Rao Distance in Statistical Distinguishability

What is Rao distance really?

Given a vector of k parameters $\vec{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta(k)$,

a contiguous point $\vec{\theta'} = (\theta_1 + d\theta_1, \ldots, \theta_k + d\theta_k) \in \Theta(k)$,

and a family $\mathcal{P}$ of probability distributions $p(x|\vec{\theta}) \in \mathcal{P}$, we can define a differential element δ:

$$\delta = -\sum_{i=1}^{k} \frac{\partial \ln p(x|\vec{\theta})}{\partial \theta_i} d\theta_i = \frac{-1}{p(x|\vec{\theta})} \sum_{i=1}^{k} \frac{\partial p(x|\vec{\theta})}{\partial \theta_i} d\theta_i,$$

where the negative is used to think of surprisals. $\langle \cdot \rangle$ is an expectation value with respect to $p(x|\vec{\theta})$.

We show the variance $\Delta\delta^2$ to be a distance measure on $\mathcal{P}$ in the following way:

$$\Delta\delta^2 = \left\langle \delta^2 \right\rangle - \langle \delta \rangle^2 \tag{1}$$

$$\text{abbreviate } p = p(x|\vec{\theta}) \tag{2}$$

$$\text{notice } \langle \delta \rangle = \sum_{i=1}^{k} \left\langle -\frac{\partial \ln p}{\partial \theta_i} \right\rangle d\theta_i = 0 \tag{3}$$

$$\therefore \quad \Delta\delta^2 = \left\langle \left(\sum_{i=1}^{k} \frac{\partial \ln p}{\partial \theta_i} d\theta_i \right) \left(\sum_{j=1}^{k} \frac{\partial \ln p}{\partial \theta_j} d\theta_j \right) \right\rangle \tag{4}$$

$$= \left\langle \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\frac{\partial \ln p}{\partial \theta_i} \right) \left(\frac{\partial \ln p}{\partial \theta_j} \right) d\theta_i d\theta_j \right\rangle \tag{5}$$

$$= \left\langle \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{1}{p^2} \left(\frac{\partial p}{\partial \theta_i} \right) \left(\frac{\partial p}{\partial \theta_j} \right) d\theta_i d\theta_j \right\rangle \tag{6}$$

$$\text{For convenience, say } x \text{ has } N \text{ states indexed } \omega : \tag{7}$$

$$\therefore \quad \Delta\delta^2 = \sum_{\omega=1}^{N} \frac{[p(x_\omega|\vec{\theta'}) - p(x_\omega|\vec{\theta})]^2}{p(x_\omega|\vec{\theta})} \tag{8}$$

Define the Fisher information matrix

$$[\mathcal{I}_{\mathcal{F}}]_{ij} = \left[\left(\frac{\partial \ln p}{\partial \theta_i} \right) \left(\frac{\partial \ln p}{\partial \theta_j} \right) \right],$$

so Equation 5 can be written as an inner product

$$\Delta\delta^2 = (\vec{\theta} - \vec{\theta}')^T \mathcal{I}_{\mathcal{F}} (\vec{\theta} - \vec{\theta}') \geq 0,$$

which, being non-negative due to the properties of the variance, makes $\mathcal{I}_{\mathcal{F}}$ a positive semi-definite Riemannian metric tensor. As such, we can write a line element in Einstein notation

$$ds^2 = \mathcal{I}_{\mathcal{F}\mu\nu} d\theta^\mu d\theta^\nu,$$

and recognize $ds = \Delta\delta$, so the line element in $\mathcal{P}$ is the standard deviation of the differential element δ.

The Rao distance $\mathcal{D}_{\mathcal{R}}(\vec{\theta}_0, \vec{\theta}_f)$ between initial and final parameter vectors is simply the integral over the line element,

$$\mathcal{D}_{\mathcal{R}}(\vec{\theta}_0, \vec{\theta}_f) = \int_{p(x|\vec{\theta}_0)}^{p(x|\vec{\theta}_f)} ds = \int_{\vec{\theta}_0}^{\vec{\theta}_f} \sqrt{\mathcal{I}_{\mathcal{F}\mu\nu} d\theta^\mu d\theta^\nu} \tag{9}$$

$$= \int_{\vec{\theta}_0}^{\vec{\theta}_f} \Delta\delta \tag{10}$$

Distance and Distinguishability

According to Wootters and others, $\mathcal{D}_{\mathcal{R}}(\vec{\theta}_0, \vec{\theta}_f)$ has something to do with the distinguishability of probability distributions by their standard deviations. Specifically, we can "keep track" of how many standard errors $\Delta\hat{\theta}_i$ of an unbiased estimator $\hat{\theta}_i$ have been exceeded along the path of integration $p(x|\vec{\theta}_0) \to p(x|\vec{\theta}_f)$ by differences in scalar elements, $\theta_i' - \theta_i$.

As one avenue to explore this concept, let's diagonalize the metric

$$\mathcal{I}_{\mathcal{F}} = V \Lambda V^T,$$

with V being an orthogonal transformation used to define a new basis for parameter vector differences $\vec{\theta} \to \vec{\eta}$

$$(\vec{\eta} - \vec{\eta}') = V^T (\vec{\theta} - \vec{\theta}')$$

and Λ being a diagonal Fisher matrix of uncorrelated scores for this choice of coordinates:

$$\Lambda = \begin{bmatrix} \left\langle \left(\frac{\partial \ln p}{\partial \eta_1} \right)^2 \right\rangle & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left\langle \left(\frac{\partial \ln p}{\partial \eta_k} \right)^2 \right\rangle \end{bmatrix}$$

We can substitute this spectral decomposition into the line element like so:

$$ds^2 = (\vec{\theta} - \vec{\theta}')^T V \Lambda V^T (\vec{\theta} - \vec{\theta}') = (\vec{\eta} - \vec{\eta}')^T \Lambda (\vec{\eta} - \vec{\eta}') \tag{11}$$

$$= \begin{bmatrix} d\eta_1 & \dots & d\eta_k \end{bmatrix} \begin{bmatrix} \left\langle \left(\frac{\partial \ln p}{\partial \eta_1} \right)^2 \right\rangle & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left\langle \left(\frac{\partial \ln p}{\partial \eta_k} \right)^2 \right\rangle \end{bmatrix} \begin{bmatrix} d\eta_1 \\ \vdots \\ d\eta_k \end{bmatrix} \tag{12}$$

$$= \sum_{i=1}^{k} \left\langle \left(\frac{\partial \ln p}{\partial \eta_i} \right)^2 \right\rangle d\eta_i^2 \geq \sum_{i=1}^{k} \frac{d\eta_i^2}{\Delta \hat{\eta}_i^2}, \tag{13}$$

invoking the Cramer-Rao inequality on unbiased estimators $\hat{\eta}_i$ of the transformed parameters in $\vec{\eta}$.

We now see $ds \geq \sqrt{\sum_{i=1}^{k} \frac{d\eta_i^2}{\Delta \hat{\eta}_i^2}}$, so the line element is at least as large as the Euclidean distance of the vector $\begin{bmatrix} \frac{d\eta_1}{\Delta \hat{\eta}_1} \\ \vdots \\ \frac{d\eta_k}{\Delta \hat{\eta}_k} \end{bmatrix}$, which "keeps track" of the number of standard errors $\Delta \hat{\eta}_i$ exceed by contiguous differences $d\eta_i$ in the change represented by the line element. This Euclidean distance, in a rather different space of $\frac{d\eta_k}{\Delta \hat{\eta}_k}$'s, is a clear sense of the distinguishability between $p(x|\eta)$ and $p(x|\eta')$, and therefore between $p(x|\theta)$ and $p(x|\theta')$. It does not quite "count the distinguishable distributions," but it does accumulate the differences in parameter values and how many times they exceed the standard error of estimating. In one parametric dimension $\vec{\theta} = \theta_1$, this is simply a number of standard deviations: $\frac{\theta_1' - \theta_1}{\Delta \hat{\theta}_1}$.

In short, the Rao distance is a distinguishability measure in the sense of the following expression:

$$\mathcal{D}_{\mathcal{R}}(\vec{\theta_0}, \vec{\theta_f}) \geq \int_{p(x|\vec{\theta}_0)}^{p(x|\vec{\theta}_0)} \sqrt{\sum_{i=1}^{k} \frac{d\eta_i^2}{\Delta \hat{\eta}_i^2}},$$

upper-bounding this Euclidean distance of $\begin{bmatrix} \frac{d\eta_1}{\Delta \hat{\eta}_1} \\ \vdots \\ \frac{d\eta_k}{\Delta \hat{\eta}_k} \end{bmatrix}$, which is basically keeping track of the number of times the transformed parameter differences $d\eta_i$ extend past the standard errors $\Delta \hat{\eta}_i$ of their respective unbiased estimators. **If the difference between two parameter values does not exceed the standard error, the distributions of those parameter values are not distinguishable.** In this sense, accumulating each infinitesimal difference with summation and integration is indeed evaluating the distinguishability of the distributions $p(x|\vec{\theta_0})$ and $p(x|\vec{\theta_f})$. So it goes!

References

1. Atkinson, C., & Mitchell, A. F. (1981). Rao's distance measure. Sankhyā: The Indian Journal of Statistics, Series A, 345-365.
2. Wootters, W. K. (1981). Statistical distance and Hilbert space. Physical Review D, 23(2), 357.