Description

1 Problem Setup

Denote the design (input) space as $\mathcal{X}$. Given M models $\{\mathcal{M}_i\}_{i=1}^M$, each with parameters $\theta_i \subseteq \Theta_i$ and prior distribution $p(\mathcal{M}_i)$, we first give each θ_i a (multivariate Gaussian) prior $p(\theta_i|\mathcal{M}_i)$. For simplicity, we assume that one of $\{\mathcal{M}_i\}_{i=1}^M$ is the ground-truth, i.e., $\mathcal{M}_{\text{true}}$.

2 Input Selection Criterion: Model Selection

2.1 Method 1: `getSelCritLogDet.m`

We first draw several samples from the density $p(\theta_i|\mathcal{M}_i)$, denoted by $\{\theta_i^s\}_{s=1}^{K_i}$, using MCMC (specifically, HMC). (As an alternative approach, we can find a local minimum of $p(\theta_i|\mathcal{M}_i)$, denoted by θ_i^{MAP}, using HMC.) Then, we estimate the response $y_i^s(x) \triangleq \mathcal{M}_i(x;\theta_i^s) + \epsilon_i^s$, where $\{\epsilon_i^s\}_{i\in[M],s\in[K_i]} \overset{iid}{\sim} \mathcal{N}(0,\sigma_{\text{n}}^2)$. Thus $y_i^s(x) \sim \mathcal{N}(\mathcal{M}_i(x;\theta_i^s),\sigma_{\text{n}}^2)$. For any $(i,s) \in [M]\times[K_i]$ and $(j,t) \in [M]\times[K_j]$, compute

$$\begin{aligned} D_{(i,s),(j,t)}(x) &\triangleq D_{\text{KL}}\left(\mathcal{N}(\mathcal{M}_i(x;\theta_i^s),\sigma_{\text{n}}^2),\mathcal{N}(\mathcal{M}_j(x;\theta_j^t),\sigma_{\text{n}}^2)\right) \\ &= \frac{\left(\mathcal{M}_i(x;\theta_i^s),\sigma_{\text{n}}^2) - \mathcal{M}_j(x;\theta_j^t),\sigma_{\text{n}}^2)\right)^2}{2\sigma_{\text{n}}^2}. \end{aligned}$$

We choose the design point x^* to be a local minimum of

$$S(x) \triangleq -\log\det D(x).$$

2.2 Method 2: `getSelCritJSDiv.m`

This method was proposed in Vanlier et al. [2014]. The first step is the same as those in Section 2.1, i.e., we draw several samples from the density $p(\theta_i|\mathcal{M}_i)$, denoted by $\{\theta_i^s\}_{s=1}^{K_i}$, using HMC. For each model $\mathcal{M}_i$, we aim to find the distribution of the its predicted response given x, i.e.,

$$p(y|\mathcal{M}_i,x) = \int_{\Theta_i} p(y|\theta_i,\mathcal{M}_i,x)p(\theta_i|\mathcal{M}_i)\,\mathrm{d}\theta_i, \tag{2.1}$$

where (assuming the noise variance σ_{n}^2 is known)

$$p\left(y|\theta_i,\mathcal{M}_i,x\right) = \frac{1}{\sqrt{2\pi\sigma_{\text{n}}^2}}\exp\left\{-\frac{(y-\mathcal{M}_i(x;\theta_i))^2}{2\sigma_{\text{n}}^2}\right\}.$$

We can approximate this density using the samples $\{\theta_i^s\}_{s=1}^{K_i}$, i.e.,

$$p(y|\mathcal{M}_i, x) \approx \widehat{p}(y|\mathcal{M}_i, x) \triangleq \frac{1}{K_i}\sum_{s=1}^{K_i} p(y|\theta_i^s, \mathcal{M}_i, x) = \frac{1}{K_i}\sum_{s=1}^{K_i} \mathcal{N}(y|\mathcal{M}_i(x;\theta_i^s), \sigma_{\mathrm{n}}^2). \tag{2.2}$$

Such an approximation also gives us

$$\begin{aligned}\nabla_x p(y|\mathcal{M}_i, x) &\approx \nabla_x \widehat{p}(y|\mathcal{M}_i, x)\\ &= \frac{1}{K_i}\sum_{s=1}^{K_i} \mathcal{N}(y|\mathcal{M}_i(x;\theta_i^s), \sigma_{\mathrm{n}}^2)\left(\frac{y-\mathcal{M}_i(x;\theta_i^s)}{\sigma_{\mathrm{n}}^2}\right)\nabla_x \mathcal{M}_i(x;\theta_i^s).\end{aligned} \tag{2.3}$$

Let us define the averaged predictive distribution $p(y|x)$ from all the M models, i.e.,

$$p(y|x) = \sum_{i=1}^{M} p(\mathcal{M}_i)p(y|\mathcal{M}_i, x). \tag{2.4}$$

The OED criterion is based on the Jensen-Shannon divergence (JSD), i.e.,

$$D_{\mathrm{JS}}(x) \triangleq \sum_{i=1}^{M} p(\mathcal{M}_i) D_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\|p(y|x)\big). \tag{2.5}$$

To approximate the KL-divergence in (2.5), for each $i \in [M]$, we first draw N_i samples from $p(y|\mathcal{M}_i, x)$ (in fact, $\widehat{p}(y|\mathcal{M}_i, x)$) using MCMC, which are denoted by $\{y_i^t\}_{t=1}^{N_i}$, so that

$$D_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\|p(y|x)\big) \approx \widehat{D}_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\|p(y|x)\big) \triangleq \frac{1}{N_i}\sum_{t=1}^{N_i} \frac{p(y_i^t|\mathcal{M}_i, x)}{p(y_i^t|x)}. \tag{2.6}$$

This gives us

$$\begin{aligned}&\nabla_x D_{\mathrm{JS}}(x)\\ &= \sum_{i=1}^{M} p(\mathcal{M}_i)\nabla_x D_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\|p(y|x)\big)\\ &\approx \sum_{i=1}^{M} p(\mathcal{M}_i)\nabla_x \widehat{D}_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\|p(y|x)\big)\\ &= \sum_{i=1}^{M} \frac{p(\mathcal{M}_i)}{N_i}\sum_{t=1}^{N_i} \frac{p(y_i^t|x)\nabla_x p(y_i^t|\mathcal{M}_i, x) - p(y_i^t|\mathcal{M}_i, x)\nabla_x p(y_i^t|x)}{p(y_i^t|x)^2}\\ &= \sum_{i=1}^{M} \frac{p(\mathcal{M}_i)}{N_i}\sum_{t=1}^{N_i}\sum_{j=1}^{M} p(\mathcal{M}_j)\frac{p(y_i^t|\mathcal{M}_j, x)\nabla_x p(y_i^t|\mathcal{M}_i, x) - p(y_i^t|\mathcal{M}_i, x)\nabla_x p(y_i^t|\mathcal{M}_j, x)}{p(y_i^t|x)^2}.\end{aligned} \tag{2.7}$$

Thus ideally, given any $x \in \mathcal{X}$, if we can (approximately) evaluate $\{p(y|\mathcal{M}_i, x)\}_{i=1}^M$ and $\{\nabla_x p(y|\mathcal{M}_i, x)\}_{i=1}^M$ for any $y \in \mathcal{Y}$, we can approximate both $D_{\mathrm{JS}}(x)$ and $\nabla_x D_{\mathrm{JS}}(x)$. Then we can use first-order methods to find a local maximum of $D_{\mathrm{JS}}(x)$ on $\mathcal{X}$, denoted by x^*. However, evaluating these values and gradients in turn requires drawing (typically a large number of) samples of θ_i for each model $\mathcal{M}_i$, as shown in (2.2) and (2.3).

2.3 Method 3: `getSelCritJSDivU.m`

Note that Method 1 in Section 2.1 is ad-hoc and not well-justified. A more principled approach would be as follows. We first approximate $p(y|\mathcal{M}_i, x)$ for each model $\mathcal{M}_i$ as in (2.2). Then, instead of using the JSD criterion as in (2.5), we use the weighted sum of pairwise KL divergences of $\{p(y|\mathcal{M}_i, x)\}_{i=1}^M$. Specifically, define

$$\widetilde{D}_{\mathrm{KL}}(x) \triangleq \sum_{i,j=1, i\neq j}^{M} p(\mathcal{M}_i)p(\mathcal{M}_j)D_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, x)\| \, p(y|\mathcal{M}_j, x)\big), \tag{2.8}$$

and we find a local maximum of $\widetilde{D}_{\mathrm{KL}}(x)$ on $\mathcal{X}$. Note that by Jensen's inequality, $\widetilde{D}_{\mathrm{KL}}(x) \geq D_{\mathrm{JS}}(x)$, for any $x \in \mathcal{X}$.

Indeed, this criterion was proposed in Box and Hill [1967], and represents the expected entropy reduction from performing the experiment at $x \in \mathcal{X}$.

2.4 Method 4: Based on Mutual Information

This approach was proposed in Drovandi et al. [2014]. For any $x \in \mathcal{X}$, define its response by

$$y(x) \triangleq \mathcal{M}^*(x) + \epsilon, \tag{2.9}$$

where $\mathcal{M}^*$ denotes the (unknown) true model and $\epsilon \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$. Let $\mathcal{M} \in \{\mathcal{M}_i\}_{i=1}^M$ be the estimate of $\mathcal{M}^*$. We aim to select $x \in \mathcal{X}$ to maximize the mutual information between $\mathcal{M}$ and $y(x)$ (written as y in the sequel), i.e.,

$$x^* \in \operatorname*{arg\,max}_{x\in\mathcal{X}} \; \big\{ I(\mathcal{M}; y|x) = H(\mathcal{M}|x) - H(\mathcal{M}|y, x) = H(\mathcal{M}) - H(\mathcal{M}|y, x) \big\}. \tag{2.10}$$

Equivalently, we have

$$x^* \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \; H(\mathcal{M}|y, x). \tag{2.11}$$

This means we choose $x \in \mathcal{X}$ such that given its response y, the remaining uncertainty in the model estimate $\mathcal{M}$ is minimized. By definition,

$$\begin{aligned} -H(\mathcal{M}|y, x) &= \int_{\mathcal{Y}} \left\{ \sum_{i=1}^{M} p(\mathcal{M}_i|y, x) \log p(\mathcal{M}_i|y, x) \right\} p(y|x) \mathrm{d}y \\ &= \sum_{i=1}^{M} \int_{\mathcal{Y}} p(\mathcal{M}_i, y|x) \log p(\mathcal{M}_i|y, x) \mathrm{d}y \\ &= \sum_{i=1}^{M} p(\mathcal{M}_i) \int_{\mathcal{Y}} p(y|\mathcal{M}_i, x) \log p(\mathcal{M}_i|y, x) \mathrm{d}y. \end{aligned} \tag{2.12}$$

Note that in (2.12), $p(y|\mathcal{M}_i, x)$ is the predicative distribution of $\mathcal{M}_i$, given in (2.1), and $p(\mathcal{M}_i|y, x)$ is the posterior distribution of $\mathcal{M}_i$ given the data point (x, y), which can be obtained from the set of predictive distributions $\{p(y|\mathcal{M}_i, x)\}_{i=1}^M$ as

$$p\left(\mathcal{M}_i|y, x\right) = \frac{p\left(y|\mathcal{M}_i, x\right) p(\mathcal{M}_i)}{\sum_{i=1}^{M} p\left(y|\mathcal{M}_i, x\right) p(\mathcal{M}_i)}. \tag{2.13}$$

Therefore, given $\{p(y|\mathcal{M}_i, x)\}_{i=1}^M$ and $\{p(\mathcal{M}_i)\}_{i=1}^M$, (2.12) can serve as another input selection criterion.

3 Input Selection Criterion: Joint Model Selection and Parameter Estimation

We consider designing experiments not only for model selection, but also for estimating the parameters in each model. A simple way to achieve this is to consider the model-parameter pair, i.e., $\{(\mathcal{M}_i, \theta_i)\}_{\theta_i \in \Theta_i, i \in [M]}$ and their predictive distributions $\{p(y|\mathcal{M}_i, \theta_i, x)\}_{\theta_i \in \Theta_i, i \in [M]}$.

3.1 Method 1: Jensen-Shannon Divergence

The criterion in Section 2.2 can be straightforwardly extended here. Specifically, we obtain the averaged predictive distribution $p(y|x)$ in the same way as in (2.4). Then the criterion is

$$D_{\mathrm{JS}}(x) \triangleq \sum_{i=1}^{n} p(\mathcal{M}_i) \int_{\Theta_i} D_{\mathrm{KL}}\big(p(y|\mathcal{M}_i, \theta_i, x) \| p(y|x)\big) p(\theta_i|\mathcal{M}_i) \,\mathrm{d}\theta_i.$$

3.2 Method 2: `getSelCritMI.m` (Mutual Information)

We can similarly extend the criterion in Section 2.4 here, i.e., we select $x \in \mathcal{X}$ to maximize the mutual information between $(\mathcal{M}, \theta)$ and y:

$$x^* \in \operatorname*{arg\,max}_{x \in \mathcal{X}} \big\{ I(\mathcal{M}, \theta; y|x) = H(\mathcal{M}, \theta) - H(\mathcal{M}, \theta|y, x) \big\}. \tag{3.1}$$

Indeed, this is the "total entropy" criterion used in Borth [1975]. By definition,

$$\begin{aligned} -H(\mathcal{M}, \theta|y, x) &= \int_{\mathcal{Y}} \left\{ \sum_{i=1}^{M} \int_{\Theta_i} p(\mathcal{M}_i, \theta_i|y, x) \log p(\mathcal{M}_i, \theta_i|y, x) \,\mathrm{d}\theta_i \right\} p(y|x) \mathrm{d}y \\ &= \sum_{i=1}^{M} \int_{\mathcal{Y}} \int_{\Theta_i} p(\mathcal{M}_i, \theta_i, y|x) \log p(\mathcal{M}_i, \theta_i|y, x) \,\mathrm{d}\theta_i \mathrm{d}y \\ &= \sum_{i=1}^{M} p(\mathcal{M}_i) \int_{\mathcal{Y}} \int_{\Theta_i} p(y|\mathcal{M}_i, \theta_i, x) p(\theta_i|\mathcal{M}_i) \log p(\mathcal{M}_i, \theta_i|y, x) \,\mathrm{d}\theta_i \mathrm{d}y. \end{aligned}$$

To obtain $p(\mathcal{M}_i, \theta_i|y, x)$, we simply invoke the Bayes' rule, i.e.,

$$p(\mathcal{M}_i, \theta_i|y, x) = \frac{p(y|\mathcal{M}_i, \theta_i, x) p(\theta_i|\mathcal{M}_i) p(\mathcal{M}_i)}{p(y|x)}, \tag{3.2}$$

where $p(y|x)$ is given by (2.4).

To approximate $-H(\mathcal{M}, \theta|y, x)$, we first write it as

$$-H(\mathcal{M}, \theta|y, x) \overset{\mathrm{c}}{=} \int_{\mathcal{Y}} \left\{ \sum_{i=1}^{M} p(\mathcal{M}_i) \int_{\Theta_i} p(\theta_i|\mathcal{M}_i, y, x) \log p(\theta_i|\mathcal{M}_i, y, x) \,\mathrm{d}\theta_i \right\} p(y|x) \mathrm{d}y,$$

where $\overset{\mathrm{c}}{=}$ omits constants that are independent of x. Next, from (2.2) and (2.4), we can approximate $p(y|x)$ as

$$\widehat{p}(y|x) \triangleq \sum_{i=1}^{M} p(\mathcal{M}_i) \widehat{p}(y|\mathcal{M}_i, x).$$

We then draw N samples from $\widehat{p}(y|x)$, denoted by $\{y_t\}_{t=1}^N$, so

$$-H(\mathcal{M},\theta|y,x) \approx \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{M} p(\mathcal{M}_i)\int_{\Theta_i} p(\theta_i|\mathcal{M}_i, y_t, x)\log p(\theta_i|\mathcal{M}_i, y_t, x)\,\mathrm{d}\theta_i.$$

We the draw K_i samples of θ_i (denoted by $\{\theta_i^s\}_{s=1}^{K_i}$) from $\widehat{p}(\theta_i|\mathcal{M}_i, y_t, x)$, where

$$\widehat{p}(\theta_i|\mathcal{M}_i, y_t, x) \triangleq \frac{p(y_t|\mathcal{M}_i,\theta_i,x)p(\theta_i|\mathcal{M}_i)}{\widehat{p}(y_t|x)},$$

so $-H(\mathcal{M},\theta|y,x)$ can be further approximate by

$$-H(\mathcal{M},\theta|y,x) \approx \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{M}\frac{p(\mathcal{M}_i)}{K_i}\sum_{s=1}^{K_i}\log\widehat{p}(\theta_i^s|\mathcal{M}_i, y_t, x). \tag{3.3}$$

If $K_i = K$, for each $i \in [M]$, then the computational complexity is $O(NMK)$, which is huge.

4 Posteriors of Model and Model Parameters

Then we simulate the response at x^*, i.e., $y(x^*)$ according to (2.9). With the data pair $(x^*, y(x^*))$, we can update the model posterior distribution $p\left(\mathcal{M}_i|x^*, y(x^*)\right)$ according to (2.13).

5 Test Model

We take equation (I.24.6) from Feynman's lecture notes, which is

$$E = cm^{e_1}(\omega^{e_2} + \omega_0^{e_3})z^{e_4}, \tag{5.1}$$

where $c = 1/4$, $e_1 = 1$ and $e_2 = e_3 = e_4 = 2$. This model has four inputs $x \triangleq (m, \omega, \omega_0, z)$ and five parameters $\theta \triangleq (c, e_1, e_2, e_3, e_4)$. We use three candidate models, the first of which is the ground-truth model in (5.1). The other two models are

$$E = cm^{e_1}\omega^{e_2}\omega_0^{e_3}z^{e_4}, \tag{5.2}$$

$$E = cm^{e_1}(\omega^{e_2} + z^{e_4})\omega_0^{e_3}. \tag{5.3}$$

We can encode the initial values of the parameters of each model, say θ_i in $\mathcal{M}_i$, in the prior distribution $p(\theta_i|\mathcal{M}_i)$.

References

D. M. Borth. A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. Royal Stat. Soc. Ser. B*, 37(1):77–87, 1975.

G. E. P. Box and W. J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A sequential monte carlo algorithm to incorporate model uncertainty in bayesian sequential design. *J. Comput. Gr. Stat.*, 23(1):3–24, 2014.

J. Vanlier, C. A. Tiemann, P. A. Hilbers, and N. A. van Riel. Optimal experiment design for model selection in biochemical networks. *BMC Syst. Biol.*, 8(1), 2014.