Dual Geometries in Statistics

Paul Vos
East Carolina University
vosp@mail.ecu.edu

29 July 2004

Goals

My Primary Goal: By describing some important statistical results from theoretical and applied statistics using ideas and concepts from dual geometries you will have some understanding of the important role these geometries play, and perhaps, motivate some of you to study these geometries further.

Another Goal: Introduce some basic terminology of (dual) differential geometry. [Terms]

Outline

Curved Exponential Family (CEF)

Consider a regular exponential family of order k ($y \in \mathcal{Y} \subset R^k$) having density

$$p(y \mid \eta) = \exp\{y'\eta - \psi(\eta)\}h(y)$$

A subset

$$\mathcal{M} = \{p : p(y \mid \eta(\theta)) \mid \theta \in \Theta\} \subset \mathcal{S} = \{p(y \mid \eta)\}$$

is a (one-parameter) curved exponential family (CEF) if-f the map $\theta \mapsto \eta(\theta)$ having domain $\Theta \subset R^1$ is an imbedding. [Manifold]

Relative Information Loss

$$\text{Limiting Relative Info Loss} = \lim I(\theta)^{-1}[nI(\theta) - I^T(\theta)]$$

Result: Let $\bar{y} = \frac{1}{n}(y_1 + \ldots + y_n) \in \mathcal{Y} \subset R^k$ have density in a CEF $\mathcal{M}$ and let $T = T(\bar{y})$ be a statistic. Then

$$\text{Rel. Info. Loss } (T \text{ for } \mathcal{M}) = \gamma^2 + \frac{1}{2}\beta^2$$

where

γ^2	is the statistical curvature of $\mathcal{M}$
β^2	is the curvature for T

Comments

$$\text{Rel. Info. Loss } (T \text{ for } \mathcal{M}) = \gamma^2 + \frac{1}{2}\beta^2$$

- Duality of γ^2 and β^2
- Parameter Invariance of γ^2 and β^2
- Efficient Notation (especially multivariate generalization)
- Geometric Intuition

Bib. Notes: Relative Information Loss

- Statistical geometry has been called Fisher-Efron-Amari theory.
- Fisher (1922, 1925)
- Efron (1975, 1978)
- Amari (1990)

Log Likelihood Function

Defn: Let $\mathcal{M}$ be a set of densities with common support $\mathcal{Y} \subset R^k$ and let $y \in \mathcal{Y}$. The log likelihood is the function $\ell_y : \mathcal{M} \mapsto R^1$ defined by $\ell_y(p) = \log(p(y))$.

Exp Fam: $\ell_y(p) = y'\eta - \psi(\eta) + \log(h(y))$

Statistical Importance:

- Fundamental inference procedure
- Generates superior inference procedures.

Maximum Likelihood

Defn: If $\hat{p} \in \mathcal{M}$ such that $\ell_y(\hat{p}) = \max_{p \in \mathcal{M}} \ell_y(p)$ then $\hat{p}$ is a maximum likelihood estimate (mle) of $\mathcal{M}$ (based on y).

Note 1: The mle is usually expressed in terms of some parameterization (co-ordinate system). Eg, $\hat{\theta} = \theta(\hat{p})$ or $\hat{\beta} = \beta(\hat{p})$.

Properties of $\hat{\theta}$ are expressed by considering $\hat{\theta}$ as a random variable. (Bias, variance, and exact or asymptotic distribution.)

Note 2: Many times the parameterization is not arbitrary (θ could be a mean or a slope).

Maximum Likelihood (cont)

Geometry: For CEF $\mathcal{M}$,

$\arg\max_{p\in\mathcal{M}} \ell_y(p) = \arg\min_{p\in\mathcal{M}} D(p_y, p)$

Maximum Likelihood = Minimum Divergence

Likelihood $\longleftrightarrow$ Divergence

Statistical Concepts $\longleftrightarrow$ Geometric Quantities

Divergence

Defn: Let Ξ be an open subset of R^k and let $\xi_1, \xi_2 \in \Xi$. A divergence on Ξ, denoted by $D_\Xi(\xi_1, \xi_2)$, is a smooth function taking values in $[0, \infty]$ such that for any $\xi_1, \xi_2 \in \Xi$ the following hold:

1. $D_\Xi(\xi_1, \xi_2) \geq 0$ with equality holding if-f $\xi_1 = \xi_2$.
2. $\frac{\partial}{\partial \xi_1^r} D_\Xi(\xi_1, \xi_2) \mid_{\xi_1 = \xi_2} = \frac{\partial}{\partial \xi_2^r} D_\Xi(\xi_1, \xi_2) \mid_{\xi_1 = \xi_2} = 0$
3. $g_{rs} = \frac{\partial^2}{\partial \xi_1^r \partial \xi_1^s} D_\Xi(\xi_1, \xi_2)$ is positive definite and a function of ξ_1 alone.

Divergence (cont)

Defn: For a space of points $\mathcal{S}$, $D : \mathcal{S} \times \mathcal{S} \mapsto R^1$ is a divergence on $\mathcal{S}$ if there exists a parameterization $f : \mathcal{S} \mapsto \Xi$ such that $D(f^{-1}(\xi_1), f^{-1}(\xi_2))$ is a divergence on Ξ.

Note 1: The parameterization exists when $\mathcal{S}$ is globally flat with respect to a nonmetric connection.

connection $\leftrightarrow$ geometry [Connection] [Metric]

metric connection $\leftrightarrow$ "straight lines" minimize distance

flat $\leftrightarrow$ "nice geometry" (R^n)

Note 2: $\mathcal{S}$ globally flat in another nonmetric connection.

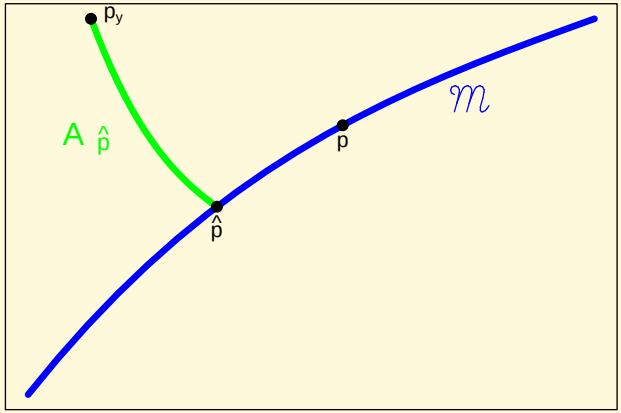
Note 3: Dual connections: 1-connection -1-connection.

Divergence (Dual) Geometry

- $D(p_1, p_2)$ behaves like a squared distance
- D is not symmetric
- Two Geometries used to describe D
- Pythagorean Relationship:

$$D(p_y, p) = D(p_y, \hat{p}) + D(\hat{p}, p)$$

Divergence Geometry (cont)



Law of Cosines

Triangle: having sides of length a, b, c and angle C opposite side of length c.

$$c^2 = a^2 + b^2 - 2ab\cos C$$

Squared Distances in R^k: $p_1, p_2, p_3 \in R^k$, so

$$\begin{aligned} c^2 &= \| p_1 - p_3 \|^2 \\ a^2 &= \| p_1 - p_2 \|^2 \\ b^2 &= \| p_2 - p_3 \|^2 \\ C &= \text{angle between } p_3 - p_2 \text{ and } p_1 - p_2 \end{aligned}$$

$$\begin{aligned} \| p_1 - p_3 \|^2 &= \| p_1 - p_2 \|^2 + \| p_2 - p_3 \|^2 \\ &\quad -2 \| p_1 - p_2 \| \cdot \| p_2 - p_3 \| \cos\theta \\ &= \| p_1 - p_2 \|^2 + \| p_2 - p_3 \|^2 - 2\langle p_1 - p_2, p_2 - p_3 \rangle \end{aligned}$$

Setting $\| p - q \|^2 = 2D(p, q)$ and $p - q = v(q, p)$, we have

$$\begin{aligned} 2D(p_1, p_3) &= 2D(p_1, p_2) + 2D(p_2, p_3) - 2\langle v(p_2, p_1), v(p_2, p_3) \rangle \\ \text{iff } D(p_1, p_3) &= D(p_1, p_2) + D(p_2, p_3) - \langle v(p_2, p_1), v(p_2, p_3) \rangle \end{aligned}$$

Divergences: For points $p_1, p_2, p_3 \in \mathcal{S}$,

$$D(p_1, p_3) = D(p_1, p_2) + D(p_2, p_3) - \langle v(p_2, p_1), v^*(p_2, p_3) \rangle_{p_2}$$

Divergence in Statistics

- Log Likelihood Function
  (properties of estimators)
- Quasi-likelihood Function
  (weaken distributional assumptions)
- Kullback-Leibler Information (Distance) (Divergence for Exp Fam, approximating Exp Fam)

Sufficiency

Defn: Let $\mathcal{M}$ be a family densities and let Y be an observation from $p \in \mathcal{M}$. A statistic $T(Y)$ is sufficient for $\mathcal{M}$ if the conditional distribution of Y given $T = t$ is the same for all $p \in \mathcal{M}$.

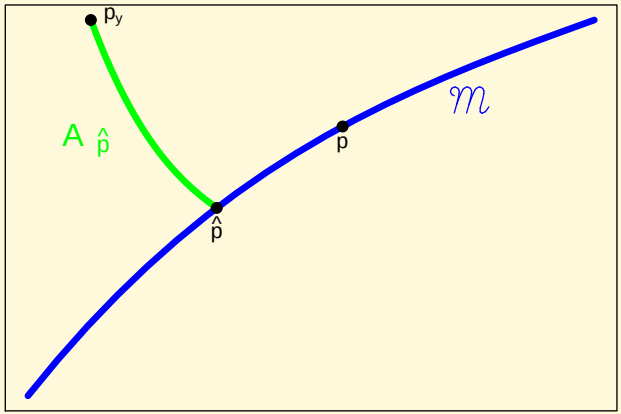
Factorization Theorem: T is sufficient for $\mathcal{M}$ if-f there exists functions u_1 and u_2 such that

$$\log(p(y)) = u_1(y) + u_2(p, T(y))$$

Note 1: For CEF $\mathcal{M}$, $\log(p(y)) = C(y) - D(p_y, p)$.

Note 2: When the Pythagorean Relationship holds $\log(p(y)) = C(y) - D(p_y, \hat{p}) - D(\hat{p}, p)$.

Sufficiency (cont)



$$\begin{aligned}\log(p(y)) &= C(y) - D(p_y, \hat{p}) - D(\hat{p}, p) \\ &= u_1(y) + u_2(p, T(y)), T(y) = \hat{p}\end{aligned}$$

Sufficiency (cont)

- Pythagorean Relationship requires 3 conditions
  - 1-geodesic connecting $\hat{p}$ to p
  - -1-geodesic connecting p_y to $\hat{p}$
  - Right angle between these geodesics
- Geometry of Maximum Likelihood Estimation
- Geometry of Exponential Families
- Information Loss and Recovery

Bib. Notes: MLE and Sufficiency

- Ideas related to MLE (divergence): contrast functions (Eguchi, 1983), Csiszar's (1967) f-divergence; minimum divergence estimation (Vos, 1997; Kass and Vos, 1997).
- Ideas Related to sufficiency: information recovery (Efron and Hinkley, 1978; Amari, 1990; Skovgaard, 1985) and higher order efficiency (Eguchi, 1983, 1985; Amari, 1990).
- Global Information Recovery: Marriott and Vos, 2004.

Global Information Recovery

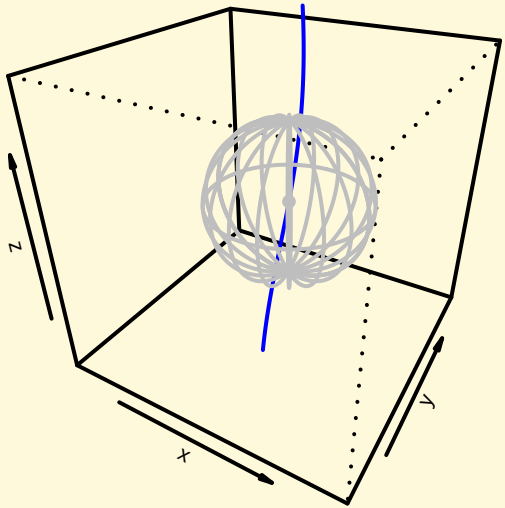
Helix Model: $X \sim N(\mu, I_{3\times 3})$ where $\mu = (r\cos\theta, r\sin\theta, \theta d)'$, r and d are fixed.

Information in a statistic T**:** Ability to approximate the likelihood function using a function that depends on X only through T. **Information Recovery:** $T = (\hat{\theta}, A)$, find A.

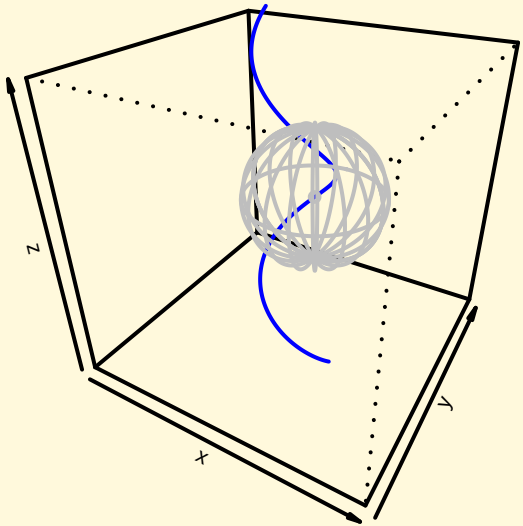
Principles of (Local) Information Recovery

- **Minimize Expected Info Loss:** minimize $I^X - I^T$
- **Observed Fisher Information:** $A = I_{obs}$
- **Ancillarity:** Choose A having distn not a function of θ

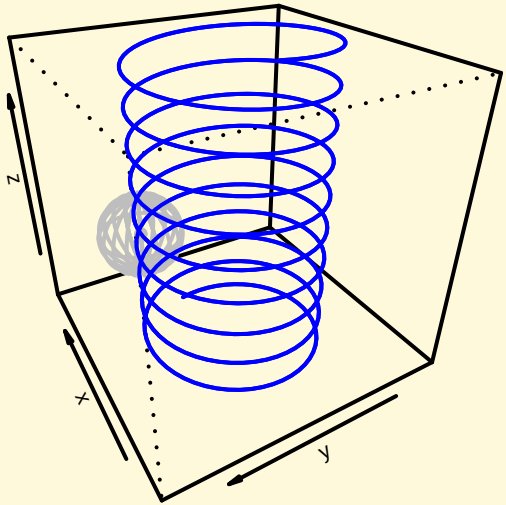
Helix Model: r = 0.5 , d = 3



Helix Model: r = 1 , d = 1



Helix Model: r = 5 , d = 0.3

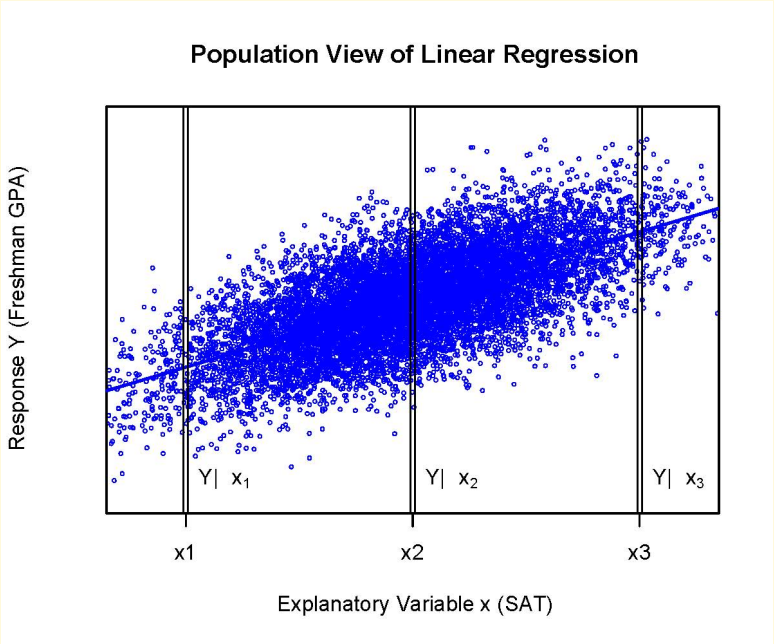


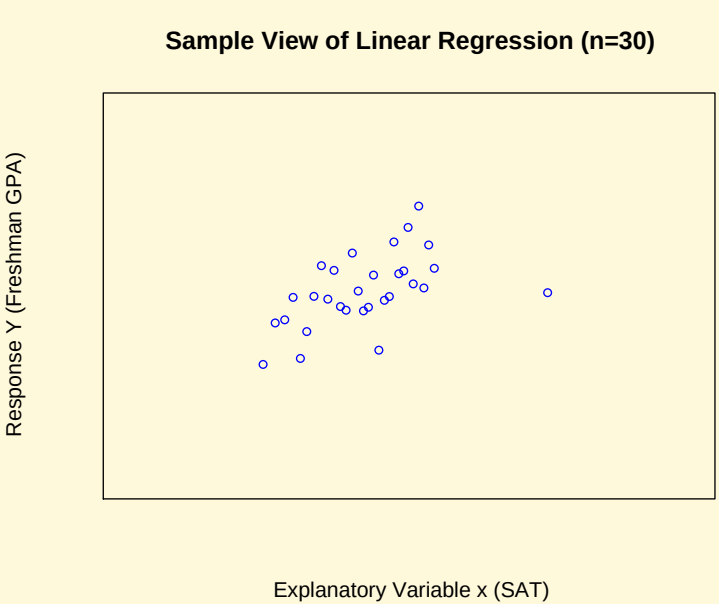
Remarks on Helix Models/Info Recovery

- Local methods of information recovery can fail.
- Global method (spectral decomposition) does not fail: indicates number of required "information directions".
- Explains why local methods (asymptotics) work.

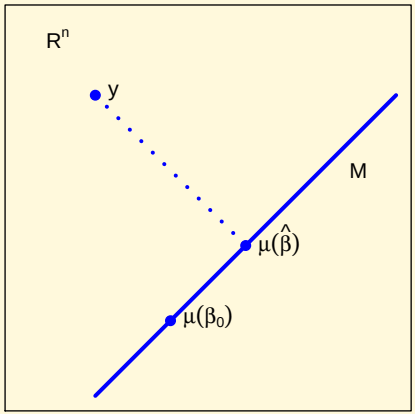
Generalized Linear Models

- Model the relationship between a response variable Y and one or more explanatory variables $x_1, \ldots, x_k$.
- Example Y is college freshman GPA and x is SAT score.
- Model $Y \mid x$; see graph.
  - Mean $(Y \mid x) = \mu(x) = \beta_1 x + \beta_0$
  - Var $(Y \mid x) = \sigma^2(\mu(x))$
  - Shape $(Y \mid x)$: exponential family or not specified

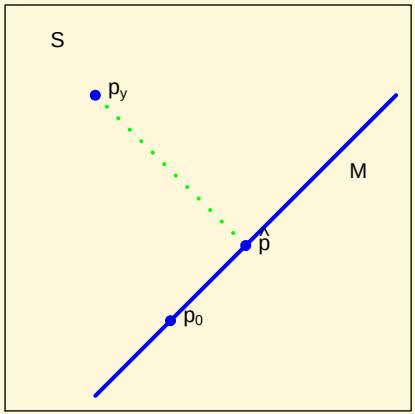

Population View of Linear Regression
Response Y (Freshman GPA)
Y| x1
Y| x2
Y| x3
x1
x2
x3
Explanatory Variable x (SAT)


Sample View of Linear Regression (n=30)
Response Y (Freshman GPA)
Explanatory Variable x (SAT)

Model View of Linear Regression



Model View of a GLM



Geometry of GLMs

- Structure
  - Dual Geometries
  - $\mathcal{S}$ is of dimension n
  - $\mathcal{M} \subset \mathcal{S}$ has dimension $k+1$
  - $\mathcal{M}$ is often flat in $+1$ connection
  - Riemannian metric often diagonal in divergence parameter
- Role
  - Estimation Algorithm
  - Model checking

Summary

- Parameter Invariance
- Economical Notation
- Intuition
- Importance of Pythagorean Relationship

Geometry Concepts Summary

Terminology	Concept
manifold	set endowed with geometric structure
metric	orthogonality
angle	departures from ortho.
α-connection	straight lines
α-curvature	departures from strght. lines
divergence	"distance" of interest
Pythagorean Relationship (Law of Cosines)	Relationship among divergence, strght lines, and ortho. (angle)

References

Amari, S.I. (1990) *Differential-Geometrical Methods in Statistics*, Springer, New York.

Csiszar, I. (1967) On topological properties of f-divergence. *Studia Sci. Math. Hungar.*, **2**, 329-339.

Efron, B. (1975) Defining the curvature of a statistical problem, *Ann. Statist.*, **3**, 1189-1217.

Efron, B. (1978) The geometry of exponential families, *Ann. Statist.*, **6**, 362-376.

Efron, B., and Hinkley, D.V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**, 457-481.

Eguchi, S. (1983) Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, **11**, 793-803.

Eguchi, S. (1985) A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.*, **15**, 341-391.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. Royal Soc. London, Ser. A*, **222** , 309-368.

Fisher, R.A. (1925) Theory of statistical estimation. *Proc. Camb. Phil Soc.*, **22**, 700-725.

Kass, R.E, and Vos, P.W. (1997) *Geometrical Foundations of Asymptotic Inference*, John Wiley & Sons, New York.

Marriott, P. and Vos, P.W. (2004) On the global geometry of parametric models and information recovery, *Bernoulli*, **10** (2), 1-11.

Skovgaard, I.M. (1985) A second-order investigation of asymptotic ancillarity. *Ann. Statist.*, **13**, 534-551.

Vos, P.W. (1997) Minimum divergence estimation. In *Encyclopedia of Statistical Sciences Update II* (ed. S. Kotz and C. Read) Wiley, New York.

Manifolds

Consider a two binomial problem with n_1 and n_2 fixed. There is a **set** of densities:

$$\mathcal{S} = \Big\{ f(x_1, x_2) : f = \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1 - x_1} \times \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2 - x_2},\ p_1, p_2 \in (0,1) \Big\}$$

In fact, $\mathcal{S}$ can be given additional structure: when (p_1, p_2) is close to (p_1', p_2') the models (densities) are not too different – statistically, a **family**; mathematically a **topological manifold**.

NOTE: There are other parameterizations besides $p = (p_1, p_2)'$, but the properties of closeness still hold.

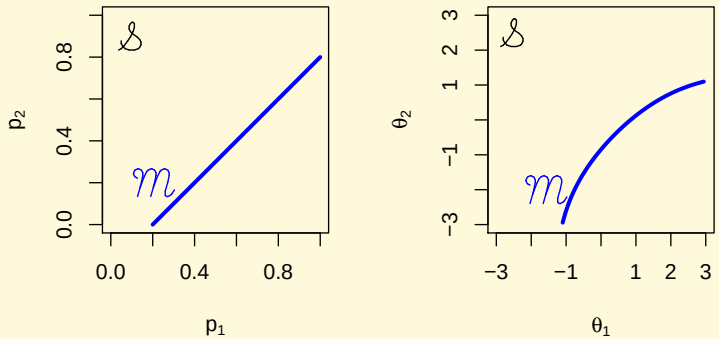
- The hypothesis $H_0 : p_1 - p_2 = .2$ defines a subfamily (imbedded submanifold)

$$\mathcal{M} = \Big\{ f \in \mathcal{S} : p_1 - p_2 = .2 \Big\}$$

- The paremeterization $p : \mathcal{M} \mapsto R^2$ defines an image, $p(\mathcal{M})$, in the real plane. ($p(\mathcal{S})$ is the open unit square.)

- Another parameteriation, say the log odds $\theta = \log(p/(1-p))$, defines another image, $\theta(\mathcal{M})$. ($\theta(\mathcal{S})$ is R^2.)

Two Visualizations



Comments on Visualizations

- Different parameterizations provide different pictures – different geometries – different connections.

  $p(\mathcal{M})$ is a straight line (geodesic); $\theta(\mathcal{M})$ is not.

- There is no one *correct* geometry.

  (straight line $\neq$ curve of shortest length)

- Geometries judged on their utility.

  For studying sufficiency, should $\mathcal{M}$ be considered a straight or curved line?

- A statistical manifold is a topological manifold (parameterization into the reals) endowed with the geometric properties of
  - straight lines – departures measured by curvature
  - orthogonality – departures measured by angle
  - divergence (replaces the notion of length)

RETURN

Connections

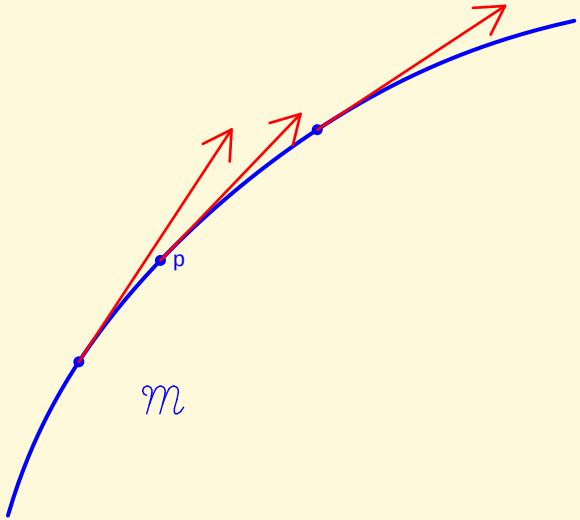
- **Concept:** straight lines are ones whose tangent vector does not change direction (or length).

  In order compare tangent vectors along a curve, there needs to be a specified mapping or connection between tangent spaces – this is provided by the connection.

- **Definition:** based on derivatives of the log likelihood or divergence functions.

- **Notation:** $\nabla^{\alpha}_{X_p} Y$, $\Gamma^{(\alpha)i}_{jk}$; $\nabla^{\alpha}_{X} Y(p)$, $\Gamma^{(\alpha)i}_{jk}(p)$

  $\nabla^{\alpha}_{\partial_j} \partial_k = \sum_i \Gamma^{(\alpha)i}_{jk} \partial_i = \Gamma^{(\alpha)i}_{jk} \partial_i$ where $\partial_i = \frac{\partial}{\partial \xi^i}$


p
$\mathcal{M}$

Metric

- **Concept:** Metric defines an inner product which in turn provides the idea of orthogonal vectors (and curves). Also gives lengths of vectors.

  Metric is actually a (smooth) collection of inner products – depending on p – that gives lengths of curves.

  However, we do not use an α family of metrics.

- **Definition:** based on derivatives of the log likelihood (Fisher information) or divergence functions.

- **Notation:** $\langle X_p, Y_p \rangle_p$, $\langle X, Y \rangle$, g_{ij}, g^{ij}; $g_{ij}(p)$, $g^{ij}(p)$

  $\langle \partial_i, \partial_j \rangle = g_{ij}$ where $\partial_i = \frac{\partial}{\partial \xi^i}$

  $\langle \partial^i, \partial^j \rangle = g^{ij}$ where $\partial^i = \frac{\partial}{\partial \xi_i^*}$

  Or, matrix $[g^{ij}]$ is the matrix inverse of $[g_{ij}]$ – there is no ambiguity because of the properties of the dual parameter ξ^* if it exists.