Examples paper 4

P1 (a) $J = b^T x = b_1 x_1 + b_2 x_2 + \dots + b_N x_N$

$$\frac{\partial J}{\partial x} = \begin{bmatrix} \partial J / \partial x_1 \\ \vdots \\ \partial J / \partial x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ | \\ b_N \end{bmatrix} = b$$

(b) $B = \begin{bmatrix} b_{11} & b_{12} & | \\ b_{21} & \ddots & \\ & & b_{NN} \end{bmatrix}$

$$\hookrightarrow x^T B x = x^T \begin{bmatrix} \sum_{k=1}^{N} b_{1k} x_k \\ \vdots \\ \sum_{k=1}^{N} b_{Nk} x_k \end{bmatrix} = x_1 \sum_k b_{1k} x_k + \dots x_N \sum b_{Nk} x_k$$

$$\frac{\partial (x^T B x)}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x_1}\left(x_1 \sum_k b_{1k} x_k + \dots + x_N \sum_k b_{Nk} x_N \right) \\ \vdots \\ \frac{\partial}{\partial x_N}\left(x_1 \sum_k b_{1k} x_k + \dots + x_N \sum_k b_{Nk} x_N \right) \end{bmatrix}$$

$$\overbrace{[b_{11}, b_{12}, \dots, b_{1N}]}^{1\text{st row of } B} \cdot x \qquad \overbrace{[b_{11}, b_{21}, \dots]}^{1\text{st column}} x$$

$$= \sum_k b_{1k} x_k + \sum_k b_{k1} x_k = \begin{bmatrix} 2 b_{11} x_1 + \sum_{k \neq 1} b_{1k} x_k + \sum_{k \neq 1} x_k b_{k1} \\ \vdots \\ 2 b_{NN} x_N + \sum_{k \neq N} b_{Nk} x_k + \sum_{k \neq 1} x_k b_{kN} \end{bmatrix}$$

and same for all components $k = 2, \dots, N$

$$= \boxed{Bx + B^T x}$$ (B is any square matrix)

(c) B is symmetric: same as (b) but since $B = B^T$ $\quad \frac{\partial J}{\partial x} = 2Bx$

P2 (a) Show that any matrix $B = G^T G$ is positive semi-definite:

$x^T G^T G x \geq 0$ for any vector $x \in \mathbb{R}^N$

Let v be $v = Gx$. Then, $x^T G^T G x$ can be written as $v^T v$, which is simply the sum of squared terms:

$$\boxed{x^T G^T G x = v^T v = v_1^2 + v_2^2 + \dots + v_N^2 \geq 0}$$

(b) Using the term v introduced in part (a), the condition to make $x^T G^T G x$ equal to 0 is $v_1 = v_2 = v_3 = \dots = v_N = 0$. To make $v = 0$ we need $Gx = 0$. Aside from the trivial case where $x = 0$, the condition $Gx = 0$ can be satisfied by some $x \neq 0$ whenever the system of equations it defines is undetermined, which is equivalent to saying that G has lower rank than the dimension of x.

We'll give some examples covering the relevant cases (we assume x is $N \times 1$ and G is $P \times N$):

1. If $P < N$, the system of equations is undetermined (notice it has more unknowns than equations), therefore it defines a complete region where $Gx = 0$ (some variables are "free" and $Gx = 0$ holds no matter what value these variables take).

Example: $G = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 2 & -1 \end{bmatrix}$. from Gx we have

$$\left.\begin{array}{l} 3x_1 + 4x_2 + 5x_3 = 0 \\ x_1 + 2x_2 - x_3 = 0 \end{array}\right\} \Rightarrow \cdot x_3 = x_1 + 2x_2$$

replacing in eq. 1:

$$8x_1 + 14x_2 = 0$$

$$x_1 = -14/8\, x_2.$$

Then, picking any value of x_2 and computing $x_1 = -14/8$ and $x_3 = x_1 + 2x_2$ yields $Gx = 0$

2. P = N | If the matrix G has full rank, then B will be positive definite since $x^T B x = x G^T G x$ will be 0 only when $x = 0$. (The system of equations imposed by Gx is fully determined, so there's just one solution for each right hand side). If G is NxN but not full rank, B is not positive definite

3. P > N | In this case, B is positive definite if G has rank N (recall that the rank of a non-square matrix is also defined as the number of independent rows/columns, so it's at most N). If the rank of G is less than N, B is not positive definite.

i.e. Matrix G needs to be full rank to make B positive definite

(c) $J = x^T \overbrace{B}^{\text{symmetric positive definite}} x - 2b^T x.$

Using the properties derived in P1.

$$\frac{\partial J}{\partial x} = 2Bx - 2b \Rightarrow x_{min} = \underbrace{B^{-1}}_{\text{If } B \text{ is positive definite then it's invertible}} b$$

(d) Find M, m, c such that $[x-m]^T M [x-m] + c$
$= x^T B x - 2b^T x.$

Expanding the expression above:

$$x^T M x - \underbrace{x^T M m - m^T M x}_{2m^T M x} + m^T M m$$

$$\Rightarrow M = B, \quad m = B^{-1} b \qquad c = -m^T M m = -b^T \overbrace{\cancel{B^{-1} B}}^{I} B^{-1} b$$
$$= -b^T B b$$

(e) In the generalised linear model with gaussian prior, we have: $p(\theta|y) \propto p(\theta)\,p(y|\theta)$. Taking log() to handle what's inside the exponential more easily:

$$\log(p(\theta|y)) = c - \frac{1}{2}[\theta - m_\theta]^T C_\theta^{-1}[\theta - m_\theta] - [y - G\theta]^T[y - G\theta]$$

$$= c' + \frac{-1}{2}\Big\{\frac{y^T y}{\sigma^2} - \frac{2y^T G\theta}{\sigma^2} + \frac{\theta^T G^T G\theta}{\sigma^2} + \theta^T C_\theta^{-1}\theta - 2m_\theta^T C_\theta^{-1}\theta + \overbrace{m_\theta^T C_\theta^{-1} m_\theta}^{\text{const}}$$

"absorbed" into the constant

$$= c'' + \frac{-1}{2}\Big\{\theta^T\Big(\frac{G^T G}{\sigma^2} + C_\theta^{-1}\Big)\theta - 2\Big[\frac{y^T G}{\sigma^2} + m_\theta^T C_\theta^{-1}\Big]\theta\Big\}$$

Using part (d), the form "$x^T B x - 2b^T x$" can be recognised: $x = \theta$, $B = \frac{G^T G}{\sigma^2} + C_\theta^{-1}$

$b = \frac{G^T y}{\sigma^2} + C_\theta^{-1} m_\theta$. Completing the squares (as with M, m, C in part d) $\log p(\theta|y)$ can be written as:

$$\log p(\theta|y) = c'' + -\frac{1}{2}\left\{[\theta - \mu_\theta]^T \Sigma_\theta^{-1} [\theta - \mu_\theta] - \underbrace{\mu_\theta^T \Sigma_\theta^{-1} \mu_\theta}_{\text{constant}}\right\}$$

with μ_θ and Σ_θ^{-1} playing the role of m and M of part (d):

$$M = B = \frac{G^T G}{\sigma^2} + C_\theta^{-1} \Rightarrow \Sigma_\theta = \left[\frac{G^T G}{\sigma^2} + C_\theta^{-1}\right]^{-1}$$

$$m = B^{-1} b = \left[\frac{G^T G}{\sigma^2} + C_\theta^{-1}\right]^{-1}\left[\frac{G_y^T}{\sigma^2} + C_\theta^{-1} m_\theta\right] = \mu_\theta$$

- Regarding $\mu_\theta^T \Sigma_\theta^{-1} \mu_\theta$ as constant and taking exp:

$$p(\theta|y) \propto \exp\left\{-\frac{1}{2}[\theta - \mu_\theta]^T \Sigma_\theta^{-1} [\theta - \mu_\theta]\right\}$$

which has the standard form of a Gaussian multivariate random variable. Then,

$$p(\theta|y) = N(\theta \mid \mu_\theta, \Sigma_\theta)$$

$$X = [x_1, x_2, \dots, x_N]^T$$

P3] $$\log p(X|\mu) = \sum_{m=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{(x_m-\mu)^2}{2\sigma_m^2} \Big/ \frac{\partial(\cdot)}{\partial\mu}$$

$$\frac{\partial(\cdot)}{\partial\mu} = \sum_{m=1}^{N} \frac{x_m-\mu}{\sigma_m^2} = \left(\sum \frac{x_m}{\sigma_m^2}\right) - \mu\sum 1/\sigma_m^2 = 0$$

- $$\mu_{ML} = \left(\sum_{m=1}^{N} x_m/\sigma_m^2\right) \Big/ \underbrace{\left(\sum 1/\sigma_m^2\right)}_{S}$$

- $$\mu_{OLS} = \frac{1}{N}\sum x_m$$

- Bias and variance amalisys:

$$\mathbb{E}\,\mu_{OLS} = \frac{1}{N}\sum \mathbb{E}\,x_m = \frac{\cancel{N}}{\cancel{N}}\,\mu$$

$$\mathbb{E}\,\mu_{ML} = \left(\sum \mathbb{E}\,x_m/\sigma_m^2\right) \Big/ \left(\sum 1/\sigma_m^2\right) = \mu \cdot \cancel{\frac{\sum 1/\sigma_m^2}{\sum 1/\sigma_m^2}}$$

Both are umbiased estimatons //

Variance:

ML

$$\mathbb{E}\left\{\left(\frac{\sum \frac{x_m}{\sigma_m^2}}{S} - \mu\right)^2\right\} = S/S = \underline{1}$$

$$= \mathbb{E}\left\{\sum \frac{x_m/\sigma_m^2}{S} - \left(\frac{\sum \frac{1}{\sigma_m^2}}{S}\right)\mu\right\}^2$$

$$= \mathbb{E}\left(\frac{1}{S}\sum \frac{x_m - \mu}{\sigma_m^2}\right)^2 = \frac{1}{S^2}\sum_{m,m} \underbrace{\mathbb{E}\left\{\frac{(x_m-\mu)}{\sigma_m^2}\frac{(x_{m}-\mu)}{\sigma_{m}^2}\right\}}_{\text{due to indepemdemce}}$$

$$= \begin{cases} 0 & m \neq m \\ \sigma_m^2 & m = m \end{cases}$$

$$= \frac{1}{S^2}\sum\left(1/\sigma_m^2\right)^2 \sigma_m^2 = \frac{1}{S^2}\cdot S = \left(\sum 1/\sigma_m^2\right)^{-1}$$

- If $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2 = \sigma^2$ the $\hat{\sigma}_{ML}$ reduces to $\hat{\sigma}_{ML} = \frac{\sigma^2}{N} = \hat{\sigma}_{OLS}$.

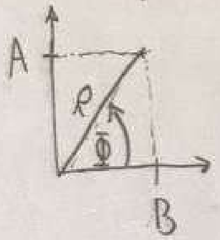
- M_{ML} is more robust to outliers since each observation is weighted appropriately, whereas M_{OLS}, being a simple average, overestimates the importance of noisier observations (i.e. those with higher variances).

P4) $X_n = A \sin(n\Omega) + B \cos(n\Omega)$

Show that an equivalent model can be expressed as $X_n = R\cos(n\Omega + \Phi)$

$\Phi \sim U[0, 2\pi)$, $R \sim$ another distribution.

- First, note that each pair of real numbers can be expressed in polar coordinates:



$R = \sqrt{A^2 + B^2}$ $\quad$ $A = R \cdot \sin\Phi$

$\Phi = \tan^{-1}(A/B)$ $\quad$ $B = R \cos\Phi$

then: $X_n = R \cdot \{ \sin\Phi \sin n\Omega + \cos\Phi \cos n\Omega \}$

Using identity $\cos(\alpha - \beta) = \cos\alpha \cos\beta + \sin\alpha \sin\beta$

$$\boxed{X_n = R\cos(n\Omega - \Phi) = R\cos(n\Omega + \theta)}$$

$\theta = -\Phi$ introduced just to make this formula match the form proposed in the handout.

- However, we don't know how R and Φ distribute. Using change of variables:

$$P_{R,\theta}(R, \theta) = P_{A,B}(\underbrace{R\sin(-\theta)}_{A}, \underbrace{R\cos(-\theta)}_{B}) \cdot \det(J)$$

$$= \frac{1}{2\pi\sigma^2} \det(J) \cdot \exp\left\{ \frac{-1}{2\sigma^2}\left(R^2 \sin^2(-\theta) + R^2\cos^2(-\theta)\right)\right\}$$

$$p_{R,\theta}(R,\theta) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2} \frac{R^2}{\sigma^2} \right\} \cdot \underbrace{R}_{\det(J)}$$

$$= \begin{vmatrix} \frac{\partial A}{\partial R} & \frac{\partial A}{\partial \theta} \\ \frac{\partial B}{\partial R} & \frac{\partial \beta}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & R\sin\theta \\ \sin\theta & -R\cos\theta \end{vmatrix}$$

$$= \left| -R\cos^2\theta - R\sin^2\theta \right|$$

$$= R$$

Then θ and R are independent:

$$p_{R,\theta}(R,\theta) = \underbrace{\left(\frac{1}{2\pi}\right)}_{p(\theta)} \underbrace{\left(\frac{R}{\sigma^2} \exp\left\{ -\frac{1}{2} \frac{R^2}{\sigma^2} \right\}\right)}_{p(R)}$$

Rayleigh distribution

P5 $t_1, \dots, t_N$ lenghts of the intervals

$$p(t_n) = \theta^{-1} \exp(-t_n/\theta) \qquad t_n > 0$$

$$\text{Likelihood} = p(t_1) \cdot p(t_2) \cdots p(t_N)$$

$$= \left(\theta^{-1} \exp -\frac{t_1}{\theta}\right)\left(\theta^{-1} \exp -\frac{t_2}{\theta}\right) \cdots \left(\theta^{-1} \exp -\frac{t_N}{\theta}\right)$$

$$= \theta^{-N} \exp\left\{-\theta^{-1} \sum_{n=1}^{N} t_n\right\}$$

$\hat{\theta}_{ML}$

$$\log(\cdot) = -N \log\theta - \theta^{-1} \sum t_n$$

$$\frac{\partial(\cdot)}{\partial\theta} = \frac{-N}{\theta} + \theta^{-2} \sum t_n = 0 \quad / \; \theta^2$$

$$-N\theta + \sum t_n = 0 \Rightarrow \hat{\theta}_{ML} = \frac{\sum t_n}{N}$$

$$\mathbb{E}\,\hat{\theta}_{ML} = \frac{1}{N} \sum \mathbb{E}\, t_n = \frac{1}{N} N\theta = \theta$$

$$\mathrm{Var}\,\hat{\theta}_{ML} = \mathbb{E}\Big(\frac{1}{N}\sum t_n - \underbrace{\theta}_{\frac{N}{N}\theta}\Big)^2 = \frac{1}{N^2}\, \mathbb{E} \sum_{n,m} (t_n - \theta)(t_m - \theta).$$

$$= \frac{1}{N^2} \cdot N \,\mathrm{Var}\, t_n = \frac{\theta^2}{N} \xrightarrow[N\to\infty]{} 0$$

We used:

$$\mathbb{E}\, t = \theta, \quad \underbrace{\mathbb{E}(t-\theta)^2}_{\mathrm{Var}(t)} = \theta^2, \quad \mathbb{E}(t_n - \theta)(t_m - \theta) = \begin{cases} 0 & n \neq m \\ \theta^2 & n = m \end{cases}$$

MAP:

$$p(\theta | \vec{t}) \propto \left(\prod_m p(t_m | \theta)\right) \cdot p(\theta) \quad / \log(\cdot)$$

$$\log p(\theta | t_1, \ldots, t_N) = C + -N \log\theta - \theta^{-1} \sum t_m - 2\log\theta - \theta^{-1}$$

$$\log(p(\theta | t_1 - t_N) = 0 \Rightarrow \frac{-(N+2)}{\theta} + \theta^{-2}\left(1 + \sum t_m\right)$$

$$\Rightarrow \boxed{\hat{\theta}_{MAP} = \frac{1 + \sum t_m}{N+2}}$$

As $N \to \infty$, the relative weight of the prior becomes negligible, $\hat{\theta}_{MAP} \approx \hat{\theta}_{ML}$

P6 $X_0, X_1, \ldots, X_{N-1}$ AR(1)

ML $$p(X_0, \ldots, X_{N-1}) = \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{X_0^2}{2}\right\}}_{p(X_0)} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_1 - aX_0)^2}{2}\right\}}_{p(X_1|X_0)} \cdots$$

easily obtained with change of variables (from $p(e_0 - e_{N-1}) = N(e_0|0,1) \cdot N(e_1|0,1) \ldots$)

$$\frac{1}{\sqrt{2\pi}} \exp\left\{-(X_{N-1} - aX_{N-2})^2\right\}$$

$$\log p(X_0, \ldots, X_{N-1}) = C - \frac{X_0^2}{2} - \sum_{m=1}^{N-1} \frac{(X_m - aX_{m-1})^2}{2}$$

$$\frac{\partial(\cdot)}{\partial a} = \sum_{m=1}^{N-1} X_{m-1}(X_m - aX_{m-1}) = 0 \Rightarrow \sum_{m=1}^{N-1} X_{m-1} X_m = a \sum_{m=1}^{N-1} X_{m-1}^2$$

$$\Rightarrow \boxed{\hat{a}_{ML} = \frac{\sum_{m=1}^{N-1} X_{m-1} X_m}{\sum_{m=0}^{N-2} X_m^2}}$$

MAP $$p(a | X_0 - X_{N-1}) \propto p(a) \cdot p(X_0 - X_{N-1})$$

$$\log(\cdot) = -\log\sqrt{2\pi\, 0.1^2} - \frac{1}{2}\frac{(a-0.9)^2}{0.1^2} - \frac{X_0^2}{2} - \sum_{m=1}^{N-1} \frac{(X_m - aX_{m-1})^2}{2} + C$$

(expanding) $$= C - \frac{1}{2}\left\{\frac{a^2}{0.1^2} - \frac{2a \cdot 0.9}{0.1^2} + a^2 \sum_{m=0}^{N-2} X_m^2 - 2a \sum_{m=1}^{N-1} X_{m-1} X_m\right\}$$

Identifying the form of a Gaussian $$= C - \frac{1}{2}\left\{a^2\left[\frac{1}{0.1^2} + \sum_{m=0}^{N-2} X_m^2\right] - 2a\left[\frac{0.9}{0.1^2} + \sum_{m=1}^{N-1} X_{m-1} X_m\right]\right\}$$

$$\Rightarrow \sigma^2_{a_{post}} = \left[\frac{1}{0.1^2} + \sum_{m=0}^{N-2} X_m^2\right]^{-1} \qquad \mu_{a_{post}} = \frac{\left[\frac{0.9}{0.1^2} + \sum_{m=1}^{N-1} X_{m-1} X_m\right]}{\left[1/0.1^2 + \sum_{m=0}^{N-2} X_m^2\right]}$$

$$\Rightarrow p(a \mid X_0 - X_{N-1}) = N(a \mid \mu_{apost}, \sigma^2_{apost})$$

Evaluating $S_1(x) = 95, S_2(x) = 97$:

$\mu_{apost} = 0.9391$

$\sigma^2_{apost} = 0.0051$

$$P(\text{"filter is unstable"}) = \int_1^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{apost}} \exp\left\{\frac{-(x-\mu_{post})^2}{2\sigma^2_{post}}\right\} dx + \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}\sigma_{post}} \exp\left\{\frac{-(x-\mu_{post})^2}{2\sigma^2_{post}}\right\} dx$$

$$= \overbrace{1 - \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}\sigma_{apost}} \exp\left\{\frac{-(x-\mu_{post})^2}{2\sigma^2_{post}}\right\} dx}^{1} + \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}\sigma_{post}} \exp\left\{\frac{-(x-\mu_{post})^2}{2\sigma^2_{post}}\right\} dx$$

(change of variables) $$= 1 - \int_{-\infty}^{\frac{1-\mu_{post}}{\sigma_{post}}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt + \int_{-\infty}^{\frac{-1-\mu_{post}}{\sigma_{post}}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt$$

(Φ: standard normal CDF) $$= 1 - \Phi\left(\frac{1-\mu_{post}}{\sigma_{post}}\right) + \Phi\left(\frac{-1-\mu_{post}}{\sigma_{post}}\right)$$

$$1 - \sim 0.8 + \sim 0$$

$$P(\text{"unstable"}) \approx 0.2.$$

Truncated (gaussian) prior, filter is stable, i.e., a lies in $(-1, 1)$

posterior:

$$p_{trunc}(a \mid X) = \begin{cases} \frac{1}{C} N(a \mid \mu_{post}, \sigma^2_{post}) & \text{if } a \in (-1, 1) \\ 0 & \text{if not} \end{cases}$$

To normalise the distribution, we need to calculate

$C = \int_{-1}^{1} N(a \mid \mu_{post}, \sigma^2_{post})\, da$, which is simply $C = 1 - P(\text{"filter unstable"}) \approx 0.8$

· Since the maximum is a point in the stable region, the Bayesian MAP estimates of the original (unconstrained) and the truncated PDF are the same

· MMSE: $$\hat{a}_{MMSE} = \frac{1}{0.8} \int_{-1}^{1} a \frac{1}{\sqrt{2\pi\sigma_{post}^2}} \cdot \exp\left\{-\frac{1}{2}\frac{(a-\mu_{post})^2}{\sigma_{post}^2}\right\} da$$

Change of variables: $\alpha = \frac{a-\mu_{post}}{\sigma_{post}} \Rightarrow a = \sigma_{post}\alpha + \mu_{post}$

$da = \sigma_{post} \cdot d\alpha$

$$\hat{a}_{MMSE} = \frac{1}{0.8}\left[\underbrace{\int_{-\frac{1-\mu_{post}}{\sigma_{post}}}^{\frac{1-\mu_{post}}{\sigma_{post}}} \frac{1}{\sqrt{2\pi}\,\cancel{\sigma_{post}}} \cancel{\sigma_{post}}\,\alpha \cdot \exp\left\{-\frac{1}{2}\alpha^2\right\}(\sigma_{post}\,d\alpha)}_{(*)} + \mu_{post}\underbrace{\int_{-\frac{1-\mu_{post}}{\sigma_{post}}}^{-\frac{1-\mu_{post}}{\sigma_{post}}} \frac{1}{\sqrt{2\pi}\,\cancel{\sigma_{post}}} \exp\left\{-\frac{1}{2}\alpha^2\right\}(\cancel{\sigma_{post}}\,d\alpha)}_{= C = 0.8031}\right]$$

(already calculated; it's just the prob. that the filter is stable)!

(*) From an inspection/guess, it's easy to tell that the primitive of $\int \alpha \exp(-1/2\,\alpha^2)\, d\alpha$ is $(1 - \exp(-1/2\,\alpha^2))$. then:

$$\hat{a}_{MMSE} = \underbrace{\frac{C}{0.8}\mu_{post}}_{0.9391} + \frac{\sigma_{post}}{\sqrt{2\pi}}\left[1 - \exp -\frac{1}{2}\alpha^2\right]\Bigg|_{\frac{1-\mu_{post}}{\sigma_{post}}}^{\frac{1-\mu_{post}}{\sigma_{post}}}$$

$$= 0.9391 + \frac{\sqrt{0.0051}}{\sqrt{2\pi}}\left[-\exp\left(-\frac{1}{2}\left(\frac{1-\mu_{post}}{\sigma_{post}}\right)^2\right) + \exp\left(-\frac{1}{2}\left(-\frac{1-\mu_{post}}{\sigma_{post}}\right)^2\right)\right] = 0.9144$$