Notes on Maximum Entropy Design

Douglas P. Wiens[1]

April 7, 2019

1 Information and entropy

Shannon (1948), as discussed in Lindley (1956), showed that subject to reasonable conditions the information about a parameter Θ taking values in a space Ω, with prior density $p(\theta)$ with respect to some dominating measure (taken here as Lebesgue measure, for simplicity) is measured by

$$I_0 = \int_\Omega p(\theta) \log p(\theta)\, d\theta = E_\Theta \left[\log p(\Theta)\right].$$

Suppose that a r.v. $Y \in \mathbb{R}^n$ has a density $p(y|\theta)$ possibly depending upon θ, and that Y is observed with the intention of acquiring information about θ. After an experiment ξ is performed, resulting in an observation y, the posterior distribution of θ is

$$p(\theta|y) = p(y|\theta)\, p(\theta) / p(y)$$

and the information is now

$$I_1(y) = \int_\Omega p(\theta|y) \log p(\theta|y)\, d\theta.$$

Thus the amount of information provided by the experiment is

$$I(\xi, y) = I_1(y) - I_0,$$

and the average amount of information provided by the experiment is

$$I(\xi) = E_Y\left[I(\xi, Y)\right] = E_Y E_\Theta \left[\log \frac{p(\Theta|Y)}{p(\Theta)}\right],$$

alternate expressions (following from the above) being

$$I(\xi) = \left\{ \begin{array}{c} E_Y E_\Theta \left[\log \frac{p(Y|\Theta)}{p(Y)}\right], \\ \int_\Omega \int_{\mathbb{R}^n} p(y,\theta) \log \frac{p(y|\theta)}{p(\theta)p(y)} dy d\theta. \end{array} \right.$$

Following Sebastiani and Wynn (2000), the *Shannon entropy* (also known as the *Boltzmann-Shannon entropy* - see Lee (2002)) of a random vector $Z \in \mathbb{R}^N$ is the negative of information:

$$Ent(\Theta) = E_\Theta\left[-\log p(\Theta)\right],$$

[1]Department of Mathematical and Statistical Sciences; University of Alberta, Edmonton, Alberta;Canada T6G 2G1. e-mail: `doug.wiens@ualberta.ca`

so that the average amount of information provided by the experiment is

$$I(\xi) = Ent(\Theta) - E_Y\left[Ent(\Theta|Y,\xi)\right].$$

If, as is typically assumed, $Ent(\Theta)$ does not depend on the experimental design, then an experiment is optimal, in the sense of maximizing $I(\xi)$, if it minimizes $E_Y\left[Ent(\Theta|Y,\xi)\right]$. Under further conditions, among them that $Ent(Y|\xi)$ and $E_Y\left[Ent(\Theta|Y,\xi)\right]$ are bounded, Theorem 1 of Sebastiani and Wynn (2000) applies and yields that minimization of $E_Y\left[Ent(\Theta|Y,\xi)\right]$ is equivalent to maximization of

$$Ent(Y|\xi) = -\int_{\mathbb{R}^n} (\log p(y|\xi))\, p(y|\xi)\, dy.$$

This motivates the name 'Maximum Entropy Sampling', as in Shewtry and Wynn (1987).

If the experiment yields observations $\mathbf{y} = (Y_1, ..., Y_n)' \in \mathbb{R}^n$, with joint density $f_n(y|\theta)$, then in the above

$$p(\mathbf{y}|\xi) = \int_{\Omega} f_n(\mathbf{y}|\theta)\, p(\theta)\, d\theta.$$

Suppose the Y_i are independent, with densities $p(y_i|\theta)$ parameterized by their means $\mu_i = \mu(\mathbf{x}_i)$ for design variables $\mathbf{x}_i$ ranging over $\chi = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$. There may be nuisance parameters as well. If $n_i = n\xi_i$ observations are made at $\mathbf{x}_i$ $(i = 1, ..., N)$ then (with $\prod_{j=1}^{0} = 1$) we have

$$f_n(\mathbf{y}|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{n_i} p(y_j; \mu(\mathbf{x}_i)|\theta).$$

Example: Suppose that
(i) for independent variables $\mathbf{x}$ belonging to a design space $\chi \subset \mathbb{R}^q$,
(ii) for regressors $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$ and a function $\psi(\mathbf{x})$, arbitrary but with $\int_\chi \psi^2(\mathbf{x})\, d\mathbf{x} = 1$ and $\int_\chi \mathbf{f}(\mathbf{x})\, \psi(\mathbf{x})\, d\mathbf{x} = \mathbf{0}$,

the conditional density is $p(y|\boldsymbol{\theta}, \eta) = \left(\sigma_\varepsilon\sqrt{2\pi}\right)^{-1} e^{-\frac{\left(y - \mathbf{f}'(\mathbf{x})\boldsymbol{\theta} - \frac{\eta}{\sqrt{n}}\psi(\mathbf{x})\right)^2}{2\sigma_\varepsilon^2}}$. Then if y_i is observed at the design point $\mathbf{x}_i$, if $\mathbf{F}_{n\times d}$ has rows $\{\mathbf{f}'(\mathbf{x}_i)\}_{i=1}^n$, and if $\boldsymbol{\psi} = (\psi(\mathbf{x}_1), ..., \psi(\mathbf{x}_n))'$

$$f_n(\mathbf{y}|\boldsymbol{\theta}, \eta) = \left(2\pi\sigma_\varepsilon^2\right)^{-n/2} e^{-\frac{1}{2}\left\|\frac{\mathbf{y} - \mathbf{F}\boldsymbol{\theta} - \frac{\eta}{\sqrt{n}}\boldsymbol{\psi}}{\sigma_\varepsilon}\right\|^2}.$$

The interpretation is that the experimenter will take $\eta = 0$, under the mistaken assumption that the true mean value is adequately specified by $\mathbf{f}'(\mathbf{x})\,\boldsymbol{\theta}$. If $\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \mathbf{R}^{-1})$, with

$$p(\boldsymbol{\theta}) = \left|2\pi\mathbf{R}^{-1}\right|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{R}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)},$$

and if η has (true) prior $p(\eta)$, then

$$\begin{aligned} p(\mathbf{y}|\xi) &= \int_0^\infty \int_{\mathbb{R}^d} f_n(\mathbf{y}|\boldsymbol{\theta},\eta)\, p(\boldsymbol{\theta})\, p(\eta)\, d\boldsymbol{\theta} d\eta \\ &= \left(2\pi\sigma_\varepsilon^2\right)^{-n/2} \left|2\pi\mathbf{R}^{-1}\right|^{-1/2} \int \left\{ \int_{\mathbb{R}^d} e^{-\frac{1}{2}\left[\left\|\frac{\mathbf{y}-\mathbf{F}\boldsymbol{\theta}-\eta\boldsymbol{\psi}/\sqrt{n}}{\sigma_\varepsilon}\right\|^2 + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{R}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]} d\boldsymbol{\theta} \right\} p(\eta)\, d\eta. \quad (1) \end{aligned}$$

With

$$\begin{aligned} \mathbf{c}_{n\times 1} &\overset{def}{=} \frac{\mathbf{y}-\mathbf{F}\boldsymbol{\theta}_0-\eta\boldsymbol{\psi}/\sqrt{n}}{\sigma_\varepsilon}, \\ \mathbf{b}_{n\times 1} &\overset{def}{=} \frac{\mathbf{F}'\mathbf{c}}{\sigma_\varepsilon}, \\ \mathbf{V}_{d\times d} &\overset{def}{=} \left(\frac{\mathbf{F}'\mathbf{F}}{\sigma_\varepsilon^2}+\mathbf{R}\right)^{-1}, \end{aligned}$$

we have that the term in square brackets in (1) is

$$\begin{aligned} & \left\|\mathbf{c}-\frac{\mathbf{F}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{\sigma_\varepsilon}\right\|^2 + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{R}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\ =\ & \mathbf{c}'\mathbf{c} - 2\mathbf{b}'(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\ =\ & (\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b}) - \mathbf{b}'\mathbf{V}\mathbf{b} + \mathbf{c}'\mathbf{c} \\ =\ & (\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b}) + \mathbf{c}'\left(\mathbf{I}_n - \frac{\mathbf{F}\mathbf{V}\mathbf{F}'}{\sigma_\varepsilon^2}\right)\mathbf{c} \\ =\ & (\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b}) + \mathbf{c}'\left(\mathbf{I}_n + \frac{\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'}{\sigma_\varepsilon^2}\right)^{-1}\mathbf{c}; \end{aligned}$$

here we use that

$$\mathbf{I}_n - \frac{\mathbf{F}\mathbf{V}\mathbf{F}'}{\sigma_\varepsilon^2} = \mathbf{I}_n - \frac{\mathbf{F}\mathbf{R}^{-1}}{\sigma_\varepsilon}\left(\frac{\mathbf{F}'\mathbf{F}\mathbf{R}^{-1}}{\sigma_\varepsilon^2}+\mathbf{I}_d\right)^{-1}\frac{\mathbf{F}'}{\sigma_\varepsilon} = \left(\mathbf{I}_n + \frac{\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'}{\sigma_\varepsilon^2}\right)^{-1}.$$

The integral in braces in (1) is

$$\begin{aligned} & |2\pi\mathbf{V}|^{1/2} e^{-\frac{1}{2}\mathbf{c}'\left(\mathbf{I}_n+\frac{\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'}{\sigma_\varepsilon^2}\right)^{-1}\mathbf{c}} \cdot \int_{\mathbb{R}^d} |2\pi\mathbf{V}|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0-\mathbf{V}\mathbf{b})} d\boldsymbol{\theta} \\ =\ & |2\pi\mathbf{V}|^{1/2} e^{-\frac{1}{2}\mathbf{c}'\left(\mathbf{I}_n+\frac{\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'}{\sigma_\varepsilon^2}\right)^{-1}\mathbf{c}}, \end{aligned}$$

and so

$$\begin{aligned} p(\mathbf{y}|\xi) &= \left(2\pi\sigma_\varepsilon^2\right)^{-n/2}\left|2\pi\mathbf{R}^{-1}\right|^{-1/2}|2\pi\mathbf{V}|^{1/2}\int e^{-\frac{1}{2}\mathbf{c}'\left(\mathbf{I}_n+\frac{\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'}{\sigma_\varepsilon^2}\right)^{-1}\mathbf{c}} p(\eta)\, d\eta \\ &= \left|2\pi\left(\sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}'\right)\right|^{-1/2} \cdot \int e^{-\frac{1}{2}\sigma_\varepsilon\mathbf{c}'\left(\sigma_\varepsilon^2\mathbf{I}_n+\mathbf{F}\mathbf{R}^{-1}\mathbf{F}'\right)^{-1}\sigma_\varepsilon\mathbf{c}} p(\eta)\, d\eta. \end{aligned}$$

i.e.

$$\mathbf{y}|\eta \sim N\left(\mathbf{F}\boldsymbol{\theta}_0 + \eta\boldsymbol{\psi}/\sqrt{n}, \sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}'\right), \tag{2}$$

in agreement with Sebastiani and Wynn (2000) when $\eta = 0$.

- The case in which $\eta \sim N(0, \sigma_\eta^2)$, independently of $\boldsymbol{\theta}$, can be derived by putting $\eta = 0$ in (2) but then making the replacements

$$\begin{aligned}
\mathbf{F} &\rightarrow \left[\mathbf{F} \vdots \boldsymbol{\psi}/\sqrt{n}\right], \\
\boldsymbol{\theta}_0 &\rightarrow \begin{pmatrix} \boldsymbol{\theta}_0 \\ \eta \end{pmatrix}, \\
\mathbf{R}^{-1} &\rightarrow \begin{pmatrix} \mathbf{R}^{-1} & \mathbf{0} \\ \mathbf{0}' & \sigma_\eta^2 \end{pmatrix},
\end{aligned}$$

  obtaining

$$\mathbf{y}|\xi \sim N\left(\mathbf{F}\boldsymbol{\theta}_0, \sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}' + \frac{\sigma_\eta^2}{n}\boldsymbol{\psi}\boldsymbol{\psi}'\right),$$

  with, up to an additive constant,

$$Ent\left(\mathbf{y}|\xi\right) = \frac{1}{2}\log\left|\sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}' + \frac{\sigma_\eta^2}{n}\boldsymbol{\psi}\boldsymbol{\psi}'\right|.$$

  Thus a maximum entropy design will maximize

$$\left|\sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}' + \frac{\sigma_\eta^2}{n}\boldsymbol{\psi}\boldsymbol{\psi}'\right| = \left|\sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}'\right|\left(1 + \frac{\sigma_\eta^2}{n}\boldsymbol{\psi}'\left(\sigma_\varepsilon^2\mathbf{I}_n + \mathbf{F}\mathbf{R}^{-1}\mathbf{F}'\right)^{-1}\boldsymbol{\psi}\right).$$

  - BUT - should this really be called the 'information about $\begin{pmatrix} \boldsymbol{\theta} \\ \eta \end{pmatrix}$?
  - The moments above (with or without the normality) also follow from first conditioning on η and calculating the expectations in stages.

References

Lee, Jon (2002), "Maximum Entropy Sampling," in *Encyclopedia of Environmetrics*, eds. Abdel H. El-Shaarawi and Walter W. Piegorsch, Chichester: Wiley, pp. 1229-1234.

Lindley, D. V. (1956), "On a Measure of Information Provided by an Experiment," *Annals of Mathematical Statistics*, 27, 986-1005.

Sebastiani, P., and Wynn, H. P. (2000), "Maximum Entropy Sampling and Optimal Bayesian Experimental Design," *Journal of the Royal Statistical Socieity, Series B*, 1, 145-157.

Shewry, M. C. and Wynn, H. P. (1987), "Maximum Entropy Sampling," *Journal of Applied Statistics*, 14, 165-170.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379-423 & 623-656.