Example Sheet 1 **Solutions**
Example Class: 16 Feb 2017, 2:30pm, MR5

Part III Astrostatistics

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 $\sigma^2_{\rm int}$.

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

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 $s = 1 \ldots N$ 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 $P(\hat{M}_s | \, M_s)$. **Solution: The sum of independent Gaussian random variables (r.v.s) is also a Gaussian r.v., with a mean equal to the sum of the means, and the variance equal to the sum of the variances. Note that Eqs. 2 & 4 can be written**

$$\hat{\mu}_s = \mu_s + \epsilon_{\mu,s}$$

$$\hat{m}_s = m_s + \epsilon_{m,s}$$

where $\epsilon_{\mu,s} \sim N(0, \sigma^2_{\mu,s})$ and $\epsilon_{m,s} \sim N(0, \sigma^2_{m,s})$ are independent, zero-mean Gaussian random variables. Therefore $\hat{M}_s = \hat{m}_s - \hat{\mu}_s = m_s - \mu_s + \epsilon_{m,s} - \epsilon_{\mu,s}$ is a Gaussian

random variable with mean $m_s - \mu_s = M_s$ and variance equal to the sum of the variances $\sigma_s^2 \equiv \sigma_{\mu,s}^2 + \sigma_{m,s}^2$. More explicitly:

$$\begin{aligned}\mathbb{E}[\hat{M}_s|M_s] &= \mathbb{E}[\hat{m}_s - \hat{\mu}_s|M_s] = \mathbb{E}[m_s - \mu_s|M_s] - \mathbb{E}[\epsilon_{\mu,s}] + \mathbb{E}[\epsilon_{m,s}] \\ &= \mathbb{E}[M_s|M_s] + 0 + 0 = M_s\end{aligned}$$

$$\begin{aligned}\mathbf{Var}[\hat{M}_s|M_s] &= \mathbf{Var}[M_s|M_s] + \mathbf{Var}[\epsilon_{\mu,s}] + \mathbf{Var}[\epsilon_{m,s}] \\ &= 0 + \sigma_{\mu,s}^2 + \sigma_{m,s}^2\end{aligned}$$

Therefore,

$$P(\hat{M}_s|M_s) = N(\hat{M}_s|M_s, \sigma_s^2 \equiv \sigma_{\mu,s}^2 + \sigma_{m,s}^2)$$

is the sampling distribution of $\hat{M}_s$ around M_s.

2. Write down the joint distribution $P(\hat{M}_s, M_s|\bar{M}, \sigma_{\text{int}}^2)$. **Solution: Factoring the joint density into a conditional and a marginal,**

$$\begin{aligned}P(\hat{M}_s, M_s|\bar{M}, \sigma_{\mathbf{int}}^2) &= P(\hat{M}_s|M_s, \bar{M}, \sigma_{\mathbf{int}}^2)P(M_s|\bar{M}, \sigma_{\mathbf{int}}^2) \\ &= P(\hat{M}_s|M_s)P(M_s|\bar{M}, \sigma_{\mathbf{int}}^2) \\ &= N(\hat{M}_s|M_s, \sigma_s^2)N(M_s|\bar{M}, \sigma_{\mathbf{int}}^2).\end{aligned}$$

The second line follows because, given M_s the sampling distribution of $\hat{M}_s$ does not depend on $\bar{M}, \sigma_{\mathbf{int}}^2$. The third line follows from Part 1 and Eq. 1.

3. Derive the observed data likelihood function $L(\bar{M}, \sigma_{\text{int}}^2) = \prod_{s=1}^N P(\hat{M}_s|\bar{M}, \sigma_{\text{int}}^2)$. Show all steps, evaluate all integrals, and maximally simplify. **Solution: We could evaluate the integral:**

$$\begin{aligned}P(\hat{M}_s|\,\bar{M}, \sigma_{\mathbf{int}}^2) &= \int P(\hat{M}_s, M_s|\bar{M}, \sigma_{\mathbf{int}}^2)\, dM_s \\ &= \int (2\pi\sigma_s^2)^{-1/2} \exp\left[-\frac{(\hat{M}_s - M_s)^2}{2\sigma_s^2}\right] (2\pi\sigma_{\mathbf{int}}^2)^{-1/2} \exp\left[-\frac{(M_s - \bar{M})^2}{2\sigma_{\mathbf{int}}^2}\right] dM_s\end{aligned}$$

However, we can reason from the properties of Gaussians r.v.s. Note that we can write:

$$\hat{M}_s|M_s = M_s + \epsilon_s$$

$$M_s|\bar{M}, \sigma_{\mathbf{int}}^2 = \bar{M} + \epsilon_{\mathbf{int},s}$$

where $\epsilon_s \sim N(0, \sigma_s^2)$ and $\epsilon_{\mathbf{int},s} \sim N(0, \sigma_{\mathbf{int},s}^2)$ are independent, zero-mean Gaussian r.v.s. Combining these,

$$\hat{M}_s|\bar{M}, \sigma_{\mathbf{int}}^2 = \bar{M} + \epsilon_{\mathbf{int},s} + \epsilon_s$$

which is the sum of Gaussian r.v.s, so its mean is equal to the sum of the means, and its variance is equal to the sum of the variances. Therefore,

$$P(\hat{M}_s|\,\bar{M}, \sigma_{\mathbf{int}}^2) = N(\hat{M}_s|\,\bar{M}, \sigma_{\mathbf{tot},s}^2 \equiv \sigma_{\mathbf{int}}^2 + \sigma_{\mu,s}^2 + \sigma_{m,s}^2) \tag{5}$$

We could also find this result by explicitly evaluating the integral. Note this is a special case of the following general lemma.

Lemma 1: Suppose $x \sim N(\mu, \sigma_x^2)$ and $y|x \sim N(x, \sigma_y^2)$. Then $y \sim N(\mu, \sigma_x^2 + \sigma_y^2)$:

$$\begin{aligned}
P(y|\,\mu, \sigma_x^2, \sigma_y^2) &= \int N(y|\,x, \sigma_y^2) N(x|\,\mu, \sigma_x^2) \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \int \exp\left(-\frac{1}{2}\left[\frac{(y-x)^2}{\sigma_y^2} + \frac{(x-\mu)^2}{\sigma_x^2}\right]\right) dx \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \int \exp\left(-\frac{1}{2}\left[Ax^2 + Bx + C\right]\right) dx \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \int \exp\left(-\frac{1}{2}\left[A(x-h)^2 + k\right]\right) dx \\
&= \frac{1}{2\pi\sigma_x\sigma_y} e^{-k/2} \int \exp\left(-\frac{1}{2}A(x-h)^2\right) dx \\
&= \frac{1}{2\pi\sigma_x\sigma_y} e^{-k/2} \sqrt{\frac{2\pi}{A}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2 + \sigma_y^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma_x^2 + \sigma_y^2}\right) \\
&= N(y|\,\mu, \sigma_x^2 + \sigma_y^2).
\end{aligned}$$

In line 3, we expand the exponent in powers of x with coefficients, e.g. $A = \sigma_y^{-2} + \sigma_x^{-2}$, and B, C defined accordingly. In line 4, we "complete the square", so that $h = -B/2A$ and $k = C - B^2/4A = (y-\mu)^2/(\sigma_x^2 + \sigma_y^2)$. In line 6, we evaluate a Gaussian integral and then we simplify. Finally, the likelihood function of the parameters $\bar{M}, \sigma_{\rm int}^2$ for all the data is the product over all supernovae:

$$L(\bar{M}, \sigma_{\rm int}^2) = \prod_{s=1}^{N} P(\hat{M}_s|\,\bar{M}, \sigma_{\rm int}^2)$$

where each term is derived above.

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 $s = 1 \ldots N$ 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. **Solution: see code**.

5. Apply your code to the data provided online for this problem and report MLE estimates and uncertainties. **Solution: see code**.

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. **Solution: see code**.

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:

$$C_s \sim N(\mu_C, \sigma_C^2) \tag{6}$$

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

$$P(E_s|\,\tau) = \tau^{-1} \exp(E_s/\tau) \tag{7}$$

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.

$$\hat{O}_s = C_s + E_s + \epsilon_s \tag{8}$$

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 $P(\hat{O}_s, C_s, E_s|\tau, \mu_c, \sigma_C^2)$. **Solution: The joint distribution can be factored into a conditional and marginal:**

$$\begin{aligned} P(\hat{O}_s, C_s, E_s|\tau, \mu_c, \sigma_C^2) &= P(\hat{O}_s|\,C_s, E_s; \tau, \mu_C, \sigma_C^2) \times P(C_s, E_s|\tau, \mu_C, \sigma_C^2) \\ &= P(\hat{O}_s|\,C_s, E_s) \times P(C_s|\mu_C, \sigma_C^2) \times P(E_s|\tau) \\ &= N(\hat{O}_s|\,C_s + E_s, \sigma_s^2) \times N(C_s|\mu_C, \sigma_C^2) \times P(E_s|\tau). \end{aligned}$$

   On the second line, we note that the sampling distribution of $\hat{O}_s$ given C_s, E_s is independent of the population parameters τ, μ_C, σ_C^2, and also that C_s and E_s are independent random variables, so their joint distribution factors accordingly. On the third line, we note that the first factor is given by Eq. 8, the second factor by Eq. 6, and the third factor by Eq. 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. **Solution: We are asked to derive:**

$$\begin{aligned} P(\hat{O}_s|\,\tau, \mu, \sigma_C^2) &= \int dE_s \int dC_s P(\hat{O}_s, C_s, E_s|\tau, \mu_C, \sigma_C^2) \\ &= \int dE_s\, P(E_s|\tau) \int dC_s\, N(\hat{O}_s|C_s + E_s, \sigma_s^2) \times N(C_s|\mu_C, \sigma_C^2) \end{aligned}$$

   It is easiest to do the integral over C_s first. Note that

$$P(\hat{O}_s|\,E_s, \tau, \mu_C, \sigma_C^2) = \int dC_s\, N(\hat{O}_s|C_s + E_s, \sigma_s^2) \times N(C_s|\mu_C, \sigma_C^2).$$

   This can be evaluated using similar reasoning as in Problem 1, Step 3 or application of Lemma 1 (with $x \to C_s$, $y \to \hat{O}_s - E_s$, $\sigma_y \to \sigma_s$, and $\sigma_x \to \sigma_C$). This results in:

$$P(\hat{O}_s|\,E_s, \tau, \mu_C, \sigma_C^2) = N(\hat{O}_s|\,\mu_C + E_s, \sigma_{\mathbf{tot},s}^2 \equiv \sigma_s^2 + \sigma_C^2)$$

Now we can perform the integral over E_s. Let $Y_s = \hat{O}_s - \mu_C$. We essentially play the same game of "completing the square" inside the exponent:

$$\begin{aligned}
P(\hat{O}_s|\,\tau,\mu,\sigma_C^2) &= \int dE_s\, P(E_s|\tau) \times N(\hat{O}_s|\,\mu_C + E_s, \sigma^2_{\mathbf{tot},s}) \\
&= \frac{1}{\tau\sigma_{\mathbf{tot},s}\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{(Y_s - E_s)^2}{2\sigma^2_{\mathbf{tot},s}}\right) \exp\left(-E_s/\tau\right) dE_s \\
&= \frac{1}{\tau\sigma_{\mathbf{tot},s}\sqrt{2\pi}} \int_0^\infty \exp\left[-\frac{1}{2\sigma^2_{\mathbf{tot},s}}\left(E_s^2 - 2(Y_s - \sigma^2_{\mathbf{tot},s}/\tau)E_s + Y_s^2\right)\right] dE_s \\
&= \frac{1}{\tau\sigma_{\mathbf{tot},s}\sqrt{2\pi}} e^{-k/2\sigma^2_{\mathbf{tot},s}} \int_0^\infty \exp\left[-\frac{(E_s - h)^2}{2\sigma^2_{\mathbf{tot},s}}\right] dE_s \\
&= \frac{\sigma_{\mathbf{tot},s}\sqrt{2\pi}}{\tau\sigma_{\mathbf{tot},s}\sqrt{2\pi}} e^{-k/2\sigma^2_{\mathbf{tot},s}} \int_0^\infty N(E_s|\,h, \sigma^2_{\mathbf{tot},s})\, dE_s
\end{aligned}$$

where $h = Y_s - \sigma^2_{\mathbf{tot},s}/\tau$ and $k = Y_s^2 - h^2$. Now we change variables inside the integral to $Z = (E_s - h)/\sigma_{\mathbf{tot},s}$.

$$\begin{aligned}
P(\hat{O}_s|\,\tau,\mu,\sigma_C^2) &= \tau^{-1} e^{-k/2\sigma^2_{\mathbf{tot},s}} \int_{-h/\sigma_{\mathbf{tot},s}}^\infty \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ \\
&= \tau^{-1} e^{-k/2\sigma^2_{\mathbf{tot},s}} \int_{-\infty}^{h/\sigma_{\mathbf{tot},s}} N(Z|\,0,1)\, dZ \\
&= \tau^{-1} \exp\left[\frac{1}{2}\frac{\sigma^2_{\mathbf{tot},s}}{\tau^2} - \frac{(\hat{O}_s - \mu_C)}{\tau}\right] \times \Phi\left(\frac{(\hat{O}_s - \mu_C)}{\sigma_{\mathbf{tot},s}} - \frac{\sigma_{\mathbf{tot},s}}{\tau}\right).
\end{aligned}$$

In the second line, we recognise the integrand as the probability density function of a standard (mean 0, variance 1) Gaussian, and note that due to the symmetry of the Gaussian, the integral from a constant to infinity is the same as the integral from negative infinity to the negative of the constant. This allows us to use the cumulative distribution function (CDF) of the standard Gaussian, $\Phi(x) = \int_\infty^x N(z|\,0,1)\, dz$ in the last line, and then we simplify. Finally, the likelihood function of the parameters τ, μ_c, σ_C^2 for all the data is the product over all supernovae:

$$L(\tau, \mu_c, \sigma_C^2) = \prod_{s=1}^{N} P(\hat{O}_s|\,\tau,\mu,\sigma_C^2)$$

where each term is derived above.

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. **See code.**

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. **See code.**

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. **See code.**

6. Now fixing the parameters τ, μ, σ_C^2 to your MLE estimated values $\hat{\tau}, \hat{\mu}, \hat{\sigma}_C^2$, plot the sampling density $P(\hat{O}_s | \hat{\tau}, \hat{\mu}, \hat{\sigma}_C^2)$ as a function of $\hat{O}_s$. Compare against a histogram of the numerical $\hat{O}_s$ values. **See code.**

7. Derive an expression for the posterior density of the reddening E_s of each SN s, $P(E_s | \hat{O}_s; \hat{\tau}, \hat{\mu}, \hat{\sigma}_C^2)$. **Solution: As we didn't cover this in the example class, we will postpone this to a future example sheet**.

8. 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). **Solution: As we didn't cover this in the example class, we will postpone this to a future example sheet**.