The Prediction

At time $t-1$: to predict $\mathbf{X}_t$: $P(\mathbf{X}_t|\mathbf{Z}_{1:t-1})$

$$P(\mathbf{X}_t) = \sum_{\forall \mathbf{X}_{t-1}} P(\mathbf{X}_t|\mathbf{X}_{t-1})P(\mathbf{X}_{t-1})$$

$$P(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \sum_{\forall \mathbf{X}_{t-1}} P(\mathbf{X}_t|\mathbf{X}_{t-1},\mathbf{Z}_{1:t-1})P(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})$$

Second term: the previous updated density!
First term: no dependence on any observation
State/Process Dynamics: e.g., 1-order Markov:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{n_t} + \mathbf{C}$$

$$\text{Gaussian: } \mathbf{n} = \mathbf{B}^{-1}(\mathbf{X}_t - \mathbf{A}\mathbf{X}_{t-1} - \mathbf{C})$$

$P(\mathbf{X}_t|\mathbf{X}_{t-1})$: Gaussian. Drawing from $P(\mathbf{X}_t|\mathbf{X}_{t-1})$: $\equiv$ using the State/Process Dynamics Model on $\mathbf{X}_{t-1}$)
Not general enough? e.g., Constant Velocity Model:

$$\begin{bmatrix} x_t \\ v_t \end{bmatrix} = \begin{bmatrix} 1 & \triangle t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ v_{t-1} \end{bmatrix}$$

The Updating/Correction

Relate $P(\mathbf{Z}_t|\mathbf{X}_t)$ (State/Process Dynamics Model) & $P(\mathbf{X}_t|\mathbf{Z}_{1:t-1})$ (Prediction) to the Updating/Correction

Consider $P(\mathbf{Z}_t|\mathbf{X}_t)$. The Bayes Rule:

$$P(\mathbf{X}_t|\mathbf{Z}_t) = \frac{P(\mathbf{Z}_t|\mathbf{X}_t)P(\mathbf{X}_t)}{\sum_{\forall \mathbf{X}_t} P(\mathbf{Z}_t|\mathbf{X}_t)P(\mathbf{X}_t)}$$

$$P(\mathbf{X}_t|\mathbf{Z}_t) \propto P(\mathbf{Z}_t|\mathbf{X}_t)P(\mathbf{X}_t)$$

$$P(\mathbf{X}_t|\mathbf{Z}_t, \mathbf{Z}_{1:t-1}) \propto P(\mathbf{Z}_t|\mathbf{X}_t, \mathbf{Z}_{1:t-1})P(\mathbf{X}_t|\mathbf{Z}_{1:t-1})$$

$$P(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto P(\mathbf{Z}_t|\mathbf{X}_t)P(\mathbf{X}_t|\mathbf{Z}_{1:t-1})$$

Updated/Corrected $\propto$ Observation/Measurement $\cdot$ Prediction

3. (a) Let $X = \{x_1, \ldots, x_N\}$

Likelihood:

$$\mathcal{L} = p(X \mid \mu, \sigma) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$\log \mathcal{L} = -\frac{N}{2} \log(2\pi) - N \log\sigma - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$ $\left(\frac{1}{2}\right)$

For maximise: (w.r.t. μ)

$$\frac{\partial \log \mathcal{L}}{\partial \mu} = 0$$

$$\frac{-1}{2\sigma^2}\left[\sum_{n=1}^{N} -2(x_n-\mu)\right] = 0$$

$$\Rightarrow \sum_n (x_n - \mu) = 0$$

$$\Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{\sum_n x_n}{N}}$$ (1)

For given data:

$$\hat{\mu}_{MLE} = \frac{60+49+30+40+53+0}{6}$$

$$= 38.67$$ $\left(\frac{1}{2}\right)$

This is probably an under-estimate, due to the last data point being an outlier.

3. (b) Treat μ as a random variable, with prior:

$$p(\mu) = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(\mu-\alpha)^2}{2\beta^2}}$$ $\left(\frac{1}{2}\right)$

Then:

$$p(\mu \mid X) = p(X \mid \mu) \cdot p(\mu) \;/\; p(X)$$

$$= \frac{\left(\prod_n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(\mu-\alpha)^2}{2\beta^2}}}{p(X)}$$ $\left(\frac{1}{2}\right)$

Page No.
Date: / /

For MAP estimate:

$$\frac{\partial p(\mu|X)}{\partial \mu} = 0$$

First let us take the log posterior:

$$\log p(\mu|X) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \sum_n \frac{(x_n-\mu)^2}{2\sigma^2}$$
$$-\frac{1}{2}\log(2\pi) - \log\beta - \frac{(\mu-\alpha)^2}{2\beta^2} - \log p(X)$$

Take derivative w.r.t. μ and set to 0:

$$-\frac{1}{2\sigma^2}\left[\sum_n -2(x_n-\mu)\right] - \frac{\mu-\alpha}{\beta^2} = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_n (x_n-\mu) - \frac{\mu-\alpha}{\beta^2} = 0$$

$$\frac{\sum_n x_n}{\sigma^2} + \frac{\alpha}{\beta^2} = \frac{N\mu}{\sigma^2} + \frac{\mu}{\beta^2}$$

$$\Rightarrow \boxed{\hat{\mu}_{MAP} = \frac{\frac{\sum_n x_n}{\sigma^2} + \frac{\alpha}{\beta^2}}{\frac{N}{\sigma^2} + \frac{1}{\beta^2}}} \quad \left(1\frac{1}{2}\right)$$

Now plug in given values:

$$\hat{\mu}_{MAP} = \frac{\frac{232}{\sigma^2} + \frac{50}{25}}{\frac{6}{\sigma^2} + \frac{1}{25}} = \frac{232 + 2\sigma^2}{6 + \frac{1}{25}\sigma^2} \quad \left(\frac{1}{2}\right)$$

So we see that MAP estimate depends on the variance of the actual marks distribution.
(2) The more this is, the stronger the effect of the prior. Limiting cases:

$\sigma = 0$: $\hat{\mu}_{MAP} = \hat{\mu}_{MLE} = 38.67$ (no effect of prior)

$\sigma \to \infty$: $\hat{\mu}_{MAP} = 50$ (same as prior α)