How to Classify a Document using Information Geometry?

Notes on [Lebanon2005c] and more

H.Luo

(3rd revised 0404)

2016-04-05

Question: How to classify a document?

Question 1: How to represent a document?

•Classical idea of document classification[Madigan, David, et al.]
To regard **frequency vectors** of a specific document as embedded into the Euclidean space $\mathbb{R}^n$ and use the **Euclidean metric** defined on Euclidean space $\mathbb{R}^n$ to derive a similarity measure to classify documents.

•Lebanon's idea of document classification [Lebanon2005c]
To regard **frequency vectors** of a specific document as embedded into the multinomial manifold simplex $\mathbb{P}_{n-1}$ via *homogenization* and use the **Fisher metric** defined on simplex $\mathbb{P}_{n-1}$ to derive a similarity measure to classify documents.

Question: How to classify a document?

Question 2: What is the statistical model of this representation?

Multinomial model with parameters θ being the frequencies of words in a specific document.

With this model, there are some existing results:

(1) Use the logistic regression to find a fitted frequency vector and determine which category this document belongs to. [Shalizi]

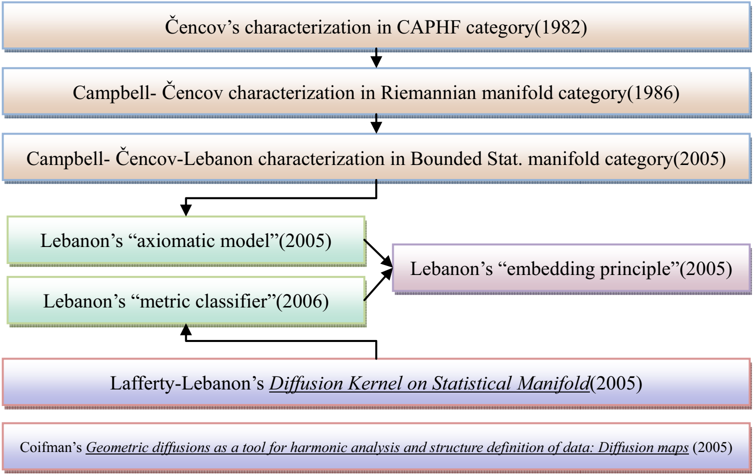
(2) Use the Bayesian logistic regression to impose a prior on the parameter θ. [Madigan, David, et al.]

Solution: Lebanon's new ideas on document-classification.

(1)Use the cosine measure induced by Fisher metric directly. [Lebanon2005b]
(2)Use the diffusion kernel's parametrix expansion on the simplex $\mathbb{P}_{n-1}$. [Lafferty&Lebanon]
(3)Use a different model other than multinomial distribution. [Lebanon2005a]
And all these generalizations can be understood using the **embedding principle** proposed in [Lebanon2005c]

The organization of [Lebanon2005c]

Figure: Overview



Čencov's characterization in CAPHF category(1982)

Campbell- Čencov characterization in Riemannian manifold category(1986)

Campbell- Čencov-Lebanon characterization in Bounded Stat. manifold category(2005)

Lebanon's "axiomatic model"(2005)

Lebanon's "metric classifier"(2006)

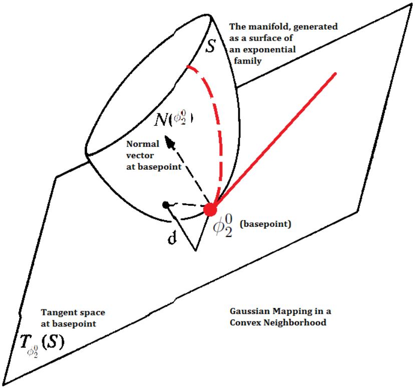
Lebanon's "embedding principle"(2005)

Lafferty-Lebanon's *Diffusion Kernel on Statistical Manifold*(2005)

Coifman's *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps* (2005)

Basics about manifolds

Figure: [Do Carmo] Fig 3-16


The manifold, generated as a surface of an exponential family
S
N(ϕ_2^0)
Normal vector at basepoint
ϕ_2^0
(basepoint)
d
Tangent space at basepoint
$T_{\phi_2^0}(S)$
Gaussian Mapping in a Convex Neighborhood

Exponential Families are naturally manifolds.

Theorem. *(Transitional mappings in exponential families, modified from 1.7 in [Lawrence Brown])*
Consider a standard exponential family $f_X(x) = exp\left(\theta \cdot x - \psi(\theta)\right)$ *with natural parameter space* Θ*. Let the linear mapping* $M_1 : \mathbb{R}^k \to \mathbb{R}^m$ *be of rank m and its orthogonal complement in* $\mathbb{R}^k$ *be* M_2*. Construct their product mapping as* $M := \left(\begin{array}{c} M_1 \\ M_2 \end{array}\right), Z := M'X = \left(\begin{array}{c} M_1'X \\ M_2'X \end{array}\right)$*, and find a base point* $\phi_2^0 \in M_2'^{-1}(\Theta)$*, then the family of distributions of* $M_1'X$ *over the parameter space* $M_1'^{-1}\left(M_2'\left(\phi_2^0\right)\right)$ *form another standard exponential family* $f_Z(z) = exp\left(\left[M'^{-1}(\theta)\right] \cdot z - \psi(\theta)\right)$ *with natural parameter space* $M_1'^{-1}\left(M_2'(\Theta)\right)$*.*

Statistical Manifolds

Definition.(Statistical manifold, [Murray & Rice] pp.76)
A statistical manifold $\mathcal{S}$ is a subset of $\mathcal{P}$, the space of probability measures in a collection of measures $\mathcal{M}$ on a sample space Ω.
This means that $(\Omega, M), \forall M \in \mathcal{M}$ is a measurable space whilst $(\Omega, P), \forall P \in \mathcal{P} \subset \mathcal{M}$ must be a probability space.
It is also required that
(1) The log-likelihood function[1]
$\ell : P \to \mathbb{R}_\Omega, p = e^f \mu \mapsto f, \forall \mu \in \mathcal{M}$ is smooth mapping on manifold.[2]
(2) For any point $p \in \mathcal{S}$ and the intrinsic coordinates θ about p the random variable $\frac{\partial \ell}{\partial \theta^i}(p)$ are linearly independent. [3]

[1] Log mapping is inverse to exponential mapping which maps the tangent vectors to geodesics.

[2] Here the e^f plays the role of normalizing factor, I think we can directly draw from $\mathcal{P}$ without much complication.

[3] Using this log-likelihood embedding, we actually embed a family of probability measures into the random variable space $\mathbb{R}_\Omega$ with the canonical coordinate frames $\frac{\partial \ell}{\partial \theta^i}(p)$.

Fisher metric.

What is a Fisher metric?

Fisher information matrix is a positively definite matrix and hence it can define a metric in the tangent bundle of the manifold Θ.

•Tangent bundle is a family of linear spaces over the statistical manifold. Collect all tangent spaces together we have a natural *linear transitional structure* on it, called this manifold of dimension $2 \times (dim\mathcal{X} + 1)$ the tangent bundle of the statistical manifold Θ.

•Statistical manifold here is defined by the parameterization of the parameterization given by the parametric family $p(x|\theta), \theta \in \Theta$. A surface can be defined by $z = f(x, y)$, the parametric family does the same job of defining a "surface" in the space of $\mathcal{X} \times \mathbb{R}$.

•Fisher metric simply impose the Fisher Informative metric g onto each of its tangent spaces. The whole stuff defined above is (Θ, g), a Riemannian manifold.

Why should we use Fisher metric?

Theorem.*(Cencov-Campbell-Lebanon Theorem, [Lebanon2005b], pp.1290-1293) Let* $\left\{\left(\mathbb{R}_+^{k\times m}, g^{(k,m)}\right) : k \geq 1, m \geq 2\right\}$ *be a sequence of Riemannian manifolds with the property that every* ***congruent embedding*** *by Markov morphism is an isometry iff the metrics defined on the manifold are of the form* $g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = A(|M|) + \delta_{ac}\left(\frac{|M|}{|M_a|}B\,|M| + \delta_{bd}\frac{|M|}{|M_{ab}|}C\,|M|\right), A, B, C \in C^\infty(\mathbb{R}^+)$ *where the notation* $|M| = \sum_i |M_i| = \sum_{ij} |M_{ij}|$ *means the row norm and entry-wise norm.*

Let's introduce some notions needed for proving and understanding this result.

Definition. ($\mathcal{A}$-stochastic matrix) $\mathcal{A}$-stochastic matrix is a Markov matrix whose rows' nonzero element is indexed by the partition $\mathcal{A}$ of $1, \cdots$ *number of rows*.

(**Congruent mapping** by Markov morphism, [Lebanon2005b] p.1287)

Let $\mathcal{B}$ be a k sized [4] partitioned of $\{1, \cdots, l\}$ and $\left\{\mathcal{A}^{(i)}\right\}_{i=1}^{k}$ be a set of m sized partitions of $\{1, \cdots, n\}$ Furthermore, let $R \in \mathbb{R}_{+}^{k \times l}$ be a $\mathcal{B}$-stochastic matrix and $Q = \left\{Q^{(i)}\right\}_{i=1}^{k}$ a sequence of $\mathcal{A}^{(i)}$-stochastic matrices in $\mathbb{R}_{+}^{m \times n}$. Then the map $f : \mathbb{R}_{+}^{k \times m} \to \mathbb{R}_{+}^{l \times n}, M \mapsto R'(M \otimes Q)$ is termed as a congruent embedding by a Markov morphism $\mathbb{R}_{+}^{k \times l} \to \mathbb{R}_{+}^{m \times n}$ and the set of all such maps is denoted by $\mathfrak{F}_{m,k}^{l,n}$.

[4] Which means there are k subcollection in such a partition.

i.e.

$$\begin{tikzcd}
\mathcal{A}^{(1)} \arrow[d] & \mathcal{A}^{(2)} \arrow[d] & \cdots & \mathcal{A}^{(k)} \arrow[d] & m - partitions \\
Q^{(1)} \arrow[rrd] & Q^{(2)} \arrow[rd] & \cdots & Q^{(k)} \arrow[ld] & (m \times n) \\
 & & M \otimes Q \arrow[dd] & & (k \times n) \\
 & & & & M\, is\ (k \times m) \\
 & & R'(M \otimes Q) & & (l \times n)
\end{tikzcd}$$

$M \otimes Q$ is defined as: i-th row of $M \otimes Q$ is exactly the i-th row of $MQ^{(i)}$ where $Q = \left\{Q^{(i)}\right\}_{i=1}^{k}$.

$$\text{i.e.} \begin{pmatrix} & \vdots & \\ \cdots & i - th\, row\, of\ \left[MQ^{(i)}\right] & \cdots \\ & \vdots & \end{pmatrix}$$

This is actually a (linear) mapping between tangent bundles of manifolds in sense of Riemannian geometry which is induced by the pull-back of a Markov matrix M.

The i-th member of this set of basis was transformed by the Markov mapping/morphism $Q^{(i)}$. The new matrix $M \otimes Q$ is the a matrix whose column vectors are lying in the tangent space of $T_{Qx}\mathcal{N}$ where $\mathcal{N}$ is another manifold of dimension k globally/pointwisely transformed by $\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_m}\right)_{1\times m} \mapsto$ $\left(\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_m}\right) Q^{(1)}, \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_m}\right) Q^{(2)}, \cdots, \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_m}\right) Q^{(k)}\right)_{1\times k}$,

The $\mathcal{B}$-stochastic matrix R since it is just a regularizing transform which is frequently encountered in geometers' books, if you want you can simply set it to a permutation matrix without loss of generality.

Devils in details.

Although it is easy to understand the multinomial distribution as a manifold, in order to apply geometric results, we have some details to consider:

(1) The $\mathbb{P}_{n-1}$ is a manifold yet not compact nor smooth across the boundary.

•Solution: [Lebanon2005b] uses a smoothing method to overcome this, statistically it introduces some kind of association.

(2) The diffusion kernel[5]/other operator kernel defined on a (compact) manifold is generally nor computationally tractable.

•Solution: [Sogge, Lafferty&Lebanon] exposed how to calculate the kernel using the parametrix expansion.

Both solutions are NOT new in differential geometry/pseudo-differential calculus. They can be traced back to 1970s.

[5] Here the kernel is the solution to the diffusion equation defined on the manifold.

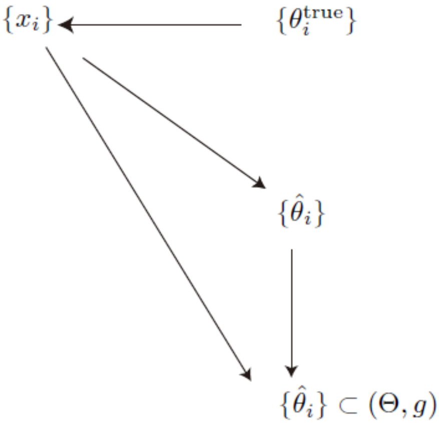
What are further details of the correspondence between geometric objects and statistic objects?

(1) [McCullagh] tells the differential aspect of the story;
(2) [Cencov] tells the categorical correspondence;
(3) [Watanabe] tells the algebraic aspect of the story.
If time permits, I will say a few words about the latter two stories.

The Embedding Principle

Theorem. *(The embedding principle, [Lebanon2005c])*
Assume that the data $x_1, \cdots, x_n$ is drawn from n distinct distributions $p(x; \theta_1^{TRUE}), \cdots, p(x; \theta_n^{TRUE})$ that lie in the same family $\theta_1^{TRUE}, \cdots \theta_n^{TRUE} \in \Theta$.
As the sampling simulates noisy corruption,
we can replace the data $x_1, \cdots, x_n$ with the underlying distributions $\theta_1^{TRUE}, \cdots \theta_n^{TRUE}$ thus obtaining an embedding of the data in a Riemannian manifold (Θ, g) where g is the Fisher information metric.
In most cases, we do not know the underlying distributions $\theta_1^{TRUE}, \cdots \theta_n^{TRUE}$. An approximate version of the embedding principle is to embed the data in (Θ, g) using estimates $\hat{\theta}_1, \cdots \hat{\theta}_n$ obtained by estimators such as MLE,MAP or empirical Bayes.

Figure: [Lebanon2005c] Fig.1



$$\begin{tikzcd}
\{x_i\} \arrow[dr] \arrow[ddr] & \{\theta_i^{\text{true}}\} \arrow[l] \\
& \{\hat{\theta}_i\} \arrow[d] \\
& \{\hat{\theta}_i\} \subset (\Theta, g)
\end{tikzcd}$$

Let's see how the document classification using the multinomial model can be put into geometric framework using our embedding principle.

Model 1: Logistic Regression on $\mathbb{P}_{n-1}$

Only (1) (5) is relevant to embedding principle, (2) (3) (4) are typical method for classification.

(1) Embed the data points into simplex $\mathbb{P}_{n-1}$ with Fisher metric using the **embedding principle**.

(2) Embed the $\mathbb{P}_{n-1}$ into hemisphere $\mathbb{S}^{+}_{n-1}$ using the centered spherical projection ι.

(3) Determine the linear margin classifier using a plane $P \subset \mathbb{R}^n$ cutting through the hemisphere $\mathbb{S}^{+}_{n-1}$.

(4) Pull-back the intersection $P \cap \mathbb{S}^{+}_{n-1}$ using the inverse spherical projection ι^{-1}.

(5) $\iota^{-1}\left(P \cap \mathbb{S}^{+}_{m-1}\right)$ is the logistic margin classifier on simplex $\mathbb{P}_{n-1}$.

Model 1: A simplest example

(1) Given a simplest document containing only one sentence: "Rose is a rose is a rose is a rose." Suppose that there are two possible authors of this sentence [6]: A robot who can only repeat "rose" and a great poet.
The frequency vector of robot is $\{1, 0, 0\}$ for $(\mathit{rose}, \mathit{is}, a)$.
The frequency vector of this sentence is $\left\{\frac{4}{10}, \frac{3}{10}, \frac{3}{10}\right\}$ for $(\mathit{rose}, \mathit{is}, a)$.
We do not necessarily know the parameter of the poet's multinomial model.
Using embedding principle,
$\mathcal{X} \to \Theta \cong \mathbb{P}_{3-1}, \left\{\frac{4}{10}, \frac{3}{10}, \frac{3}{10}\right\} \mapsto \mathit{Multinomial}\left(\frac{4}{10}, \frac{3}{10}, \frac{3}{10}\right)$.

[6] The source of this sentence is actually (*Sacred Emily*, Geography and Plays) by G.Stein.

(2) Embed the $\mathbb{P}_{3-1}$ into hemisphere $\mathbb{S}^{+}_{3-1}$ using the centered spherical projection ι.
(3) Determine the linear margin classifier using a plane $P \subset \mathbb{R}^{n}$ cutting through the hemisphere $\mathbb{S}^{+}_{n-1}$.
(4) Pull-back the intersection $P \cap \mathbb{S}^{+}_{n-1}$ using the inverse spherical projection ι^{-1}.
(5) $\iota^{-1}\left(P \cap \mathbb{S}^{+}_{m-1}\right)$ is the logistic margin classifier on simplex $\mathbb{P}_{n-1}$. After doing this, we can directly compare $\left\{\frac{4}{10}, \frac{3}{10}, \frac{3}{10}\right\}$ to some other parameter vector to determine whether this sentence is conducted by a poet or a robot.

Model 2: Diffusion kernel SVM on $\mathbb{P}_{n-1}$

Only (1) is relevant to embedding principle, (2) (3) are typical method for SVM on $\mathbb{P}_{n-1}$ given by the principle.

(1) Embed the data points into simplex $\mathbb{P}_{n-1}$ with Fisher metric using the **embedding principle**.

(2) Use the kernel function obtained by parametrix expansion of the diffusion kernel on simplex $\mathbb{P}_{n-1}$ to map these data points in the SVM methods.

(3) Determine the linear margin classifier using Lagrange method with the kernel specified.[7]

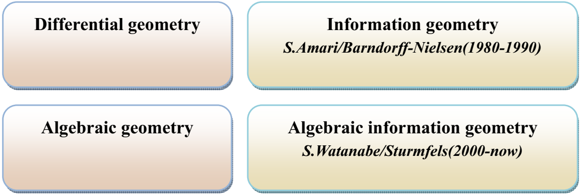
I recommend the paper [Coifman et.al] for those who are interested in choosing a correct kernel. Actually this paper extends the details in [Lafferty&Lebanon] paper.

[7] The kernel function can be regarded as a measure of similarity of two data points. We know that we have to project the data points to higher dimensional space in order to find a satisfying SVM separating surface using the Lagrange multiplier method including a term of higher dimensional inner product. The higher dimensional inner product is replaced by kernel function.

Why we should refer to our geometric intuition?

“Vision, I understand from friends who work in neurophysiology, uses up something like 80 or 90 percent of the cortex of the brain. There are about 17 different centers in the brain, each of which is specialized in a different part of the process of vision: some parts are concerned with vertical, some parts with horizontal, some parts with colour, perspective, finally some parts are concerned with meaning and interpretation. Understanding, and making sense of, the world that we see is a very important part of our evolution. **Therefore spatial intuition or spatial perception is an enormously powerful tool, and that is why geometry is actually such a powerful part of mathematics**-not only for things that are obviously geometrical, but even for things that are not. We try to put them into geometrical form because that enables us to use our intuition. ”
-by *Sir Michael Atiyah*, Fields Lecture at the World Mathematical Year 2000 Symposium, Toronto, June 7-9, 2000.

What can we do with algebraic informative geometry?:Advantages



Differential geometry

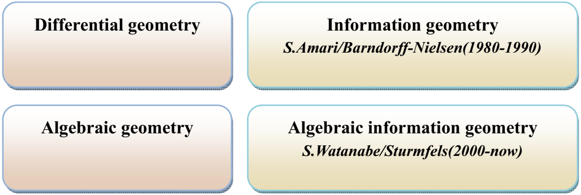
Information geometry
S.Amari/Barndorff-Nielsen(1980-1990)

Algebraic geometry

Algebraic information geometry
S.Watanabe/Sturmfels(2000-now)

(1) Calculable, elegant framework.
(2) Very convenient in dealing with infinite-dimensional problem.

What can we do with algebraic informative geometry?:Disadvantages



Differential geometry

Information geometry
S.Amari/Barndorff-Nielsen(1980-1990)

Algebraic geometry

Algebraic information geometry
S.Watanabe/Sturmfels(2000-now)

(1) It is often done on an algebraically closed field while statistics are done on $\mathbb{R}^n$.
(2) There are still some gaps between complex geometry and usual differential geometry.
(3) The categorical functor is not completed yet, actually it stumbled ever since its birth.

Thank you for your attention.

"Most AI workers are responsible people who **are aware of the pitfalls of a difficult field** and produce good work in spite of them. However, to say anything good about anyone is beyond the scope of this paper."
-McDermott, Drew. "Artificial intelligence meets natural stupidity." ACM SIGART Bulletin 57 (1976): 4-9.

Now it is discussion and question time.

References

Madigan, David, et al. "Author identification on the large scale." Proc. of the Meeting of the Classification Society of North America. 2005.

Do Carmo, Manfredo Perdigao, and Manfredo Perdigao Do Carmo. Differential geometry of curves and surfaces. Vol. 2. Englewood Cliffs: Prentice-hall, 1976.

Johnson, Mark. Probabilistic Models for Computational Linguistics, Macquarie University Machine Learning summer school, 2010.

Brown, Lawrence D. "Fundamentals of statistical exponential families with applications in statistical decision theory." Lecture Notes-monograph series (1986): i-279.

A.Stuart&J.K.Ord&M.Kendall, The Advanced Theory of Statistics(Volume 2: Classical Inference and Relationship), 5ed, Oxford University Press, 1991

Bhattacharyya, A. "On some analogues of the amount of information and their use in statistical estimation." Sankhyā: The Indian Journal of Statistics (1946): 1-14.

Fend, A. V. "On the attainment of Cramér-Rao and Bhattacharyya bounds for the variance of an estimate." The Annals of Mathematical Statistics (1959): 381-388.

Watanabe, Sumio. Algebraic geometry and statistical learning theory. Vol. 25. Cambridge University Press, 2009.

Lebanon, Guy. Riemannian geometry and statistical machine learning. Carnegie Mellon University, Language Technologies Institute, School of Computer Science, 2005. (Ph.D Thesis)

McCullagh, Peter. Tensor methods in statistics. Vol. 161. London: Chapman and Hall, 1987.

Mumford, David. Elastica and computer vision. Springer New York, 1994.

Cao, Yan, and David Mumford. "Geometric structure estimation of axially symmetric pots from small fragments." Signal Processing, Pattern Recognition, and Applications, IASTED International Conference. Vol. 2. 2002.

Gu, Xianfeng, and Shing-Tung Yau. "Computing conformal structure of surfaces." arXiv preprint cs/0212043 (2002).

Murray, Michael K., and John W. Rice. Differential geometry and statistics. Vol. 48. CRC Press, 1993.

Lebanon, Guy. "Axiomatic geometry of conditional models." Information Theory, IEEE Transactions on 51.4 (2005): 1283-1294.

Campbell, L. L. "An extended Čencov characterization of the information metric." Proceedings of the American Mathematical Society 98.1 (1986): 135-141.

Cencov, N. N. "Statistical decision rules and optimal inferences, Translation of Math." (1982).

Drton, Mathias, Bernd Sturmfels, and Seth Sullivant. Lectures on algebraic statistics. Vol. 39. Springer Science & Business Media, 2008.

Améndola, Carlos, Jean-Charles Faugère, and Bernd Sturmfels. "Moment Varieties of Gaussian Mixtures." arXiv preprint arXiv:1510.04654 (2015).

Grenander, Ulf. Probabilities on algebraic structures. Courier Corporation, 2008.

Lebanon, Guy. "Information geometry, the embedding principle, and document classification." Proceedings of the 2nd International Symposium on Information Geometry and its Applications. 2005.

- McCullagh, Peter, and John A. Nelder. Generalized linear models. Vol. 37. CRC press, 1989.

- Shalizi, Cosma. uADA, Chapter 12: Logistic Regression, 2012. http://www.stat.cmu.edu/∼cshalizi/uADA/12/lectures/ch12.pdf

- Beran, Rudolf. "Minimum Hellinger distance estimates for parametric models." The Annals of Statistics (1977): 445-463.

- Sogge, Christopher D. Fourier integrals in classical analysis. Vol. 105. Cambridge University Press, 1993.

- Lafferty, John D., and Guy Lebanon. "Diffusion kernels on statistical manifolds." (2005).

- Coifman, Ronald R., et al. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps." Proceedings of the National Academy of Sciences of the United States of America 102.21 (2005): 7426-7431.