Minimum Predictive Discrepancy Parameter Estimation

Kevin S. Van Horn
Leuther Analytics
ksvanhorn@kvanhorn.com

December 4, 2003

1 The Problem

Given a state of information I for which the the sequence of values x_i, $i \geq 1$, is infinitely exchangeable, we would like to estimate x_m from $x_1^n \triangleq x_1, \ldots, x_n$, for $m > n$. Our prior information is that

- $x_i \in X$ for all i;
- there is an additional model variable $\theta \in \Theta$, possibly vector valued;
- $p(\theta|I)\,d\theta = \pi(\theta)\,d\theta$;[1]
- $p(x_i \mid \theta, I)\,d\theta = f(x_i; \theta)\,dx_i$.

Bayes' rule tells us that our posterior pdf for θ is

$$p(\theta \mid x_1^n, I) = g(\theta; x_1^n) \triangleq c(x_1^n)\pi(\theta)\prod_{i=1}^{n} f(x_i; \theta),$$

where $c(x_1^n)$ is a normalizing constant. The posterior predictive pdf for x_m is then

$$p(x_m \mid x_1^n, I) = h(x_m; x_1^n) \triangleq \int_{\Theta} f(x_m; \theta) g(\theta; x_1^n) d\theta.$$

This predictive distribution incorporates both the uncertainty about x_m inherent in the pdf $f(x_m; \theta)$, and our uncertainty about the value of θ.

[1] We assume continuous domains, parameters, and pdfs in this article, but the results generalize to apply also to discrete domains, parameters, and probability mass functions if we allow the use of delta functions in constructing our pdfs.

In some cases the above integral evaluates analytically to a simple form, but often there is no closed form for the integral, or there are practical constraints that lead us to seek a good approximation to this predictive distribution using our given functional form $f(x;\theta)$. That is, we seek to find a specific value $\hat{\theta}$ for which $f(x;\hat{\theta})$ most closely approximates $h(x;x_1^n)$ (as functions of x).

It has been common practice to instead use MAP (maximum *a posteriori*) estimation: simply choose the value $\hat{\theta}$ that maximizes $g(\hat{\theta};x_1^n)$. The rationale is that if n is very large, then the pdf is sharply peaked about the MAP estimate $\hat{\theta}$, and so $f(x;\hat{\theta})$ is a good approximation to $h(x;x_1^n)$.

There are problems with MAP, however. First of all, n is usually not as large as we would like, and the posterior pdf for θ has a significant width. Secondly, the posterior pdf can be significantly skewed, so that its peak is not very representative of the distribution as a whole. Finally, MAP estimation is sensitive to the particular parameterization we choose; for example, if θ is one-dimensional the MAP estimate of θ^{-1} generally does not correspond to the MAP estimate of θ.

Having decided to try to approximate $h(x;x_1^n)$ as closely as we can, what is the proper criterion of closeness to use? For reasons discussed in [1, §2.7.1–2.7.3], the *discrepancy* (a.k.a. *directed divergence* or *Kullback-Liebler divergence*) between the approximating and target distributions is the natural criterion to use. The discrepancy between an approximating pdf q and a target pdf p is

$$\delta(q;p) = \int_X p(x)\log(p(x)/q(x))dx.$$

The discrepancy is always nonnegative, and is zero when q and p are the same distribution.

We can rewrite the discrepancy as a function of q as

$$\delta(q;p) = \text{constant} - \int_X p(x)\log q(x)\,dx.$$

Thus our problem reduces to the following: find the value $\hat{\theta}$ that *maximizes*

$$D(\hat{\theta}) \triangleq \int_X h(x;x_1^n)\log f(x;\hat{\theta})\,dx.$$

Note that this is just the decision-theoretic problem of maximizing expected utility with a utility function of $\log f(x;\hat{\theta})$.

2 Solution for Exponential Models

The above problem has a rather nice solution for exponential models. An exponential model has the form

$$p(x \mid \theta)\,dx = \frac{1}{Z(\theta)}\psi(x)\exp(-\theta\cdot s(x))\,dx$$

where $s(x)$ and θ are both vectors of real values, of the same length. Exponential models arise when one applies Jaynes's principle of maximum entropy [2, 3] to the construction of probability distributions. Defining

$$S(\theta) \triangleq E[s(x) \mid \theta] = \int_X s(x) f(x;\theta) dx,$$

we find that

$$\begin{aligned}
D(\hat{\theta}) &= \int_X \int_\Theta g(\theta; x_1^n) f(x;\theta) \log f(x;\hat{\theta}) d\theta dx \\
&= \int_\Theta g(\theta; x_1^n) \left(\int_X f(x;\theta) \log f(x;\hat{\theta}) dx \right) d\theta \\
&= \int_\Theta g(\theta; x_1^n) E[\log f(x;\hat{\theta}) \mid \theta, I] d\theta \\
&= \int_\Theta g(\theta; x_1^n) (-\log Z(\hat{\theta}) - \hat{\theta} \cdot S(\theta)) d\theta \\
&= -\log Z(\hat{\theta}) - \hat{\theta} \cdot E[S(\theta) \mid x_1^n, I]
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial D}{\partial \hat{\theta}_i} &= -\frac{\partial \log Z}{\partial \hat{\theta}_i} - E[S_i(\theta) \mid x_1^n, I] \\
&= S_i(\hat{\theta}) - E[S_i(\theta) | x_1^n, I].
\end{aligned}$$

(The last equality uses [3, p. 358, (11.60)].) So at an extremum $\hat{\theta}$ of $D(\theta)$ we have

$$S_i(\hat{\theta}) = E[S_i(\theta) \mid x_1^n, I] = \int_\Theta S_i(\theta) g(\theta; x_1^n) d\theta. \tag{1}$$

We must now show that if $\hat{\theta}$ satisfies the above equation it is in fact a maximum, and not a minimum or saddle point, of D. We have a maximum if the Hessian $\partial^2 D / \partial \hat{\theta}^2$ is negative definite. We find

$$\begin{aligned}
\frac{\partial^2 D}{\partial \hat{\theta}_j \partial \hat{\theta}_i} &= \frac{\partial^2 \log Z(\hat{\theta})}{\partial \hat{\theta}_j \partial \hat{\theta}_i} \\
&= E[s_i(x) \mid \hat{\theta}, I] E[s_j(x) \mid \hat{\theta}] - E[s_j(x) s_i(x) \mid \hat{\theta}, I].
\end{aligned}$$

(The last equality uses [3, p. 361, (11.73)]. Thus element (i, j) of the Hessian matrix is the negative covariance between $s_i(x)$ and $s_j(x)$ for a given value of $\hat{\theta}$. Since covariance matrices are positive definite, we then know that the Hessian is negative definite, and we have indeed found a maximum.

2.1 Example: The Exponential Distribution

Let $X = \Theta = \mathcal{R}^+$ (the set of positivve reals) and let $f(x;\theta) = \theta \exp(-\theta x)$. Then

$$\begin{aligned}
s(x) &= x \\
S(\theta) &= \theta^{-1}.
\end{aligned}$$

Let us choose an uninformative prior over θ. Since θ is a scale parameter, this would be

$$\pi(\theta) \propto \theta^{-1}$$

(an improper prior). The posterior pdf for θ is then a gamma distribution $\mathrm{Ga}(\theta; n, t)$, where

$$\begin{aligned} \mathrm{Ga}(\theta; \alpha, \beta) &\triangleq \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\ t &\triangleq \sum_{i=1}^{n} x_i \end{aligned}$$

(see [1, p. 438].) Then

$$\begin{aligned} & E[S(\theta) \mid x_1^n, I] \\ = \; & E[\theta^{-1} \mid x_1^n, I] \\ = \; & \int_0^\infty \theta^{-1} \frac{t^n}{\Gamma(n)} \theta^{n-1} e^{-t\theta} d\theta \\ = \; & \frac{t^n}{\Gamma(n)} \frac{\Gamma(n-1)}{t^{n-1}} \\ = \; & \frac{t}{n-1} \end{aligned}$$

So the optimal estimate is given by $s(\hat{\theta}) = t/(n-1)$, i.e.

$$\hat{\theta} = \frac{n-1}{t},$$

which in this case coincides with the MAP estimate for θ.

2.2 Example: The Normal Distribution

A normal distribution on $X = \mathcal{R}$ has the form

$$f(x; \theta) = \mathcal{N}(x; \mu, \lambda) \triangleq \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right)$$

where μ is the mean and λ the precision (inverse variance) of the distribution. Since

$$f(x; \mu, \lambda) \propto \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) \propto \exp\left(-\frac{\lambda}{2}x^2 + \lambda\mu x\right)$$

we have

$$\begin{aligned} \theta_1 &= \lambda/2 \\ \theta_2 &= -\lambda\mu \\ s(x) &= (x^2, x). \end{aligned}$$

Furthermore, the optimality conditions

$$\begin{aligned} E[x_m \mid \hat{\theta}, I] &= E[x_m \mid x_1^n, I] \\ E[x_m^2 \mid \hat{\theta}, I] &= E[x_m^2 \mid x_1^n, I] \end{aligned}$$

are equivalent to

$$\begin{aligned} E[x_m \mid \hat{\theta}, I] &= E[x_m \mid x_1^n, I] \\ V[x_m \mid \hat{\theta}, I] &= V[x_m \mid x_1^n, I]. \end{aligned}$$

Thus we need only find the mean and variance of the predictive distribution, and choose our point estimates $\hat{\mu}$ and $\hat{\lambda}$ to match these.

If we use the uninformative, improper prior

$$\pi(\mu, \lambda) \propto \lambda^{-1}$$

then the posterior pdf for (μ, λ) is a normal-gamma distribution

$$p(\mu, \lambda \mid x_1^n, I) = \mathcal{N}(\mu; \overline{x}, n\lambda)\mathrm{Ga}\left(\lambda; \frac{1}{2}(n-1), \frac{1}{2}ns^2\right)$$

where $\overline{x} \triangleq \frac{1}{n}\sum_{i=1}^n x_i$ and $s^2 \triangleq \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$. The predictive distribution for x_m is a Student distribution:

$$p(x_m \mid x_1^n, I) = \mathrm{St}(x_m;\, \overline{x},\, (n-1)(n+1)^{-1}s^{-2},\, n-1).$$

Since a Student distribution $\mathrm{St}(x; \mu', \lambda', \alpha)$ has mean μ' and variance $(\lambda')^{-1}\alpha(\alpha-2)^{-1}$, this gives us the MPD estimate

$$\begin{aligned} \hat{\mu} &= \overline{x} \\ \hat{\lambda}^{-1} &= \frac{n+1}{n-3}\,s^2. \end{aligned}$$

This compares with $\hat{\lambda}^{-1} = n(n-3)^{-1}s^2$ for the MAP estimate.

3 Extensions

3.1 Going beyond exponential models

In the general case we can find the MPD estimate by Monte Carlo methods. In particular, consider Monte Carlo estimation of $D(\hat{\theta})$:

$$\begin{aligned} D(\hat{\theta}) &= \int_X h(x; x_1^n) \log(f(x; \hat{\theta}))\, dx \\ &\approx \frac{1}{N}\sum_{i=1}^N \log(f(\tilde{x}_i; \hat{\theta})), \end{aligned}$$

for some suitably large N, where the values $\tilde{x}_i$ are sampled from the posterior predictive distribution $h(x; x_1^n)$. This sampling can be done by first sampling from the posterior distribution over θ, using Markov Chain Monte Carlo if necessary, then sampling $\tilde{x}_i$ from the the sampling distribution $f(x;\theta)$. Maximizing this approximation to $D(\hat{\theta})$ is then the same as maximizing

$$\prod_{i=1}^{N} f(\tilde{x}_i; \hat{\theta}),$$

that is, finding the *maximum likelihood* estimate for θ given the values $\tilde{x}_i$ as data.

One should choose N large enough so that the sampling error in the above approximation is around $\epsilon/2$, where ϵ is the acceptable optimization error tolerance—that is, coming within a difference ϵ of maximizing $D(\theta)$ is acceptable. From the properties of the discrepancy, ϵ should be reasonably small compared to 1.

3.2 Regression

A more interesting case is when the variables x_i have a logical dependence on predictor variables w_i; that is, we have x_i independent of x_j and w_j given θ for $j \neq i$, and

$$p(x_i \mid w_i, \theta, I) = f(x_i; w_i, \theta).$$

This is the usual case in pattern recognition / machine learning problems. For the same reasons as before we choose $\log f(x; w_i, \hat{\theta})$ for our utility function, where $\hat{\theta}$ is our point estimate for θ. Maximizing expected utility then amounts to maximizing

$$D(\hat{\theta}) \triangleq \int_{X \times W} h(x, w; x_1^n, w_1^n) \log f(x; w, \hat{\theta})\, dx\, dw$$

where

$$h(x_m, w_m; x_1^n, w_1^n)\, dx\, dw \triangleq p(x_m, w_m \mid x_1^n, w_1^n, I)\, dx\, dw$$

is the joint posterior predictive distribution for (x_m, w_m) given (x_1^n, w_1^n), for any $m > n$. Unlike the usual complete-data regression problem—but like regression problems with some missing data—we find that we need to infer a distribution for w as well as x conditioned on w.

Again, we can make a Monte Carlo estimate of $D(\hat{\theta})$:

$$D(\hat{\theta}) \quad \approx \quad \frac{1}{N} \sum_{i=1}^{N} \log f(\tilde{x}_i; \tilde{w}_i, \hat{\theta}),$$

where $(\tilde{x}_i, \tilde{w}_i)$, $1 \leq i \leq N$, are sampled from the posterior predictive distribution for (x, w). Again, maximizing this approximation to $D(\hat{\theta})$ is equivalent to finding the maximum likelihood estimate for θ given the values $(\tilde{x}_i, \tilde{w}_i)$ as data.

3.3 Decoupling training and prediction

We can generalize the above even more by allowing prediction to be done using a different family of (conditional) distributions than those used in parameter estimation. That is, we try to find the value $\psi \in \Psi$ maximizing

$$D^\star(\psi) \triangleq \int_{X \times W} h(x, w; x_1^n, w_1^n) \log f^\star(x; w, \psi) \, dx \, dw,$$

where computational or other limitations require us to use a conditional distribution for x given w from a family $f^\star(x; w, \psi)$, $\psi \in \Psi$. Again we proceed by sampling from the posterior distribution for (x, w) and finding the maximum-likelihood estimate for ψ given this artificial data.

Such an approach may be useful in pattern recognition applications where there is a great disparity between the computational resources available for training the recognizer (which is done once) and the computational resources available for running the recognizer. For example, a speaker recognition system must return its answer within a second or so, and may be required to run on hardware of limited capabilities, but it may be reasonable to spend weeks of computation on multiple high-end computers to train the recognizer. Thus during training we may focus on building an elaborate model (or set of models) that best represents all the information we have about our recognition problem, with little concern for computational efficiency, then sample from the posterior predictive distribution to obtain training data for a more computationally-efficient model to be used for actual recognition.

Finally, to generalize even further we can replace the $\log f(x; w, \theta)$ utility function with an arbitrary utility (or negative loss) function $u(\alpha(w; \psi); x, w)$, where

- ψ indicates which of a family of decision rules to use;
- $\alpha(w; \psi)$ is the action the system takes (decision made) upon receiving input w; and
- $u(a; x, w)$ is the utility of action a when $w_m = w$ and $x_m = x$.

Our goal is again to maximize expected utility; that is, we wish to find a value ψ maximizing

$$U(\psi) \triangleq \int_{X \times W} h(x, w; x_1^n, w_1^n) u(\alpha(w; \psi); x, w) \, dx \, dw,$$

Again we can solve this by sampling $(\tilde{x}_i, \tilde{w}_i)$, $1 \le i \le N$, from the posterior predictive distribution for (x_m, w_m), $m > n$, and then finding the value ψ that maximizes

$$\frac{1}{N} \sum_{i=1}^{N} u(\alpha(\tilde{w}_i; \psi); \tilde{x}_i, \tilde{w}_i)$$

or, equivalently, minimizes the negative of this quantity, which is known as the *empirical risk.*

In the terminology of frequentist classification and pattern-recognition methods, the above is just the method of *empirical risk minimization*, only applied to data drawn from the posterior predictive distribution for (x, w) instead of using the original training data (x_i, w_i), $1 \leq i \leq n$. In the frequentist approach a lot of effort is put into dealing with the *bias-variance dilemma*—the fact that, although considering a larger space of decision rules may allow one to achieve a lower empirical risk (lower bias), it also increases the probability that the empirical risk and actual risk (negative expected utility) of the decision rule thus found differ significantly (higher variance). The variance increases as the Vapnik-Chervonenkis dimension of the space of decision rules (a measure of its complexity) increases, and decreases as the size of the training set increases.

The standard example of this phenomenon is trying to model x_i as a polynomial function of w_i plus some noise. Given training data (x_i, w_i), $1 \leq i \leq n$, it is clear that if we expand our set of hypotheses to include all polynomials of order n or larger, we can always find a polynomial that exactly matches the data. However, this polynomial will generally be useless for predicting future values x given m, as it fits the noise rather than the underlying relation between x and w.

In our Bayesian approach, this bias-variance dilemma disappears because we can choose N as large as necessary to ensure that the empirical risk and actual risk of all decision rules considered are, with high probability, close in value. (Note that, in the Bayesian context, "actual risk" is the negative expected utility of the decision rule, with the expectation taken over the posterior predictive distribution, and *not* over some imaginary "true probability distribution" of (x, w).)

References

[1] Bernardo, J. M., & Smith, A. F. M., 1994. *Bayesian Theory*, John Wiley & Sons Ltd., West Sussex, England.

[2] Jaynes, E. T., 1957. "Information Theory and Statistical Mechanics," *Physical Review* 106, pp. 620–630.

[3] Jaynes, E. T., 2003. *Probability Theory: The Logic of Science*, Cambridge University Press.