HBR! 2022/04/23

BAYES FACTOR AND MODEL SELECTION

OVERVIEW

- ONE OF THE BEST FEATURES OF A PROGRAM OF BAYESIAN INFERENCE IS THE BUILT-IN TOOL FOR MODEL COMPARISON AND APPLYING OCCAM'S RAZOR: THE MARGINAL LIKELIHOOD AND THE BAYES FACTOR. WE'LL SHOW HOW THESE ARE DEFINED IN A MOMENT, AND LATER, HOW TO CALCULATE THEM. BUT THE PRIMARY ISSUES ADDRESSED IN THESE NOTES INVOLVE THE ASSOCIATED ISSUES/PROBLEMS W/ THE TECHNIQUE, AND (HOPEFULLY) HOW TO SOLVE THEM:

  - MARGINAL LIKELIHOOD DEPENDS ON THE CHOSEN PRIOR DISTRIBUTION OF PARAMETERS. WHILE A POSTERIOR IS GENERALLY INSENSITIVE TO THE PRIOR (LEADING TO WIDE USE OF "UNINFORMATIVE" / UNIFORM / IMPROPER PRIORS), THE MARGINAL LIKELIHOOD, AND THUS THE ASSOCIATED BAYES FACTORS ARE NOT. THEREFORE A PROPER, WELL-FOUNDED PRIOR MUST BE USED.

  - WHILE THE THEORETICAL EXPRESSION FOR MARGINAL LIKELIHOOD IS STRAIGHTFORWARD, ITS COMPUTATION IN MOST SITUATIONS INVOLVES A MC INTEGRATION USING THE SAMPLE GENERATED BY AN MCMC PROCESS (SAMPLING THE POSTERIOR). THIS IS FROUGHT WITH POTENTIAL PITFALLS.

  - ADDRESSING THESE ISSUES CAN BE A VERY INVOLVED PROCESS — IN SOME SENSE CONTRADICTING THE SIMPLICITY OF THE THEORETICAL IDEA — AND THERE ARE SIMPLER METHODS OF MODEL COMPARISON, E.G., "BAYESIAN INFORMATION CRITERION" (BIC), AIC, DIC ...

- **Marginal likelihood of a model**: We assume an observed data set $\vec{d}$ which is considered one realization of a real-valued vector of random variables $\vec{D}$. As we see below, the marginal likelihood is the evaluation of the conditional probability density of $\vec{D}$, conditioned on the assumption of a particular model hypothesis, $\mathcal{H}$:

In statistical physics the analogous problem is determining the partition function, Z, or the <u>free energy</u>

$$F = -\tau \log(Z)$$

using the Boltzmann factor (unnormalized density):

$$g(\omega|\tau) = \exp\left[-\frac{H(\omega)}{\tau}\right]$$

See refs in Gelman & Meng (SS, 1998).

$$\mathcal{L}(\vec{d}\,|\mathcal{H}) = f_{\vec{D}}(\vec{d}\,|\mathcal{H}) \qquad \left(\begin{array}{c}\text{Marginal}\\ \text{likelihood of } \vec{d}\\ \text{assuming } \mathcal{H}\end{array}\right)$$

Bayesian inference for a particular choice of model hypothesis involves calculating the "posterior" density of that model's parameters — also considered a vector of real-valued random variables, $\vec{\Theta}$ — meaning the conditional density given the data and the model hypothesis (which incorporates all other hidden assumptions in applying that model, including the statistical error model assumed for the likelihood of $\vec{\Theta}$ and the prior distribution of the parameters). This is then expressed using Bayes theorem for densities:

$$f_{\vec{\Theta}}(\vec{\theta}\,|\,\{\vec{D}=\vec{d}\}\cap\mathcal{H}) = \frac{f_{\vec{D}}(\vec{d}\,|\,\{\vec{\Theta}=\vec{\theta}\}\cap\mathcal{H}) \cdot f_{\vec{\Theta}}(\vec{\theta}\,|\mathcal{H})}{f_{\vec{D}}(\vec{d}\,|\mathcal{H})}$$

The event: $\{\vec{D}=\vec{d}\}$ is $\{\omega\in\Omega\,|\,\vec{D}(\omega)=\vec{d}\}$ where the prob. space is $(\Omega, \mathcal{F}, P)$

Where we see that the marginal likelihood is the normalization constant for the posterior, i.e.:

From the law of large numbers (see below), this expression tells us one estimate of $f(\vec{d}|\mathcal{H})$:

$$f(\vec{d}|\mathcal{H}) \approx \frac{1}{N}\sum_{i=1}^{N} f(\vec{d}|\vec{\theta}_i, H)$$

with samples $\vec{\theta}_i$ taken from the <u>prior</u> $f(\vec{\theta}|H)$. We will instead derive an expression in terms of <u>posterior</u> samples, which is of course more efficient.

$$f_{\vec{D}}(\vec{d}\,|\mathcal{H}) = \int f_{\vec{D}}(\vec{d}\,|\,\{\vec{\Theta}=\vec{\theta}\}\cap\mathcal{H}) \cdot f_{\vec{\Theta}}(\vec{\theta}\,|\mathcal{H})\; d\vec{\theta}$$

In these notes we might drop the identifying subscripts on the densities and write the conditioning events in shorthand:

$$f(\theta\,|\,d,\mathcal{H}) \propto f(d\,|\,\theta,\mathcal{H}) \cdot f(\theta\,|\mathcal{H})$$

- Given two model hypotheses, $\mathcal{H}_1$ and $\mathcal{H}_2$, for the same observed data set, we can compute from Bayes theorem:

See my notes on Kolmogorov's Probability → Bayes

$$P(\mathcal{H} \mid \{\vec{D} = \vec{d}\}) = \frac{f_{\vec{D}}(\vec{d} \mid \mathcal{H}) \cdot P(\mathcal{H})}{f_{\vec{D}}(\vec{d})}$$

And therefore the ratio:

$$\frac{P(\mathcal{H}_1 \mid \{\vec{D} = \vec{d}\})}{P(\mathcal{H}_2 \mid \{\vec{D} = \vec{d}\})} = \frac{f_{\vec{D}}(\vec{d} \mid \mathcal{H}_1)}{f_{\vec{D}}(\vec{d} \mid \mathcal{H}_2)} \cdot \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}$$

So, we see that the ratio of the marginal likelihoods serves to adjust the "prior odds" of the two models $P(\mathcal{H}_1)/P(\mathcal{H}_2)$ in the face of new data. In the case that nothing is known a priori, and this ratio is one, the ratio of the marginal likelihoods is the "posterior odds".

- <u>Def'n</u>: Given two model hypotheses, $\mathcal{H}_1$ and $\mathcal{H}_2$, with two associated parameter vectors $\vec{\Theta}_1$ and $\vec{\Theta}_2$, and an observed data set $\vec{d}$ (realization of the random var. $\vec{D}$), the <u>Bayes Factor</u> is defined to be the ratio of the marginal likelihoods:

$$B_{12}(\vec{d}) := \frac{f_{\vec{D}}(\vec{d} \mid \mathcal{H}_1)}{f_{\vec{D}}(\vec{d} \mid \mathcal{H}_2)}$$

$$= \frac{\int f_{\vec{D}}(\vec{d} \mid \{\vec{\Theta}_1 = \vec{\theta}_1\} \cap \mathcal{H}_1) \cdot f_{\vec{\Theta}_1}(\vec{\theta}_1 \mid \mathcal{H}_1)\, d\vec{\theta}_1}{\int f_{\vec{D}}(\vec{d} \mid \{\vec{\Theta}_2 = \vec{\theta}_2\} \cap \mathcal{H}_2) \cdot f_{\vec{\Theta}_2}(\vec{\theta}_2 \mid \mathcal{H}_2)\, d\vec{\theta}_2}$$

A value $B_{12}(\vec{d}) > 1$ suggests that the data provides evidence in favor of $\mathcal{H}_1$ over $\mathcal{H}_2$ and $B_{12}(\vec{d}) < 1$ vice versa.

- **INTERPRETATION OF BAYES FACTORS**: KASS & RAFTERY (1995) PROVIDE A TABLE (FOLLOWING JEFFREYS' ORIGINAL SPECIFICATION IN HIS 1961 BOOK) THAT GIVES THE GENERALLY ACCEPTED QUANTITATIVE INTERPRETATION OF B_{12}:

*— SEE, E.G. HELD & OTT (ARSA 2018)

$\log_{10}(B_{12})$	$2\ln(B_{12})$	B_{12}	EVIDENCE ($\mathcal{H}_1$ OVER $\mathcal{H}_2$)
$0 - 1/2$	$0-2$	$1-3$	NOT WORTH MENTIONING [APPROXIMATELY*: $p > 0.05$]
$1/2 - 1.3$	$2-6$	$3-20$	POSITIVE / SUBSTANTIAL [$p < 0.05$]
$1.3 - 2$	$6-10$	$20-150$	STRONG [$p < 0.01$]
>2	>10	>150	DECISIVE [$p < 0.001$]

- **CALCULATION OF MARGINAL LIKELIHOOD**: FROM BAYES THEOREM FOR DENSITIES WE CAN WRITE

$$f(\vec{d}\,|\,\mathcal{H}) \cdot \frac{f(\vec{\theta}\,|\,\vec{d},\mathcal{H})}{f(\vec{d}\,|\,\vec{\theta},\mathcal{H})} = f(\vec{\theta}\,|\,\mathcal{H})$$

WHICH WE CAN INTEGRATE TO FIND

$$f(\vec{d}\,|\,\mathcal{H}) \int \left[\frac{f(\vec{\theta}\,|\,\vec{d},\mathcal{H})}{f(\vec{d}\,|\,\vec{\theta},\mathcal{H})} \right] d\vec{\theta} = \int f(\vec{\theta}\,|\,\mathcal{H})\, d\vec{\theta} = 1$$

SUCH THAT THE MARGINAL LIKELIHOOD IS THE INVERSE OF THE THE EXPECTATION (UNDER THE POSTERIOR) OF THE INVERSE LIKELIHOOD:

$$f(\vec{d}\,|\,\mathcal{H}) = \left\{ \int \frac{f(\vec{\theta}\,|\,\vec{d},\mathcal{H})}{f(\vec{d}\,|\,\vec{\theta},\mathcal{H})}\, d\vec{\theta} \right\}^{-1}$$

$$= \left\{ \int \left[\frac{1}{f(\vec{d}\,|\,\vec{\theta},\mathcal{H})} \right] f(\vec{\theta}\,|\,\vec{d},\mathcal{H})\, d\vec{\theta} \right\}^{-1} = \left\{ E_{\vec{\Theta}|\vec{D}} \left[\frac{1}{f(\vec{d}\,|\,\vec{\theta},\mathcal{H})} \right] \right\}^{-1}$$

DISCRETIZATION OF AN INTEGRAL YIELDS THE APPROXIMATION:

$$I = \int_\Omega F(\vec{\theta})\, d\vec{\theta} \approx \sum_{i=1}^{N} F(\vec{\theta}_i)\, \Delta V = \sum_{i=1}^{N} F(\vec{\theta}_i)\, \frac{V(\Omega)}{N} \quad \text{“} \int_\Omega d\vec{\theta}$$

$$\cong \frac{1}{N} \sum_{i=1}^{N} F(\vec{\theta}_i) \cdot V(\Omega)$$

WE CAN VIEW THIS AS AN EXPECTATION OF $(F \cdot V(\Omega))$ OVER UNIFORM DISTRIBUTION:

$$p(\vec{\theta}) = \begin{cases} 1/V(\Omega) & \vec{\theta} \in V(\Omega) \\ 0 & \text{OTHERWISE} \end{cases}$$

WRITING:

$$I = E\left[F \cdot V(\Omega)\right] = \int \left[F(\vec{\theta}) \cdot V(\Omega)\right] p(\vec{\theta})\, d\vec{\theta} \approx \frac{1}{N} \sum_{i=1}^{N} F(\vec{\theta}_i)\, V(\Omega)$$

WHERE THE INTEGRAL IS NOW OVER ALL SPACE. THIS RESULT HOLDS FOR <u>ANY</u> PROBABILITY DENSITY, $p(\vec{\theta})$, AND IS CALLED THE <u>LAW OF LARGE NUMBERS</u>:

$$\boxed{E_p[X] := \int X(\vec{\theta}) \cdot p(\vec{\theta})\, d\vec{\theta} = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} X(\vec{\theta}_i)}$$

<u>WHERE THE $\vec{\theta}_i$ ARE SAMPLES FROM $p(\vec{\theta})$</u>.

THUS, WE CAN APPROXIMATELY CALCULATE THE MARGINAL LIKELIHOOD DIRECTLY FROM POSTERIOR SAMPLES, $\{\vec{\theta}_i\}$:

SAMPLES $\vec{\theta}_i$ FROM THE POSTERIOR DIST.

$$f(\vec{d} \mid \mathcal{H}) = \left\{ E_{\vec{\theta} \mid \vec{d}}\left[\frac{1}{f(\vec{d} \mid \vec{\theta}, \mathcal{H})}\right] \right\}^{-1} \approx \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{f(\vec{d} \mid \vec{\theta}_i, \mathcal{H})} \right\}^{-1}$$

THE ESTIMATOR OF THE MARGINAL LIKELIHOOD IS THE <u>HARMONIC MEAN</u> OF THE (PARAMETER) LIKELIHOOD VALUES AS SAMPLED FROM THE POSTERIOR DENSITY FUNCTION.

WE CAN FIND THE SAME EXPRESSION BY IMPORTANCE SAMPLING (SEE APP. A) — AS KASS & RAFTERY DO — USING:

$$f(d \mid H) = E_\theta\left[f(d \mid \theta, H)\right]$$
$$\approx \frac{1}{N} \sum f(d \mid \theta_i, H) \quad \text{w/ } \theta_i \sim f(\theta \mid H)$$
$$\approx \frac{\sum w_i\, f(d \mid \theta_i, H)}{\sum w_i} \quad \text{w/ } \theta_i \sim \text{POST.}$$

WHERE

$$w_i = \frac{f(\theta \mid H)}{\text{POST}}, \quad \text{POST} = f(\theta \mid d, H)$$

SERIOUS ISSUES IN CALCULATING THE MARGINAL LIKELIHOOD:

ALTHOUGH WE FOUND AN EXPRESSION FOR $f_{\vec{D}}(\vec{d} \mid \mathcal{H})$ USING MCMC SAMPLES OF THE POSTERIOR DENSITY, THIS HAS TWO MAJOR ISSUES THAT NEED TO BE ADDRESSED:

① THE HARMONIC MEAN ESTIMATOR IS GUARANTEED TO CONVERGE, BUT HAS INFINITE VARIANCE... IT MAY REQUIRE MORE SAMPLES THAN THE "NUMBER OF ATOMS IN THE OBSERVABLE UNIVERSE"

- <u>THE HARMONIC MEAN OF THE LIKELIHOOD: WORST MC METHOD EVER</u>, RADFORD NEAL

  radfordneal.wordpress.com/2008/08/17/...

- <u>CROSS VALIDATED: COMPUTATION OF MARGINAL LIKELIHOOD FROM MCMC SAMPLES</u> (ANSWER BY XI'AN)

  stats.stackexchange.com/questions/209810

@XI'AN ON STACKEXCHANGE IS: CHRISTIAN ROBERT (WARWICK, PARIS-DAUPHINE) <u>THE BAYESIAN CHOICE</u>

THE SECOND ENTRY HERE PROVIDES OPTIONS ON HOW TO ESTIMATE THE MARGINAL LIKELIHOOD FROM MCMC SAMPLES.

② THE MARGINAL LIKELIHOOD DEPENDS STRONGLY ON THE PRIOR DENSITY $f_{\vec{\Theta}}(\vec{\theta} \mid \mathcal{H})$. SPECIFICALLY, BROAD IMPROPER OR UNINFORMATIVE PRIORS WILL SCREW IT UP.

- GELMAN, CARLIN ET AL <u>BAYESIAN DATA ANALYSIS</u> SECTION 7.4 (3RD ED, 2014) ⟶ DON'T USE BAYES FACTORS

- KRUSCHKE <u>DOING BAYESIAN DATA ANALYSIS</u> SECTION 10.6 (2ND ED, 2014)

THE SECOND ENTRY (KRUSCHKE) SUMMARIZES SOME OPTIONS (OTHER THAN ABANDONING BAYES FACTORS), WHICH COME DOWN TO: PICK A GOOD PRIOR!

IN THE NEXT TWO SECTIONS, WE ADDRESS THESE TWO ISSUES, SPECIFYING STANDARD SOLUTIONS FROM THE LITERATURE.

Estimating Marginal Likelihood from MCMC Samples

- **Harmonic Mean of Likelihood**: We already saw that, by the law of large numbers

$$f_{\vec{D}}(\vec{d} \mid \mathcal{H}) = \lim_{N \to \infty} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[f_{\vec{D}}\left(\vec{d} \mid \{\vec{\Theta} = \vec{\theta}_i\} \cap \mathcal{H}\right) \right]^{-1} \right\}^{-1}$$

where the $\vec{\Theta}_i$ are drawn from the posterior density $f_{\vec{\Theta}}(\vec{\theta} \mid \{\vec{D} = \vec{d}\} \cap \mathcal{H})$. But this estimator can have infinite variance and fail to converge.

It is also said that "the occasional occurance of a value of $\vec{\theta}_i$ w/ small likelihood" has a large effect on the result. —Kass & Raftery 1995

- **Importance Sampling Technique I**: Newton and Raftery (1994) suggest using an importance sampler (see App. A) that is a mixture of the prior and posterior densities, i.e.,

$$f_D(\vec{d}) = \int f_D(\vec{d} \mid \vec{\theta}) \cdot f_\Theta(\vec{\theta})\, d\vec{\theta} = E_{\Theta}\left[f_D(\vec{d} \mid \vec{\theta})\right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} f_D(\vec{d} \mid \vec{\theta}_i) \quad \text{w/ } \vec{\theta}_i \sim f_\Theta(\vec{\theta})$$

$$\approx \frac{\sum_{i=1}^{N} w_i\, f_D(\vec{d} \mid \vec{\theta}_i)}{\sum_{j=1}^{N} w_j} \quad \text{w/ } w_i = \frac{f_\Theta(\vec{\theta}_i)}{p(\vec{\theta}_i)} \text{ and } \vec{\theta}_i \sim p(\vec{\theta})$$

Using

$$p(\vec{\theta}) = \delta \cdot f_\Theta(\vec{\theta}) + (1-\delta)\, f_\Theta(\vec{\theta} \mid \vec{d}) \quad \text{w/ } 0 < \delta < 1$$

"Avoids the instability [of the harmonic mean] and does satisfy a Gaussian central limit theorem." The further suggest that you can just use N samples from the posterior and "imagine" sampling $m \cdot \delta/(1-\delta)$ more from the prior, all w/ likelihoods equal to their expectation (i.e., $f_D(\vec{d} \mid \mathcal{H})$).

- **Adjusted Harmonic Mean**: Gelfand & Dey (JRSSB 1994)

See also:
- Kruschke's *Doing Bayesian* (2nd, 2014), Section 10.3
- Carlin & Louis's *Bayesian Methods* (3rd, 2009), Sec 4.4.1

Proposed an alternative form of the harmonic mean estimator. From Bayes theorem, we can immediately write the marginal likelihood as:

$$\frac{1}{f(\vec{d})} = \frac{f(\vec{\theta}|\vec{d})}{f(\vec{d}|\vec{\theta}) \cdot f(\vec{\theta})}$$

We can then multiply both sides by __any__ proper density, $h(\theta)$, and integrate:

$$\int \frac{h(\vec{\theta})}{f(\vec{d})} \, d\vec{\theta} = \int \frac{f(\vec{\theta}|\vec{d}) \cdot h(\vec{\theta})}{f(\vec{d}|\vec{\theta}) \cdot f(\vec{\theta})} \, d\vec{\theta}$$

$$\frac{1}{f(\vec{d})} = \int \left[\frac{h(\vec{\theta})}{f(\vec{d}|\vec{\theta}) \cdot f(\vec{\theta})}\right] f(\vec{\theta}|\vec{d}) \, d\vec{\theta}$$

Which allows us to estimate

$$f(\vec{d}) = \left\{ E_{\Theta|D}\left[\frac{h(\vec{\theta})}{f(\vec{d}|\vec{\theta}) \, f(\vec{\theta})}\right]\right\}^{-1}$$

$$\simeq \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{h(\vec{\theta}_i)}{f(\vec{d}|\vec{\theta}_i) \, f(\vec{\theta}_i)} \right\}^{-1} \quad \text{w/ } \vec{\theta}_i \sim f(\vec{\theta}|\vec{d})$$

Notice that taking $h(\vec{\theta}) = f(\vec{\theta})$ yields the harmonic mean estimator. A better choice is something with narrower tails, e.g., a multivariate normal that well fits the posterior samples (apparently this does not work well as the dimension grows too large).

- Reverse Logistic (Geyer) $\longrightarrow$

- Bridge Sampling (Gelman & Meng 1998)

  $\longrightarrow$ Gronau "Tutorial on Bridge Sampling" (JMP 2017)

APPENDIX A: LAW OF LARGE NUMBERS AND IMPORTANCE SAMPLING

- I HAVEN'T TRACKED THIS DOWN (PROVEN BY KOLMOGOROV, APPARENTLY, BUT I CAN'T PARSE IT FROM HIS CHAPTER VI) BUT THE <u>LAW OF LARGE NUMBERS</u> SAYS:

  FOR A REAL-VALUED FUNCTION X OF A VECTOR OF RANDOM VARIABLES $\vec{\Theta}$, WITH PROBABILITY DENSITY, $p(\vec{\theta})$, THE ARITHMETIC AVERAGE OF X FOR A SAMPLE $\{\vec{\theta}_i\}$ DRAWN FROM p WILL CONVERGE TO $E_f[X]$ AS THE NUMBER OF SAMPLES GOES TO INFINITY. I.E.,

$$\boxed{\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X(\vec{\theta}_i) = E_p[X] = \int X(\vec{\theta})\, p(\vec{\theta})\, d\vec{\theta} \quad (\vec{\theta}_i \sim p)}$$

  SO, GENERALLY, WE CAN APPROXIMATE AN EXPECTATION WITH AN AVERAGE OVER THE SAMPLE EVALUATIONS. BUT, AS WE SEE IN THESE NOTES ABOUT THE HARMONIC AVERAGE, THERE IS NO GUARANTEE THAT THIS AVERAGE WILL CONVERGE IN A "REASONABLE NUMBER" OF SAMPLES!

- <u>IMPORTANCE SAMPLING</u>: IT IMMEDIATELY FOLLOWS FROM THE ABOVE THAT WE CAN CHANGE FROM ONE SAMPLING DISTRIBUTION TO ANOTHER. FOR $q(\vec{\theta})$ ANOTHER PROBABILITY DENSITY:

$$E_p[X] = \int X(\vec{\theta})\, p(\vec{\theta})\, d\vec{\theta} = \int \left[\frac{X(\vec{\theta})\, p(\vec{\theta})}{q(\vec{\theta})}\right] q(\vec{\theta})\, d\vec{\theta}$$

$$= E_q\left[\frac{X\, p}{q}\right]$$

$$= \lim_{N\to\infty} \sum_{i=1}^{N} \frac{X(\vec{\theta}_i)\, p(\vec{\theta}_i)}{q(\vec{\theta}_i)} \qquad \text{w/ } \vec{\theta}_i \sim q$$

Importance Sampling with Proportional Density

There could be situations where the distribution $p(\vec{\theta})$ is known only up to an unknown constant:

$$p(\vec{\theta}) = c\, p_0(\vec{\theta})$$

where p_0 can be calculated exactly. Let's write

$$E_p[X] = E_q\left[\frac{X p}{q}\right]$$

$$\cong \frac{1}{N}\sum_{i=1}^{N} \frac{X(\vec{\theta}_i)\, p(\vec{\theta}_i)}{q(\vec{\theta}_i)} \qquad (\vec{\theta}_i \sim q)$$

$$\cong \frac{1}{N}\sum_{i=1}^{N} w_i\, X(\vec{\theta}_i) \qquad (\vec{\theta}_i \sim q)$$

If we instead write the weight in terms of p_0

$$w_i := \frac{p(\vec{\theta}_i)}{q(\vec{\theta}_i)} = \frac{c\, p_0(\vec{\theta}_i)}{q(\vec{\theta}_i)} =: c\, w_{0i}$$

Then:

$$E_p[X] \cong \frac{1}{N}\sum_{i=1}^{N} c\, w_{0i}\, X(\vec{\theta}_i) \qquad (\vec{\theta}_i \sim q)$$

And note that:

$$\sum_{i=1}^{N} w_{0i} = \sum_{i=1}^{N} \frac{p_0(\vec{\theta}_i)}{q(\vec{\theta}_i)} \qquad (\vec{\theta}_i \sim q)$$

$$= \frac{N}{c}\left[\frac{1}{N}\sum_{i=1}^{N} 1 \cdot \frac{p(\vec{\theta}_i)}{q(\vec{\theta}_i)}\right] \qquad (\vec{\theta}_i \sim q)$$

$$\cong \frac{N}{c}\, E_q\left[\frac{1 \cdot p}{q}\right] = \frac{N}{c}\, E_p[1] = \frac{N}{c}$$

Thus we can write:

$$\boxed{E_p[X] \cong \frac{\sum_{i=1}^{N} w_{0i}\, X(\vec{\theta}_i)}{\sum_{j=1}^{N} w_{0j}} \qquad \left(w_{0i} = \frac{p_0(\vec{\theta}_i)}{q(\vec{\theta}_i)}\ ;\ \vec{\theta}_i \sim q\right)}$$

SIMILARLY, IT IS EASY TO SHOW THAT, IF q IS ONLY KNOWN UP TO A CONSTANT

$$q(\vec{\theta}) = \tilde{c}\, q_0(\vec{\theta})$$

THEN:

$$E_p[X] \cong \frac{\sum_{i=1}^{N} \tilde{w}_{0i}\, X(\vec{\theta}_i)}{\sum_{j=1}^{N} \tilde{w}_{0j}} \qquad \left(\tilde{w}_{0i} = \frac{p(\vec{\theta}_i)}{q_0(\vec{\theta}_i)} \,;\, \vec{\theta}_i \sim q \right)$$

SO THE ESTIMATOR WORKS WHETHER THE "PROPORTIONAL DENSITY" IS p OR q.