Statistical Machine Learning
Manifold Learning

Horacio Gómez-Acevedo
Department of Biomedical Informatics

February 22, 2022

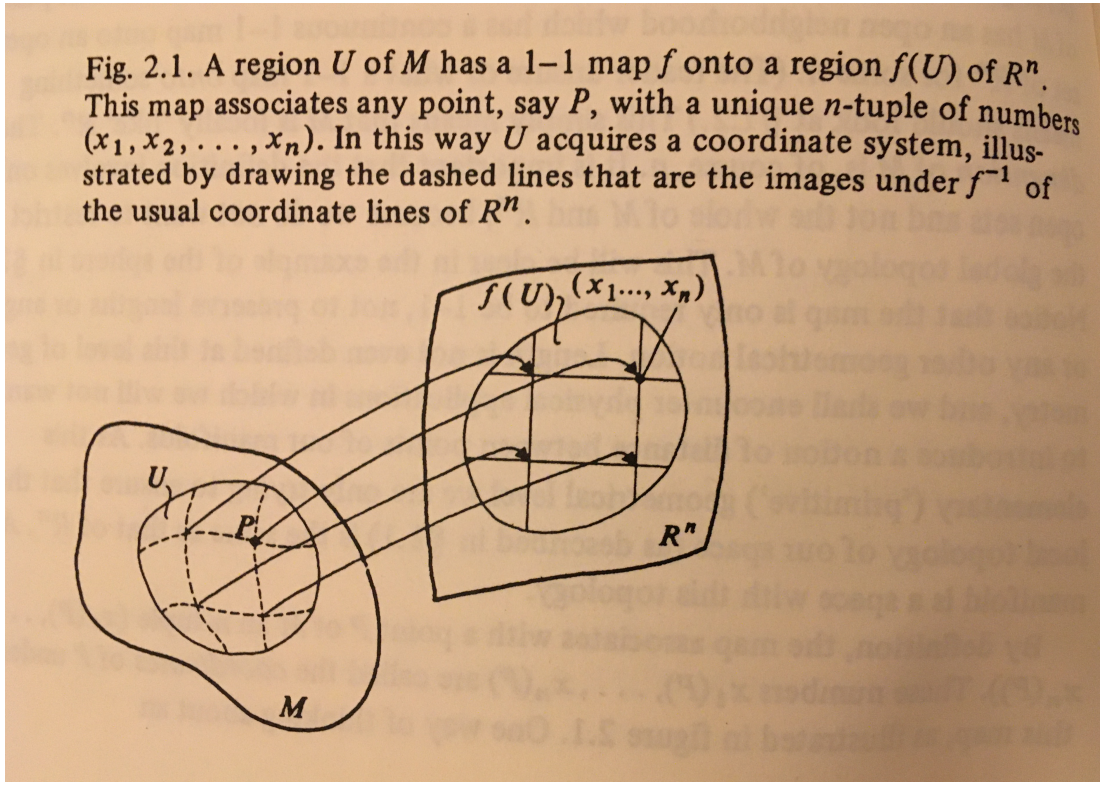
UAMS

Linear Models

Popular methods of data analysis make the assumption that the data lies on a linear k-dimensional subspace of $\mathbb{R}^n$ where $k < n$.
The general problem reduces to search for a linear transformation $\theta : \mathbb{R}^n \to \mathbb{R}^k$ such that $\{\theta(x_i)\}$ retains something about the structure of $\{x_i\}$.
Linear models are ubiquitous in sciences, but the assumption of linearity is often unrealistic. One alternative, is to consider that the data has been sampled from a (compact Riemannian) **manifold** $\mathcal{M} \subset \mathbb{R}^n$ of much lower dimension than n.

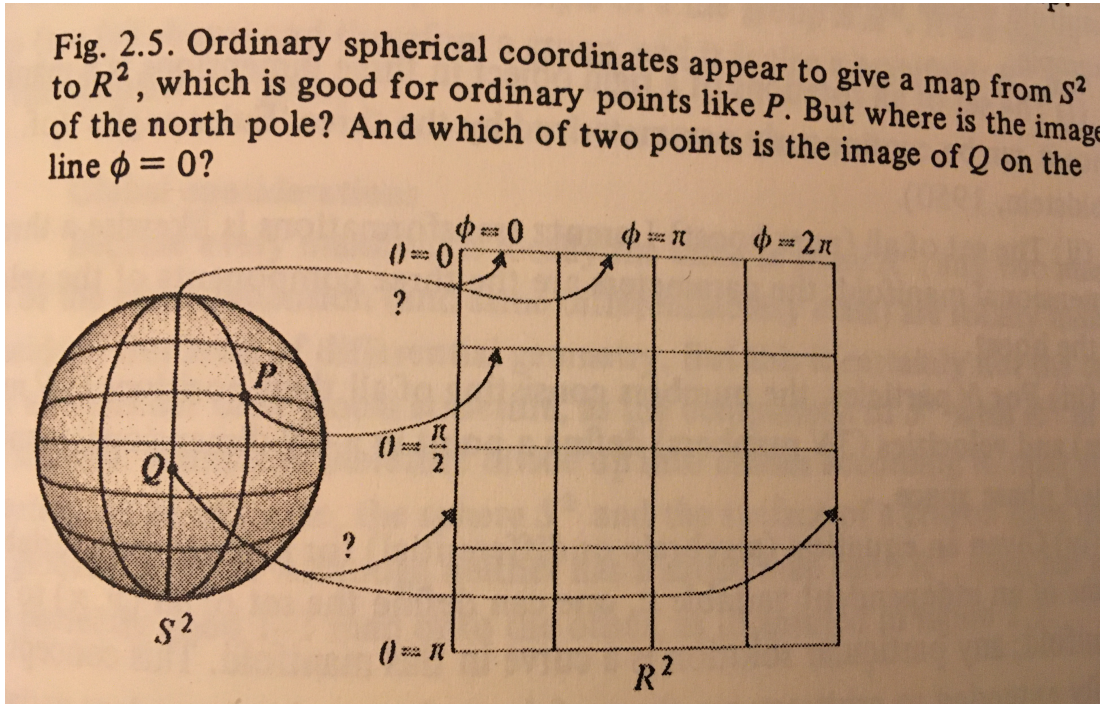
What is a manifold?

Roughly speaking, a manifold (of dimension 2) is a surface that locally behaves as a regular space in $\mathbb{R}^2$. In geometry, we refer to local properties as intrinsic.



The Sphere as a manifold

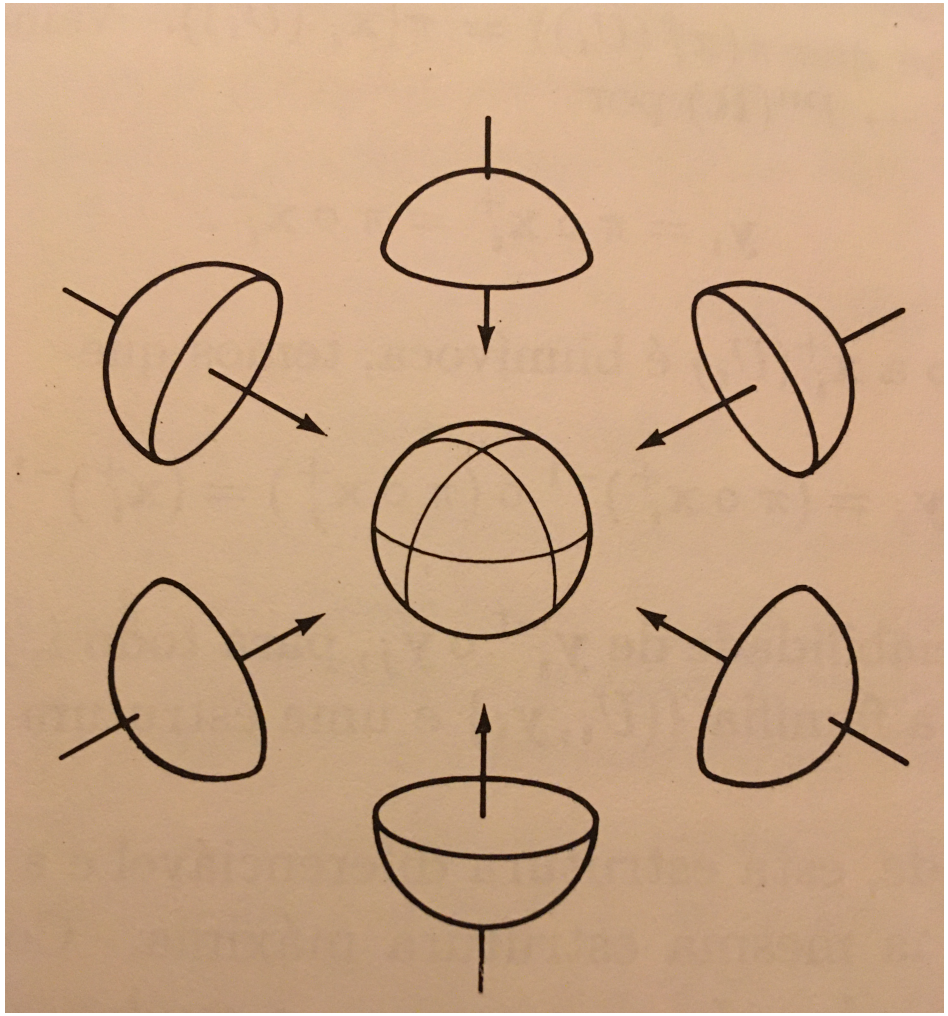
Another typical example is the sphere (often called S^2).



Intuitively, we cannot cover the whole sphere (in a unique way) with only one sheet. The north and south poles are a problem.

Sphere as a patch work

Instead, we use several overlapping sheets



Manifold Learning

Loosely speaking, the basic idea of most manifold learning techniques is to take the k-nearest neighbors of a point x and use the vectors specified by the line segments from x to its neighbors as an approximation for the tangent plane at x.
Global optimization then sews these local approximations together to produce a low-dimensional representation of the data.

Manifold learning is also known as dimensionality reduction.

PCA

In a general setup, we can think that we have m vectors in $\mathbb{R}^n$, namely $\{x_1, \ldots, x_m\}$. We try to find the "optimal" linear projection $\theta\colon \mathbb{R}^n \to \mathbb{R}^k$ where $k \ll n$ (meaning k much lower than n)

1. $\tilde{x_j} = x_j - \mu$ where $\mu = \frac{1}{m}\sum x_j$
2. We find the variance-covariance matrix

$$C = \frac{1}{m}\sum_{j=1}^{m} \tilde{x_j}\tilde{x_j}^t$$

3. We compute the top k-eigenvectors $\{v_1, \ldots, v_k\}$ of C.
4. These eigenvectors span a hyperplane (subspace) of $\mathbb{R}^n$. The projection $\theta\colon \mathbb{R}^n \to \mathbb{R}^k$ is precisely the orthogonal projection onto this plane followed by a choice of identification of the plane with $\mathbb{R}^k$.
5. We can think θ as

$$y_j = \theta(\tilde{x_j}) + \mu$$

PCA cont

The process chooses the basis which maximizes the variance captured by the representation; the eigenvector v_1 with the largest eigenvalue is the single direction that captures the maximal amount of information about the variance; the plane spanned by $\{v_1, v_2\}$ is the plane with the most variance, etc.
Another characterization of PCA is to find a projection that minimizes the error function

$$E = \sum_{j=1}^{m} \|x_j - y_j\|^2,$$

where $\|\cdot\|$ denotes some distance. For PCA this means that it produces the points $\{y_i\}$ that minimize the reconstruction error among all projections onto a k dimensional subspace.

Multidimensional scaling (MDS)

In this methodology, we search for a mapping $\theta\colon \mathbb{R}^n \to \mathbb{R}^k$ that minimizes

$$\mathcal{E} = \sum_{x_i, x_j} (\|x_i - x_j\| - \|\theta(x_i) - \theta(x_j)\|)^2$$

The procedure is as follows.

1. Consider the matrix of the original distances $D = (D_{ij})$ where $D_{ij} = \|x_i - x_j\|$.
2. Set $H = I_n - \frac{1}{n}\mathbb{1} \cdot \mathbb{1}^t$, where $\mathbb{1} = (1, 1, \ldots, 1)$. Set $Z = -\frac{1}{2}HDH$.
3. The function that minimizes $\mathcal{E}$ is then given by finding the eigenvectors v_j of Z. The $\theta(x_j)$ is specified by normalizing so that $\|v_j\|^2 = \lambda_j$.

What we have learned?

- The power of resampling methods for estimation of parameters.
- Bootstrap is considered one of the most powerful statistical techniques for estimation.
- Methods for subset selection of covariates in a regression setting.
- We explore the forward and backward selection.

References

Some of the pictures are from (Do Carmo, 1988), and (Schutz, 1980), and (Géron, 2019). Main reference for manifold learning (Rabadán and Blumberg, 2020).

- Do Carmo, M. P. (1988). *Geometria Riemanniana*. Projeto Euclides. IMPA. ISBN: 85-244-0036-6.
- Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. Edition. O'Reilly. ISBN: 978-1-492-03264-9.
- Rabadán, R. and A. J. Blumberg (2020). *Topological Data Analysis for Genomics and Evolution*. Cambridge University Press. ISBN: 978-1107-159549.
- Schutz, B. (1980). *Geometrical methods in mathematical physics*. Cambridge University Press. ISBN: 978-0521-298872.

I have used some of the graphs by hacking TiKz code from StakExchange, Inkscape for more aesthetic plots and other old tricks of TEX