Stat260: Bayesian Modeling and Inference **Lecture Date: February 24, 2010**

Reference Priors and Maximum Entropy

Lecturer: Michael I. Jordan *Scribe: Alan Malek*

1 Monte Carlo Estimation of Priors

Recall from last lecture that a reference prior is of the form:

$$f_k(\theta) = \exp\left\{\int p(t^k|\theta)\log\left(\frac{p(t^k|\theta)h(\theta)}{\int p(t^k|\theta)h(\theta)d\theta}\right)dt^k\right\}, \tag{1}$$

where we have included an arbitrary pseudo-prior $h(\theta)$. The integral inside the exponential is often very difficult to evaluate (it has high-dimension) and often does not have a closed form. Thus, we are interested in approximating it.

1.1 Monte Carlo Integration

The basic idea behind Monte Carlo integration is to sample $x^{(i)}$ i.i.d. from $p(x)$ then use the approximation

$$\int f(x)p(x)dx \approx \frac{1}{m}\sum_{i=1}^{m} f(x^{(i)}).$$

where we are asymptotically guaranteed the correct value by law of large numbers.

Applying this idea to the integral in equation (1) yields the following algorithm:

1. pick large k
   pick $h(\theta)$
   pick m

2. for $j = 1, \ldots, m$,
   simulate $\{x_{1_j}, \ldots, x_{k_j}\}$ from $p(x|\theta)$
   compute $c_j = \int p(t^k|\theta)h(\theta)d\theta$
   evaluate $r_j(\theta) = \log\prod_{i=1}^{k} p(x_{ij})h(\theta)d\theta$

3. compute $\pi(\theta) \propto \exp\{\frac{1}{m}\sum_{j=1}^{m} r_j(\theta)\}$
   for a grid of different values of θ

Remark 1. Since we are free to chose $h(\theta)$, it is generally a good idea to chose $h(\theta)$ such that $\int p(t^k|\theta)h(\theta)d\theta$ is easy to compute (e.g. let $h(\theta)$ be the conjugate prior).

2 Relation Between Reference Priors and Maximum Entropy

2.1 Discrete Case

Assume that we have a finite parameter space $\Theta = \{\theta_1, \ldots, \theta_M\}$. In this case, any prior is equivalent to a vector in the M-dimensional simplex:

$$\pi(\theta) = (p_1, \ldots, p_M).$$

We also assume that the $\{\theta_i\}$'s are discrimable, that is

$$D\big(p(x|\theta_i)||p(x|\theta_j)\big) = \int p(x|\theta_i) \log \frac{p(x|\theta_i)}{p(x|\theta_j)} dx > 0 \qquad \forall i \neq j$$

where $D\big(p(x|\theta_i)||p(x|\theta_j)\big)$ is the KL divergence. The posterior is a consistent estimator in the following sense. If $x = (x_1, x_2, \ldots, x_n)$ where x_i are n i.i.d. copies, then:

Lemma 2. *Without loss of generality, assume that θ_1 is the correct parameter. Then*

$$\lim_{n\to\infty} p(\theta_j|x) = \delta_{1,j}$$

where $\delta_{i,j}$ is the Kronecker delta.

Proof.

$$\begin{aligned} p(\theta_j|x) &= \frac{p(x|\theta_j)p_j}{p(x)} \\ &= \frac{p_j \frac{p(x|\theta_j)}{p(x|\theta_1)}}{\sum_k p_k \frac{p(x|\theta_k)}{p(x|\theta_1)}} \\ &= \frac{\exp\{\log p_j + S_j\}}{\sum_k \exp\{\log p_k + S_k\}} \end{aligned}$$

for

$$S_j = \sum_{i=1}^{n} \log \frac{p(x_i|\theta_j)}{p(x_i|\theta_1)}.$$

Since $\frac{1}{n}S_j$ is the average of i.i.d. random variables, the law of large numbers tells us that since θ_1 is the true parameter,

$$\frac{1}{n}S_j \xrightarrow{\text{a.s.}} \int p(z|\theta_1) \log \frac{p(z|\theta_j)}{p(z|\theta_1)} dz = -D\big(p(z|\theta_1)||p(z|\theta_j)\big)$$

where $D\big(p(z|\theta_1)||p(z|\theta_j)\big) = 0$ iff $j = 1$ and $D\big(p(z|\theta_1)||p(z|\theta_j)\big) > 0$ otherwise. Therefore, for $j \neq 1$, we have that $\exp\{\log p_j + S_j\} \xrightarrow{\text{a.s.}} 0$, and for $j = 1$, we have that $\exp\{\log p_j + S_j\} \xrightarrow{\text{a.s.}} p_j$ which implies that $p(\theta_j|x) \xrightarrow{\text{a.s.}} \delta_{1,j}$. □

Thus, in the discrete case, the reference prior is

$$f_k(\theta_i) = \exp\left\{\int p(t^k|\theta_i) \log p(\theta_i|t^k) dt^k\right\} \propto \text{constant}. \tag{2}$$

2.2 Maximum Entropy Principle

We begin with a definition of entropy:

Definition 3. Consider a discrete space of cardinality M. The entropy of a probability density $p = (p_1, p_2, \ldots, p_M)$, denoted $H(p)$, is given by

$$H(p) = -\sum_{i=1}^{M} p_i \log(p_i). \tag{3}$$

The maximum entropy principle states that given some constrains on the prior, the prior should be chosen to be the distribution with the largest entropy which follows these constraints. The most basic constraint is that p lie in the probability simplex; that is, $\sum_i p_i = 1$ and $p_i \geq 0$ for all i. We give two examples of this principle.

Example 4 (no constraints)**.** Without constraints, the maximum entropy principle yields the prior which solves the maximization problem:

$$\begin{aligned} \text{maximize} \quad & -\sum_{i=1}^{M} p_i \log(p_i) \\ \text{subject to} \quad & \sum_i p_i = 1. \end{aligned}$$

Using Lagrange multipliers, we obtain

$$p^* = \operatorname*{argmax}_p H(p) + \lambda(\sum p_i - 1)$$

which has solution

$$p^* \propto \text{constant}.$$

Hence, subject to no moment constraints, the uniform distribution has maximum entropy.

Example 5 (moment constraints)**.** Suppose we have the constraints $f_\ell(p) = c_\ell$ for $\ell = 1, \ldots, L$. Our optimization problem is then:

$$\begin{aligned} \text{maximize} \quad & -\sum_{i=1}^{M} p_i \log(p_i) \\ \text{subject to} \quad & \sum_i p_i = 1 \\ & f_\ell(p) = c_\ell \quad \forall \ell = 1, \ldots, L. \end{aligned}$$

In this case, the Lagrange functional is

$$H(p) + \lambda(\sum p_i - 1) + \sum_{\ell=1}^{L} \mu_\ell (f_\ell(p) - c_\ell).$$

The solution to this optimization problem will be an exponential family with sufficient statistics $f_\ell(p)$.

(4)

The second example illustrates the disadvantage of the maximum entropy principle: it always produces priors in the exponential family, but the exponential family is not broad enough to include some useful priors (such as those that scale like $1/\theta$).

2.3 Continuous Case

There exist analogs of the above results for continuous random variables. We first define the corresponding notion of entropy.

Definition 6. The differential entropy of a probability density p, denoted $h(p)$, is given by

$$h(p) = -\int p(\theta)\log(p(\theta))d\theta. \tag{5}$$

Optimizing over p as functions requires calculus of variations, but the solutions will be in the exponential family and thus suffer the same drawbacks as the discrete case.

Remark 7. Note that we have implicitly used the Lebesgue measure in the definition of differential entropy. If we use other base measures, we get different results in the exponential family. However, this approach detracts from the objectivity of the maximum entropy procedure as we are left with the problem of choosing a base measure.

3 Restricted Reference Priors

We can modify the reference prior framework to allow constraints much like the kind imposed on the maximum entropy priors. Thus, we want the prior that maximizes the mutual information between the prior and posterior while satisfying the constraints

$$\mathbb{E}_p[\theta^i] = \beta_i, i \in \tilde{I}$$

for some index set $\tilde{I}$. A very similar derivation to the standard reference prior derivation yields

$$\pi(\theta) = \pi_0(\theta)\exp\left\{\sum_{i\in\tilde{I}}\lambda_i\theta^i\right\} \tag{6}$$

where $\pi_0(\theta)$ is the prior for the unconstrained case. This prior is known as a tilted distribution.

4 Reference Priors with Nuisance Parameters

Assume that our model is indexed by two parameters:

$$\mathcal{M} = \{p(x|\theta,\lambda) : \theta \in \Theta, \lambda \in \Lambda\}$$

where θ is a parameter of interest and λ is a nuisance parameter - that is, we are not interested in recovering its value. Note that nuisance parameters are quite common in statistics (for example, scale parameters are usually nuisance parameters), and being able to deal with them effectively is a strength of reference priors.

To calculate a reference prior,

1. Condition on θ and find a reference prior $\pi(\lambda|\theta)$ (that is, hold θ constant and find a reference prior for λ by the same procedure as before)
2. Integrate out λ according to this conditional prior (requires the prior to be proper):

$$p(x|\theta) = \int p(x|\theta,\lambda)\pi(\lambda|\theta)d\lambda \tag{7}$$

3. Find $\pi(\theta)$ based on $p(x|\theta)$
4. Set $\pi(\theta, \lambda) = \pi(\theta)\pi(\lambda|\theta)$

Note that $\pi(\lambda|\theta)$ must be proper otherwise the integral in equation (7) will not exist. If $\pi(\lambda|\theta)$ is improper, a common work around is to take a sequence of proper approximations and pass to the limit. For example, this can be done by limiting the domain of λ, e.g. choosing some $\Lambda_1 \subseteq \Lambda_2 \subseteq \cdots \subseteq \Lambda$ and generating a corresponding sequence of priors.

Remark 8. Unfortunately, the procedure for nuisance priors breaks the likelihood principle. Recall that the likelihood principle requires the inference to be dependent only on the likelihood function and the observed data. This is clearly broken as choosing difference nuisance parameters will generate different priors.