# Bayesian Inference and Model Regression

**Please submit your homework in pdf form, which can be either a scanned copy of your hand-written answers, or, computer generated documents (e.g., via word/latex etc). In either case, please include computer calculated results and necessary scientific figures added to your documents.**

**1. Is CDM substructure a plausible theory to explain gravitational "flux-ratio anomaly" observations? (6 pts)**

The cold dark matter (CDM) cosmology has been a prevailing theory describing the matter content of the Universe. Together with a constant term – the "cosmological constant $\Lambda$" describing the cosmos mass-energy density budget in form of dark energy, the theory of $\Lambda$CDM can successfully explain distributions of galaxies on large scales[1]. CDM predicts a great number of dark matter substructures surviving through the assembly process during galaxy formation, more abundant than any "warmer"-flavored dark matter theories. Their "dark" presence adds density perturbation/fluctuation to the smoothly-distributed gravitational potential of their host galaxies. A strong gravitational lensing system can be used to test their presence and constrain their abundance. In this case, the light from a background galaxy (or a quasar) can be splitted into several paths as it travels through the gravitational potential of a foreground galaxy! This simply causes the background source to cast into multiple copies of images, each being magnified or demagnified depending on where it appears in the (projected) potential of the lens galaxy (like an optical lens magnifing objects behind)! Astronomers can fit the *positions* of the multiple images within their measurement uncertainties (error!) to constrain the mass density distribution model of the lens galaxy in the foreground. Interestingly, the best estimate model often fails to reproduce the relative magnifications (*flux ratios*) among the multiple images. This phenomenon has been observed over decades and is referred to as the "flux-ratio anomaly" problem. Theorists predict that the CDM substructures sitting in the lens galaxy may be the major cause of the "flux-ratio anomaly" problem. Here is the question: using CDM cosmology $N$-body simulations, theorists calculated the CDM substructure abundance and through the so-called "ray-tracing" numerical experiment, they predicted a 25% upper limit of the fraction of lensing systems in which "flux-ratio anomalies" are expected to happen due to the presence of CDM substructures. Observerations have shown 8 out of 10 multiply lensed quasar systems exhibiting evidence of "flux-ratio anomalies". Using the Bayesian analysis assuming a flat prior for your model, can you rule out the proposed theoretical explanation at a confidence level of $\alpha = 5\%$? How about $\alpha = 1\%$? What would a maximum likelihood estimate (assuming Gaussian approximation) tell you? Please write down the key Bayesian formula for this case, the relevant statistical distribution, your calculation, reasonings and results.

**2. Model Regression (18 pts)**

1. A polynomial function of degree 3 is given by:

$$y \equiv f(x) = 7 + 2(x - 0.2) - 3(x - 0.5)^2 - 6(x - 0.8)^3, \tag{1}$$

defined on $x \in [0, 1]$. Use Monte Carlo method to generate $N = 20$ points to randomly (uniformly) sample $x \in [0, 1]$, and work out the expected $y$ values at these $x$ locations.

[1] On small scales, however, the CDM theory (along) is facing a few chellanges. Various alternative dark matter theories have been proposed to solve the small-scale issues.

2. Now let us assume a Gaussian error behavior that at each sampled location $x_i$, the error in $y_i$ around $y_i^* = f(x_i)$ has a Gaussian standard deviation of $\sigma = 0.1$. Use Monte Carlo method to generate an error $\delta_{y_i}$ according to such a distribution for every data pair $(x_i,\, y_i = y_i^* + \delta_{y_i})$.

3. The 20 data pairs $(x_i,\, y_i)$ generated above now compose a measurement sample, with $x$ being the independent variable and $y$ the dependent variable. Suppose you have no idea about the true model through which this sample is generated. *The only information that we have is the Gaussian error behavior which has a standard deviation of* $\sigma = 0.1$ *(in reality, this can be a known measurement uncertainty).* Now use what you have learnt to play the linear regression game, and find out the most optimal models (and their parameters) which can describe the existing data and make reliable predictions for future measurements (elsewhere within the domain).

In your final answer, please give: (1) the mathematical forms of your most optimal models (you are not limited to use the polynomial format); (2) the procedure that you obtain the *best* models (and model parameters); (3) the reason how do you justify such models are most optimal.

In your final answer, please also plot: (1) the polynomial distribution given by Eq.1; (2) the Monte-Carlo generated 20 data pairs $(x_i,\, y_i)$; (3) the best regression model to fit the data.

* Change your $\sigma = 0.1$ to $\sigma = 1.0$ , do you get the same *best* models? Why not?