Introducing Doubt in Bayesian Model Comparison

Assignment -3

Suprio Dubey
(2013036)

1°

Main Question) what is the main message of the paper? (NOTE: usually good papers have 1 message per paper!)
Q1) what is the χ2-per-degree-of-freeedom rule-of-thumb? How is the χ2distribution related to this test?
Q2) what is the Jeffrey scale?
Q3) how it is used and what is the Bayesian Information Criterion (BIC)?
Q4) it is always possible to Taylor expand the likelihood around the maximum? What does it represent the "curvature" term of this expansion?
Q5) What is the Fisher information matrix and how is related to the likelihood?
Q6) are you convinced by this paper? Report some idea to improve the use of Doubt or to avoid it if deemed unnecessary.

Ans 1. χ2-distribution

For statistically independent data $\vec{d} = (d_1 \ldots d_n)$ the likelihood of $\vec{d}$ is factorised into the likelihoods of single data

$$\mathcal{L}(\vec{d}\,|\,\vec{n}, I) = \prod_{i=1}^{n} \mathcal{L}(d_i\,|\,\vec{n}, I) \qquad \text{where } \vec{n} = \text{parameter vector}$$

If the noise ($n_i = d_i - \mu_i$) is gaussian distributed with variance σ^2 and average μ_i where $\mu_i = \mu_i(\vec{n})$ (depends on $\vec{n}$). The expected value can still be a f^n of the parameters. We can write the likelihood as:

$$\mathcal{L}(\vec{d}\,|\,\vec{n}, I) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-1}{2}\frac{(d_i - \mu_i(\vec{n}))^2}{\sigma_i^2}\right]$$

Now, taking an improper uniform prior i.e $P(\vec{n}\,|\,\sigma, I) = \begin{cases} \text{const} & \forall\ \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$

The Posterior ($\mathcal{P}$) will be directly proportional to the likelihood ($\mathcal{L}$)

$$\mathcal{P}(\vec{n}\,|\,\vec{d}, I) \propto \mathcal{L}(\vec{d}\,|\,\vec{n}, I) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-1}{2}\frac{(d_i - \mu_i(\vec{n}))^2}{\sigma_i^2}\right]$$

Now taking the log posterior -

$$L = \log \mathcal{P}(\vec{n}\,|\,\vec{d}, I) = \text{const} - \frac{1}{2}\sum_{i=1}^{n} \frac{(d_i - \mu_i(\vec{n}))^2}{\sigma_i^2}$$

So the posterior is max when $\sum_{i=1}^{n} \frac{(d_i - \mu_i(\vec{n}))^2}{\sigma_i^2} = \chi^2$ is min. → (A)

Now, as we have seen χ^2 distribution let us see how it is related to χ^2-per-degree-of-freedom-rule of thumb

χ2-per-degree-of-freeedom rule-of-thumb

Let us first define DOF.

In the above eq^n (A) σ_i (std. deviation) acts as a scale of the difference b/w d_i & $\mu_i(\vec{n})$. If $\mu_i(\vec{n})$ is a reasonable approximation to d_i then, $\frac{d_i - \mu_i(\vec{n})}{\sigma_i}$ will be close to ± 1, we add up those "1" and anticipate it to be equal to $\sum_{i=1}^{n} 1 = n$ which is the number of data points. But there are free

parameters (let us say for gaussian fit, the amplitude, std deviation, mean) which we vary to minimise the χ^2. So, the net effect is reduction of the total number of DOF (degree of freedom) which we define as

$$dof = \text{number of data pts.} - \text{number of free parameters}$$

(χ2/dof) rule-of-thumb :

It says that for normally distributed data points. $\chi^2_{/dof}$ is distributed as a χ^2 distribution where $\forall$

- ☆ $\chi^2/dof \approx 1$ — The model is appropriate for the data
- ☆ $\chi^2/dof << 1$ — The model is overfitting (too many free parameters)
- ☆ $\chi^2/dof >> 1$ — The fit is unsatisfactory (The model that's assumed doesn't have enough freedom or doesn't fit the data correctly)

Ans 2. **Jeffrey's Scale**

It is an emperical scale for evaluating the strength of evidence when we are comparing two models M_0 vs M_1. It is an emperical prescription for translating the values Bayes factor into strength of Belief.

$\lvert \ln B_{01} \rvert$	Odds.	Strength Of Evidence
< 1.0	$\lesssim 3:1$	Inconclusive
1.0	$\sim 3:1$	Weak Evidence
2.5	$\sim 12:1$	Moderate Evidence
5.0	$\sim 150:1$	Strong Evidence.

In the above table we are comparing two Models M_0 & M_1

Now, I will mention what Bayes Factor is; For that let us recall Bayes theorem

$$P(M_j|d) = \frac{P(d|M_j)\,P(M_j)}{P(d)} \quad -1)$$

$P(M_j|d)$: Posterior probability

$P(M_j)$: Prior.

$P(d) = \sum_i P(d|M_i)\,P(M_i)$ - normalising const.

The Bayesian Evidence $\boxed{p(d|M_j) = \int d\theta\, p(d|\theta_j, M_j)\, p(\theta_j|M_j)}$

$\theta_j \rightarrow$ free parameter.

$P(\theta_j|M_j) \rightarrow$ prior probability distribution

$P(d|\theta_j, M_j) \rightarrow$ likelihood.

So, given two competing models M_0, M_1

$$\text{Bayes Factor} = B_{01} = \frac{p(d|M_0)}{p(d|M_1)}$$

large B_{01} means preference of M_0 and small B_{01} means preference of M_1

Ans 3. & 4. Bayesian Information Criterion (BIC)

(The answer to question 4. Will be highlighted.)

It is a criteria for selecting models based on the likelihood function.

$$P(d|M_j) = \int P(d|\theta_j, M_j)\, P(\theta_j|M_j)\, d\theta \qquad -3.1)$$

$d = n$ independent and identically distributed observation $(n_1, n_2 \ldots n_n)$ each of which may be a vector.

Let us rewrite the likelihood without the model for simplicity

$$P(d) = \int P(d|\theta)\, P(\theta)\, d\theta \qquad -3.2)$$

Let us define:

$$g(\theta) = \log \{ P(d|\theta)\, P(\theta) \}_{|\theta=\bar{\theta}} \qquad \text{writing } \bar{\theta} = \theta_{max}. \qquad -3.3)$$

$$g(\theta) = g(\bar{\theta}) + (\theta-\bar{\theta})^T g'(\bar{\theta}) + \frac{1}{2}(\theta-\bar{\theta})^T g''(\bar{\theta})(\theta-\bar{\theta}) \qquad -3.4)$$

$$g'(\bar{\theta}) = \frac{dg(\theta)}{d\theta_i} = 0 \because g(\bar{\theta}) \text{ is max at } \bar{\theta}. \qquad -3.5)$$

$$g''(\bar{\theta}) = \frac{d^2 g(\theta)}{d\theta_i d\theta_j} \qquad -3.6)$$

$$g(\theta) = g(\bar{\theta}) + \frac{1}{2}(\theta-\bar{\theta})^T g''(\bar{\theta})(\theta-\bar{\theta}) \qquad -3.7)$$

Ans 4) Yes

The above appx done to Taylor enpand the likelihood around θ_{max} is only good when θ is close to θ_{max}. However, when n (the number of data) is large $P(d|\theta)$ the likelihood is concentrated about its maximum and declines sharply as one moves away. So, only values θ close to θ_{max} contributes more to the likelihood.

The "curvature" term $g''(\bar{\theta})$ is the Hessian Matrix H which is $(K \times K)$ matrix where K is the number of parameters. The -ve inverse of the Hessian Matrix i.e $(-H^{-1})$ gives the covariance matrix. The diagonal terms of which tells about the variance of each parameter and the off diagonal terms tell us the linear combination b/w the parameters, and if 0, it implies that they are uncorrelated.

Now, using eq^n 3.7) we can write eq 3.2) as:

$$P(d) = \int \exp[g(\theta)]\, d\theta$$

$$\approx \exp[g(\bar{\theta})] \int \exp\left(\frac{1}{2} (\theta - \bar{\theta})^T g''(\bar{\theta}) (\theta - \bar{\theta}) \right) d\theta \quad -3.8)$$

Let us see how can we solve the integral

$$Z = \int d\theta \; \exp\left\{ -\frac{1}{2} \theta^T H \theta \right\}$$

Let us do a change of variable. $\vec{\theta} = O\vec{y}$ (rotation). Such that O are normal eigen values of H.

- $\det(O) = 1$
- $(O^T O)_{ij} = \delta_{ij}$
- $(O^T H O) = \lambda$

$\lambda = \text{diagonal}(\lambda_1, \lambda_2 \dots \lambda_n)$

So,

$$\theta^T H \theta = (Oy)^T H (Oy) = y^T (O^T H O) y = \sum_{j=1}^{n} \lambda_j y_j^2$$

$\det(O^T O) = 1 \Rightarrow \det(O^T) \det(O) = (\det[O])^2 = 1$

Now,

$$\prod_{j=1}^{N} \lambda_j = \det(O^T H O) = \det(H)$$

So, we get $Z = \int dy_j \exp\left(-\frac{1}{2}\sum_{j=1}^{N} \lambda_j y_j^2\right) =$

$$Z = \prod_{j=1}^{N} \int \exp\left\{-\tfrac{1}{2}\lambda_j y_j^2\right\} dy_j = \prod_{j=1}^{n} \sqrt{\frac{2\pi}{\lambda_j}}$$

$$\left(\int_{-\infty}^{\infty} \exp\left[\frac{-x^2}{2\sigma^2}\right] dx = \sigma\sqrt{2\pi}\right)$$

on comparing we get.

$$Z = (2\pi)^{n/2} (\det(H))^{-1/2}$$

$$\therefore P(d) = \exp[g(\bar{\theta})] (2\pi)^{K/2} |H|^{-1/2} \quad - 3.9)$$

the error in the above eq^n is $O(n^{-1})$

$$\log P(d) = \log P(d|\bar{\theta}) + \log P(\bar{\theta}) + K/2 \log(2\pi) - \frac{1}{2}\log|H| + O(n^{-1})$$

For large samples

→ $O(n^{-1}) \rightarrow$ const as $n \rightarrow \infty$.

→ $\bar{\theta} \approx \hat{\theta}$ where $\hat{\theta}$ is Maximum Likelihood Estimator.

→ $H = nI$ where I is Fisher Information Matrix.

$|H| = n^K I$ $\forall$ $(K \times K)$ Matrix K number of parameters

$$\log P(d) = \log p(d|\hat{\theta}) + \log P(\hat{\theta}) + \frac{K}{2}\log(2\pi) - \frac{K}{2}\log|n|$$

$$- \frac{1}{2}\log(I) + O(n^{-1/2}) \qquad 3.10)$$

The two approx ie $H = nI$
$\&|H| = n^K I$ introduce this error.

Complexity: $O(n)$ $O(\log n)$ rest terms are $O(1)$ and lower

$$\log P(d) = \log p(d|\hat{\theta}) - (K/2)\log n + O(1) \quad -3.11)$$

log integrated likelihood = Max likelihood − correction term. The error doesn't vanish even if there are ∞ no. of data.

A/c to empirical experience it is found the above eqn to be more accurate in practice than the $O(1)$ error term would suggest. Actually, the error is of much smaller order of magnitude for a particular, reasonable choice of normal distribution with mean $\hat{\theta}$ & variance I^{-1}.

The prior distribution contains the same amount of info as would on average, a single observation.

$$P(\hat{\theta}) = (2\pi)^{-K/2}|i|^{1/2}$$

$$\log P(\hat{\theta}) = -\frac{K}{2}\log(2\pi) + \frac{1}{2}|i| \quad -3.12)$$

using it in eqn 3.10) we get.

$$\boxed{\log(d) = \log P(d|\hat{\theta}) - (K/2)\log n + O(n^{-1/2})}.$$

$$\Rightarrow \boxed{\log(d|M_j) = \ln L_{max} - \frac{K}{2}\log n + O(n^{-1/2})}$$ - Std expression for BIC.

So, for the above given prior method the error in the appx of the log integ. likelihood is $O(n^{-1/2})$ rather than $O(1)$ which is smaller for moderate to large sample sizes and tends to 0 as $n \to \infty$.

We can use the std. expression for BIC to estimate the evidence.

Ans 5. **Fisher Information Matrix**

It is a $(K \times K)$ matrix whose $(i,j)^{th}$ element is

$$I_{ij} = \left\langle -\frac{\partial^2 \log P(d/\theta)}{d\theta_i d\theta_j} \Big|_{\theta=\hat{\theta}} \right\rangle \qquad d = (n_1, \ldots n_n)$$

The expectations are taken over values x, with θ held fixed.

It represents the average of (minus) of the curvature of the log likelihood computed in the maximum likelihood value.

The reason it is important, is because its inverse defines the smallest achievable error bars ie the likelihood curvature in the maximum defines the best possible precision we can attain.

6. The paper is convincing because it introduces an extension to Bayesian model selection called Bayesian doubt where instead of relative performance of two models given by the Bayes Factor we can calibrate the absolute value of the evidence. It is advantageous over the relative performance of two models because in the later case we always get a preferred model even if both the models doesn't fit the data well.
It is useful in case of huge data set with multi-dimensional parameter which is difficult to visualise. We can compute the doubt giving the trustworthiness of the model

In the paper (arXiv:0811.2415v1 [physics.data-an] 14 Nov 2008)

we saw that measure the relative change of doubt wrt to the prior doubt rather than directly looking at doubt.

ie $\mathcal{R} = \frac{D}{P(x)}$

So to a large extent our inference depends on the assumed prior of the unknown models.
So choosing appropriate prior to known models. might help in estimating the doubt.

We can avoid the use of doubt in models which are complex with more free parameters than the true models.

Overall Message of the paper:

We can use doubt to discover model. It helps us to enlarge the space of known models and avoid unnecessary complexity.

2.

The follow-up paper is: https://arxiv.org/abs/1005.3655

Main Question) what it is the main message of the paper?

Q1) What is the purpose of MultiNest algorithm?

Q2) What is the Savage-Dickey density ratio?

Q3) What are CosmoMC and CAMB, and what are their differences?

Q4) In your opinion this work added something new to the problem of Dark Energy? Please explain.

Bonus Question -not required to answer) To my knowledge there are no other attempts, from the Bayesian perspective, to address the problem of unknow model assessment. Can you find something related to this problem other than the previous references?

Ans 1: MultiNest Algorithm:

The MultiNest Algorithm implements the Nested Sampling Algorithm

Let us define Nested Sampling.

It is a Monte Carlo Technique which efficiently evaluate the Bayesian Evidence and also produces posterior inferences as by-product.

Suppose we have a complex model M_1 with prior $P(\theta|M_1)$ which reduces to simpler model (M_0) for a certain value of the parameter,

eg at $\theta = \theta^*$

So let us take the multidemensional evidence integral

$$P(d|M_i) = \int d\theta_i \, p(d|\theta_i, M_i)\, \pi(\theta_i|M_i) \quad - 1.1$$

& recast into a 1-D integral.

We do that by defining prior volume x as $dx = \pi(\theta)d\theta$

Such that

$$x(\lambda) = \int_{\mathcal{L}(\theta) > \lambda} \pi(\theta)\, d\theta \quad - 1.2$$

The integral is over the parameter space enclosed by the iso-likelihood contour $\mathcal{L}(\theta) = \lambda$.

So, $\pi(\lambda)$ gives the volume of parameter space above a certain level λ of the likelihood.

Then the Bayesian Evidence eq (2) can be written as

$$p(d) = \int_0^1 \mathcal{L}(\pi)\, d\pi \qquad -1.3$$

$\mathcal{L}(\pi)$ is inverse of eq (1.2)

Samples from $\mathcal{L}(x)$ can be obtained by drawing uniformly samples from the likelihood volume within the iso-contour surface defined by λ.

1-D integral of eq (1.3) can be obtained via quadrature (determining the area)

$$p(d) = \sum_i \mathcal{L}(\pi_i) w_i$$

$$w_i = \text{weight} = \frac{1}{2}(\pi_{i-1} - \pi_{i+1})$$

MultiNest Algorithm:

1. Set $i=0$; initial conditions $X=0$, $E=0$
2. Sample N points $\{\theta_j\}$ randomly from $\pi(\theta)$ and calculate likelihood.
3. Set $i = i+1$
4. Find point with the lowest likelihood value (L_i)
5. Remaining Prior Volume $X_i = t_i X_{i-1}$ where $\langle t_i \rangle = N/N+1$ or $P(t_i|N) = N t_i^{N-1}$
6. Increase the evidence $E \to E + L_i w_i$
7. Remove lowest pt. from active set.
8. Replace the new pt sampled from $\pi(\theta)$ within hard-edged region $\mathcal{L}(\theta) > L_i$
9. If $L_{max} X_i < \alpha E$ (where α is some tolerance)
   $\Rightarrow E \to E + X_i \sum_{j=1}^{N} L(\theta_j)/N$; stop
   else go to (3).

Ans 2. Savage-Dickey Density Ratio:

The SDDR method is used for model selection involving nested models. It reduces the computational effort needed to calculate the Bayes factor of two nested models.

The marginal likelihood for a Model M is given by

$$p(d|M) = \int p(d|\theta, M)\, \pi(\theta|M)\, d\theta$$

If we are comparing two parameters model M_1 with a restricted submodel M_0 with only one free parameter ψ and fixed parameter $\omega = \omega_*$. We assume the prior is seperable i.e

$$\pi(\omega, \psi|M_1) = \Pi(\omega|M_1)\, \pi(\psi|M_0)$$

then the Bayes factor $B_{01} \equiv \dfrac{P(d|M_0)}{P(d|M_1)}$

let us compute the integral

$$P(d|M_0) = \int d\psi\, \pi_0(\psi)\, P(d|\psi, \omega_*) \qquad -1)$$

$$P(d|M_1) = \int d\psi\, d\omega\, \pi_1(\psi, \omega)\, P(d|\psi, \omega) = q \qquad -2)$$

$\because$ Models are nested the likelihood fn for M_0 is just a slice at const $\omega = \omega_*$ of likelihood of model M_1 i.e $P(d|\psi, \omega)$

Now we multiply and divide B_{01} by $p(\omega_*|d)$.

$\equiv p(\omega = \omega_*|d, M_1) \rightarrow$ marginalised posterior for ω under M_1, evaluated at ω_*.

Using that

$$P(\omega_*, \psi|d) = P(\omega_*|d)\, P(\psi|\omega, d)$$

we get $P(\omega_*|d) = P(\omega_*, \psi|d) \,/\, P(\psi|\omega_*, d)$ at any ψ.

So $B_{01} = P(\omega_* | d) \int \frac{d\psi\, \pi_0(\psi)\, p(d|\psi, \omega_*)\, p(\psi|\omega_*, d)}{q\, p(\omega_*, \psi | d)}$ -3

$= p(\omega_* | d) \int d\psi\, \frac{\pi_0(\psi)\, p(\psi | \omega_*, d)}{\pi_1(\omega_*, \psi)}$ -4)

using $\boxed{p(\omega_*, \psi | d) = \frac{p(d|\omega_*, \psi)\, \pi_1(\omega_*; \psi)}{q}}$

Now assuming $\pi_1(\psi | \omega_*) = \pi_0(\psi)$ which holds in the case of seperable priors (usual in cosmology)

$\pi_1(\omega, \psi) = \pi_1(\omega)\, \pi_0(\psi)$

$\therefore$ eq 4) is normalised marginal posterior it becomes -

$= \int d\psi\, \frac{\pi_0(\psi)\, p(\psi | \omega_*, d)\, p(\omega_* | d)}{\pi_1(\omega_*)\, \pi_0(\psi)}$

$B_{01} = \boxed{\frac{p(\omega | d, M_1)}{\pi(\omega | M_1)}}$ - SDDR -5)

So, for the nested model we only need properly normalised value of marginal posterior at $\omega = \omega_*$ under extended model M_1,

For a Gaussian Prior centered on ω_* with std. deviation Δw and a gaussian likelihood with mean $\hat{\mu}$ & width $\hat{\sigma}$ gives .

$$\ln B_{01}(\beta, \lambda) = \frac{1}{2}(1 + \beta^{-2}) - \frac{\lambda^2}{2(1+\beta^2)} \quad -6)$$

where
- $\lambda = \dfrac{|\hat{\mu} - \omega_*|}{\hat{\sigma}}$ SIGMA DISCREPANCY
- $\beta = \hat{\sigma}/\Delta\omega$ → Volume Reduction Factor. (the factor by which the accessible parameter space is reduced after the arrival of the data)

∀ uninformative data: $\lambda = 0$ & $\beta = 1$

unless $\Lambda >> 1$ (null hypothesis is rejected with many σ & there is hardly any need of model comparison)

we can measure informative data: $I = -\ln\beta$

∀ strongly informative data $I = -\ln\beta \gtrsim 0$
i.e $\beta^{-1} >> 1$

we can write eqⁿ 6) as

$$\ln B_{01} \approx I - \lambda^2/2 \quad \text{(informative data)}$$

→ large I signals a large volume of wasted parameter space under the prior and favours the simple Model.

→ large λ favours more complex model because of the mismatch b/w measured and predicted value of the extra parameter.

Answer 3.

CAMB(Code for Anisotropies in the Microwave Background):
Integrates (using the Runge-Kutta integration scheme) the Boltzmann equations(in the CDM gauge) and generates predictions for observables.

It is a Boltzmann code (linear perturbation theory i.e. evolve each mode independently of all others) to solve a realization for a given cosmological model and generate observables to be fitted with data, e.g., the expansion rate of the universe, anisotropy power spectrum of perturbations of the cosmic microwave background and the matter power spectrum.

CosmoMc:
Markov-Chain Monte-Carlo (MCMC) engine for exploring cosmological parameter space. Uses the likelihoods and data made available to find the model that better represents the data.
By default CosmoMC uses a simple Metropolis- Hasting algorithm or an optimized fast-slow sampling method (which works for likelihood with many fast nuisance parameters like Planck).

Metropolis-Hastings Algorithm :

- ☐ We start from data, d, and a model, M(θ), that describes the data and depends on parameters θ = {θ1,θ2,...,θN}
- ☐ Do a random walk, instead, where in each step from θ_i we move to θ_{i+1} at a random direction
- ☐ Compute the likelihood at both points (e.g. by running CAMB) and compare them, i.e. compute $r_{i+1|i} \equiv L(d|M, \theta_{i+1})/L(d|M, \theta_i)$; i.e. for Gaussian likelihood: $-2\ln L \propto \chi 2 \equiv (d-M(\theta))^T \Sigma^{-1} (d-M(\theta))$
- ☐ If $r_{i+1|i} \geq 1$ accept θ_i, i.e. add it to the chain
- ☐ If $r_{i+1|i} < 1$ do not move to new point, but add old point to the chain again instead (i.e. the multiplicity of that point will increase)
- ☐ Choose a new direction and repeat

Answer 4.

The work done in the paper didn't add something new to the problem of Dark Energy.
A list of known models including the possible extension of Dark Energy Sector and non- zero curvature of the universe was adapted.
The findings were unaffected by addition of new models to the list of known models. We see that a weak evidence is provided in favor of the unknown model by the upper bound of the Bayesian evidence for a currently unknown dark energy model against the ΛCDM. The ΛCDM remains a sufficient qualitative description of the currently availaible observation.

So, the main message of the paper was how we apply the methodology of Bayesian Doubt introduced by Starkman et al. (2008). to the Dark Energy equation of state, by comparing an absolute upper bound on the Bayesian evidence for a presently unknown dark energy model against a collection of known models including a flat ΛCDM scenario. ΛCDM remains a sufficient qualitative description of the currently availaible observation.

We saw the extended application of Bayesian model selection to define an absolute scale of goodness of fit for models, rather than just a relative one, such as the Jeffreys' scale.