A Note on Bayesian Design for the Normal Linear Model with Unknown Error Variance

Isabella Verdinelli

Department of Statistics Carnegie Mellon University, Pittsburgh, PA

isabella@stat.cmu.edu

Summary

The Bayesian theory of optimal experimental design for the normal linear model has been developed under the assumption of known variance. The insensitivity of specific design criteria to prior assumptions on the variance distribution has been demonstrated in special cases, but a general result showing the way in which Bayesian optimal designs are affected by prior information on the variance is lacking. This note proves that Bayesian designs are insensitive to information about the variance in a more general way than previously thought.

Some Key Words: Bayesian design criteria; Design for prediction; Linear model; Expected Utility Function.

1. Introduction

Bayesian optimal experimental designs for the normal linear model when the sampling variance is unknown have been examined by several authors; see for example Sinha (1970), Guttman (1971), Brooks (1974, 1976, 1977) and Smith & Verdinelli (1980). The solution proposed in these papers is to replace the unknown value of σ^2 by its prior mean in the final expressions for the criterion considered. Other authors (Hill & Hunter, 1969; Borth, 1975; Spezzaferri, 1988; Meyer, Steinberg & Box, 1996) examined some optimality criteria with specific prior assumptions for σ^2, sometimes using approximations or noninformative prior distributions, and obtained the same criteria as derived in the case of known variance.

This paper shows that the above results are quite general. More specifically, let $\phi(X)$ denote an optimality criterion to be minimised with respect to a design X and assume that ϕ can be written in the form $\phi(X) = H(X) + (\sigma^2)^\delta\, G(X)$, with H and G appropriate functions. It will be shown that, when σ^2 is unknown with prior density $p(\sigma^2)$, the Bayesian optimal design X^* depends on $p(\sigma^2)$ only through $E\left\{(\sigma^2)^\delta\right\}$. In the special case of $H(X) = 0$, the design X^* that minimises $\phi(X)$ for σ^2 known coincides with the one that minimizes $\phi(X)$ for σ^2 unknown.

A number of optimal design criteria $\phi(X)$, considered in the Bayesian literature on design for normal linear models, are listed in §2; §3 presents a condition that shows the influence that a prior distribution on σ^2 has on design criterion $\phi(X)$; §4 contains a short discussion.

2. Bayesian Optimal Design Criteria for the Normal Linear Model.

Let θ be a vector of k parameters in a linear regression model, where the data $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are normally distributed, with mean $X\theta$ and covariance matrix $\sigma^2 I_n$, and assume first that σ^2 is known. The choice of an experimental design consists of the choice of the matrix X, whose columns are n values at which k regressors are fixed for observing y. Let the prior distribution for θ conditional on σ^2 be conjugate normal with mean θ_0 and covariance matrix $\sigma^2 R^{-1}$, with R known. The posterior distribution for θ is then normal with mean $E(\theta|y, X) = \theta^* = (X^{\mathrm{T}}X + R)^{-1}(X^{\mathrm{T}}y + R\theta_0)$ and covariance matrix $V(\theta|y, X) = \sigma^2 D(X) = \sigma^2(X^{\mathrm{T}}X + R)^{-1}$. When using this model, many authors (Owen, 1970; Giovagnoli & Verdinelli, 1983, 1985; Pilz, 1991) defined Bayesian optimal design criteria as functions $\phi(X)$ of the posterior covariance matrix $\sigma^2 D(X)$, by analogy with the non-Bayesian theory of optimal design. Other authors (Box & Hill, 1967; Borth, 1975) proposed Bayesian optimality criteria built to meet specific requirements of an experiment. Still other authors, following the decision-theoretic framework advocated by Lindley's work in Bayesian statistics, presented utility based approaches for solving specific design problems; see for example Brooks (1972, 1974, 1976, 1977), Tiao &Afonja (1976), Smith & Verdinelli (1980), Spezzaferri (1988), Verdinelli, Polson & Singpurwalla (1993). These different approaches have beenillustrated in Chaloner & Verdinelli (1995), and in I. Verdinelli unpublished 1996 Ph.D. Thesis from Carnegie Mellon University, where a contribution towards their unification is also given.

Whatever approach is chosen, Bayesian optimal design criteria for the linear model are functions $\phi(\cdot)$ of the matrix X. For completeness, we list here some of the Bayesian design criteria more often considered in the literature. In particular there is a group of criteria that has been named Bayesian alphabetical optimal design criteria, by analogy with the corresponding non-Bayesian criteria (Box, 1982). They include D-, A- and c-optimality.

Bayesian D-optimality is a design criterion that minimises

$$\phi_D(X) = \det\{\sigma^2(X^{\mathrm{T}}X + R)^{-1}\}, \tag{1}$$

with respect to X. As noted in Chaloner & Verdinelli (1995), besides the expected gain in Shannon information on θ, considered by many authors, there are other expected utility functions that generate Bayesian D-optimality. Some are based on the use of proper scoring rules as in Spezzaferri (1988), or Eaton Giovagnoli & Sebastiani (1996). Tiao & Afonja (1976) adopted instead a two-valued utility function for deriving this criterion.

Bayesian A-optimality is the design criterion that minimises

$$\phi_A(X) = \mathrm{tr}\{A\ \sigma^2(X^{\mathrm{T}}X + R)^{-1}A^{\mathrm{T}}\}\ , \tag{2}$$

where A is a non-negative definite matrix. A-optimality is obtained from an expected quadratic loss function for estimating the parameter vector θ.

Bayesian c-optimality is a special case of Bayesian A-optimality, obtained if $\mathrm{rank}(A) = 1$. In this case A can be written as cc^{T}, where c is a vector of k components. Then $\phi_A(X)$ reduces to

$$\phi_c(X) = \sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X + R)^{-1}c. \tag{3}$$

This criterion can also be derived in different ways. Duncan & DeGroot (1976) and Chaloner (1984) found Bayesian c-optimality using expected squared loss for estimating a linear combination of the parameters: $\psi = c^{\mathrm{T}}\theta$ with c fixed. Also Verdinelli *et al.* (1993) obtained c-optimality when designing for prediction about a future observation y_{n+1}, at a preassigned point x_{n+1}, with y_{n+1} normally distributed with mean $x_{n+1}^{\mathrm{T}}\theta$, variance σ^2, and independent of y given θ and σ^2. They used as expected utility the expected gain in Shannon information about y_{n+1} and derived c-optimality with $c = x_{n+1}$.

More generally, when x_{n+1} is not fixed in advance, the use of Shannon information, as utility for prediction, leads to the intuitive notion of minimising the predictive variance of a future observation y_{n+1}, to be taken at some point x_{n+1}, and the function

$$\phi_1(X, x_{n+1}) = \sigma^2_{n+1}(X, x_{n+1}) = \sigma^2\left\{x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X + R)^{-1}x_{n+1} + 1\right\} \tag{4}$$

is the criterion to be minimised. This also yields Bayesian c-optimality when x_{n+1} is preassigned. The criterion ϕ_1 in (4) depends on both X and x_{n+1}. When interest is in prediction, in fact, the design problem might be regarded as choosing either the matrix X or the point x_{n+1} where the next observation will be taken. The criterion in (4) can be generalised when the design choice covers not only X, but also the $(n_F \times k)$ follow-up design matrix X_F where a new vector y_F of n_F observations will be taken. This leads to the criterion

$$\phi_1(X, X_F) = \det\left[\sigma^2\{X_F(X^{\mathrm{T}}X + R)^{-1}X_F^{\mathrm{T}} + I_{n_F}\}\right], \tag{5}$$

a generalisation of (4). This criterion might involve only the choice of X_F, or x_{n+1}, if a vector y has been obtained from a previous design matrix X.

Other criteria have been devised for prediction, in particular when a previous experiment has shown that several different models might be appropriate and the follow-up design matrix X_F has to be chosen for model selection. Hill, Hunter & Wichern (1968), and Borth (1975) proposed methods for these types of design. In his review of design procedures for model selection, Hill (1978) stressed the importance of experimental design for the dual problem of model discrimination and inference on the parameters. Meyer *et al.* (1996), in a follow-up fractional factorial experiment, used the method proposed by Box & Hill (1967) for choosing a design X_F to discriminate among m candidate models $M_1, M_2, \ldots, M_m$. The procedure is based on the expected total Kullback-Leibler distance, or the expected gain in Shannon information, between the predictive distributions of each pair of candidate models M_i and M_j. This leads to the design criterion

$$\begin{aligned}\phi_2(X_F) = & \sum_{i \neq j = 1}^{m} P(M_i|y_F)P(M_j|y_F) \times \qquad (6)\\ & \times \left[\mathrm{tr}\left\{(I_{n_F} + X_{i,F}D_i(X)X_{i,F}^{\mathrm{T}})(I_{n_F} + X_{j,F}D_j(X)X_{j,F}^{\mathrm{T}})^{-1}\right\} - n_F + \sigma^{-2}y_F^{\mathrm{T}}Q_{i,j}y\right].\end{aligned}$$

In (6) $P(M_h|y_F)$ is the posterior probability of the hth model ($h = 1, 2, \ldots, m$) once the vector y_F is available, $X_{h,F}$ is the sub matrix of X_F related to the regressors in the hth model, $D_h(X)$ is the posterior covariance matrix obtained for the hth model after observing y, and the matrix $Q_{i,j}$ in the quadratic form is given by

$$\begin{aligned}Q_{i,j} = & \left\{X_{i,F}D_i(X)X_i^{\mathrm{T}} - X_{j,F}D_j(X)X_j^{\mathrm{T}}\right\}^{\mathrm{T}} \times \\ & \left\{I_{n_F} + X_{j,F}D_j(X)X_{j,F}^{\mathrm{T}}\right\}^{-1} \left\{X_{i,F}D_i(X)X_i^{\mathrm{T}} - X_{j,F}D_j(X)X_j^{\mathrm{T}}\right\},\end{aligned}$$

where X_h is the sub matrix of the design X relative to the regressors of the hth model.

A different approach was pursued by Spezzaferri (1988), who adopted a utility function based on a proper quadratic scoring rule, first considered by de Finetti (1962) and by Buehler (1971). This author showed that such a utility function can be used in designing for a variety of problems. In particular, when designing for model discrimination and parameter estimation for two linear models M_1 and M_2, with M_1 nested in M_2, the criterion to be minimised with respect to X can be written as

$$\begin{aligned} \phi_3(X) &= \det\left\{(X^{\mathrm{T}}X + R_2^{-1})\right\}^{-1/2} \times \qquad (7) \\ &\times \left[\int_{\mathcal{Y}} \frac{p(y|M_1)}{P(M_1)\ \sqrt{[\ \det(W_2)\ \det\{S(X)\}]}\ +\ P(M_2)\ \exp\{(2\sigma^2)^{-1}y^{\mathrm{T}}Q(X)y\}}\ dy\right]^{-1}, \end{aligned}$$

where $p(y|M_1)$ is the marginal density of the data under model M_1. The prior covariance matrix M_2 is $\sigma^2 R_2^{-1}$, and $\sigma^2 W_2^{-1}$ is a submatrix of it, representing the prior covariance matrix of the extra parameters in M_2. The integral in (7) is computed under the assumption that M_1 is the true model. The expressions for the matrices $S(X)$ and $Q(X)$, not given explicitly here, depend on the design matrix X and on the matrices R_2 and W_2. It can be shown that $\sigma^{-2}y^{\mathrm{T}}Q(X)y \sim \chi^2_{k_1}$, where k_1 is the number of parameters in M_1, and therefore the integral in (7) is not a function of σ^2.

Yet another expected utility function, that can be expressed as a mixture of two components, was proposed by Verdinelli & Kadane (1992), when, as well as the issue of optimal design, there is also concern for the total output of an experiment, $1^{\mathrm{T}}y = y_1 + \ldots + y_n$. The optimal criterion derived from such a composite expected utility can be expressed as:

$$\phi_4(X) = (\sigma^2)^{-k\omega/2}\left[\mathrm{e}^{1^{\mathrm{T}}X\theta_0}\ \det\{(X^{\mathrm{T}}X + R)^{-1}\}\right], \qquad (8)$$

where ω is the ratio between the weights in the two components of the expected utility.

The optimality criteria in (5), (6), (7), and (8) do not correspond to any non-Bayesian criterion. Similarly, there are non-Bayesian design criteria, used in designs for normal linear models, that do not have parallel Bayesian criteria, with a utility theoretical justification. The main point of similarity of the two approaches to the design choice for a normal model is that criteria are function of a covariance matrix: either the covariance matrix of the least squares estimator, or the posterior covariance matrix for θ. It is known that any difference between the same alphabetical Bayesian and non-Bayesian design criterion is unimportant if n is large, or when there is little prior information. When, instead, n is small the prior distribution has a larger influence on the posterior distribution and on the design. Of special concern, here, is the fact that when the sample size is small it might not be reasonable to assume that the error variance is known.

The next section shows that, apart from $\phi_2(X_F)$ in (6), all other optimality criteria listed here are insensitive to the knowledge of σ^2. The only prior quantity that affects the optimal design minimising $\phi_2(X_F)$ is the prior mean. When the planned sample size is small, and Bayesian optimal designs can perform better than non-Bayesian designs, the only relevant information regarding the sample variability, so far as the design choice is concerned, might be the knowledge of an appropriate moment.

3. Optimal Design Criteria for σ^2 unknown.

As in §2, let the prior distribution for θ, conditional on σ^2, belong to the conjugate family, $\theta|\sigma^2 \sim N(\theta_0, \sigma^2 R^{-1})$, and assume now that σ^2 is unknown, with prior density $p(\sigma^2)$. A typical choice for $p(\sigma^2)$ is the conjugate inverted gamma density $p(\sigma^2|\alpha,\beta) \propto (\sigma^2)^{-(\alpha+1)}\exp(-\beta\sigma^{-2})$, that includes, in the limit, the noninformative prior $p(\sigma^2) \propto \sigma^{-2}$. When this prior for σ^2 is used, both the prior and the posterior marginal distributions for θ are multivariate Student t distributions. As shown in I. Verdinelli's Thesis, and contrary to what was said in Chaloner & Verdinelli (1995), this does not involve intractable algebraic difficulties, nor does it require approximations for deriving the appropriate optimal design criteria generated by these assumptions for the distribution of $(y, \theta, \sigma^2|X)$.

More generally, whatever the prior density $p(\sigma^2)$, the following propositions hold.

PROPOSITION 1. *Consider a normal linear regression model, with mean $X\theta$ and covariance matrix $\sigma^2 I_n$. Let the prior distribution for $\theta|\sigma^2$ be conjugate normal and let $\phi(X)$ be an optimal design criterion. Let $H(\cdot)$ and $G(\cdot)$ be functions of the matrix X, not depending on σ^2, and let δ be a positive real number. Denote by X^* the optimal design matrix for ϕ when σ^2 is known. If $\phi(X)$ can be written in the form*

$$\phi(X) = H(X) + (\sigma^2)^\delta G(X), \tag{9}$$

for σ^2 known, then, when σ^2 is unknown with a prior $p(\sigma^2)$, the optimal design matrix has the same structure as X^* with $(\sigma^2)^\delta$ replaced by $E\{(\sigma^2)^\delta\}$.

Proof.

When σ^2 is unknown, a criterion $\phi(X)$ can be seen as a conditional criterion $\phi(X|\sigma^2)$, depending on σ^2. Since (9) holds for σ^2 known, the conditional criterion can be written in the same form. The unconditional criterion $\phi(X)$, when σ^2 is unknown, is then obtained by integrating out the nuisance parameter with prior $p(\sigma^2)$, yielding

$$\phi(X) = H(X) + G(X) \int (\sigma^2)^\delta d\sigma^2 = H(X) + E\left\{(\sigma^2)^\delta\right\}\ G(X).$$

Proposition 2. *All the criteria of §2, can be written in the form (9). Only in criterion $\phi_2(X_F)$, where the follow-up design matrix X_F has to be chosen for model selection, is the function $H(X_F)$ nonzero. Therefore, the optimal follow-up design maximising ϕ_2 depends on the δ moment of $p(\sigma^2)$.*

Proof.

The algebraic expressions for the optimal design criteria of §2 lead to the following specific results.

The D-optimality criterion in (1) has $G(X) = \det(X^{\mathrm{T}}X + R)$ and $\delta = k$, the number of parameters.

The A-optimality criterion in (2), or in (3), has $G(X) = \mathrm{tr}\{A^{\mathrm{T}}(X^{\mathrm{T}}X + R)^{-1}A\}$ and $\delta = 1$.

The criterion $\phi_1(X, X_F)$ in (5) has $G(X, X_F) = \det[\sigma^2\{X_F(X^{\mathrm{T}}X + R)^{-1}X_F^{\mathrm{T}} + I_{n_F}\}]$ and $\delta = n_F$.

The criterion $\phi_2(X_F)$ in (6) is a sum of quantities $\phi_2^{(ij)}(X_F)$ and

$$\begin{aligned} H(X_F) \;=\; & \sum_{i \neq j=1}^{m} H^{(ij)}(X_F) = \sum_{i \neq j=1}^{m} P(M_i|y_F)P(M_j|y_F) \times \\ & \times \left[\mathrm{tr}\left\{\ (I_{n_F} + X_{i,F}D_i(X)X_{i,F}^{\mathrm{T}})\ (I_{n_F} + X_{j,F}D_j(X)X_{j,F}^{\mathrm{T}})^{-1}\right\} - n_F\right]\ ; \end{aligned}$$

$G(X_F) = \sum_{i \neq j=1}^{m} G^{(ij)}(X_F) = \sum_{i \neq j=1}^{m}\ P(M_i|y_F)\ P(M_j|y_F)\ y_F^{\mathrm{T}}\ Q_{i,j}\ y$, and $\delta = -1$.

The criterion $\phi_3(X)$ in (7) has $G(X) = \phi_3(X)$, $\delta = 0$.

The criterion $\phi_4(X)$ in (8), has $G(X) = e^{1^{\mathrm{T}}X\theta_0}\left\{\det(X^{\mathrm{T}}X + R)^{-1/2}\right\}$, and $\delta = \frac{1}{4}\omega k$.

Hence there is no need to specify the prior information on σ^2 when criteria other than $\phi_2(X_F)$ are considered for experimental planning. The criterion $\phi_2(X_F)$ produces an optimal follow-up design X_F^* that depends only on the expectation of σ^{-2}.

4. Discussion.

This paper has shown that, under the assumption of normality, Bayesian experimental designs are, at worst, only mildly affected by the knowledge of data variability, so long as the relevant criteria for optimal design choice can be expressed as in (9). An open question is the possible use of criteria for which condition (9) does not hold. However, this concern does not seem to be very serious for current applications of the theory of Bayesian design in practice. In this light, it is reassuring that a type of robustness to knowledge about unknown variance holds in most of the cases proposed in the literature.

Acknowledgment

I wish to thank the Associate Editor and a referee for helpful comments, Jay Kadane for useful conversations on this topic and Kathryn Chaloner for having the right insight, long before I did. This work was partially supported by grants from the National Institute of Health, from the National Science Foundation and from the Pennsylvania Department of Health.

References

Borth, D.M. (1975). A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. R. Statist. Soc.* B **37**, 77-87.

Box, G.E.P. (1982). Choice of response surface design and alphabetic optimality. *Utilitas Math.* B **21** 11-55.

Box, G.E.P. & Hill, W.J. (1967). Discriminating Among Mechanicistic Models. *Technometrics*, **9**, 57-71.

Brooks, R.J. (1972). A decision theory approach to optimal regression designs. *Biometrika* **59** 563-71.

Brooks, R.J. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika* **61** 303-11.

Brooks, R.J. (1976). Optimal regression designs for prediction when prior knowledge is available. *Metrika* **23** 217-21.

Brooks, R.J. (1977). Optimal regression design for control in linear regression. *Biometrika* **64** 319-25.

Buehler, B.H. (1971). Measuring information and uncertainty. In *Foundations of Statistical Inference*, Ed. V.P. Godambe and D.A. Sprott. Holt, pp. 330-41. Toronto: Rinehart and Winston of Canada.

Chaloner, K. (1984). Optimal Bayesian experimental designs for linear models. *Ann. Statist.* **12** 283-300.

Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10**, 237-304.

de Finetti, B. (1962). Does it make sense to speak of "Good Probability Appraisers?" In *The Scientist Speculates.* Ed. I.J. Good, pp. 357-364, London: Heinemann.

Duncan, G. & DeGroot, M.H. (1976). A mean squared error approach to optimal design theory. In *Proceedings of the 1976 Conference on Information: Science and systems.* The John Hopkins University. 217-21.

Eaton, M.L., Giovagnoli, A. & Sebastiani, P. (1996). A predictive approach to the Bayesian design problem with application to normal regression models. *Biometrika* **83** 111-25.

Geisser, S. (1993). *Predictive Inference: An Introduction*, London: Chapman and Hall.

Giovagnoli, A. & Verdinelli, I. (1983). Bayesian D-optimal and E-optimal block designs. *Biometrika* **70** 695-706.

Giovagnoli, A. & Verdinelli, I. (1985). Optimal block designs under a hierarchical linear model. In *Bayesian Statistics 2.* Ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, pp. 655-61. Amsterdam: North Holland.

Guttman, I. (1971). A remark on the optimal regression designs with previous observations of Covey-Crump and Silvey. *Biometrika*, **58** 683-85.

Hill, D.H. (1978). A review of experimental design procedures for regression model discrimination. *Technometrics*, **20** 15-21.

Hill, W.J. & Hunter, W.G. (1969). A note on designs for model discrimination: Variance unknown case. *Technometrics* **11**, 396-400.

Hill, W.J., Hunter, W.G. & Wichern, D.W. (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics* **10**, 145-60.

MEYER, A.B., STEINBERG, C.D. & BOX, G.E.P. (1996). Follow-up designs to resolve confounding in multi-factor experiments. *Technometrics* **38** 303-332.

OWEN, R.J. (1970). The optimum design of a two-factor experiment using prior information. *Ann. Statist.* **41** 1917-34.

PILZ, J. (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models*, 2nd ed. New York: Wiley.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley.

SINHA, B.K. (1970). A Bayesian approach to optimum allocation in regression problems. *Calcutta Statist. Assoc. Bull.* **19** 45-52.

SMITH, A.F.M. & VERDINELLI, I. (1980). A note on Bayesian design for inference using a hierarchical linear model. *Biometrika* **67** 613-9.

SPEZZAFERRI, F. (1988). Non-sequential designs for model discrimination and parameter estimation. In *Bayesian Statistics 3*, Ed. J.M. Bernardo *et al.* Oxford: Oxford University Press. 777-83.

TIAO, G. C. & AFONJA, B. (1976). Some Bayesian considerations of the choice of design for ranking, selection and estimation. *Ann. Inst. Statist. Math.*, **28**, 167-86.

VERDINELLI, I. & KADANE, J. B. (1992). Bayesian designs for maximizing information and outcome. *J. Am. Statist. Assoc.* **87**, 510-515.

VERDINELLI, I., POLSON, N. & SINGPURWALLA, N. (1993). Shannon Information and Bayesian design for prediction in accelerated life testing. In *Reliability and Decision Making.* Ed. R.E. Barlow, C.A. Clarotti & F. Spizzichino, pp 247-56. London: Chapman and Hall.