# Example Sheet 2
# Example Class: Friday, 02 Mar 2017, 2:00pm, MR5

Part III Astrostatistics

## 1 Linear Regression with $(x, y)-$measurement error and intrinsic dispersion: Quasar X-ray Spectral Slopes vs. Eddington Ratios

In class we examined the problem of linear regression of the quasar X-ray spectral index vs. bolometric luminosity in the presence of measurement error in both quantities and intrinsic dispersion. (Regression is also described in Feigelson & Babu, Chapter 7, Ivezic et al., Chapter 8, and Kelly et al. 2007, The Astrophysical Journal, 665, 1506). Consider the probabilistic generative model described in class:

$$\xi_i \sim N(\mu, \tau^2) \tag{1}$$

$$\eta_i | \xi_i \sim N(\alpha + \beta \xi_i, \sigma^2) \tag{2}$$

$$x_i | \xi_i \sim N(\xi_i, \sigma_x^2) \tag{3}$$

$$y_i | \eta_i \sim N(\eta_i, \sigma_y^2) \tag{4}$$

The astronomer measures values $\mathcal{D} = \{x_i, y_i\}$ with known measurement error variances $\{\sigma_{x,i}^2, \sigma_{y,i}^2\}$, for $i = 1, \dots, N$ quasars.

1. Write down the joint distribution $P(x_i, y_i, \xi_i, \eta_i | \alpha, \beta, \sigma^2, \mu, \tau^2)$ for a single quasar.

2. Derive the observed data likelihood function for all the quasars:

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) = \prod_{s=1}^{N} P(x_i, y_i | \alpha, \beta, \sigma^2, \mu, \tau^2). \tag{5}$$

   Show all steps and maximally simplify.

3. Write a code to find the maximum likelihood estimate, if given $\{x_i, y_i\}$ and their known measurement variances for $i = 1 \dots N$ quasars. (If you were unable to complete steps 1 & 2, use Eqs 19 - 23 of Kelly et al. 2007 with K= 1, $\gamma_1 = \pi_1 = 1$). Test your code on simulated data you generate from the model with known true parameter values. Find an approximate 95% confidence interval for each parameter using the observed Fisher information. (Use a generic optimisation library or toolbox to numerically minimise a given function, e.g. scipy.optimize in Python, fmincon in Matlab, or optim in R, or equivalent).

4. Using the dataset provided online ("example_sheet2_prb1_data.txt"), find the maximum likelihood estimates (MLEs) of the parameters $\alpha, \beta, \sigma^2, \mu, \tau^2$, and their uncertainties.

5. Suppose the distribution of the latent (true) independent variables $\{\xi_i\}$ is assumed to be "non-informative" or flat. Take the limit $\tau \to \infty$ of Eq. 5 to derive $L_{\tau\to\infty}(\alpha, \beta, \sigma^2)$.

6. Compare your MLE for $\beta$ using Eq 5 against what you get using ordinary least squares (OLS), minimum $\chi^2$, FITEXY modified $\chi^2$ methods, and the MLE with $L_{\tau\to\infty}(\alpha, \beta, \sigma^2)$.

   (a) Ordinary Least Squares minimises the residual sum of squares (RSS) with respect to the parameters:
   $$RSS = \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 \,. \tag{6}$$

   (b) Minimum $\chi^2$ or Weighted Least Squares minimises the following with respect to the parameters:
   $$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma_{y,i}^2}. \tag{7}$$
   What minimum value do you find for the reduced $\chi^2_\nu = \chi^2/(N-2)$?

   (c) The FITEXY methods (Press et al. *Numerical Recipes in C*) minimise an "effective" $\chi^2$ statistic that takes in account $x$-measurement errors
   $$\chi^2_{EXY} = \sum_{i=1}^{N} \frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma_{y,i}^2 + \beta^2 \sigma_{x,i}^2}. \tag{8}$$
   What minimum value do you find for the reduced $\chi^2_{EXY,\nu} = \chi^2_{EXY}/(N-2)$?

   (d) The maximum likelihood solution assuming a non-informative distribution on the $\{\xi_i\}$ is obtained by minimising:
   $$-\log L_{\tau\to\infty}(\alpha, \beta, \sigma^2) = \lim_{\tau\to\infty} -\log L(\alpha, \beta, \sigma^2, \mu, \tau^2). \tag{9}$$

7. State and employ appropriate non-informative priors on the parameters $\alpha, \beta, \sigma^2, \mu, \tau^2$ defined in Eqs. 1 - 5. Construct and implement a MCMC algorithm to sample from the posterior probability density:
   $$P(\alpha, \beta, \sigma^2, \mu, \tau^2 |\, \mathcal{D}) \propto L(\alpha, \beta, \sigma^2, \mu, \tau^2) \times P(\alpha, \beta, \sigma^2, \mu, \tau^2) \tag{10}$$
   Run 4 independent chains to diagnose convergence using the Gelman-Rubin ratio. Remove "burn-in" and use the combined chains to compute the marginal distributions of the parameters, and compare against the point estimates you previously obtained with the other methods.

# 2 Importance Sampling for Bayesian Estimates of the Milky Way Mass using Angular Momentum Measurements

Look up the paper Patel et al. 2017, "Orbits of massive satellite galaxies II. Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations." *Monthly Notices of the Royal Astronomical Society*, 468, 3428. Use the measurements in Table 1 and the online data from the Illustris simulation to estimate the Milky Way mass using angular momentum $j$ and the rotational velocity $v_{\max}$ of the Large Magellanic Cloud (LMC). In this context, the Milky Way is the central (host) galaxy of the system, and the LMC is a "satellite" galaxy.

1. Let $\boldsymbol{x} = (v_{\max}, j)$ be the latent parameters, and let $\boldsymbol{d} = (v_{\max}^{\mathrm{obs}}, j^{\mathrm{obs}})$ be their measured values, with uncertainties shown in Table 1. Write down the likelihood function $P(\boldsymbol{d}|\,\boldsymbol{x})$, assuming Gaussian measurement errors.

2. The Illustris simulation implicitly encodes a joint distribution between these latent dynamical parameters of satellites and the $\log_{10}$ masses of central (or host) galaxies, $P(\boldsymbol{x}, \log_{10} M)$. Assuming this exists, write down an expression for the normalised posterior probability density of the Milky Way $\log_{10}$ mass.

3. Write down an expression for the posterior mean estimate of the $\log_{10}$ MW mass in terms of integrals involving the likelihood and prior.

4. Using an arbitrary importance sampling distribution $Q(\boldsymbol{x}, \log_{10} M)$ from which we can easily draw samples, rewrite this expression in terms of expectations with respect to $Q$.

5. Rewrite this expression now assuming now that the importance sampling distribution is the same as the prior $Q(\boldsymbol{x}, \log_{10} M) = P(\boldsymbol{x}, \log_{10} M)$. Approximate this expression with weighted sums over the prior samples, suitable for the Monte Carlo method, and derive the importance weights.

6. Use the Illustris host-satellite data in the online file "Patel17b_Illustris_Data_KM.txt" as samples from the prior. Use the columns labelled "MVIR", "SATVMAX" and "SATJMAG". Compute the importance weights, and estimate the posterior mean and standard deviation of the $\log_{10}$ MW mass, given the LMC data $\boldsymbol{d}$. Also compute an effective sample size using Eq. B2 in the paper, and compare against the number of samples from the prior.

7. Using the bandwidth Eq. B1, create a weighted KDE representation of the posterior distribution $P(\log_{10} M|\boldsymbol{d})$. Plot it over a KDE representation of the prior $P(\log_{10} M)$.