This work is a field guide for Applied Machine Learning. These pages have variously served as my personal notes and software capability demonstrations over the years and it is my hope that . A certain amount of mathematical sophistication is assumed at points.

Contents

This work addresses

> *the application and implementation of techniques in Machine Learning, Statistical Pattern Recognition, & Spectral Graph Theory*

to *problems in real world data analysis.*

Analytic Learning Theory

This section covers basic definitions and concepts of statistical learning theory. we approach this from a functional analytic viewpoint. Supervised learning in its most abstract setting requires finding a function $f(x)$ given instances $(x_i, f(x_i))$. Typically x_i is an independent and identially distribute (iid) sample from some unknown distribution. A loss function is a random variable

$$L : Ran(f) \times Ran(f) \to \mathbb{R}^+$$

defining the cost of misclassification. The risk associated with a candidate function f' is defined to be the expectation of the loss over the sample space Ω,

$$R(f') = \int L(f(\omega), f'(\omega)) d\omega$$

. Statistical learning theory is concerned with assessing the approximations to f given by minimizing the empirical loss associated with a sample $x_i, f(x_i)$.

The notion of a loss function goes back to the roots of modern probability theory and economics. The St. Petersburg paradox is an example of a random variable $S : \mathbb{N} \to \mathbb{R}^=$ with infinite expectation limited utility. Let $W(k)$ be the winnings after k plays of from a game with outcome S that pays 2^{i-1} with probability $p_i = 1/2^i$. The expected payout is given by

$$\lim_{k \longrightarrow \infty} W(k)/k = E(S) = \sum_{i=1}^{\infty} p_i 2^{i-1} = \infty$$

The implication for a decision theory based only on expected value is that a rational player would pay an infinite amount of money to play this game. Bernoulli introduced the notion of expected utility which takes into account the fact that a payout of 2^i may not have twice the utility of a payout of 2^{i+1} when i gets large. The utility U is a random variable on the sample space representing preferences of an agent. Loss represents the aversion of an agent to the outcomes of the sample space, $L(\omega)+U(\omega) = \alpha \; \forall \; \omega \in \Omega$ where α is constant. Expected loss $R(f')$ is the risk associated with choosing the approximation f'. Restricting the class of functions to consider when minimizing the risk for a candidate approximation to f is a key aspect of classifier design.

(BBCREVISIT GP RKHS)Gaussian processes provide a class of models and learning algorithms for real world problems that have a long history and are well characterized. Learning algorithms are cast as minimization problems $min_(H)R()$ in a Hilbert space $\mathcal{H}$ with a dot product that encapsulates a model and sample data. Bayesian methods are often employed for estimation and inference with Gaussian processes. They allow an intuitive approach to incorporating prior knowledge in classification problems and the ability to obtain confidence intervals for predictions. Many common regression and classification algorithms can be cast as minimization problems in a Reproducing Kernel Hilbert Space (RKHS).

Learning With Kernels

?, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?** Kernel learning is a paradigm for classification and regression where prior belief is expressed in the construction of a similarity matrix of distances between points in a feature space Ω by embedding via a non linear map ϕ in a higher [often infinite] dimensional Hilbert space using the kernel as an inner product.

$$K(x,x') =< \phi(x), \phi(x') >$$

$$K \succeq 0$$

$$SPD \Rightarrow \sum_{x \in \Omega} \sum_{x' \in \Omega} f(x)K(x,x')f(x') \geq 0 \forall f \ \in \ell^2(\Omega)$$

Recall that infinitely divisible probability distributions aries as the sum of *iid* random variables. Infinitely divisible kernels have the representation

$$K = K^{\frac{1}{n}} \dots K^{\frac{1}{n}}$$

$$K = e^{\beta H}$$

$$e^{\beta H} = \lim_{n \to \infty} (1 + \frac{\beta H}{n})^n$$

We construct a mutli-resolution representation of the data with exponentiated kernels. The sequence of kernels $K(\beta)$ represents a one parameter group associated with a diffusion on the graph of the data. A $\beta \to 0\infty$ the kernel moves from the identity to one that represents the clusters in the off diagonal components. The local structure of Ω is preserved in H while the global geometry of the data set is progressively revealed in $K(\beta)$ as we push the diffusion forward with the one parameter group. We can construct exponentiated kernels over direct products of sets $\Omega_1 \bigotimes \Omega_2$ that will allow for the class conditional representation [bbcrevisit term use multiclass]. Simply set $H = H_1 \bigotimes I_{\Omega_1} + H_2 \bigotimes I_{\Omega_2}$.

$$K(\beta) = e^{\beta H} = e^{(\beta H_1 \bigotimes I_{\Omega_1} + H_2 \bigotimes I_{\Omega_2})} \Rightarrow$$
$$\frac{d}{d\beta} K(\beta) = H(K_1(\beta) \bigotimes K_2(\beta))$$

The kernels thus constructed can be used to drive a diffusion on a graph by letting H be the familiar graph Laplacian. Furthermore, the continuum limit of infinite data can be analyzed in within the framework of a discreet stochastic process much the way the convergence of finite element solutions of PDE's takes place.

Kernel Density Estimation

To define the empirical distribution function of a sample of size N - place mass $1/N$ at each member of the sample. This forms a nonparametric estimate of the marginal density $P(X)$. This is a singular form of kernel smoothing for density estimation. If ψ belongs to some nice class of function, and $\int\limits_{\infty}^{-\infty} \psi(x)dx = 1$, we can form a parametric estimator for the pdf of a process from a sample population of size N by calculating

$$p(x;\theta) = \frac{1}{N\theta} \sum_{i-1}^{N} \phi(\frac{x - X_n}{\theta})$$

If ϕ happens to be a density then $p(x,\theta)$ is also a density. Letting $\theta \to 0$ for the right kernel, we get the empirical density of the sample population. The mean squared error of the estimator expressed as a bias term and a variance term is

$$Err_p(x;\theta) = E[p(x;\theta) - p(x)]^2 = E[p(x;\theta) - p(x)]^2 = (E[p(x;\theta)] - p(x))^2 + Var[p(x;\theta)]$$

Distance Functions & The Affinity Matrix

There are two key ingredients to forming the affinity matrix; a distance function, and a convention for which pairs to consider. If all or too many pairs are compared, sparse methods will not be possible. We can include n nearest points, all points within ϵ of x, or some other criteria. For instance when working with spatial data, the diffusion can be done on a graph formed by a tessellation of the locations. This is exactly how numerical solutions of heat diffusions are done.

If the feature vector is a histogram, the L^2 distance is not meaningful, in this case use the χ^2

The Graph Cut

Spectral clustering is a relaxation to an NP-hard problem of finding an optimal way to cut a graph. Here we describe some various cut criteria. Let

$$\begin{aligned} X_1, \ldots, X_n \ &\in \mathbb{R}^n \\ W_{ij} &= s(X_i, X_j) \\ d_i &= \sum_j W_{ij} \\ |A| &= card(A) \\ vol(A) &= \sum_{i \in A} d_i \end{aligned}$$

Definition 0.0.1 (Min Cut)**.**

$$mincut(A, B) = \sum_{i \in A\ j \in B} W_i j$$

Definition 0.0.2 (Balanced Cut / Ratio Cut)**.**

$$mincut(A, B) \frac{1}{|A|} \frac{1}{|B|}$$

Definition 0.0.3 (n Cut).

$$mincut(A,B)\frac{1}{vol(A)}\frac{1}{vol(B)}$$

Graph Spectra

Graph spectral methods are some of the most successful heuristic approaches to partitioning algorithms in solving sparse linear systems, clustering and, ranking problems. Eigenvalues of the graph Laplacian are used to transform a combinatorial optimization problem to a continuous one, typically a SDP problem. Recent advances in SDP optimization techniques have opened new avenues of research in combinatorial optimization. For instance, isoperimetric properties of a graph are used to find efficient communication networks, and fast convergence of Markov Chains.

Matrix Factorization

Many forms of matrix factorization can be cast as an optimization problem that involves minimization of generalized Bregman divergences**?**. Factorization algorithms such as NNMF, Weighted SVD, Exponential Family PCA, , pLSI, Bregman co-clustering **?** can be cast in this framework. The approach uses an alternating projection algorithm for solving the optimization problem which allows for generalizations that include row, column, or relaxed cluster constraints. A brief description of the algorithm is given below. The description of a generalized Bregman divergence can be found in **?**.

PCA and its generalization to the Exponential Family

PCA finds linear combinations of the variables that correspond to directions of maximal variance in the data. Typically this is performed via a singular value decomposition (SVD) of the data matrix $A \in R^{n,m}$, or via an eigenvalue decomposition if A is a covariance matrix in which case $A \in R^{n,n}$. Representing the data in the directions of maximum variance allows for a dimension reduction that preserves information. Principal component directions are uncorrelated which can be useful. PCA has the disadvantage that components are usually linear combinations of all variables. Weights in the linear combination data elements are non-zero. Sparse PCA is an attempt to find a low dimension representation of the data that explainers most of the variance.

Here we describe a generalization of Principal component analysis (PCA) to the Exponential Family of probability distributions. PCA is a popular dimensionality reduction technique that seeks to find a low-dimensional subspace passing close to a given set of points

$$\{x_i\} \subset \mathbb{R}^n$$

. The procedure is to solve the optimization problem that minimizes the sum of squared differences of the data points to the projections on a subspace spanned by the empirical variance after centering the data to have mean 0;

$$\sum_{i=i}^{n} \|x_i - \theta_i\|_{\ell^2}^2$$

. The choice of ℓ^2 norm here codifies the assumption of Gaussian data. An alternate interpretation of the algorithm is finding the parameters θ_i that maximizes the log likelihood of the data which corresponds to

$$\sum_{i=i}^{n} \|x_i - \theta_i\|_{\ell^2}^2$$

. The goal of PCA is to find the the true low dimensional distribution of the data given the assumption that data is corrupted by Gaussian noise. Bregman divergences

$$D_\phi(A, B) = \phi(A) - \phi(B) - \nabla\phi(B)(A - B)$$

offer a framework to extend PCA [and other spectral dimension reduction techniques] to the entire Exponential Family. Here ϕ is a striclty convex function. The roles of

Let θ_i be the natural parameter for dimension i, with Exponential distribution P_θ. Then the conditional expectation is given by

$$logP_\theta(x|\theta) = logP_0(x) + x\theta - G(\theta)\dot{}G \ni \int P_\theta dx = 1$$

We can model multivariate data where the conditional distribution can vary along the feature space. The common feature of this PCA model and GLZ regression is the derivative of G which is familiar link function and the loss function which is appropriate for $P_\theta(x|\theta)$. The non linear relationship in the GLZ regression model data is captured by the link function $h = \frac{d}{d\theta}G(\theta)'$. This feature is also passed on to the generalized PCA. Instead of projecting on to a linear subspace, a Bregman divergence is used as the distortion measure. This gives a convex optimization problem to solve which can be shown to converge. In **?** a dual function to ϕ is defined by the relationship $\phi(g(\theta)) + G(\theta) = h(\theta)\theta$ which is used to write the log likelihood as a Bregman divergence

$$\log P(x|\theta) = -logP_0(x) - \phi(x) + D_\phi(x, h(\theta))$$

. Typically x is a vector but extending to matrices is straightforward.

Sparse PCA [CS Notes from http://ugcs.caltech.edu/ srbecker/

Classic PCA is sensitive to outliers. Convex methods can be use to address this by attempting to factor the data X as a sum of a low rank component L and a sparse component S by solving

$$\underset{min}{S, L} \left\|L\right\|_* + \lambda \left\|S\right\|_{\ell_1} : X = L + S$$

See http://cvxr.com/tfocs/demos/rpca/ for a demo of this in action with the popular cvx software package.

Kernel Feature Selection via Conditional Covariance Minimization

Manifold Learning

There are numerous machine learning techniques which accomplish some form of dimensionality reduction. Manifold learning uses principal curves and manifolds to encode a natural geometric framework for nonlinear dimensionality reduction. These methods construct low-dimensional data representation using a cost function that retains local properties. Contrasting methods such as MDS employ proximity data via a similarity or distance matrices. The important ISOMAP **?**algorithm extends MDS by capturing geodesic measurements of non-local pairs on the data manifold M via an multi-scale approximation. Non-local distances are approximated via a shortest path on a K nearest neighbor clustering of the data. Effectively a ball in data space is used to represent a cluster, and a graph is then constructed to encode the non-local information. The connectivity of the data points in the neighborhood graph are the nearest k Euclidean neighbors in the feature space. Dijkstra's algorithm for computing shortest paths with weights is used to construct the proximity matrix from the neighborhood graph. The top n eigenvectors encode the coordinates in the low dimensional Euclidean space. Choosing the correct number of neighbors is an essential component to an accurate representation. Other shortest path algorithms that may be employed to calculate the geodesic distances are listed below:

- Dijkstra's algorithm finds the single-pair, single-source, and single-destination shortest path.
- Johnson's algorithm finds all pairs shortest paths
- Bellman-Ford algorithm single source problem and allows negative edge weights.
- Floyd-Warshall algorithm solves all pairs shortest paths.
- A* search algorithm solves the single pair shortest path problem.

In **?** a sampling condition is given which bounds the quality of the manifold embedding based on the quality of the neighborhood graph.

Graph Laplacian

?, ?, ?, ?, ?, ?, ?, ?

Let G be a connected simple graph with vertex set $V = 1, 2, ..., n$, edge set E and let each edge be associated with a positive number, called the weight of the edge. The above graph is called a weighted graph. An unweighted graph is just a weighted graph with each of the edges bearing weight 1. The weight $w(i)$ of a vertex V_i is the sum of the weights of the edges incident with it. There are a number of ways in which the Laplacian matrix L is defined; the combinatorial Laplacian, the normalized Laplacian and the unsigned Laplacian. Spectra from graph matrix representations may be obtained from the adjacency matrix A and the various Laplacian discretizations. Spectra can also be derived from the heat kernel matrix and path length distribution matrix.

The matrix representation of the graph Laplacian has a significant effect on the spectrum. Attributes may be accounted for by by a complex number that encodes the edge attributes. The node attributes may be encoded in the diagonal elements. The complex graph Laplacian matrix is Hermitian, and hence it has real eigenvalues and complex eigenvectors. Graph feature vectors can be embedded in a pattern space by PCA, MDS, and LDA(linear discriminant analysis). Attribute graphs may be characterized by the application of symmetric polynomials to the real and complex components of the eigenvectors. **?** This gives rise to permutation invariants that can be used for pattern vectors. Partitioning a graph into three pieces, with two of them large and connected, and the third a small separator set can be accomplished using the second eigenvector [the Feidler Vector] of the graph Laplacian. In the case or sparse graphs, the first few eigenvectors can be efficiently computed using the Lanczos algorithm [see section below on ARPAC]. This graph partitioning algorithm can be extended to give a hierarchical subdivision of the graph.

Distance Metrics

A measure of similarity between data points is a vital component to clustering algorithms. The suitability of any given measure is dependent on the generative process providing the data.

Markov Chains

Let $X \in \mathcal{M}$, $P(x,y)$ the transition probability for an irreducible Markov chain. If P is reversible relative to π then we have that $Q(x,y) = \Pi(x)P(x,y) = \Pi(y)P(y,x) \forall x, y, \in X$. This implies that π is the stationary distribution. Bounding the rate of convergence to the stationary distribution is related to L_G and isoperimetric problems.

Detailed balance is a property of a Markov chain where $\pi(x)P(x,y) = \pi(y)P(y,x) \Rightarrow$ reversibility. This is stronger than the requirement of a stationary distribution. It is the property that implies for every closed cycle of states there is no flow of probability.

There are four aspects of Markov chain stability.

1. π-irreducible

2. small set

3. Harris recurrence

4. Geometric Ergodicity

For reversible Markov chains, the rate of convergence to Π for a finite state chain is determined by the second eigenvalue of P.

There are two types of bounds for the rates of convergence

1. graph theoretic Cheeger type - see Diaconis and Stook

2. Spectral

The rate of convergence to π for finite state Markov chains is determined by the second eigenvalue λ_2 of P.

BBCREVISIT CLEAN UP LOTS

Theorem 0.0.1 (Perron Frobenius - $P \in \mathbb{P}^{n \times n}$)**.** *There is a simple largest eigenvalue, and it's eigenvector is positive.*

The growth of P^k is determined by the Perron eigenvalue. Frobenious extended the theorem to include a class of non negative matrices. This is achieved through the concept of reducibility. P is irreducible if

$$\neg\exists\ \Lambda\ \ni\ \Lambda^\dagger L \Lambda = \begin{pmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{pmatrix}$$

where Λ is a projection $\Lambda^2 = \Lambda$ The irreducible matrices L have a simple largest eigenvalue λ_n, now called the Perron-Frobenius eigenvalue whose right eigenvector components are all positive. We can put this in Markov chain terminology; irreducibility is equivalent to the existence of a unique stationary distribution.

BBC Make the Markov Chain section nice.

Hammerlsy Clifford

The Hammersley–Clifford theorem gives necessary and sufficient conditions for when a positive probability distribution can be represented as a Markov Random Field. It states that a probability distribution that has a positive mass or density satisfies one of the Markov properties with respect to an undirected graph G if and only if it is a Gibbs random field. The density of the Gibbs random field can be factorized over the cliques (complete subgraphs) of G.

Spectral Graph Theory

Modern spectral graph theory increasingly takes insights from geometry. Discrete analogues of isoperimetry results and heat flow on manifolds are just a few examples being put to use in modern applications. The normalized graph Laplacian is used to aid in consistency between spectral geometry and stochastic processes. We consider connected graphs $G = (E, V)$ in this work, in which case we can define the normalized graph Laplacian as $\mathcal{L} = T^{\frac{1}{2}} L T^{\frac{-1}{2}} = I - T^{\frac{1}{2}} A T^{-\frac{1}{2}}$, where A is the adjacency matrix, L is defined by

$$L(u,v) = \begin{cases} d_v & : \ u = v \\ -1 & : u \sim v \\ 0 & : u \nsim v \end{cases}$$

and $T = diag\{d_1, \cdots, d_n\}$ where d_v is the degree of vertex v.

$\mathcal{L}$ is a difference operator :

$$\mathcal{L} = \frac{1}{\sqrt{d_u}} \sum^{v:u\sim v} (\frac{g(u)}{\sqrt{d_u}} - \frac{g(v)}{\sqrt{d_v}})$$

$$Vol(G) = \sum_{v\in V}^{d_v} = Tr(T) \tag{0.0.1}$$

$$\sigma(\mathcal{L}) \in \mathbb{R}^+ \tag{0.0.2}$$

$$ker(\mathcal{L}) = span\{T^{\frac{1}{2}}\nVdash\} \tag{0.0.3}$$

Diffusion Maps

?, **?**, **?**, **?**, **?**, **?**.

Spectral clustering involves constructing a Markov chain over a graph is constructed over the graph of the data and using the sign of the first non-constant eigenvector for graph cuts and cluster localization. This approach can be generalized to higher-order eigenvectors yielding a multi-resolution view of the data. Using multiple eigenvectors allows one to embed and parameterize the data in a lower dimensional space. Examples of this procedure include LLE, Laplacian & Hessian Eigenmaps. The common theme among these approaches is that eigenvectors of a Markov process can encode coordinates of the data set on a low dimensional manifold in a Euclidian space. The advantage over conventional methods is that the representation is non-linear and they preserve local structure. Kernel eigenmap embeddings can be generalized into a diffusion framework where a discrete Laplacian acts on a low dimensional representation space. This allows for a true multi-scale parametrization. Iterating a Markov process involves computing power of the transition matrix to run a random walk of the graph forward in time. By construction a one parameter map defining the diffusion and specifying boundary conditions the full power of diffusions on a smooth manifold may be brought to bear on parameterizing the geometry of the data. Different boundary conditions and diffusion operators give rise to a discrete approximations of familiar stochastic PDE's.

Let $(X, \mathcal{A}, \mu)$ be a measure space and $k : X \times Y \longrightarrow \mathbb{R}$ a kernel function.

$$d(x) = \int\limits_X k(x,y) d\mu(y) \tag{0.0.4}$$

$$P(x,y) = \frac{k(x,y)}{d(x)} \tag{0.0.5}$$

$$(D_t(x,y))^2 = \|P_t(x,\cdot) - P_t(y,\cdot)\|_{L^2(X,\frac{d\mu}{\pi})} \tag{0.0.6}$$

$$\pi(\mu) = \frac{d(y)}{z \in Z^{d(z)}} \tag{0.0.7}$$

$$\pi(x)p(x,y) = \pi(y)p(y,x) \tag{0.0.8}$$

$D_t(x,y)$ is the functionally weighted L^2 distance between the 2 posteriors $\mu \to P_t(x,u)$ and $\mu \to P_t(y,u)$. This is related to isoperimetry. Think about what happens as the cardinality of paths connecting x and y is increased. D_t can be computed using the eigenvalues of P.

$$D_t(x,y) = \sqrt{\sum_{\lambda \geq 1} \lambda_l (\phi_l(x) - \phi_l(y))^2}$$

We can define an embedding in Euclidian space via

$$\Psi_t(x) = \{\lambda_1^t \phi_1(x), ...\lambda_{s(\delta,t)}^t \phi_{s(\delta,t)}(x)\}$$

Spectral Geometry

Spectral Geometry concerns itself with the relationships between a geometric structure and the spectra of a differential operator, typically the Laplacian. Inferring the geometry from the spectra is a type of inverse problem since two non isometric manifolds may share the same spectra. Going the other way, we encounter isoperimetric inequalities and spectral gap theorems. "Can One Hear the Shape of a Drum?" was the of an article by Mark Kac in the American

Mathematical Monthly 1966. The frequencies at which a drum vibrate depends on its shape. The elliptic PDE $\nabla^2 A + kA = 0$ tells us the frequencies if we know the shape. These frequencies are the eigenvalues of the Laplacian in the region. Can the spectrum of the Laplacian tell us the shape if we know the frequencies? Hermann Weyl showed the eigenvalues of the Laplacian in the compact domain Ω are distributed according to $N(\lambda) \sim (2\pi)^{-d)\omega_d \lambda^{\frac{d}{2}} vol(\Omega}$

The Laplace Beltrami operator is the generalization of $\nabla \circ \nabla = \Delta$ to $\mathcal{M}$

$$\Delta f = tr(H(f))$$

In the exterior calculus we have $\Delta f = d^* d\, f$.

//BBCREVISIT - Fill this out and check The Laplacian of a Gaussian has well known applications in image processing. Given $f(x, y)$, we get a scale space representation when we convolve by

$$g(x, y, t) = \frac{e^{x^2+y^2}}{2\pi t}$$

$$L(x, y, t) = g(x, y, t) * f(x, y)$$

Applying Δ to $L(x, y, t)$ gives response to blobs of extent $\sqrt{t}$

There is a well known connection between diffusion processes and Schrodinger operators;

$$H = \nabla^2 + V(x)\Phi \in L^2(\mathbb{R}^n)$$

$$H\Phi = E\Phi$$

$$E = \sigma(H)$$

Concentration of Measure

?, ?, ?, ?, ?, ?, ?, ?.

Familiar tools used when when dealing with additive functions of independent random variables are the CLT, LLN, and the inequalities of Markov, Chebychev, and Chernoff. When the differences are not independent we rely on the theory of martingales and use inequalities like Azuma's to provide concentration bounds.

Let (X, Σ, μ) be a measure space and let f be an real-valued measurable function defined on X. Then for any number $t > 0 \in (R$ we have

$$\mu(x \in X | f(x) \geq t) < \frac{1}{t} \int_X |f(x)| d\mu$$

If we let μ be a probability measure - $\mu(X) = 1$, then the above is equivalent to $P(|X| \geq a) \leq \frac{E(|X|)}{a}$ commonly known as Markov's inequality.

Another familiar concentration inequality is the Chebychev inequality has it's origins as a measure theoretic inequality;

$$\mu(x \in X : |f(x)| \geq t) \geq \frac{1}{t^2} int_X \left|f\right|^2 d\mu$$

When X has finite first moment μ and non-zero second moment σ, we have the more familiar

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \; \forall k > 0$$

The isoperimetric inequality concerns the relationship between the length l of a closed curve and the area a of the planar region that it encloses. Specifically, $4\pi a \leq l^2$. Equality holds in the case that the curve is a circle. The isoperimetric problem is to determine a plane figure of the largest area whose boundary has a given length. F

Federer **?** is a good reference for a measure theoretic generalization to higher dimensions. We make a few remarks here that will be expanded on later.

Proposition 0.0.2 (Isoperimetric Inequality In $\mathbb{R}^n$)**.** *Let μ be Lebesgue measure in $\mathbb{R}$ and X in$\mathbb{R}$ $\ni$ $\mu(cl(X) < \infty$, then*

$$n\omega_n^{\frac{1}{n}} \mu(cl(X))^{\frac{n-1}{n}} \leq M^{n-1}(\partial X)$$

M is the Minkowski content, which is the Hausdorff measure of ∂X for rectifiable ∂X. The proof relies on the Brunn–Minkowski theorem which states that $\mu(A + B)^{\frac{1}{n}} \geq \mu(A)^{\frac{1}{n}} + \mu(A)^{\frac{1}{n}}$ where set addition in $\mathbb{R}$ is in the sense of Minkowski. This addition behaves well with respect to the convex hull; $\forall A, B \in \mathbb{R} \; co(A + A) = co(A) + co(A)$

For smooth domains general Isoperimetry inequality is equivalent to a Sobolev inequality on $\mathbb{R}$.

Proposition 0.0.3 (Sobolev inequality)**.** *Let u be a continuously differentiable real-valued function on $\mathbb{R}$ with compact support. Then for $1 \leq p < n$ there is a constant C depending only on n and p such that*

$$||u||_{L^{\frac{pn}{n-p}}} \leq C||Du||_{L^p}$$

The Sobolev embedding theorem relies on the

Theorem 0.0.4 (Hardy Littlewood Sobolev fractional integration theorem)**.** *Let $0 < \alpha < n$ and $1 < p < q < \infty$ and let $I\alpha = (-\Delta)\alpha/2$ be the Riesz potential*

$$I_\alpha f(x) = \frac{1}{C_\alpha} \int_{\mathbb{R}^n} \frac{f(y)}{|x-y|^{n-\alpha}} dy$$

Then, for $q = \frac{pn}{n-\alpha p}$ there exists a constant C depending only on p such that $||I_\alpha f||_q \leq C||f||_p$

The Hardy–Littlewood–Sobolev lemma implies the Sobolev embedding by the relationship between the Riesz transforms and the Riesz potentials. The Riesz potential defines an inverse for a power of the Laplace operator on Euclidean space.

The Chernoff and Hoeffding bounds tell us that the average of n iid random variables $X_1, X_2, \ldots, Xn$ is tightly concentrated around its mean if X_i are bounded and n is sufficiently large. hat about $G(X_1, X_2, \ldots, X_n)$? The feature of the average which gives rise to tight concentration is that is is Lipschitz. The following concentration bound applies to any Lipschitz function of iid normal random variables. See Ledoux (2001, page 41, 2.35).

High dimensional space is mostly empty. This is more commonly called the *"curse of dimensionality"*. One way to get around the curse of dimensionality is to find interesting projections. Many common algorithms such as principal components, multidimensional scaling, and factor analysis fall into this category. Huber **?** placed many of these in to a common framework called projection pursuit.

Logarithmic Sobolev inequalities have a close relationship with the concentration of measure phenomena. There are two major types of concentration; Gaussian and Exponential. [see Ledoux]

Let $(e^{-At})_{t\geq 0} = (T_t)_{t\geq 0}$ be a symmetric Markov semigroup on $L^2(X, d\mu)$ with generator A defined on a σ-finite measure space $(X, d\mu)$. $(T_t)_{t\geq 0}$ is ultracontractive if for any $t > 0$, there exists a finite positive number $a(t)$ such that, for all $f \in L^1$:

$$\|T_t f\|_\infty \leq a(t)\|f\|_1.$$

An equivalent formulation (by interpolation) of ultracontractivity is that for any $t > 0$, there exists a finite positive number $c(t)$ such that, $\forall f \in L^2$,

$$\|T_t f\|_\infty \leq c(t)\|f\|_2$$

Also by duality, the inequality () is equivalent to

$$\|T_t f\|_2 \leq c(t)\|f\|_1$$

It is known that, under the assumptions on the semigroup $(T_t)_{t\geq 0}$, () implies () with $a(t) \leq c^2(t/2)$ and () implies () with $c(t) \leq \sqrt{a(t)}$.

We say that the generator A satisfies LSIWP (logarithmic Sobolev inequality with parameter) if there exist a monotonically decreasing continuous function $\beta : (0, +\infty) \to (0, +\infty)$ such that

$$\int f^2 \log f \, d\mu \leq \epsilon Q(f) + \beta(\epsilon)\|f\|_2^2 + \|f\|_2^2 \log \|f\|_2$$

for all $\epsilon > 0$ and $0 \leq f \in \text{Quad}(A) \cap L^1 \cap L^\infty$ where $\text{Quad}(A)$ is the domain of $\sqrt{A}$ in L^2 and $Q(f) = (\sqrt{A}f, \sqrt{A}f)$.

This inequality is modeled on the Gross inequality .

In **?**,**?**, the authors show that LSIWP implies ultracontractivity property under an integrability condition on β. This condition can be enlarged and be stated as follows:

Theorem 0.0.5. *Let $\beta(\epsilon)$ be a monotonically decreasing continuous function of ϵ such that*

$$\int f^2 \log f \, d\mu \leq \epsilon Q(f) + \beta(\epsilon)\, \|f\|_2^2 + \|f\|_2^2 \log \|f\|_2$$

for all $\epsilon > 0$ and $0 \leq f \in Quad(A) \cap L^1 \cap L^\infty$. Suppose that for one $\eta > -1$,

$$M_\eta(t) = (\eta + 1)t^{-(\eta+1)}) \int_0^t s^\eta \beta \left(\frac{s}{\eta + 1} \right) ds$$

is finite for all $t > 0$. Then e^{-At} is ultracontractive and

$$\|e^{-At}\|_{\infty,2} \leq e^{M_\eta(t)}$$

for all $0 < t < \infty$.

The Condition Number of a Markov Chain

Generalized Chebyshev Bounds on Quadratic Sets via Semidefinite Programming

Boyd et al **?** provide a simplified development of an algorithm to compute the lower bound on the probability of a set which is defined by quadratic inequalities. That algorithm is discussed here.

$$\min(1 - \sum_{i=1}^{m} \lambda_i) \ni Tr(A_i z_i) + 2b_i^T z_i + c_i \lambda_i \geqslant 0 \quad \forall i = 1, ..., m$$

$$\sum_{i=1}^{m} [\begin{array}{cc} z_i & z_i \\ z_i & \lambda_i \end{array}] \succeq 0$$

$$C = \{x \in \mathbb{R} : x^T A_i x + 2b_i^T + c_i < 0 : i = 1, ..., m\}$$

$$\min E[f_0(X)] \ni E[f_i(X)] = a_i : i = 1, ..., m$$

moment constraints

Let

$$\bar{x} \in \mathbb{R}^n S \subset S^n \ni S \succeq \bar{x}\bar{x}^T$$

and define

$$P(C, \bar{x}, S) = inf_{\mathcal{P}(\mathbb{R}^n)} \{P(X \in C) \mid E[X] = \bar{x} E[XX^T] = S\}$$

The optimization problem is to find $P \in \mathcal{P}(\mathbb{R}^n)$ - a probability density function which maximizes the probability of the convex set C and satisfies the moment constraints.

Bregman Divergences

NNMA is the approximation of a non-negative matrix A by a low rank matrix BC where $B \succ 0$ and $C \succ 0$. Bregman divergences are a robust distortion measure for this matrix factorization. Formally $D_\phi(A, BC) = \phi(A) - \phi(BC) - \nabla\phi(BC)(A - BC)$ measures the quality of the factorization relative relative to a convex penalty function.

Modeling of relational data can be abstracted out to the factorization in a low dimensional representation of a data matrix (X_ij) where links [or relations] are represented as an nxm matrix X where $X_{i,j}$ indicates whether a relation exists between entities of type i, j. Let f be a link function and X be a factorization of X into a low rank approximation $X \approx UV^T : U \in R^{mxk}, v \in R^{mxk}$. The link function f can be interpreted as in GLM which gives extends exponential models to matrices. A simple example is choosing the identity link which and minimizing in the ℓ^2 norm gives rise to the SDV and the Gaussian model for the data X_ij. Similarly we can extend to Bernoulli, Poisson, Gamma, error distributions.

Many forms of matrix factorization can be cast as an optimization problem that involves minimization of generalized Bregman divergences **?**. Factorization algorithms such as NNMF, Weighted SVD, E xponential Family PCA, , pLSI, Bregman co-clustering **?** can be cast in this framework. The approach uses an alternating projection algorithm for solving the optimization problem which allows for generalizations that include row, column, or relaxed cluster constraints. A brief description of the algorithm is given below. The description of a generalized Bregman divergence can be found in **?**.

Let $\phi S \in \mathbb{R}^n \to \mathbb{R}$, $D_\phi(x, y) = \phi(x) - \phi(y) - < x - y, \nabla\phi(y) >$ be the Bregman divergence. Let $\chi = X_i \in S \in dblr^d$ be a random variable and take the encoding $X_i \to S$ so the rate is zero and the code book is 1. The rate distortion is $E_\nu[D_\phi(\chi, s)] = minSinS \sum_{i=1}^{n} \nu_i D_\phi(x_i, x)$ which we call the Bregman information for the random variable X.

Theorem 0.0.6. *$\mu = E_\nu[X]$ is the unique minimizer*

Sparse Representation

A Gaussian distribution is often an accurate density model for low dimensional data, but very rarely for high-dimensional data. High dimensional data is less likely to be Gaussian, because of the high degree of independence this demands. Recall the a Gaussian is a rotation of a distribution with completely independent coordinates. In a typical high dimensional application, one may be able to find a few features that are approximately independent, but generally as more features are added the dependencies between them will grow.

Diaconis and Freedman showed that for *most* high dimensional point clouds, *most* low dimensional orthogonal projections are a mixture of normal spherically symmetric distributions.

Lemma 0.0.7 (Poincare Lemma)**.** *If σ_n is uniform on $\sqrt{n}S_{n-1} \in \mathbb{R}^n$, $d < n$ and*

$$\Pi_{d,n}(x_1, \ldots, x_n) \to (x_1, \ldots, x_n)$$

is the canonical projection, then for fixed d, as $n \to \infty$, we have that $\Pi_{d,n}$ converges weakly towards a centered reduced Gaussian distribution on $\mathbb{R}^d$

Proof [See pp55 Some Aspects of Brownian Motion : Some Recent Martingale Problems]. Uee LLN. If $(X_1, X_2, \ldots, X_n)$ iid $N(0, 1)$, then

$$\frac{1}{n}\rho_n^2 =: \frac{1}{n}\sum_{i=0}^{n} x_i^2 \to 1 \to \infty$$

If we define $\tilde{X}_{(n)} = (X_1, X_2, \ldots, X_n) = \frac{1}{\sqrt{n}}\rho_n\theta_n$ where $\theta_n \sim \sigma_n$ a uniform distribution on $\sqrt{n}S_{n-1}$. Then the lemma follows from the equation $\tilde{X}_{(n)} = \frac{1}{\sqrt{n}}\rho_n\Pi_{d,n}(\theta_n)$.

Sparse PCA [CS Notes from http://ugcs.caltech.edu/ srbecker/

Classic PCA is sensitive to outliers. Convex methods can be use to address this by attempting to factor the data X as a sum of a low rank component L and a sparse component S by solving

$$\underset{min}{S, L} \left\| L \right\|_* + \lambda \left\| S \right\|_{\ell_1} : X = L + S$$

See http://cvxr.com/tfocs/demos/rpca/ for a demo of this in action with the popular cvx software package.

Compressed Sensing

Bound on Limiting Probabilities for a Perturbation of a Markov Chain

In **?** bounds are established on the relative error for the limiting probabilities for a perturbation of a finite Markov chain. This improves on the traditional eigenvector perturbation approach by exploiting the constraints of the problem. We collect some of these results in this section and see if they can be applied to analyse the stability of spectral clustering algorithms.

Let T be the transition matrix of a finite Markov chain $\mathcal{C}$. $A = I - T$, and $\mathcal{C}$ be a perturbation to $\mathcal{C}$ where $\tilde{T} = T - E$ is the transition matrix of $\mathcal{C}$. Let ω be the limiting probability, $\omega = \lim_{n \to \infty} T^n x$. Define $A = I - T$ and $A^\sharp$ the generalized inverse of A. Then $W = \lim_{n \infty} \frac{I+T+T^2+\ldots+T^{n-1}}{n} = I - AA^\sharp$ is the limiting matrix of $\mathcal{C}$. Every row of W is ω

We state a few relations without proof from Meyers paper

$$(A + E^\sharp) = A - A^\sharp E A^\sharp (I + EA^\sharp)^{-1} - W(I + EA^\sharp)^{-1} A^\sharp (I + EA^\sharp)^{-1}$$

This combined with the expression for $\tilde{W}$ yields

$$\tilde{W} = W(I + EA^\sharp)^{-1} = W - WEA^\sharp(I + EA^\sharp)^{-1}$$

The above gives a condition for the limiting matrix to be invariant under a perturbation; $W = \tilde{W} \iff range(E) \in range(A)$. It also allows us to write $\omega - \tilde{\omega} = \omega E A^\sharp (I + EA^\sharp)^{-1}$, so

$$\|\omega - \tilde{\omega}\| \leq \left\|EA^\sharp\right\| \left\|(I + EA^\sharp)^{-1}\right\|$$

and when $\left\|EA^\sharp\right\| \leq 1$ we can use the familiar Taylor expansion

$$\left\|(I + EA^\sharp)^{-1}\right\| \leq \frac{1}{1 - \|EA^\sharp\|}$$

to obtain an expression for the relative error in ω for a given relative error in A

$$\frac{\|\omega - \tilde{\omega}\|}{\|\omega\|} = \frac{\frac{\|E\|}{\|A\|}\kappa(\mathcal{C})}{1 - \frac{\|E\|}{\|A\|}\kappa(\mathcal{C})}$$

Ultracontractivity

Generalized Uncertainty Principles

The sparse recovery problem can be stated as follows. Given an unknown signal $f\ in \mathbb{C}^n$, when can we recover f from a set of k linear measurements $\Phi f\ in \mathbb{C}^k\ k < n$? This problem is underdetermined and we are interested in the sparsest solution.

$$min\, \|f*\|_0 : \Phi f* = \Phi f$$

This problem is not convex, and in fact is generally NP-Hard **?** and **?**. Note that the norm here is not the usual ℓ_0 norm, here we abuse notation $\|g\|_0$ to mean $card g_i ! = 0$, the size of the support. This problem can be relaxed to a convex problem by using the ℓ_1 norm.

$$min\, \|f*\|_1 : \Phi f* = \Phi f$$

This problem is not differentiable at points where $f*_i = 0$ and will require a general convex solver. Usually the problem is solved in practice by recasting the minimization as a linear program

$$min < f*, f >: \Phi f* = f\ f* \succeq 0$$

bbcrevisit

Much work has been done to determine when these two problems have the same solution, i.e. when is exact recovery possible. One sufficient condition for exact recovery is the RIP condition of Candes and Tao (see below).

Definition 0.0.4 (RIP (Candes & Tao))**.** Let $A \in M_{m\times n}(\mathbb{R})$ and $p \in [1, n]$, then we say that A has the restricted isometry property if

$$\exists \delta_p \ni \forall m, pA_s \in A$$
$$(1-\delta_p)\|y\|_{\ell^2}^2 \leq \|A_s y\|_{\ell^2}^2 \leq (1+\delta_p)\|y\|_{\ell^2}^2 \leq$$

RIP is a property which classifies a matrix as being close to orthogonal. δ_p is referred to as the RIC, and is NP-Hard to compute. Compressed sensing decoders are guaranteed to recover the sparsest solution to $Y = Ax$ when A is close to an isometry. Bounds on the RIC are available for some classes of random matrix ensembles.

Random Matrix Theory & OP

RMT concerns itself with the eigenvalue statistics of large matrices with random entries. We define the eigenvalue counting measure of a matrix Hn as;

$$\mu_H(A) = \frac{|\lambda_i \in A|}{n} = N_{1_A,H}$$

More generally a eigenvalue statistic $N_{f,H} = \frac{trf(H)}{n}$ For many types of random matrices we have a CLT;

$$\frac{N_{f,H} - \int f(\lambda)dN(\lambda)}{\sigma_{f,n}} \to_d N(0,1)$$

$GUE(n)$ Start with Gaussian measure

$$\gamma^n(A) - \frac{1}{(\sqrt{2\pi})^n}\int_A e^{\frac{-1}{2}}\left\|X\right\|^2 d\lambda^n$$

Weiener space is the Hilbert space $L^{2,0}[0,1]$ of upon which a Gaussian measure can be defined. The inner product on Weiner space is

$$<\sigma_i, \sigma_j> = \int\limits_0^1 <\dot{\sigma_i}, \dot{\sigma_j}> dt$$

where $\dot{\sigma_i} = \frac{d\sigma)i}{dt}$ The Wiener measure is a Gaussian measure.

Let H be Hermitian, $Z_{GUE(n)} = 2^{\frac{n}{2}}\pi^{\frac{n^2}{2}}$ We can write

$$\gamma^n(A) - \frac{1}{Z_{GUE(n)}}\int_A e^{\frac{-1}{2}}tr(H^2)d\mu$$

There are two domains of eigenvalue statistics, local and bulk. Locally we are concerned with level spacing ,$\Delta\lambda = \lambda_i - \lambda_{i-1}$, and edge statistics $P_{TW}(\lambda_1)$, $\underset{n\to\infty}{lim} P_{TW}(\lambda_n)$. The Tracy Widom law P_{TW} gives edge statistics.

The empirical spectral measure of H is

$$\mu_H = \frac{|eigsH \in A|}{n}$$

The CDF of $H(n)$ is $N_n(\lambda)$ and as $n \to \infty$ is $N_n(\lambda) \to W(\lambda)$ where W is the Winger Semi-Circle Law.

Bulk statistics look at the determinantal point process $E(\lambda_0) = \sum_j \delta(n\rho(\lambda_0(\lambda_j - \lambda_0)$. The kernel of the process is the sine kernel $K(x,y) = \frac{sin(\pi(x-y))}{\pi(x-y)}$. $E(\lambda)$ captures the statistics of the eigenvalues in the vicinity of λ_0. Recall the joint densities of a determinantal point process with kernel K are given by $\rho_n(x_1, ..., x_n) = det(K(x_i, x_j)_{1\leq i,j \leq j})$

The Hamburger Moment Problem - HMP

Given a sequence m_i the HMP seeks to find a measure that generates the moments.

$$\exists \mu : m_n = \int\limits_{-\infty}^{+\infty} x^n d\mu$$

The answer is affirmative when the Hankel kernel $A \succ 0$, which is equivalent to $\sigma(A) \in \mathbb{R}^+$

$$A = \begin{pmatrix} m_0 & m_1 & \dots & \\ m_1 & m_2 & m_3 & \\ m_2 & m_3 & m_4 & \\ \vdots & & & \ddots \end{pmatrix}$$

Recall that A Hankel matrix is a square matrix with constant skew-diagonals; $A_{i,j} = Ai-1, j+1$ A Hankel matrix is an upside-down Toeplitz matrix.

The Hilbert matrix $Hi, j = \frac{1}{i+j-1}$ is a special case of a Hankel matrix. The Hilbert matrices are canonical examples of ill-conditioned matrices. They arise in the expression for the Grammian matrix of powers of x; $H_{ij} = \int 01x^{i+j-2}dx$ which shows up in the least squares approximation by polynomials.

The solutions to the HMP are either unique or infinite in number and form a convex set in $\mathcal{H}$

Let

$$\Delta_n = \begin{pmatrix} m_0 & m_1 & \dots & m_n \\ m_1 & m_2 & \dots & m_{n-1} \\ & & \ddots & \\ m_n & m_{n+1} & & m_{2n} \end{pmatrix}$$

, then $A \succ 0 \Rightarrow det(\Delta_n) \geq 0 \forall n$ If $det(\Delta_n) = 0$ then $((H), <, >)$ is finite dimensional and T is self adjoint.

A gives us a sesquilinear form on $(H) = (l)^2$ via $< x, y >= \hbar x^T Ay$.

Let T be a shift operator. The HMP is closely related to OP in $\mathbb{R}$. Gram Schmidt gives basis ϕ_i in which T has tridiagonal Jacobi.

The Caley transform $Q(T) = (I - T)(I + T)^{-1}$ shows the connection to the Nevanlina class of functions (sub-harmonic).

Random Matrix Ensembles

?, ?, ? , ?, ?, ?

The classical ensembles of random matrix theory are GOE, GUE, GSE, Wishart, and MANOVA. These correspond to the weight functions of the equilibrium measure of the orthogonal polynomials Hermite, Laguerre,and Jacobi. The Jacobians of the well known matrix factorizations are used to compute the joint eigenvalue densities of these ensembles. The distribution of eigenvalues of the GOE ensemble follow the well know Winger Semi-circle distribution. The joint densities up to a constant factor are listed below:

- Hermite
- Laguerre
- Jacobi

We generated histograms in Matlab for samples from the GOE, GUE, GSE, Wishart, and MANOVA ensembles. The joint PDF of a generic Gaussian ramdom matrix is given by,

$$P(M) = G_\beta(n, m) = \frac{1}{2\pi^{\frac{\beta nm}{2}}} \exp^{\frac{-1}{2}\|M\|_F}$$

where β encodes the dimension of the field. Note this leaves open the possibility to generalize to non integer β.

The table below describes how to generate from the common ensembles starting from a sample $A \in G_\beta(n, n)$

$$GOE\{M|M = \frac{A + A^T}{2}, A \in G_1(n, n)\}$$
$$GUE\{M|M = \frac{A + A^\dagger}{2}, A \in G_2(n, n)\}$$
$$GSE\{M|M = \frac{A + A^\ddagger}{2}, A \in G_4(n, n)\}$$

The CS decomposition is a matrix factorization equivalent to four SVD's which correspond to rotation problems $\begin{pmatrix} X \to Y & X^\perp \to Y \\ X \to Y^\perp & X^\perp \to Y^\perp \end{pmatrix}$ Which can be compactly written $[X|X^\perp]^T[Y|Y^\perp] = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} * \begin{pmatrix} C & S \\ -S & C \end{pmatrix} * \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}$ Where U, S are unary.

The Tracy-Widom law of order one is the limiting distribution of the largest eigenvalue of a Wishart matrix with identity covariance when properly scaled. This has some application to weighted directional graphs. The largest eigenvalue of the adjacency matrix of a random d-regular directed graph follows the Tracy-Widom law. The kernels of integrable operators describe the asymptotic eigenvalue distribution of self-adjoint random matrices from the unitary ensembles. Consider the discreet operator $K(n, m) : {}^2(N) \to {}^2(M)$ where $K(n, m) = \frac{(<Ja(m), a(n)>}{m-n}$ the discrete Bessel kernel and kernels arising from the almost Mathieu equation. The celebrated paper of Tracy and Widom **?** investigated integral kernels of the form

$$K(x, y) = \frac{f(x)g(y) - f(y)g(x)}{x - y} : x \neq y f(x), g(x) \in L^2(0, \infty)$$

are solutions to the system of ODE's

$$\frac{d}{dx}\begin{pmatrix} f(x) \\ g(x) \end{pmatrix} = \begin{pmatrix} \alpha(x) & \beta(x) \\ -\gamma(x) & -\alpha(x) \end{pmatrix} * \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$$

Let $\phi_i(x)$ be an orthogonal basis in a Hilbert Space $\mathcal{H}$ where

$$\Gamma_\phi = \{\phi_(j + k - 1)\}_{j,k=1}^{\infty}$$

is the induced Hankel Matrix.

Let $(L) : \mathcal{H} \to \mathcal{H}$ be compact. Then $(L) : f \mapsto \sum_{n=1}^{N} \omega_n < \phi_n, f > \psi_n$ where $\{\phi_i\}_{i=1}^{N}$

The Tracy Widom Law

The Tracy-Widom distribution is related to to determinantal stochastic processes. A process following this law is distributed as the largest point of a point process on the real line where the kernel K is the so-called Airy kernel. In addition to describing the edge spectrum of random matrices, it arises in several place in combinatorial for instance the longest increasing subsequences of random permutations is described by the Tracy Widom law. In addition to the eigenvalues of random matrices, this type of point process is used in models such as fluctuations in first and last passage percolation, and the asymmetric exclusion process.

The path configuration of random viscous walkers is related to the Young tableaux. Statistical problems related to the Young tableaux include random growth, point processes, random permutation, and the random word problem. the asymptotic distribution of scaled variables from these models are described by the Tracy Widom distribution which is the limit distribution for the largest eigenvalue of $X \in GUE$.

Nyström Method

The Nystrom method is a technique to speed up large-scale learning applications by generating low-rank approximations to large matrices. The origins of the Nystrom method lie in the numerical solution of the integral eigenvalue problem

$$\int p(y)k(x,y)\phi(y) = \lambda\phi(y)$$

(BBCREVISIT talk about the functional analysis and physics origin of this method for solving integral problems)

Problems in computer vision, natural language processing, computational biology and other areas can involve data sets containing more training examples than can be held in memory. Low rank approximations to the kernel matrices found in kernel based machine learning algorithms such as spectral clustering, manifold learning,support vector machines offer the opportunity to train on large data sets. The performance of this technique relies on the ability of a matrix to be approximated by a subset of its columns. Coherence bounds from the field of compressed sensing and matrix completion can be used to bound the error of the Nystrom approximation. **?**

The Nystrom method can be evaluated by comparing the approximation to a low rank approximation generated from the first n singular vectors from the SVD. If the process generating the data is $N(\mu, \Sigma)$ random sampling can be used and exact error bounds can be calculated based on the number of columns sampled. (BBCREVISIT confirm and cite).

Choosing the columns based on the underlying probability density $p(y)$ has the potential to improve the quality of the approximation **?**.

Statistical Leverage

Michael Mahoney - videolectures.net Statistical Leverage Given an m x n matrix A and a rank parameter k, define the leverage of the i-th row of A to be the i-th diagonal element of the projection matrix onto the span of the top k left singular vectors of A. In this case, ”high leverage” rows have a disproportionately large amount of the ”mass” in the top singular vectors. Historically, this statistical concept (and generalizations of it) has found extensive applications in diagnostic regression analysis. Recently, this concept has also been central in the development of improved randomized algorithms for several fundamental matrix problems that have broad applications in machine learning and data analysis. Two examples of the use of statistical leverage for improved worst-case analysis of matrix algorithms will be described. The first problem is the least squares approximation problem, in which there are n constraints and d variables.

Classical algorithms, dating back to Gauss and Legendre, use O(nd2) time. We describe a randomized algorithm that uses only O(n d log d) time to compute a relative-error, i.e., 1+/-epsilon, approximation. The second problem is the problem of selecting a ”good” set of exactly k columns from an m x n matrix, and the algorithm of Gu and Eisenstat provides the best previously existing result. We describe a two-stage algorithm that improves on their result. Recent applications of statistical leverage ideas in modern large-scale machine learning and data analysis will also be briefly described. This concept has proven to be particularly fruitful in large data applications where modeling decisions regarding what computations to perform are made for computational reasons, as opposed to having any realistic hope that the statistical assumptions implicit in those computations are satisfied by the data.

How are differential operators related to semi group

Sobolev inequalities relate norms of Sobolev spaces and can be used to prove embedding of $W^{k,p}(\mathbb{R}^n) \in W^{l,q}(\mathbb{R}^n)$ Logarithmic Sobolev inequalities. The Rellich–Kondrachov theorem makes this precise for the L^p spaces. A is an operator on a L^p Banach space. We define this in an operator sense as the sum given by the Taylor expansion of $e^x = \sum \frac{x^n}{n!}$

$$e^{-tA} : L^p \to L^q$$

is contractive if this holds for $p = 2$ and $q = 4$. e^{-tA} is a positive preserving contractive semi-group if

$$\begin{aligned} e^{-tA} f &\geq 0 \\ \parallel e^{-tA} f \parallel_p &\leq \parallel f \parallel_p \end{aligned}$$

When e^{-tA} maps L^2 to L^∞ we call the semi group untracontractive.

Covering number

Union bound

Kashin representation

Candes and Tao extend Uncertainty Principle to discrete domain For $x \in \mathbb{C}^n$

$$|supp(x)||supp(\hat{x}) \geq N$$

Uniform uncertainty principle is related to the restricted isometry property.

The Λ problem of Talingrad relates Fourier matrix to HUP question. Consider matrices which are euclidian projection of cube $Q^N = x : \parallel x \parallel_\infty \leq 1$, a problem form geometric functional analysis. Let $B^n = x : \parallel x \parallel_2 \leq 1$

Kashin; $\exists$ orthogonal projection P

$$P : Q^N \rightarrow U \in \mathbb{C}^N \ni U \approx B^n$$

This implies

$$\exists A \in M^{n,N} \ni B^n \in \frac{K}{\sqrt{N}} AQ^N \in KB^n whereK = K(\lambda)\lambda = \frac{N}{n}$$

λ is the redundancy. This statement is essentially telling us that we can project the cube into the ball when N and n allow for it.

Learning with integral operators

Machine learning methods such as Principal Components Analysis (PCA), Laplacian-based spectral clustering, and manifold methods rely on estimating eigenvalues and eigenvectors of the underlying data-dependent matrices. These matrices can be interpreted as empirical versions of underlying integral operators or closely related objects, such as continuous Laplacian operators. Establishing a connection between the finite dimensional empirical operator represented by the data matrix and the continuous counterpart can yield insight, help with convergence theorems, and even provide the basis for new machine learning methods.

Theoretical analysis of many problems, such as low-rank matrix recovery, covariance estimation and approximate matrix multiplication, is built upon exponential bounds for $P(\| \sum X_i \| > t)$ where X_i is a finite sequence of self-adjoint random matrices and k · k is the operator norm. Moment-generating function techniques can be used to produce generalizations of Chernoff, Bernstein and Friedman inequalities to the non-commutative case.

Concentration inequalities extend classical exponential inequalities for sums of independent random variables to functions of independent random variables. In machine learning theory we are interested in functions such as the supremum of the deviations between the true risk and the empirical risk, the empirical VC-dimension, the empirical VC entropy, and the eigenvalues of the Gram matrix. $E[\hat{\theta} - \theta_0]$

High Dimensional Probability Topics

Verysin's Book :

High-dimensional probability offers insight into the behavior of random vectors, random matrices, random subspaces, and objects used to quantify uncertainty in high dimensions. Drawing on ideas from probability, analysis, and geometry, it lends itself to applications in mathematics, statistics, theoretical computer science, signal processing, optimization, and more. It is the first to integrate theory, key tools, and modern applications of high-dimensional probability. Concentration inequalities form the core, and it covers both classical results such as Hoeffding's and Chernoff's inequalities and modern developments such as the matrix Bernstein's inequality. It then introduces the powerful methods based on stochastic processes, including such tools as Slepian's, Sudakov's, and Dudley's inequalities, as well as generic chaining and bounds based on VC dimension. A broad range of illustrations is embedded throughout, including classical and modern results for covariance estimation, clustering, networks, semidefinite programming, coding, dimension reduction, matrix completion, machine learning, compressed sensing, and sparse regression.

- Concentration of sums of independent random variables - Random vectors in high dimensions - Random matrices - Concentration without independence - Quadratic forms, symmetrization and contraction - Random processes - Chaining - Deviations of random matrices and geometric consequences - Sparse recovery - Dvoretzky-Milman's theorem

WDA - Wasserstein Discriminant Analysis

Rémi Flamary, Marco Cuturi, Nicolas Courty, Alain Rakotomamonjy (Submitted on 29 Aug 2016 (v1), last revised 23 May 2018 (this version, v2)) Wasserstein Discriminant Analysis (WDA) is a new supervised method that can improve classification of high-dimensional data by computing a suitable linear map onto a lower dimensional subspace. Following the blueprint of classical Linear Discriminant Analysis (LDA), WDA selects the projection matrix that maximizes the ratio of two quantities: the dispersion of projected points coming from different classes, divided by the dispersion of projected points coming from the same class. To quantify dispersion, WDA uses regularized Wasserstein distances, rather than cross-variance measures which have been usually considered, notably in LDA. Thanks to the underlying principles of optimal transport, WDA is able to capture both global (at distribution scale) and local (at samples scale) interactions between classes. Regularized Wasserstein distances can be computed using the Sinkhorn matrix scaling algorithm; We show that the optimization of WDA can be tackled using automatic differentiation of Sinkhorn iterations. Numerical experiments show promising results both in terms of prediction and visualization on toy examples and real life datasets such as MNIST and on deep features obtained from a subset of the Caltech dataset.

Chapter 1

Appendix Statistics

Concepts from Mathematical Statistics

The sampling distribution of an estimator is the probability distribution of the estimator under repeated sampling. The standard error of a measurement is essentially the standard deviation of the process by which the measurement was generated. When the underlying probability distribution of the generating process is known the standard error can be used to calculate confidence intervals. Otherwise Chebyshev's inequality can be used. The standard error of a sample from a population is the standard deviation of the sampling distribution and may be estimated as $\frac{\sigma}{\sqrt{n}}$

Completeness is the ability of a statistical estimator to ensures that the parameters of the underlying probability distribution representing the model can be estimated from the statistic.

Consider a family of probability distributions for a random variable X parameterized by a parameter θ. A quantity $T(X)$ that depends on the random variable X but not on the parameter θ is called a statistic. A statistic that captures all of the information in X that is relevant to the estimation of θ is called a sufficient statistic. Since the conditional distribution of X given $T(X)$ does not depend on θ, neither does the conditional expected value of $g(X)$ given $T(X)$. The conditional expected value is itself a statistic and so is available for use in estimation. If $g(X)$ is an estimator of θ, then typically the conditional expectation of $g(X)$ given $T(X)$ is a better estimator of θ. The Rao-Blackwell theorem makes this precise.

L-Estimators & M-Estimators

Order statistics of a sample X_i can be used to estimate quantiles. An L statistic is a linear combination of an order statistic. L statistics are used in Box plots. An M-Estimator is the minima of sums of functions of the sample data. Maximum Likelihood and Least Squares are examples of M-Estimators.

Testing for normality and other distributions

Powerful inference methods can be employed when data is generated by a Gaussian process. This section describes techniques for testing the normality of a sample and comparing two samples. Kolmogorov-Smirnov test uses the fact that the empirical cumulative distribution function is normal in the limit. It is a non-parametric and distribution free test. Given the empirical distribution

$$F_n(x) = \frac{1}{n} \sum_{n}^{i=1} \left\{ \begin{array}{l} 1 : x_i \leq x \\ 0 : x_i > x \end{array} \right.$$

, and a test CDF

$$F(x)$$

the K-S test statistics are $D_n^+ = max(F_n(x) - F(x))$ and $D_n^- = min(F_n(x) - F(x))$ The generality of this test comes at a loss in precision near the tails of a distribution. The K-S statistics are more sensitive near points close to the median, and are only valid for continuous distributions. The Kuipers test uses the statistic $D_n^+ + D_n^-$ and is useful for detecting changes in time series since the statistic is invariant in ???? transformation of the dependent variable F_n. The Anderson-Darling test is based on the K-S test and uses the specific distribution to specify the ????critical values??? of the test. The chi-squared is based on the sample histogram and allows comparison against a discrete distribution, but has the potential drawback of being sensitive to how the histogram is binned and requires more samples to be valid. The Shapiro-Wilk test uses the expected values of the order statistics of $F(x)$ to calculate the test statistic. It is sensitive to data that are very close together, and numerical implementations may suffer from a loss of accuracy for large sample sizes.

K-S [Chakravarti, Laha, and Roy, (1967). Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, pp. 392-394]. Shapiro-Wilk [Shapiro, S. S. and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)", Biometrika, 52, 3 and 4, pages 591-611.]

Regression Methods

Standard least squares regression consists in fitting a line through the data points (training points in learning theory) that minimizes the sum of square residuals. The underlying assumption is that the data and the response can be modeled by a linear relationship. In the event that the model accurately captures the functional dependence of the response generated by the data, and under the assumptions that the data is corrupted by Gaussian noise, precise statistical inferences can be made on the model parameters. Modifications to this standard model include nonlinear mapping of the input data, local fitting, biased estimators, subset selection, coefficient shrinking, weighted least squares, and basis expansion transformations.

The multiple regression model in matrix notation can be expressed as (BBCREVISIT)

These equations hold for the univariate and the multivariate case.

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$.

where $\mathbf{e} =_d N(\mu, \mathbf{\Sigma})$

$E\mathbf{Y} = \mu X\beta$

$var\mathbf{Y} = \sigma^2\mathbf{I}$

Maximum Likelihood and Least Squares give same estimator;

$\widehat{\beta} = (X^TX)^{-1}X^TY$

$\hat{\sigma}^2 = frac1n||Y - \hat{\mu}||^2$

Multivariate regression is the extension to $YM = Xb + e$

Here Y, X, b, and e are as described for the multivariate regression model and M is an mxs matrix of coefficients defining linear transformation of the dependent variables. The normal equations are $X'Xb = X'YM$ and a solution for the normal equations is given by $b = (X'X) - X'YM$. Here the inverse of $X'X$ is a generalized inverse if $X'X$ contains redundant columns.

Generalized Linear Models

Suppose we have n observations of k dimensional data denoted $\{x_i\}_{i=1}^{k}$ and for each observation we have a response y_i. We wish to fit the observations to the responses. Generalized Linear Regression is a modeling technique that allows for non normal distributions and models non-linear relationships in the training data. M-estimators are used to fit a generalized linear model Ref Huber (1964). A linear model $Y = \Lambda(X) = X\beta + \epsilon$ fits a linear relationship between the dependent variables Y_i and the predictor variables X_i

$$Y_i = \Lambda(X_i) = b_o + b \circ X_i. \tag{1.0.1}$$

A generalized linear model $Y = g(\Lambda(X)) + \epsilon$ fits the data to $Y = g(X \circ W)$. Fitting the model consists of minimizing the objective function $\sum_{i=1}^{n} g(e_i) = \sum_{i=1}^{n} g(y_i - x_i\beta)$, where e_i are the residuals $y_i - x_i\beta$. We see that for ordinary least squares $g(e_i) = e_i^2$, and the usual matrix equations fall out by differentiating with respect to β. Carrying this out for general g

$$\sum_{i=1}^{n} \frac{\partial g(y_i - x_i\beta)}{\partial \beta} = 0 \tag{1.0.2}$$

gives the system of $k + 1$ equations to solve for estimating the coefficients b_i. If we set$\alpha(x) = \frac{g'(x)}{x}$ and calculate the derivative above, we have to solve

$$\sum_{i=1}^{n} \omega(e_i)(y_i - x_i\beta)x_i = 0 \tag{1.0.3}$$

. Which gives rise to a weighted least squares where the weights depend on the residuals - which depend on the coefficients - which depend on the weights. This suggests an iterative algorithm;

$$\beta^{\tau} = (X^t W^{(\tau-1)} X)^{-1} X^t W^{\tau-1} y \tag{1.0.4}$$

where $W_{ij}^{(\tau-1)} = \alpha(e_i^{(\tau-1})$. Several parameterizations are popular for the exponential family. The most general form of the distribution

$$p(x, \theta) = f(x, \theta) e^{g(x,\theta)} \in C^2(\mathbb{R} \otimes \mathbb{R}) \otimes C^2(\mathbb{R} \otimes \mathbb{R})$$

. The estimators derived below assume that f and g are separable,

$$p(x, \theta) = f(x) h(\theta) e^{\alpha(x)\beta(\theta)} \in C^2(\mathbb{R}) \otimes C^2(\mathbb{R}) \otimes C^2(\mathbb{R}) \otimes C^2(\mathbb{R})$$

. From

$$\int\limits_{x=-\infty}^{x=+\infty} p(x, \theta) dx \ = 1$$

we get

$$\frac{d}{d\theta}\, p(x, \theta) = 0 = \frac{d^2}{d^2\theta}\, p(x, \theta) d$$

Since the parametrization we have chosen for the exponential family allows, in the sequel we drop the notation for dependent variable and denote the derivative with a prime.

$$\frac{d}{d\theta}\, p(x, \theta) = \frac{d}{d\theta}\, f h e^{\alpha\beta} = h' f e^{\alpha\beta} + f h \alpha \beta' e^{\alpha\beta} = \big(\frac{h'}{h} + \alpha\beta'\big) p(x, \theta)$$

which gives

$$\int \frac{d}{d\theta} p(x, \theta) dx = \int \big(\frac{h'}{h} + \alpha\beta'\big) p(x, \theta) dx = \frac{h'}{h} \int p(x, \theta) dx + \beta' \int \alpha(x) p(x, \theta) dx = \frac{h'}{h} + \beta' E[\alpha(x)]$$

so that

$$E[\alpha(x)] = -\frac{h'}{h\beta'}$$

. Continuing along this vein,

$$0 = \int \frac{d^2}{d^2\theta}\, p(x,\theta)dx = \int \frac{d}{d\theta}\,\big(\frac{h'}{h} + \alpha\beta'\big)p(x,\theta)dx =$$
$$\int \big(\frac{h''}{h} - \frac{(h')^2}{h^2} + \alpha\beta''\big)p(x,\theta) + \big(\frac{h'}{h} + \alpha\beta'\big)\frac{d}{d\theta}\, p(x,\theta)\; dx =$$
$$\int \big(\frac{h''}{h} - \frac{(h')^2}{h^2} + \alpha\beta''\big)p(x,\theta) + \big(\frac{h'}{h} + \alpha\beta'\big)^2 p(x,\theta)\; dx =$$
$$\int \big(\frac{h''}{h} - \frac{(h')^2}{h^2} + \alpha\beta''\big)p(x,\theta) + \big(\frac{h'}{h} + \alpha\beta'\big)^2 p(x,\theta)\; dx =$$
$$\int \big(\frac{h''}{h} - \frac{(h')^2}{h^2} + \alpha\beta''\big)p(x,\theta) + (\alpha\beta' - E[\alpha(x)]\beta')^2 p(x,\theta)\; dx$$

Keeping in mind that

$$Var[ax] = E[(ax - E(ax)^2] = a^2 E[(x - E[x])^2] = a^2 Var[x]$$

we get the variance via

$$\big(\frac{h''}{h} - \frac{(h')^2}{h^2} + E[\alpha(x)]\beta''\big) + Var[\alpha(x)\beta'(\theta)] = \big(\frac{h''}{h} - \frac{(h')^2}{h^2} + E[\alpha(x)]\beta''\big) + (\beta')^2 Var[\alpha(x)] = 0$$

The score $U(x)$ is given by

$$U(x) = \frac{\partial}{\partial\theta}L(\theta,x) = \frac{\partial}{\partial\theta}\log p(x,\theta) = \frac{\partial}{\partial\theta}\big(\log h(\theta) + \log f(x) + \alpha(x)\beta(\theta)\big) = \frac{h'}{h} + \alpha\beta'$$

so

$$E[U(x)] = \beta' E[\alpha(x)] + \frac{h'}{h} = 0$$

. The Fisher Information $\mathcal{F}$ is defined

$$\mathcal{F} = Var[U(x)] = Var[\alpha\beta' + \frac{h'}{h}] = Var[\alpha\beta']$$

So from above we have

$$Var[U(x)] = Var[\alpha\beta'] = \big(-\frac{h''}{h} + \frac{(h')^2}{h^2} - E[\alpha(x)]\beta''\big)$$

. Now differentiating,

$$\frac{d}{d\theta}\, U(\theta,x) = \frac{h''}{h} - \frac{(h')^2}{h^2} + \alpha\beta''$$

$$E[U'(\theta,x)] = \frac{h''}{h} - \frac{(h')^2}{h^2} + E[\alpha]\beta'' = \frac{h''}{h} - \frac{(h')^2}{h^2} - \frac{\beta'' h'}{\beta'} = -Var[U(x)]$$

. Note that if we write the parametrization of the separable exponential family as

$$p(x,\theta) = e^{\alpha(x)\beta(\theta)+\log(f(x))+\log(h(\theta))}$$

then,

$$\frac{d^2}{d^2\theta}\log(h(\theta)) = \frac{d}{d\theta}\frac{h'}{h} = \frac{h''}{h} - \frac{(h')^2}{h^2}$$

. (BBCREVISIT - this duplicates and has notation clash with material above).

A general form of the exponential distribution

$$\rho(x;\theta) = exp(\frac{x\theta - \xi(\theta)}{\sigma})\nu(x) \tag{1.0.5}$$

has a log likelihood for a random sample $\{X_i\}_{i=1\ldots N}$ given functionally by

$$\mathcal{L}(\theta) = \sum_{i=1}^{N}[X_i\theta - \xi(\theta) + log(\nu(X_i))] \tag{1.0.6}$$

The scale parameter σ and θ are orthogonal parameters in that E [ ] The Generalized Linear model can ρ' is referred to as a link function in the statistical literature. If $\rho'(x) = x(1)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are iid $N(\mu, \sigma)$ we have multiple linear regression. In classification problems or binomial models the logit $\rho'(x) = log(x/(1-x))$ link function is used. The logit is extended to the k category case by

$$\rho'(x_i|x_j j \neq i) = log(\frac{x_i}{1 - \sum_{j\neq i} x_j}) \tag{1.0.7}$$

. The posterior probability densities $p_i(?)$ bbcrevisit (or p_i the probability of observing class i) of k classes are modeled by linear functions of the input variables x_i.

Fitting the GLM

Iteratively re-weighted least squares (IRLS) is used to for fitting generalized linear models and in finding M-estimators. The objective function

$$J(\beta^{i+1}) = argmin \sum w_i(\beta)|y_i - f_i(\beta)| \tag{1.0.8}$$

is solved iteratively using a Gauss-Newton or Levenberg-Marquardt (LM) algorithm. LM is an iterative technique that finds a local minimum of a function that is expressed as the sum of squares of nonlinear functions. It is a combination of steepest descent and the Gauss-Newton method. When the current solution is far from the minimum the next iterate is in the direction of steepest descent. When the current solution is close to the minimum the next iterate is a Gauss-Newton step.

Linear least-squares estimates can behave badly when the error is not normal. Outliers can be removed, or accounted for by employing a robust regression that is less sensitive to outliers than least squares. M-Estimators introduced by Huber generalize maximum likelihood estimation and are less biased and more efficient. Instead of trying to minimize the log likelihood

$$L(\theta) = \sum -log(p(x_i, \theta) \tag{1.0.9}$$

Huber proposed minimizing

$$M(\theta) = \sum \rho(x_i, \theta) \tag{1.0.10}$$

where ρ reduces the effect of outliers. Common loss function are the Huber, and Tukey Bisquare. For $\rho(x) = x^2$ we have the familiar least squares loss.

M estimators arise from the desire to apply Maximum Likelihood Estimators to noisy normal data, and to model more general distributions. They provide a regression that is robust against outliers in the training set, and allow for modeling of non-Gaussian processes. When ρ above is a probability distribution, we are preforming a maximum likelihood estimation.

The Huber function which is a hybrid L^2 L^1 norm

$$\rho_\eta(e_i) = \begin{cases} \frac{e_i^2}{2} & |e_i| \leq \eta \\ \eta|e_i| - \frac{\eta^2}{2} & |e_i| > \eta \end{cases} \tag{1.0.11}$$

The Tukey Bisquare estimator is given by

$$g_\eta(e_i) = \begin{cases} \frac{\eta^2}{6}(1 - [1 - \frac{e_i}{\eta}{}_2]^3) & |e_i| \leq \eta \\ \frac{\eta^2}{6} & |e_i| > \eta \end{cases} \tag{1.0.12}$$

Numerical procedures for doing this calculation are the Newton-Raphson method [see the section on root finding below], and Fisher-Scoring method [replace $\frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta\partial\theta^t}$ with $E[\frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta\partial\theta^t}]$. For high dimensional data, many models may be fit in an attempt to find the simplest one that can explain the data. In the language of statistical learning theory, the choice of a norm ρ is tantamount to choosing a loss function. Restricting the admissible functions to the one parameter family of exponential probability distributions defines the capacity via a functional form of the law of large numbers. **?**

Feature Subset Selection (FSS)

The goal of feature selection techniques to to improve the model building process by eliminating features that do not have discriminative power. Algorithms for feature selection either rank features or create subsets of increasing optimality. FSS should be contrasted with feature extraction techniques such as PCA, LLE, or Laplacian eigenmaps. The goal of feature extraction is to transform data from a high dimensional space to a low dimensional one while preserving the relevant information.

The statistical approach to feature selection most commonly used is stepwise regression. Common optimality criteria are FSS schemes the Kolmogorov-Smirnov Test ,the t-test, the f-test, the Wilks Lambda Test and Wilcoxon Rank Sum Test.

It's important to distinguish the FSS process from a data dimension reduction process such as PCA which requires all the original measurements to compute the projection. The better FSS algorithms are recursive

Construct a pxM basis matrix H^T and transform feature vector $x' = H^T x$.

Generalize to L^2 with smoothing splines

Smoothing spline $RSS(f, \lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$. where $f \in C^2(\mathbb{R})$ This is minimized in L^2 the first term measuring closeness of fit, and the second term penalizes curvature. $\lambda \to 0$ gives any function interpolating the data points $x_{i i \in 1,\ldots N}$ an $\lambda \to \infty$ constrains f to be linear.

Longitudinal Data Analysis

Longitudinal data analysis is the observation of multiple subjects over repeated intervals. Binary repeated responses are typically modelled with a marginal or random effects model, which will be made precise below. Marginal Models are a generalization of the GLM presented above for correlated data. Here, the correlation is inter subject across time. Statistical analysis of longitudinal data must take into account that serial observations of a subject are likely to be correlated, time may be an explanatory variable, and that missing response data my induce a bias in the results. Let X_{ij} be time varying or fixed covariates for the binary response Y_{ij} of subject $i \in 1, ...n$ at time intervals $j \in t_1, ...t_m$. By convention $X_{ij} \in \mathbb{R}x\mathbb{R}^{'}$ where the first dimension is the intercept. The marginal model is; $logit(E(Y_{ij}|X_{ij})) = X^{\dagger}_{ij}\beta$ and enforces the assumption that the relationship between the covariates and the response is the same for all subjects. Recall that for a binary response, $E(Y_{ij}|X_{ij}) = P(Y_{ij} = 1|X_{ij})$. The random effects model takes into account that the relationship between the covariates and response varies between subjects; $logit(P(Y_{ij} = 1|X_{ij})) = X^{\dagger}_{ij}\beta_i$ If it is know that only a subset of the covariates are involved in the inter-subject variability, we can set $\beta_i = \beta + \beta_i$ and write $logit(P(Y_{ij} = 1|X_{ij}, \beta_i)) = X^{\dagger}_{ij}\beta + OX_{ij}\beta_i$ Where the kernel of $O : \mathbb{R}^n \to \mathbb{R}^{n'}$ is the span of the covariates that do not change between subjects. If $\lambda_i =_d N(0, \sigma)$ then the difference in the parameter vectors β in the two models differ according to σ.

The GEE method of fitting the marginal model is described in: **???** The Survival Analysis is a form of longitudinal analysis that takes into consideration the amount of time an observation is made on a subject. GLM's can be used to fit discrete longitudinal hazard models derived from survival analysis, see **?**. **??** generalized that approach to account for an unobserved subject heterogeneity. **??** applied the hazard model of **??** to the takeover hazard of large firms. A negative relationship between dual class ownership and value is empirically known, and that relationship can be explained by the lower takeover probability of the dual class firms. Dual class entities had a higher risk for takeover, but the hazard is lower since these firms use the dual class structure to change the capital structure in a way that allows the controlling shareholders to remain in control by reducing firm value. The proportional hazards model can be discretized, but it is important to identify whether the process is truly a discrete process. In that case the link function should be the logit as the Marginal Model above specifies, rather than the log-log function of the discretized proportional model. The difference is the modelling of a probability transition in the former case versus a rate for the latter case. Variable selection techniques for longitudinal data are relatively limited and most seem to rely on Wald type tests. Wald tests to include a variable are based on already computed maximum likelihood values. The Rao score test is used to include a covariate in the model building process. The Wald test calculates

$$z^2 = \frac{\widehat{\beta}}{stderr} =_d \chi^2$$

The likelihood ratio statistic for comparing two models $L_0 \in L_1$

$$-2\frac{L_0}{L_1} =_d \chi^2$$

is useful for backward stepwise variable subset selection. The degrees of freedom of the of the statistic is equal to the difference in dimension of the two models.

Discretization & Sheppard's Correction

W. Sheppard (1898) Derived an approximate relationship between the moments of a continuous distribution and it's discrete approximation. This provides a transformation to statistical estimators that correct for the binning of continuous data. As the scale at which datum are collected is increased, the variance of an estimate can become biased. It is important to assess bias caused by grouping and to correct it if necessary. The bias of the approximate maximum likelihood estimator where observations are approximated by interval midpoints $O(w^2)$, where w is the bin width. A Sheppards correction can be used to reduce the bias to order $O(w^3)$, Signal processing engineers often have to deal with such a quantization effect when designing finite precision systems, image processing being a particularly relevant example. The engineering community typically models the quantization noise $Q = [X] - X$, where $[X]$ is the quantized realization of X. One might be tempted to apply a Sheppard's correction to the moments of the quantized data, thinking that $Var(X) < Var([X])$ but it is possible to construct examples where Q and $[X]$ are independent, or where $Cov(X, Q)$ is such that $Var(X) > Var([X])$. Shepard's correction is limited in that is doesn't apply to the first moment, and the frequencies of the first and last bins need to be low. Expand $p(x; \theta)$ in a Taylor series and substitute in the Maximum Likelihood equations. **??** Suppose we have n realizations of iid RV's $X_1, \ldots, X_n$ and the data is collected on a discrete grid on the range of X $Ran(X) = \{[y_i - d_i/2, y_i + d_i/2]\}_{i=1}^{i=m}$ where the intervals are centered on the location where a measurement. The realized values $y_1, \ldots, y_m$ have probabilities $p_i = \int\limits_{y_i - d_i/2}^{y_i + d_i/2} p(x; \theta) \quad dx$ Expanding $p(x; \theta)$ in a Taylor series about y, $p(x; \theta) = \sum\limits_{i=0}^{\infty} \frac{p^{(i)}(y)}{i!}(x - y)^i$.

Multidimensional Scaling

Multidimensional scaling (MDS) is an alternative to factor analysis. The aim of MDS and factor analysis is to detect meaningful underlying dimensions that explain similarities or dissimilarities data points. In factor analysis, the similarities between points are expressed via the correlation matrix. With MDS any kind of similarity or dissimilarity matrix may be used. Given n observations $x_{i}{}_{i=1}^{n} \in \mathbb{R}^k$ and n^2 distances d_{ij} between them, MDS looks for n points $\xi_{i}{}_{i=1}^{n}$ in $dblr^l : l < k$ that preserve the distance relations. When a metric $\rho()$ exists for the similarity measure, gradient descent is used to minimize the MDS functional $S(\xi_1, \ldots, \xi_l) = \left(\sum_{i \neq j} d_{ij} - ||\xi_i - \xi_j||_\rho \right)^{\frac{1}{2}}$.

Principal Components

For a data set $\mathbf{X} \in M_{(N,m)}(\mathbb{R}) = x_1, x_2, \dots x_N | x_i \in \mathbb{R}^m$, the first k principal components provided the best k dimensional linear approximation to that data set. Formally, we model the data via $f(\theta) = \mu + \mathbf{V}_k \theta | \mu \in \mathbb{R}^m, V_k \in O_{m,k}(\mathbb{R}), \theta \in \mathbb{R}^k$ so $f(\theta)$ is an affine hyperplane in $\mathbb{R}^m$

Evaluating classifier performance

Multi-class problems can be treated simultaneously or broken in to a sequence of two class problems. Cross validation is used both for classifier parameter tuning and for feature subset selection. Student-t and ANOVA can be used to evaluate the performance of classifiers against one another. The Student-t test compares two classifiers, while the ANOVA test can compare multiple classifiers against one another. Confusion matrices and ROC graphs are commonly employed visualization tools for assessing classifier performance. The rows of a confusion matrix add to the total population for each class, and the columns represent the predicted class. An ROC curve plots the TP rate against the FP rate. Often a curve in ROC space is drawn using classifier parameters for tuning purposes.

TN	FP
FN	TP

Table 1.1: Two class confusion matrix where the proportions are specified

Common performance metrics for the two class problem are sensitivity (TP), specificity (TN), precision (the proportion of predicted cases within a class that were correct), and accuracy (the overall proportion of correct predictions). These metric can be extended to more than two classes by defining $A = tr(C)/||C||_{L^\infty}$ where C is the confusion matrix. TP, FN, FP, TN are proportions defined for the two class problem.

Covariance Matrix Estimation

For numerical stability in regression algorithms, the covariance matrix needs to be positive definite. An well conditioned estimator for the covariance matrix of a process can be obtained by mixing the sample covariance with the identity matrix. This is a linear shrinkage estimator based on a modified Frobenius norm for $A \in M_{mn}$

$$||\mathbf{A}||_{\mathcal{F}} = \sqrt{\frac{tr(AA^t)}{n}} \tag{1.0.13}$$

Without loss of generality, set $\mu = 0$ and let $\widehat{\Sigma} = \alpha\mathbb{I} + \beta\mathbf{S}$ where $\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{n}$ is the sample covariance. We seek to minimize $E(||\widehat{\Sigma} - \Sigma||^2)$, but since we don't know the true population covariance matrix, we have to form an approximation.

Many applications in statistics and machine learning require an estimate of the covariance matrix or it's inverse. Generally we avoid taking inverses of matrices in practice for stability reasons. The sample covariance matrix usually performs poorly as a proxy for the underlying covariance matrix. There is a large literature on this subject; particularly fomr the finance industry where this is an important part of portfolio theory.two major challenges in covariance estimation are the positive-defniteness constraint and the high-dimensionality where the number of parameters grows quadratically in the dimension.

Testing For Normality

In this section we will use the term $EPDF_X$ to mean the empirical probability density function. There are a variety of univariate tests to help determine which parametric distribution your data belongs to. These fall under the category of Goodness of Fit testing. For a parametric family the null hypothesis $H_o : X =_d p(x|\theta)$ is tested against the alternative that X does not belong to the family $p(x|\theta)$ There are also family of test to determine whether two $EPDF$'s come from the same distribution. Keep in mind that there are many transformations (polynomial, logarithmic, & rational) to transform data to look more Normal. The presence of tails and skew will be the most problematic to deal with.

The Anderson-Darling test determines whether a sample comes from a specified distribution. The sample data can is transformed to a uniform distribution and then a uniformity test is then done on the transformed data. The test statistic is compared against pre-computed values for the assumed probability distribution.

The Kolmogorov-Smirnov is non-parametric a form of minimum distance estimation. It can be used to test a sample against a reference or to compare two samples against each other. In the one sided case the KS statistic calculates the distance between the $EPDF$ of a sample and a reference. In the two sided case the distance between the $EPDS$'f of the two samples are calculated. The KS test is robust to location and shape, making it Omnibus tests evaluate whether the explained variance in a set of data is significantly greater than the unexplained variance. For example is the F-test in ANOVA. Omnibus tests of normality based on the likelihood ratio outperform the Anderson-Darling test statistic.

Chapter 2

Appendix Probability

Probability Basics

Let μ be a non-negative countably additive set function over a sigma algebra Ω of sets from a sample space S. In probability theory, Ω is the set of possible events E. Sets of measure zero denote impossible outcomes. An important feature of measures ν and μ that agree on sets of measure zero is the ability to define a derivative $\frac{d\nu}{d\mu}$, the Radon-Nikodym derivative. When $\nu(E) = 0 \; \forall \; E \in \Omega \mid \mu(e) = 0$ Alternatively, given a measure μ and a nonnegative measurable function f, a new measure can be defined by $\nu(E) = \int E \in \Omega f d\mu$. A random variable is a real valued function on a sample space into a metric space, $X : S \to \mathbb{R}^1$. Associated with a random variable is it's probability density function $f_X(x) = P(\{s \in S | X(s) = x\})$ operating on an algebra of sets generated by the sample space S. By definition, $f_X(x)$ is the sum of probabilities of the events in S that get mapped to $x \in \mathbb{R}$ by X. Let $\mathcal{B}(S)$ be the Borel sets on S, then X has density f if $P(X \in A) = \int_{A \in \mathcal{B}(S)} f(x)dx$ and distribution function $F(x) = P(X < x) = \int\limits_{-\infty}^{x} f(y)dy$ so $F'(x) = f(x) \;\; a.e.$. $E(X) = \int x f(x)dx$ is the expectation of X. The characteristic function $\phi(x) = E(e^{itX})$ determines the distribution and is used in the proof of the CLT theorem, testing for symmetry, and conditional independence. We denote samples with lower case in this section. $x_1 \ldots x_n$ is a random sample of size n In $\mathbb{R}^n$ the random variate X has distribution function $F(x_i, \ldots, x_n) = P(X_1 < x_1, \ldots, X_n < x_n)$ and density $f(x_1, \ldots, x_n)$.

For parametric distributions, one is interested in the question of what value of a parameter best describes the data at hand. This obviously requires the assumption that the data derives from a family of distributions parameterized by one or more variables θ_k. If $x_1 \dots x_n$ is a random sample from X with a distribution given by $p(x; \theta_1 \dots \theta_k)$, we can think of the joint pdf of of the sample $L(x_i, \dots, x_n) = \prod_{i=1}^{n} p(x_i; \theta_1 \dots \theta_k)$ as being explained by the parameters. L is the likelihood of the data given the parameters. The maximum likelihood estimate is obtained by solving the set of equations;

$$\frac{\partial L(\theta_1 \dots \theta_k)}{\partial \theta_1} = 0$$
$$\vdots$$
$$\frac{\partial L(\theta_1 \dots \theta_k)}{\partial \theta_k} = 0$$

The statistical moments of a random variable X are defined $\mu_n = E[X^n] = \int_\Omega X^n P(X) dX$. The characteristic function is the fourier transform of $P(X)$

$$\Phi(\omega) = F\mathcal{F}[P(X)](\omega) = \int_{\infty}^{\infty} e^{i\omega X} P(X) dX.$$

Taking the logarithm and expanding in a MacLauren series, we can relate the statistical moments to the coefficients. Statistical moments are central or raw.

Common sample descriptive statistics relating to location, scale, tail size, and peakedness;

$$mean = \hat{\mu_1} = \frac{1}{n} \sum x_i$$
$$variance = sdd^2 = \hat{\mu_2} = \frac{1}{n-1} \sum (x_i - \hat{\mu_1}$$
$$skewness = \frac{\hat{\mu_3}}{\hat{\mu_2}^{3/2}}$$
$$kurtosis = \frac{\hat{\mu_4}}{\hat{\mu_2}^{2}}$$

The linear association between X_i and X_j is measured by the covariance

$$Cov_{ij} = \sigma_{ij} = \frac{1}{n-1} \sum_k (x_k i - \hat{\mu_1}_i)(x_{kj} - \hat{\mu_1}_j).$$

For a measure without dependence on units the scaled covariance is the correlation

$$C_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

Univariate Probability Distributions

This section covers the properties of common univariate probability distributions.

2.0.1 Uniform, $U(\alpha,\beta)$

$X =_d U(\alpha,\beta)$ if $p(x;\alpha,\beta) = \frac{\chi_{[\alpha,\beta]}}{\beta-\alpha}$.

2.0.2 Exponential Class of Distributions

The exponential class of distributions are characterized by the functional for of the pdf;

$$p(x;\theta) = exp(\alpha(x)\beta(\theta) + \gamma(\theta) + \delta(x))$$

. This class of distributions form the basis of Generalized Linear Model Theory that is discussed below.

There is an alternate parametrization of the exponential family that explicitly includes a dispersion parameter ϕ. This is useful for count data where $E[X] = E[X^2] = \theta$ In general if $E[X^2] > E[X]$ we say the process or data is over dispersed. The parameter ϕ is usually fixed in practice. If we write

$$p(x;\theta,\phi) = exp(\frac{x\theta - \beta(\theta)}{\alpha(\phi)} + \gamma(x,\phi))$$

, the dispersion parameter ϕ for some common distributions;

$p(x;\theta,\phi)$	ϕ
$N(\mu,\sigma)$	σ^2
$IG(\mu,\sigma)$	σ^2
$Gamma(\theta,\phi)$	$\frac{1}{\phi}$
$Poisson(\theta)$	1
$Binomial(\theta)$	1
$NegativeBinomial(\theta,r)$	r

Normal/Gaussian, $N(\mu, \sigma)$

$X =_d N(\mu, \sigma)$ if

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

. This can be re-written in exponential form

$$p(x; \theta) = exp(\frac{x\theta}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{1}{2} log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2})$$

Binomial

Setting

$$\alpha(x) = x \ \ , \beta(\theta) = log(\frac{\theta}{1-\theta}) \ \ , \gamma(\theta) = nlog(1-\theta) \ \ , \delta(x) = log(\left(\begin{array}{c} n \\ y \end{array} \right))$$

gives us the binomial distribution $p(x; \theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{(n-x)}$

Negative Binomial

$$p(x; \theta, r) = \left(\begin{array}{c} x+r-1 \\ r-1 \end{array} \right) \theta^r (1-\theta)^x$$

Poisson

X is Poisson distributed if $p(x; \theta) = \frac{\theta^x e^{-\theta}}{x!}$ Setting $\beta(\theta) = log(\theta) \ \ , \gamma(\theta) = -\theta \ \ , \delta(x) = -log(x!)$ we get the exponential form of the Poisson distribution,

$$p(x; \theta) = exp(xlog(\theta) - \theta - log(x!))$$

The expected value and the variance of a Poisson distributed random variable is equal to θ. The higher moments of the Poisson distribution are the Touchard polynomials in θ. There is a combinatorial interpretation. When $E[X] = 1$ for a Poisson random variate then the i-th moment of X is equal to the number of partitions of a set of size n $\frac{1}{e} \sum\limits_{n=0}^{\infty} \frac{n^i}{n!}$ via Dobinski. The normal distribution with mean θ and variance θ is a good approximation to the Poisson distribution for large θ.

Pareto

X is Pareto distributed if

$$p(x;\theta) = \theta x^{-\theta}.$$

Gamma

X is Gamma distributed if

$$p(x;\theta,\phi) = \frac{x^{\phi^{-1}}\theta^{\phi}e^{-x\theta}}{\Gamma(\phi)}.$$

Weibull

X follows the Weibull distribution if

$$p(x;\theta,\lambda) = \frac{\lambda x^{\lambda-1}}{\theta^{\lambda}}e^{(\frac{x}{\theta})^{\lambda}}.$$

Inverse Gaussian/ Wald Distribution

X follows the Inverse Gaussian distribution if

$$p(x;\theta) = \sqrt{\frac{\theta}{2\pi x^3\sigma}}exp(-\frac{\lambda(x-\theta)^2}{2x\theta^2\sigma})$$

2.0.3 Generalized Extreme Value Distribution $GEV(\theta,\phi,\xi)$

This class of distributions includes the three limiting extreme value distributions of **?** and **?**. $X =_d GEV(\theta,\phi,\xi)$ if

$$p(x;\theta,\phi,\xi) = exp(-max\Big(\big(1+\xi\frac{x-\theta}{\phi}\big)^{-\frac{1}{\xi}},\ \ 0\Big)$$

. Where

2.0.4 Multinomial

Let

$$\Omega = \mathcal{B}(\prod_{i=0}^{i=\infty}\mathbb{Z}(K))$$

be the Borel Algebra generated by $\prod 1, 2, \ldots, K$. This is the sample space of all realizations of experiments with K categorical outcomes. Equip $\mathbb{Z}(K) =$

$i \in 1, \ldots, K$ with a measure $P(i) = \theta_i$. Let X_i be n iid copies of $X =_d p(i; \pi_1, \ldots, \pi_K)$. Now map $\mathbf{X} = (X_1, \ldots, X_n) \in \Omega \to \mathbf{Y} \in \mathbb{N}^K$ Then Y_i are counts of the number of elements of category i in the experiment with n observations. The multinomial distribution is given by

$$p(\mathbf{Y}; n) = \frac{n!}{y_1! \ldots y_K!} (\theta_1)^{y_1} \ldots (\theta_K)^{y_K}$$

. This in not a member of the exponential family, but we can show that the multinomial distribution is the joint distribution of $Y_i =_d Poission(\theta_i')_{i=1,\ldots K}$ random variables conditional to their sum.

$$p(\mathbf{Y}; \theta_1', \ldots, \theta_K') = \prod_{i=1}^{K} \frac{(\theta_i')^{y_i} \; e^{-\theta_i'}}{y_i!}$$

, set $n = Y_1 + \ldots + Y_K$. Writing $p(\mathbf{Y}|n) = p(\mathbf{Y}; \theta_1', \ldots, \theta_K')/p(n)$ and noting that $n =_d \; Poisson(\sum_{i=1}^{K} \theta_i')$, we recover the multinomial distribution by simplifying and setting $\theta_i = \frac{\theta_i'}{\sum_{i=1}^{K} \theta_i'}$

2.0.5 $\chi^2(n)$

If X_i iid $N(0,1)$ and $Y_i = X_i^2$ then $Y = \sum_{i=1}^{n} Y_i =_d \chi^2(n)$. $E[Y] = n$ and $Var(Y) = E[(Y - \mu_Y)^2] = E[(Y - E[Y])^2] = 2n$. More generally, if $Y_i = X_i + \mu_i$ then

$$Y = \sum_{i=1}^{n} (Y_i)^2 = \sum_{i=1}^{n} X_i^2 + 2 \sum_{i=1}^{n} X_i \mu_i + \sum_{i=1}^{n} \mu_i^2 =_d \chi^2(n, \lambda)$$

. Where $\lambda = \sum_{i=1}^{n} \mu_i$ is non-centrality parameter.

See the section on multivariate probability distributions for further information, but it is worth noting that if $X =_d N(\mu, \sigma)$ is multivariate and the variance covariance matrix $s\sigma$ is non-singular, then $(y - \mu)^T \sigma^{-1} (y - \mu) =_d \chi^2(n)$ and setting $\lambda = \mu^T \sigma^{-1} \mu$ we have $y^T \sigma^{-1} y =_d \chi^s(n, \lambda)$

2.0.6 Student-t $t(\nu)$

$X =_d \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\Gamma(\nu/2)} (1 + x^2)^{-\frac{\nu+1}{2}}$

2.0.7 Generalized Inverse Gaussian $GIG(\lambda, \alpha, \beta)$

The GIG distributions are characterized by

$$p(x;\lambda,\theta,\sigma) = \big(\frac{\theta}{\sigma}\big)^{\frac{\lambda}{2}} x^{\lambda-1}\,\frac{1}{2K_\lambda(\sqrt{\theta\sigma})}\,exp(-\frac{1}{2}(\theta x^{-1}+\sigma x)$$

.

Note, $GIG(-1/2,\theta,\sigma) = IG(\theta,\sigma)$

The GIG family members arise as first passage time distributions of ordinary Brownian diffusions to a constant boundary.

2.0.8 Normalized Inverse Gaussian $NIG(\mu, \alpha, \beta, \delta)$

$X =_d NIG(\mu,\alpha,\beta,\delta)$ if

$$p(x;\mu,\beta,\alpha,\delta) = \frac{\delta\alpha}{\pi} exp\big(\delta\sqrt{\alpha^2-\beta^2}+\beta(x-\mu)\big)\frac{K_1(\alpha\; s_\delta(x-\mu))}{s_\delta(x-\mu)}$$

where $x \in \mathbb{R} \quad \mu \in \mathbb{R} \quad \delta > 0 \quad 0 \leq |\beta| \leq \alpha$ and $s_\delta(x) = \sqrt{\delta^2+x^2}$ and $K_1(x) = \frac{x}{4}\int\limits_0^\infty exp - \big(y+\frac{x^2}{4y}\big)y^{-2}\;dy$ is the modified Bessel function of the third kind. This family of distributions is infinitely divisible, $\exists \quad X_t$, a Levy process, $X_{t+\Delta t} - X_t =_d X_{\Delta t} =_d NIG(\mu,\alpha,\beta,\delta)$. X_t is a pure jump process, and

$$p(t;\alpha,\beta,\delta) = \left(\frac{\delta\alpha}{\pi|t|}\right)e^{\beta t}K_1(\alpha|t|)$$

. See Eberlein and Keller (1995) and Barndorff-Nielsen (1998)

2.0.9 Generalized Hyperbolic $GH(\lambda, \alpha, \beta, \delta, \mu)$

The parameters $\lambda, \alpha, \beta, \delta, \mu$ have the respective interpretation of tail heaviness, kurtosis, skewness, and scale, and location. The distribution includes the important classes GIG, NIG, IG, and can be characterized as a Normal variance-mean mixture NVMM parameterized by a GIG distribution. Formally set $U =_d GIG()$, then $X =_d NVMM()$ if $P(X|U=u) =_d N(\mu+\beta, u\Delta)$. This gives a stochastic representation $X = \mu + \beta Z + \sqrt{Z}Y$ where $Y =_d N(0,1)$ and $Z =_d GIG()$.

Limit Theorems

Limit Theorems use the notion of a basin of attraction for pdf's in some functional space $\mathcal{H}$. In $L^2(\Omega)$, we have $N(\mu, \sigma) \subset L^2(\Omega, \nu)$ is the basin of attraction for all pdf's satisfying the conditions of the CLT.

The CLT says that the series $\frac{\sum\limits_{i=1}^{n} X_i}{n}$ converges in in probability to the mean of x_i. Cramer's theorem gives a bound on the probability of large deviation away from the mean in the series $\frac{\sum\limits_{i=1}^{n} X_i}{n}$. The probability decays exponentially with a rate given by the Legendre transform of the cumulant generating function for X_i

Theorem 2.0.1 (Cramer's Theorem)**.** *Let $X_1, X_2, \ldots$ be iid $E(X_i) = 0$, $E(X_i^2) = \sigma^2$, and $F_n(x) = P(\frac{1}{\sigma n^{\frac{1}{2}}} \sum\limits_{i=1}^{n} X_i < x)$, then if $x > 1$ and $x = O(\sqrt{n})$ as $n \to \infty$ we have*

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = exp(\frac{x^3}{\sqrt{n}} \lambda(\frac{x}{\sqrt{n}}))[1 + O(\frac{x}{\sqrt{n}})]$$

$\lambda(x) = \sum\limits_{i=0}^{\infty} c_i x_i$ where the c_i depend on the moments of X_i. $\Phi(x)$ is the distribution function of $N(0, 1)$.

Theorem 2.0.2 (Law of Large Numbers)**.**

The CLT tells us that the pdf of the scaled mean of a sample approaches the normal distribution and the Berry–Esseen theorem specifies the rate at which that happens. The CLT requires X_i to be iid, and with finite second moment and the Berry-Esseen theorem additionally requires a finite third moment.

Theorem 2.0.3 (Berry–Esseen)**.** *Let X_i be iid, $E(X_i^2) = \sigma$, $E(X_i^3) = \rho$, $Y_n = \frac{X_1 + \ldots + X_n}{n}$, and $F_n = \int \frac{Y_n \sqrt{n}}{\sigma}$ and Φ the CDF of $N(0, 1)$, then*

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3 \sqrt{n}} \tag{2.0.1}$$

Multivariate Probability Distributions

Let S_n be the unit sphere in $\mathbb{R}^n$ A random variate X is uniformly distributed on S_n when X is radially symmetric and $||X||_{L^2} = 1 a.s.$ The pdf of a radially symmetric random variable is necessarily of the form $f(x_1, ..., x_n) = g(||x||)$ for some $g \in [0, \infty) \ni \int\limits_0^\infty nV_n r^{n-1} g(r)dr = 1$ where $V_n = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of S_n. If X is radially symmetric , then $\frac{X}{|||X||}$ is uniformly distributed on S_n. If X is uniformly distributed on S_n then $(X_1^2, \ldots, X_n^2) =_{dist} (\frac{Y_1}{\kappa}, \ldots, \frac{Y_n}{\kappa})$ where Y_i iid $\Gamma(\frac{1}{2})$ with sum κ. If $N_1, \ldots, N_n$ iid normal, then $(N_1, \ldots, N_n)$ is radially symmetric with density $g(r) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{\frac{-r^2}{2}}$ This leads us to an algorithm for generating pseudo random variants on uniformly distributed on S_n ;

- Generate n iid $N(0, 1)$
- Compute $\kappa = (\sum\limits_i = 1^n N_i^2)^{\frac{1}{2}}$
- Return $(\frac{N_1}{\kappa}, \ldots, \frac{N_n}{\kappa})$

?

With a little linear algebra, the above can be generalized to a generator for $N(\mu, \Sigma) \in \mathbb{R}^n$. Consider $f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2} x^T \dot{x}} x \in \mathbb{R}_n$, f has density of n iid $N(0, 1)$ rv if

2.0.10 Multivariate Normal $N(\mu, \mathbf{\Sigma})$

$N(\mu, \mathbf{\Sigma})$ is arguably the most important and tractable multivariate probability distribution. The Gaussian distribution is separable via rotation. Precisely the rotation induced by PCA.

2.0.11 Wishart Distribution

The Wishart distribution $W(n)$ is the multivariate generalization of the $\chi^2(n)$ distribution. If $X_{(i)} \sim N(\mu, \mathbf{\Sigma})$, then $XX^t = S \sim W(N)$.

2.0.12 Elliptic $E(\mu, \mathbf{\Sigma})$

Elliptical distributions $E(\mu, \mathbf{\Sigma})$ extend the multivariate normal $N(\mu, \mathbf{\Sigma})$. They can be characterized as affine maps of spherical distributions. The density functions are defined by $p(x) = cg((x - \mu)'\Sigma^{-1}(x - \mu))$ Where $g : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$

and $\Sigma \succ 0$ is positive definite. Many of the properties of the multivariate normal distribution are shared by the elliptical distributions. Linear combinations, marginal distributions and conditional distributions of elliptical random variables can largely be determined by linear algebra using knowledge of covariance matrix, mean and generator.

Statistical Dependence

Linear correlation is a natural dependence measure for multivariate normally and elliptically distributed random variables. Other dependence concepts include rank correlation, comonotonicity, and Brownian covariance.

Distance measures for probability distribution functions.

A number of distance measures for probability distance functions exist. Kullbak Lieber divergence:

$$J_D = \int_{\mathrm{x}} \left[p(\mathrm{x} \mid \omega_1) - p(\mathrm{x} \mid \omega_2)\right] \log \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \mathrm{dx}$$

which simplifies to:

$$J_D = \tfrac{1}{2}\left(\mu_2 - \mu_1\right)^T \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)\left(\mu_2 - \mu_1\right) + \tfrac{1}{2}\mathrm{tr}\left\{\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right\}$$

when

$$X_1 =_d N(\mu_1, \Sigma_1) \;\;,\;\; X_2 =_d N(\mu_2, \Sigma_2).$$

The Bhattacharyya distance :

$$J_B = -\log \int \left[p(\xi \left|\omega_1\right.) p(\xi \left|\omega_2\right.)\right]^{1/2} \mathrm{d}\xi$$

which simplifies to:

$$J_B = \tfrac{1}{8}\left(\mu_2 - \mu_1\right)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \left(\mu_2 - \mu_1\right) + \tfrac{1}{2}\log\frac{\left|\frac{1}{2}(\Sigma_1 + \Sigma_2)\right|}{(\left|\Sigma_1\right| \left|\Sigma_2\right|)^{1/2}}$$

when

$$X_1 =_d N(\mu_1, \Sigma_1) \;\;,\;\; X_2 =_d N(\mu_2, \Sigma_2).$$

The Matusita distance:

$$J_T = \left\{ \int \left[\sqrt{p(\xi\,|\omega_1\,)} - \sqrt{p(\xi\,|\omega_2\,)} \right]^2 \mathrm{d}\xi \right\}^{1/2}$$

which simplifies to:

$$J_T = \left\{ 2\left[1 - \exp(-J_B)\right] \right\}^{1/2}$$

where J_B is the Bhattacharyya distance, when

$$X_1 =_d N(\mu_1, \Sigma_1) \ \ , \ \ X_2 =_d N(\mu_2, \Sigma_2).$$

The Patrick-Fisher distance:

$$J_P = \left\{ \int \left[p(\xi\,|\omega_1\,)P_1 - p(\xi\,|\omega_2\,)P_2 \right]^2 \mathrm{d}\xi \right\}^{1/2}$$

which simplifies to:

$$J_P = \begin{array}{c} (2\pi)^d \,|2\Sigma_1|)^{-1/2} + ((2\pi)^d \,|2\Sigma_2|)^{-1/2} - \\ 2((2\pi)^d \,|\Sigma_1 + \Sigma_2|)^{-1/2} \exp\left\{ -\frac{1}{2}(\mu_2 - \mu_1)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_2 - \mu_1) \right\} \end{array}$$

when

$$X_1 =_d N(\mu_1, \Sigma_1) \ \ , \ \ X_2 =_d N(\mu_2, \Sigma_2).$$

Reference: pp257, et sqq., Devijver, P.A. & Kittler, J (1982) "Pattern Recognition: A Statistical Approach", Prentice Hall International, Englewood Cliffs, NJ.

$S_\alpha(\sigma, \beta, \mu)$ Stable Random Variates

A Levy process is a stochastic process with a drift, a diffusion, and a jump component. The Lévy–Khinchine representation of a Levy process X_t with parameters (a, σ^2, W) is given by

The Levy-Ito decompositon of a Levy process X_t is a decomposition of X_t into singular, absolutely continuous, and discrete processes

$$\begin{aligned} X_{ac} &: X_{ac} \ll X \\ X_s &: X_s \perp X \\ X_d &: card(suppX_d) = \aleph_0 \end{aligned}$$

via Lebesgue's decomposition theorem.

2.0.13 4 definitions of stable

- If

  $$\exists\, C, D\;\forall\; A, B s.t. AX_1 + BX_2 =_d CX + D\;\forall\; X_1, X_2$$

  independent copies of X, then $X \in S_\alpha(\sigma, \beta, \mu)$. Furthermore $\exists\;\alpha \in (0, 2] s.t. C$ satisfies $C^\alpha = A^\alpha + B^\alpha$, for any stable RV and $\forall\; A, B$

- Stable RV's satisfy a general CLT. If

  $$\forall\; n \geq 2\;\exists\; C_n > 0 D_n \in \mathbb{R} s.s. X_i + X_@ \ldots X_n =_D C_n X + D_n$$

  where X_i iid, then $X \in S_\alpha(\sigma, \beta, \mu)$

- If $\exists$ *iid RV* Y_i and

  $$d_n, a_n \in BBCREVISIT^n \mathbb{R}^n s.t \frac{\sum_1^n Y_i}{d_n} + a_n =_d X$$

  then $X \in S_\alpha(\sigma, \beta, \mu)$.

- If $\exists\;\alpha \in (0, 2],\; \sigma \geq 0\;\beta \in [-1, 1],\; \mu \in \mathbb{R}$ such that

  $$E[e^{i\theta X}] = \int_\Omega e^{i\theta X} dX = exp\big(-\sigma^\alpha |\theta|^\alpha (1 - i\beta\; sgn(\theta)\; tan\big(\frac{\pi\alpha}{2}\big) + i\mu\theta)\big)$$

  when $\alpha \neq 1$ and when $\alpha = 1$ we have

  $$E[e^{i\theta X}] = exp(-\sigma|\theta|(1 + i\frac{2}{\pi}\beta\; sgn(\theta)\; ln|\theta| + i\mu\theta))$$

2.0.14 Variance Gamma Process

$X_{VG}(t;\sigma,\nu,\theta)\theta\gamma(t;\nu) + \sigma W_{Y(t;\nu)}$ where $\gamma(t;\nu)$is a Γ process.

$$P_{\gamma(t;\nu)}(x) = \frac{x^{\frac{t}{\nu - 1}} e^{-\frac{x}{\nu}}}{\nu^{t/\nu}\Gamma(t/\nu)}$$

$$\Phi_{VG}(\omega) = E(e^{i\omega X_{VG}}) = \frac{1}{(1 - i\omega\nu\theta + \sigma^2\nu\mu^2/2)^{t/\nu}}$$

We can show that the Variance Gamma process is the difference of two independent Gamma processes $X_{VG} = \gamma_p - \gamma_n$ to obtain a new pdf

$$P_{\gamma(t;\nu)} = \begin{cases} \frac{1}{\nu|x|} e^{-|x|/\eta_p} & x < 0 \\ \frac{1}{\nu|x|} e^{-|x|/\eta_n} & x > 0 \end{cases}$$

.

Maximum Entropy

Entropy in the context of information theory is expressed in units of bits, the amount of uncertainty in a yes or no question. Formally, for a sequence $\{X_i\} \ni p_i$ is a priori/posteriori probability of observing X_i we define $H = -\sum_i p_i log_2(p_i)$. We can define the entropy of a probability distribution by $H = \int_{infty}^{\infty} p(x)log(p(x))dx$. We see the uniform distribution maximizes the entropy; if $p_i = \alpha \ \forall \ i$ then $\frac{\partial H}{\partial} = logp + 1/p = 0$

Gaussian Processes

X_t is a Gaussian process (GP) if $\forall(t_1, ...t_n)$ we have that $(X_{t_1}, ..., X_{t_n}) \approx N(\mu, \Sigma)$. $\Sigma \sim K(x, x')$ the covariance functions. Gaussian processes are defined by second order statistics (BBCREVISIT). $K(x, x')$ specifies a distribution on functions (in the distribution sense of functional analysis). Every covariance function is a scalar product of vectors. Orenstein-Uhlenbeck (OU) and Brownian Motion (BM) are Gaussian processes. If a GP depends on $|x - x'|$ it is stationary. The OU process is a stationary GP and BM is not. The covariance function of the OU process is not differentiable. If a Gaussian process depends only on $|x - x'|$ it's called isotropic, these form an important class of Here we list the common covariance functions (kernels) constant, linear, Gaussian noise, square exponential, and OU.

$$K_c(x, x') = const$$

$$K_{\updownarrow}(x, x') =< x, x' >$$

$$K_{GN}(x, x') = \sigma^2 \delta_{x,x'}$$

$$K_{SE}(x, x') = e^{\frac{-|d|^2}{2\updownarrow}}$$

$$K_{OU}(x, x') = e^{\frac{-|d|}{\updownarrow}}$$

Here $\updownarrow$ is the length scale and $d = |x - x'|$

Discretizing Continuous Variables

R discretization library implements Fayyad and Irani

Python https://github.com/navicto/Discretization-MDLPC

Most methods used for discretizing a continuous variable use its relationship to another variable to determine the partitions. This is often found in classification procedures, such as decision trees6] 10 and in naive Bayesian classifiers.11,12 An entropy based method, proposed by Fayyad and Irani,6 chooses the partitioning point s in a sorted set of continuous values to minimize the joint Ž . entropy Ž Ž .. H X, Y of the continuous variable and the classification variable. This is applied to the creation of decision tree structures for classification by 66 CLARKE AND BARTON recursively finding more partitioning points top-down discretization . The Ž . method is expanded to minimize a MDL metric to choose the partitioning points. Another MDL based method for discretization is described by Pfahringer.7 A set of the best partitioning points is determined by recursively partitioning the Ž D sorted variable values to a depth D 2 y 1 partitions in a binary tree. Then . the MDL metric is used in a best first search in this set to determine the best partitions for decision tree classification. A method that merges adjacent partitions of sorted variable values, according to the x2 statistical test, is described by Liu and Setiono.8 The variable values are sorted and initially partitioned into, at most, N intervals. The intervals are first recursively merged according to the lowest x2 value until a significance level of 0.05 is reached for each partition. The intervals are further merged until a preset error rate with the classification variable is reached. If there is only one resulting interval, the variable is not relevant to the classification problem and is dropped. This method combines the discretization of continuous variables with feature selection for classification. Dougherty, Kohavi, and Sahami9 compared several discretization techniques with decision trees and with naive Bayesian classifiers. They found that a MDL metric, similar to that used by Fayyad and Irani,6 provided slightly better classifications in both methods. A metric for discretization, based upon a variable's classification in relation to other variables, is described by Hong.10 This metric is based upon a K nearest neighbor clustering technique and is used to generate decision trees. An interesting feature of this method is that it returns an optimal number of partitions according to the metric. This is done by finding the ‘‘knee’’ of the plotted curve of the score as a function of the number of partitions. The plot is a concave function of the metric; the knee is the point on the plot where the changes in the number of variable values Ž . X axis become greater than the changes in the metric value Ž . Y axis . The concavity of a plot is exploited in the information theoretic discretization methods developed in Section IV. Subramonian, Venkata, and Chen,13 describe a visual framework for interactive discretization for decision tree classification. A user can choose between several algorithms and metrics for a classification problem, instead of being limited to one method and metric. The choice of metrics includes cross-entropy and the L y 1 norm between two distributions. Pazzani11 describes a technique for iterative discretization of continuous variables for naive Bayesian classifiers. Each continuous variable is initially divided into five partitions. For each variable, two partitions are then merged or a partition is divided into two partitions to find a lower classification error. This procedure is repeated for each continuous variable until the error rate can no longer be reduced. Another method for constructing naive Bayesian classifiers using a MDL metric is presented by Friedman and Goldszmidt.12 This method begins by finding the best initial partition of a continuous variable by dividing the range BAYESIAN BELIEF NETWORKS 67 into two partitions and then iterating the partitioning until there is no further improvement in the MDL score top-down partitioning . The MDL metric Ž . includes all of the variables used for classification and is repeated for each continuous variable. Given a BBN structure, this method discretizes each continuous variable in the Markov blanket of each classification variable. This procedure is iterated until there is no improvement in each local MDL score. Friedman and Goldszmidt propose that this method can be adopted to learning BBN structure by starting with some initial discretization of each continuous variable, learning an initial structure, and then rediscretizing as described above. While this technique optimizes the conditional probabilities, it depends upon the initial approximate discretization of continuous variables to learn the correct network structure. Extensions to current discretization techniques for classification methods have been proposed for BBNs, but none have currently been published. Methods for both static done in data preprocessing and dynamic done during BBN Ž .Ž construction discretization of continuous variables are presented in Sections IV . and V. Often, a continuous variable has a normal, or Gaussian, frequency distribution. The characteristics of the normal distribution are well understood and a unified heuristic method for finding a BBN structure with both discrete variables and continuous variables with a Gaussian distribution is described by Heckerman and Geiger.14 This method requires the mean and variance of each continuous variable's values to parameterize the distribution

Chapter 3

Appendix Numerical Linear Algebra

$M_{m,n}(\mathbb{F})$ denotes the vector space of matrices over the field $\mathbb{F}$.

$$A \in M_{n,n}(\mathbb{F})\ b \in \mathbb{F}^n,\ \exists\ x \ni\ Ax = b \texttt{ iff } det(A) = 0$$

$\mathbf{X} \in M_{mn}(\mathbb{R})$ is positive definite if $v^t X v > 0\ \forall\ v \in \mathbb{R}^n$. We can construct positive definite symmetric matrices by by forming $\mathbf{X}^t\mathbf{X}$ where $\mathbf{X}$ is an orthogonal (full rank) matrix.

Horner's rule is a method for evaluating a polynomial at a point in $O(n)$ time. Straightforward evaluation of a n degree polynomial is done in $O(n^2)$ time. Simply rewrite the function $f(x) = \sum_{i=0}^{n-1} a_i x^i$ as $f(x) = (\ldots(a_{n-1}x + a_{n-2})x + \ldots + a_1)x + a_0$

The min max characterization of eigenvalues

This is a variational characterization of the eigenvalues of compact operators on a Hilbert space. Let $H\ \ni\ H = H^\dagger$ The Rayleigh quotient is defined by $R(x) = \frac{<Hx,x>}{\|x\|^2}$ Let $\sigma(H) = \lambda_i$, then $\forall S_k \in \mathbb{R}^n$ we have

$$\underset{x \in S_k \|x\|=1}{max} < Hx, x > \geq \lambda_k$$

which implies

$$\underset{S_k}{inf} \underset{x \in S_k \|x\|=1}{max} < Hx, x >\geq \lambda_k$$

Equality above is achieved when $S_k = spn\mu_k$ where μ_k is the kth eigenvector of H.

The min max theorem is that

$$\lambda_1 \leq R(x) \leq \lambda_n$$

The Discrete Fourier Transform on $\ell^2(\mathbb{Z}_{N_1})$

Let $z \in \mathbb{Z}_{N_1}$,$z = (z(0), z(1), \ldots, z(N_1 - 1))$. We index from 0 instead of 1 for convenience of presenting the FFT. Define

$$\widehat{z(m)} = \sum_{k=0}^{N_1-1} z(k)e^{\frac{-2\pi\imath km}{N_1}}$$

The map$\hat{}$: $\ell^2(\mathbb{Z}_{N_1}) \to \ell^2(\mathbb{Z}_{N_1})$ is the Fourier Transform. The vectors

$$E_0, E_1, \ldots, E_{N_1-1} : E_m(n) = \frac{e^{\frac{-2\pi\imath mn}{N_1}}}{\sqrt{N_1}}$$

form an orthonormal basis for $\ell^2(\mathbb{Z}_{N_1})$. The vectors $\frac{E_0}{N_1}, \frac{E_1}{N_1}, \ldots, \frac{E_{N_1-1}}{N_1}$ form an orthogonal basis called the Fourier Basis.

Extend the indices over $\mathbb{Z}_{N_1}$ to $\mathbb{Z}$ by considering $\mathbb{Z}_{N_1}$ to be the algebraic group $\mathbb{Z} mod N_1$. Then we can define the translation operator

$$(R_l z)(n) = z(n - l).$$

We can also define the convolution operator with this extended notion of $\mathbb{Z}_{N_1}$;

$$z * w = \sum_{k=0}^{N_1-1} z(m - n)W(n)$$

The Fourier Multiplier Operator $T_{(m)}$ where $m \in \ell_2\mathbb{Z}_{N_1}$ is given by

$$T_{(m)} = (m\hat{z})\check{}$$

Fourier Inversion Formula:

$$z(m) = \frac{1}{N_1} \sum_{k=0}^{N_1-1} z\hat{(k)}e^{\frac{2\pi\imath km}{N_1}}$$

Parsevall's Relation:

$$< z, w >= \frac{1}{N_1} < \hat{z}, \hat{w} >$$

Plancherel's Formula: Parsevall's relation with $w = z$.

Representation in the Fourier Basis:

$$z = \sum_{k=0}^{N_1-1} z(\hat{k}) F_k$$

The effect of the translation operator is to rotate the phase of the Fourier Transform:

$$(R_l z)\hat{(}k) = e^{\frac{2\pi\imath kl}{N_1}} \widehat{z(k)}$$

The effect of conjugation is to reflect the Fourier Transform:

$$(\overline{z})\hat{(}k) = \overline{\hat{z}(-k)}$$

The Convolution Operator is equivalent to a Fourier Multiplier Operator:

$$b * z = (m\hat{z})\check{}: m = \hat{b}$$

Multiresolution analysis

Basis functions of a linear subspace $V_j \subset L^2(\Omega)$ are defined by a scaling function ϕ

via the following procedure;

$$\phi_{ij}(x) = \phi(2^{-j}x - i)$$

$$V_j = span\{\phi_{ij}\}$$

$$W_{j+1} = V_j \ V_{j+1}^{\perp}$$

$$\ldots V_{j+1} \subset V_j \subset \ldots \subset V_0 \subset \ldots V_{-j} \subset \ldots$$

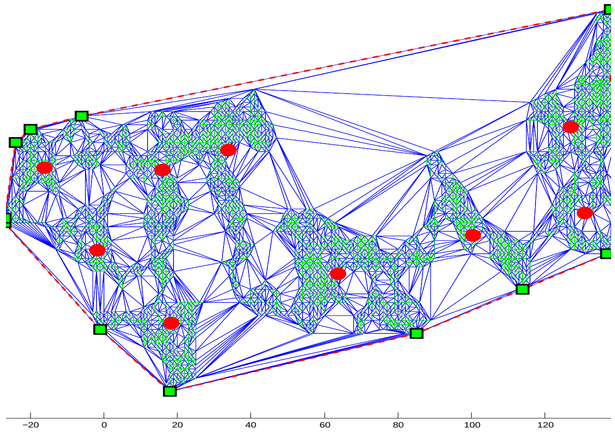
$$V_j = V_{j+1} \oplus W_{j+1} x \in W_{j+1} \Rightarrow \ \exists \ a_l \ \ x = \sum_l \{a_l\} \phi_{jl}$$

A basis for W_{ij} is constructed from a mother wavelet ψ.

Voroni Tesselations

A centroidal Voronoi tessellation is a Voronoi tessellation where the generating points are the centroids of the corresponding regions. Applications Voronoi tessellations can be found in image compression, clustering, quadrature, and finite difference methods. distribution of resources. The dual of the Voroni tessellation in $\mathbb{R}^2$ is the Delaunay triangulation.

The example below is a simulated example of resource allocation in $\mathbb{R}^2$ A partition of a Random walk in $\mathbb{R}^2$ obtained by calculating the Voroni tessellation and associated Delaunay triangulation on the k-means centroids.



The Matrix Exponential

The matrix exponential of a matrix $\mathbf{A}$ is defined as

$$\begin{aligned} e^{\mathbf{A}} &= \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \dots \\ &= \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}. \end{aligned}$$

The Pade approximation to $e^{\mathbf{A}}$ is

$$e^{\mathbf{A}} \approx R(\mathbf{A}),$$

with

$$R_{pq}(\mathbf{A}) = (D_{pq}(\mathbf{A}))^{-1} N_{pq}(\mathbf{A})$$

where

$$D_{pq}(\mathbf{A}) = \sum_{j=1}^{p} \frac{(p+q-j)!p!}{(p+q)!j!(p-j)!}\, \mathbf{A}^j$$

and

$$N_{pq}(\mathbf{A}) = \sum_{j=1}^{q} \frac{(p+q-j)!q!}{(p+q)!j!(q-j)!}\, \mathbf{A}^j.$$

See **?** for a detailed accounting of this and other matters regarding the calculation of the matrix exponential.

Chapter 4

Appendix Numerical Software

Computers are getting more powerful over time but size of the problems we're solving scales with the increased performance. Tools for acquiring and storing data are improving at an even faster pace than processors. It turns out the communication is the real bottleneck to scaling many algorithms. The capacity of fast memory close to the computing resource (cache) grows very slowly in time. Commodity hardware is now being used to deploy large distributed computing systems running Apache Hadoop and Spark. The possibility for computing on ever larger data sets is tantalizing for the machine learning community. These notes discuss the a number of numerical computing libraries the author has used in various machine learning and image processing projects. There is a lot of effort underway to integrate and distribute some of the algorithms in these libraries to the Hadooop ecosystem.

Many of the libraries discussed here are available in Matlab. Intel MLK & IPP, Arpack, UMFPACK, and SDPA have been integrated in the autor's klMatrix library see the section on the KL Libraries below for specifics.

BLAS MLK & Atlas

Graphviz - Graph Visualization Software

Graphviz (http://www.graphviz.org) is open source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains.

Graphviz is used to generate collaboration, inheritance, and call diagrams the KL documentation. There is an API that is used in the KL framework to facilitate graph visualization.

ARPACK

ARPACK++ is an object-oriented version of the Fortran ARPACK package. ARPACK is designed to compute a few eigenvalues and eigenvectors of large scale sparse matrices and pencils via the Arnoldi process for finding eigenvalues called. These methods utilize Krylov Subspace Projections for iterative solution that avoids matrix multiplication. ARPACK implements the implicit restarted Arnoldi method which reduces the storage requirements of the traditional Lanczos iteration for Hermitian matrices and Arnoldi iteration for general matrices. The key to the Krylov method is to calculate the linear subspace of $\mathbb{R}^{(n,n)}$ induced by span of the first m powers of the image of b under a linear operator A, $\kappa_m(A, b)|A \in \mathbb{R}^{(n,n)}b \; in\mathbb{R}^n = \{b, Ab(A)^2b, \ldots (A)^m b\}$. This avoids direct matrix matrix operations when finding the first few eigenvector, eigenvalue pairs in a large system of linear equations.

ATLAS

Automatically Tuned Linear Algebra software.

METIS

METIS is a software library for finite element analysis and graph partitions. It also can be used to reduce the fill order of sparse matrices.

SDPA

SDPA is a software library for solving SDPs using on the Mehrotra-type predictor-corrector infeasible primal-dual interior-point method. It is implemented C++ language and utilizes the machine dependent BLAS such as Intel MKL, ATLAS. LAPACK routines are used for matrix computations. Efficient methods to compute the search directions exploiting the sparsity of the data matrices are implemented. Sparse or dense Cholesky factorization for the Schur complemetn matrix is automatically selected. The calculation of the Schur complement matrix is implemented in reentrant code. A sparse version of SDPA is available that uses METIS and SPOOLES libraries for finding a proper sparse structure of the problem.

SPOOLS

SPOOLES is a library for solving sparse real and complex linear systems of equations. SPOOLES can factor and solve square linear systems of equations with symmetric structure, and it can compute multiple minimum degree, generalized nested dissection and multisection orderings of matrices with symmetric structure. SPOOLES utilizes a variety of Krylov iterative methods. The preconditioner is a drop tolerance factorization.

SuperLU

SuperLU (http://crd-legacy.lbl.gov/ xiaoye/SuperLU/) is a general purpose library for the direct solution of large, sparse, nonsymmetric systems of linear equations on high performance machines. The library is written in C and is callable from either C or Fortran. The library routines will perform an LU decomposition with partial pivoting and triangular system solves through forward and back substitution. The LU factorization routines can handle non-square matrices but the triangular solves are performed only for square matrices. The matrix columns may be preordered (before factorization) either through library or user supplied routines. This pre-ordering for sparsity is completely separate from the factorization. Working precision iterative refinement subroutines are provided for improved backward stability. Routines are also provided to equilibrate the system, estimate the condition number, calculate the relative backward error, and estimate error bounds for the refined solutions.

SuiteSparse

Tim Davis' (http://www.cise.ufl.edu/ davis/welcome.html) collection of sparse matrix software. Tim is also the curator of The University of Florida Sparse Matrix Collection (http://www.cise.ufl.edu/research/sparse/matrices/) a must see for anyone interested in sparse matrices and visualization.

AMD: symmetric approximate minimum degree BTF: permutation to block triangular form CAMD: symmetric approximate minimum degree CCOLAMD: constrained column approximate minimum degree COLAMD: column approximate minimum degree CHOLMOD: sparse supernodal Cholesky factorization and update/downdate CSparse: a concise sparse matrix package CXSparse: an extended version of CSparse KLU: sparseLU factorization, for circuit simulation LDL: a simple LDL^T factorization UMFPACK: sparse multifrontal LU factorization RBio: MATLAB toolbox for reading/writing sparse matrices UFconfig: common configuration for all but CSparse SuiteSparseQR: multifrontal sparse QR

SuiteSparse AMD

AMD is a set of routines for pre-ordering a sparse matrix prior to numerical factorization. It uses an approximate minimum degree ordering algorithm to find a permutation matrix P so that the Cholesky factorization $PAP^{\dagger} = LL^{\dagger}$ has fewer (often much fewer) nonzero entries than the Cholesky factorization of A. The algorithm is typically much faster than other ordering methods and minimum degree ordering algorithms that compute an exact degree . Some methods, such as approximate deficiency [Rothberg and Eisenstat 1998] and graph-partitioning based methods [Hendrickson and Rothberg 1999; Karypis and Kumar 1998; Pellegrini et al. 2000; Schulze 2001] can produce better orderings, depending on the matrix. The algorithm starts with an undirected graph representation of a symmetric sparse matrix . Node i in the graph corresponds to row and column i of the matrix, and there is an edge (i, j) in the graph if a_{ij} is nonzero. The degree of a node is initialized to the number of off diagonal non-zeros in row i, which is the size of the set of nodes adjacent to i in the graph.

SuiteSparse UMFPACK

UMFPACK is a set of routines for solving systems of linear equations, $Ax = b$, when A is sparse and unsymmetric. It is based on the Unsymmetric-pattern MultiFrontal method. UMFPACK factorizes PAQ, $PRAQ$ and $PR^{-1}AQ$, into the product LU, where L and U are lower and upper triangular, respectively, P and Q are permutation matrices, and R is a diagonal matrix of row scaling factors (or $R = I$ if row-scaling is not used). Both P and Q are chosen to reduce fill-in (new nonzeros in L and U that are not present in A). The permutation P has the dual role of reducing fill-in and maintaining numerical accuracy (via relaxed partial pivoting and row interchanges). The sparse matrix A can be square or rectangular, singular or non-singular, and real or complex (or any combination). Only square matrices A can be used to solve $Ax = b$ or related systems. Rectangular matrices can only be factorized.

fftw

fftw is a highly optimized library for calculating the discrete Fourier transform. ?. Generally fftw performs better than Intel IPP. fftw can be sped up by compiling with the Intel Compiler.

Simulation and modeling with the kl Software Framework

Class, interaction and collaboration diagrams are presented below for a modeling framework implemented by the author. The framework is implemented in C++. The simulation of various univariate and multivariate random number generators along with the distribution tests from the CDHC library are included as well. Features of this framework include:

- utilizing optimized BLAS libraries
- the up to date methods for univariate random number generation
- wrappers for Intel performance primitives and GSL
- multiple memory management facilities

We will use the term $EPDF_X$ to mean the empirical probability density function. There are a variety of univariate tests to help determine which parametric distribution your data belongs to. These fall under the category of Goodness of Fit testing. For a parametric family the null hypothesis $H_o : X =_d p(x|\theta)$ is tested against the alternative that X does not belong to the family $p(x|\theta)$ There are also family of test to determine whether two $EPDF$'s come from the same distribution.