EXAMPLE SHEET 1
EXAMPLE CLASS: 16 FEB 2017, 2:30PM, MR5

PART III ASTROSTATISTICS

1. Problem 1: Calibrating Supernova Magnitudes: Inferring an intrinsic distribution with measurement error

Type Ia supernovae (SNe Ia) are thermonuclear explosions of white dwarf stars. They are used as "standard candles," objects with a narrow range of absolute magnitude (log luminosity), so their distances can be judged from their apparent magnitudes (log apparent brightness or flux). Suppose the absolute magnitudes of SNe Ia come from an intrinsic Gaussian distribution with unknown population mean $\bar{M}$ and variance σ^2_{int}.

$$M_s \sim N(\bar{M}, \sigma^2_{\text{int}}) \tag{1}$$

To "calibrate" the SNe Ia, and determine these parameters, we need a "training set" of SNe Ia with independent distance estimates. Astronomers use the "distance modulus", a logarithmic measure of distance. Suppose for $s = 1 \dots N$ SNe Ia, we have independent estimates $\{\hat{\mu}_s\}$ of the distance moduli $\{\mu_s\}$, with known Gaussian uncertainties $\sigma_{\mu,s}$:

$$\hat{\mu}_s | \, \mu_s \sim N(\mu_s, \sigma^2_{\mu,s}). \tag{2}$$

The astronomer measures the apparent magnitudes $\{m_s\}$ of the supernovae using telescopes on Earth. These estimates $\hat{m}_s$ have Gaussian uncertainties known variance $\sigma^2_{m,s}$.

$$\hat{m}_s | \, m_s \sim N(m_s, \sigma^2_{m,s}). \tag{3}$$

The true quantities are related by the inverse square law, which in logarithmic form is:

$$m_s = M_s + \mu_s \tag{4}$$

Therefore an estimator of M_s is $\hat{M}_s = \hat{m}_s - \hat{\mu}_s$.

(1) What is the sampling distribution of the estimator $\hat{M}_s$ around the true value? Derive $P(\hat{M}_s | \, M_s)$.
(2) Write down the joint distribution $P(\hat{M}_s, M_s | \bar{M}, \sigma^2_{\text{int}})$.
(3) Derive the observed data likelihood function $L(\bar{M}, \sigma^2_{\text{int}}) = \prod_{s=1}^N P(\hat{M}_s | \bar{M}, \sigma^2_{\text{int}})$. Show all steps, evaluate all integrals, and maximally simplify.
(4) Write a code (in Python, Matlab, or R, etc.) to find the maximum likelihood solution of the above, if given data $\{\hat{m}_s, \hat{\mu}_s\}$ and their known variances for $s = 1 \dots N$ SNe Ia. Test your code on simulated data you generate from the model

Date: February 10, 2018.

with known true parameter values. Your code should also find an approximate 95% confidence interval for each parameter using the observed Fisher information.

(5) Apply your code to the data provided online for this problem and report MLE estimates and uncertainties.

(6) Bootstrap the dataset 100 times, and apply your code to each bootstrap samples. Compare the bootstrap distribution of MLE estimates to the uncertainty you found using the Fisher information on the original dataset. See Ivezic, §4.5 & F&B §3.6.2 to read about bootstrap.

2. Problem 2: Correcting for Interstellar Dust with Empirical Bayes

One problem with using Type Ia supernovae as distance indicators is dust. A random, unknown amount of interstellar dust along the line of sight in the supernova's galaxy absorbs, scatters, and therefore dims the light, so the supernova appears farther away. The dust also makes the colour of the SN look redder. By estimating the *reddening* E_s of the supernova from its apparent colours, we can correct for the dust effect. However, one obstacle is that we do not ever observe the *intrinsic* colour C_s of any supernova, so we do not know exactly what the original colour of the SN was before the reddening effect of dust. We can however, build a probabilistic generative model for the observed SN colour distribution. Suppose the intrinsic colour C_s of a SN s comes from a Gaussian distribution with mean μ_C and variance σ_C^2:

$$C_s \sim N(\mu_C, \sigma_C^2) \tag{5}$$

A commonly used model for the reddening distribution is exponential: $E_s \sim \text{Exponen}(\tau)$, i.e.

$$P(E_s|\,\tau) = \tau^{-1} \exp(E_s/\tau) \tag{6}$$

for $E_s \geq 0$ or zero otherwise. The observed, apparent colour $\hat{O}_s$ is the sum of the intrinsic colour, the reddening, and measurement error.

$$\hat{O}_s = C_s + E_s + \epsilon_s \tag{7}$$

where the measurement erro $\epsilon_s \sim N(0, \sigma_s^2)$ is a mean-zero Gaussian random variable with known variance.

(1) Write down the joint distribution $P(\hat{O}_s, C_s, E_s|\tau, \mu_c, \sigma_C^2)$.

(2) Derive the observed data likelihood function $L(\tau, \mu, \sigma_C^2) = \prod_{s=1}^{N} P(\hat{O}_s|\,\tau, \mu, \sigma_C^2)$. Show all steps, evaluate all integrals, and maximally simplify.

(3) Write a code to find the maximum likelihood estimate of the above, if given apparent colour measurements $\{\hat{O}_s\}$ and their known measurement variances for $s = 1 \ldots N$ SNe Ia. Test your code on simulated data you generate from the model with known true parameter values. Your code should also find an approximate 95% confidence interval for each parameter using the observed Fisher information.

(4) Apply your code to estimate the parameters τ, μ, σ_C^2 from the Table 3 dataset from Jha, Riess & Kirshner. (2007), *"Improved Distances to Type Ia Supernovae with*

Multicolor Light-Curve Shapes: MLCS2k2". The Astrophysical Journal, 659, 122. This will be provided online in ASCII form.

(5) Bootstrap the dataset 100 times, and apply your code to each bootstrap samples. Compare the bootstrap distribution of MLE estimates to the uncertainty you found using the Fisher information on the original dataset. See Ivezic, §4.5 & F&B §3.6.2 to read about bootstrap.
(6) Now fixing the parameters τ, μ, σ_C^2 to your MLE estimated values $\hat{\tau}, \hat{\mu}, \hat{\sigma}_C^2$, and derive an expression for the posterior density of the reddening E_s of each SN s, $P(E_s | \hat{O}_s; \hat{\tau}, \hat{\mu}, \hat{\sigma}_C^2)$.
(7) For each SN s in the Jha dataset, compute the posterior mean and mode using this expression. Also find the 68% credible interval containing the highest posterior density (HPD).

3. Quasar X-ray spectral index vs. bolometric luminosity Linear Regression with (x, y) measurement error and intrinsic dispersion

Consider the probabilistic generative model described in class:

$$\xi_i \sim N(\mu, \tau^2) \tag{8}$$

$$\eta_i | \xi_i \sim N(\alpha + \beta \xi_i, \sigma^2) \tag{9}$$

$$x_i | \xi_i \sim N(\xi_i, \sigma_x^2) \tag{10}$$

$$y_i | \eta_i \sim N(\eta_i, \sigma_y^2) \tag{11}$$

The astronomer measures values (x_i, y_i) with known measurement variances (σ_x^2, σ_y^2).

(1) Write down the joint distribution $P(x_i, y_i, \xi_i, \eta_i | \alpha, \beta, \sigma^2, \mu, \tau^2)$.
(2) Derive the observed data likelihood function

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) = \prod_{s=1}^{N} P(x_i, y_i | \alpha, \beta, \sigma^2, \mu, \tau^2) \tag{12}$$

Show all steps, evaluate all integrals, and maximally simplify.
(3) Write a code to find the maximum likelihood estimate of the above, if given $\{x_i, y_i\}$ and their known measurement variances for $i = 1 \ldots N$ quasars. Test your code on simulated data you generate from the model with known true parameter values. Your code should also find an approximate 95% confidence interval for each parameter using the observed Fisher information.
(4) Using the data provided online, find the maximum likelihood estimate of the slope β using this model. Compare your MLE against what you get using ordinary least squares (OLS) and the FITEXY modified χ^2 algorithm (Press et al. *numerical recipes in C*, or Tremaine et al. (2002), The Astrophysical Journal, 574, 740.