Part III Astrostatistics: Example Sheet 4
Example Class: Tuesday, 30 April 2019, 12pm in MR13

So far includes solutions to problem 1, 3, 5.

1 Warm-Up

Prove that the product of m multivariate Gaussian densities in random d-dimensional vector $\boldsymbol{x}$:

$$I(\boldsymbol{x}) = \prod_{i=1}^{m} N(\boldsymbol{x}|\,\boldsymbol{\mu}_i, \boldsymbol{C}_i) \tag{1}$$

is proportional to a single Gaussian density in $\boldsymbol{x}$. Here the $\{\boldsymbol{\mu}_i\}$ and $\{\boldsymbol{C}_i\}$ are m pairs of constant mean d-vectors and $d \times d$ covariance matrices. Find the mean and covariance matrix of the single resulting Gaussian.

Solution: First, we derive a vectorial "complete the square" lemma:

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{b} + c &= (\boldsymbol{x}-\boldsymbol{d})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{d}) + e = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{d}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{d} + \boldsymbol{d}^T\boldsymbol{A}\boldsymbol{d} + e \\ &= \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{d}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{d}^T\boldsymbol{A}\boldsymbol{d} + e\end{aligned}$$

Therefore, $\boldsymbol{d} = -\frac{1}{2}\boldsymbol{A}^{-1}\boldsymbol{b}$ and $e = c - \boldsymbol{d}^T\boldsymbol{A}\,\boldsymbol{d} = c - \frac{1}{4}\boldsymbol{b}^T\,\boldsymbol{A}^{-1}\boldsymbol{b}$.

Now, by definition,

$$I(\boldsymbol{x}) = \prod_{i=1}^{m} |2\pi\boldsymbol{C}_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{C}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

$$\begin{aligned}-2\log I(\boldsymbol{x}) &= \textbf{const} + \sum_{i=1}^{m} \boldsymbol{x}^T\boldsymbol{C}_i^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{C}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^T\boldsymbol{C}_i^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{C}_i^{-1}\,\boldsymbol{\mu}_i \\ &= \textbf{const} + \boldsymbol{x}^T\left[\sum_{i=1}^{m}\boldsymbol{C}_i^{-1}\right]\boldsymbol{x} - 2\boldsymbol{x}^T\sum_{i=1}^{m}\boldsymbol{C}_i^T\boldsymbol{\mu}_i\end{aligned}$$

Let $\boldsymbol{A} = \sum_{i=1}^{m}\boldsymbol{C}_i^{-1}$ and $\boldsymbol{b} = -2\sum_{i=1}^{m}\boldsymbol{C}_i^{-1}\,\boldsymbol{\mu}_i$. We have

$$I(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^T\boldsymbol{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_x)\right] \propto N(\boldsymbol{x}|\,\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$

where the resulting precision matrix is the sum of the individual precision matrices,

$$\boldsymbol{\Sigma}_x^{-1} \equiv \sum_{i=1}^{m}\boldsymbol{C}_i^{-1}$$

and the resulting mean is the precision-weighted mean of the individual means,

$$\boldsymbol{\mu}_x \equiv \boldsymbol{\Sigma}_x \sum_{i=1}^{m} \boldsymbol{C}_i^{-1} \boldsymbol{\mu}_i = \left[\sum_{i=1}^{m} \boldsymbol{C}_i^{-1} \right]^{-1} \sum_{i=1}^{m} \boldsymbol{C}_i^{-1} \boldsymbol{\mu}_i$$

2 A Hierarchical Bayesian Model for Supernovae and Dust

Consider the following hierarchical Bayesian generative model for supernova colours. The latent intrinsic colour of a supernova s is C_s and is drawn from a Gaussian distribution with mean colour μ_C and variance $\sigma^2_{\rm int}$: $C_s \sim N(\mu_C, \sigma^2_{\rm int})$. The latent reddening due to interstellar dust in the supernova's galaxy is E_s, and is drawn from an exponential distribution with mean τ: $E_s \sim \text{Exponen}(\tau)$, i.e.,

$$P(E_s|\,\tau) = \tau^{-1} \exp(-E_s/\tau) \times H(E_s), \tag{2}$$

where $H(x)$ is the Heaviside step function:

$$H(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{3}$$

Because the presence of interstellar dust can only redden the colours, the probability density is only positive for $E_s > 0$. The measured, observed colour $\hat{O}_s$ results from the sum of the intrinsic colour, reddening, and measurement error with known variance $\sigma^2_{O,s}$: $\hat{O}_s|E_s, C_s \sim N(C_s+E_s, \sigma^2_{O,s})$. There are $s = 1, \ldots, N$ independent supernovae in our sample. For hyperpriors, you may use the improper, noninformative $P(\mu_C) \propto 1$, $P(\tau) \propto H(\tau)$, and $P(\sigma^2_{\rm int}) \propto H(\sigma^2_{\rm int})$.

1. Write down the joint probability distribution of the observed data $\{\hat{O}_s\}$, latent variables $\{C_s, E_s\}$, and hyperparameters $\mu_C, \sigma^2_{\rm int}, \tau$ for the sample of N supernovae.

   Solution: Let us change notation slightly to save typing: $\mu = \mu_C$, $\sigma^2 \equiv \sigma^2_{\rm int}$. The joint probability density is:

   $$\begin{aligned} &P(\{\hat{O}_s, E_s, C_s\}, \mu, \sigma^2, \tau) \\ &\quad = \left[\prod_{s=1}^{N} N(\hat{O}_s|\, C_s + E_s, \sigma^2_{O,s}) N(C_s|\,\mu, \sigma^2) \mathbf{Expon}(E_s|\tau) \right] \times P(\mu) P(\sigma^2) P(\tau) \end{aligned}$$

   For positive τ, σ^2, this is

   $$\begin{aligned} &P(\{\hat{O}_s, E_s, C_s\}, \mu, \sigma^2, \tau) \\ &\quad = \left[\prod_{s=1}^{N} N(\hat{O}_s|\, C_s + E_s, \sigma^2_{O,s}) N(C_s|\,\mu, \sigma^2) \mathbf{Expon}(E_s|\tau) \right] \end{aligned}$$

   and zero otherwise. QED.

2. Draw a probabilistic graphical model or directed acyclic graph representing this joint distribution.

3. Construct a Gibbs sampler that generates an MCMC to sample the joint posterior probability density of the unknown latent variables and hyperparameters given the observed colours, $P(\{C_s, E_s\}, \mu_C, \sigma^2_{\rm int}, \tau | \{\hat{O}_s\})$, by deriving the $2N+3$ conditional posterior densities that one can directly sample from. You may assume that you have access to algorithms that allow you to directly sample random variates from the following probability densities:

   - Gaussian $N(x|\,\mu, \sigma^2)$.
   - truncated Gaussian $\propto H(x) \times N(x|\,\mu, \sigma^2)$.
   - Inverse gamma: Inv-Gamma$(x|\,a, b) \propto x^{-(a+1)} \exp(-b/x)$, $x > 0$.

   Briefly describe how you would implement the sampler, and analyse and assess the convergence of the MCMC.

   Solution: We need to derive the following conditionals from the posterior: $P(\{E_s, C_s\}, \mu, \sigma^2, \tau | \mathcal{D})$, where the data are the measurements $\mathcal{D} = \{\hat{O}_s\}$.

   (a) $P(C_s| \ldots; \mathcal{D})$

   (b) $P(E_s| \ldots; \mathcal{D})$

   (c) $P(\mu, \sigma^2| \ldots, \mathcal{D})$

   (d) $P(\tau| \ldots, \mathcal{D})$**.**

4. Apply your sampler to analyse the data from the Table 3 dataset from Jha, Riess & Kirshner. (2007), "Improved Distances to Type Ia Supernovae with Multicolor Light-Curve Shapes: MLCS2k2." Compute posterior summaries of the hyperparameters $\mu_C, \sigma^2_{\rm int}, \tau$ and the latent variables $\{C_s, E_s\}$. The data is provided online.

3 Harmonic Mean Estimator for Bayesian Evidence

Consider a general Bayesian inference problem with data y, parameter θ, likelihood function $P(y|\,\theta)$ and a proper prior $P(\theta)$. We wish to compute the evidence or marginal likelihood $Z \equiv P(y) = \int P(y|\,\theta)\, P(\theta)\, d\theta$.

1. Suppose you have $i = 1, \ldots, m$ independent, random samples from the posterior distribution, $\theta_i \sim P(\theta|\,y)$. Consider the estimator

$$\hat{I} \equiv \frac{1}{m} \sum_{i=1}^{m} P(y|\,\theta_i)^{-1}. \tag{4}$$

   Show that $\mathbb{E}_{\theta|y}[\hat{I}] = Z^{-1}$, where the expectation is taken with respect to the posterior density $P(\theta|\,y)$. Thus, by LLN, as $m \to \infty$, this converges to the inverse of the evidence.
   Solution: From Bayes' Theorem, the posterior density is:

$$P(\theta|\,y) = \frac{P(y|\,\theta) P(\theta)}{Z}$$

Thus, the expectation with respect to the posterior density is:

$$\begin{aligned}\mathbb{E}_{\theta|y}[\hat{I}] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\theta|y}[P(y|\theta)^{-1}] \\ &= \frac{1}{m}\sum_{i=1}^{m}\int \frac{P(y|\theta)P(\theta)}{P(y|\theta)Z}\,d\theta \\ &= Z^{-1}\int P(\theta)\,d\theta = Z^{-1}\end{aligned}$$

since the prior is proper and normalised. QED.

2. Suppose the sampling distribution of the data is $y \sim N(\theta, \sigma^2)$ and the proper prior is $\theta \sim N(0, \tau^2)$. The measurement variance σ^2 and the prior variance τ^2 are known, and $\tau \gg \sigma$.

   (a) What is the posterior density $P(\theta|\, y)$? **Solution:**

$$\begin{aligned}P(\theta|y) &\propto P(y|\theta)P(\theta) \\ &\propto N(y|\,\theta, \sigma^2) \times N(\theta|\,0, \tau^2) \\ &= N(\theta|\,\tilde{\theta}, \sigma_\theta^2),\end{aligned}$$

   where $\sigma_\theta^{-2} = \sigma^{-2} + \tau^{-2}$ **and** $\tilde{\theta} = \sigma_\theta^2(\sigma^{-2}\, y)$**.**

   (b) What is the exact evidence Z? **Solution:**

$$Z = P(y) = \int P(y|\theta)P(\theta)\,d\theta = \int N(y|\theta, \sigma^2)\, N(\theta|0, \tau^2)\,d\theta = N(y|\,0, \sigma^2 + \tau^2)$$

   (c) What is the expectation of the estimator $\mathbb{E}_{\theta|y}[\hat{I}]$? **Solution:**

$$\mathbb{E}_{\theta|y}[\hat{I}] = N(y|\,0, \sigma^2 + \tau^2)^{-1}$$

   (d) What is the variance of the estimator $\mathrm{Var}_{\theta|y}[\hat{I}]$? **Solution:**

$$\mathbf{Var}_{\theta|y}[\hat{I}] = \mathbf{Var}[\frac{1}{m}\sum_{i=1}^{m} P(y|\theta_i)^{-1} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{i=1}^{m}\mathbf{Var}[P(y|\theta)^{-1}] = \frac{1}{m}\mathbf{Var}[P(y|\theta)^{-1}].$$

   We need to compute

$$\begin{aligned}V \equiv \mathbf{Var}[P(y|\theta)^{-1}] &= \mathbb{E}[P(y|\,\theta)^{-2}] - \mathbb{E}[P(y|\,\theta)^{-1}]^2 \\ &= \int \frac{P(y|\theta)P(\theta)}{P(y|\theta)^2 Z}\,d\theta - Z^{-2} \\ &= Z^{-1}\int \frac{P(\theta)}{P(y|\theta)}\,d\theta - Z^{-2} \\ &= Z^{-1}\int \frac{N(\theta|0,\tau^2)}{N(y|\theta,\sigma^2)}d\theta - Z^{-2} \\ &= Z^{-1}\frac{\sigma}{\tau}R - Z^{-2}.\end{aligned}$$

The integral of the ratio of the prior to the likelihood is:

$$R = \int \exp\left(-\frac{1}{2}\theta^2/\tau^2\right) \exp\left(\frac{1}{2}(y-\theta)^2/\sigma^2\right).$$

By completing the square in θ inside the exponent, we have

$$R = \exp\left(\frac{1}{2}y^2\sigma^{-2}(1+\sigma^{-2}/a)\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}a(\theta-h)^2\right) d\theta$$

where $a \equiv (\tau^{-2} - \sigma^{-2})$ and $h = y\sigma^{-2}/a$. Note that the integral diverges if $a < 0$, i.e. if $\tau^2 > \sigma^2$. Otherwise,

$$R = \exp\left(\frac{1}{2}y^2\sigma^{-2}\left[1 + \frac{\sigma^{-2}}{\tau^{-2}-\sigma^{-2}}\right]\right) \frac{\sigma\sqrt{2\pi}}{\sqrt{1-\tau^2/\sigma^2}},$$

Therefore as $\tau \to \sigma$ from below, this integral diverges. For $\tau^2 > \sigma^2$, the variance of the estimator is infinite.

4 Bayesian Model Comparison

1. Data points $\{x_i\}$ come independently from a probability distribution $P(x)$. According to model H_0, $P(x)$ is a uniform distribution $P(x|H_0) = \frac{1}{2}$ for $x \in (-1, 1)$. According to model H_1, $P(x)$ is a nonuniform distribution with an unknown parameter $m \in (-1, 1)$:

$$P(x|\, m, H_1) = \frac{1}{2}(1 + m\, x), \tag{5}$$

for $x \in (-1, 1)$. Given the data $\mathcal{D} = \{0.3, 0.5, 0.7, 0.8, 0.9\}$, what is the evidence for H_0 and H_1?

2. Datapoints $\{(x_i, t_i)\}$ are believed to come from a straight line. The experimenter chooses x_i, and t_i is Gaussian-distributed about $y_i = w_0 + w_1\, x_i$ with variance σ^2. According to model H_1, the straight line is horizontal, so $w_1 = 0$. According to model H_2, w_1 is a parameter with prior distribution Normal(0, 1). Both models assign a prior distribution Normal(0, 1) to w_0. Given the data set $D = \{(-8, 8), (-2, 10), (6, 11)\}$, and assuming the noise level is $\sigma = 1$, what is the evidence for each model?

5 Gaussian Processes as Infinite Basis Expansions

Functions drawn from a Gaussian process prior often have an equivalent description as arising from a linear combination of an infinite set of basis functions. Consider a finite set of $J > 2$ basis functions with a Gaussian shape centred at values c_i,

$$\phi_i(x) = \exp\left[-\frac{(x-c_i)^2}{l^2}\right] \tag{6}$$

defined on the real line $x \in \mathbb{R}$. The centres span a distance $c_J - c_1 = h$, and the centres are spaced so that $\Delta c = c_{i+1} - c_i = h/(J-1)$. Suppose a function is formed as a linear combination of these functions:

$$f(x) = \sum_{i=1}^{J} w_i\, \phi_i(x). \tag{7}$$

Suppose we put a Gaussian prior on the coefficients, $w_i \sim N(0, \sigma^2 h/J)$.

1. What is the mean $\mathbb{E}[f(x)]$ and the covariance function $k(x, x') = \text{Cov}[f(x), f(x')]$?

   Solution: The expectation is

$$\mathbb{E}[f(x)] = \sum_{i=1}^{J} \phi_i(x)\mathbb{E}(w_i) = 0 \tag{8}$$

   The kernel is

$$\begin{aligned} k(x, x') = \mathbf{Cov}[f(x), f(x')] &= \mathbf{Cov}\left[\sum_{i=1}^{J} w_i\phi_i(x), \sum_{j=1}^{J} w_i\phi_i(x')\right] \\ &= \sum_{i=1}^{J}\sum_{j=1}^{J} \phi_i(x)\phi_j(x')\mathbf{Cov}[w_i, w_j] \\ &= \sum_{i=1}^{J}\sum_{j=1}^{J} \phi_i(x)\phi_j(x')\,\delta_{ij}\,\sigma^2 h/J \\ &= \sum_{i=1}^{J} \phi_i(x)\phi_i(x')\sigma^2 h/J \\ &= \sigma^2 \sum_{i=1} \phi_i(x)\phi_i(x')\frac{J-1}{J}\Delta c \end{aligned} \tag{9}$$

   where δ_{ij} is a Kronecker delta function.

2. Derive the kernel function $k(x, x')$ in the limit of an infinite number of basis functions spanning the real line: $J \to \infty$ and $c_1 \to -\infty$, $h \to \infty$.

   Solution: In the limit of $J \to \infty$, this Riemann sum becomes the integral

$$\begin{aligned} k(x, x') &= \sigma^2 \int_{c_1}^{c_J = c_1 + h} \phi_i(x)\phi_i(x')\, dc \\ &= \sigma^2 \int_{c_1}^{c_1+h} e^{-(x-c)^2/l^2} e^{-(x'-c)^2/l^2}\, dc \end{aligned} \tag{10}$$

   Now letting the basis span the real line, $c_1, h \to \infty$, we have

$$k(x, x') = \sigma^2 \int_{-\infty}^{+\infty} e^{-(x-c)^2/l^2} e^{-(x'-c)^2/l^2}\, dc \tag{11}$$

   Noting that

$$e^{-(x-c)^2/l^2} = \frac{l}{\sqrt{2}}\sqrt{2\pi}N(x|c, l^2/2) = \frac{l}{\sqrt{2}}\sqrt{2\pi}N(c|x, l^2/2), \tag{12}$$

   we find

$$\begin{aligned} k(x, x') &= \sigma^2 \left(\frac{l}{\sqrt{2}}\sqrt{2\pi}\right)^2 \int_{-\infty}^{+\infty} N(x|\, c, l^2/2)N(c|\, x', l^2/2)\, dc \\ &= \sigma^2 \left(\frac{l}{\sqrt{2}}\sqrt{2\pi}\right)^2 N(x|\, x', l^2) \end{aligned} \tag{13}$$

We recognised the integral from previous Gaussian marginalisation examples. Finally,

$$k(x, x') = \frac{l\sigma^2}{2}\sqrt{2\pi}e^{-(x-x')^2/2l^2} \tag{14}$$

Therefore, the squared exponential kernel generates functions that are linear combinations of Gaussian functions of width l, distributed densely along the real line.

3. What is the variance of the resulting Gaussian process at any x?

   Solution: $\mathbf{Var}(f(x)) = k(x, x) = \sqrt{\frac{\pi}{2}}l\sigma^2$.