Ezra Miller : stats with samples on stratified spaces (Ezra 1)

stat methods: mean, variance, PCA, CLN, CLT, ...

Def: manifold - 2nd countable, Hausdorff topological space s.t. every pt has a neighborhood homeomorphic to an open ball in $\mathbb{R}^n$.

(2nd countable = M is covered by a finite # balls)

chart: $\pi : U \in \mathbb{R}^n \xrightarrow[\text{injection}]{} M$.

$M_{\alpha\beta}$ transition for chart α & β

Category: topological manifold
smooth / C^∞ "
analytic "
algebraic variety

Def: A topologically stratified space is a Hausdorff topol space X that is a disjoint union of manifolds $M_1 \cup M_2 \cup \dots M_\ell$ s.t.

1/ $M_1 \cup \dots \cup M_k$ is a closed subset $\forall k \leq \ell$

2/ $\forall x, y \in M_i$ (stratum), there exist a homeomorphism $\varphi : X \to X$ with $\varphi(M_k) = M_k \ \forall k$ and $\varphi(x) = y$
stratum preserving (transitive action

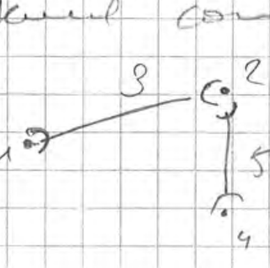
3/ Each stratum has finitely many connected components.

Examples

dim 0: finite discrete set of pts (each pt is closed).

stratification: as one disconnected manifold or as several connected.

dim 1: graphs. 2 pts + one edge,



or trees

Trees = connected graphs with no cycles. Egpa (2)

Spider or stars:

Dim 2+:

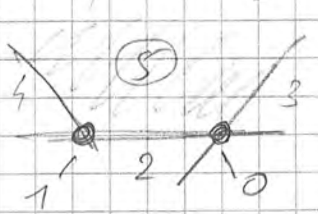
open book = spider $\times \mathbb{R}^d$

spine $\cup$ open pages - no

GM example

dim 0: 1
dim 1: open circle.
dim 2: torus + disk

Polyhedron: intersection of finitely many closed half spaces in $\mathbb{R}^d$.



Whitney cusp

dim 0: 1 Sing(X) = |

not equisingular

dim 1: 2 (or 1 disconnected) dim 2: 4 (or less if disconnected)

Shape spaces A = $d \times n$ matrix (n labeled landmarks in $\mathbb{R}^d$).

$A \sim gA$ $g \in$ Transf Group G.

$\Rightarrow$ $G = SE(d)$, similarity, projective

$X = G \backslash \mathbb{R}^d$

mod out the left = algebraic variety

Ezra (3)

⚠ Remove unstable orbits to avoid having non-Hausdorff quotients. (free action is needed).

Ex Phylogenetic tree is a tree with labeled leaves (vertices of deg 1) an edge length ≥ 0

eg 

removing an edge creates 2 disconnected comp: an edge can be identified to the set of leaves on both sides.

internal edges: $\mathbb{R}^2_{>0}$

→ a tree is a union of positive orthants.

[Billera - Holmes - Vogtmann 1998].

- facets of T_n are orthants $\mathbb{R}^{n-2}_{>0}$ (binary tree topologies).
- faces are orthants $\mathbb{R}^d_{>0}$, $d \leq n-2$

tree $T \longleftrightarrow z_T \in \mathbb{R}^{edge(T)}_{>0}$ ← splits $= O_T$

$O_T \subseteq \overline{O_{T'}}$ ⟺ T is a contraction of T'

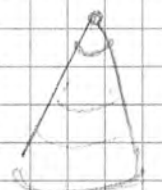
$\overline{O_T} = \overline{O_{T'}} \cap \overline{O_{T''}}$ ⟺ T is the biggest common contraction

Epa 4

PSD matrices

stratification by subsets of equal eigenvalues

Cone:



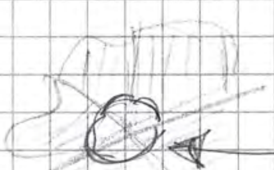
homeomorphic to $\mathbb{R}^2$

and locally isometric to except at the apex.

Volume of the sphere at the vertex is $2\pi r$ in the plane

but only $\alpha r < 2\pi r$ at the vertex of the cone

$\alpha < 2\pi$: positive curvature.



circle length $5\pi/2 \cdot r$

$\alpha > 2\pi$ negatively curved (Kale).

$Kale_{\alpha = 5\pi/2} \subset T_4$

Thm: Tubular neighborhood Thm

X stratified $\Longrightarrow$ a tubular neighborhood N of each stratum S is a fiber bundle over S whose fiber N_S is homeomorphic to a cone over a stratified space L with $\dim L = \dim N - 1$. N_pX

Tg cone $T_pX = \underbrace{T_pS}_{\text{tg space of } S} \times \underbrace{N_pX}_{\text{Normal slice to } X\text{, i.e. going to another stratum}}$

Ezia 5

open book

$T_pX = \mathbb{R} \times$ spider

Def: $\dim(X) = \max_\alpha (\dim \Pi_\alpha)$.

for $p \in S$ stratum of dim $d-1$, where $d = \dim X$.

$$\underset{\mathbb{R}^{d-1}}{T_pX = T_pS} \times \underset{\text{spider dim}=1}{N_p\cancel{X}} \quad \text{with } N_pX = \text{cone over the finite set} = \text{spider}$$

Thm: if Π_i and Π_j are strata then $\Pi_i \cap \overline{\Pi}_j \neq \emptyset \Rightarrow \Pi_i \subseteq \overline{\Pi}_j$

Corollary: the strata are partially ordered.

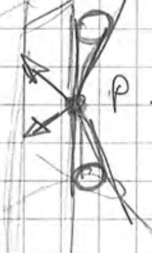
$$\Pi_i \leq \Pi_j \Leftrightarrow \Pi_i \subseteq \overline{\Pi}_j$$

Rem: the open book is universal in the sense that every co-dim ~~one~~ 1 singularity of a stratified space is an open book.

eg for the Whitney cusp

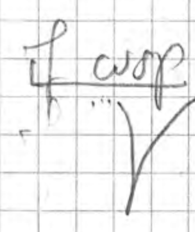
T_pS N_pX

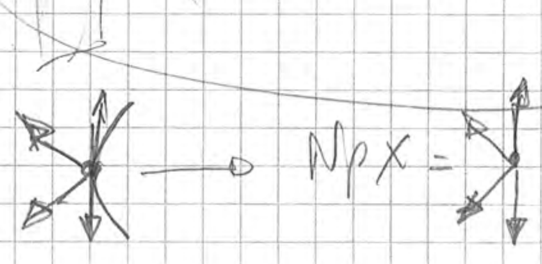
$\dim(S) = 2-1 = 1$.

if discontinuous  $p \Rightarrow T_pS = 0$

$N_pX = T_pX =$ 

if cusp

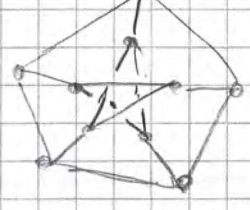


$\longrightarrow N_pX =$ 

tangent space for the tree space T_4

T_pX at a lane =

$T_0\overline{T_4} = N_0T_4$ = cone over petersen graph



Metrics, smoothness, etc

geodesic metric

a Riemannian stratified space should require at least:
a topologically stratified metric space such that each stratum is Riemannian.

sheaf of smooth functions: [book by Pflaum].
→ link with the manifold with corners and germs of manifolds of Peter Michor.

Exercise: embed T_3 into the Kale K_α for $\alpha > 3\pi$

discrete hyperbolic saddle

Thm: Whitney stratified spaces (embedded in an Eucl sp)
a topologically stratified space is a WSS if limits of secant lines joining strata R_i and R_j with $R_i \subseteq \overline{R_j}$ is included in the limit of tg planes to R_j as points cv to R_i.

Examples: real or complex algebraic
(or semi-analytic, or sub analytic)

Thm: Whitney stratified $\Rightarrow$ triangulable

X is a metric stratified space.

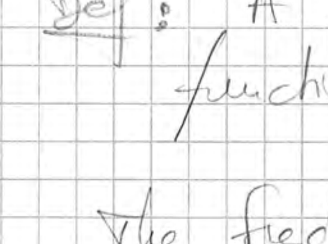
Def: A probability distrib. p on X has a Frechet function $F(y) = \int_X d^2(y, x)\, p(dx)$

The Frechet mean of p is $\operatorname{argmin}_{y \in X} F(y)$.

What can go wrong:

wiggle point $\Rightarrow$ ~~wiggle mean~~ the mean can stick

Example: 3 spider



$a = b = c = 1 \Rightarrow$ mean is 0.

$F(\mu) = a^2 + b^2 + c^2$

$f(\mu_\varepsilon) = (a-\varepsilon)^2 + (c+\varepsilon)^2 + (b+\varepsilon)^2$

$= a^2 + b^2 + c^2 + 2(c+b-a)\varepsilon + 3\varepsilon^2$

$F(\mu_\varepsilon) - F(\mu) = 2\varepsilon\underbrace{(c+b-a)}_{} + 3\varepsilon^2$.

positive for $a < b + c$ for small ε.

Thus the means sticks at 0 for small ε.

On a general stratum, the mean wiggles along the Tg space but may be sticky along the N/pt

Epra (8)

<u>LLN</u>: $x_1, x_2, \ldots$ indep random variables
with value in X distrib according to p.

then $\overline{x}_n \longrightarrow \mu(p)$. as $n \to +\infty$.

fix X = open-book = $S \times$ spider, where $S = \mathbb{R}^d$
given a proba p, there are 3 possibilities:

1/ $\mu(p) \notin S$ (it is in a page).

2/ $\mu(p) \in S$ and the same is true for $\mu(p')$ for
~~all~~ p' near p.

3/ $\mu(p) \in S$ but $\exists p'$ arbitrarily close to p such
that $\mu(p') \notin S$

<u>Rem</u> the open-book X is $CAT(0)$ so the Frechet mean
is unique.

<u>Def</u> for these three cases:

1/ mean is non-sticky

2/ mean is ~~non~~ sticky

3/ mean is partly sticky.
(the definition implies that there exist directions
in which it stays sticky).

Egra (9)

Thm [BATTBSI WG].

If x_1, x_2 are iid in X distrib according to p with supp(p) in at least 3 pages.
then $\bar{x}_n \xrightarrow[n\to+\infty]{} \mu(p)$ (std LLN)

Moreover: (•) if $\mu(p)$ is non-sticky or partly sticky (case 1 or 3), then there is a page L and a random integer N s.t. $\bar{x}_n \in L$ $\forall n \geq N$ almost surely

(•) $\mu(p)$ sticky $\Rightarrow$ $\exists$ N random such that $\bar{x}_n \in S$ $\forall n \geq N$ a.s.

CLT: $x_1 \dots x_n$ iid $\to$ $\bar{x}_n$ is a random variable

$\bar{x}_n \to \delta_{\mu(p)}$ and $\sqrt{n}(\bar{x}_n - \mu(p)) \to \tilde{p}$,

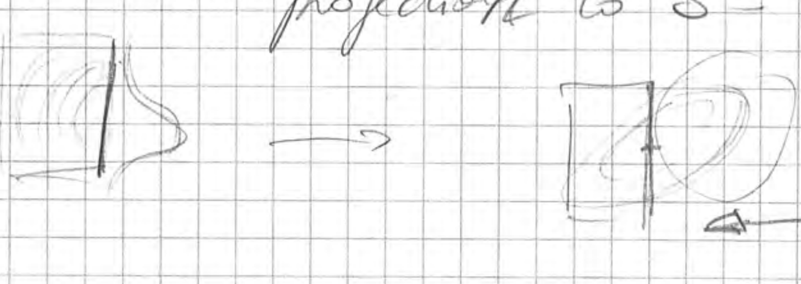
where $\tilde{p}$ is a limiting distribution (classically Gaussian).

Thm [BATTBSI WBG] A CLT for square integrable p holds with limiting distrib:

1/ non-sticky mean: Gaussian on $\mathbb{R}^{d+1} \simeq T_{\mu(p)}L = T_{\mu(p)}X$

2/ sticky case: Gaussian on $\mathbb{R}^d = S = T_p S$

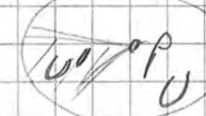
3/ partly sticky: Gaussian on L + its reflective projection to S



Symetric part projected on the sphere

Def Fix a metric space X — Ezra (10)

and a topologized set of $\mathcal{P}$ of proba distrib on X.
The mean $\mu(p)$ of $p \in \mathcal{P}$ <u>sticks</u> to a subset $K \subseteq X$
if every neighborhood U of $p \in \mathcal{P}$ contains a non-empty open set $U' \subseteq U$ with $\mu(p') \subseteq K$
for all $p' \in U'$



[partly sticky + sticky case].

<u>Example</u> [Huckemann, Mattingly, Nolan, Miller].

the kale (isolated planar hyperbolic singularity)

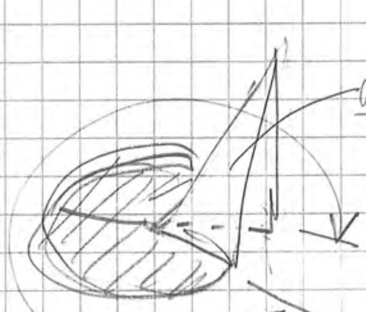
Kale with $\alpha > 2\pi$

projection on the sections.

Gaussian = visible part

less than π

the shadows of the kale



less than π.

or combinations thereof / "medical networks"?

02/05/19

Ezra Miller:

- Goal: statistics where sample space is a stratified space X

  e.g. trees, shapes, persistence diagrams (inf. dimensional ⇝ not covered), diffusion tensors

- ordinary statistics: object = vector, sample space = vector space
  methods: mean, variance, PCA, LLN, CLT
- stratified stat.: all of these ⇝ raise fundamental problems in geometric probability
- Def. A manifold is a second-countable, Hausdorff topological space M s.t. every point $x \in M$ has a neighborhood ≅ open ball in $\mathbb{R}^n$.
- chart: $\pi_\alpha : U_\alpha \hookrightarrow M$, $M_{\alpha\beta} = \pi_\alpha(U_\alpha) \cap \pi_\beta(U_\beta)$
  open ball in $\mathbb{R}^d$

  $\pi_\beta : U_\beta$

  ⇝ $\pi_{\alpha\beta} : \pi_\alpha^{-1}(M_{\alpha\beta}) \to \pi_\beta^{-1}(M_{\alpha\beta})$, $\pi_{\alpha\beta}(x) = \pi_\beta^{-1} \circ \pi_\alpha(x)$

  $\pi_{\alpha\beta}$ s topological / smooth / analytic / algebraic ⇝ topological / smooth / analytical manifold / algebraic variety

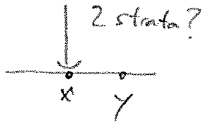
  $\{\text{all } \pi_\alpha\}$ = atlas

- data examples: angles (circle), rotations ($SO(3)$), lines ($\mathbb{RP}^{n-1}$), subspaces ($Gr_{k,n}$)

- stratified spaces: **Def.** A <u>topologically</u> <u>stratified</u> <u>space</u> is a Hausdorff topological space X that is a disjoint union $X = M_1 \cup M_2 \cup \dots \cup M_\ell$ of manifolds called <u>strata</u> s.t.
  0) each stratum has finitely many connected components
  1) $M_1 \cup \dots \cup M_k$ is closed in X $\forall k \le \ell$
  2) $\forall x, y \in M_i$ $\exists$ homeomorphism $\varphi : X \to X$ with
     - $\varphi(M_k) = M_k$ $\forall k$
     - $\varphi(x) = \varphi(y)$

  ⇝ points "look the same" on the same stratum

  ⇝ not:



- examples:

  dim 0: finite set with discrete topology

  ⇝ 1 disconnected stratum or lots of connected strata or ...

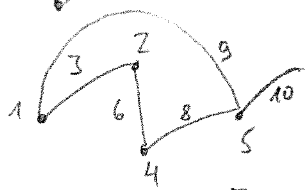
  ⇝ same space but different stratifications

  dim 1: graphs ⇝ first vertices (condition 1) then edges

  e.g.

not

(but without 3 ok)



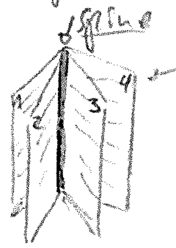
e.g. trees: connected, no cycles

•7 spiders, or star:

⇝ no vertices at the end

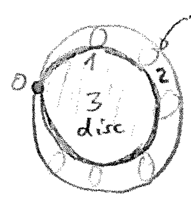
higher dim.: e.g. <u>open book</u> = spider $\times \mathbb{R}^d$, $d = 0$: spider

$d = 1$:



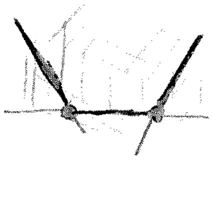
= spine ∪ open pages
M_0 $M_1, \dots, M_\ell$

GM example:
Goretsky & McPherson (?) ⇝ book



dim.	0	1	2
#	1	0	2

polyhedron: intersection of finitely many half-spaces in $\mathbb{R}^d$

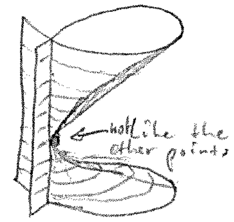


dim	0	1	2
#	2	3	1

Geometric Statistics, Toulouse

2

Whitney cusp:



sing(X) = |

but $X = \text{sing}(X) \cup (X \setminus \text{sing}(X))$

is not a (top.) stratification

shape spaces, eg. face recognition $(d=3)$

$$\begin{array}{cc} & n \text{ (landmarks, labelled)} \\ d & \left[\begin{array}{ccccc} | & | & | & | & | \end{array}\right] \end{array} = A \in \mathbb{R}^{d \times n}$$

, A' same shape if $A' = gA$,

$g \in G$ = group of transformations, e.g. G = rigid motions

or G could have rotations, translations, projective transformations, reflections

$\rightsquigarrow$ $X = G \backslash \mathbb{R}^{d \times n}$ $\rightsquigarrow$ (typically) algebraic variety $\rightsquigarrow$ stratified (see below)

after removing unstable orbits (keep stable and semi-stable orbits)

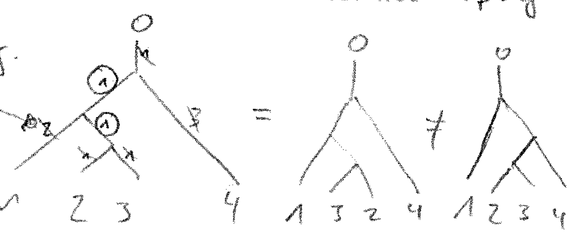
e.g. $\Sigma^n_d = \mathbb{R}^{d \times n} \setminus \{\text{all columns equal}\} / \text{similarities}$

$d = 2$: $\Sigma^n_2 \cong \mathbb{CP}^{n-2}$, else $(d>2)$: "non-free shapes" $\rightsquigarrow$ lower dim strata

Def: A phylogenetic n-tree is a tree with $n+1$ labelled leaves (vertices of degree 1) and edge lengths >0 on internal edges (no pendant edges)

$\mathcal{T}_n = \{n\text{-trees}\}$

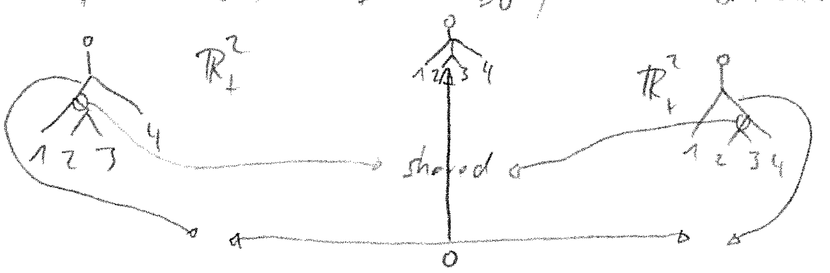
E.g.



$\rightsquigarrow$ removing edge gives partition of leave set into two connected components

$\rightsquigarrow$ fixed "combinatorics": specify internal edge lengths, e.g. $\mathbb{R}^2_+$

$(\mathbb{R}_+ = \mathbb{R}_{>0}, \ \overline{\mathbb{R}}_+ = \mathbb{R}_{\geq 0})$ $\rightsquigarrow$ orthants $\mathbb{R}^2_+$ are manifolds



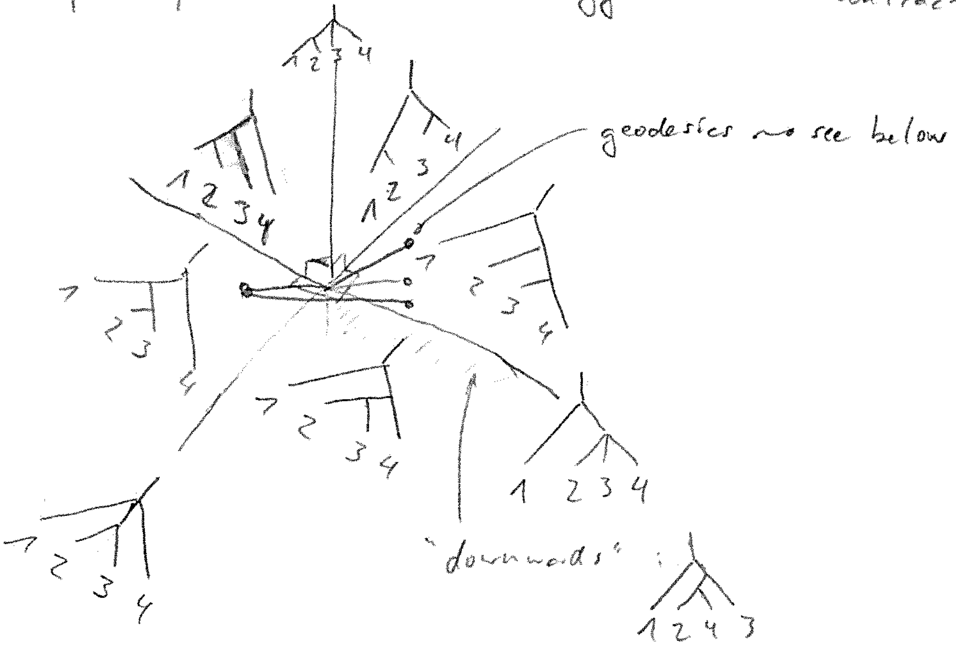
(Billera-Holmes-Vogtmann 1998)

facets of $\mathcal{T}_n$: orthants $\overline{\mathbb{R}}^{n-2}_+$ in bijection with (indexed by) binary tree topologies

faces of $\mathcal{T}_n$: orthants $\overline{\mathbb{R}}^d_+$ for $d \leq n-2$

tree T: $x_T \in \mathbb{R}^{e(T)}_+$ ← edges $\cong$ splits, O_T $\rightsquigarrow$ strata, $\overline{O_T} \subseteq \overline{O_{T'}} \iff T$ is a contraction of edges of T'

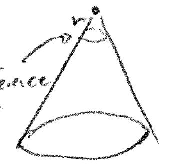
$\bar{O}_T = \bar{O}_{T'} \cap \bar{O}_{T''} \iff T$ is the biggest common contraction



diffusion tensors: positive semidefinite matrices of size 3

stratification by subsets of eigenvalues that are equal

ice-cream cone:

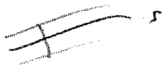


homeom. $\cong \mathbb{R}^2$

Local isometry away from apex

$\alpha > 2\pi$, e.g. $\underset{4}{5}\frac{\pi}{2}$ ~ negatively curved: <u>Kale</u>

contained in $\mathcal{T}_4$, see above

- <u>Tubular neighbourhood Theorem</u>: X is topologically stratified
  $\Rightarrow$ a tubular neighbourhood of each stratum S is a fiber bundle over S (i.e.  it is locally a cross-product of S and N_S) with fiber N_S where N_S is homeomorphic to a cone over a stratified space L with $\dim L = \dim N_S - 1$.

  ~ tangent cone at $p \in S$: $T_pX = \underbrace{T_pS}_{\text{tangent space}} \times \underbrace{N_p\cancel{X}}_{\text{normal slice} \cong N_S}$

  manifold ← a copy of $\mathbb{R}^{\dim S}$

  even more: the tangent bundle is locally trivial

e.g. open book: on pages: N_pS = one point = cone over $\emptyset$

on spine: N_pS = spider = cone over ℓ points

- Def. $\dim X = \max_k (\dim M_k)$
- Suppose $p \in S$ stratum of dim. $d-1$ where $d = \dim X$.

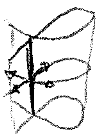
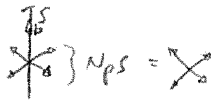
  $N_p X = ?$ $T_p X = ?$ $\quad T_p X = \underbrace{T_p S}_{\dim: d-1} \times \underbrace{N_p X}_{1}$

Thm. If M_i and M_j are strata, $M_i \cap \overline{M}_j \neq \emptyset \Rightarrow M_i \subseteq \overline{M}_j$

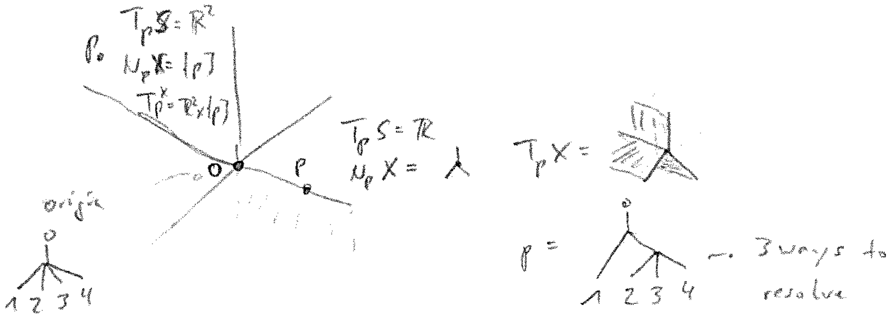
Cor. Strata are partially ordered: $M_i \leq M_j \Leftrightarrow M_i \subseteq \overline{M}_j$

$p \in \overline{S'} \Rightarrow \overline{S'} \geq S \rightsquigarrow N_p X$ is cone over a finite set (conn. components of intersection of tub. neigh. with d-dim strata containing S) $\rightsquigarrow$ spider

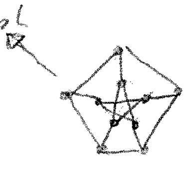
$\rightsquigarrow T_p X$: d-dim. open book $\rightsquigarrow$ universal for points in co-dim. 1 strata

e.g.  

therefore at the Whitney cusp the point must be its own stratum

e.g. $\mathcal{T}_4$ 

$T_0 \mathcal{T}_4 = N_0 \mathcal{T}_4 =$ cone over Peterson graph



- Metrics, smoothness etc.:

  "Riemannian stratified space" when properly defined should require top. strat. geodesic* metric space s.t. each stratum is Riemannian

  ["smooth stratified space" $\rightsquigarrow$ see [Pflaum] (book)]

  e.g. icecream cone / Kale: apex has to be its own stratum to get Riemannian strat. space

  exercise: embed $\mathcal{T}_3$ into Kale K_α for $\alpha > 3\pi$ $\longrightarrow$ 3 angles each larger than π

* geodesic has to be required (at least intrinsic), else extrinsic metric on λ would be allowed

- **Def.** A Whitney stratified space is a top. strat. space X embedded in a vector space where limits of secant lines joining strata M_i and M_j with $M_i \subseteq \overline{M}_j$ are contained in limits of tangent planes to M_j as points converge to M_i

  e.g. real or complex alg./semianalytic/subanalytic varieties

  Thm. Whitney stratification $\Rightarrow$ triangulation exists

- X strat. metric space

  Def. A prob. distr. ρ on X has Fréchet function

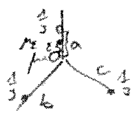
$$F(y) = \int_X d(x,y)^2 \, \rho(dx)$$

  ρ has Fréchet mean $\operatorname{argmin}_{y\in X} F(y)$.

- usual stat.: data wiggles $\Rightarrow$ mean wiggles

  strat. stat.: mean can stick

  e.g.



$a = b = c = 1 \Rightarrow \mu = \text{mean} = 0$

$$F(\mu) = a^2 + b^2 + c^2$$

$$F(\mu_\varepsilon) = F(\mu) - 2\varepsilon a + 2\varepsilon b + 2\varepsilon c + 3\varepsilon^2$$

$$\rightsquigarrow F(\mu_\varepsilon) - F(\mu) = 2\varepsilon\underbrace{(b+c-a)}_{>0} + 3\varepsilon^2 \quad \text{if } a < b+c \text{ for small } \varepsilon$$

  higher dim.: e.g. open book

- What is a law of large numbers (LLN)?

  $X_1, \dots, X_n$ iid. rand. var. with values in X distributed according to ρ

  LLN: $\overline{X}_n \longrightarrow \mu(\rho)$ as $n \to \infty$.

  $\mu(\rho)$ is sticky $\Rightarrow$?

  Fix X = open book = $K \times$ spider where $K = \mathbb{R}^d$ is spine. X is CAT(0) $\Rightarrow$ unique Fréchet mean

  An integrable prob. distr. ρ has 3 possibilities for mean $\mu(\rho)$:

  1) $\mu(\rho) \notin K$
  2) $\mu(\rho) \in K$ and $\mu(\rho') \in K$ for all ρ' "near" ρ
  3) $\mu(\rho) \in K$ but $\mu(\rho') \notin K$ for some ρ' "arbitrarily near" ρ

  Def. Corresponding to these cases, $\mu(\rho)$ is 1) non-sticky, 2) sticky, 3) partly sticky (there is also a perturbation which makes it sticky)

Thm [SAMSI WG] iid $X_1, X_2, \ldots$ in $X \sim \rho \Rightarrow \bar{X}_n \to \mu(\rho)$ as $n \to \infty$

Moreover • $\mu(\rho)$ is not sticky or partly sticky (1)+3))
is supp ρ spreads over 3 pages $\Rightarrow \exists$ page L and rand. $N \in \mathbb{N}$ s.t. $\bar{X}_n \in \bar{L}$ $\forall n \geq N$ a.s.

• $\mu(\rho)$ sticky $\Rightarrow \exists$ rand. $N \in \mathbb{N}$ s.t. $\bar{X}_n \in K$ $\forall n \geq N$ a.s.

• What is a central limit theorem (CLT)?

$X_1, X_2, \ldots \overset{iid}{\sim} \rho$, $\bar{X}_n \xrightarrow{\mathcal{D}} \delta_{\mu(\rho)}$

idea: rescale: $\sqrt{n}\,\bar{X}_n \xrightarrow{\mathcal{D}} \tilde{\rho}$ limiting distr. (classical: $\tilde{\rho}$ gaussian)

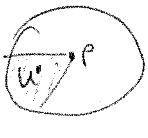
Thm [SAMSI WG] A CLT holds for square-int. ρ holds with limiting distribution

1) gaussian on $\mathbb{R}^{d+1} = T_{\mu(\rho)}X = T_{\mu(\rho)}L$

2) gaussian on $\mathbb{R}^d = K = T_{\mu(\rho)}K$

3) half gaussian on L plus its reflected projection to K ($\frac{1}{2} *$ gaussian on $T_{\mu(\rho)}K$)

• Def. Fix a metric space X and a top set $\mathcal{P}$ of int. prob. distr. on X as well as a subset $K \subseteq X$.



The mean of $\mu(\rho)$ of $\rho \in \mathcal{P}$ sticks to K if every neighborhood U of ρ in $\mathcal{P}$ contains a nonempty open set $U' \subseteq U$ with $\mu(\rho') \subseteq K$ $\forall \rho' \in U'$.

!

• Example (Huckemann, Mattingly, Nolan, Miller):

X = isolated planar hyperbolic singularity (Kale with $\alpha > 2\pi$)

$\Rightarrow$ CLT has limiting distribution:

$\rightsquigarrow$ induces further stratification (boundary of sector)