KL computation

Martin D. Weinberg

July 22, 2019

Abstract

Some comments on your notes dated July 14, 2019. In short, it does not seem that you have computed the cross-validation-like KL divergence. The idea is to find the minimum signal-to-noise cut such that does not overfit the noise. It is possible that I misunderstand your explanation, but the reference distribution *should not* be the untruncated density estimate but the distribution as a function of signal-to-noise cut. Finally, I do strongly recommend computing the true likelihood cross-validation using the same ideas that I suggested for KL, as I will outline below.

1 Introduction

The overall procedure for KL divergence, I think, should be as follows:

1. Partition the ensemble of particles randomly into M partitions.
2. For each partition, compute the density estimate from the basis-function expansion for some particular signal-to-noise cut
3. For each density estimate, compute the Kullback-Leibler divergence for the other $M-1$ samples.
4. Examine the run of mean of the $M-1$ samples as function of the signal-to-noise cut.

This is a cross-validation-like procedure. Intuitively, I expect the diverge to be larger for no signal-to-noise cut, because the noise from one subsample will not be present in the other subsamples. As the signal-to-noise cut increases, the density features that represent the noise will be truncated away.

As you make the signal-to-noise ratio larger, I would expect divergence to level off and decrease more slowly. Note, the divergence is not estimating the goodness of fit! So as the signal-to-noise ratio cut increases, the divergence should not increase.

2 Definitions

1. Partition your particles in to M groups $\{N_i\}$ where $\{N_i\}$ denotes the ith partition with N_i particles. The total ensemble is $\{N\} = \bigcup_{i=1} M\{N_i\}$

2. For each partition $\{N_i\}$, compute the basis-function expansion for some signal-to-noise cut: $\hat{\rho}(\mathbf{x}|\{N_i\},\gamma)$ where γ indexes the signal-to-noise level cut on the coefficients.

3. The KL divergence for subset i relative to subset j for some particular signal-to-noise cut is:

$$D_{KL}(i||j)[\gamma] = \int d\mathbf{x}\hat{\rho}(\mathbf{x}|\{N_i\},\gamma)\log\left(\frac{\hat{\rho}(\mathbf{x}|\{N_i\},\gamma)}{\hat{\rho}(\mathbf{x}|\{N_j\},\gamma)}\right)$$

   This integral could be done by explicit quadrature. But as before, we can estimate the integral using a Monte Carlo procedure based on the particle distribution itself as:

$$D_{KL}(i||j)[\gamma] \approx \tilde{D}_{KL}(i||j)[\gamma] = \sum_{k=1}^{N_i} m_k \log\left(\frac{\hat{\rho}(\mathbf{x}_k|\{N_i\},\gamma)}{\hat{\rho}(\mathbf{x}_k|\{N_j\},\gamma)}\right)$$

   where $\mathbf{x}_k$ is one of the particles in sample i. It is probably best to adjust the masses so that $\sum_{k=1}^{N_i} m_k = 1$.

4. As you pointed out, for every i there are $M-1$ values $j \neq i$ so we may compute:

$$\langle D_{KL}(i||\cdot)\rangle[\gamma] = \frac{1}{M-1}\sum_{i\neq j}\langle D_{KL}(i||j)$$

   and finally:

$$\langle\langle D_{KL}[\gamma]\rangle\rangle \equiv \langle\langle D_{KL}(\cdot||\cdot)[\gamma]\rangle\rangle = \frac{1}{M}\sum_{i=1}^{M}\langle D_{KL}(i||\cdot)[\gamma]\rangle$$

3 Goodness of fit

We can play the same game to evaluate the fit using the mean-integrated square error. If we knew true density $\rho(\mathbf{x})$ then we could estimate how close any signal-to-noise truncated estimate is as:

$$L[\gamma] = \int d\mathbf{x}\left(\hat{\rho}(\mathbf{x}|\gamma) - \rho(\mathbf{x})\right)^2$$

Expanding we find:

$$L[\gamma] = \int d\mathbf{x}\hat{\rho}^2(\mathbf{x}|\gamma) - 2\int d\mathbf{x}\rho(\mathbf{x})\hat{\rho}(\mathbf{x}|\gamma) + \int d\mathbf{x}\rho^2(\mathbf{x}). \tag{1}$$

3.1 The first term of equaton (1)

The first term can be reduced to sum of things we know by biorthogonality of the series expansion used to construct $\hat{\rho}$ itself. Assume that our basis constructed from the eigenfunctions of the Laplacian have the following form:

$$\hat{\rho}(\mathbf{x}) = \sum_{lm}\sum_{n} c_{l,m,n} Y_{l,m}(\theta,\phi) d_{l,m,n}(r) \tag{2}$$

$$\hat{\Psi}(\mathbf{x}) = \sum_{lm}\sum_{n} c_{l,m,n} Y_{l,m}(\theta,\phi) p_{l,m,n}(r) \tag{3}$$

$$\tag{4}$$

where

$$\nabla^2 Y_{l,m}(\theta,\phi) p_{l,m,n}(r) = 4\pi G Y_{l,m}(\theta,\phi) d_{l,m,n}(r) \tag{5}$$

with

$$\int d\mathbf{x} Y_{l,m}(\theta,\phi) p_{l,m,n}(r) Y_{l'-m'}(\theta,\phi) d_{l',-m',n'}(r) \tag{6}$$

$$= 4\pi G \delta_{l,l'}\delta_{m,m'} \int dr r^2 p_{l,m,n}(r) d_{l,m,n'}(r) \tag{7}$$

$$= 4\pi G \delta_{l,l'}\delta_{m,m'}\delta_{n,n'}. \tag{8}$$

These properties allow us to reduce the three-dimensional integral into a sum of one-dimensional integrals as follows:

$$\int d\mathbf{x}\hat{\rho}^2(\mathbf{x}|\gamma) = \int d(\cos\theta)d\phi \int dr r^2 \hat{\rho}(\mathbf{x}|\gamma)\hat{\rho}(\mathbf{x}|\gamma) \tag{9}$$

$$= \sum_{n,n'} c_{l,m,n} c_{l,-m,n'} \int dr r^2 d_{l,m,n}(r) d_{l,-m,n'}(r). \tag{10}$$

Note that the $d_{l,m,n}$ are not mutually orthogonal, so this term is a sum of one dimensional quadratures. To evaluate this, one uses the fact that $Y_{l,m}(\theta,\phi) = (-1)^m Y^*_{l,m}(\theta,\phi)$ to get $d_{l,-m,n}(r)$ from $d_{l,m,n}(r)$ with the appropriate parity since $\hat{\rho}$ must be real. In your notation with the C and S terms, the S terms have opposite sign for $-m$ when m is odd. That is, my $c_{l,m,n} = C_{l,m,n} + iS_{l,m,n}$ where $c_{l,-m,n} = C_{l,m,n} + (-1)^m iS_{l,m,n}$ with $d_{l,m,n}(r) = d_{l,m,n'}(r)$. Therefore, $c^*_{l,m,n} = c_{l,-m,n}$ so that the density is real. the sum over $\pm m$ will give: terms like:

$$\begin{aligned} c_{l,m,n}c_{l,-m,n'} + c_{l,-m,n}c_{l,m,n'} &= [C_{l,m,n} + iS_{l,m,n}][C_{l,m,n'} - iS_{l,m,n'}] + \\ &\quad [C_{l,m,n} - iS_{l,m,n}][C_{l,m,n'} + iS_{l,m,n'}] \\ &= 2[C_{l,m,n}C_{l,m,n'} + S_{l,m,n}S_{l,m,n'}]. \end{aligned}$$

3.2 The second term of equation (1)

We can estimate the second term for some particular subset of particles drawn from the phase space to estimate the integral $\int d\mathbf{x}\rho(\mathbf{x})$. That is, the integral in the second term may be estimated by:

$$\int d\mathbf{x}\rho(\mathbf{x})\hat{\rho}(\mathbf{x}|\gamma) \approx \sum_{k=1}^{N} m_k \hat{\rho}(\mathbf{x}_k|\gamma).$$

So, just as in the previous section, we can estimate $\hat{\rho}(\mathbf{x}|\gamma)$ for one of the partitions i and compute the first term by direct quadrature and the second term as the mean over the other partitions $j \neq i$. The details are the same as for KL.

3.3 The third term of equaton (1)

The third term is an unknown constant because we do not know $\rho(\mathbf{x})$. However, because it is a constant, we can just drop it.

3.4 Discussion

At γ, the noise will increase L. At very large values of γ, the divergence will be small but the value of L will not look like $\rho(\mathbf{x})$. So, presumably, $L[\gamma]$, the cross-validation likelihood, will have a minimum at some value of γ when the two densities are closest. Using KL and $L[\gamma]$ together may be very helpful.