Example Sheet 1
Example Class: 19 Feb 2019, 1pm, MR5

Part III Astrostatistics

1 Combining Multiple Distance Estimates

Suppose N Cepheid variable stars are observed in the same external galaxy, and analyses of their brightness time series yields unbiased, independent estimates of their distance moduli, $\hat{\mu}_i$, $i = 1 \ldots N$. The distance modulus is a logarithmic measure of their distance d from us on Earth.

$$\mu = 25 + 5 \log_{10}[d \text{ Mpc}^{-1}], \tag{1}$$

where Mpc is a mega-parsec, a unit of distance. Since the distances between galaxies are very much larger than the distances within galaxies, the Cepheid stars must all be effectively at the same true distance modulus μ. However, the estimates $\hat{\mu}_i$ have different sampling variances σ_i^2 around the true μ, because of observational heteroskedastic measurement error. We wish to combine the N independent estimates from the N individual stars to determine the "best" single estimate of the distance modulus μ to the galaxy.

1. Consider $N = 2$ stars. Consider all estimators that are linear combinations of the data $\hat{\mu}_1, \hat{\mu}_2$: $\hat{\mu} = \alpha_1 \hat{\mu}_1 + \alpha_2 \hat{\mu}_2$. What restriction is required of all *unbiased* linear estimators of μ?

2. For $N = 2$, what is the sampling variance $\text{Var}[\hat{\mu}]$ of the unbiased linear estimators in part 1? Find the *minimum variance* unbiased linear estimator by solving for the appropriate coefficients. Show that they can be expressed as $\alpha_i = K \sigma_i^\gamma$, and determine K and γ. What is the variance of the minimum variance unbiased linear estimator?

3. Now generalise to $N > 2$. Consider all linear estimators of the form $\hat{\mu} = \sum_{i=1}^{N} \alpha_i \hat{\mu}_i$. What are the coefficients of the minimum variance unbiased linear estimator? Verify that they satisfy the first- and second-derivative conditions for a local minimum.

4. For $N > 2$, suppose all the uncertainties of the individual estimates are the same, $\sigma_i = \sigma$ for $i = 1, \ldots, N$. What is the variance of the minimum variance unbiased linear estimator, and how does it scale with the number of stars N?

5. Now suppose, because of systematic uncertainties, the distance errors for $N > 2$ stars are jointly Gaussian and correlated between stars, with known pairwise covariances $\text{Cov}[\hat{\mu}_i, \hat{\mu}_j] \equiv C_{ij} = \sigma_i \, \sigma_j \, \rho_{ij}$, and correlation coefficients $|\rho_{ij}| < 1$. What is required for the matrix $\boldsymbol{C}$ to be a valid covariance matrix? Assuming $\boldsymbol{C}$ is a valid covariance matrix, derive the maximum likelihood estimator $\hat{\mu}_{\text{MLE}}$. Compute the bias and variance of the MLE. Compare the variance to the Cramér-Rao bound. You may leave your answers in terms of elements Λ_{ij} of the inverse of the covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{C}^{-1}$.

2 Estimating the Hubble Constant on the Local Distance Ladder

Type Ia supernovae (SNe Ia) are thermonuclear explosions of white dwarf stars. They are used as "standard candles," objects with a narrow range of peak absolute magnitude (log luminosity), so their distances can be judged from their apparent magnitudes (log apparent brightness or flux). Suppose the peak absolute magnitudes of SNe Ia come from an intrinsic Gaussian distribution with population mean M_0 and variance σ^2_{int}:

$$M_s \sim N(M_0, \sigma^2_{\text{int}}) \tag{2}$$

for every supernova s. The true absolute magnitude is related to the true apparent magnitude m_s via the true distance modulus μ_s, which is a logarithmic measure of the true distance d_s.

$$m_s = M_s + \mu_s. \tag{3}$$

The definition of the distance modulus is $\mu = 25 + 5\log_{10}[d \text{ Mpc}^{-1}]$, where Mpc is a megaparsec, a unit of astronomical distance. For every supernova s, we obtain an estimate of its peak apparent magnitude $\hat{m}_s$ with known error variance $\sigma^2_{m,s}$.

$$\hat{m}_s | \, m_s \sim N(m_s, \sigma^2_{m,s}) \tag{4}$$

Type Ia supernovae can be observed at great distances, and are used to estimate the current expansion rate of the Universe, the Hubble constant H_0. However, to do this, their luminosities (absolute magnitudes) must be calibrated. *Calibration* is the statistical task of estimating M_0 and σ^2_{int}. (In most of this problem we will assume known values of σ^2_{int} for simplicity). Ideally, calibration is done simultaneously with the estimation of H_0 to ensure proper propagation of uncertainties.

Suppose we have a calibrator set of $k = 1, \ldots, K$ supernovae located in nearby galaxies in which we can observe Cepheid variable stars. We can use the observations of the apparent brightness and periodicity of the variable stars, along with the Cepheid period-luminosity relation, to estimate their distances. For each calibrator supernova, we have an unbiased distance modulus estimate $\hat{\mu}_{C,k}$ with a Gaussian error with variance $\sigma^2_{C,k}$, obtained from the Cepheid stars in the same galaxy.

$$\hat{\mu}_{C,k} | \, \mu_k \sim N(\mu_k, \sigma^2_{C,k}) \tag{5}$$

We also have a much larger, "Hubble Flow" set of $i = 1, \ldots, N$ supernovae which are much further away, so the Cepheids stars cannot be observed in their galaxies. However, they are far enough away that they participate in the smooth, overall expansion of the Universe. Thus, they follow the Hubble Law, the linear relation between their recession velocities $v_i = c\, z_i$ and their distances d_i:

$$d_i = \frac{c}{H_0} z_i \tag{6}$$

where c is the speed of light and z_i is the redshift. Assume the redshift is measured exactly for each supernova i. In terms of the distance modulus,

$$\mu_i = 25 + 5\log_{10}\left[\frac{c\, z_i}{H_0} \text{ Mpc}^{-1}\right]. \tag{7}$$

The units of the Hubble Constant are km s^{-1} Mpc^{-1}. Let

$$\theta = 5\log_{10}[H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})] \tag{8}$$

1. Write down the likelihood function of (M_0, θ) in terms of the data of the calibrator set $\{\hat{m}_k, \hat{\mu}_{C,k}\}$ and the Hubble flow sample $\{\hat{m}_i, z_i\}$, and the relevant variances.

2. Assume that all error variances are known, as well as the intrinsic dispersion σ_{int}. Derive the maximum likelihood estimators for (M_0, θ).

3. Evaluate the bias and variance your estimators $(\hat{M_0}, \hat{\theta})$, and compare against the Cramér-Rao bound.

4. Simplify for the case where each source of error is homoskedastic, i.e. $\sigma_{C,k} = \sigma_C$ for all calibrators, and $\sigma_{m,s} = \sigma_m$ for all supernovae. Derive an expression for the variance of $\hat{\theta}$, and propagate the error to derive the standard deviation of $\hat{H_0}$.

5. Suppose $K = 19, N = 300$ and $\sigma_C = 0.1$, $\sigma_m = 0.05$, and $\sigma_{\text{int}} = 0.1$. What uncertainty (standard deviation of error) would you expect for $\hat{H_0}$? What would decrease the uncertainty the most: obtaining one more calibrator supernova, or one more Hubble flow supernova?

6. Return to the heteroskedastic errors. Look up the paper "Measuring the Hubble constant with Type Ia supernovae as near-infrared standard candles" by Dhawan, Jha & Leibundgut (2018, Astronomy & Astrophysics 609, A72). For the calibrator sample use the "m_J","σ_{fit}", "μ_{Ceph}" and "σ_{Ceph}" columns in Table 1 for $\{\hat{m}_k, \sigma_{m,k}, \hat{\mu}_{C,k}, \sigma_{C,k}\}$, respectively. For the Hubble flow sample, use the "z_{CMB} (flow-corrected)", "m_J" and "σ_{fit}" columns in Table 2 for $\{z_i, \hat{m}_i, \sigma_{m,i}\}$. Assuming a value of $\sigma_{\text{int}} = 0.10$, what estimate do you obtain for the Hubble constant and its standard error? What if $\sigma_{\text{int}} = 0.16$?

7. (Bonus) Now allowing σ_{int} to be a free parameter, compute the joint maximum likelihood estimates $(\hat{M_0}, \hat{\theta}, \hat{\sigma}_{\text{int}})$, and estimate the standard errors from the inverse of the observed Fisher matrix at the MLE.

3 Likelihood with Observational Selection Effects

Suppose the masses of N stars $\{y_i\}$ observed in a star-forming region follow a Pareto distribution:

$$P(y|\gamma) = \begin{cases} 0, & y < t_0 \\ A\, y^{-\gamma}, & y \geq t_0 \end{cases}, \tag{9}$$

where $t_0 = 1$ is a lower limit, A is a normalisation constant, and γ is the exponent of astrophysical interest.

1. Solve for A. What conditions on γ must hold for y to have finite expectation and variance?

2. Derive the maximum likelihood estimator for γ. Using the dataset of $\{y_i\}$ provided, compute the maximum likelihood estimate. Bootstrap the dataset 100 times to compute its standard deviation.

3. After you've done all that work, the astronomer who obtained the data now tells you that she would have been unable to see any stars with masses $y > t_1$ due to observational selection effects. This can be modeled by introducing an *inclusion* vector $\boldsymbol{I}$ with binary components

$$I_i = \begin{cases} 0, & y_i \text{ is not observed,} \\ 1, & y_i \text{ is observed.} \end{cases} \tag{10}$$

The inclusion vector has a probability representing the data collection process:

$$P(I_i = 1|\, y_i, t_1) = \begin{cases} 1, & y_i \le t_1 \\ 0, & y_i > t_1 \end{cases} \tag{11}$$

where $t_1 = 5$ is the known observational limit. Let $y_i^{\rm obs}$ indicate the the mass y_i was observed, $I_i = 1$. Derive the likelihood function $P(y_i^{\rm obs}|I_i = 1, \gamma)$ for one star, and then the likelihood for the full sample.

4. Compute the corrected MLE on the dataset provided for $\boldsymbol{y}^{\rm obs}$. Bootstrap the dataset 100 times to compute its standard deviation.

5. Using the corrected MLE value $\hat{\gamma}$ you found, generate 100 new samples of the same size by drawing random numbers from the observationally-truncated distribution $P(y_i^{\rm obs}|I_i = 1, \hat{\gamma})$. Compute the corrected MLE on each simulated dataset and compute its standard deviation over simulations.

4 Bayesian Inference for Gaussian data with unknown mean and variance

Suppose instead that log stellar masses $\{y_i\}$ are from Gaussian distribution with unknown population mean μ and variance σ^2:

$$y_i \sim N(\mu, \sigma^2) \tag{12}$$

for $i = 1, \ldots, N$.

1. Derive the likelihood function $P(\boldsymbol{y}|\, \mu, \sigma^2)$, expressed in terms of the sufficient statistics: the sample mean,

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{13}$$

and sample variance:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2. \tag{14}$$

2. Adopt a "non-informative" improper prior density $P(\mu, \sigma^2) \propto \sigma^{-2}$ for $\sigma^2 > 0$. Derive the posterior density $P(\mu, \sigma^2|\, \boldsymbol{y})$. Show that:

$$P(\mu|\, \sigma^2, \boldsymbol{y}) = N(\mu|\, \bar{y}, \sigma^2/n) \tag{15}$$

and

$$P(\sigma^2|\, \boldsymbol{y}) = \text{Inv-}\chi^2(\sigma^2|n-1, s^2) \tag{16}$$

where the scaled inverse χ^2 distribution has an unnormalised density:

$$\text{Inv-}\chi^2(\theta\,|\, n-1, s^2) \propto \theta^{(-\nu/2+1)} \exp(-\nu s^2/(2\theta)). \tag{17}$$

3. Show that the marginal $P(\mu|\, \boldsymbol{y})$ is a t-distribution and derive its parameters. A t-random variable has unnormalised density:

$$t_\nu(\theta|\, \mu, \sigma^2) \propto \left[1 + \frac{1}{\nu}\Big(\frac{\theta - \mu}{\sigma}\Big)^2\right]^{-(\nu+1)/2}. \tag{18}$$

4. Now adopt the informative conjugate prior from class:

$$P(\mu, \sigma^2) = P(\mu|\, \sigma^2)P(\sigma^2) = N(\mu|\, \mu_0, \sigma^2/\kappa_0) \times \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \tag{19}$$

where μ_0 is a prior mean, σ_0^2 is the prior scale of the variance, and κ_0 and ν_0 quantify the strength of the prior information. Derive the posterior $P(\mu, \sigma^2|\, \boldsymbol{y})$. Show that is of the same form as the prior, and derive its parameters $\mu_n, \kappa_n, \nu_n, \sigma_n^2$.

5. Derive the marginal density $P(\mu|\boldsymbol{y})$. What is the limiting form of this density as $n \to \infty$? Use the fact that the t-distribution tends to a Gaussian as its degrees of freedom tends toward infinity. Show that the limiting posterior is independent of the prior.