1 a) Model 2: $S := S_1 = S_2$

$$g_i(x) = \frac{-d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \log(P(C_i))$$

$$\mathcal{L}(\Sigma^{-1}|x) \equiv \sum_{t=1}^{N}\log(g_i(x^t)) \equiv \sum_{t=1}^{N} g_i(x^t)$$

$$= \sum_{t=1}^{N}\left[\frac{-d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x^t-\mu_i)^T\Sigma^{-1}(x^t-\mu_i) + \log(P(C_i))\right]$$

$$\equiv \sum_{t=1}^{N}\left[-\frac{1}{2}\log(|\Sigma|) + \frac{-1}{2}(x^t-\mu_i)^T\Sigma^{-1}(x^t-\mu_i)\right]$$

$$= -\frac{N}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{t=1}^{N}(x^t-\mu_i)^T\Sigma^{-1}(x^t-\mu_i)$$

$$\frac{\partial\mathcal{L}}{\partial\Sigma^{-1}} = -\frac{N}{2}\frac{\partial(\log(|\Sigma|))}{\partial\Sigma^{-1}} - \frac{1}{2}\sum_{t=1}^{N}\frac{\partial\left[(x^t-\mu_i)^T\Sigma^{-1}(x^t-\mu_i)\right]}{\partial\Sigma^{-1}}$$

$$= -\frac{N}{2}(-\Sigma^T) - \frac{1}{2}\sum_{t=1}^{N}(x^t-\mu_i)^T(x^t-\mu_i) = 0$$

$$\frac{N}{\cancel{2}}\Sigma = \frac{1}{\cancel{2}}\sum_{t=1}^{N}(x^t-\mu_i)^T(x^t-\mu_i)$$

$$\boxed{S = \Sigma = \frac{\sum_{t=1}^{N}(x^t-\mu_i)^T(x^t-\mu_i)}{N}}$$

1a) <u>Model 3:</u> $S_1 = \alpha_1 I$ $\quad S_2 = \alpha_2 I$

$$g_i(x) = \frac{-d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_i|)$$
$$- \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \log(P(C_i)).$$

$$\frac{-1}{2}\log(|\Sigma_i|) = \frac{-1}{2}\log(|\alpha_i I|) = \frac{-1}{2}\log(\alpha_i^d) = \frac{-d}{2}\log(\alpha_i)$$

$$\Sigma_i^{-1} = (\alpha_i I)^{-1} = (\alpha_i^{-1}) I$$

$$g_i(x) = \frac{-d}{2}\log(2\pi) - \frac{d}{2}\log(\alpha_i) - \frac{1}{2}\alpha^{-1}(x-\mu_i)^T(x-\mu_i)$$
$$+ \log(P(C_i))$$

$$\mathcal{L}(\alpha_i|x) \equiv \sum_{t=1}^{N} \log(g_i(x^t))$$
$$\equiv \frac{-1}{2}\sum_{t=1}^{N} \log(d\log(\alpha_i) + \alpha_i^{-1}(x^t-\mu_i)^T(x^t-\mu_i))$$
$$\equiv \frac{-1}{2}\sum_{t=1}^{N}\left[d\log(\alpha_i) + \alpha_i^{-1}(x^t-\mu_i)^T(x^t-\mu_i)\right]$$
$$= \frac{-1}{2} \cdot Nd\log(\alpha_i) + \alpha_i^{-1}\sum_{t=1}^{N}(x^t-\mu_i)^T(x^t-\mu_i)$$

$$\frac{\partial\mathcal{L}}{\partial\alpha_i} = \frac{-1}{2} \cdot Nd\,\frac{\partial(\log(\alpha_i))}{\partial\alpha_i} + \frac{\partial(a_i^{-1})}{\partial a_i}\sum_{t=1}^{N}(x^t-\mu_i)^T(x-\mu_i)$$
$$= \frac{-Nd}{2\alpha_i} + \frac{-1}{\alpha_i^2} \cdot \sum_{t=1}^{N}(x^t-\mu_i)^T(x-\mu_i) = 0.$$

$$\boxed{\alpha_i = \frac{2}{Nd}\sum_{t=1}^{N}(x^t-\mu_i)^T(x-\mu_i)}$$

1c) **Error Rates:**

	#	error
Model 1:	test_data1	0.22
	test-data2	0.23
	test-data3	0.11
Model 2:	test_data1	0.17
	test_data2	0.55
	test_data3	0.45
Model 3:	test_data1	0.34
	test_data2	0.38
	test_data3	0.07

ata 3
model3

Based on the results, it seems most likely that test_data3 was distributed with S_1 very different from S_2, which is why models 1 and 3 were much more effective than model 2. It was also likely to have $S_i \approx \alpha_i I$, which meant that Model 3, a less complex model, had less error than model 1.

ata 1
model2

It also likely that test-data1 had $S_1 \approx S_2$, which meant that Model 2 (less complex than model 1) had less error. It is also likely that there were many non-zero covariances, which would explain model 3's high error rate.

data2
model1

None of the models were particularly good at learning test_data 2, which means that $S_1 \neq S_2$ and $S_1 \neq \alpha_i I$, but it is possible that a different simplification would allow better fit than model 1.