$D-$Optimal designs for Antoine's Equation: with homoscedastic and heteroscedastic response

Carlos de la Calle-Arroyo[†], Jesús López-Fidalgo[⋆], and Licesio J. Rodríguez-Aragón[†]

[†]Universidad de Castilla-La Mancha, Escuela de Ingeniería Industrial y Aeroespacial de Toledo. Instituto de Matemática Aplicada a la Ciencia y la Ingeniería
[⋆]Universidad de Navarra, DATAI

March 2021

Abstract

Vapor pressure is a temperature-dependent characteristic of pure liquids, and also of their mixtures. This thermodynamic property can be characterized through a wide range of models. Antoine's equation stands out among them for its simplicity and precision. Its parameters are estimated via maximum likelihood with experimental data. Once the parameters of the equation have been estimated, vapor pressures between known values of the curve can be interpolated. Other physical properties such as heat of vaporization can be predicted as well.

The probability distribution of a physical phenomenon is often hard to know in advance, as it depends on the phenomenon itself as well as the procedures to carry on the experiments and the measurements. Hence, assuming a probability distribution for such events has to be done with caution, as it affects the Fisher Information Matrix and consequently the optimal designs. This work presents $D-$optimal designs to estimate the unknown parameters of the Antoine's equation as accurately as possible for homoscedastic and heteroscedastic normal distribution of the response. In both cases, the aim is to improve the precision of inferences using $D-$optimality criterion.

Finally, the effects of common transformations towards a partial linearisation of either models on the optimal designs are explored. Experimenters should have in mind that fitting those models may have implications that need to be considered when using these procedures.

Keywords— Heteroscedasticity, D-optimality, Logarithmic Transformation, Efficiency

1 Antoine's Equation

The influence of temperature on the vapor pressure of pure liquids and their mixtures is derived from a class of semi-empirical equations known as Antoine's equation. This equation was developed and presented in 1888 by Louis Charles Antoine, $P(T) = \eta(T,\theta) + \varepsilon = 10^{a-\frac{b}{c+T}} + \varepsilon$. It represents the non-linear thermodynamic relationship between equilibrium vapor pressure, P, and temperature, T [3]. The unknown parameters $\theta = (a,b,c)^t$ are numerical constants related to the enthalpy and entropy of vaporization. The parameters vary for different pure substances.

1.1 Homoscedastic model

The first considerations regarding this model are done with the normal homoscedastic distribution. This is the baseline and common suspect when no information about the distribution or enough experimental data are available. The model to be considered with this assumption would be

$$P(T) = \eta(T,\theta) + \varepsilon = 10^{a-\frac{b}{c+T}} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma^2), \tag{1}$$

which, with the usual first order Taylor expansion, commonly used to work with non-linear models in optimal experimental design, would have the one-point Fisher Information Matrix (FIM)

$$M_o(\xi) = \frac{1}{\sigma^2}\frac{\partial\eta(T,\theta)}{\partial\theta}\frac{\partial\eta(T,\theta)}{\partial\theta^T}.$$

Often, in experimental procedures [1], logarithms are taken to obtain a linearly separable problem. To have an actual model, the expectation of the logarithm of the response is approximated by the logarithm of its mean, i.e., the logarithm of the model

$$\log[P(T)] = \mathbb{E}(\log[P(T)]) + \varepsilon \approx log[\eta(T,\theta)] + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma^2/\eta(T,\theta)^2), \tag{2}$$

and in an equivalent way this produces the one-point FIM

$$M_{lo}(\xi) = \frac{\partial\eta(T,\theta)}{\partial\theta}\frac{\partial\eta(T,\theta)}{\partial\theta^T}\left(\frac{2}{\eta(T,\theta)^2} + \frac{1}{\sigma^2}\right).$$

1.2 Heteroscedastic model

Empirical experiences suggest that the errors are, indeed, normal. However, it is remarked that the *relative error* is deemed constant, instead of the *absolute error* [1]. This is equivalent to error proportional to the response. The implications are that the response variance on this scenario is heteroscedastic, an hence the model results

$$P(T) = \eta(T,\theta) + \varepsilon = 10^{a-\frac{b}{c+T}} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\lambda^2\eta(T,\theta)^2), \tag{3}$$

with a FIM expression given by

$$M_e(\xi) = \frac{\partial\eta(T,\theta)}{\partial\theta}\frac{\partial\eta(T,\theta)}{\partial\theta^T}\left(\frac{2}{\eta(T,\theta)^2} + \frac{1}{\eta(T,\theta)^2\lambda^2}\right) = \frac{2\lambda^2+1}{\lambda^2}\frac{\partial\eta(T,\theta)}{\partial\theta}\frac{\partial\eta(T,\theta)}{\partial\theta^T}\frac{1}{\eta(T,\theta)^2}.$$

As in the homoscedastic case, logarithm can be taken, with an analogous approximation

$$\log[P(T)] = \mathbb{E}(\log[P(T)]) + \varepsilon \approx log[\eta(T,\theta)] + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\lambda^2), \tag{4}$$

which has a FIM proportional to the heteroscedastic model

$$M_{le}(\xi) = \frac{\partial \eta(T,\theta)}{\partial \theta} \frac{\partial \eta(T,\theta)}{\partial \theta^T} \frac{1}{\eta(T,\theta)^2},$$

and therefore the optimal designs for both the original heteroscedastic model and logarithmically transformed one are the same.

2 $D-$Optimal Designs

Regarding $D-$optimality there are some theoretical results that can be proved for these models. Considering 1.1 or 1.2 leads to sensitive changes either in the estimators or the optimal designs.

In the homoscedastic case, given by Equation (1), the following result holds:

Theorem 1. *The D-optimal design for Antoine's Equation homoscedastic model is supported at three points, one of them in the boundary of $\mathcal{X}$.*

The analytical expression of the $D-$optimal design has been calculated. Depending on the values of the unknown parameters b and c and the superior extreme of the design space $\mathcal{X}$ there are two options.

1. Two interior points of the design space and the superior extreme

$$\xi_D^\star = \left\{ \begin{array}{ccc} T_1^\star & T_2^\star & T_{max} \\ 1/3 & 1/3 & 1/3 \end{array} \right\},$$

2. Both extremes of the design space and an interior point

$$\xi_D^\star = \left\{ \begin{array}{ccc} T_{min} & T_2^{\star\star} & T_{max} \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

Analytical expressions for both cases have been obtained.

While for the heteroscesdastic case, given by Equation (3), a similar result holds:

Theorem 2. *The D-optimal design for Antoine's Equation heteroscedastic model is supported at three points. Both of the extremes of $\mathcal{X}$ are support points of the design.*

The inner support point of the $D-$optimal design is

$$T_2 = \frac{cT_{max} + cT_{min} + 2T_{max}T_{min}}{2c + T_{max} + T_{min}}.$$

3 Example: Water in liquid state

Considering the case of water at liquid state, that is, $\mathcal{X} = [1, 100]$, and best guesses of the unknown parameters $a = 8.07131$, $b = 1730.63$ and $c = 233.426$ [2], the $D-$optimal designs have been computed. For the homoscedastic model, Equation (1), the $D-$optimal design is

$$\xi_o^\star = \left\{ \begin{array}{ccc} 44.90 & 83.20 & 100 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}, \tag{5}$$

and its logarithm model's $D-$optimal design, Equation (2), with a value of $\sigma^2 = 9$, is

$$\xi^{\star}_{lo} = \left\{ \begin{array}{ccc} 44.89 & 83.20 & 100 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}. \tag{6}$$

The heteroscedastic model, displayed in Equation (3) has the same $D-$optimal design as its logarithm, Equation (4), which is

$$\xi^{\star}_{e} = \left\{ \begin{array}{ccc} 1 & 41.76 & 100 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}. \tag{7}$$

The goodness of a design can be measured with their efficiency. This can be interpreted as how much information, with the same number of replications, a design gives compared to the other. The expression of the efficiency for $D-$optimality is $\text{eff}_D(\xi, \xi^{\star}) = (|M(\xi)|/|M(\xi^{\star})|)^{1/m}$, with m the number of unknown parameters.

The cross efficiencies table of the designs, that allows a comparison of them, is

Efficiency	$\xi^{\star}_{o}$	$\xi^{\star}_{lo}$	$\xi^{\star}_{e}$
$\text{eff}(\cdot, \xi^{\star}_{o})$	100	99.9	25.4
$\text{eff}(\cdot, \xi^{\star}_{lo})$	99.9	100	30.7
$\text{eff}(\cdot, \xi^{\star}_{e})$	18.7	18.7	100

Table 1: Cross efficiencies of the $D-$optimal designs for the homoscedastic, heteroscedastic and logarithm models of the Antoine's Equation.

4 Conclusions

Sometimes overlooked, probability distributions are a capital part of Optimal Design of Experiments. As it is showcased in this work, a different variance structure of the response can substantially change the optimal design for the same model, with a resulting impoverished efficiency when there is a misspecification of the probability distribution of the response. When possible, different reasonable assumptions must be taken to account by the scientists as well as make comparisons.

There are also strategies to find designs with a compromise between different criteria or, as in this case, different probability distribution. These allow to find robust designs with a reasonable efficiency for the different scenarios considered. Future work will delve into how to develop and apply the strategies to this and other real problems.

References

[1] Brozena, A., Davidson, C.E, Ben-David, A., Shindler, B., Tevault, D.E. (2016) Vapor Pressure Data Analysis and Statistics. Tech. rep., U.S. Army Edgewood Chemical Biological Center.

[2] Dortmund Data Bank (2021) www.ddbst.com.

[3] Wisniak, J. (2001) Historical Development of the Vapor Pressure Equation from Dalton to Antoine J. Phase Eq., 22:622–630.