-en-Josas, France, F78352

²Veterinary School of Alfort, Alfort, France, F94700

³Corresponding author: Jean-Pierre Gauchi.

e-mail: jean-pierre.gauchi@jouy.inra.fr

Abstract

Several statistical predictive models have been used in the food microbiology field in recent years to describe the growth or the inactivation of bacterial populations. These dynamic primary models (PMs) are nonlinear relative to their parameters. In addition, these parameters can themselves be modelled by nonlinear secondary models (SMs) as functions of environmental factors (EFs) such as temperature, pH, water activity, nitrites and phenol concentrations, etc. For example, the μ_{\max} parameter (representing the maximum growth rate of the PMs) can be modelled by several SMs. Typically, one efficient class of SMs deals with the nonlinear interactions between the EFs. A major objective for the microbiology community is to have the model that will lead to the best parameter estimates for the purpose of ensuring food safety, notably the minimal temperature and pH levels for which the growth of bacterial populations start. However, it remains difficult to obtain adequately accurate estimates because of the high variability of microbiological data. Obviously, some optimal experiments must be conducted in order to reach this objective. In this paper, for a published specific model, we propose an experimental strategy based on the use of some local and robust statistical optimal designs, followed by computer simulations, to compare the different designs.

Key Words: D and X-optimal designs, bayesian D and X-optimal designs, accurate parameter estimation, food safety.

1. Introduction

Nowdays, the mathematical and statistical models of the predictive microbiology field are essential tools for food safety. Therefore, the parameter estimates of these models must be of good quality – typically with parameter estimates of small bias and small mean square error (MSE) – and the estimated model must lead to good predictions. In this paper we focus on a very specific nonlinear model, a so-called secondary model (SM) of the growth rate, referred to as μ_{\max} (Augustin and Carlier, 2000), the AC model, and on accurate parameter estimates of this AC model. In this publication the AC model is a function of only two experimental factors, the temperature, T , and the pH , keeping constant the other environmental factors (a_w effect, CO_2 effect, sodium nitrite effect, ...). Since performing real microbiological experiments is cost and labor intensive, small-size optimal experimental designs are needed. Moreover, since a high variability – not always uniform over the whole experimental domain – of microbiological data characterises this field, determining robust designs, relatively to a misspecification of the θ_0 *a priori* parameter value and also relatively to an overestimation or an underestimation of the experimental variance, is a very crucial goal. Therefore we propose two classes of well-known robust optimal designs: (i) the first class, derived from the D-optimality criterion, is based on a well-known criterion that

is the maximisation, relatively to the design factors (here T and pH), of the expectation of the log determinant of the Fisher information matrix, that is a bayesian approach. See details about this criterion for instance in Atkinson and Donev (1992); (ii) the second class, the X-optimality criteria family, is based on the minimisation of the expectation of an exact parametric confidence region, that includes also a bayesian approach (Vila and Gauchi, 2007). Until now, these types of designs, although already well defined and described in the literature, have never been used in the food microbiology field. In this publication we give some designs based on these criteria for the AC model, and we compare the parametrical absolute bias (PAB) and parametrical MSE (PMSE) estimates by means of a simulation study, only considering here the misspecification relatively to θ_0 . Many other aspects about the simulation results will be find in a publication to come.

2. Description of the model

The AC model takes into account interactions between environmental factors. Its function (limited here to the T and pH factors), based on p ($= 7$) parameters is:

$$\mu_{\max} = \mu_{opt} \times CM_2(T) \times CM_1(pH) \times \xi(T, pH)$$

with :

- $CM_2(T) = 0$ if $T \leq T_{\min}$ and

$$CM_2(T) = \frac{(T - T_{\max})(T - T_{\min})^2}{(T_{opt} - T_{\min}) \{ (T_{opt} - T_{\min})(T - T_{opt}) - (T_{opt} - T_{\max}) [T_{opt} + T_{\min} - 2T] \}}$$

if $T_{\min} < T < T_{\max}$.

- $CM_1(pH) = 0$ if $pH \leq pH_{\min}$ and

$$CM_1(pH) = \frac{(pH - pH_{\max})(pH - pH_{\min})}{(pH_{opt} - pH_{\min})(pH - pH_{opt}) - (pH_{opt} - pH_{\max})(pH_{\min} - pH)}$$

if $pH_{\min} < pH < pH_{\max}$.

- and

$$\xi(T, pH) = \begin{cases} 1 & \text{if } \psi \leq 0.5 \\ 2(1 - \psi) & \text{if } 0.5 < \psi < 1 \\ 0 & \text{if } \psi \geq 1 \end{cases}$$

where

$$\psi = \frac{\left(\frac{T_{opt}-T}{T_{opt}-T_{\min}}\right)^3}{2\left(1 - \left(\frac{pH_{opt}-pH}{pH_{opt}-pH_{\min}}\right)^3\right)} + \frac{\left(\frac{pH_{opt}-pH}{pH_{opt}-pH_{\min}}\right)^3}{2\left(1 - \left(\frac{T_{opt}-T}{T_{opt}-T_{\min}}\right)^3\right)}$$

The vectorial parameter to estimate has seven elements :

$$\theta = \{ \mu_{opt}, T_{\min}, T_{\max}, T_{opt}, pH_{\min}, pH_{\max}, pH_{opt} \}$$

Initial guesses, θ_0 , and/or their ranges, Δ_θ , are needed for the construction of optimal designs, according to the criteria used. They are given in Table 1, taken in Augustin et al. (2005). For T_{\max} , T_{opt} , pH_{\max} , pH_{opt} , the lower and upper

bounds of Δ_θ were built as $\theta_0 \pm C\theta_0$ where $0 < C < 1$. For some optimal design criteria C took the typical value of 0.1. For T_{\min} , pH_{\min} , μ_{opt} , their lower and upper bounds of Δ_θ bounds were defined by $\theta_0 \pm SD$, where SD is the standard deviation (of the sample) also given in Augustin et al. (2005). A crucial goal is to obtain accurate estimates for T_{\min} , pH_{\min} , typically with small PAB and PMSE estimates.

Parameter	θ_0 and lower and upper bounds of Δ_θ
μ_{opt}	[0.969 ← 1.154 → 1.339]
T_{\min}	[-2.44 ← -0.95 → 0.54]
T_{\max}	[40.95 ← 45.5 → 50.05]
T_{opt}	[33.30 ← 37.0 → 40.70]
pH_{\min}	[4.15 ← 4.44 → 4.73]
pH_{\max}	[8.64 ← 9.6 → 10.56]
pH_{opt}	[6.39 ← 7.1 → 7.81]

Table 1. Initial guesses, θ_0 , and lower and upper bounds of Δ_θ , for the seven parameters. In this Table $C = 0.1$ for illustration.

We can see in Figure 1 the graph and the isocontours of the AC model on the $Temp \times pH$ experimental domain, for θ_0 of Table 1.

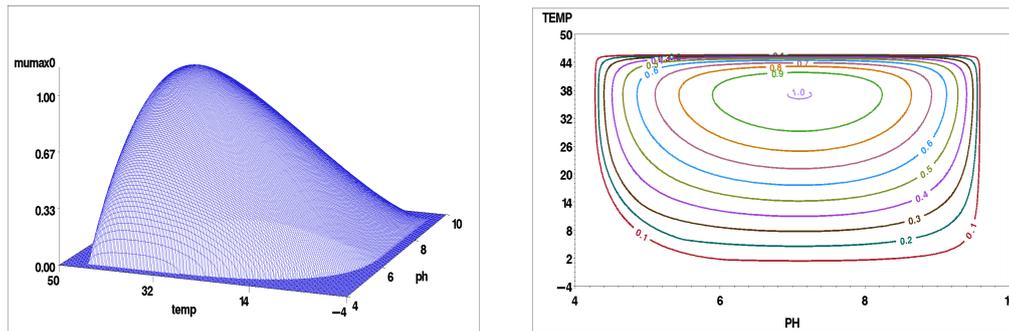


Fig.1. Graph and Isocontours of the AC model on the $Temp \times pH$ experimental domain.

3. The optimal designs

We constructed local optimal designs, $D(\theta_0)$ -optimal designs and $X(\theta_0)$ -optimal designs. They are named local because they only depend on one θ_0 value. Notice that, in a practical context, if a θ_0 good guess of the unknown θ is available then it seems reasonable to use local $D(\theta_0)$ -optimal designs based on such guess. Otherwise, it is always useful to compute a local $D(\theta_0)$ -optimal design because it is a natural initial design for the iterative procedures involved in the computation of a bayesian D-optimal design (see subsection 3.2) or an $X(\theta_0)$ -optimal design.

3.1. Local approach

Local D-optimal design

A discrete or continuous (non local) D-optimal design is strictly defined in the case of linear models, typically the polynomial models. It is obtained by the maximisation, relatively to the experimental factors, of the information matrix determinant, $|X^T X|$, where X is the model matrix. Since, in the case of nonlinear models this matrix does not longer exist, the common use is to replace it by the $J^T(\theta)J(\theta)$ matrix, where $J(\theta)$ is the jacobian matrix of the nonlinear model,

depending on the unknown θ . It is computed for $\theta = \theta_0$. Hence, for nonlinear models, a D-optimal design is only a first-order approximation criterion, generally referred to as $D(\theta_0)$. The OPTTEX procedure of the SAS/QC software was used for finding a discrete $D_7(\theta_0)$ -optimal design (the number 7 indicates the support point number of the design). We checked, thanks to the computation of the D-optimal continuous measure, that its D-efficiency was of 100% (the corresponding graph of this measure will shown in the talk and in the future publication). We can see in Figure 2 the seven, N_S , support points of this $D_7(\theta_0)$. Moreover, we checked also that the all experiment repetitions must be made only on these seven support points, thanks to the checking of the sufficient and quasi-necessary condition of Vila (1991).

Local X-optimal design

The second local optimal design we used was an $X(\theta_0)$ -optimal design. The computation of a local $X(\theta_0)$ -optimal design is a much more heavy computing task than that of a $D(\theta_0)$ -optimal design. It was computed by an iterative procedure based on alternating phases of (deterministic or Monte Carlo type) integration and deterministic (or stochastic) minimisation (see Vila and Gauchi, 2007). We can see in Figure 2 the seven support points of an $X_7(\theta_0)$.

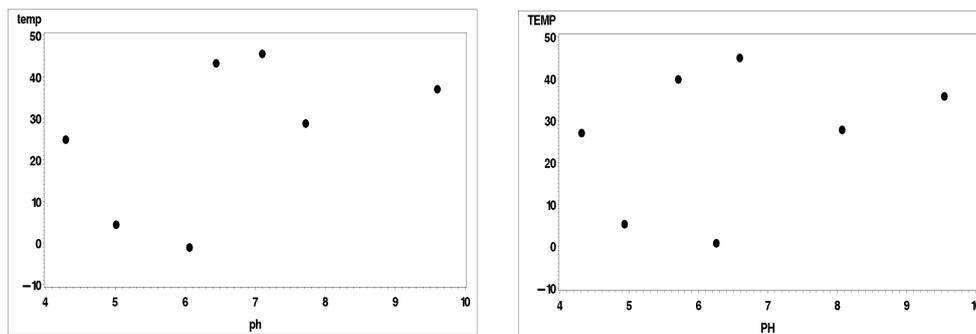


Fig.2. The seven support points of the local $D(\theta_0)$ -optimal (left graph) and $X(\theta_0)$ -optimal (right graph) designs located on the $Temp \times pH$ experimental domain.

3.2. Robust approach

There are several robust optimal design types characterised by the taking into account of an *a priori* probability density function for the unknown θ parameter, it is a bayesian framework. In this study we used the maximisation of the expectation of the logarithm of $|J^T(\theta)J(\theta)|$, its corresponding design was referred to as an ELD-optimal design (see Atkinson and Donev, 1992). We use also another robust approach based on bayesian X-optimal designs, referred to as X' -optimal designs. We computed these two robust optimal designs along a multivariate gaussian probability density function (pdf) for θ centered on θ_0 and with a diagonal variance-covariance (7×7) -matrix where the diagonal terms were $\sigma_{\theta_j}^2 = ((|\theta_{0j}| + (\delta/100) \times |\theta_{0j}|)/2)^2$. δ took the 5%, 10%, or 20% values. We can see in Figure 3 the seven support points of the gaussian ELD_7 , and the seven support points of the gaussian X'_7 , both for $\delta = 20\%$.

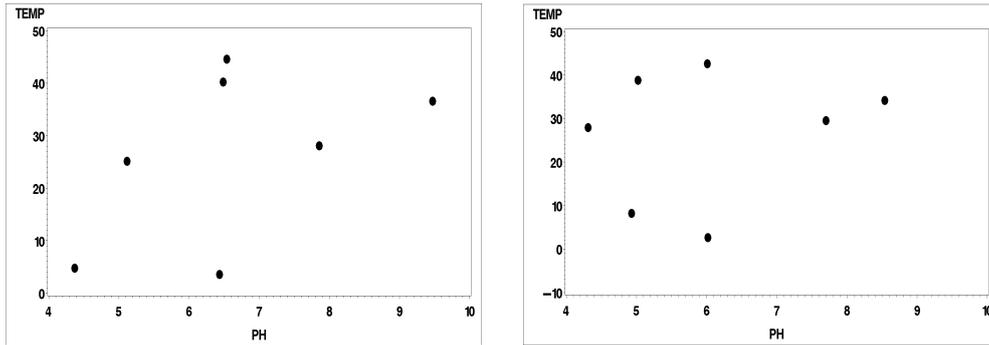


Fig.3. Coordinates of the $ELD_7(\delta = 20\%)$ -optimal design (left graph) and the $X'_7(\delta = 20\%)$ -optimal design (right graph) support points on the $Temp \times pH$ experimental domain.

4. Simulation results

For each design, the model response was simulated on each support point with a gaussian noise of a variance corresponding to 20% of the response value. In these conditions five repetitions of one thousand response sets were obtained. For assessing the respective robustness level of the designs, relatively to a misspecification of θ_0 , 50 random initial values around the θ_0 value used in the construction of the design were tested. An uncertainty of 20% on θ_0 was chosen for generating these 50 values. The robustness comparison can be quantified by several statistics, notably the PAB and PMSE estimates. The robustness, relatively to a misspecification of θ_0 , was assumed to be better if these PAB's and PMSE's were smaller. They are given in Tables 2 and 3.

Design	T_{\min}	T_{\max}	T_{opt}	pH_{\min}	pH_{\max}	pH_{opt}	μ_{opt}
$D_7(\theta_0)$	0.1200	0.0205	0.0859	0.0002	0.0323	0.0147	0.0106
$X_7(\theta_0)$	0.0201	0.1243	0.0493	0.0024	0.0186	0.0055	0.0119
ELD_7	0.0332	0.0535	0.0548	0.0102	0.0038	0.0009	0.0077
X'_7	0.0374	0.0250	0.0620	0.0124	0.0199	0.0092	0.0074

Table 2. PAB estimates obtained with four different designs: the local optimal designs are $D_7(\theta_0)$ and $X_7(\theta_0)$, and the bayesian optimal designs are ELD_7 and X'_7 . For these four designs we took $\delta = 20\%$.

We can observe in Table 2 that the PAB estimates are globally smaller with the bayesian X'_7 design and therefore this design is the globally most robust design, relatively to the bias.

Design	T_{\min}	T_{\max}	T_{opt}	pH_{\min}	pH_{\max}	pH_{opt}	μ_{opt}
$D_7(\theta_0)$	0.5336	0.3118	0.4996	0.0036	0.0455	0.0279	0.0039
$X_7(\theta_0)$	0.0580	0.8930	0.6870	0.0027	0.0330	0.0139	0.0032
ELD_7	0.2030	0.8770	0.5870	0.0260	0.0360	0.0270	0.0036
X'_7	0.0460	0.4920	0.4390	0.0213	0.0238	0.0146	0.0027

Table 3. PMSE estimates obtained with four different designs: for the bayesian optimal designs, ELD_7 and X'_7 , we took $\delta = 20\%$.

We can observe in Table 3 the following behaviors relatively to the PMSE's: (i) the specific $X_7(\theta_0)$ design, even though it is local, is more robust than the (local) $D_7(\theta_0)$ design, notably relatively to the crucial T_{\min} and pH_{\min} parameters; (ii)

the bayesian ELD_7 design is globally not really better than the $D_7(\theta_0)$ design; (iii) the bayesian design X'_7 appears to be globally the most robust. Otherwise, we checked that the advantage of the bayesian designs X' type was confirmed or improved if $N_S > 7$, the detailed results about this aspect will be given in a publication to come.

5. Conclusions

According to the large observed variation in the PAB and PMSE estimates, the great influence of the design for this SM is confirmed, and therefore the choice of an adequate optimal design is a crucial choice. The next step of this study will be to undertake real experiments at the laboratory for confirming the advantage of these small-size ($N_S = p$), and robust, designs over the usual (naive) factorial design encountered in this field, that needs at least the redhibitory number of 13 levels for each factor (we checked that a smaller level number lead to a non-convergent minimisation of the parameter estimation procedure). We advocate that a significant improvement could be obtained in food safety by using such efficient tools. At last, notice that it is an heavy computing task to determine these designs (typically in a bayesian approach) for a so sophisticated (nonlinear) model, where seven parameters have to be estimated. To our knowledge no such study was published elsewhere.

Acknowledgements The support of the French National Agency (ANR) through the "ALID/Greenswimmers" project, and the Modelisation Group, MEXICO, of the MIA department of the National Institute for Agronomical Research (INRA), are gratefully acknowledged. They partly supported this work.

References

- Atkinson, A.C., Donev, A.N. (1992) *Optimum experimental designs*, Clarendon Press, Oxford, UK.
- Augustin, J.C., Carlier, V. (2000) "Modelling the growth rate of *Listeria monocytogenes* with a multiplicative type model including interactions between environmental factors", *International Journal of Food Microbiology*, 56, 1, 53–70.
- Augustin, J.C, Zuliani, V., Cornu, M., Guillier, L. (2005) "Growth rate and growth probability of *Listeria monocytogenes* in dairy, meat and seafood products in suboptimal conditions", *Journal of Applied Microbiology*, 99, 1019-1042.
- Vila, J.P. (1991) "Local optimality of replications from a minimal D-optimal design in regression: a sufficient and a quasi-necessary condition", *Journal of Statistical Planning and Inference*, 29, 261-277.
- Vila, J.-P., Gauchi, J.-P. (2007) "Optimal designs based on exact confidence regions for parameter estimation of a nonlinear regression model", *Journal of Statistical Planning and Inference*, 137, 9, 2935-2953.