Lecture-21: Local behavior for parametrized family

1 Parametrized family

Consider a statistical experiment $\mathcal{P}(\Theta) \triangleq \{P_\theta \in \mathcal{M}(\mathcal{X}) : \theta \in \Theta\}$ for the parameter space $\Theta \subseteq \mathbb{R}^d$ to be an open subset. We assume that there exists a measure $\mu \in \mathcal{M}(\mathcal{X})$ such that $P_\theta \ll \mu$ for all $\theta \in \Theta$ and denote the relative density by $p_\theta \triangleq \frac{dP_\theta}{d\mu}$.

1.1 Fisher information

Definition 1.1 (Fisher matrix). If the map $\theta \mapsto p_\theta(x)$ is smooth for each $x \in \mathcal{X}$, then we can define *score* $V \triangleq \nabla_\theta \ln p_\theta(X)$ and the *Fisher information matrix* with respect to the parameter θ as

$$J_F(\theta) \triangleq \mathbb{E}[VV^\top \mid \theta] \triangleq \mathbb{E}_{X\sim P_\theta} VV^\top.$$

Lemma 1.2. *Under suitable regularity conditions, we have the identity* $\mathbb{E}[V \mid \theta] = 0$ *and several equivalent expressions for the Fisher information matrix, such as*

$$J_F(\theta) = \operatorname{cov}[V \mid \theta] = 4\int_{\mathcal{X}} d\mu \nabla_\theta \sqrt{p_\theta}(\nabla_\theta \sqrt{p_\theta})^\top = -\mathbb{E}[\operatorname{Hess}_\theta(\ln p_\theta(X))] \mid \theta].$$

Proof. We observe that $V = \nabla_\theta \ln p_\theta(X) = \frac{1}{p_\theta(X)} \nabla_\theta p_\theta(X)$.

(a) Under suitable regularity conditions, we can exchange integration and derivative to obtain the following identity

$$\mathbb{E}[V \mid \theta] = \int_{\mathcal{X}} d\mu(x) p_\theta(x) \nabla_\theta \ln p_\theta(x) = \int_{\mathcal{X}} d\mu(x) \nabla_\theta p_\theta(x) = \nabla_\theta \int_{\mathcal{X}} d\mu(x) p_\theta(x) = 0.$$

(b) Under the same regularity conditions, we obtain $V - \mathbb{E}[V \mid \theta] = V$, and hence we can write

$$\operatorname{cov}(V \mid \theta) = \mathbb{E}[(V - \mathbb{E}[V \mid \theta])(V - \mathbb{E}[V \mid \theta])^\top \mid \theta] = \mathbb{E}[VV^\top \mid \theta] = J_F(\theta).$$

(c) Since $V = \frac{1}{p_\theta(X)} \nabla_\theta p_\theta(X)$, we obtain $VV^\top = \frac{1}{p_\theta(X)^2} \nabla_\theta p_\theta(X)(\nabla_\theta p_\theta(X))^\top$. Furthermore, we have $\nabla_\theta \sqrt{p_\theta(x)} = \frac{1}{2\sqrt{p_\theta(x)}} \nabla_\theta p_\theta(x)$. Combining these two facts, we obtain

$$J_F(\theta) = \mathbb{E}[VV^\top \mid \theta] = \int_{\mathcal{X}} d\mu(x) \frac{1}{p_\theta(x)} \nabla_\theta p_\theta(x)(\nabla_\theta p_\theta(x))^\top = 4\int_{\mathcal{X}} d\mu(x) \Big(\nabla_\theta \sqrt{p_\theta(x)}\Big)\Big(\nabla_\theta \sqrt{p_\theta(x)}\Big)^\top.$$

(d) From the definition of Hessian, we can write

$$[\operatorname{Hess}_\theta \ln p_\theta(X)]_{ij} = \frac{\partial}{\partial \theta_j}\Big[\frac{1}{p_\theta(X)} \frac{\partial}{\partial \theta_i} p_\theta(X)\Big] = -\frac{1}{p_\theta(X)^2} \frac{\partial}{\partial \theta_j} p_\theta(X) \frac{\partial}{\partial \theta_i} p_\theta(X) + \frac{1}{p_\theta(X)} \frac{\partial^2}{\partial \theta_j \partial \theta_i} p_\theta(X).$$

Under suitable regularity conditions, we can exchange integration and derivative to obtain the following identity

$$-\mathbb{E}[(\operatorname{Hess}_\theta \ln p_\theta(X))_{ij} \mid \theta] = \mathbb{E}[(VV^\top)_{ij} \mid \theta] + \int_{\mathcal{X}} d\mu(x) \frac{\partial^2}{\partial \theta_j \partial \theta_i} p_\theta(x) = (J_F(\theta))_{ij} + \frac{\partial^2}{\partial \theta_j \partial \theta_i} \int_{\mathcal{X}} d\mu(x) p_\theta(x).$$

□

1.2 Local behavior of divergence for parametrized family

The significance of Fisher information matrix arises from the fact that it gauges the local behavior of divergence for smooth parametric families.

Lemma 1.3. *Under suitable technical conditions*[1]*,*

$$D(P_{\theta_0}\|P_{\theta_0+\xi}) = \frac{1}{2}\xi^\top J_F(\theta_0)\xi + o(\|\xi\|^2). \tag{1}$$

Proof. Let $P_{\theta_0} \ll \mu$ for some measure $\mu \in \mathcal{M}(\mathcal{X})$ such that there exists relative density $p_{\theta_0} \triangleq \frac{P_{\theta_0}}{\mu}$. We can write the Taylor series expansion for $\ln p_{\theta_0+\xi}(x)$ for first two terms, in the neighborhood of $\ln p_{\theta_0}(x)$, as

$$\ln p_{\theta_0+\xi}(x) = \ln p_{\theta_0}(x) + \xi^\top \nabla_\theta \ln p_{\theta_0}(x) + \frac{1}{2}\xi^\top \operatorname{Hess}_\theta(\ln p_{\theta_0}(x))\xi + o(\|\xi\|^2).$$

Recall that $D(P_{\theta_0}\|P_{\theta_0+\xi}) \triangleq \mathbb{E}_{X\sim P_{\theta_0}} \ln \frac{p_{\theta_0}(X)}{p_{\theta_0+\xi}(X)}$, and the result follows from the fact that $\mathbb{E}[V \mid \theta_0] = 0$. □

Remark 1. We will establish this fact rigorously later. Property (1) is of paramount importance in statistics. We should remember it as *Divergence is locally quadratic on the parameter space, with Hessian given by the Fisher information matrix*.

Exercise 1.4 (KL Divergence for GLM). Consider the Gaussian location model, where the parametrized distribution for observations is given by $P_\theta \triangleq \mathcal{N}(\theta,\Sigma)$.
(a) Show that $J_F(\theta) = \Sigma^{-1}$.
(b) Consider unconstrained parameter space $\Theta \subseteq \mathbb{R}^d$, and Gaussians $P_{\theta_i} \triangleq \mathcal{N}(\theta_i,\Sigma_i)$ for $i \in \{0,1\}$. Assuming $\det\Sigma_0 \neq 0$, show that

$$D(P_{\theta_0}\|P_{\theta_1}) = \frac{1}{2}(\theta_0-\theta_1)^\top\Sigma_1^{-1}(\theta_0-\theta_1) + \frac{1}{2}\Big(\ln\det\Sigma_1 - \ln\det\Sigma_0 + \operatorname{tr}(\Sigma_1^{-1}\Sigma_0 - I_d)\Big).$$

Remark 2. As another example, note that the result that KL divergence is locally like χ^2 divergence, is a special case of Property (1) by considering $P^\lambda = \bar\lambda Q + \lambda P$ parametrized by $\lambda \in [0,1]$. In this case, the Fisher information at $\lambda = 0$ is simply $\chi^2(P\|Q)$. Nevertheless, local behavior of KL divergence of convex combination of two measures is completely general while the asymptotic expansion (1) holds under certain regularity conditions.

Exercise 1.5. Let $P,Q \in \mathcal{M}(\mathcal{X})$ and $\Theta \triangleq [0,1]$ and define $P_\theta \triangleq \theta P + \bar\theta Q \in \mathcal{M}(\mathcal{X})$. Show that $\lim_{\theta\downarrow 0} J_F(\theta) = \chi^2(P\|Q)$.

Remark 3. Some useful properties of Fisher information are as follows.

- **Reparametrization:** It can be seen that if one introduces another parametrization $\tilde\theta \in \tilde\Theta$ by means of a smooth invertible map $\tilde\Theta \to \Theta$, then denoting the Jacobian of this map $A \triangleq \frac{d\theta}{d\tilde\theta}$, we can write the Fisher information matrix for reparametrization as $J_F(\tilde\theta) = A^\top J_F(\theta) A$. So we can see that J_F transforms similarly to the metric tensor in Riemannian geometry. This idea can be used to define a Riemannian metric on the parameter space Θ, called the *Fisher-Rao metric*.
- **Additivity:** Suppose we are given a sample of *i.i.d.* observations $X : \Omega \to \mathcal{X}^m$ under common distribution P_θ. As such, consider the parametrized family of product distributions $\{P_\theta^{\otimes m} : \theta \in \Theta\}$, whose Fisher information matrix is denoted by $J_F^{\otimes m}(\theta)$. For each $\theta \in \Theta$, let $P_\theta \ll \mu$ for some dominating measure $\mu \in \mathcal{M}(\mathcal{X})$, then the relative density is denoted by $p_\theta \triangleq \frac{d}{d\mu}P_\theta$. Recall that $p_\theta^{\otimes m}(X) = \prod_{i=1}^m p_\theta(X_i)$, and hence the score for this sample is $V = \nabla_\theta \ln p_\theta^{\otimes m}(X) = \sum_{i=1}^m \nabla_\theta \ln p_\theta(X_i)$, and $\operatorname{Hess}_\theta \ln p_\theta^{\otimes m}(X) = \sum_{i=1}^m \operatorname{Hess}_\theta \ln p_\theta(X_i)$. From the linearity of expectation, and equivalent expression for Fisher information, we obtain $J_F^{\otimes m}(\theta) = mJ_F(\theta)$.

[1] To illustrate the subtlety here, consider a scalar location family, i.e. $p_\theta(x) = f_0(x-\theta)$ for some density f_0. In this case, Fisher information $J_F(\theta_0) = \int \frac{(f_0')^2}{f_0}$ does not depend on θ_0 and is well-defined even for compactly supported f_0, provided f_0 vanishes at the endpoints sufficiently fast. But at the same time the left-hand side of Property 1 is infinite for any $\xi > 0$. Thus, a more general interpretation for Fisher information is as the coefficient in expansion $D(P_{\theta_0}\|\frac{1}{2}P_{\theta_0} + \frac{1}{2}P_{\theta_0+\xi}) = \frac{\xi^2}{8}J_F(\theta_0) + o(\xi^2)$. We will discuss this in more detail.

Lemma 1.6 (Matrix inversion). *Let A, C be invertible matrices and $UCVA^{-1}$ has spectral radius smaller than unity, then*

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}.$$

Proof. If B, D are two invertible matrices then $B^{-1}D^{-1} = (DB)^{-1}$, and hence

$$C^{-1}(C^{-1}+VA^{-1}U)^{-1} = (I+VA^{-1}UC)^{-1}, \qquad A^{-1}(I+UCVA^{-1})^{-1} = (A+UCV)^{-1}.$$

Further, for any matrix B with spectral radius smaller than unity, we have $(I-B)^{-1} = \sum_{n\in\mathbb{Z}_+} B^n$, and

$$(I+VA^{-1}UC)^{-1} = \sum_{n\in\mathbb{Z}_+}(-VA^{-1}UC)^n = I - VA^{-1}\sum_{n\in\mathbb{Z}_+}(-UCVA^{-1})^n UC = I - VA^{-1}(I+UCVA^{-1})^{-1}UC.$$

Pre-multiplying with $A^{-1}UC$ and post-multiplying with VA^{-1}, we obtain

$$\begin{aligned} &A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} = A^{-1}UCVA^{-1} - A^{-1}UCVA^{-1}(I+UCVA^{-1})^{-1}UCVA^{-1} \\ &= A^{-1}(I+UCVA^{-1}-I)(I+UCVA^{-1})^{-1} = A^{-1}(I-(I+UCVA^{-1})^{-1}) = A^{-1}-(A+UCV)^{-1}. \end{aligned}$$

The result follows by rearranging terms. □

Lemma 1.7. *Following identities are true for determinants.*
(a) For commutative matrices C, D with spectral radius of CD smaller than unity, $\det(I+CD) = \det(I+DC)$.
(b) Let A be an invertible matrix and $y^\top A^{-1}x \leqslant 1$, then $\det(A+xy^\top) = (1+y^\top A^{-1}x)\det(A)$.

Proof. We will show them separately.
(a) Let $((x_i,\lambda_i) : i\in[d])$ be eigenvector and eigenvalue pairs for $I+CD$, such that $(I+CD)x_i = \lambda_i x_i$ for each $i\in[d]$. Defining matrix $X \triangleq \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}$ and $\Lambda \triangleq \operatorname{diag}(\lambda_1,\ldots,\lambda_d)$, we obtain

$$(I+CD)X = \begin{bmatrix} \lambda_1 x_1 & \ldots & \lambda_d x_d \end{bmatrix} = X\Lambda.$$

Pre-multiplying both sides by D, we obtain $(I+DC)DX = DX\Lambda$, i.e. (DX,Λ) is the eigenvector matrix and diagonal eigenvalue matrix for $I+DC$. In particular, matrices $I+CD$ and $I+DC$ have same eigenvalues and hence the same determinant.
(b) We observe that $A+xy^\top = A(I+A^{-1}xy^\top)$. Applying the product property of determinants for commutative matrices $A^{-1}x$ and $y^\top$, we obtain

$$\det(A+xy^\top) = \det(A)\det(I+A^{-1}xy^\top) = \det(A)(1+y^\top A^{-1}x).$$

□

Example 1.8. Let the input space $\mathcal{X} \triangleq \{0,\ldots,d\}$ and consider a stochastic model $P_\theta \in \mathcal{M}(\mathcal{X})$ that generates the observation $X : \Omega \to \mathcal{X}$ generated by parameter θ. Since $\mathcal{X}$ is discrete, each P_θ is a probability mass function of the form $\theta_x \triangleq P_\theta(x)$ for $x\in\mathcal{X}$. It follows that the parameter space $\Theta \triangleq \mathcal{M}(\mathcal{X})$. We observe that $\theta_0 - 1 - \sum_{i\in[d]}\theta_i$, and hence we can take all derivatives only with respect to free parameters $\theta \triangleq (\theta_1,\ldots,\theta_d)$. In particular, $\frac{d}{d\theta_i}P_\theta(x) = \mathbb{1}_{\{x=i\}} - \mathbb{1}_{\{x=0\}}$ for $i\in[d]$. In terms of unit vectors $(e_i : i\in[d])$ and all one vector $\mathbf{1} \triangleq \sum_{i=1}^d e_i$, we can write $\nabla_\theta P_\theta(x) = \sum_{i=1}^d I_{xi}e_i - I_{x0}$. We can write the score as $V \triangleq \nabla_\theta \ln P_\theta(X) = \frac{1}{P_\theta(X)}\nabla_\theta P_\theta(X)$. It follows that

$$VV^\top = \frac{1}{(P_\theta(X))^2}(\sum_{i=1}^d I_{Xi}e_i - I_{X0})^2 = \frac{1}{\theta_X^2}(\sum_{i=1}^d I_{Xi}e_ie_i^\top + I_{X0}).$$

We can write the Fisher information matrix in terms of all one $d\times 1$ vector $\mathbf{1}$, as

$$J_F(\theta) \triangleq \mathbb{E}[VV^\top \mid \theta] = \sum_{x=1}^d \frac{1}{\theta_x}e_xe_x^\top + \frac{1}{\theta_0} = \operatorname{diag}\Big(\frac{1}{\theta_1},\ldots,\frac{1}{\theta_d}\Big) + \frac{1}{1-\sum_{i=1}^d\theta_i}\mathbf{1}\mathbf{1}^\top.$$

Let $D \triangleq \operatorname{diag}(\theta_1,\ldots,\theta_d)$ and $\alpha = 1-\mathbf{1}^\top D\mathbf{1}$, then $(D^{-1}+\alpha^{-1}\mathbf{1}\mathbf{1}^\top)^{-1} = D - D\mathbf{1}(\alpha+\mathbf{1}^\top D\mathbf{1})^{-1}\mathbf{1}^\top D$ by Lemma 1.6, and hence $J_F^{-1}(\theta) = \operatorname{diag}(\theta) - \theta\theta^\top$. For the determinant, it follows from Lemma 1.7 that $\det J_F(\theta) = (1-\theta^\top D^{-1}\theta)^{-1}\det(D)^{-1} = \prod_{i=0}^d \frac{1}{\theta_i}$.

Example 1.9 (Location family). In statistics and information theory it is common to talk about Fisher information of a (single) random variable or a distribution without reference to a parametric family. In such cases one is implicitly considering a location parameter. Specifically, for any density p_0 on $\mathbb{R}^d$ we define a location family of distributions on $\mathbb{R}^d$ by setting $dP_\theta(x) \triangleq p_0(x-\theta)dx$, for all $\theta \in \mathbb{R}^d$. Note that $J_F(\theta)$ here does not depend on θ. For this special case, we will adopt the standard notation. Let $X \sim p_0$, then

$$J(X) \triangleq J(p_0) \triangleq \mathbb{E}_{X\sim p_0}[(\nabla \ln p_0(X))(\nabla \ln p_0(X))^\top] = -\mathbb{E}_{X\sim p_0}[\mathrm{Hess}(\ln p_0(X))],$$

where the second equality requires applicability of integration by parts.