A TOY EXAMPLE OF CLASSIFICATION MODULO INVARIANCE

ANDREW M. FRASER

1. Summary of Fraser et al.

In the framework of Fraser et al. [], measurements from a single object Y vary with nuisance parameters ϕ. The set of possible nuisance parameters maps any single object to a low dimensional manifold in the space of possible measurements $\mathcal{G}$. They write the j^{th} measurement of object Y_i as

$$g_{ij} = \tau(Y_i, \phi_{ij}) + \varepsilon_{ij}, \tag{1}$$

in which $\varepsilon_{ij} \sim \mathcal{N}(0, \Sigma_w)$ denotes an independent measurement error.

Consider the Taylor series

$$\tau(Y, \phi) = \tau(Y, \bar{\phi}) + V(\phi - \bar{\phi}) + (\phi - \bar{\phi})^t H(\phi - \bar{\phi}) + R,$$

where

$$V_i = \left.\frac{\partial \tau(Y,\phi)}{\partial \phi_i}\right|_{\phi=\bar{\phi}} \text{ and } H_{i,j} = \left.\frac{\partial^2 \tau(Y,\phi)}{\partial \phi_i \partial \phi_j}\right|_{\phi=\bar{\phi}}. \tag{2}$$

If the excursions of ϕ_{ij} from its mean $\bar{\phi}_i$ are small, then

$$\tau(Y_i, \phi_{ij}) \approx \mu_i + V_i(\phi_{ij} - \bar{\phi}_i) \tag{3}$$

where $\mu_i \equiv \tau(Y_i, \bar{\phi}_i)$. And if ϕ_{ij} and ε_{ij} are independent and Gaussian, then

$$g_{ij} \sim \mathcal{N}\left(\mu_i, V_i \Sigma_{\phi,i} V_i^t + \Sigma_w\right), \tag{4}$$

i.e.,

$$\mathbb{P}(g|Y_i) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} \exp\left(-\frac{1}{2}(g-\mu_i)^t \Sigma_i^{-1} (g-\mu_i)\right), \tag{5}$$

where N is the dimension of $\mathcal{G}$. Fraser et al. specify the parameter values as follows:

Class mean, μ_i: Use the mean of the measurements available for individual Y_i.

Class-conditional covariance matrix Σ_i:

$$\Sigma_i \equiv V_i \Sigma_{\phi,i} V_i^t + \Sigma_w \tag{6}$$

Prior probabilities for classes, $\pi(i)$: Fraser et al. set $\pi(i) = \sqrt{|\Sigma_i|}$ so that classification could be done on the basis of Mahalanobis distance alone.

Pooled within-class covariance, Σ_w: Fit to the training data.

Tangent to manifold, V_i: Differentiate the function τ with respect to the nuisance parameters ϕ and evaluate at μ_i.

Covariance of nuisance parameters, $\Sigma_{\phi,i}$: They use a formula, Eq. (11), for this matrix that depends on the second derivative of the function τ and the within-class covariance matrix Σ_w.

With these parameters, the Bayes classifier minimizes a sum of the Mahalanobis distance and a data independent constant,

$$\hat{Y}(g) = \operatorname*{argmax}_i \mathbb{P}(g|Y=i)\pi(i) \tag{7a}$$

$$= \operatorname*{argmin}_i (g-\mu_i)^t \Sigma_i^{-1} (g-\mu_i) + \log \frac{|\Sigma_i|^{\frac{1}{2}}}{\pi(i)}\,. \tag{7b}$$

Fraser et al. drop the last term, saying that $\pi(i) = |\Sigma_i|^{\frac{1}{2}}$ is an acceptable lie.

Choosing Σ_ϕ for each individual Y_i is a compromise between the two contradictory goals: (1) allow large excursions in the tangent directions; (2) limit the error of approximating the manifold with its tangent. Figure 1 illustrates the situation. The intuition is that in order to constrain the distance from points on the tangent to the true manifold, the bounds on displacements along the tangent should be inversely proportional to the second derivative.

In terms of the eigen-decomposition

$$\Sigma_w = \sum_d e_d \lambda_d e_d^t$$

Fraser et al. calculate Σ_ϕ as follows. Break the $\dim(\Phi) \times \dim(\mathcal{G}) \times \dim(\Phi)$ tensor H into components

$$H_d \equiv e_d^t H\,. \tag{8}$$

For each component, define the $\dim(\Phi) \times \dim(\Phi)$ positive definite matrix

$$H_d^+ \equiv \sqrt{H_d H_d}\,, \tag{9}$$

and take the average

$$\bar{H} \equiv \sum_d H_d^+ \lambda_d^{-1}\,. \tag{10}$$

Finally set

$$\Sigma_\phi = \alpha \left(\bar{H}\right)^{-1}, \tag{11}$$

where α functions as a Lagrange multiplier that balances the contradictory goals.

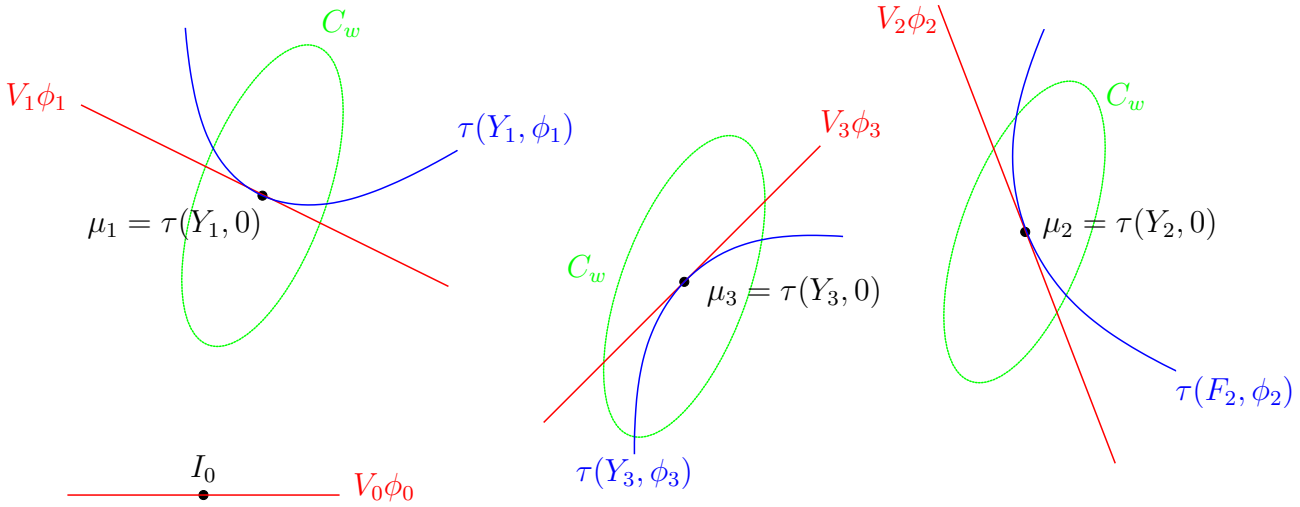

FIGURE 1. A geometric view of the model. For each class $k \in \{1,2,3\}$, the black dot labeled μ_k represents the class mean, the blue curve labeled $\tau(Y_k,\phi_k)$ represents the manifold generated by the nuisance parameters, the red line labeled $V_k\phi_k$ represents the tangent to the manifold at μ_k, and the green ellipse labeled Σ_w represents level sets of $(g-\mu_k)^t\Sigma_w^{-1}(g-\mu_k)$. To obtain class-specific covariance matrices Σ_k, they augment the pooled covariance with a term whose direction is given by the tangent to the manifold and whose magnitude is proportional to the inverse of the curvature of the manifold.

2. A Worked Example

Consider a contrived task in which we are asked to classify measurements of a scalar field on a circle $\theta \in [0, 2\pi)$:

- The four equiprobable classes are:

$$\begin{aligned} f_1(\theta) &= \cos(\theta) \\ f_2(\theta) &= \cos(2\theta) \\ f_3(\theta) &= \cos(3\theta) \\ f_4(\theta) &= \cos(4\theta) \end{aligned}$$

- The components of the measurement vectors g are taken at 64 equally spaced discrete locations

$$\left\{ g(\theta_t) : \theta_0 = 0, \theta_1 = \frac{2\pi}{64}, \theta_2 = \frac{2\cdot 2\pi}{64}, \ldots \theta_2 = \frac{63\cdot 2\pi}{64} \right\}.$$

  The vectors consist of correlated Gaussian noise $\eta \sim \mathcal{N}(0,\Sigma_\eta)$ added to one of the classes shifted by a random time ϕ, i.e.,

$$\begin{aligned} \tau(Y_n,\phi) &= \cos(n(\theta+\phi)) \text{ and} \\ g(\theta_t) &= f_n(\theta_t+\phi) + \eta(\theta_t) \end{aligned}$$

- Σ_η is diagonalized by the discrete Fourier transform $\mathcal{F}$ and has eigenvalues σ_k^2.

The Bayes classifier is

$$\hat{n} = \underset{n}{\operatorname{argmax}}\, p(g|n) \tag{12}$$

$$= \underset{n}{\operatorname{argmax}} \int p(g|n,\phi)p(\phi)d\phi. \tag{13}$$

Note that for $G \equiv \mathcal{F}g$

$$p(g|n,\phi) = \frac{1}{\sqrt{(2\pi)^{64}\prod_k \sigma_k^2}} e^{-\frac{1}{2}\sum_{k\neq n}\frac{G_k^* G_k}{\sigma_k^2}} \times e^{-\frac{1}{2}\frac{(G_n - e^{-in\phi})^*(G_n - e^{-in\phi})}{\sigma_n^2}}.$$

If σ_n^2 is small then $\mathrm{Re}\,(G_n) \approx \cos(\phi)$ and $\mathrm{Im}\,(G_n) \approx \sin(\phi)$, i.e., G_n is near the unit circle in $\mathbb{C}$

The approach of Fraser et al. makes sense if the density $p(\phi)$ is concentrated near $\phi = 0$. In that case, one might simplify Eqn. (13) using the approximation[1]

$$\begin{aligned} -\log \int e^{-\frac{(G_n - e^{-in\phi})^*(G_n - e^{-in\phi})}{2\sigma_n^2}} p(\phi)d\phi \\ \approx \frac{(\mathrm{Re}\,(G_N) - 1)^2}{2\sigma_n^2} + \epsilon\frac{(\mathrm{Im}\,(G_N))^2}{2\sigma_n^2} \end{aligned}$$

where ϵ is inversely proportional to the variance of ϕ. The result is similar to the approach of Fraser et al. as the following illustrates.

[1] I have not verified this integral.

The first and second derivatives of $\tau(Y_n, \phi)$ with respect to ϕ are

$$V_n(t) = \left(-\sin\left(\frac{2\pi nt}{64}\right)\right)\frac{2\pi n}{64} \text{ and}$$
$$H_n(t) = \left(-\cos\left(\frac{2\pi nt}{64}\right)\right)\left(\frac{2\pi n}{64}\right)^2$$

respectively. Working in the frequency domain

$$V_n(\omega_m) = \frac{1}{\sqrt{64}}\frac{-2\pi im}{64}\delta_{|m|,n}$$
$$H_n(\omega_m) = \frac{-1}{\sqrt{64}}\left(\frac{2\pi m}{64}\right)^2\delta_{|m|,n}$$
$$\bar{H} = \frac{(2\pi n)^2}{8(64)^2\sigma_n^2},$$

and the augmentation term (see Eqn. (6)) for the covariance matrix of class n has four components

$$\left(V_n\Sigma_{\phi,n}V_n^t\right)_{j,k} = \begin{cases} \frac{\alpha}{8}\sigma_n^2 & \text{if } j = k = n \text{ or } -j = -k = n \\ -\frac{\alpha}{8}\sigma_n^2 & \text{if } j = -k = n \text{ or } -j = k = n \\ 0 & \text{otherwise.} \end{cases}$$

The resulting classifier

$$\hat{n} = \underset{n}{\operatorname{argmin}}\left(\frac{\left(\operatorname{Re}(G_n) - 1\right)^2}{\sigma_n^2} + \frac{\left(\operatorname{Im}(G_n)\right)^2}{(1+\frac{\alpha}{8})\sigma_n^2} + \sum_{k=0:k\neq n}^{\frac{64}{2}}\frac{|G_k|^2}{\sigma_k^2}\right)$$

deemphasizes $\operatorname{Im}(G_n)$ (which to first order corresponds to time shifts) by the factor $1 + \frac{\alpha}{8}$.