This work if a field reference for Machine Learning, and Numerical Linear Algebra.

I am grateful to my parents and family for their patience and *love.* Without them this work would never have come into existence (literally).

My graduate studies in the US were supported in part by the National Petroleum Research Council.

Finally, I wish to thank the following: Beth Orton and Sarah McLachlan; *and* my brother (because he asked me to).

Pittsburgh, PA Bruce Campbell
February 2012

Contents

1 StagingArea **4**

1.0.1 Distance Metrics 4

1.0.2 Primer on Spectral Graph Theory 4

1.0.3 Generalized Chebyshev Bounds on Quadratic Sets via Semidefinite Programming 5

1.0.4 Diffusion Map 5

1.0.5 The Matrix Exponential 6

1.0.6 Spectral Geometry 7

1.0.7 Sparse Representation 7

1.0.8 Concentration of Measure 8

1.0.9 Primal Dual Theory 9

2 Probability & Statistics **10**

2.1 Univariate Probability Distributions 12

2.1.1 Uniform, $U(\alpha, \beta)$ 12

2.1.2 Exponential Class of Distributions 12

2.1.3 Generalized Extreme Value Distribution $GEV(\theta, \phi, \xi)$ . . 14

2.1.4 Multinomial 14

2.1.5 $\chi^2(n)$ 15

2.1.6 Student-t $t(\nu)$ 15

2.1.7 Generalized Inverse Gaussian $GIG(\lambda, \alpha, \beta)$ 15

2.1.8 Normalized Inverse Gaussian $NIG(\mu, \alpha, \beta, \delta)$ 15

2.1.9 Generalized Hyperbolic $GH(\lambda, \alpha, \beta, \delta, \mu)$ 16

2.2 Theorems 16

2.3 Multivariate Probability Distributions 16

2.3.1 Multivariate Normal $N(\mu, \mathbf{\Sigma})$ 17

2.3.2 Wishart Distribution 17

2.3.3 Elliptic $E(\mu, \mathbf{\Sigma})$ 17

2.4 Statistical Dependence 17

2.5 Distance measures for probability distribution functions. 18

2.6 $S_\alpha(\sigma, \beta, \mu)$ Stable Random Variates 19

2.6.1 4 definitions of stable 19

2.6.2 Variance Gamma Process 20

2.7 Maximum Entropy 20

2.8 Simulation and modeling with the kl Software Framework 21

3 **MachineLearning** **23**

3.1 Kernel Density Estimation 23

3.2 Covariance Matrix Estimation 23

3.3 Learning Theory and Functional Analysis 24

3.4 Testing for normality and other distributions 25

3.5 Regression Methods 25

3.6 Generalized Linear Models 26

3.7 Fitting the GLM 29

3.8 Feature Subset Selection (FSS) 30

3.9 Longitudinal Data Analysis 30

3.10 Discretization & Sheppard's Correction 32

3.11 Multidimensional Scaling 32

3.12 Principal Components 33

3.13 Evaluating classifier performance 33

3.14 Graph Spectra 33

3.15 Matrix Factorization 34

3.16 PCA and its generalization to the Exponential Family 34

3.17 Manifold Learning 35

3.18 Graph Laplacian 36

3.19 Learning With Kernels 36

3.20 Bregman Divergences 38

3.21 Matrix Factorization via Generalized Bregman Divergences . . . 38

3.22 Spectral Clustering 38

3.23 Co-Clustering 38

3.24 Proximity Measurement - central versus pairwise grouping 38

4 **Numerical Linear Algebra Background** **40**

4.1 The Discrete Fourier Transform on $\ell^2(\mathbb{Z}_{N_1})$ 40

4.2 Multiresolution analysis 42

4.3 Voroni Tesselations 42

5 **Random Matrix Theory** **43**

5.0.1 RSK and Young Tableaux - A combinatorial application of The Tracy Widom Distribution - Panleve 44

5.1 Generating Random Matrices $A \in U(n), P(n), O(n)$ 45

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.*

Chapter 1

StagingArea

1.0.1 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.

1.0.2 Primer on 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

$$L(u,v) = \begin{cases} d_v & : \ u = v \\ -1 & : u \sim v \\ 0 & : u \nsim v \end{cases} \tag{1.0.1}$$

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

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

$$\mathcal{L} = \frac{1}{\sqrt{d_u}} \sum^{v:u\sim v} (\frac{g(u)}{\sqrt{d_u}} - \frac{g(v)}{\sqrt{d_v}}) \tag{1.0.2}$$

$$Vol(G) = \sum_{v\in V}^{d_v} = Tr(T) \tag{1.0.3}$$

$$\sigma(\mathcal{L}) \in \mathbb{R}^+ \tag{1.0.4}$$

$$ker(\mathcal{L}) = span\{T^{\frac{1}{2}} \nVdash\} \tag{1.0.5}$$

1.0.3 Generalized Chebyshev Bounds on Quadratic Sets via Semidefinite Programming

Boyd et al Vandenberghe et al. [59] 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.

$$\min(1 - \sum_{i=1}^{m} \lambda_i) \ni Tr(A_i z_i) + 2b_i^T z_i + c_i \lambda_i \geqslant 0 \quad \forall i = 1, ..., m \tag{1.0.6}$$

$$\sum_{i=1}^{m} [\begin{matrix} z_i & z_i \\ z_i & \lambda_i \end{matrix}] \succeq 0 \tag{1.0.7}$$

$$C = \{x \in \mathbb{R} : x^T A_i x + 2b_i^T + c_i < 0 : i = 1, ..., m\} \tag{1.0.8}$$

$$\min E[f_0(X)] \ni E[f_i(X)] = a_i : i = 1, ..., m \tag{1.0.9}$$

moment constraints

Let

$$\bar{x} \in \mathbb{R}^n S \subset S^n \ni S \succeq \bar{x}\bar{x}^T \tag{1.0.10}$$

and define

$$P(C, \bar{x}, S) = inf_{\mathcal{P}(\mathbb{R}^n)} \{P(X \in C) \mid E[X] = \bar{x} E[XX^T] = S\} \tag{1.0.11}$$

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.

1.0.4 Diffusion Map

[15], [17], [19], [36], [37], [43].

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.

$$d(x) = \int\limits_X k(x,y)d\mu(y) \tag{1.0.12}$$

$$P(x,y) = \frac{k(x,y)}{d(x)} \tag{1.0.13}$$

$$(D_t(x,y))^2 = \|P_t(x,\cdot) - P_t(y,\cdot)\|_{L^2(X,\frac{d\mu}{\pi})} \tag{1.0.14}$$

$$\pi(\mu) = \frac{d(y)}{z \in Z^{d(z)}} \tag{1.0.15}$$

$$\pi(x)p(x,y) = \pi(y)p(y,x) \tag{1.0.16}$$

$D_t(x,y)$ 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.

$$D_t(x,y) = \sqrt{\sum_{\lambda \geq 1} \lambda_l(\phi_l(x) - \phi_l(y))^2} \tag{1.0.17}$$

We can define an embedding in Euclidian space via

$$\Psi_t(x) = \{\lambda_1^t \phi_1(x), ...\lambda_{s(\delta,t)}^t \phi_{s(\delta,t)}(x)\} \tag{1.0.18}$$

1.0.5 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 [41] for a detailed accounting of this and other matters regarding the calculation of the matrix exponential.

1.0.6 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 $N(\lambda) \sim (2\pi)^{-d)\omega_d \lambda^{\frac{d}{2}} vol(\Omega}$

1.0.7 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 1.0.1 (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 $\tilde{X}_{(n)} = \frac{1}{\sqrt{n}}\rho_n\Pi_{d,n}(\theta_n)$.

1.0.8 Concentration of Measure

[6], [10], [12], [24], [39], [42], [52], [56].

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 [31] 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 $f \in L^1$:

$$\|T_t f\|_\infty \leq a(t)\|f\|_1. \tag{1.0.19}$$

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

$$\|T_t f\|_\infty \leq c(t)\|f\|_2 \tag{1.0.20}$$

Also by duality, the inequality (1.0.20) is equivalent to

$$\|T_t f\|_2 \leq c(t)\|f\|_1 \tag{1.0.21}$$

It is known that, under the assumptions on the semigroup $(T_t)_{t\geq 0}$, (1.0.20) implies (1.0.19) with $a(t) \leq c^2(t/2)$ and (1.0.19) implies (1.0.20) 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

$$\int f^2 \log f \, d\mu \leq \epsilon Q(f) + \beta(\epsilon) \|f\|_2^2 + \|f\|_2^2 \log \|f\|_2 \tag{1.0.22}$$

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

$$\int f^2 \log f \, d\mu \leq \epsilon Q(f) + \beta(\epsilon) \, \|f\|_2^2 + \|f\|_2^2 \log \|f\|_2 \tag{1.0.23}$$

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

$$M_\eta(t) = (\eta + 1) t^{-(\eta+1)}) \int_0^t s^\eta \beta \left(\frac{s}{\eta + 1} \right) ds \tag{1.0.24}$$

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

$$\|e^{-At}\|_{\infty,2} \leq e^{M_\eta(t)} \tag{1.0.25}$$

for all $0 < t < \infty$.

1.0.9 Primal Dual Theory

[**?**]

Chapter 2

Probability & Statistics

μ is 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 $\frac{d\nu}{d\mu}$. 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, $X : S \rightarrow \mathbb{R}^1$. 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.*. $E(X) = \int xf(x)dx$ 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 $x_1 \ldots x_n$ is a random sample from X with a distribution given by $p(x; \theta_1 \ldots \theta_k)$, we can think of the joint pdf of of the sample $L(x_i, \ldots, x_n) = \prod\limits_{i=1}^{n} p(x_i; \theta_1 \ldots \theta_k)$ 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

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

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 $\frac{\sigma}{\sqrt{n}}$

2.1 Univariate Probability Distributions

This section covers the properties of common univariate probability distributions.

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

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

2.1.2 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 $\frac{1}{e}\sum_{n=0}^{\infty}\frac{n^i}{n!}$ 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

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

2.1.3 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

2.1.4 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 $\theta_i = \frac{\theta_i'}{\sum_{i=1}^{K} \theta_i'}$

2.1.5 $\chi^2(n)$

If X_i iid $N(0,1)$ and $Y_i = X_i^2$ then $Y = \sum\limits_{i=1}^{n} Y_i =_d \chi^2(n)$. $E[Y] = n$ 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 $\lambda = \sum\limits_{i=1}^{n} \mu_i$ 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 $y^T \sigma^{-1} y =_d \chi^s(n, \lambda)$

2.1.6 Student-t $t(\nu)$

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

2.1.7 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.

2.1.8 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, $\exists \;\; X_t$, 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)

2.1.9 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 $Z =_d GIG()$.

2.2 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 $\frac{\sum_{i=1}^{n} X_i}{n}$ 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 $\frac{\sum_{i=1}^{n} X_i}{n}$. The probability decays exponentially with a rate given by the Legendre transform of the cumulant generating function for X_i

Theorem 2.2.1 (Cramer's Theorem)**.** *Let $X_1, X_2, \dots$ be iid $E(X_i) = 0$, $E(X_i^2) = \sigma^2$, and $F_n(x) = P(\frac{1}{\sigma n^{\frac{1}{2}}} \sum_{i=1}^{n} X_i < x)$, then if $x > 1$ and $x = O(\sqrt{n})$ as $n \to \infty$ we have*

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

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

2.3 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 $g \in [0, \infty) \ni \int_0^\infty nV_n r^{n-1} g(r)dr = 1$ 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 $(X_1^2, \dots, X_n^2) =_{dist} (\frac{Y_1}{\kappa}, \dots, \frac{Y_n}{\kappa})$ where Y_i iid $\Gamma(\frac{1}{2})$ with sum κ. If $N_1, \dots, N_n$ iid normal, then $(N_1, \dots, N_n)$ 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 $\kappa = (\sum_i = 1^n N_i^2)^{\frac{1}{2}}$
- 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

2.3.1 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.

2.3.2 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 $XX^t = S \sim W(N)$.

2.3.3 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.

2.4 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.

2.5 Distance measures for probability distribution functions.

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

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

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 :

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

which simplifies to:

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

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:

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

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.

2.6 $S_\alpha(\sigma, \beta, \mu)$ Stable Random Variates

A Levy process is a stochastic process with a drift, a diffusion, and a jump component. The LvyKhinchine 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.

2.6.1 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
  $$E[e^{i\theta X}] = exp(-\sigma|\theta|(1 + i\frac{2}{\pi}\beta\ sgn(\theta)\ ln|\theta| + i\mu\theta))$$

2.6.2 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

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

.

2.7 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 $\frac{\partial H}{\partial} = logp + 1/p = 0$

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.

Below we present a table of test and results from the KL libraries via a CDH port from the original Fortran Statlib libray. [1]

[1] I don't know who did the port. The following is a comment from the GRASS GIS tutorial on goodness of fit: *"The cdhc library was inspired by Johnson's STATLIB collection of FORTRAN routines for testing distribution assumptions. Some functions in cdhc are loosely based on Johnson's work (they have been completely rewritten, reducing memory requirements and number of computations and fixing a few bugs). Others are based on algorithms found in Applied Statistics, Technometrics, and other related journals."*

Omnibus Moments Test for Normality

Geary's Test of Normality

Calculate Extreme Normal Deviates

D'Agostino's D-Statistic Test of Normality

Kuiper V-Statistic Modified to Test Normality

Watson U^2-Statistic Modified to Test Normality

Durbin's Exact Test (Normal Distribution

Anderson-Darling Statistic Modified to Test Normality

Cramer-Von Mises W^2-Statistic to Test Normality

Kolmogorov-Smirnov D-Statistic to Test Normality

Kolmogorov-Smirnov D-Statistic (Lilliefors Critical Values)

Chi-Square Test of Normality (Equal Probability Classes)

Shapiro-Wilk W Test of Normality for Small Samples

Shapiro-Francia W' Test of Normality for Large Samples

Shapiro-Wilk W Test of Exponentiality

Cramer-Von Mises W^2-Statistic to Test Exponentiality

Kolmogorov-Smirnov D-Statistic to Test Exponentiality

Kuiper V-Statistic Modified to Test Exponentiality

Watson U^2-Statistic Modified to Test Exponentiality

Anderson-Darling Statistic Modified to Test Exponentiality

Chi-Square Test for Exponentiality(with E.P.C.)

Kotz Separate-Families Test for Lognormality vs. Normality

We only discuss the Anderson Darling and Kolmogorov-Smirnov tests.

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.

2.8 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.

Chapter 3

MachineLearning

3.1 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 $\int_{\infty}^{-\infty} \psi(x)dx = 1$, we can form a parametric estimator for the pdf of a process from a sample population of size N by calculating

$$p(x;\theta) = \frac{1}{N\theta}\sum_{i-1}^{N}\phi(\frac{x-X_n}{\theta}) \tag{3.1.1}$$

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

$$Errp(x;\theta) = E[p(x;\theta)-p(x)]^2 = E[p(x;\theta)-p(x)]^2 = (E[p(x;\theta)]-p(x))^2+Var[p(x;\theta)] \tag{3.1.2}$$

3.2 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 $A \in M_{mn}$

$$||\mathbf{A}||_{\mathcal{F}} = \sqrt{\frac{tr(AA^t)}{n}} \tag{3.2.1}$$

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.

3.3 Learning Theory and Functional Analysis

Supervised learning in it's most abstract setting requires finding a function $f(x)$ given instances $(x_i, f(x_i))$. Typical assumptions are that x_i is an 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,

$$R(f') = \int L(f(\omega), f'(\omega))d\omega \tag{3.3.1}$$

. 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, but where ???. Let $W(k)$ be the winnings after k plays of from a game with outcome S that pays 2^{i-1} with probability $p_i = 1/2^i$. $\lim_{k\longrightarrow\infty} W(k)/k = E(S) = \sum\limits_{i=1}^{\infty} p_i 2^{i-1} = \infty$ The implication for a decision theory based only 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.

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).

3.4 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

$$F_n(x) = \frac{1}{n} \sum_{n}^{i=1} \begin{cases} 1 : x_i \leq x \\ 0 : x_i > x \end{cases}$$

, 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.]

DAgostino-Pearson

3.5 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.

3.6 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

$$Y_i = \Lambda(X_i) = b_o + b \circ X_i. \tag{3.6.1}$$

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 $\sum\limits_{i=1}^{n} g(e_i) = \sum\limits_{i=1}^{n} g(y_i - x_i\beta)$, 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

$$\sum_{i=1}^{n} \frac{\partial g(y_i - x_i\beta)}{\partial \beta} = 0 \tag{3.6.2}$$

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

$$\sum_{i=1}^{n} \omega(e_i)(y_i - x_i\beta)x_i = 0 \tag{3.6.3}$$

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;

$$\beta^{\tau} = (X^t W^{(\tau-1)} X)^{-1} X^t W^{\tau-1} y \tag{3.6.4}$$

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.

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

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,

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

Keeping in mind that

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

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,

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

.

A general form of the exponential distribution

$$\rho(x;\theta) = exp(\frac{x\theta - \xi(\theta)}{\sigma})\nu(x) \tag{3.6.5}$$

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

$$\mathcal{L}(\theta) = \sum_{i=1}^{N}[X_i\theta - \xi(\theta) + log(\nu(X_i))] \tag{3.6.6}$$

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

$$\rho'(x_i|x_j j \neq i) = log(\frac{x_i}{1-\sum\limits_{j\neq i} x_j}) \tag{3.6.7}$$

. 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.

3.7 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

$$J(\beta^{i+1}) = argmin \sum w_i(\beta)|y_i - f_i(\beta)| \tag{3.7.1}$$

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 were introduced by Huber as a generalization to maximum likelihood estimation. Instead of trying to minimize the log likelihood

$$L(\theta) = \sum -log(p(x_i, \theta) \tag{3.7.2}$$

Huber proposed minimizing

$$M(\theta) = \sum \rho(x_i, \theta) \tag{3.7.3}$$

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

$$\rho_\eta(e_i) = \begin{cases} \frac{e_i^2}{2} & |e_i| \leq \eta \\ \eta|e_i| - \frac{\eta^2}{2} & |e_i| > \eta \end{cases} \tag{3.7.4}$$

The Tukey Bisquare estimator is given by

$$g_\eta(e_i) = \begin{cases} \frac{\eta^2}{6}(1 - [1 - \frac{e_i}{\eta}{}_2]^3) & |e_i| \leq \eta \\ \frac{\eta^2}{6} & |e_i| > \eta \end{cases} \tag{3.7.5}$$

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. [**?**]

3.8 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 FS schemes the Kolmogorov-Smirnov Test ,the t-test, the f-test, the Wilks Lambda Test and Wilcoxon Rank Sum Test.

Feature subset selection (FSS) is the process of determining which measurements will be used for classification. It's important to distinguish this 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 $RSS(f,\lambda) = \sum\limits_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$. 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 $x_{i i \in 1,...N}$ an $\lambda \to \infty$ constrains f to be linear.

3.9 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; $logit(E(Y_{ij}|X_{ij})) =$

$X_{ij}^{\dagger}\beta$ 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; $logit(P(Y_{ij} = 1|X_{ij})) = X_{ij}^{\dagger}\beta_i$ 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 $logit(P(Y_{ij} = 1|X_{ij}, \beta_i)) = X_{ij}^{\dagger}\beta + OX_{ij}\beta_i$ 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.

3.10 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 $p_i = \int\limits_{y_i - d_i/2}^{y_i + d_i/2} p(x;\theta) \ \ dx$ Expanding $p(x;\theta)$ in a Taylor series about y, $p(x;\theta) = \sum\limits_{i=0}^{\infty} \frac{p^{(i)}(y)}{i!}(x-y)^i$.

3.11 Multidimensional Scaling

Multidimensional scaling (MDS) is an alternative to factor analysis. The aim of MDS and factor analysis of the 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 $S(\xi_1, \ldots, \xi_l) = \left(\sum_{i \neq j} d_{ij} - ||\xi_i - \xi_j||_\rho\right)^{\frac{1}{2}}$.

3.12 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 $\mathbb{R}^m$

3.13 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 3.1: 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.

3.14 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.

3.15 Matrix Factorization

Many forms of matrix factorization can be cast as an optimization problem that involves minimization of generalized Bregman divergences[53]. Factorization algorithms such as NNMF, Weighted SVD, Exponential Family PCA, , pLSI, Bregman co-clustering [8] 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 [25].

3.16 PCA and its generalization to the Exponential Family

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

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

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 [7] 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.

3.17 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 [32]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 [11] a sampling condition is given which bounds the quality of the manifold embedding based on the quality of the neighborhood graph.

3.18 Graph Laplacian

[13], [18], [21], [26], [13], [18], [38], [35]

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. [46] 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.

3.19 Learning With Kernels

[16], [34], [45], [49], [48], [51], [58], [61], [50]

Kernel learning is a paradigm for classification and regression where prior belief is expressed in the construction of a similarity matrix of distanced 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

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

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.

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.

3.20 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 to 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.

3.21 Matrix Factorization via Generalized Bregman Divergences

Many forms of matrix factorization can be cast as an optimization problem that involves minimization of generalized Bregman divergences[53]. Factorization algorithms such as NNMF, Weighted SVD, E xponential Family PCA, , pLSI, Bregman co-clustering [8] 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 [25].

3.22 Spectral Clustering

[14], [22], [33], [44], [47], [60]

3.23 Co-Clustering

[3], [1], [55], [9], [8], [22], [23], [30], [29], [27], [28], [40], []

3.24 Proximity Measurement - central versus pairwise grouping

Central grouping with K-means of GMM via EM relies on the assumption that feature vectors for each group have gaussian distribution. This justifies the use of a Euclidian or Mahalanobis distance metric.

A standard procedure in Machine Learning to to compute a matrix of pairwise similarity measurements that represent proximity of data points. This is a familiar aspect of multivariate feature data where data dimensions effectively capture class representation. When presented with graph or network data, the situation becomes a little more complicated.

Propagating pairwise similarity in a transitive fashion avoids the requirement that all members of a cluster are close to some prototype.

Connection subgraph methods were developed for this purpose.

Chapter 4

Numerical Linear Algebra Background

$M_{m,n}(\mathbb{F})$ 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 $f(x) = \sum\limits_{i=0}^{n-1} a_i x^i$ as $f(x) = (\ldots(a_{n-1}x + a_{n-2})x + \ldots + a_1)x + a_0$

4.1 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:

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

4.2 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;

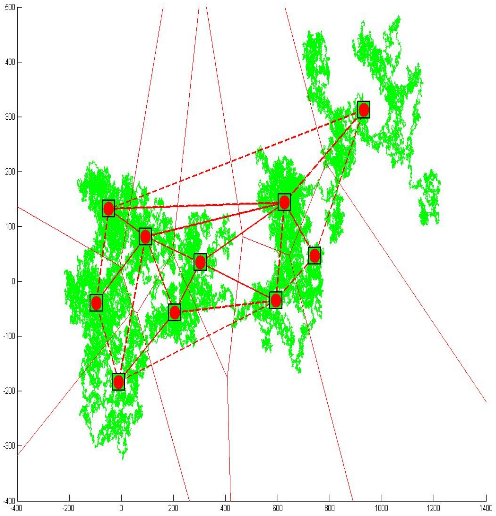
$$\phi_{ij}(x) = \phi(2^{-j}x - i)$$
$$V_j = span\{\phi_{ij}\}$$
$$W_{j+1} = V_j \; V_{j+1}^{\perp}$$
$$\ldots V_{j+1} \subset V_j \subset \ldots \subset V_0 \subset \ldots V_{-j} \subset \ldots$$
$$V_j = V_{j+1} \oplus W_{j+1} x \in W_{j+1} \Rightarrow \; \exists \; a_l \;\; x = \sum_l \{a_l\}\phi_{jl}$$

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

4.3 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.



Chapter 5

Random Matrix Theory

[2], [4], [5] , [20], [54], [57]

The distribution of eigenvalues of the GOE ensemble follow the well know Winger Semi-circle distribution. 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 joint densities up to a constant factor are listed belelo:

- 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 $\left(\begin{array}{cc} X \rightarrow Y & X^{\perp} \rightarrow Y \\ X \rightarrow Y^{\perp} & X^{\perp} \rightarrow Y^{\perp} \end{array} \right)$ Which can be compactly written $[X|X^{\perp}]^T[Y|Y^{\perp}] = \left(\begin{array}{cc} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{array} \right) \left(\begin{array}{cc} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{array} \right) = \left(\begin{array}{cc} U_1 & 0 \\ 0 & U_2 \end{array} \right) * \left(\begin{array}{cc} C & S \\ -S & C \end{array} \right) * \left(\begin{array}{cc} V_1 & 0 \\ 0 & V_2 \end{array} \right)$ 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 $K(n,m) : l^2(N) \rightarrow l^2(M)$ 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 [57] 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

$$\frac{d}{dx} \left(\begin{array}{c} f(x) \\ g(x) \end{array} \right) = \left(\begin{array}{cc} \alpha(x) & \beta(x) \\ -\gamma(x) & -\alpha(x) \end{array} \right) * \left(\begin{array}{c} f(x) \\ g(x) \end{array} \right)$$

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.

Definitions and results form Operator Theory

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

5.0.1 RSK and Young Tableaux - A combinatorial application of The Tracy Widom Distribution - Panleve

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.

This kind of fluctuations arises (or is believed to arise) in a surprising variety of models: eigenvalues of random matrices, , shape fluctuations in first and last passage percolation, polynuclear growth models, frozen region of a random domino tiling of the aztec diamond, totally asymmetric exclusion process

5.1 Generating Random Matrices $A \in U(n), P(n), O(n) \dots$

COSturm99usingsedumi

Bibliography

[1] A general framework for fast co-clustering on large datasets using matrix decomposition.

[2] D Achlioptas. Random matrices in data analysis. In *Proceedings of the 15 th European Conference on Machine Learning*, pages 1–8, 2004.

[3] E Achtert, C Bahm, P Krager, and A Zimek. Mining hierarchies of correlation clusters. In *In Proc. SSDBM*, 2006.

[4] N Alon and B Sudakov. Bipartite subgraphs and the smallest eigenvalue. *Combinatorics, Probability And Computing*, (9):1–12, 2000.

[5] N Alon, M Krivelevich, and V H Vu. On the concentration of eigenvalues of random symmetric matrices. Technical report, Israel J. Math, 2000.

[6] S Arora, S Rao, and U Vazirani. Expander flows, geometric embeddings and graph partitioning. In *In Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 222–231, 2004.

[7] K Azoury and M Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. In *In In Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, 1999.

[8] A Banerjee, I Dhillon, J Ghosh, S Merugu, and D Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *In KDD*, pages 509–514, 2004.

[9] Arindam Banerjee, Inderjit Dhillonjoydeep, and Ghosh Srujana Merugu.

[10] P L Bartlett, M I Jordan, and J D McAuliffe. Convexity, classification, and risk bounds. Technical report, Journal of the American Statistical Association, 2003.

[11] M Bernstein, V de Silva, J C Langford, and J B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, 2000.

[12] S Boucheron, O Bousquet, and G Lugosi. Concentration inequalities. In *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

[13] S Boyd. Convex optimization of graph laplacian eigenvalues. In *in International Congress of Mathematicians*, pages 1311–1319.

[14] M Brand and K Huang. A unifying theorem for spectral embedding and clustering. In *In Proc. of the Ninth International Workshop on AI and Statistics*, 2003.

[15] James C Bremer, Ronald R Coifman, Mauro Maggioni, and Arthur D Szlam. Abstract diffusion wavelet packets.

[16] C J C Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, (2):121–167, 1998.

[17] John Lafferty Carnegie, John Lafferty, and Guy Lebanon. Information diffusion kernels.

[18] F R K Chung. Laplacians of graphs and cheeger's inequalities. In *Proc. Int. Conf. Combinatorics, Paul Erdos is Eighty*, pages 1–16, 1993.

[19] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates and low dimensional representation of stochastic systems.

[20] C Cooper. On the rank of random matrices. *Random Struct. Algorithms*, (16):2000, 2000.

[21] P Crescenzi, R Silvestri, and L Trevisan. To weight or not to weight: Where is the question. In *In Proc. of 4th Israel Symp. on Theory of Computing and Systems*, pages 68–77, 1996.

[22] I S Dhillon, S Mallela, and D S Modha. Information-theoretic co-clustering. In *In KDD*, pages 89–98. ACM Press, 2003.

[23] Yiling Chen Frederico. A bipartite graph co-clustering approach to ontology mapping.

[24] J FRIEDMAN. Computing betti numbers via combinatorial laplacians. In *In Proc. 28th Ann. ACM Sympos. Theory Comput*, pages 386–391, 1996.

[25] G Gordon. Approximate solutions to markov decision processes. Technical report, 1999.

[26] S Guattery and G L Miller. Graph embeddings and laplacian eigenvalues. *SIAM Journal on Matrix Analysis and Applications*, (21).

[27] T Hofmann. Probabilistic latent semantic indexing. pages 50–57, 1999.

[28] Tianming Hu and Sam Yuan Sung. Preserving patterns in bipartite graph partitioning.

[29] Tianming Hu, Chew Lim Tan, Yong Tang, Sam Yuan Sung, Hui Xiong, and Chao Qu. Co-clustering bipartite with pattern preservation for topic extraction.

[30] J Huang, T Zhu, R Greiner, D Zhou, and D Schuurmans. Information marginalization on subgraphs. In *In PKDD*, 2006.

[31] Peter J. Huber. Projectionpursuit. *Source: Ann. Statist.*, 13, 1985.

[32] et al. Joshua B. Tenenbaum. A global geometric framework for nonlinear dimensionality reduction. *Science 290, 2319*, (290), 2000.

[33] R Kannan, S Vempala, and A Vetta. On clusterings: good, bad and spectral. In *Journal of the ACM*, pages 367–377, 2000.

[34] S S Keerthi, S K Shevade, C Bhattacharyya, and K R K Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, (13):637–649, 1999.

[35] Mitchel T. Keller. Signed graph laplacians.

[36] T Kubota and F Espinal. Reaction-diffusion systems for hypothesis propagation. In *In Int. Conf. Pattern Recognition*, page 3547, 2000.

[37] J Lafferty and G Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, (6):2005, 2005.

[38] M Ledoux. Structural, syntactic, and statistical pattern recognition, joint iapr international workshops, sspr 2004 and spr 2004, lisbon, portugal, august 18-20, 2004 proceedings. In *SSPR/SPR*. Springer, 2004.

[39] M Ledoux. Spectral gap, logarithmic sobolev constant, and geometric bounds. In *Surveys in Diff. Geom., Vol. IX, 219240, Int*, page 2195409. Press, 2004.

[40] S Merugu, A Banerjee, and S Basu. Multi-way clustering on relation graphs. Technical report, In Proc. of the 7th SIAM Intl. Conf. on Data Mining, 2006.

[41] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20:801–836, 1978.

[42] Pinar Muyan and Nando De Freitas. A blessing of dimensionality: Measure concentration and probabilistic inference.

[43] B Nadler, S Lafon, R R Coifman, and I G Kevrekidis. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. In *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, 2006.

[44] A Ng, M Jordan, and Y Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.

[45] With Semi-Definite Programming, Gert Lanckriet, Nello Cristianini, and Laurent El Ghaoui. Learning the kernel matrix.

[46] Edwin R HANCOCK Richard C WILSON. Spectral analysis of complex laplacian matrices. In *Structural, syntactic, and statistical pattern recognition: Joint IAPR international workshops, SSPR 2004 and SPR 2004, Lisbon, Portugal, August 18-20, 2004*, 2004.

[47] B Sch?lkopf, A Smola, and K-R M?ller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, (10):1299–1319, 1998.

[48] B Scholkopf. Statistical learning and kernel methods. Technical report, 2000.

[49] B Scholkopf, A Smola, R C Williamson, and P L Bartlett. New support vector algorithms. *Neural Computation*, (12):112–1, 2000.

[50] M Schultz and T Joachims. Learning a distance metric from relative comparisons. In *In NIPS*. MIT Press, 2003.

[51] S K Shevade, S S Keerthi, C Bhattacharyya, and K R K Murthy. Improvements to smo algorithm for svm regression. Technical report, IEEE Transactions on Neural Networks, 1999.

[52] A Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Prob., Comput*, (1), 1992.

[53] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases (Proc. ECML PKDD), volume 5212/2008 of Lecture Notes in Computer Science, pages 358-373. Springer Berlin / Heidelberg, 2008*, 2008.

[54] A Soshnikov. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Statist. Phys*, (108): 1033–1056, 2002.

[55] For Summarization, , Krishna Kummamuru, and Karan Singal. A hierarchical monothetic document clustering algorithm.

[56] M Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. IHES*, (81):73–205, 1995.

[57] C A Tracy and H Widom. Correlation functions, cluster functionsand spacing distribution for random matrices. *J. Statist. Phys*, (92), 1998.

[58] I W Tsang and J T Kwok. Distance metric learning with kernels. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 126–129, 2003.

[59] Lieven Vandenberghe, Stephen Boyd, and Katherine Comanor. Generalized chebyshev bounds via semidefinite programming. *SIAM Review*, 49.

[60] Y Weiss. Segmentation using eigenvectors: a unifying view. In *In International Conference on Computer Vision*, pages 975–982, 1999.

[61] J Weston, S Mukherjee, O Chapelle, M Pontil, T Poggio, and V Vapnik. Feature selection for svms. In *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2000.