Lecture 10

Math 178

Nonlinear Data Analytics

Introduction of
Fisher Information
and
Fisher Information Metric

Prof. Weiqing Gu

1 Maximum likelihood estimation and Fisher information

Suppose we have observe $\{x_1, \dots, x_n\}$, independently drawn from a random variable X with PDF $p(x;\theta)$, where θ is an unknown parameter. Which parameter θ is most likely, given this observation? One approach is known as maximum likelihood estimation.

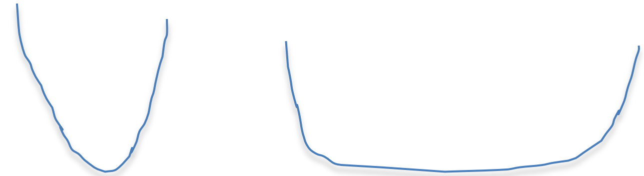
First, we observe that the *likelihood* that we observe this set of observations $\{x_1, \dots, x_n\}$ is

$$L(\theta; x_1, \dots, x_n) \equiv \prod_{i=1}^{n} p(x_i; \theta).$$

It is often more convenient to consider the logarithm of the likelihood[1]:

$$\ell(\theta; x_1, \dots, x_n) \equiv \log L(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \log p(x_i; \theta),$$

which will be maximal when the likelihood is maximal.



What's the θ that gives the maximum likelihood? We would differentiate the log-likelihood and set it to 0:

$$0 = \frac{\partial \ell}{\partial \theta}\,.$$

Then, we solve for θ to find the parameter that most likely generated the observations we observed. This optimal parameter θ_{MLE} is called the *maximum likelihood estimator*, and it means that we believe that $\{x_1, \ldots, x_n\}$ were drawn from the PDF $p(x; \theta_{\text{MLE}})$.

Once we have a maximum likelihood estimator θ, we might be interested in *how* optimal this estimate is – how much can we trust this estimate? In the following, we will quantify how optimal a given maximum likelihood estimator is. The resulting quantity is called the *Fisher information* $I(\theta)$.

Since θ is a maximum likelihood estimator, θ is a local maximum of the log-likelihood function. If the log-likelihood function is sharply peaked around θ, then the values surrounding θ are extremely unlikely compared to θ, in which case θ is a good estimate. By constrast, if the log-likelihood function is relatively flat around θ, then surrounding parameters are less likely than θ, but still comparatively likely. In this case θ, is a poor estimate.

In calculus, the second derivative gives a measure of how sharply a function is curving; therefore, the second derivative of the log-likelihood function will be a good measure of how sharply peaked the log-likelihood is. Thus, we define

$$I(\theta; x_1, \ldots, x_n) \equiv -\frac{\partial^2 \ell}{\partial \theta^2} = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(x_i; \theta),$$

where the minus sign is a convention to ensure that $I \geq 0$ for a maximum. The better the estimate, the greater I is.

This is clearly related to the curvature at this point! Bring the geometry into the picture of Fisher Information!

Since θ is a maximum likelihood estimator, θ is a local maximum of the log-likelihood function. If the log-likelihood function is sharply peaked around θ, then the values surrounding θ are extremely unlikely compared to θ, in which case θ is a good estimate. By constrast, if the log-likelihood function is relatively flat around θ, then surrounding parameters are less likely than θ, but still comparatively likely. In this case θ, is a poor estimate.

In calculus, the second derivative gives a measure of how sharply a function is curving; therefore, the second derivative of the log-likelihood function will be a good measure of how sharply peaked the log-likelihood is. Thus, we define

$$I(\theta; x_1, \ldots, x_n) \equiv -\frac{\partial^2 \ell}{\partial \theta^2} = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(x_i; \theta),$$

where the minus sign is a convention to ensure that $I \geq 0$ for a maximum. The better the estimate, the greater I is.

Now we move to the limit where the number of observations $n \to \infty$. In this limit, by the law of large numbers,

$$\frac{1}{n} I(\theta; x_1, \ldots, x_n) \to -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log p(X; \theta)\right],$$

where $\mathbb{E}$ denotes the expectation value with respect to X, which is distributed according to $p(x;\theta)$. With this in mind, we define

$$\boxed{I(\theta) \equiv -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log p(X; \theta)\right] = -\int p(x;\theta)\, \frac{\partial^2}{\partial\theta^2} \log p(x;\theta)\, \mathrm{d}x.}$$

This is the *Fisher information* at the parameter θ, and it measures how sharply the likelihood is peaked at the parameter θ in the limit of an infinite number of observations.

It is useful to derive an alternate expression for the Fisher information. The *efficient score* $V(\theta; x)$ is defined as

$$V(\theta; x) = \frac{\partial}{\partial \theta} \log p(x; \theta).$$

Recall that the maximum likelihood condition was setting

$$0 = \frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} V(\theta; x_i).$$

Therefore, it is not surprising that if θ is the maximum likelihood estimator, then

$$\mathbb{E}[V(\theta; X)] = 0.$$

We can confirm this by calculating:

$$\begin{aligned}
\mathbb{E}[V(\theta;X)] &= \int V(\theta;x)\,p(x;\theta)\,\mathrm{d}x \\
&= \int \left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] p(x;\theta)\,\mathrm{d}x \\
&= \int \frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)} \cdot p(x;\theta)\,\mathrm{d}x \\
&= \int \frac{\partial}{\partial\theta}\,p(x;\theta)\,\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\int p(x;\theta)\,\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\,1 \\
&= 0.
\end{aligned}$$

Now differentiate both sides of

$$\mathbb{E}[V(\theta; X)] = \int V(\theta; x)\, p(x; \theta)\, \mathrm{d}x = 0$$

with respect to θ to get

$$\int \left[\frac{\partial}{\partial\theta} V(\theta; x)\right] p(x; \theta)\, \mathrm{d}x + \int V(\theta; x) \left[\frac{\partial}{\partial\theta}\, p(x; \theta)\right] \mathrm{d}x = 0.$$

The first term is

$$\int \left[\frac{\partial}{\partial\theta} V(\theta;x)\right] p(x;\theta)\,\mathrm{d}x = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right]$$

$$= -I(\theta).$$

The second term is

$$\int V(\theta;x)\left[\frac{\partial}{\partial\theta} p(x;\theta)\right]\mathrm{d}x = \int V(\theta;x)\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] p(x;\theta)\,\mathrm{d}x$$

$$= \int \left(V(\theta;x)\right)^2 p(x;\theta)\,\mathrm{d}x$$

$$= \mathbb{E}\left[\left(V(\theta;X)\right)^2\right].$$

Additionally, because $\mathbb{E}[V(\theta;X)] = 0$, we also have

$$\boxed{I(\theta) = \mathrm{Var}(V(\theta;X))},$$

where Var denotes the variance with respect to X.

Now suppose that there is more than one parameter θ; replace θ with k parameters $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_k)$. We can look at the mixed derivatives

$$g_{ij} \equiv -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X;\boldsymbol{\theta})\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i} \log p(X;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j} \log p(X;\boldsymbol{\theta})\right)\right]$$

or

$$g_{ij} \equiv -\mathbb{E}\left[\frac{\partial V_i}{\partial\theta_j}\right] = \mathbb{E}\left[V_i(\boldsymbol{\theta};X)\, V_j(\boldsymbol{\theta};X)\right],$$

with

$$V_i(\boldsymbol{\theta};x) \equiv \frac{\partial}{\partial\theta_i} \log p(x;\boldsymbol{\theta}).$$

It is also the covariance

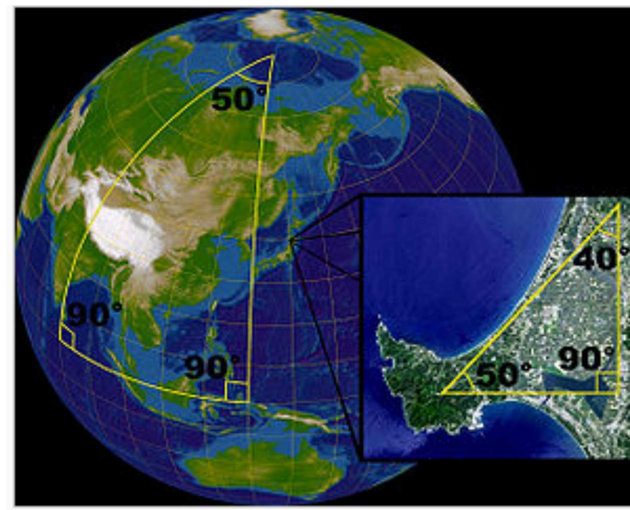
$$g \equiv \mathrm{Cov}\big(V_1(\boldsymbol{\theta};X), \ldots, V_k(\boldsymbol{\theta};X)\big).$$

This matrix or rank-two tensor is the *Fisher information metric.*

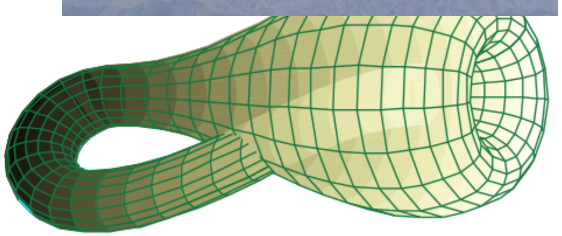
In abstract level, we view a set of all distributions as elements of a manifold (call **the statistical manifold**), then we put a Riemannian metric on it. One of the most famous Riemannian metric is the above Fisher information metric! *This Riemannian metric is similarly defined as before: each g_{ij} is a ``Statistic inner product version" of two tangent vectors V_i and V_j:* $g_{ij} \equiv -\mathbb{E}\left[\frac{\partial V_i}{\partial \theta_j}\right] = \mathbb{E}[V_i(\theta;X)\,V_j(\theta;X)]$,

Recall, what is a manifold?

- An n-dimensional manifold locally “looks like” a piece of $\mathbf{R}^n$.
- For examples, sphere and torus.
- Key features of a **manifold**: **curved**



The sphere (surface of a ball) is a two-dimensional manifold since it can be represented by a collection of two-dimensional maps.



- Only manifolds can capture UAV’s dynamical behaviors

2 The manifold of normal distributions

Recall that a normal distribution with mean μ and variance σ^2 is defined by the probability distribution function

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We can therefore view the space of normal distributions as a two-dimensional manifold $\mathcal{N}$, parameterized by μ and $\sigma > 0$.

Moreover, the Fisher information metric defines a natural metric on this space with $\theta_1 = \mu$ and $\theta_2 = \sigma$. We can imagine this as scaling each direction so that maximum likelihood estimation will generate identical likelihood plots.

Actually, we will take $\theta_1 = \mu$ and $\theta_2 = \sqrt{2}\sigma$ to simplify calculations. The probability distribution function becomes

$$p(x; \theta_1, \theta_2) = \frac{1}{\sqrt{\pi}\theta_2} \exp\left(-\frac{(x-\theta_1)^2}{\theta_2^2}\right).$$

Now we calculate the Fisher information metric from the expression

$$g_{ij} \equiv -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X; \boldsymbol{\theta})\right].$$

Using a program like Mathematica, we can calculate that

$$g_{11} = g_{12} = \frac{2}{\theta_2^2} \qquad g_{12} = g_{21} = 0.$$

We can express this concisely as

$$\mathrm{d}s^2 \equiv \frac{2\,\mathrm{d}\theta_1^2 + 2\,\mathrm{d}\theta_2^2}{\theta_2^2}.$$

Incidentally, this is a well-known situation in non-Euclidean geometry. The *Poincaré half-plane model* is the upper half-plane

$$\mathbb{H}^2 = \{(x, y) \in \mathbb{R}^2 \mid y > 0\},$$

with the metric

$$\mathrm{d}s^2 = \frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}.$$

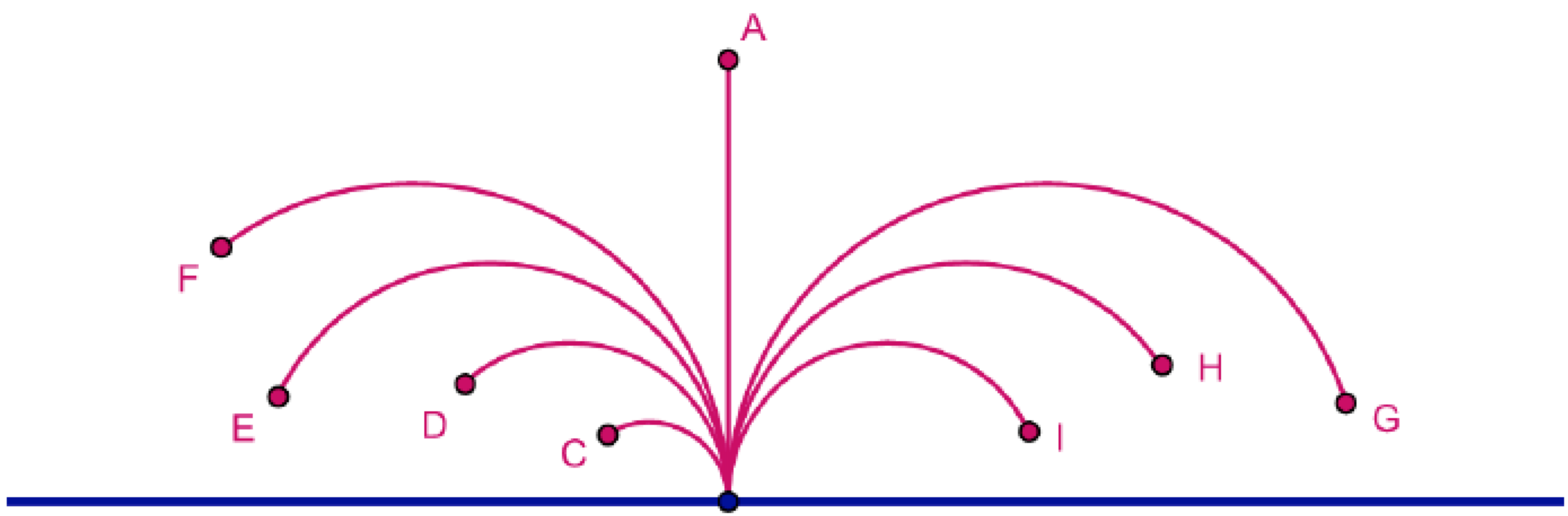


Figure 1: Geodesics in the Poincaré half-plane model.

Compare this to our situation:

$$\mathcal{N} = \{(\theta_1, \theta_2) \in \mathbb{R}^2 \mid \theta_2 > 0\}$$

with the metric

$$\mathrm{d}s^2 \equiv \frac{2\,\mathrm{d}\theta_1^2 + 2\,\mathrm{d}\theta_2^2}{\theta_2^2}.$$

Since a scaling by 2 of metric does not affect geodesics, the geodesics in $\mathcal{N}$, when parameterized by $\theta_1 = \mu$ and $\theta_2 = \sqrt{2}\sigma$, are the same as those of the half-plane model.

Backup slides

Poincare half-plane model

In non-Euclidean geometry, the **Poincaré half-plane model** is the upper half-plane, denoted below as **H** $\{(x, y) | y > 0; x, y \in \mathbb{R}\}$, together with a metric, the Poincaré metric, that makes it a model of two-dimensional hyperbolic geometry.

Equivalently the Poincaré half-plane model is sometimes described as a complex plane where the imaginary part (the *y* coordinate mentioned above) is positive.

Metric

The metric of the model on the half-plane, $\{\langle x, y\rangle | y > 0\}$, is:

$$(ds)^2 = \frac{(dx)^2 + (dy)^2}{y^2}$$

where *s* measures the length along a (possibly curved) line. The *straight lines* in the hyperbolic plane (geodesics for this metric tensor, i.e., curves which minimize the distance) are represented in this model by circular arcs perpendicular to the *x*-axis (half-circles whose origin is on the *x*-axis) and straight vertical rays perpendicular to the *x*-axis.

Distance calculation

In general, the *distance* between two points measured in this metric along such a geodesic is:

$$\begin{aligned}\operatorname{dist}(\langle x_1, y_1\rangle, \langle x_2, y_2\rangle) &= \operatorname{arcosh}\left(1 + \frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{2y_1 y_2}\right) \\ &= 2 \operatorname{arsinh} \frac{1}{2}\sqrt{\frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{y_1 y_2}} \\ &= 2 \ln \frac{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} + \sqrt{(x_2 - x_1)^2 + (y_2 + y_1)^2}}{2\sqrt{y_1 y_2}},\end{aligned}$$

where *arcosh* and *arsinh* are inverse hyperbolic functions

$$\begin{aligned}\operatorname{arsinh} x &= \ln\left(x + \sqrt{x^2 + 1}\right), \\ \operatorname{arcosh} x &= \ln\left(x + \sqrt{x^2 - 1}\right) \qquad x \geq 1.\end{aligned}$$

Distance for Special cases

Some special cases can be simplified:

$$\operatorname{dist}(\langle x, y_1 \rangle, \langle x, y_2 \rangle) = \left| \ln \frac{y_2}{y_1} \right| = |\ln(y_2) - \ln(y_1)|.$$ [1]

$$\operatorname{dist}(\langle x_1, y \rangle, \langle x_2, y \rangle) = \operatorname{arcosh}\left(1 + \frac{(x_2 - x_1)^2}{2y^2}\right) = 2 \operatorname{arsinh}\left(\frac{|x_2 - x_1|}{2y}\right)$$

$$\operatorname{dist}(\langle x, r \rangle, \langle x \pm r \sin\phi, r \cos\phi \rangle) = \operatorname{arsinh}(\tan\phi) = \operatorname{arcosh}\left(\frac{1}{\cos\phi}\right) = \ln\left(\frac{1 + \sin\phi}{\cos\phi}\right)$$