Advanced Analytics

Business Analytics – Business Intelligence – Predictive Analytics – Predictive Intelligence

Review by L. Querella, PhD, MSc, Data Scientist, Mar 2023

Contents

1 Machine Learning

1.1 Louis Dorard

1.1.1 How to improve your business by predicting churn

Example of a company that sells SaaS for a monthly fee.

Three possible strategies to increase the revenue:

1. Acquire more customers
2. Upsell existing customers
3. Increase customer retention

Actions taken for these strategies have a cost; one is interested in the ROI (difference between cost and extra revenue).

Churn: is this customer going to leave us within the next X months?

== binary classification (yes/no)

Process:

1. Gather historical customer data (csv file)
2. Upload these data to a prediction service (automatically creates a model)
3. Use the model for each customer

Feature types:

- Customer features (age, income, education, etc.)
- Support features (characterisations of interaction with customer support: nbr of calls, etc.)
- Usage features (characterisations of customer usage of the service)
- Contextual features (misc)

Time frames:

- Features average over last 2-3 months e.g.

DATA EXTRACTION

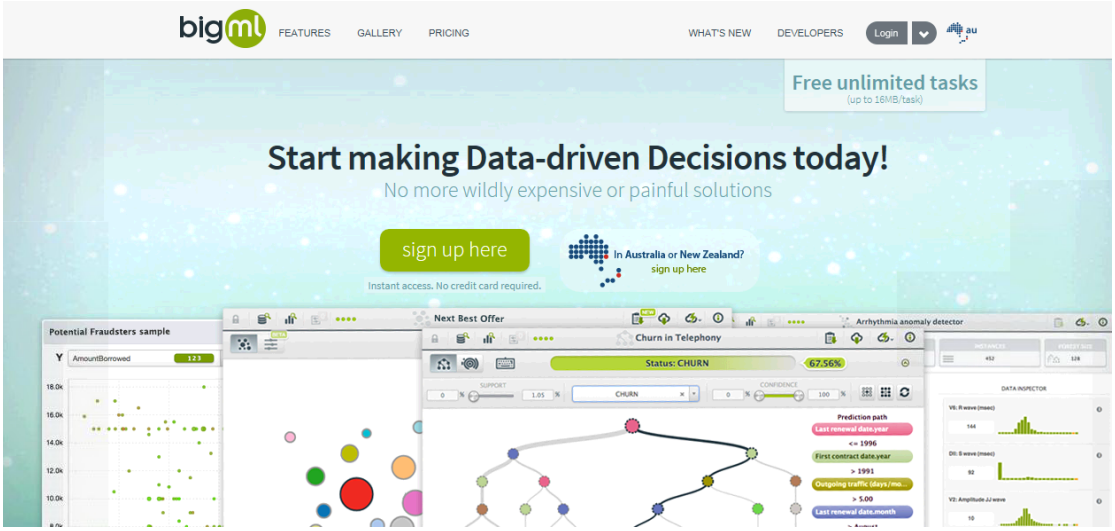
Example in csv:

Talk	Text	Purchases	Data	Age	Churn?
148	72	0	33.6	50	TRUE
85	66	0	26.6	31	FALSE
183	64	0	23.3	32	TRUE

DATA UPLOAD

Prediction Services (via WebUI or API):

- BigML



- Google Prediction API

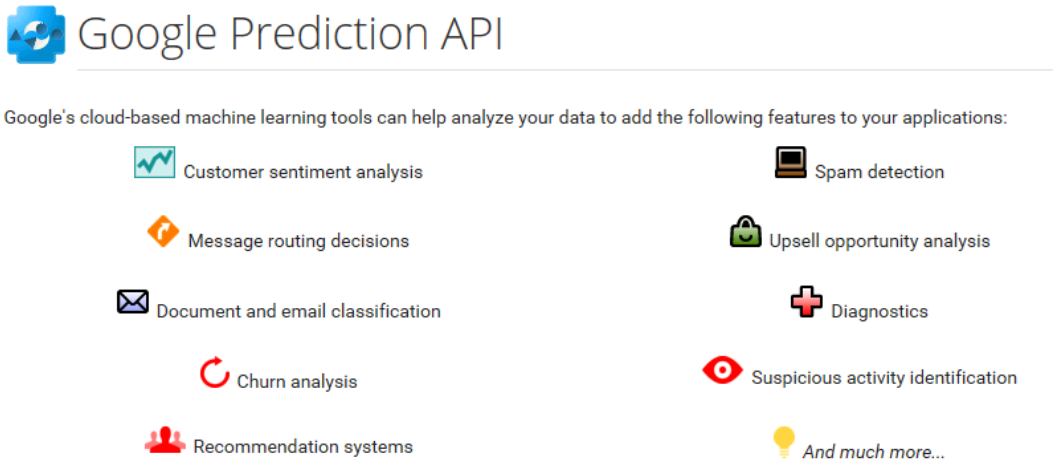


ILLUSTRATION WITH BIGML

(…)

The take-away message for us is that you don't need more than a service like BigML to do churn prediction and to start exploiting the value of your business's data.

1.1.2 When Machine Learning fails

ML is a set of AI techniques where intelligence is built by referring to examples.

Providing more examples works usually better than using more sophisticated algorithms.

Noise perturbs the ability to learn (things look more random).

1.1.3 Bootstrapping Machine Learning

Sample of the book

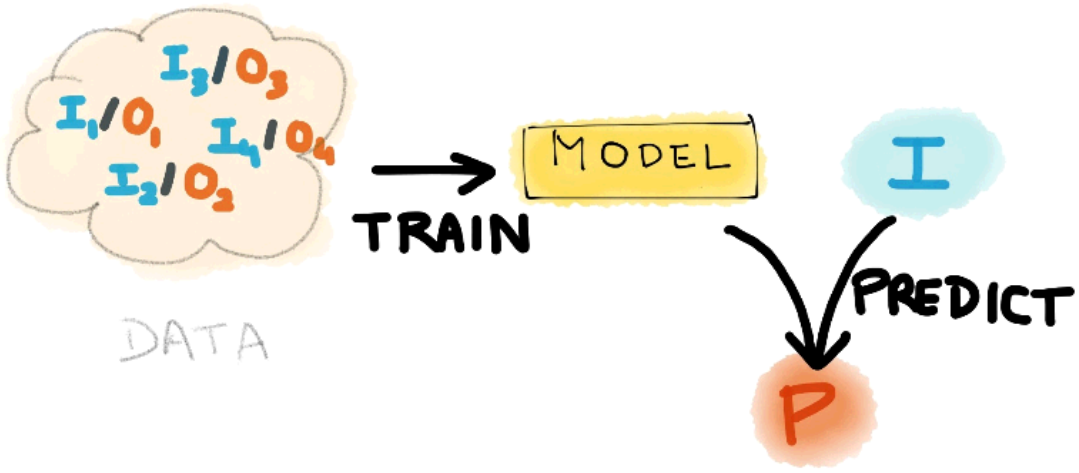
Main goal: using ML via Prediction APIs without any knowledge/expertise in ML algorithms.

Framework for ML:

- **Who**: who is using the application / where is it being used?
- **Description**: what is the context and what are we trying to do?
- **Question asked**: how would you write, in plain English, the question that the predictive model should give answers to?
- **Input**: what are we doing predictions on?
- **Features**: which aspects of the input are we considering / what kind of information do we have in its representation?
- **Output**: what does the predictive model return?
- **Data collection**: how are example input-output pairs obtained to train the predictive model?
- **Predictions**: when are predictions being made and what do we do once we have them?
- **Value**: how is value created for the end user?

1.1.4 Machine Learning as a Service (MLaaS)

Predictive models can be automatically created from data you send (automatic choice of algorithms and parameters) and they are run on the cloud — no need to worry about infrastructure. MLaaS can either be accessed programmatically through an API, or graphically through a web interface, which makes it usable both by programmers and non-technical types.



Example of code to use BigML API (Python):

```
from bigml.api import BigML

# create a model
api = BigML()
source = api.create_source('training_data.csv')
dataset = api.create_dataset(source)
model = api.create_model(dataset)

# make a prediction
prediction = api.create_prediction(model, new_input)
print "Predicted output value: ",prediction['object']['output']
```

Check IPython notebook on Wakari

1.2 Deep Learning

See

- Deep Learning in a Nutshell - what it is, how it works, why care?
- Where to Learn Deep Learning - Courses, Tutorials, Software
- Deep Learning - important resources for learning and understanding
- KDnuggets stories Tagged: Deep Learning

1.3 Top 10 data mining algorithms

http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

2 Information Geometry

2.1 Definition

http://en.wikipedia.org/wiki/Information_geometry

Information geometry is a branch of mathematics that applies the techniques of differential geometry to the field of probability theory. This is done by taking probability distributions for a statistical model as the points of a Riemannian manifold, forming a *statistical manifold*.

The Fisher information metric provides the Riemannian metric.

Information geometry reached maturity through the work of S. Amari and other Japanese mathematicians in the 1980s. Amari and Nagaoka's book, Methods of Information Geometry,[1] is cited by most works of the relatively young field due to its broad coverage of significant developments attained using the methods of information geometry up to the year 2000.

http://bactra.org/notebooks/info-geo.html

This [Information geometry] is a slightly misleading name for applying differential geometry to families of probability distributions, and so to statistical models. Information does however play two roles in it: **Kullback-Leibler information**, or **relative entropy**, features as a measure of divergence (not quite a metric, because it's asymmetric), and **Fisher information** takes the role of **curvature**. One very nice thing about information geometry is that it gives us very strong tools for proving results about statistical models, simply by considering them as well-behaved geometrical objects. Thus, for instance, it's basically a tautology to say that a manifold is not changing much in the vicinity of points of low curvature, and changing greatly near points of high curvature. Stated more precisely, and then translated back into probabilistic language, this becomes the **Cramer-Rao inequality**, that the variance of a parameter estimator is at least the reciprocal of the Fisher information.

2.1.1 Information and probability

2.1.1.1 Entropy (information theory)

http://en.wikipedia.org/wiki/Entropy_(information_theory)

In information theory, entropy is the average amount of information contained in each message received. Here, message stands for an event, sample or character drawn from a distribution or data stream. Entropy thus characterizes our uncertainty about our source of information. (Entropy is best understood as a measure of uncertainty rather than certainty as entropy is larger for more random sources). The source is also characterized by the probability distribution of the samples drawn from it. The idea here is that the less likely an event is, the more information it provides when it occurs. For some other reasons (explained below) it makes sense to define information as the negative of the logarithm of the probability distribution. The probability distribution of the events, coupled with the information amount of every event, forms a random variable whose average (a.k.a. expected) value is the average amount of information, a.k.a. entropy, generated by this distribution. The units of entropy are commonly referred to as bits, but entropy is also measured in shannons, nats, or hartleys, depending on the base of the logarithm used to define it.

The logarithm of the probability distribution is useful as a measure of information because it is additive. For instance, flipping a coin provides 1 Shannon of information whereas m tosses gather m bits. Generally, you need log2(n) bits to represent a variable that can take one of n values. Since 1 of n outcomes is possible when you apply a scale graduated with n marks, you receive log2(n) bits of information with every such measurement. The log2(n) rule holds only while all outcomes are equally probable. If one of the events occurs more often than others, observation of that event is less informative. Conversely, observing rarer events compensate by providing more information when observed. Since observation of less probable events occurs more rarely, the net effect is that the entropy (thought of as the average information) received from non-uniformly distributed data is less than log2(n). Entropy is zero when only one certain outcome is expected. Shannon entropy quantifies all these considerations exactly when a probability distribution of the source is provided. It is important to note that the meaning of the events observed (a.k.a. the meaning of messages) do not matter in the definition of entropy. Entropy only takes into account the probability of observing a specific event, so the information it encapsulates is information about the underlying probability distribution, not the meaning of the events themselves.

Generally, "entropy" stands for "disorder" or uncertainty. The entropy we talk about here was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication".[1] We also call it Shannon entropy to distinguish from other occurrences of the term, which appears in various parts of physics in different forms. Shannon entropy provides an absolute limit on the best possible average length of lossless encoding or compression of any communication, assuming that[2] the communication may be represented as a sequence of independent and identically distributed random variables.

Definition of the entropy of a random variable in terms of information content (Shannon's).

Information content (probability distribution p(v) of a random variable V):

$$I(v) = -\log_2 p(v)$$

Entropy of a random variable V with occurrences v:

$$H(V) = -\sum p(v) \log p(v)$$

2.1.1.2 Statistical model, parameters

Observation context (depicted by statistical model) identified by a set of parameters ksi (n-dim). Every ksi determines one probability distribution for V.

Mixture distribution:

A parameterization of the form

$$p(v) = \sum \xi^i p_i(v) = \xi^i p_i$$

with

$$\sum p_i(v_j) = 1 \text{ and } \sum \xi^i = 1,$$

that mixes different distributions, is called a mixture distribution.

All such parameterizations are related through an affine transformation

$$\rho = A\xi + B$$

A parameterization with such a transformation rule is called flat.

Exponential family/class:

http://en.wikipedia.org/wiki/Exponential_family

= set of probability distributions of a certain form:

$$p(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right)$$

Or alternatively,

$$f_X(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})\cdot\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})\right)$$

Where

- Theta are the natural parameters
- T(x) is a *sufficient statistic* of the distribution
    - The dimension of T(x) equals the number of parameters of θ and encompasses all of the information regarding the data related to the parameter θ
- η is called the natural parameter
- A(η) is called the log-partition function because it is the logarithm of a normalization factor

Exponential families have a large number of properties that make them extremely useful for statistical analysis.

Example:

As a first example, consider a random variable distributed normally with unknown mean μ and *known* variance σ^2. The probability density function is then

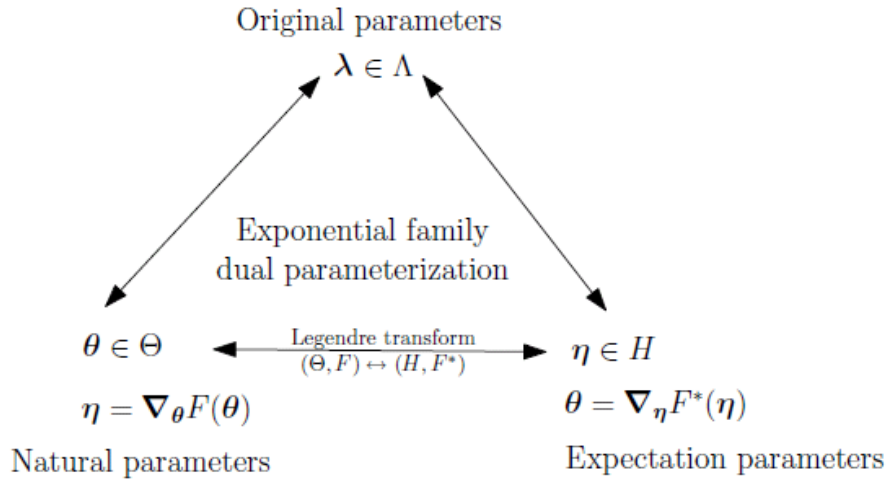
$$f_\sigma(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This is a single-parameter exponential family, as can be seen by setting

$$h_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$
$$T_\sigma(x) = \frac{x}{\sigma}$$
$$A_\sigma(\mu) = \frac{\mu^2}{2\sigma^2}$$
$$\eta_\sigma(\mu) = \frac{\mu}{\sigma}.$$

If σ = 1 this is in canonical form, as then $\eta(\mu) = \mu$.

Dual parameterizations



Geometry of exponential families: Riemannian and information geometries

A family of parametric distributions {p(x;θ)} (exponential or not) may be thought as a smooth manifold that can be studied using the framework of differential geometry [Lau87]. We review two main types of geometries: (1) Riemannian geometry defined by a bilinear tensor with an induced Levi-Cevita connection, and non-metric geometry induced by a symmetric affine connection.

The only Riemannian metric that "makes sense" for statistical manifolds is the Fisher information metric:

$$I(\theta) = \left[\int \frac{\partial \log p(x;\theta)}{\partial \theta_i} \frac{\partial \log p(x;\theta)}{\partial \theta_j} p(x;\theta) \mathrm{d}x \right] = [g_{ij}]$$

The infinitesimal length element is given by

$$\mathrm{d}s^2 = \sum_{i=1}^{d} \sum_{i=1}^{d} \mathrm{d}\theta_i^T \nabla^2 F(\theta) \mathrm{d}\theta_j$$

Equipped with the tensor I(θ), the metric distance between two distributions on a statistical manifold can be computed from the geodesic length (e.g., shortest path):

The Fisher information matrix can be interpreted as the Hessian of the Shannon entropy:

$$g_{ij}(\theta) = -E\left[\frac{\partial^2 \log p(x;\theta)}{\partial \theta_i \partial \theta_j} \right] = \frac{\partial^2 H(p)}{\partial \theta_i \partial \theta_j},$$

with $H(p) = -\int p(x;\theta) \log p(x;\theta) \mathrm{d}x$.

For an exponential family, the Kullback-Leibler divergence is a Bregman divergence on the natural parameters. Using Taylor approximation with exact remainder, we get $\mathrm{KL}(\theta||\theta + \mathrm{d}\theta) = \frac{1}{2}\mathrm{d}\theta^T \nabla^2 F(\theta) \mathrm{d}\theta$. Moreover, the infinitesimal Rao distance is $\sqrt{\mathrm{d}\theta^T I(\theta) \mathrm{d}\theta}$ for $I(\theta) = \nabla^2 F(\theta)$. We deduce that $D(\theta, \theta + \mathrm{d}\theta) = \sqrt{2\mathrm{KL}(\theta||\theta + \mathrm{d}\theta)}$.

Furthermore, in an exponential family manifold, the geometry is flat and theta/eta are dual coordinate systems.

Amari [iAN00] focused on a pair of dual affine mixture/exponential connections ∇m and ∇e induced by a contrast function F (also called potential function).

(…)

Bregman divergences are the canonical divergences of dually flat Riemannian manifolds [iAN00]. Bregman divergences extend naturally the quadratic distances (squared Euclidean and squared Mahalanobis distances), and generalize the notions of orthogonality, projection, and Pythagoras' theorem.

Banerjee et al. [BMDG05] formally proved the duality between exponential families and Bregman divergences for regular exponential families using Legendre transform:

2.2 PhD Course on Information Geometry and Machine Learning (Copenhague, 2014)

2.2.1 Amari

We consider the manifolds of probability distributions on which dual affine connections are defined. These are Riemannian manifolds; divergence between two points very close to each other is given in terms of a Riemannian metric tensor g_ij.

Riemannian Structure

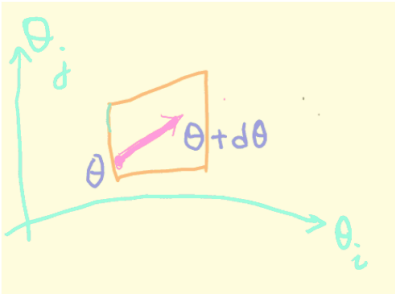
$$ds^2 = \sum g_{ij}(\theta) d\theta^i d\theta^j = d\theta^T G(\theta) d\theta$$

Fisher information

$$g_{ij} = E\left[\frac{\partial}{\partial \theta_i}\log p \frac{\partial}{\partial \theta_j}\log p\right]$$

$$G(\theta) = (g_{ij})$$

Euclidean $G = E$



K-L divergence:

$$D[p(x) : q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Bregman divergence and exponential families

2.3 Information Criteria and Statistical Modeling (Springer, 2008)

Konishi S., Kitagawa G. (2008), *Information Criteria and Statistical Modeling*, Springer

2.3.1 Concept of Statistical Modelling

2.3.1.1 Role of statistical models

The key to solving complex real-world problems lies in the development and construction of a suitable model.

Focus on model not on examples (“more data is better than algorithm”).

2.3.1.1.1 Description of stochastic structures by statistical models

A **statistical model** is a probability distribution that uses observed data to approximate the true distribution of probabilistic events. As such, the purpose of statistical modeling is to construct a

model that approximates the true structure as accurately as possible through the use of available data.

These models must be considered as an approximation that represents only one aspect of complex phenomena. The important issue here is whether we should pursue a structure that is as close as possible to the true model.

2.3.1.1.2 Predictions by statistical models

Akaike considered that the purpose of statistical modeling is not to accurately describe current data or to infer the "true distribution." Rather, he thought that the purpose of statistical modeling is to predict future data as accurately as possible. In this book, we refer to this viewpoint as the *predictive point of view*.

There may be no significant difference between the point of view of inferring the true structure and that of making a prediction if an infinitely large quantity of data is available or if the data are noiseless. However, in modelling based on a finite quantity of real data, there is a significant gap between these two points of view, because an optimal model for prediction purposes may differ from one obtained by estimating the "true model."

In fact, as indicated by the information criteria given in this book for evaluating models intended for making predictions, simple models, even those containing biases, are often capable of giving better predictive distributions than models obtained by estimating the true structure.

2.3.1.1.3 Extraction of information by statistical models

A recent trend that has been gaining popularity is the idea that models are tools of convenience that are used for extracting information and discovering knowledge. In this viewpoint, a statistical model is not something that exists in the objective world; rather, it is something that is constructed based on the prior knowledge and expectations of the analyst concerning the modeling objective, e.g., his knowledge based on past experience and data and based on the purpose of the analysis, such as the specific type of information to be extracted from the data and what is to be accomplished by the analysis.

2.3.1.2 Constructing statistical models

2.3.1.2.1 Evaluation of statistical models – Road to the Information Criterion

If the role of a statistical model is understood as being a tool for extracting information, it follows that a model is not something that is uniquely determined for a given object, but rather that it can assume a variety of forms depending on the viewpoint of the modeler and the available information. In other words, the purpose of statistical modeling is not to estimate or identify the "unique" or "perfect" model, but rather to construct a "good" model as a tool for extracting information according to the characteristics of the object and the purpose of the modelling.

Herein lies the importance of model evaluation criteria for assessing the "goodness" of a subjective model.

How shall we set about evaluating the goodness of a model? In considering the circumstances under which statistical models are actually used, Akaike considered that a model should be evaluated in terms of the goodness of the results when the model is used for prediction. Furthermore, for the

general evaluation of the goodness of a statistical model, he thought that it is important to assess the closeness between the predictive distribution f(x) defined by the model and the true distribution g(x), rather than simply minimizing the prediction error. Based on this concept, he proposed evaluating statistical models in terms of **Kullback–Leibler** information (divergence) [Akaike (1973)].

In this book, we refer to the model evaluation criterion derived from this fundamental model evaluation concept based on Kullback–Leibler information as the **information criterion**. This information criterion is derived from three fundamental concepts:

(1) a prediction-based viewpoint of modeling
(2) evaluation of prediction accuracy in terms of distributions
(3) evaluation of the closeness of distributions in terms of Kullback–Leibler information

2.3.1.2.2 Modeling methodology

The information criterion suggests several concrete methods for developing good models based on a limited quantity of data. First, it is obvious that the larger its log-likelihood, the better the model. The information criterion indicates, however, that given a finite quantity of data available for modeling, a model having an excessively high degrees of freedom will lead to an increase in the instability of the estimated model (ndlr, overfitting), and this will result in a reduced prediction ability. In other words, it is not beneficial to needlessly increase the number of free parameters without any restriction. Under these considerations, several methods are appropriate for assessing a good model based on a given set of data.

Point estimation and model selection

Applying the information criterion directly to determine the number of unknown parameters to be estimated and to select the specific model to use.

Regularization and Bayesian modelling

Another method for obtaining a good model involves imposing appropriate restrictions on parameters using a large number of parameters, without restricting the number of parameters.

It has been suggested that in many cases these model construction methods can be implemented in terms of a Bayesian model that combines information from prior distribution and data. In Bayesian modeling, a model can be constructed by obtaining the posterior distribution

$$\pi(\theta|\boldsymbol{x}_n) = \frac{f(\boldsymbol{x}_n|\theta)\pi(\theta)}{\int f(\boldsymbol{x}_n|\theta)\pi(\theta)d\theta}$$

by introducing an appropriate prior distribution π(θ) for an unknown parameter vector θ that defines the data distribution f(x|θ).

Hierarchical Bayesian modelling

(Generalizing Bayesian modelling.)

2.3.2 Statistical Models

2.3.2.1 Modeling of probabilistic events and statistical models

Given a random variable X defined on the sample space Ω, for any real value x(∈ R), the probability Pr({ω ∈ Ω ;X(ω) ≤ x}) of an event such that X(ω) ≤ x can be determined. If we regard such a probability as a function of x and express it as

$$G(x) = Pr(\{\omega \in \Omega ; X(\omega) \leq x\}) = Pr(X \leq x)$$

then the function G(x) is referred to as the distribution function of X.

In particular, if there exists a nonnegative function g(t) ≥ 0 that satisfies

$$G(x) = \int_{-\infty}^{x} g(t)dt$$

then X is said to be continuous, and the function g(t) is called a probability density function. A continuous probability distribution can be defined by determining the density function g(t).

On the other hand, if the random variable X takes either a finite or a countably infinite number of discrete values x1, x2, . . ., then the variable X is said to be discrete. The probability of taking a discrete point X = xi is determined by

$$\begin{aligned} g_i = g(x_i) &= \Pr(\{\omega \in \Omega\,; X(\omega) = x_i\}) \\ &= \Pr(X = x_i), \qquad\qquad i = 1, 2, \ldots \end{aligned}$$

where g(x) is called a probability function, for which the distribution function is given by

$$G(x) = \sum_{\{i; x_i \leq x\}} g(x_i)$$

Where the Sum represents the sum of the discrete values such that xi ≤ x.

If we assume that the observations xn = {x1, x2, . . . , xn} are generated from the distribution function G(x), then G(x) is referred to as the *true distribution*, or the *true model*. On the other hand, the distribution function F(x) used to approximate the true distribution is referred to as a *model* and is assumed to have either a density function or a probability function f(x). If a model is specified by p-dimensional parameters θ = (θ1, θ2, . . . , θp)T , then the model can be written as f(x|θ). If the parameters are represented as a point in the set Θ ⊂ Rp, then {f(x|θ); θ ∈ Θ} is called a *parametric family of probability distributions or models*.

An estimated model f(x|ˆθ) obtained by replacing an unknown parameter θ with an estimator ˆθ is referred to as a *statistical model*. The process of constructing a model that appropriately represents some phenomenon is referred to as *modeling*. In statistical modeling, it is necessary to estimate unknown parameters. However, setting up an appropriate family of probability models prior to estimating the parameters is of greater importance.

2.3.2.2 Probability distribution models

Examples:

- Normal or Gaussian
- Cauchy
- Laplace
- Pearson's family
- Mixture of normal distribution models
- Binomial (Bernoulli distribution)
- Poisson
    - Rare events in short intervals
- Histogram
- Probability

Multivariate random vector X

- Multivariate normal distribution
- Multinomial

2.3.2.3 Conditional distribution models

From the viewpoint of statistical modeling, the probability distribution is the most fundamental model in the situation in which the distribution of the random variable X is independent of various other factors. In practice, however, information associated with these variables can be used in various ways. The essence of statistical modeling lies in finding such information and incorporating it into a model in an appropriate form. In the following, we consider cases in which a random variable depends on other variables, on past history, on a spatial pattern, or on prior information. The important thing is that such modeling approaches can be considered as essentially estimating conditional distributions. Thus, the essence of statistical modeling can be thought of as obtaining an appropriate conditional distribution.

In general, if the distribution of the random variable Y is determined in a manner that depends on a p-dimensional variable x = (x1, x2, . . . , xp)T , then the distribution of Y is expressed as F(y|x) or f(y|x), and this is called a *conditional distribution model*. There are several ways in which the random variable depends on the other variables x. In the following, we consider typical conditional distribution models.

2.3.2.3.1 Regression models

- Linear
    - In the linear regression model, the critical issue is to determine a set of explanatory variables that appropriately describes changes in the distribution of the response variable y; this problem is referred to as the variable selection problem
- Polynomial
    - In a polynomial regression model, the crucial task is determining the order m, which is referred to as the order selection problem
- Nonlinear
    - Splines, kernel functions, neural networks,etc.
- Changing variance (and constant mean)
    - Earthquakes, financial data

Generally, a regression model is composed of a model that approximates the mean function E[Y |x] representing the structure of phenomenon and a probability distribution model that describes the probabilistic fluctuation of the data.

In the case of a regression model expressed by a density function, we estimate the parameter vector θ of the model by using the maximum likelihood method, and we denote it as ˆθ. Then the density function in which the unknown parameters are replaced with their corresponding estimators, is called a *statistical model*.

In general, a model that is too complex overadjusts for the random fluctuation in the data, while, on the other hand, overly simplistic models fail to adequately describe the structure of the phenomenon being modeled. Therefore, the key to evaluating a model is to strike a balance between, badness of fit of the data and the model complexity (**bias/variance tradeoff**, ndlr).

Spline functions (…)

2.3.2.3.2 Time series model

Observed data, x1, . . . , xN, for events that vary with time are referred to as a time series. The vast majority of real-world data, including meteorological data, environmental data, financial or economic data, and time-dependent experimental data, constitutes time series. The main aim of time series analysis is to identify the structure of the phenomenon represented by a sequence of measurements and to predict future observations. To analyze such time series data, we consider the conditional distribution f(xn|xn−1, xn−2, . . .), given observations up to the time n – 1.

- AR model and ARMA model
    - *autoregressive* (AR) model – linear structure in finite dimensions

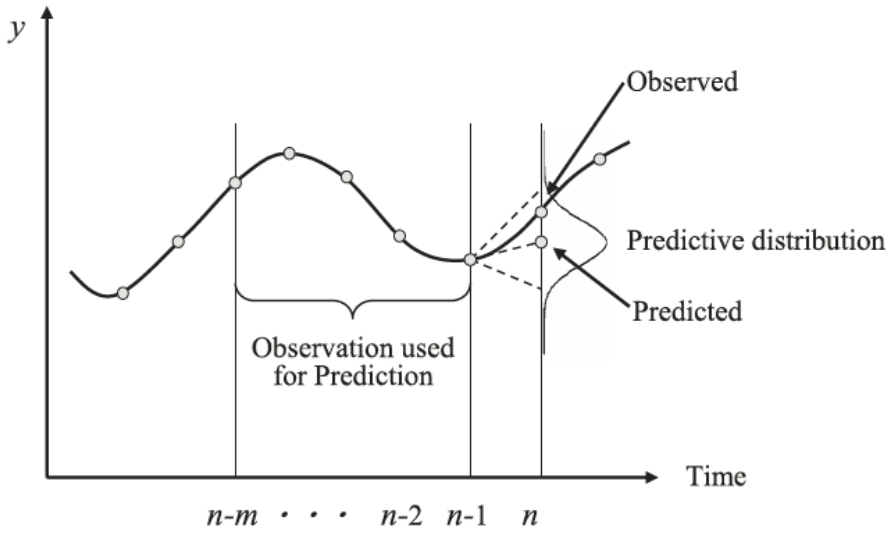


Fig. 2.9. Predictive distribution of time series.

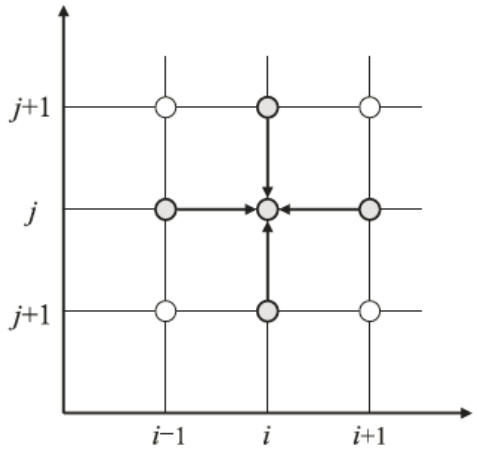
    - *autoregressive moving average* (ARMA) model
- nonlinear models
- multivariate time series models
- state-space models

2.3.2.3.3 Spatial models

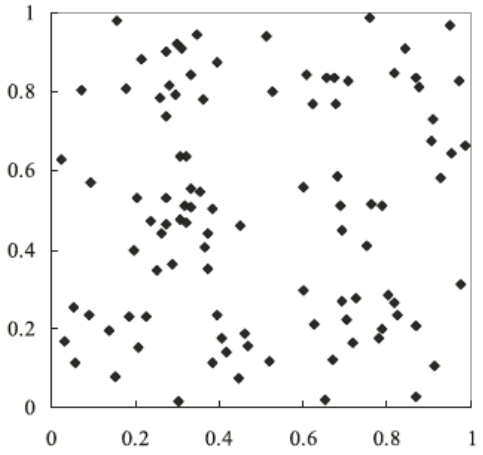
The spatial model represents the distribution of data by associating a spatial arrangement with it. For the case when data are arranged in a regular lattice, as depicted in the figure, a model such as

$$p(x_{ij} \mid x_{i,j-1}, x_{i,j+1}, x_{i-1,j}, x_{i+1,j})$$

that represents the data xij at point (i, j), for example, can be constructed as a conditional distribution of the surrounding four points.



On the other hand, in the general case in which the pointwise arrangement of data is not necessarily a lattice pattern, as illustrated in the figure, a model that describes an equilibrium state can be obtained by modeling the local interaction of the points called particles.



2.3.3 Information Criterion

In this chapter, we discuss using Kullback–Leibler information as a criterion for evaluating statistical models that approximate the true probability distribution of the data and its properties. We also explain how this criterion for evaluating statistical models leads to the concept of the **information criterion, AIC**. To this end, we explain the basic framework of model evaluation and the derivation of AIC by adopting a unified approach.

2.3.3.1 Kullback-Leibler Information

2.3.3.1.1 Definition and properties

Let $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n observations drawn randomly (independently) from an unknown probability distribution function $G(x)$. In the following, we refer to the probability distribution function $G(x)$ that generates data as the true model or the true distribution. In contrast, let $F(x)$ be an arbitrarily specified model. If the probability distribution functions $G(x)$ and $F(x)$ have density functions $g(x)$ and $f(x)$, respectively, then they are called *continuous models* (or *continuous distribution models*). If, given either a finite set or a countably infinite set of discrete points $\{x_1, x_2, \ldots, x_k, \ldots\}$, they are expressed as probabilities of events

$$\begin{aligned} g_i &= g(x_i) \equiv \Pr(\{\omega;\ X(\omega) = x_i\}), \\ f_i &= f(x_i) \equiv \Pr(\{\omega;\ X(\omega) = x_i\}), \quad i = 1, 2, \ldots, \end{aligned} \tag{3.1}$$

then these models are called *discrete models* (*discrete distribution models*).

We assume that the goodness of the model f(x) is assessed in terms of the closeness as a probability distribution to the true distribution g(x). As a measure of this closeness, Akaike (1973) proposed the use of the following **Kullback–Leibler information** [or *Kullback–Leibler divergence*, Kullback–Leibler (1951), hereinafter abbreviated as "K-L information"]:

$$I(G;F) = E_G\left[\log\left\{\frac{G(X)}{F(X)}\right\}\right]$$

where E_G represents the expectation with respect to the probability distribution G.

By unifying the continuous and discrete models, we can express the K-L information as follows:

$$\begin{aligned} I(g\,;f) &= \int \log\left\{\frac{g(x)}{f(x)}\right\} dG(x) \\ &= \begin{cases} \displaystyle\int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx, & \text{for continuous model,} \\ \displaystyle\sum_{i=1}^{\infty} g(x_i)\log\left\{\frac{g(x_i)}{f(x_i)}\right\}, & \text{for discrete model.} \end{cases} \end{aligned} \tag{3.5}$$

The smaller the quantity of K-L information, the closer the model f(x) is to g(x):

Properties of K-L information. The K-L information has the following properties:

(i) $I(g;f) \geq 0$,
(ii) $I(g;f) = 0 \iff g(x) = f(x)$.

In view of these properties, we consider that the smaller the quantity of K-L information, the closer the model $f(x)$ is to $g(x)$.

Measures of the similarity between distributions (other than K-L):

$$\chi^2(g;f) = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i} \qquad \chi^2\text{-statistics},$$

$$I_K(g;f) = \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx \qquad \text{Hellinger distance},$$

$$I_\lambda(g;f) = \frac{1}{\lambda} \int \left\{ \left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right\} g(x) dx \qquad \text{Generalized information},$$

$$D(g;f) = \int u\left(\frac{g(x)}{f(x)} \right) g(x) dx \qquad \text{Divergence},$$

$$L_1(g;f) = \int |g(x) - f(x)| dx \qquad L^1\text{-norm},$$

$$L_2(g;f) = \int \{g(x) - f(x)\}^2 dx \qquad L^2\text{-norm}.$$

In this book, following Akaike (1973), the model evaluation criterion based on the **K-L information** will be referred to generically as an **information criterion**.

2.3.3.1.2 Examples of K-L information

- K-L information for normal models

$$I(g\,;f) = E_G\left[\log g(X)\right] - E_G\left[\log f(X)\right]$$

$$= \frac{1}{2}\left\{ \log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\xi - \mu)^2}{\sigma^2} - 1 \right\}$$

- K-L information for normal and Laplace models
  (…)
- K-L information for two discrete models
  (…)

2.3.3.1.3 Topics on K-L information

Boltzmann's entropy

The negative of the K-L information, B(g ; f) = −I(g ; f), is referred to as Boltzmann's entropy.

On the functional form of K-L information

Any measure that can be decomposed into two additive terms is limited to the K-L information.

2.3.3.2 Expected Log-Likelihood and Corresponding Estimator

The preceding section showed that we can evaluate the appropriateness of a given model by calculating the K-L information. However, K-L information can be used in actual modeling only in limited cases, since K-L information contains the unknown distribution g, so that its value cannot be calculated directly.

K-L information can be decomposed into

$$I(g\,;f) = E_G\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right] = E_G\left[\log g(X)\right] - E_G\left[\log f(X)\right]. \qquad (3.16)$$

Moreover, because the first term on the right-hand side is a constant that depends solely on the true model g, it is clear that in order to compare different models, it is sufficient to consider only the second term on the right-hand side. This term is called the *expected log-likelihood*. The larger this value is for a model, the smaller its K-L information is and the better the model is.

The second term is called the **expected log-likelihood**; the larger its value is for a model, the smaller its K-L information is and the better the model is.

Since the expected log-likelihood can be expressed as

$$E_G\left[\log f(X)\right] = \int \log f(x)dG(x)$$

$$= \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x)\log f(x)dx, & \text{for continuous models,} \\ \displaystyle\sum_{i=1}^{\infty} g(x_i)\log f(x_i), & \text{for discrete models,} \end{cases} \qquad (3.17)$$

it still depends on the true distribution g and is an unknown quantity that eludes explicit computation. However, if a good estimate of the expected log-likelihood can be obtained from the data, this estimate can be used as a criterion for comparing models. Let us now consider the following problem.

As the expected log-likelihood still depends on the true distribution, it is an unknown quantity that cannot be computed explicitly; therefore one needs a mean to obtain an estimate of this term.

If we replace the unknown probability distribution G above with the empirical distribution function ^G(x), we obtain:

$$E_{\hat{G}}\left[\log f(X)\right] = \int \log f(x)d\hat{G}(x)$$

$$= \sum_{\alpha=1}^{n} \hat{g}(x_\alpha)\log f(x_\alpha) \qquad (3.18)$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha).$$

According to the law of large numbers, when the number of observations, n, tends to infinity, the mean of the random variables $Y_\alpha = \log f(X_\alpha)$ $(\alpha = 1, 2, \ldots, n)$ converges in probability to its expectation, that is, the convergence

$$\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha) \longrightarrow E_G\left[\log f(X)\right], \qquad n \to +\infty, \tag{3.19}$$

holds. Therefore, it is clear that the estimate based on the empirical distribution function in (3.18) is a natural estimate of the expected log-likelihood. The estimate of the expected log-likelihood multiplied by n, i.e.,

$$n\int \log f(x) d\hat{G}(x) = \sum_{\alpha=1}^{n} \log f(x_\alpha), \tag{3.20}$$

is the log-likelihood of the model $f(x)$. This means that the *log-likelihood*, frequently used in statistical analyses, is clearly understood as being an approximation to the K-L information.

The log-likelihood, frequently used in statistical analyses, is clearly understood as being an approximation to the K-L information.

2.3.3.3 Maximum Likelihood Method and Maximum Likelihood Estimators

2.3.3.3.1 Log-Likelihood Function and Maximum Likelihood Estimators

Let us consider the case in which a model is given in the form of a probability distribution $f(x|\boldsymbol{\theta})(\boldsymbol{\theta} \in \Theta \subset R^p)$, having unknown p-dimensional parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. In this case, given data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$, the log-likelihood can be determined for each $\boldsymbol{\theta} \in \Theta$. Therefore, by regarding the log-likelihood as a function of $\boldsymbol{\theta} \in \Theta$, and representing it as

$$\ell(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\boldsymbol{\theta}), \tag{3.21}$$

the log-likelihood is referred to as the *log-likelihood function*. A natural estimator of $\boldsymbol{\theta}$ is defined by finding the maximizer $\boldsymbol{\theta} \in \Theta$ of the $\ell(\boldsymbol{\theta})$, that is, by determining $\boldsymbol{\theta}$ that satisfies the equation

$$\ell(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}). \tag{3.22}$$

This method is called the *maximum likelihood method*, and $\hat{\boldsymbol{\theta}}$ is called the *maximum likelihood estimator*. If the data used in the estimation must be specified explicitly, then the maximum likelihood estimator is denoted by $\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$. The model $f(x|\hat{\boldsymbol{\theta}})$ determined by $\hat{\boldsymbol{\theta}}$ is called the *maximum likelihood model*, and the term $\ell(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}})$ is called the *maximum log-likelihood*.

2.3.3.3.2 Implementation of the Maximum Likelihood Method by Means of Likelihood Equations

If the log-likelihood function $\ell(\boldsymbol{\theta})$ is continuously differentiable, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is given as a solution of the likelihood equation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, p \quad \text{or} \quad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{3.23}$$

where $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a p-dimensional vector, the i^{th} component of which is given by $\partial \ell(\boldsymbol{\theta})/\partial \theta_i$, and $\mathbf{0}$ is the p-dimensional zero vector, all the components of which are 0. In particular, if the likelihood equation is a linear equation having p-dimensional parameters, the maximum likelihood estimator can be expressed explicitly.

(Examples).

Linear regression model (with Gaussian noise):

Consequently, the maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}). \tag{3.35}$$

LQ: Cf. normal equation:

$$\theta = (X^T X)^{-1} X^T y$$

2.3.3.3.3 Implementation of the Maximum Likelihood Method by Numerical Optimization

When a given likelihood equation cannot be solved explicitly, a numerical optimization method is frequently employed, which involves starting from an appropriately chosen initial value θ0 and successively generating quantities θ1, θ2, . . . , in order to cause convergence to the solution ˆθ.

Newton-Raphson method – based on Taylor expansion

where the quantity $\boldsymbol{g}(\boldsymbol{\theta}_k)$ is a gradient vector and $H(\boldsymbol{\theta}_k)$ is a Hessian matrix. By virtue of (3.38), it follows that $\boldsymbol{\theta} \approx \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1}\boldsymbol{g}(\boldsymbol{\theta}_k)$. Therefore, using

$$\boldsymbol{\theta}_{k+1} \equiv \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1}\boldsymbol{g}(\boldsymbol{\theta}_k),$$

we determine the next point, $\boldsymbol{\theta}_{k+1}$. This technique, called *the Newton–Raphson method*, is known to converge rapidly near the root, or in other words, provided an appropriate initial value is chosen.

Practical problems with this method => quasi-Newton method (algorithm) is used.

For time series:

Using the state-space representation of the model and the Kalman filter recursive algorithm.

2.3.3.3.4 Fluctuations of the Maximum Likelihood Estimators

(…)

2.3.3.3.5 Asymptotic Properties of the Maximum Likelihood Estimators

Asymptotic normality.

$$I(\boldsymbol{\theta}) = \int f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx. \tag{3.56}$$

This matrix $I(\boldsymbol{\theta})$, with $(i, j)^{th}$ component given as (3.53) under condition (3), is called the *Fisher information matrix*.

2.3.3.4 *Information Criterion AIC*

2.3.3.4.1 Log-Likelihood and Expected Log-Likelihood

The argument that has been presented thus far can be summarized as follows. When we build a model using data, we assume that the data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are generated according to the true distribution $G(x)$ or $g(x)$. In order to capture the structure of the given phenomena, we assume a parametric model $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset R^p\}$ having p-dimensional parameters, and we estimate it by using the maximum likelihood method. In other words, we construct a statistical model $f(x|\hat{\boldsymbol{\theta}})$ by replacing the unknown parameter $\boldsymbol{\theta}$ contained in the probability distribution by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. Our purpose here is to evaluate the goodness or badness of the statistical model $f(x|\hat{\boldsymbol{\theta}})$ thus constructed. We now consider the evaluation of a model from the standpoint of making a prediction.

Our task is to evaluate the expected goodness or badness of the estimated model $f(z|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $Z = z$ generated from the unknown true distribution $g(z)$. The K-L information described below is used to measure the closeness of the two distributions:

$$\begin{aligned} I\{g(z); f(z|\hat{\boldsymbol{\theta}})\} &= E_G\left[\log\left\{\frac{g(Z)}{f(Z|\hat{\boldsymbol{\theta}})}\right\}\right] \\ &= E_G\left[\log g(Z)\right] - E_G\left[\log f(Z|\hat{\boldsymbol{\theta}})\right], \end{aligned} \tag{3.69}$$

where the expectation is taken with respect to the unknown probability distribution $G(z)$ by fixing $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$.

In view of the properties of the K-L information, the larger the expected log-likelihood

$$E_G\left[\log f(Z|\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z) \tag{3.70}$$

of the model is, the closer the model is to the true one. Therefore, in the definition of the information criterion, the crucial issue is to obtain a good estimator of the expected log-likelihood. One such estimator is

$$\begin{aligned} E_{\hat{G}}\left[\log f(Z|\hat{\boldsymbol{\theta}})\right] &= \int \log f(z|\hat{\boldsymbol{\theta}}) d\hat{G}(z) \\ &= \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}), \end{aligned} \tag{3.71}$$

in which the unknown probability distribution G contained in the expected log-likelihood is replaced with an empirical distribution function $\hat{G}$. This is the log-likelihood of the statistical model $f(z|\hat{\boldsymbol{\theta}})$ or the maximum log-likelihood

$$\ell(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}). \tag{3.72}$$

It is worth noting here that the estimator of the expected log-likelihood $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ is $n^{-1}\ell(\hat{\boldsymbol{\theta}})$ and that the log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ is an estimator of $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$.

2.3.3.4.2 Necessity of Bias Correction for the Log-Likelihood

In practical situations, it is difficult to precisely capture the true structure of given phenomena from a limited number of observed data. For this reason, we construct several candidate statistical models based on the observed data at hand and select the model that most closely approximates the mechanism of the occurrence of the phenomena. In this subsection, we consider the situation in which multiple models $\{f_j(z|\boldsymbol{\theta}_j);\ j = 1, 2, \ldots, m\}$ exist, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_j$ has been obtained for the parameters of the model, $\boldsymbol{\theta}_j$.

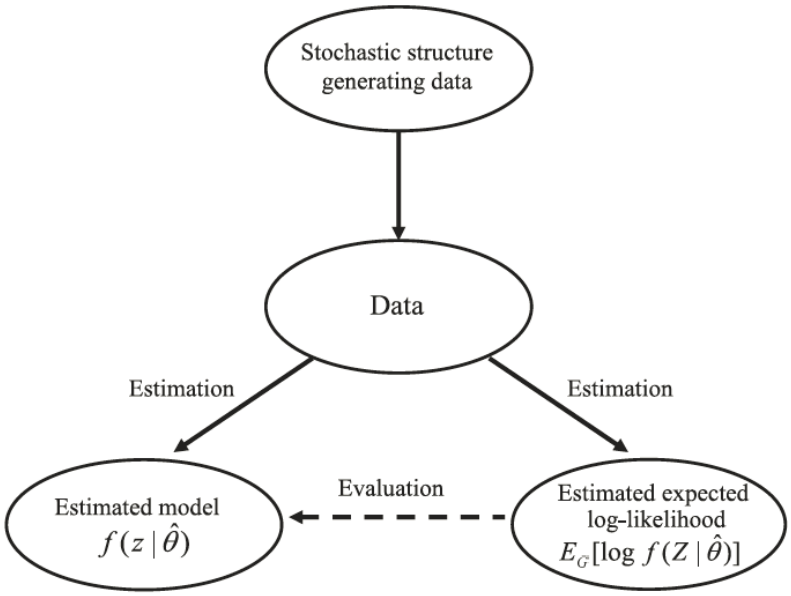


Fig. 3.5. Use of data in the estimations of the parameter of a model and of the expected log-likelihood.

From the foregoing argument, it appears that the goodness of the model specified by $\hat{\boldsymbol{\theta}}_j$, that is, the goodness of the maximum likelihood model $f_j(z|\hat{\boldsymbol{\theta}}_j)$, can be determined by comparing the magnitudes of the maximum log-likelihood $\ell_j(\hat{\boldsymbol{\theta}}_j)$. However, it is known that this approach does not provide a fair comparison of models, since the quantity $\ell_j(\hat{\boldsymbol{\theta}}_j)$ contains a bias as an estimator of the expected log-likelihood $nE_G[\log f_j(z|\hat{\boldsymbol{\theta}}_j)]$, and the magnitude of the bias varies with the dimension of the parameter vector.

This result may seem to contradict the fact that generally $\ell(\boldsymbol{\theta})$ is a good estimator of $nE_G[\log f(Z|\boldsymbol{\theta})]$. However, as is evident from the process by which the log-likelihood in (3.71) was derived, the log-likelihood was obtained by estimating the expected log-likelihood by reusing the data $\boldsymbol{x}_n$ that were initially used to estimate the model in place of the future data (Figure 3.5). The use of the same data twice for estimating the parameters and for estimating the evaluation measure (the expected log-likelihood) of the goodness of the estimated model gives rise to the bias.

Relationship between log-likelihood and expected log-likelihood:

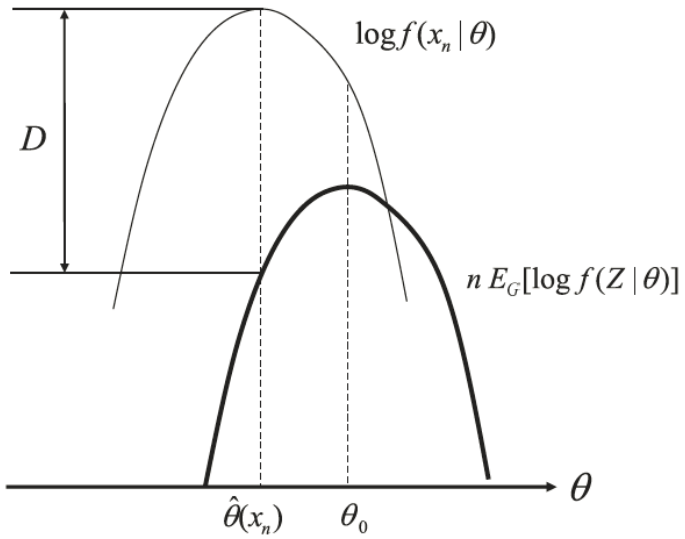


Fig. 3.6. Log-likelihood and expected log-likelihood.

Figure 3.6 shows the relationship between the expected log-likelihood function and the log-likelihood function

$$n\eta(\theta) = nE_G\left[\log f(Z|\theta)\right], \qquad \ell(\theta) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\theta), \tag{3.73}$$

for a model $f(x|\theta)$ with a one-dimensional parameter θ. The value of θ that maximizes the expected log-likelihood is the true parameter θ_0. On the other hand, the maximum likelihood estimator $\hat{\theta}(\boldsymbol{x}_n)$ is given as the maximizer of the log-likelihood function $\ell(\theta)$. The goodness of the model $f(z|\hat{\theta})$ defined by $\hat{\theta}(\boldsymbol{x}_n)$ should be evaluated in terms of the expected log-likelihood $E_G[\log f(Z|\hat{\theta})]$. However, in actuality, it is evaluated using the log-likelihood $\ell(\hat{\theta})$ that can be calculated from data. In this case, as indicated in Figure 3.6, the true criterion should give $E_G[\log f(Z|\hat{\theta})] \leq E_G[\log f(Z|\theta_0)]$ (see Subsection 3.1.1). However, in the log-likelihood, the relationship $\ell(\hat{\theta}) \geq \ell(\theta_0)$ always holds.

The log-likelihood function fluctuates depending on data, and the geometry between the two functions also varies; however, the above two inequalities always hold. If the two functions have the same form, then the log-likelihood is actually inferior to the extent that it appears to be better than the true model. The objective of the bias evaluation is to compensate for this phenomenon of reversal. Therefore, the prerequisite for a fair comparison of models is evaluation of and correction for the bias. In this subsection, we define an information criterion as a bias-corrected log-likelihood of the model.

2.3.3.4.3 Derivation of bias of the log-likelihood

(…) TIC information criterion.

2.3.3.4.4 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) has played a significant role in solving problems in a wide variety of fields as a model selection criterion for analysing actual data. The AIC is defined by

AIC = −2(maximum log-likelihood) + 2(number of free parameters)

The number of free parameters in a model refers to the dimensions of the parameter vector θ contained in the specified model f(x|θ).

The AIC is an evaluation criterion for the badness of the model whose parameters are estimated by the maximum likelihood method, and it indicates that the bias of the log-likelihood (3.80) approximately becomes the "number of free parameters contained in the model."

$$\mathrm{AIC} = -2\sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta}) + 2p$$

The AIC does not require any analytical derivation of the bias correction terms for individual problems and does not depend on the unknown probability distribution G, which removes fluctuations due to the estimation of the bias. Further, Akaike (1974) states that if the true distribution that generated the data exists near the specified parametric model, the bias associated with the log-likelihood of the model based on the maximum likelihood method can be approximated by the number of parameters. These attributes make the AIC a highly flexible technique from a practical standpoint.

2.3.3.5 Properties of MAICE

The estimators and models selected by minimizing the AIC are referred to as MAICE (minimum AIC estimators).

2.3.3.5.1 Finite Correction of the Information Criterion

In previous section, we derived the AIC for general statistical models estimated using the maximum likelihood method. In contrast, information criterion for particular models such as normal distribution models can be derived directly and analytically by calculating the bias, without having to resort to asymptotic theories such as the Taylor series expansion or the asymptotic normality.

Example of normal distribution model.

2.3.3.5.2 Distribution of Orders Selected by AIC

(…)

Consistency: In general, under the assumptions that the true model is of finite dimension and it is included in the class of candidate models, a criterion that identifies the correct model asymptotically with probability one is said to be consistent.

2.3.3.5.3 Discussion

The AIC has been criticized because it does not yield a consistent estimator with respect to the selection of orders. Such an argument is frequently misunderstood, and we attempt to clarify these misunderstandings in the following.

(…)

Although the information criterion makes automatic model selection possible, it should be noted that the model evaluation criterion is a relative evaluation criterion. This means that selecting a model using an information criterion is only a selection from a family of models that we have

specified. Therefore, the critical task for us is to set up more appropriate models by making use of knowledge regarding that object.

2.3.4 Statistical Modeling by AIC

The majority of the problems in statistical inference can be considered to be problems related to statistical modeling. They are typically formulated as comparisons of several statistical models. In this chapter, we consider using the AIC for various statistical inference problems.

NB: Checking on Google Scholar with “AIC” + “Predictive” yields a large number of research articles mainly in biology and medicine.

2.3.4.1 Checking the Equality of Two Discrete Distributions

(…)

2.3.4.2 Determining the Bin Size of a Histogram

(…)

2.3.4.3 Equality of the Means and/or the Variances of Normal Distributions

(…)

2.3.4.4 Variable Selection for Regression Model

In multiple regression analysis, all of the given explanatory variables may not be necessarily effective for predicting the response variable. An estimated model with an unnecessarily large number of explanatory variables may be unstable. By selecting the model having the minimum AIC for different possible combinations of the explanatory variables, we expect to obtain a reasonable model.

Example: daily temperature data

Table 4.4 summarizes the estimated residual variances and coefficients and AICs of various models. It shows that the model having the latitude and the altitude as explanatory variables has the smallest value for the AIC. The AIC of the model with all three explanatory variables is larger than that of the model having the lowest value for the AIC. This is because the reduction in the residual variance of the former model is miniscule compared to that of the model having the lowest value of the AIC, and it indicates that knowledge of the longitude x2 is of little value if we already know the latitude and altitude (x1 and x3).

Table 4.4. Subset regression models: AICs and estimated residual variances and coefficients.

No.	Explanatory variables	Residual variance	k	AIC	Regression coefficients a_0	a_1	a_2	a_3
1	x_1, x_3	1.490	2	88.919	40.490	−1.108	—	−0.010
2	x_1, x_2, x_3	1.484	3	90.812	44.459	−1.071	—	−0.010
3	x_1, x_2	5.108	2	119.715	71.477	−0.835	−0.305	—
4	x_1	5.538	1	119.737	40.069	−1.121	—	—
5	x_2, x_3	5.693	2	122.426	124.127	—	−0.906	−0.007
6	x_2	7.814	1	128.346	131.533	—	−0.965	—
7	x_3	19.959	1	151.879	0.382	—	—	−0.010
8	none	24.474	0	154.887	−0.580	—	—	—

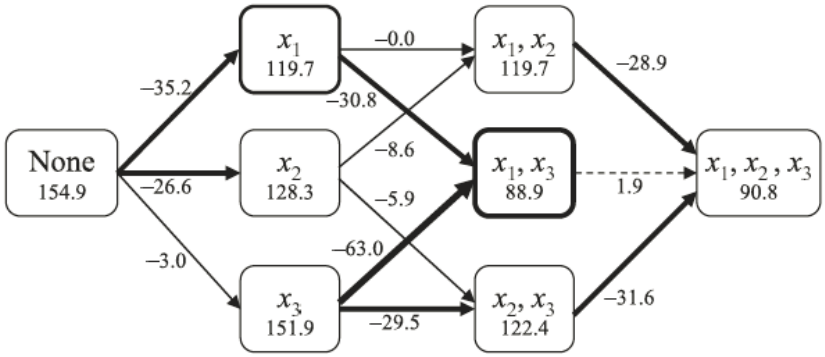


Fig. 4.3. Decrease of AIC values by adding regressors.

Figure 4.3 shows the change in the AIC value when only one explanatory variable is incorporated in a subset regression model. It is interesting to note that when only one explanatory variable is used, x3 (altitude) gives the smallest reduction in the AIC value. However, when the models with two explanatory variables are considered, the inclusion of x3 is very effective in reducing the AIC value, and the AIC best model out of these models had the explanatory variables of x1 and x3, i.e., the latitude and the altitude. The AIC of the model with x1 and x2 is the same as that of the model with x1. These suggest that x1 and x2 contain similar information, whereas x3 has independent information.

This can be understood from Figure 4.4, which shows the longitude vs. latitude scatterplot. Since the four main islands of Japan are located along a line that runs from northeast to southwest, x1 and x2 have a strong positive correlation. Thus, the information about the longitude of a city has a similar predictive ability for temperature as that of the latitude. However, when the latitude is known, knowledge of the longitude is almost redundant, whereas knowledge of the altitude is very useful and the residual variance becomes less than one third when the altitude is included.

Note that when the number of explanatory variables is large, we need to exercise care when comparing subset regression models having a different number of nonzero coefficients. This is addressed later.

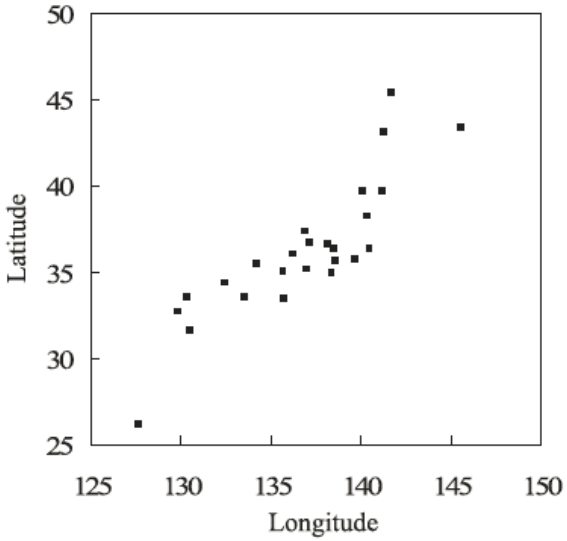


Fig. 4.4. Scatterplot: latitude vs. longitude.

2.3.4.5 Generalised Linear Models

(…)

2.3.4.6 Selection of Order of Autoregressive Model

(Times series)

(…)

2.3.4.7 Detection of structural changes

In statistical data analysis, we sometimes encounter the situation in which the stochastic structure of the data changes at a certain time or location. We consider here estimation of this change point by the statistical modeling based on the AIC.

(…)

2.3.4.8 Comparison of Shapes of Distributions

(…)

2.3.4.9 Selection of Box–Cox Transformations

(…)

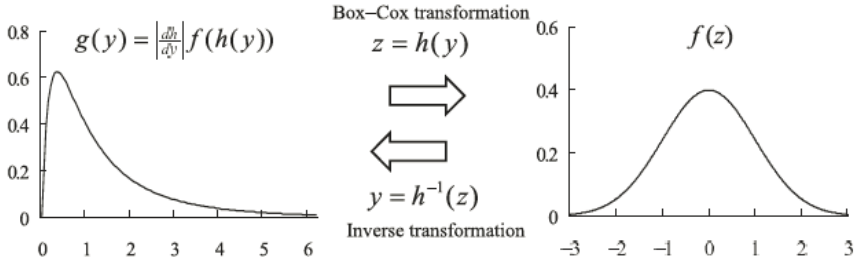


Fig. 4.11. Transformation of the probability density function by a Box–Cox transformation.

2.3.5 Generalized Information Criterion (GIC)

We have so far considered the evaluation of statistical models estimated using the maximum likelihood method, for which the AIC is a useful tool for evaluating the estimated models. However, statistical models are constructed to obtain information from observed data in a variety of ways. So if models are developed that employ estimation procedures other than the method of maximum likelihood, how should we construct an information criterion for evaluating such statistical models? With the development of other modelling techniques, it has been necessary to construct information criteria that relax the assumptions imposed on the AIC.

In this chapter, we describe a general framework for constructing information criteria in the context of functional statistics and introduce a generalized information criterion, GIC [Konishi and Kitagawa (1996)]. The GIC can be applied to evaluate statistical models constructed by various types of estimation procedures including the robust estimation procedure and the maximum penalized likelihood procedure.

(…)

2.3.6 Statistical Modeling by GIC

The current wide availability of fast and inexpensive computers enables us to construct various types of nonlinear models for analyzing data having a complex structure. Crucial issues associated with nonlinear modeling are the choice of adjusted parameters including the smoothing parameter, the number of basis functions in splines and B-splines, and the number of hidden units in neural networks. Selection of these parameters in the modeling process can be viewed as a model selection and evaluation problem. This chapter addresses these issues as a model selection and evaluation problem and provides criteria for evaluating various types of statistical models.

(…)

NB: interesting to examine methods (radial functions, etc.)

2.3.7 Theoretical Development and Asymptotic Properties of the GIC

(…)

2.3.8 Bootstrap Information Criterion

The bootstrap information criterion [Efron (1983), Wong (1983), Konishi and Kitagawa (1996), Ishiguro et al. (1997), Cavanaugh and Shumway (1997), and Shibata (1997)], obtained by applying the bootstrap methods originally proposed by Efron (1979), permits the evaluation of models estimated through complex processes.

2.3.8.1 Bootstrap Method

The bootstrap method has received considerable interest due to its ability to provide effective solutions to problems that cannot be solved by analytic approaches based on theories or formulas. A salient feature of the bootstrap method is that it uses massive iterative computer calculations rather than analytic expressions. This makes the bootstrap method a flexible statistical method that can be applied to complex problems employing very weak assumptions.

(…)

2.3.8.2 Bootstrap Information Criterion

(…)

2.3.9 Bayesian Information Criteria

This chapter considers model selection and evaluation criteria from a Bayesian point of view. A general framework for constructing the Bayesian information criterion (BIC) is described.

(…)

2.3.9.1 Bayesian Model Evaluation Criterion (BIC)

2.3.9.1.1 Definition of BIC

The Bayesian information criterion (BIC) or Schwarz's information criterion (SIC) proposed by Schwarz (1978) is an evaluation criterion for models defined in terms of their posterior probability [see also Akaike (1977)]. It is derived as follows.

> Let $M_1, M_2, \ldots, M_r$ be r candidate models, and assume that each model M_i is characterized by a parametric distribution $f_i(x|\boldsymbol{\theta}_i)$ $(\boldsymbol{\theta}_i \in \Theta_i \subset R^{k_i})$ and the prior distribution $\pi_i(\boldsymbol{\theta}_i)$ of the k_i-dimensional parameter vector $\boldsymbol{\theta}_i$. When n observations $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ are given, then, for the i^{th} model M_i, the marginal distribution or probability of $\boldsymbol{x}_n$ is given by
>
> $$p_i(\boldsymbol{x}_n) = \int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i. \qquad (9.1)$$
>
> This quantity can be considered as the likelihood of the i^{th} model and is referred to as the *marginal likelihood* of the data.

According to Bayes' theorem, if we suppose that the prior probability of the i^{th} model is $P(M_i)$, the posterior probability of the i^{th} model is given by

$$P(M_i|\boldsymbol{x}_n) = \frac{p_i(\boldsymbol{x}_n)P(M_i)}{\sum_{j=1}^{r} p_j(\boldsymbol{x}_n)P(M_j)}, \qquad i = 1, 2, \ldots, r. \qquad (9.2)$$

This posterior probability indicates the probability of the data being generated from the ith model when data xn are observed. Therefore, if one model is to be selected from r models, it would be natural to adopt the model that has the largest posterior probability. This principle means that the model that maximizes the numerator pi(xn)P(Mi) must be selected, since all models share the same denominator in (9.2).

If we further assume that the prior probabilities P(Mi) are equal in all models, it follows that the model that maximizes the marginal likelihood pi(xn) of the data must be selected. Therefore, if an approximation to the marginal likelihood expressed in terms of an integral in (9.1) can readily be obtained, the need to compute the integral on a problem-by-problem basis will vanish, thus making the BIC suitable for use as a general model selection criterion.

The BIC is actually defined as the natural logarithm of the integral multiplied by −2, and we have

$$-2\log p_i(\boldsymbol{x}_n) = -2\log\left\{\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right\}$$
$$\approx -2\log f_i(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_i) + k_i\log n,$$

where$\hat{}$θ i is the maximum likelihood estimator of the ki-dimensional parameter vector θi of the model fi(x|θi). Consequently, from the r models that are to be evaluated using the maximum likelihood method, the model that minimizes the value of BIC can be selected as the optimal model for the data. Thus, even under the assumption that all models have equal prior probabilities, the posterior probability obtained by using the information from the data serves to contrast the models and helps to identify the model that generated the data.

(…)

2.3.10 Various Model Evaluation Criteria

So far in this book, we have considered model selection and evaluation criteria from both an information-theoretic point of view and a Bayesian approach.

The **AIC**-type criteria were constructed as estimators of the Kullback–Leibler information between a statistical model and the true distribution generating the data or equivalently the expected log-likelihood of a statistical model.

In contrast, the **Bayes** approach for selecting a model was to choose the model with the largest posterior probability among a set of candidate models.

There are other model evaluation criteria based on various different points of view. This chapter describes cross-validation, generalized cross-validation, final predictive error (FPE), Mallows' Cp, the Hannan–Quinn criterion, and ICOMP. **Cross-validation** also provides an alternative approach to estimate the Kullback–Leibler information. We show that the cross-validation estimate is asymptotically equivalent to AIC-type criteria in a general setting.

http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2010.02210.x/full

Box 1: a summary of the alternatives to AIC Jump to…

Forms of the AIC, such as AIC_C (small sample size correction, Table 1) and Quasi-AIC (QAIC: controls for overdispersion), remain the most widely used information criteria for ranking models in the IT approach. However, there is debate surrounding the utility of AIC (e.g. Spiegelhalter *et al.*, 2002; Stephens *et al.*, 2007), and various alternatives have been proposed. The different criteria in use today may be appropriate in different circumstances (Murtaugh, 2009), but all information criteria are in fact approximations of Bayes Factors (BFs) (Congdon, 2006a) with certain assumptions such as large sample sizes. The BF is a ratio between two models, reflecting 'true' model probabilities given data support, i.e. posterior model probabilities (other information criteria approximate these posterior model probabilities) (Jefferys, 1961*in*Congdon, 2006b):

$$\mathrm{BF}_{1,2} = \frac{p\,(y|M_1)}{p\,(y|M_2)} \qquad (1)$$

where $p(y|M_i)$ is the marginal likelihood of model *i*. Therefore, BFs seem to be the ideal index for model selection and averaging. However, calculations of BFs directly become quickly complicated when comparing more than two models. Although several methods for using BFs for model averaging have been suggested, it seems that currently available methods are highly technical and difficult to implement (Congdon, 2006a). Practical implementations of BFs for multimodel comparisons are an active frontier of statistical research (R. Barker, personal communication) and thus advances in the area are anticipated in the near future.

Table 1. Information criteria for model selection.

Information criterion	Formula*	References
Akaike Information Criterion	$AIC = -2 \cdot \ln L + 2k$	Akaike (1973)
AIC – small sample size correction	$AIC_C = -2 \cdot \ln L + \frac{2k(k+1)}{n-k-1}$	Hurvich & Tsai (1989)
Quasi-AIC	$QAIC = \frac{-2 \cdot \ln L}{\hat{c}} + 2k$	Lebreton *et al.* (1992)
Conditional AIC	$cAIC = -2 \cdot \ln L + 2k_c$	Vaida & Blanchard (2005); Liang *et al.* (2008)
Bayesian Information Criterion	$BIC = -2 \cdot \ln L + k \ln(n)$	Schwarz (1978)
Deviance Information Criterion	$DIC = -2 \cdot \ln L + 2k_D$	Spiegelhalter *et al.* (2002)

*L = likelihood function = $p(y|\theta)$, or, if random factors are explicitly separated as parameters (as in cAIC) = $p(y|\theta,u)$. NB, $-2 \cdot \ln L$ is also known as the 'deviance'. k = number of parameters in the model; n = sample size; $\hat{c}$ = overdispersion parameter; k_c = effective number of degrees of freedom (cAIC); k_D = effective number of parameters (DIC). See listed references for additional details of formula components.

2.3.10.1 Cross-Validation

2.3.10.1.1 Prediction and cross-validation

The objective of statistical modeling or data analysis is to obtain information about data that may arise in the future, rather than the observed data used in the model construction itself. Hence, in the model building process, model evaluation from a predictive point of view implies the evaluation of the goodness of fit of the model based on future data obtained independently of the observed data. In practice, however, it is difficult to consider situations in which future data can be obtained separately from the model construction data, and if, in fact, such data can be obtained, a better model would be constructed by combining such data with the observed data. As a way to circumvent this difficulty, **cross-validation** refers to a technique whereby evaluation from a predictive point of view is executed solely based on observed data while making modifications in order to preserve the accuracy of parameter estimation as much as possible.

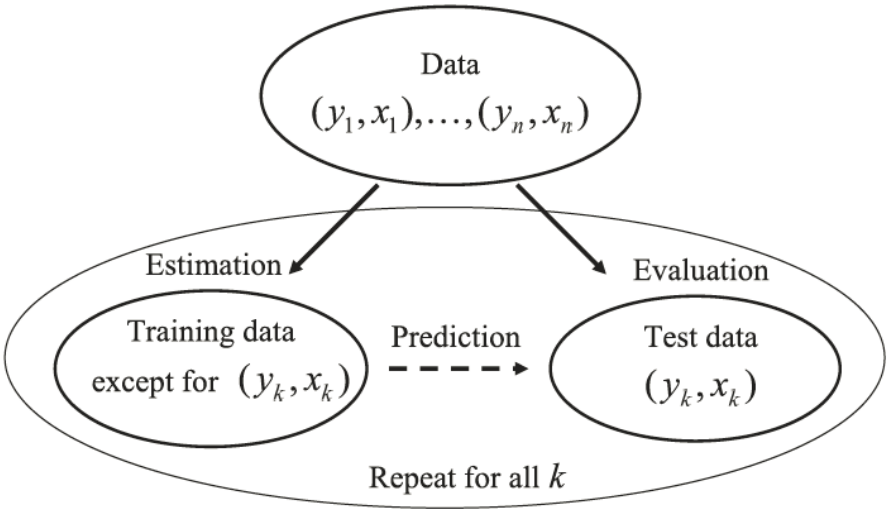


Fig. 10.1. Schematic of the cross-validation procedure.

Given a response variable y and p explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$, let us consider the regression model

$$y = u(\boldsymbol{x}) + \varepsilon, \tag{10.1}$$

where $E[\varepsilon] = 0$ and $E[\varepsilon^2] = \sigma^2$. Since $E[Y|\boldsymbol{x}] = u(\boldsymbol{x})$, the function $u(\boldsymbol{x})$ represents the mean structure. We estimate $u(\boldsymbol{x})$ based on n observations $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, 2, \ldots, n\}$ and write it as $\hat{u}(\boldsymbol{x})$. For example, when a linear regression model $y = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$ is assumed, an estimate of $u(\boldsymbol{x})$ is given by $\hat{u}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$, using least squares estimates $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ of the regression coefficients $\boldsymbol{\beta}$, where $X^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

The goodness of fit of the estimated regression function $\hat{u}(\boldsymbol{x})$ is measured using the (average) predictive mean square error (PSE)

$$\mathrm{PSE} = \frac{1}{n} \sum_{\alpha=1}^{n} E\left[\{Y_\alpha - \hat{u}(\boldsymbol{x}_\alpha)\}^2\right], \tag{10.2}$$

in terms of future observations Y_α that are randomly drawn at points $\boldsymbol{x}_\alpha$ according to (10.1) in a manner independent of the observed data. Here the residual sum of squares (RSS) is used to estimate the PSE by reusing the data y_α instead of the Y_α:

$$\mathrm{RSS} = \frac{1}{n} \sum_{\alpha=1}^{n} \{y_\alpha - \hat{u}(\boldsymbol{x}_\alpha)\}^2. \tag{10.3}$$

If $\hat{u}(x)$ is a polynomial model, for example, the greater the order of the model, the smaller this value becomes, and the goodness of fit of the model seems to

be improved. As a result, we end up selecting a polynomial of order $n-1$ that passes through all the observations, which defeats the purpose of an order selection criterion.

Cross-validation involves the estimation of a predictive mean square error by separating the data used for model estimation (training data) from the data used for model evaluation (test data). Cross-validation is executed in the following steps:

Cross-Validation

(1) From the n observed data values, remove the α^{th} observation $(y_\alpha, \boldsymbol{x}_\alpha)$. Estimate the model based on the remaining $n-1$ observations, and denote this estimate by $\hat{u}^{(-\alpha)}(\boldsymbol{x})$.
(2) For the α^{th} data value $(y_\alpha, \boldsymbol{x}_\alpha)$ removed in step 1, calculate the value of the predictive square error $\{y_\alpha - \hat{u}^{(-\alpha)}(\boldsymbol{x}_\alpha)\}^2$.
(3) Repeat steps 1 and 2 for all $\alpha \in \{1, \ldots, n\}$, and obtain

$$\mathrm{CV} = \frac{1}{n}\sum_{\alpha=1}^{n}\left\{y_\alpha - \hat{u}^{(-\alpha)}(\boldsymbol{x}_\alpha)\right\}^2 \qquad (10.4)$$

as the estimated value of the predictive mean square error defined by (10.2). This process is known as *leave-one-out cross-validation.*

CV can be considered to be an estimator of predictive mean square error.

The leave-one-out cross-validation procedure can be generalized to the method called K-fold cross-validation as follows. The observed data are divided into K subsets. One of the K subsets is used as the test data for evaluating a model, and the union of the remaining K−1 subsets is taken as training data. The average prediction error across the K trials is then calculated.

2.3.10.1.2 Selecting a Smoothing Parameter by Cross-Validation

CV can be used to select a smoothing parameter lambda (regularization).

(…)

2.3.10.1.3 Generalised CV

Selecting the optimal values of the number of basis functions and a smoothing parameter by applying cross-validation to a large data set can result in computational difficulties. If the predicted value ˆy is given in the form of ˆy = Hy, where H is a matrix that does not depend on the data y, then in cross-validation, the estimation process performed n times by removing observations one by one is not needed, and thus the amount of computation required can be reduced substantially.

(…)

2.3.10.1.4 Asymptotic Equivalence Between AIC-Type Criteria and Cross-Validation

Cross-validation offers an alternative approach to estimate the Kullback–Leibler information from a predictive point of view.

(…)

2.4 Geometric Modeling in Probability and Statistics (Springer, 2014)

Statistical manifolds are geometric abstractions used to model information, their field of study belonging to **Information Geometry**, a relatively recent branch of mathematics that uses tools of differential geometry to study statistical inference, information loss, and estimation. This field started with the differential geometric study of the manifold of probability density functions. For instance, the set of normal distributions

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R},$$

with $(\mu, \sigma) \in R \times (0,+\infty)$, can be considered as a two-dimensional surface. This can be endowed with a Riemannian metric, which measures the amount of information between the distributions. One of these possibilities is to consider the **Fisher information metric**. In this case, the distribution family $p(x; \mu, \sigma)$ becomes a space of constant negative curvature. Therefore, any normal distribution can be visualized as a point in the Poincaré upper-half plane (hyperbolic plane).

In a similar way, we shall consider other parametric model families of probability densities that can be organized as a differentiable manifold embedded in the ambient space of all density functions. Every point on this manifold is a density function, and any curve corresponds to a one-parameter subfamily of density functions. The distance between two points (i.e., distributions), which is measured by the Fisher metric, was introduced almost simultaneously by C. R. Rao and H. Jeffreys in the mid-1940s. The role of differential geometry in statistics was first emphasized by Efron in 1975, when he introduced the concept of statistical curvature. Later Amari used the tools of differential geometry to develop this idea into an elegant representation of Fisher's theory of information loss.

A fundamental role in characterizing statistical manifolds is played by two geometric quantities, called **dual connections**, which describe the derivation with respect to vector fields and are interrelated in a duality relation involving the Fisher metric. The use of dual connections leads to several other dual elements, such as volume elements, Hessians, Laplacians, second fundamental forms, mean curvature vector fields, and Riemannian curvatures. The study of dual elements and the relations between them constitute the main direction of development in the study of statistical manifolds.

Even if sometimes we use computations in local coordinates, the relationships between these geometric quantities are invariant with respect to the selection of any particular coordinate system. Therefore, the study of statistical manifolds provides techniques to investigate the intrinsical properties of statistical models rather than their parametric representations. This invariance feature made statistical manifolds useful in the study of information geometry.

The book is structured into two distinct parts. The first one is an accessible introduction to the theory of statistical models, while the second part is devoted to an abstract approach of statistical manifolds.

Part I

Chapter 1 introduces the notion of **statistical model**, which is a space of density functions, and provides the exponential and mixture families as distinguished examples. The **Fisher information** is defined together with **two dual connections** of central importance to the theory. The skewness tensor is also defined and computed in certain particular cases.

Chapter 2 contains a few important examples of statistical models for which the Fisher metric and geodesics are worked out explicitly. This includes the case of normal and lognormal distributions, and also the gamma and beta distribution families.

Chapter 3 deals with an introduction to entropy on statistical manifolds and its basic properties. It contains definitions and examples, an analysis of maxima and minima of entropy, its upper and lower bounds, Boltzmann–Gibbs submanifolds, and the adiabatic flow.

Chapter 4 is dedicated to the **Kullback–Leibler divergence** (or relative entropy), which provides a way to measure the relative entropy between two distributions. The chapter contains explicit computations and fundamental properties regarding the first and second variations of the cross entropy, its relation with the Fisher information matrix and some variational properties involving Kullback–Leibler divergence.

Chapter 5 defines and studies the concept of informational energy on statistical models, which is a concept analogous to kinetic energy from physics. The first and second variations are studied and uncertainty relations and some thermodynamics laws are presented.

Chapter 6 discusses the significance of maximum entropy distributions in the case when the first N moments are given. A distinguished role is played by the case when N = 1 and N = 2, cases when some explicit computations can be performed. A definition and brief discussion of Maxwell–Boltzmann distributions is also made.

Part II

The second part is dedicated to a detailed study of statistical manifolds and contains seven chapters. This part is an abstractization of the results contained in Part I. Instead of statistical models, one considers here differentiable manifolds, and instead of the Fisher information metric, one takes a Riemannian metric. Thus, we are able to carry the ideas from the theory of statistical models over to Riemannian manifolds endowed with a dualistic structure defined by a pair of torsion-less dual connections.

Chapter 7 contains an introduction to the theory of differentiable manifolds, a central role being played by the Riemannian manifolds. The reader accustomed with the basics of differential geometry may skip to the next chapter. The role of this chapter is to accommodate the novice reader with the language and objects of differential geometry, which will be further developed throughout the later chapters.

A formulation of the dualistic structure is given in Chap. 8. This chapter defines and studies general properties of dual connections, relative torsion tensors and curvatures, α-connections, the skewness and difference tensors. It also contains an equivalent construction of statistical manifolds starting from a skewness tensor.

Chapter 9 describes how to associate a volume element with a given connection and discusses the properties of dual volume elements, which are associated with a pair of dual connections. The properties of α-volume elements are provided with the emphasis on the relation with the Lie derivative and vector field divergence. An explicit computation is done for the distinguished examples of exponential and mixture cases. A special section is devoted to the study of equiaffine connections, i.e. connections which admit a parallel n-volume form. The relation with the statistical manifolds of constant curvature is also emphasized.

Chapter 10 deals with a description of construction and properties of dual Laplacians, which are Laplacians defined with respect to a pair of dual connections. An α-Laplacian is also defined and studied. The relation with the dual volume elements is also emphasized. The last part of the chapter deals with trace of the metric tensor and its relation to Laplacians.

The construction of statistical manifolds starting from contrast functions is described in Chap. 11. The construction of a dualistic structure (Riemannian metric and dual connections) starting from a contrast function is due to Eguchi [38, 39, 41]. Contrast functions are also known in the literature under the name of divergences, a denomination we have tried to avoid here as much as we could.

Chapter 12 presents a few classical examples of contrast functions, such as Bregman, Chernoff, Jefferey, Kagan, Kullback–Leibler, Hellinger, and f-divergence, and their values on a couple of examples of statistical models.

The study of statistical submanifolds, which are subsets of statistical manifolds with a similar induced structure, is done in Chap. 13. Many classical notions, such as second fundamental forms, shape operator, mean curvature vector, and Gauss–Codazzi equations, are presented here from the dualistic point of view. We put our emphasis on the relation between dual objects; for instance, we find a relation between the divergences of dual mean curvature vector fields and the inner product of these vector fields.

The present book follows the line started by Fisher and Efron and continued by Eguchi, Amari, Kaas, and Vos. The novelty of this work, besides presentation, can be found in Chaps. 5, 6, 9, and 13.

2.4.1 Statistical Models

2.5 Methods of Information Geometry (Amari, Nagaoka, 2000)

One of the well-known textbooks in the field is written by two of the information geometry founders, Amari and Nagaoka [8], which was published first time in Japan in 1993, and then translated into English in 2000. This book presents a concise introduction to the mathematical foundation of information geometry and contains an overview of other related areas of interest and applications.

2.6 Research Articles and Forums

2.6.1 Amari, Shun-ichi

2.6.1.1 Natural Gradient works efficiently in learning (1998)

Neural Computation 10, 251–276 (1998)

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

When a parameter space has a certain underlying structure, the ordinary gradient of a function does not represent its steepest direction, but the natural gradient does. Information geometry is used for calculating the natural gradients in the parameter space of perceptrons, the space of matrices (for blind source separation), and the space of linear dynamical systems (for blind source deconvolution). The dynamical behavior of natural gradient online learning is analyzed and is proved to be Fisher efficient, implying that it has asymptotically the same performance as the optimal batch estimation of parameters. This suggests that the plateau phenomenon, which appears in the backpropagation learning algorithm of multilayer perceptrons, might disappear or might not be so serious when the natural gradient is used. An adaptive method of updating the learning rate is proposed and analyzed.

Cf. also: http://www.hindawi.com/journals/aans/2011/407497/

Amari et al. developed the adaptive natural gradient learning (ANGL) algorithm for multilayer perceptrons [1–3]. It is similar to Newton's method in which the gradient of the error function is multiplied by the inverse of the Hessian matrix of the error function. However, Amari's algorithm replaces the inverse of the Hessian matrix with the inverse of the Fisher information matrix. The simplified natural gradient learning (SNGL) algorithm introduced in this paper uses a new formulation of the Fisher information matrix. SNGL is based on the backpropagation algorithm [4]. In addition, the SNGL algorithm also uses regularization [5] to penalize solutions with large connection weights. This regularization also prevents the Fisher information matrix from approaching a singular matrix. The learning rate is also annealed [6] to improve the general performance of SNGL.

The natural gradient learning algorithm is "natural" in the sense that it follows the manifold of the underlying parameter space of multilayer perceptrons. This parameter space has Riemannian geometry. Calculating the direction of steepest descent in a Riemannian manifold is more complicated than it is in a Euclidean space.

The parameter space of multilayer perceptrons is an affine space in which probability distributions are the points and random variables are vectors [9, 10].

An exponential family is a set of distributions with probability density functions of the form

$$p\left(x \mid \theta^1, \ldots, \theta^r\right) = c\left(x\right) \exp\left(\sum_{i=1}^{r} \theta^i \eta_i\left(x\right) - A\left(\theta^1, \ldots, \theta^r\right)\right). \tag{1}$$

Thus, the distributions in the exponential family are parameterized by $\{\theta^1, \ldots, \theta^r\}$, and these parameters are the canonical coordinates of the distribution. The number $A(\theta^1, \ldots, \theta^r)$ is chosen such that $\int p(x \mid \theta^1, \ldots, \theta^r)dx = 1$. Given an open set U in the event space of the random variable X, the probability that the result of the random variable X lies within U is

$$P\left[X \in U\right] = \int_U p\left(x \mid \theta^1, \ldots, \theta^r\right) dx. \tag{2}$$

The sufficient statistics for the exponential family are the r random variables $\eta_1, \ldots, \eta_r$ in (1). These statistics can be used to estimate the coordinates $\theta^1, \ldots, \theta^r$ from the values of samples of x.

The log-likelihood function of a probability distribution is the logarithm of its probability density function $\log p$. The score functions are the derivatives of the log-likelihood function with respect to the coordinates $\theta^1, \ldots, \theta^r$. The score for parameter i is

$$\mathbf{s}^i = \frac{\partial \log p\left(x \mid \theta^1, \ldots, \theta^r\right)}{\partial \theta^i} = \eta_i\left(x\right) - \frac{\partial A}{\partial \theta^i}. \tag{3}$$

The Fisher information of two score functions is the expected value of their product. For the score functions $\mathbf{s}^i$ and $\mathbf{s}^j$, the Fisher information is

$$g_p^{ij} = E_p\left\{\mathbf{s}^i \mathbf{s}^j\right\}, \tag{4}$$

where E_p is the expectation operator for the distribution with probability density function p. The Fisher information matrix is a real symmetric $r \times r$ matrix $\mathbf{G}_p = (g_p^{ij})$ whose elements are the Fisher information for each pair of score functions.

When finding the minimum of a function f in a Euclidean space, a descent direction [11] is a unit norm tangent vector $\mathbf{v}_p$ such that

$$\langle \mathbf{v}_p, \nabla f(p) \rangle \triangleq \sum_{i=1}^{n} v_i \left. \frac{\partial f}{\partial x_i} \right|_p < 0. \tag{5}$$

The minimum value of (5) is $-\alpha \|\nabla f(p)\|^2$ where α is an arbitrary scalar and $\|\nabla f(p)\|^2 = \sum_{i=1}^{n} (\partial f / \partial x_i \mid_p)^2$ which occurs when $\mathbf{v}_p = -\nabla f(p) / \|\nabla f(p)\|$. Thus, in a Euclidean space, the direction of steepest descent is the unit tangent vector $-\nabla f(p) / \|\nabla f(p)\|$.

Finding the direction of steepest descent turns out to be harder than it first appears to be. In a Euclidean geometry, it is relatively straightforward. In a Riemannian geometry, the coordinate system is curved, and a gradient descent algorithm should follow the curvature of the manifold. The direction of steepest descent should be compensated for this curvature.

In a Riemannian space, a metric g_p exists at every point p. If the tangent vectors $\mathbf{v}_p^1, \ldots, \mathbf{v}_p^n$ form a basis for the tangent space T_p at p, then $g_p^{ij} = \langle \mathbf{v}_p^i, \mathbf{v}_p^j \rangle$ is the Riemannian metric where $\langle \cdot, \cdot \rangle$ is an inner product defined on the tangent space T_p.

Given a tangent vector $\mathbf{w}_p = \sum_{i=1}^{n} \alpha_i \mathbf{v}_p^i$ and the gradient $\nabla f(p) = \sum_{i=1}^{n} \beta_i \mathbf{v}_p^i$, their inner product will be

$$\langle \mathbf{w}_p, \nabla f(p) \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} g_p^{ij} \alpha_i \beta_j. \tag{6}$$

If $\mathbf{w}_p$ is the steepest descent direction, then, according to the analysis above, $\langle \mathbf{w}_p, \nabla f(p) \rangle = -\alpha \|\nabla f(p)\|^2$. This occurs when $\mathbf{w}_p = -\mathbf{G}_p^{-1} \nabla f(p)$ where $\mathbf{G}_p = (g_p^{ij})$ is the matrix whose elements are equal to the inner products of the basis vectors. In other words, the coordinate system matters when doing gradient descent, and it may change at each point on the manifold. The natural gradient learning algorithm uses this "natural" descent direction at each step.

2.6.1.2 Information Geometry on Hierarchy of Probability Distributions (2001)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 47, NO. 5, JULY 2001

2.6.1.2.1 Manifold, curve and orthogonality

Let us consider a parameterized family of probability distributions $\boldsymbol{S} = \{p(x, \boldsymbol{\xi})\}$, where x is a random variable and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is a real vector parameter to specify a distribution. The family $\boldsymbol{S}$ is regarded as an n-dimensional manifold having $\boldsymbol{\xi}$ as a coordinate system. When the Fisher information matrix $G = (g_{ij})$

$$g_{ij}(\boldsymbol{\xi}) = E\left[\frac{\partial \log p(x, \boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(x, \boldsymbol{\xi})}{\partial \xi_j}\right] \tag{1}$$

where E denotes expectation with respect to $p(x, \boldsymbol{\xi})$, is nondegenerate, $\boldsymbol{S}$ is a Riemannian manifold, and $G(\boldsymbol{\xi})$ plays the role of a Riemannian metric tensor.

The squared distance ds^2 between two nearby distributions $p(x, \boldsymbol{\xi})$ and $p(x, \boldsymbol{\xi}, +d\boldsymbol{\xi})$ is given by the quadratic form of $d\boldsymbol{\xi}$

$$ds^2 = \sum g_{ij}(\boldsymbol{\xi})\, d\xi^i\, d\xi^j. \tag{2}$$

It is known that this is twice the Kullback–Leibler divergence

$$ds^2 = 2KL[p(x, \boldsymbol{\xi}) : p(x, \boldsymbol{\xi} + d\boldsymbol{\xi})] \tag{3}$$

where

$$KL[p : q] = \int p(x) \log \frac{p(x)}{q(x)}\, dx. \tag{4}$$

Let us consider a curve $\boldsymbol{\xi} = \xi(t)$ parameterized by t in $\boldsymbol{S}$, that is, a one-parameter family of distributions $p(x, \boldsymbol{\xi}(t))$ in $\boldsymbol{S}$. It is convenient to represent the tangent vector $\dot{\boldsymbol{\xi}}(t) = (d/dt)\boldsymbol{\xi}(t)$ of the curve at t by the random variable called the score

$$\dot{\boldsymbol{\xi}}(t) = \frac{d}{dt} \log p(x, \boldsymbol{\xi}(t)) \tag{5}$$

which shows how the log probability changes as t increases. Given two curves $\boldsymbol{\xi}_1(t)$ and $\boldsymbol{\xi}_2(t)$ intersecting at t, the inner product of the two tangent vectors is given by

$$\begin{aligned}\left\langle \dot{\boldsymbol{\xi}}_1(t), \dot{\boldsymbol{\xi}}_2(t) \right\rangle &= E\left[\frac{d}{dt} \log p(x, \boldsymbol{\xi}_1(t)) \frac{d}{dt} \log p(x, \boldsymbol{\xi}_2(t))\right] \\ &= \sum g_{ij}(\boldsymbol{\xi}) \frac{d}{dt} \xi_{2j}(t) \frac{d}{dt} \xi_{1i}(t).\end{aligned} \tag{6}$$

The two curves intersect at $\boldsymbol{\xi}_1(t) = \boldsymbol{\xi}_2(t)$ orthogonally when

$$\left\langle \dot{\boldsymbol{\xi}}_1(t), \dot{\boldsymbol{\xi}}_2(t) \right\rangle = 0 \tag{7}$$

that is, when the two scores are noncorrelated.

2.6.1.2.2 Dually flat manifolds

A manifold $\boldsymbol{S}$ is said to be e-flat (exponential-flat), when there exists a coordinate system (parameterization) $\boldsymbol{\theta}$ such that, for all i, j, k

$$E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x,\boldsymbol{\theta})\frac{\partial}{\partial\theta_k}\log p(x,\boldsymbol{\theta})\right]=0 \qquad (8)$$

identically. Such $\boldsymbol{\theta}$ is called e-affine coordinates. When a curve $\boldsymbol{\theta}(t)$ is given by a linear function $\boldsymbol{\theta}(t)=t\boldsymbol{a}+\boldsymbol{b}$ in the $\boldsymbol{\theta}$-coordinates, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors, it is called an e-geodesic. Any coordinate curve θ_i itself is an e-geodesic. (It is possible to define an e-geodesic in any manifold, but it is no more linear and we need the concept of the affine connection.)

A typical example of an e-flat manifold is the well-known exponential family written as

$$p(\boldsymbol{x},\boldsymbol{\theta})=\exp\left\{\sum\theta_i k_i(x)-\psi(\boldsymbol{\theta})\right\} \qquad (9)$$

where $k_i(x)$ are given functions and ψ is the normalizing factor. The e-affine coordinates are the canonical parameters $\boldsymbol{\theta}=(\theta_i)$, and (8) holds because

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p=-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\psi(\boldsymbol{\theta}) \qquad (10)$$

does not depend on x and $E[\frac{\partial}{\partial\theta_i}\log p]=0$.

Dually to the above, a manifold is said to be m-flat (mixture-flat), when there exists a coordinate system $\boldsymbol{\eta}$ such that

$$E\left[\frac{1}{p(x,\boldsymbol{\eta})}\frac{\partial^2}{\partial\eta_i\partial\eta_j}p(x,\boldsymbol{\eta})\frac{\partial}{\partial\eta_k}\log p(x,\boldsymbol{\eta})\right]=0 \qquad (11)$$

identically. Here, $\boldsymbol{\eta}$ is called m-affine coordinates. A curve is called an m-geodesic when it is represented by a linear function $\boldsymbol{\eta}(t)=\boldsymbol{a}t+\boldsymbol{b}$ in the m-affine coordinates. Any coordinate curve η_i of $\boldsymbol{\eta}$ is an m-geodesic.

A typical example of an m-flat manifold is the mixture family

$$p(x,\boldsymbol{\eta})=\sum\eta_i q_i(x)+\left(1-\sum\eta_i\right)q_0(x) \qquad (12)$$

where $q_i(x)$ are given probability distributions and $0<\eta_i<1$, $\sum\eta_i<1$.

The following theorem is known in information geometry.

Theorem 1: A manifold $\boldsymbol{S}$ is e-flat when and only when it is m-flat and *vice versa*.

This shows that an exponential family is automatically m-flat although it is not necessarily a mixture family. A mixture family is e-flat, although it is not in general an exponential family. The m-affine coordinates ($\boldsymbol{\eta}$-coordinates) of an exponential family are given by

$$\eta_i=E[k_i(x)]=\frac{\partial}{\partial\theta_i}\psi(\boldsymbol{\theta}) \qquad (13)$$

which is known as the expectation parameters. The coordinate transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is given by the Legendre transformation, and the inverse transformation is

$$\theta_i=\frac{\partial\varphi(\boldsymbol{\eta})}{\partial\eta_i} \qquad (14)$$

where $\varphi(\boldsymbol{\eta})$ is the negative entropy

$$\varphi(\boldsymbol{\eta})=E[\log p(x,\boldsymbol{\eta})] \qquad (15)$$

and

$$\psi(\boldsymbol{\theta})+\varphi(\boldsymbol{\eta})+\boldsymbol{\theta}\cdot\boldsymbol{\eta}=0 \qquad (16)$$

holds with $\boldsymbol{\theta}\cdot\boldsymbol{\eta}=\sum\theta_i\eta_i$. This was first remarked by Barndorff-Nielsen [15] in the case of exponential families.

Theorem 2: The tangent vectors (represented by random variables) of the coordinate curves θ_i

$$\boldsymbol{e}_i = \frac{\partial}{\partial \theta_i} \log p(x, \boldsymbol{\theta}) \tag{19}$$

and the tangent vectors of the coordinate curves η_j

$$\boldsymbol{e}_j^* = \frac{\partial}{\partial \eta_j} \log p(x, \boldsymbol{\eta}) \tag{20}$$

are orthonormal at all the points

$$\begin{aligned} \langle \boldsymbol{e}_i, \boldsymbol{e}_j^* \rangle &= E\left[\frac{\partial}{\partial \theta_i} \log p(x, \boldsymbol{\theta}) \frac{\partial}{\partial \eta_j} \log p(x, \boldsymbol{\eta}) \right] \\ &= \delta_{ij} \end{aligned} \tag{21}$$

where δ_{ij} is the Kronecker delta.

2.6.1.2.3 Divergence and generalised Pythagoras theorem

Let $p = p(x, \boldsymbol{\theta})$ and $p' = p(x, \boldsymbol{\theta}')$ be two distributions in a dually flat manifold $\boldsymbol{S}$, and let η and $\boldsymbol{\eta}'$ be the corresponding m-affine coordinates. They have two convex potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$. In the case of exponential families, ψ is the cumulant generating function and φ is the negative entropy. For a mixture family, φ is also the negative entropy. By using the two functions, we can define a divergence from p to p' by

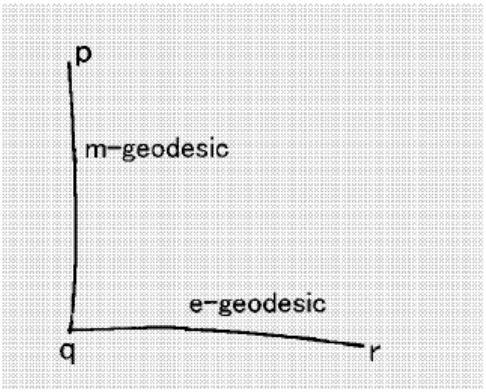
$$D[p : p'] = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}'. \tag{22}$$

The divergence satisfies $D[p : p'] \geq 0$ with equality when, and only when, $p = p'$. In the cases of an exponential family and a mixture family, this is equal to the Kullback–Leibler divergence

$$D[p : p'] = E_{\boldsymbol{\theta}}\left[\log \frac{p(x, \boldsymbol{\theta})}{p(x, \boldsymbol{\theta}')} \right] \tag{23}$$

where $E_{\boldsymbol{\theta}}$ is the expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\theta})$.

For a dually flat manifold $\boldsymbol{S}$, the following Pythagoras theorem plays a key role (Fig. 1).



Theorem 3: Let p, q, r be three distributions in $\boldsymbol{S}$. When the m-geodesic connecting p and q is orthogonal at q to the e-geodesic connecting q and r

$$D[p : q] + D[q : r] = D[p : r]. \tag{24}$$

The same theorem can be reformulated in a dual way.

Theorem 4: For p, q, $r \in \boldsymbol{S}$, when the e-geodesic connecting p and q is orthogonal at q to the m-geodesic connecting q and r

$$\overline{D}[p : q] + \overline{D}[q : r] = \overline{D}[p : r] \tag{25}$$

with

$$\overline{D}[p : q] = D[q : p]. \tag{26}$$

2.6.1.3 Information geometry of divergence functions (2010)

BULLETIN OF THE POLISH ACADEMY OF SCIENCES TECHNICAL SCIENCES Vol. 58, No. 1, 2010

Abstract. Measures of divergence between two points play a key role in many engineering problems. One such measure is a distance function, but there are many important measures which do not satisfy the properties of the distance. The Bregman divergence, Kullback-Leibler divergence and f-divergence are such measures. In the present article, we study the differential-geometrical structure of a manifold induced by a divergence function. It consists of a Riemannian metric, and a pair of dually coupled affine connections, which are studied in information geometry. The class of Bregman divergences are characterized by a dually flat structure, which is originated from the Legendre duality. A dually flat space admits a generalized Pythagorean theorem. The class of f-divergences, defined on a manifold of probability distributions, is characterized by information monotonicity, and the Kullback-Leibler divergence belongs to the intersection of both classes. The f-divergence always gives the α-geometry, which consists of the Fisher information metric and a dual pair of $\pm\alpha$-connections. The α-divergence is a special class of f-divergences. This is unique, sitting at the intersection of the f-divergence and Bregman divergence classes in a manifold of positive measures. The geometry derived from the Tsallis q-entropy and related divergences are also addressed.

2.6.1.3.1 Introduction

The present paper aims at elucidating the differential geometrical structure of a manifold equipped with a divergence function. We study the geometry induced by a divergence function, and demonstrate that it endows a Riemannian metric and a pair of dually coupled affine connections.

Bregman divergences are derived from convex functions. The Bregman divergence induces a dual structure through the Legendre transformation. It gives a geometrical structure consisting of a Riemannian metric and dually flat affine connections, called the dually flat Riemannian structure [1]. A dually flat Riemannian manifold is a generalization of the Euclidean space, in which the generalized Pythagorean theorem and projection theorem hold. These two theorems provide powerful tools for solving problems in optimization, statistical inference and signal processing. We show that the Bregman type divergence is automatically induced from the dual flatness of a Riemannian manifold.

(…) check details in article

2.6.1.3.2 Bregman divergence and Riemannian metric

Bregman divergence and Riemannian metric/Legendre transformation inducing a dual structure in S/ Two dual metrics in matrix form = same metric expressed in different coordinate systems giving the same local distance/two dual divergences.

2.6.1.3.3 Dual affine structure and Pythagorean theorem

Affine structure in S different from Riemannian structure in S

2.6.1.3.4 Geometry derived from general divergence function

concepts from differential geometry. Let us consider a manifold S having a local coordinate system $z = (z_1, ..., z_n)$. Let us consider a differentiable function $D[y : z]$, satisfying the condition of divergence. Then, the Taylor expansion,

$$D(z + dz, z) = \frac{1}{2} \sum g_{ij}(z) dz_i dz_j \qquad (114)$$

is a positive definite quadratic form. Hence, a Riemannian metric is induced in S by the second-order derivatives of D at $y = z$, that is,

$$g_{ij}(z) = \frac{\partial^2}{\partial z_i \partial z_j} D(z : y) \big|_{y=z} . \qquad (115)$$

It is easy to prove that this is a tensor.

We next consider an affine connection which is necessary for defining a geodesic. When we consider tangent spaces at all points z ∈ S, they are a collection of local linear approximations of S at different points. Such a collection is a fibre bundle called the tangent bundle. A geodesic is an extension of the "straight line", and is defined as a curve of which the tangent directions are the same along the curve. To define this "sameness", we need to connect tangent spaces defined at different points and define which directions are the same in different tangent spaces.

The basis vectors ei have always thesame directions in a flat space, provided affine coordinates are taken. In this special case, ei(z) and ei(z+dz) have the same direction. This is the case with a Bregman divergence. When a space is curved, the directions of ei(z) and ei(z +dz) are different, and we cannot have an affine coordinate system in general.

$$\delta e_j = \sum \Gamma^k_{ij}(z) dz^i e_k. \qquad (120)$$

Therefore, if we define the coefficients Γ^k_{ij}, we have a correspondence between two nearby tangent spaces. This is an affine connection, and Γ^k_{ij} are called the coefficients of the affine connection. Note that $e_i(z + dz)$ and $e_i(z)$ belong to different tangent spaces, so that we cannot subtract one from the other directly. We have defined the intrinsic difference δe_i by the above equation.

Covariant derivative operator:

$$\nabla_X Y = \sum \left(\frac{\partial Y^k}{\partial z^i} + \sum \Gamma^k_{ij} Y^j \right) X_i e_k, \qquad (121)$$

for two vector fields

$$X = \sum X_i e_i, \qquad (122)$$

$$Y = \sum Y_j e_j. \qquad (123)$$

The basis vectors are regarded as vector fields, and the covariant derivative of vector field $e_j(z)$ in the direction of $e_i(z)$ is

$$\nabla_{e_i} e_j = \sum \Gamma_{ij}^k e_k. \tag{124}$$

Affine connection derived from divergence:

$$\Gamma_{ijk}(z) = -\frac{\partial^3}{\partial z_i \partial z_j \partial y_k} D[z : y]_{|y=z}. \tag{126}$$

Geodesic equation:

$$\frac{d^2 z_k(t)}{dt^2} + \sum \Gamma_{ijk} \dot{z}_j \dot{z}_i = 0. \tag{128}$$

Since our affine connection is different from that derived from the Riemannian metric, it is not a curve of minimal distance connecting two points. But it keeps straightness in the sense of this affine connection, because the tangent direction never changes along the curve.

Dual affine connection from dual divergence:

$$\Gamma_{ijk}^* = -\frac{\partial^3}{\partial y_i \partial y_j \partial z_k} D[z : y]_{|y=z}.$$

Two affine connections are said to be mutually dual when

$$D_X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X^* Z \rangle \tag{131}$$

holds for three vector fields X, Y and Z. In terms of the components, this relation can be written as

$$\Gamma_{kij} + \Gamma_{kji}^* = \frac{\partial}{\partial z_k} g_{ij}. \tag{132}$$

The Riemannian connection (Levi-Civita connection) Γ_{ijk}^0 is given by

$$\Gamma_{ijk}^0 = \frac{1}{2}\left(\frac{\partial}{\partial z_i} g_{jk} + \frac{\partial}{\partial z_j} g_{ki} - \frac{\partial}{\partial z_k} g_{ij} \right). \tag{133}$$

It is the average of the two connections,

$$\Gamma_{ijk}^0 = \frac{1}{2}\left(\Gamma_{ijk} + \Gamma_{ijk}^* \right). \tag{134}$$

When two affine connections are dually coupled, we have a tensor

$$T_{ijk} = \Gamma_{ijk} - \Gamma^*_{ijk}, \tag{136}$$

which is symmetric with respect to the three indices i, j and k. Therefore, by using this tensor, the two dually coupled affine connections can be written as

$$\begin{aligned} \Gamma_{ijk} &= \Gamma^0_{ijk} - \frac{1}{2}T_{ijk}, \\ \Gamma^*_{ijk} &= \Gamma^0_{ijk} + \frac{1}{2}T_{ijk}. \end{aligned} \tag{137}$$

4.3. Geometry of Bregman divergence. We have already studied the geometry derived from a Bregman divergence,

$$D[y : z] = k(y) - k(z) - \{\text{Grad } k(z)\} \cdot (y - z). \tag{138}$$

From (115), the Riemannian metric is given by the Hessian

$$g_{ij}(z) = \frac{\partial^2}{\partial z_i \partial z_j} k(z). \tag{139}$$

Furthermore, we see from (126) that

$$\Gamma_{ijk}(z) = 0 \tag{140}$$

holds. Hence, the space is flat with respect to this connection, and the related coordinates z are affine. However, note that $\Gamma^*_{ijk}(z) \neq 0$. If we use the dual affine coordinate system z^*, then the dual affine connection $\Gamma^*_{ijk}(z^*)$ vanishes in the dual coordinate system.

Theorem 9. The α-divergence is the canonical divergence in the dually flat space M_n of positive measures. The KL-divergence is the canonical divergence in the dually flat space S_n of discrete probability distributions.

4.7. Symmetric divergence. A divergence is not symmetric in general. However, we can construct a symmetric divergence $D_s[z : y]$ from an asymmetric $D[z : y]$ by

$$D_s[z : y] = \frac{1}{2}\{D[z : y] + D[y : z]\}. \tag{158}$$

Its geometry is elucidated by the following theorem.

Theorem 10. The symmetrized divergence gives the same Riemannian metric as the asymmetric one. The derived affine connections are self-dual, and is the Riemannian connection,

$$\Gamma^S_{ijk} = \Gamma^{*S}_{ijk} = \Gamma^0_{ijk}. \tag{159}$$

2.6.1.4 Differential Geometry Derived From Divergence Functions – Information Geometry Approach

Abstract: *We study differential-geometrical structure of an information manifold equipped with a divergence function. A divergence function generates a Riemannian metric and furthermore it provides a symmetric third-order tensor, when the divergence is asymmetric. This induces a pair of affine connections dually coupled to each other with respect to the Riemannian metric. This is the arising emerged from information geometry. When a manifold is dually flat (it may be curved in the sense of the Levi-Civita connection), we have a canonical divergence and a pair of convex functions from which the original dual geometry is reconstructed. The generalized Pythagorean theorem and projection theorem hold in such a manifold. This structure has lots of applications in information sciences including statistics, machine learning, optimization, computer vision and Tsallis statistical mechanics. The present article reviews the structure of information geometry and its relation to the divergence function. We further consider the conformal structure given rise to by the generalized statistical model in relation to the power law.*

Divergence function:

- Symmetric => Riemannian metric and Levi-Civita connection
- Asymmetric => 3-tensor symmetric (=0 in symmetric case) => pair of affine connections dually coupled wrt Riem metric (i.e. T + g => dual gamma's) – Levi-Civita = average of dual gamma's
    - Example: K-L divergence

Gamma's are not metric connections but are dually coupled with respect to the Riemannian metric in the sense that the parallel transports of a vector by the two affine connections keep their inner product invariant with respect to the Riemannian metric. The duality can be expressed in terms of the related covariant derivatives.

When a Riemann-Christoffel curvature vanishes with respect to one affine connection, it vanishes automatically with respect to the dual affine connection. We have a **dually flat manifold** in this case, even though the Riemannian curvature with respect to the Levi-Civita connection does not vanish in general. A dually flat Riemannian manifold has nice properties such as the generalized Pythagorean theorem and projection theorem. Moreover, when a manifold is dually flat, we have two convex functions from which a canonical divergence is uniquely determined. The canonical divergence generates the original dually flat Riemannian structure. Euclidean space is a special example of the dually flat manifold which has a symmetric divergence given by the square of the Euclidean distance.

A dually flat manifold has two affine coordinate systems related to the two flat affine connections. They are connected by the Legendre transformation of the two convex functions. The canonical divergence is given as the Bregman divergence.

What is the natural divergence function to be introduced in a manifold? We study this question in the case of a manifold of probability distributions. We impose an invariant criterion such that the geometry is invariant under bijective transformations of random variables [Chentsov, 1982], [Picard, 1992] (more generally invariant by using sufficient statistics). Then, the Kullback-Leibler divergence is given as the unique canonical divergence. We further extend our notions to the manifold of positive measures [Amari, 2009]. We have invariant structure [Chentsov, 1982; Amari, 1985] and non-invariant flat structure [Amari and Ohara, 2011] which is related to the Tsallis entropy [Tsallis, 2009].

2.6.1.5 *Information Geometry of Positive Measures and Positive-Definite Matrices (2014)*

Entropy 2014, 16, 2131-2145

Related to central cluster theorem (…)

2.6.2 Baez, John

https://johncarlosbaez.wordpress.com/2010/10/22/information-geometry/

2.6.2.1 *Thermodynamic length*

Cf. other article (classical thermodynamics): http://arxiv.org/abs/0706.0559v2

Thermodynamic length:

Thermodynamic length is a natural measure of the distance between equilibrium thermodynamic states, which equips the surface of thermodynamic states with a Riemannian metric and defines the length of a quasi-static transformation as the number of natural fluctuations along that path. Unlike the entropy or free energy changes, which are state functions, the thermodynamic length explicitly depends on the path taken through thermodynamic state space. Thermodynamic length is of fundamental interest to the generalization of thermodynamics to finite time (rather than infinity slow) transformations. Minimum distance paths are geodesics on the Riemannian manifold and minimize the dissipation for slow, but finite time transformations.

There are deep connections between thermodynamic length, information theory and the statistical physics of small systems far-from-equilibium.

The configurational probability distribution is given by the Gibbs ensemble:

$$p(x|\lambda) = \frac{1}{Z} e^{-\beta H(x,\lambda)} = \frac{1}{Z} e^{-\lambda^i(t) X_i(x)} \qquad (1)$$

where x is the configuration, t is time, $\beta = 1/k_{\mathrm{B}}T$ is the reciprocal temperature (T) of the environment in natural units, (k_{B} is the Boltzmann constant), Z is the partition function, and H is the Hamiltonian of the system. This total Hamiltonian is split into a collection of collective variables X_i and conjugate generalized forces λ^i, $\beta H = \lambda^i(t) X_i(x)$. We use the Einstein convention

The λ's are the experimentally controllable parameters of the system and define the accessible thermodynamic state space. For example, in the isothermal-isobaric ensemble we have X = {U, V } and λ = {β, βp}, where U is the internal energy, V is the volume and p is the external pressure. Modern experimental techniques have broadened the range of controllable parameters beyond those considered in standard thermodynamics.

The partition function that normalizes the probability distribution, Z, is directly related to the free energy F (Gibbs potential), the free entropy ψ (Massieu potential) and entropy S:

$$\ln Z = -\beta F = \psi = S - \lambda^i \langle X_i \rangle \qquad (2)$$

Angled brackets indicate an average over the appropriate equilibrium ensemble. The first derivatives of the free entropy give the first moments of the collective variables,

$$\frac{\partial \psi}{\partial \lambda^i} = -\langle X_i \rangle \tag{3}$$

and the second derivative yields the covariance matrix,

$$g_{ij} = \frac{\partial^2 \psi}{\partial \lambda^i \partial \lambda^j} = -\frac{\partial \langle X_i \rangle}{\partial \lambda^j} = \big\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \big\rangle \,. \tag{4}$$

The covariance matrix g_{ij} is positive semi-definite and varies smoothly from point to point, except at macroscopic phase transitions. Therefore, we can use the covariance matrix as a metric tensor and naturally equip the manifold of thermodynamic states with a Riemannian metric. Recall that a metric provides a mea-

With this definition we can make an important connection to statistical estimation theory, since the thermodynamic metric tensor Eq. (4) is then identical to the Fisher information matrix

$$\begin{aligned} g_{ij}(\lambda) &= \sum_x p(x) \frac{\partial \ln p(x)}{\partial \lambda^i} \frac{\partial \ln p(x)}{\partial \lambda^j} \\ &= \sum_x p(x)(X_i + \frac{\partial \psi}{\partial \lambda^i})(X_j + \frac{\partial \psi}{\partial \lambda^j}) \\ &= \big\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \big\rangle \end{aligned} \tag{6}$$

According to the Cramér-Rao inequality the variance of any unbiased estimator is at least as high as the inverse of the Fisher information [19].

In 1945 Rao introduced the 'entropy differential metric', the distance between two distributions arising from the Riemannian metric over the parameter space with the Fisher information metric tensor. This entropy differential metric is identical to the thermodynamic length when, as here, the variables are conjugate parameters of a Gibbs ensemble.

2.6.2.2 Quantum statistical mechanics

Gibbs ensemble: the mixed state that maximizes entropy subject to the constraint that some observable have a given value.

We can do the same thing in quantum mechanics, and we can even do it for a bunch of observables at once. Suppose we have some observables X_1, …, X_n and we want to find the mixed state rho that maximizes entropy subject to these constraints:

$$\langle X_i \rangle = x_i \tag{1}$$

for some numbers x_i. Then a little exercise in Lagrange multipliers shows that the answer is the Gibbs state:

$$\rho = \frac{1}{Z}\exp(-\lambda_1 X_1 + \cdots + \lambda_n X_n)$$

This answer needs some explanation. First of all, the numbers lambda_i are called Lagrange multipliers. You have to choose them right to get (1). So, in favorable cases, they will be functions of the numbers lambda_i. And when you're really lucky, you can solve for the numbers x_i in terms of the numbers lambda_i. We call lambda_i the conjugate variable of the observable X_i. For example, the conjugate variable of energy is inverse temperature!

But third: what's that number Z ? It begins life as a humble normalizing factor. Its job is to make sure ρ has trace equal to 1:

$$Z = \mathrm{tr}(\exp(-\lambda_1 X_1 + \cdots + \lambda_n X_n))$$

However, once you get going, it becomes incredibly important! It's called the **partition function** of your system.

As an example of what it's good for, it turns out you can compute the numbers x_i as follows:

$$x_i = -\frac{\partial}{\partial \lambda_i} \ln Z$$

In other words, you can compute the expected values of the observables X_i by differentiating the log of the partition function:

$$\langle X_i \rangle = -\frac{\partial}{\partial \lambda_i} \ln Z$$

But we can go further: after the 'expected value' or 'mean' of an observable comes its variance, which is the square of its standard deviation:

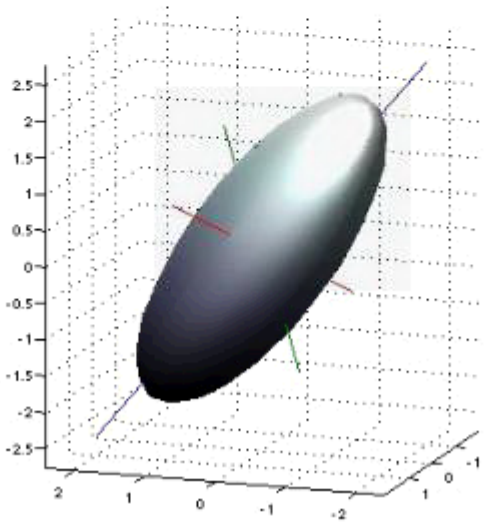
$$(\Delta A)^2 = \langle A^2 \rangle - \langle A \rangle^2$$

This measures the size of fluctuations around the mean. And in the Gibbs state, we can compute the variance of the observable X_i as the *second* derivative of the log of the partition function:

$$\langle X_i^2 \rangle - \langle X_i \rangle^2 = \frac{\partial^2}{\partial^2 \lambda_i} \ln Z$$

But when we've got lots of observables, there's something better than the variance of each one. There's the covariance matrix of the whole lot of them! Each observable X_i fluctuates around its mean value x_i ... but these fluctuations are not independent! They're *correlated*, and the covariance matrix says how.

All this is very visual, at least for me. If you imagine the fluctuations as forming a blurry patch near the point $(x_1, \dots, x_n)$, this patch will be ellipsoidal in shape, at least when all our random fluctuations are Gaussian. And then the *shape* of this ellipsoid is precisely captured by the covariance matrix! In particular, the eigenvectors of the covariance matrix will point along the principal axes of this ellipsoid, and the eigenvalues will say how stretched out the ellipsoid is in each direction!



(2)

observables X_i don't commute. However, the *real part* of the covariance matrix is symmetric, even in quantum mechanics. So let's define

$$g_{ij} = \mathrm{Re}\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

You can check that the matrix entries here are the second derivatives of the partition function:

$$g_{ij} = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

And now for the cool part: this is where information geometry comes in! Suppose that for any choice of values x_i we have a Gibbs state with

$$\langle X_i \rangle = x_i$$

Then for each point

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

we have a matrix

$$g_{ij} = \mathrm{Re}\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle = \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \ln Z$$

And this matrix is not only symmetric, it's also positive. And when it's positive definite we can think of it as an inner product on the tangent space of the point x. In other words, we get a Riemannian metric on $\mathbb{R}^n$. This is called the **Fisher information metric**.

I hope you can see through the jargon to the simple idea. We've got a space. Each point in this space describes the maximum-entropy state of a quantum system for which our observables have specified mean values. But in each of these states, the observables are random variables. They don't just sit at their mean value, they fluctuate! You can picture these fluctuations as forming a little smeared-out blob in our space. To a first approximation, this blob is an ellipsoid. And if we think of this ellipsoid as a 'unit ball', it gives us a standard for measuring the length of any little vector sticking out of our point. In other words, we've got a Riemannian metric: the Fisher information metric!

Remember the story so far: we've got a physical system that's in a state of maximum entropy. I didn't emphasize this yet, but that happens whenever our system is in thermodynamic equilibrium

This (2) weirdly warped geometry is an example of an 'information geometry': a geometry that's defined using the concept of information. This shouldn't be surprising: after all, we're talking about maximum entropy, and entropy is related to information. But I want to gradually make this idea more precise.

Crooks does the classical case — so let's do the quantum case, okay? Last time I claimed that in the quantum case, our maximum-entropy state is the **Gibbs state**

$$\rho = \frac{1}{Z} e^{-\lambda^i X_i}$$

where λ^i are the 'conjugate variables' of the observables X_i, we're using the Einstein summation convention to sum over repeated indices that show up once upstairs and once downstairs, and Z is the **partition function**

$$Z = \mathrm{tr}\left(e^{-\lambda^i X_i}\right)$$

Also last time I claimed that it's tremendously fun and enlightening to take the derivative of the logarithm of Z. The reason is that it gives you the mean values of your observables:

$$\langle X_i \rangle = -\frac{\partial}{\partial \lambda^i} \ln Z$$

But now let's take the derivative of the logarithm of ρ. Remember, ρ is an operator — in fact a density matrix. But we can take its logarithm as explained last time, and the usual rules apply, so starting from

$$\rho = \frac{1}{Z} e^{-\lambda^i X_i}$$

we get

$$\ln \rho = -\lambda^i X_i - \ln Z$$

Next, let's differentiate both sides with respect to λ^i. Why? Well, from what I just said, you should be itching to differentiate $\ln Z$. So let's give in to that temptation:

$$\frac{\partial}{\partial \lambda^i} \ln \rho = -X_i + \langle X_i \rangle$$

Hey! Now we've got a formula for the 'fluctuation' of the observable X_i — that is, how much it differs from its mean value:

$$X_i - \langle X_i \rangle = -\frac{\partial \ln \rho}{\partial \lambda^i}$$

This is incredibly cool! I should have learned this formula decades ago, but somehow I just bumped into it now. I knew of course that $\ln \rho$ shows up in the formula for the **entropy**:

$$S(\rho) = \mathrm{tr}(\rho \ln \rho)$$

But I never had the brains to think about $\ln \rho$ all by itself. So I'm really excited to discover that it's an interesting entity in its own right — and fun to differentiate, just like $\ln Z$.

Now we get our cool formula for g_{ij}. Remember, it's defined by

$$g_{ij} = \mathrm{Re}\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle$$

But now that we know

$$X_i - \langle X_i \rangle = -\frac{\partial \ln \rho}{\partial \lambda^i}$$

we get the formula we were looking for:

$$g_{ij} = \mathrm{Re}\langle \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j} \rangle$$

Beautiful, eh? And of course the expected value of any observable A in the state ρ is

$$\langle A \rangle = \mathrm{tr}(\rho A)$$

so we can also write the covariance matrix like this:

$$g_{ij} = \mathrm{Re}\,\mathrm{tr}(\rho \frac{\partial \ln \rho}{\partial \lambda^i} \frac{\partial \ln \rho}{\partial \lambda^j})$$

Lo and behold! This formula makes sense whenever ρ is *any* density matrix depending smoothly on some parameters λ^i. We don't need it to be a Gibbs state! So, we can work more generally.

Indeed, whenever we have *any* smooth function from a manifold to the space of density matrices for some Hilbert space, we can define g_{ij} by the above formula! And when it's positive definite, we get a Riemannian metric on our manifold: the **Bures information metric**.

The classical analogue is the somewhat more well-known 'Fisher information metric'. When we go from quantum to classical, operators become functions and traces become integrals. There's nothing complex anymore, so taking the real part becomes unnecessary. So the Fisher information metric looks like this:

$$g_{ij} = \int_\Omega p(\omega) \frac{\partial \ln p(\omega)}{\partial \lambda^i} \frac{\partial \ln p(\omega)}{\partial \lambda^j} \, d\omega$$

Here I'm assuming we've got a smooth function p from some manifold M to the space of probability distributions on some measure space $(\Omega, d\omega)$. Working in local coordinates λ^i on our manifold M, the above formula defines a Riemannian metric on M, at least when g_{ij} is positive definite. And that's the **Fisher information metric**!

Crooks says more: he describes an experiment that would let you measure the length of a path with respect to the Fisher information metric — at least in the case where the state $\rho(x)$ is the Gibbs state with $\langle X_i \rangle = x_i$. And that explains why he calls it 'thermodynamic length'.

2.6.2.3 Other link with statistical mechanics

Reference?

The density matrix ρ of the equilibrium states for the Hamiltonian $\mathcal{H}$ is defined as

$$\rho = \frac{1}{Z(\beta)} e^{-\beta\mathcal{H}} = e^{-\beta(\mathcal{H}-F)}, \tag{1}$$

where the parameters $\beta, Z(\beta)$ and F denote the inverse temperature $\beta = 1/k_{\mathrm{B}}T$, the partition function $Z(\beta) = \mathrm{Tr}\exp[-\beta\mathcal{H}]$ and the free energy $F = -k_{\mathrm{B}}T\log Z(\beta)$, respectively. Using the density matrix ρ, we define the matrix $\gamma(\rho)$ as

$$\gamma(\rho) = -\log\rho = \beta\mathcal{H} + \log Z(\beta), \tag{2}$$

which is called as entropy operator [10–12]. The expectation value of $\gamma(\rho)$ is equal to the thermal entropy, $S = k_{\mathrm{B}}\langle\gamma(\rho)\rangle$. According to the identity

Role of Fisher Information in Quantum Mechanics.

configurational probability distribution is given by the Gibbs ensemble, [16]

$$p(x|\lambda) = \frac{1}{Z} e^{-\beta H(x,\lambda)} = \frac{1}{Z} e^{-\lambda^i(t)X_i(x)} \tag{1}$$

where x is the configuration, t is time, $\beta = 1/k_{\mathrm{B}}T$ is the reciprocal temperature (T) of the environment in natural units, (k_{B} is the Boltzmann constant), Z is the partition function, and H is the Hamiltonian of the system. This total Hamiltonian is split into a collection of collective variables X_i and conjugate generalized forces λ^i, $\beta H = \lambda^i(t)X_i(x)$. We use the Einstein convention

The partition function that normalizes the probability distribution, Z, is directly related to the free energy F (Gibbs potential), the free entropy ψ (Massieu potential) and entropy S:

$$\ln Z = -\beta F = \psi = S - \lambda^i\langle X_i\rangle \tag{2}$$

Angled brackets indicate an average over the appropriate equilibrium ensemble. The first derivatives of the free entropy give the first moments of the collective variables,

$$\frac{\partial \psi}{\partial \lambda^i} = -\langle X_i \rangle \tag{3}$$

and the second derivative yields the covariance matrix,

$$g_{ij} = \frac{\partial^2 \psi}{\partial \lambda^i \partial \lambda^j} = -\frac{\partial \langle X_i \rangle}{\partial \lambda^j} = \big\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \big\rangle \,. \tag{4}$$

The covariance matrix g_{ij} is positive semi-definite and varies smoothly from point to point, except at macroscopic phase transitions. Therefore, we can use the covariance matrix as a metric tensor and naturally equip the manifold of thermodynamic states with a Riemannian metric. Recall that a metric provides a mea-

2.6.3 Burnham and Anderson

2.6.3.1 Kullback–Leibler information as a basis for strong inference in ecological studies (2001)

Wildlife Research, 2001, 28, 111–119

Abstract. We describe an information-theoretic paradigm for analysis of ecological data, based on Kullback–Leibler information, that is an extension of likelihood theory and avoids the pitfalls of null hypothesis testing. Information-theoretic approaches emphasise a deliberate focus on the *a priori* science in developing a set of multiple working hypotheses or models. Simple methods then allow these hypotheses (models) to be ranked from best to worst and scaled to reflect a strength of evidence using the likelihood of each model (g_i), given the data and the models in the set (i.e. $L(g_i \mid data)$). In addition, a variance component due to model-selection uncertainty is included in estimates of precision. There are many cases where formal inference can be based on all the models in the *a priori* set and this multi-model inference represents a powerful, new approach to valid inference. Finally, we strongly recommend inferences based on *a priori* considerations be carefully separated from those resulting from some form of data dredging. An example is given for questions related to age- and sex-dependent rates of tag loss in elephant seals (*Mirounga leonina*).

Null hypothesis: http://en.wikipedia.org/wiki/Null_hypothesis

Statistical inference can be done without a null hypothesis, thus avoiding the criticisms under debate. An approach to statistical inference that does not involve a null hypothesis is the following: for each candidate hypothesis, specify a statistical model that corresponds to the hypothesis; then, use model selection techniques to choose the most appropriate model. (The most common selection techniques are based on either **Akaike information criterion** or **Bayes factor**.)

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Excerpts:

The statistical null hypothesis testing approach is not wrong, but it is relatively uninformative and, thus, slows scientific progress and understanding.

Three general principles guide us in model-based inference in the sciences:

- **Simplicity and Parsimony**
  - Model selection (variable selection in regression is a special case) is a **bias v. variance trade-off** and this is the principle of parsimony (Fig. 1). Models with too few parameters (variables) have bias, whereas models with too many parameters (variables) may have poor precision or tend to identify effects that are, in fact, spurious (slightly different issues arise for count data v. continuous data). These considerations call for a balance between under- and over-fitted models –the so-called 'model selection problem' (see Forster 2000).

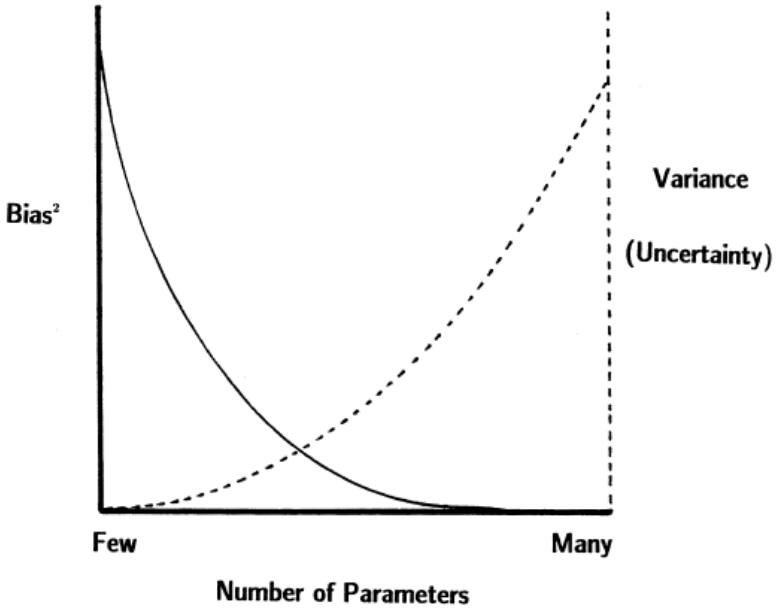


Fig. 1. The principle of parsimony: the conceptual trade-off between squared bias (solid line) and variance (i.e. uncertainty) versus the number of estimable parameters in the model. The best model has dimension (K_0) near the intersection of the two lines, while full reality lies far to the right of trade-off region.

- **Multiple Working Hypotheses**
  - Here, there is no null hypothesis; instead, there are several well-supported hypotheses (equivalently, 'models') that are being entertained. The a priori 'science' of the issue enters at this important point. Relevant empirical data are then gathered, analysed, and the results tend to support one or more hypotheses, while providing less support for other hypotheses. Repetition of this general approach leads to advances in the sciences. New or more elaborate hypotheses are added, while hypotheses with little empirical support are gradually dropped from consideration. At any one point in time, there are multiple hypotheses (models) still under consideration. An important feature of this multiplicity is that the number of alternative models should be kept small (Zucchini 2000); the analysis of, say, hundreds of models is not justified except when prediction is the only objective, or in the most exploratory phases of an investigation.
- **Strength of Evidence**

- Providing information to judge the 'strength of evidence' is central to science. Null-hypothesis- testing provides only arbitrary dichotomies (e.g. significant v. non-significant) and in the all-too-often-seen case where the null hypothesis is obviously false on a priori grounds, the test result is superfluous

Major advances.

- **Advance 1 – Kullback–Leibler information**
  - Concept related to Fisher's concept of a 'sufficient statistics'
  - K-L information is related to Boltzmann's entropy
  - Definition:

> Kullback–Leibler (K–L) information is a measure (a 'distance' in an heuristic sense) between conceptual reality, f, and approximating model, g, and is defined for continuous functions as the integral
>
> $$I(f,\ g) = \int f(x) \log_e \left(\frac{f(x)}{g(x \mid \theta)} \right) dx$$
>
> where f and g are n-dimensional probability distributions. K–L information, denoted $I(f, g)$, is the 'information' lost when model g is used to approximate reality, f. The analyst seeks an approximating model that loses as little information as possible; this is equivalent to minimising $I(f, g)$, over the set of models of interest (we assume there are R *a priori* models in the candidate set).
>
> Boltzmann's entropy H is $-I(f, g)$, although these quantities were derived along very different lines. Boltzmann derived the fundamental relationship between entropy (H) and probability (P) as
>
> $$H = \log_e(P)$$
>
> and because $H = -I(f, g)$, one can see that entropy, information and probability are linked, allowing probabilities to be multiplicative whereas information and entropy are additive.

  - K–L information can be viewed as an extension of the famous Shannon (1948) entropy and is often referred to as 'cross entropy'.
  - K–L information, by itself, will not aid in data analysis as both reality (f) and the parameters (theta) in the approximating model are unknown to us. H. Akaike made the next breakthrough in the early 1970s
- **Advance 2 – Estimation of Kullback–Leibler information (AIC)**

- Akaike (1973, 1974) found a formal relationship between K–L information (a dominant paradigm in information and coding theory) and maximum likelihood (the dominant paradigm in statistics) (see deLeeuw 1992). This finding makes it possible to combine estimation (e.g. maximum likelihood or least squares) and model selection under a single theoretical framework – optimisation. Akaike's breakthrough was the finding of an estimator of the expected, relative K–L information, based on the maximised log-likelihood function. Akaike's derivation (which is for large samples) relied on K–L information as averaged entropy and this lead to 'Akaike's information criterion' (AIC),

  $$\mathrm{AIC} = -2\log_e(L(\hat{\theta} \mid data)) + 2K,$$

  where $\log_e(L(\hat{\theta} \mid data))$ is the value of the maximised log-likelihood over the unknown parameters (θ), given the data and the model, and K is the number of estimable parameters in that approximating model. In the special case of least-
- Assuming that a set of a priori candidate models has been defined and is well supported by the underlying science, then AIC is computed for each of the approximating models in the set (i.e. gi, i = 1, 2, …, R). The model for which AIC is minimal is selected as best for the empirical data at hand. This is a simple, compelling concept, based on deep theoretical foundations (i.e. entropy, K–L information, and likelihood theory).
- When K is large relative to sample size n (which includes when n is small, for any K) there is a small-sample (secondorder) version of AIC – AIC*; this should be used unless n/K > ~40
- Both AIC and AIC* are estimates of expected, relative Kullback–Leibler information and are useful in the analysis of real data in the 'noisy' sciences. Assuming independence, AIC-based model selection is equivalent to certain cross-validation methods (Stone 1974, 1977) and this is an important property
- Akaike's general approach allows the best model in the set to be identified, but also allows the rest of the models to be easily ranked

-

> The principle of parsimony, or Occam's razor, provides a philosophical basis for model selection; Kullback–Leibler information provides an objective target based on deep, fundamental theory; and the information criteria (AIC and AIC_c), along with likelihood- or least-squares-based inference, provide a practical, general methodology for use in the analysis of empirical data. Objective data analysis can be rigorously based on these principles without having to assume that the 'true model' is contained in the set of candidate models – surely there are no true models in the biological sciences!

- Advance 3 – Likelihood of a model, given the data
  -

> The simple transformation $\exp(-\Delta_i/2)$, for $i = 1, 2, \ldots, R$, provides the likelihood of the model (Akaike 1981) given the data: $L(g_i \mid data)$. This is a likelihood function over the model set in the same sense that $L(\theta \mid data, g_i)$ is the likelihood over the parameter space (for model g_i) of the parameters θ, given the data (x) and the model (g_i). The relative likelihood of model i versus model j is $L(g_i \mid data)/L(g_j \mid data)$; this ratio does not depend on any of the other models under consideration. Without loss of generality we may assume model g_i is more likely than g_j. Then if this ratio is large (e.g. >10 is large), model g_j is a poor model to fit the data *relative* to model g_i. The expression $L(g_i \mid data)/L(g_j \mid data)$ can be regarded as an *evidence ratio* – the evidence for model i versus model j.

- Other advances – more specific (…)

NB: all information criteria are in fact approximations of Bayes Factors (BFs) (Congdon, 2006a) with certain assumptions such as large sample sizes.

Cf. Multimodel inference in ecology and evolution challenges and solutions, J . EVOL. BIOL . 24 (2011) 699–711

Table 1 Information criteria for model selection.

Information criterion	Formula*	References
Akaike Information Criterion	$AIC = -2 \cdot \ln L + 2k$	Akaike (1973)
AIC – small sample size correction	$AIC_C = -2 \cdot \ln L + \frac{2k(k+1)}{n-k-1}$	Hurvich & Tsai (1989)
Quasi-AIC	$QAIC = \frac{-2 \cdot \ln L}{\hat{c}} + 2k$	Lebreton *et al.* (1992)
Conditional AIC	$cAIC = -2 \cdot \ln L + 2k_C$	Vaida & Blanchard (2005); Liang *et al.* (2008)
Bayesian Information Criterion	$BIC = -2 \cdot \ln L + k \ln(n)$	Schwarz (1978)
Deviance Information Criterion	$DIC = -2 \cdot \ln L + 2k_D$	Spiegelhalter *et al.* (2002)

*L = likelihood function = $p(y|\theta)$, or, if random factors are explicitly separated as parameters (as in cAIC) = $p(y|\theta,u)$. NB, $-2 \cdot \ln L$ is also known as the 'deviance'. k = number of parameters in the model; n = sample size; $\hat{c}$ = overdispersion parameter; k_C = effective number of degrees of freedom (cAIC); k_D = effective number of parameters (DIC). See listed references for additional details of formula components.

2.6.4 Kass, Robert E.

2.6.4.1 Differential Geometry in Statistical Inference (1987)

Excellent introduction for the underlying ideas and seminal results of Information Geometry.

Excerpts:

Geometrical analyses of parametric inference problems have developed from two appealing ideas: that a local measure of distance between members of a family of distributions could be based on Fisher information, and that the special place of exponential families in statistical theory could be understood as being intimately connected with their loglinear structure. The first led Jeffreys (1946) and Rao (1945) to introduce a Riemannian metric defined by Fisher information, while the second led Efron (1975) to quantify departures from exponentiality by defining the curvature of a statistical model. The

One reason that the role of the mixture curvature in (1) and in the variance decomposition went unnoticed in Efron's paper was that he had not made the underlying geometrical structure explicit: to calculate statistical curvature at a given value θ_0 of a single parameter θ in a curved exponential family, Efron used the natural parameter space with the inner product defined by Fisher information at the natural parameter point corresponding to θ_0. In order to calculate the curvature at a new point θ_1, another copy of the natural parameter space with a different inner product (namely, that defined by Fisher information at the natural parameter point corresponding to θ_1) would have to be used. The appropriate gluing together of these spaces into a single structure involves three basic elements: a manifold, a Riemannian metric, and an affine connection. Riemannian geometry involves the study of geometry determined by the metric and its uniquely associated Riemannian connection. In his discussion to Efron's paper, Dawid (1975) pointed out that Efron had used the Riemannian metric defined by Fisher information, but that he had effectively used a non-Riemannian affine connection, now called the exponential connection, in calculating statistical curvature. Although Dawid did not identify the role of the mixture curvature in (1), he did draw attention to the mixture connection as an alternative to the exponential connection. (Geodesics with respect to the exponential connection form exponential families, while geodesics with respect to the mixture connection form families of mixtures; thus, the terminology.) Amari, who had much earlier researched the Riemannian geometry of Fisher information, picked up on Dawid's observation, specified the framework, and provided the results outlined above.

The manifold with the associated linear spaces is structured in what is usually called a tangent bundle, the elements of the linear spaces being tangent vectors. For curved exponential families, the linear spaces are finite-dimensional, but to analyze general families this does not suffice so Amari uses Hilbert spaces. When these are appropriately glued together, the result is a Hilbert bundle. The idea stems from Dawid's remark that the tangent vectors can be identified with score functions, and these in turn are functions having zero expectation. As his Hilbert space at a distribution P, Amari takes the subspace of the usual $L_2(P)$ Hilbert space consisting of functions that have zero expectation with respect to P. This clearly furnishes the extension of the information metric, and has been used by other authors as well, e.g., Beran (1977). Amari then defines the exponential and mixture connections and notes that these make the Hilbert bundle flat, and that the inherited connections on the usual tangent bundles agree with those already defined there. He then decomposes each Hilbert space into tangential and normal components, which is exactly what is needed to define statistical curvature in the general setting. Amari goes on to construct an "exponential bundle" by associating with each distribution a finite-dimensional linear space containing vectors defined by higher derivatives of the loglikelihood function, and using structure

In his Annals paper, Amari also noted an interesting relationship between the exponential and mixture connections: they are, in a sense he defined, mutually dual. Furthermore, a one-parameter family of connections, which Amari called the α-connections, may be defined in such a way that for each α the α-connection and the $-\alpha$-connection are mutually dual, while $\alpha=1$ and -1 correspond to the exponential and mixture connections. See Amari's Theorem 2.1. This family coincides with that introduced by Centsov (1971) for multinomial distributions. When the family of densities on which these connections are defined is an exponential family, the space is flat with respect to the exponential and mixture connections, and the natural parametrization and mean-value parameterization play special roles: they become affine coordinate systems for the two respective connections and are related by a Legendre transformation. The duality in this case can incorporate the convex duality theory of exponential families (see Barndorff-Nielsen, 1978, and also Section 2 of his paper in this volume). In Theorem 2.2 Amari points out that such a pair of coordinate systems exists whenever a space is flat with respect to an α-connection (with $\alpha \neq 0$). For such spaces, Amari defines α-divergence, a quasi-distance between two members of the family based on the relationship provided by the Legendre transformation. In Theorem 2.4 he shows that the element of a curved exponential family that minimizes the α-divergence from a point in the exponential family parameter space may be found by following the α-geodesic that contains the

As soon as the α-connections are constructed a mathematical question arises. On one hand, the α-connections may be considered objects of differential geometry without special reference to their statistical origin. On the other hand, they are not at all arbitrary. They are the simplest one-parameter family of connections based on the first three moments of the score function. What is it about their special form that leads to the many special properties of α-connections (outlined by Amari in Section 2)?

Lauritzen has posed this question and has provided a substantial part of the answer. Given any Riemannian manifold M with metric g there is a unique Riemannian connection $\bar{\nabla}$. Given a covariant 3-tensor D that is symmetric in its first two arguments and a nonzero number c, a new (symmetric) connection is defined by

$$\nabla = \bar{\nabla} + c \cdot D \tag{2}$$

which means that given vector fields X and Y,

$$\nabla_X Y = \bar{\nabla}_X Y + c \cdot \tilde{D}(X,Y)$$

where

$$g(\tilde{D}(X,Y),Z) \equiv D(X,Y,Z)$$

for all vector fields Z. Now, when M is a family of densities and g and D are defined, in terms of an arbitrary parameterization, as

$$g(\partial_i, \partial_j) = E(\partial_i \ell \partial_j \ell)$$

$$D(\partial_i, \partial_j, \partial_k) = E(\partial_i \ell \partial_j \ell \partial_k \ell)$$

where ℓ is the loglikelihood function, and if $c = -\alpha/2$, then (2) defines the α-connection.

In this statistical case, D is not only symmetric in its first two arguments, as it must be in (2), it is symmetric in all three. Lauritzen therefore defines an abstract statistical manifold to be a triple (M,g,D) in which M is a smooth m-dimensional manifold, g is a Riemannian metric, and D is a completely symmetric covariant 3-tensor. With this additional symmetry constraint alone, he then proceeds to establish a large number of basic properties, especially those relating to the duality structure Amari described. The treatment is "fully geometrical" or "coordinate-free." This is aesthetically appealing, especially to those who learned linear models in the coordinate-free setting. Lauritzen's primary purpose is to show that the appropriate mathematical object of study is one that is not part of the standard differential geometry, but does have many special features arising from an apparently simple structure. He not only presents the abstract generalities about α-connections on statistical manifolds, he also examines five examples in full detail. The first is the univariate Gaussian model, the second is the inverse Gaussian model, the third is the two-parameter gamma model, and the last two are specially constructed models that display interesting possibilities of the non-standard geometries of α-connections. In particular, the latter two statistical manifolds are not what Lauritzen calls "conjugate symmetric" and so the sectional curvatures do not determine the Riemann tensor (as they do in Riemannian geometry). He also discusses the construction of geodesic folia-

2.6.4.2 The Geometry of Asymptotic Inference (1989)

Stat. Sc. 1989, Vol. 4, No 3, p.188-234

2.6.4.2.1 Introduction

Fundamental ideas underpinning information geometry:

- Base a local measure of distance between members of a family of distributions on K-L divergence or equivalently on Fisher information
  - Fisher metric = Riemannian
    - Cramer-Rao lower bound (1945)
    - Jeffreys' invariant prior for estimation problems (1946)
- Connect the special role of exponential families in statistical theory with their loglinear structure
  - Quantify departures from exponentiality by defining the curvature of a statistical model
    - Efron (1975)
    - Dawid (1975) – comments on Efron's

Connection between these two ideas: the appropriate foundation of Efron's measure of curvature actually involves non-Riemannian geometry.

[Role of reference priors in Bayesian inference & Use of conditioning in non-Bayesian inference]

Reminder:

- Homeomorphism – mapping and its inverse are continuous
- Diffeomorphism – mapping and its inverse are smooth (Infinite diff) – Jacobians are invertible

Local vs global parameterizations => impose smooth manifold structure (diffeomorphisms).

2.6.5 Nielsen, Frank

2.6.5.1 Statistical exponential families: A digest with flash cards

http://arxiv.org/abs/0911.4863v2

Statistical exponential families: A digest with flash cards*

Frank Nielsen† and Vincent Garcia‡

May 16, 2011 (v2.0)

Abstract

This document describes concisely the ubiquitous class of exponential family distributions met in statistics. The first part recalls definitions and summarizes main properties and duality with Bregman divergences (all proofs are skipped). The second part lists decompositions and related formula of common exponential family distributions. We recall the Fisher-Rao-Riemannian geometries and the dual affine connection information geometries of statistical manifolds. It is intended to maintain and update this document and catalog by adding new distribution items.

Excerpts:

Cencov[2] proved [Čen72] (see also [Leb05] and [GS01] for an equivalent in quantum information geometry) that the only Riemannian metric that "makes sense" for statistical manifolds is the Fisher information metric:

$$I(\theta) = \left[\int \frac{\partial \log p(x;\theta)}{\partial \theta_i} \frac{\partial \log p(x;\theta)}{\partial \theta_j} p(x;\theta) \mathrm{d}x \right] = [g_{ij}] \tag{15}$$

The infinitesimal length element is given by

$$\mathrm{d}s^2 = \sum_{i=1}^{d} \sum_{i=1}^{d} \mathrm{d}\theta_i^T \nabla^2 F(\theta) \mathrm{d}\theta_j \tag{16}$$

Cencov proved that for a non-singular transformation of the parameters $\lambda = f(\theta)$, the information matrix

$$I(\lambda) = \left[\frac{\partial \theta_i}{\partial \lambda_j} \right] I(\theta) \left[\frac{\partial \theta_i}{\partial \lambda_j} \right], \tag{17}$$

is such that $\mathrm{d}s^2(\lambda) = \mathrm{d}s^2(\theta)$. Equipped with the tensor $I(\theta)$, the metric distance between two distributions on a statistical manifold can be computed from the geodesic length (e.g., shortest path):

[2] also written as Chentsov

$$D(p(x;\theta_1), p(x;\theta_2)) = \min_{\theta(t) \mid \theta(0)=\theta_1, \theta(1)=\theta_2} \int_0^1 \sqrt{\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^T I(\theta) \frac{\mathrm{d}\theta}{\mathrm{d}t}} \mathrm{d}t \tag{18}$$

Rao's geodesic distance is invariant by non-singular transformations. The multinomial Fisher-Rao-Riemannian geometry yields a spherical geometry, and the normal Fisher-Rao-Riemannian geometry yields a hyperbolic geometry [KV97, CSS05]. Indeed, the Fisher information matrix for univariate normal distributions is

$$I(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{19}$$

The Fisher information matrix can be interpreted as the Hessian of the Shannon entropy:

$$g_{ij}(\theta) = -E\left[\frac{\partial^2 \log p(x;\theta)}{\partial \theta_i \partial \theta_j} \right] = \frac{\partial^2 H(p)}{\partial \theta_i \partial \theta_j}, \tag{20}$$

with $H(p) = -\int p(x;\theta) \log p(x;\theta) \mathrm{d}x$.

For an exponential family, the Kullback-Leibler divergence is a Bregman divergence on the natural parameters. Using Taylor approximation with exact remainder, we get $\mathrm{KL}(\theta||\theta + \mathrm{d}\theta) = \frac{1}{2}\mathrm{d}\theta^T \nabla^2 F(\theta) \mathrm{d}\theta$. Moreover, the infinitesimal Rao distance is $\sqrt{\mathrm{d}\theta^T I(\theta) \mathrm{d}\theta}$ for $I(\theta) = \nabla^2 F(\theta)$. We deduce that $D(\theta, \theta + \mathrm{d}\theta) = \sqrt{2\mathrm{KL}(\theta||\theta + \mathrm{d}\theta)}$.

2.6.6 Misc

2.6.6.1 Dynamics of the Fisher Information Metric

http://arxiv.org/abs/cond-mat/0410452v1

Application of the variational principle to the functional built on Fisher Information – dynamical derivation of E-L movement equations.

J[g_μν(θ_i)]

3.2 Dynamics

We now have all the necessary tools to introduce the invariant functional which describes the dynamics of Fisher information metric. Because of [4], it seems natural to posit that the functional $J[g_{\mu\nu}]$, describing the dynamics of the metric, is invariant under general coordinate transformations $\theta^i \rightarrow \theta^{i'}$[3]. We shall start from the functional:

$$J[g_{\mu\nu}] = \frac{-1}{16\pi} \int \sqrt{g(\theta)} R(\theta) d^4\theta \qquad (13)$$

where $g = \det g^{\mu\nu}$ and $R(\theta)$ is the curvature scalar. This functional is a scalar and it is

2.6.6.2 Lost causes in theoretical physics

http://www.mth.kcl.ac.uk/~streater/lostcauses.html#VII

Physics from Fisher Information

Cf. B.R. Frieden, Science from Fisher Information, 2nd ed. (Cambridge University Press, 2004)

2.6.6.3 Extension of information geometry for modelling non-statistical systems (2014)

Extension of information geometry for modelling non-statistical systems, PhD Thesis, Antwerpen, 2014 - http://arxiv.org/abs/1501.00853v1

Maximum Entropy Method – this thesis presents a very different approach to modelling: the **data set model formalism**. (The data set model formalism is for a large part a generalisation of the results of information geometry.)

The geometry is derived from general divergence functions which quantify how well data is described by a given element of the model. An example of such a divergence function, found in statistics and statistical physics, is the relative entropy or Kullback-Leibler divergence.

Cf. quote by John Archibald Wheeler:

> *All things physical are information theoretic in origin … and information gives rise to physics*.

Example of classical attempt: Roy Frieden, who tries to base all of physics on an information theoretical foundation.

Quantum information theory: many contributions (to check).

Question asked to the author, Ben Anthonis:

> *Could you please provide me some explanation on this paragraph at page 4:*
>
> *Perhaps the most prominent of these (applications based on data set formalism) is that the parametrised family of models is no longer required to be a subset of the data.*
>
> *+ following: It is even allowed for models to be mathematical objects qualitatively different from those data. This is to be contrasted with both the maximum entropy method and with information geometry. For this reason, a possibly less obvious but still rather promising field for applications is that of machine learning.*

2.6.6.3.1 Chapter 2 – Elementary differential geometry

(…) good summary

2.6.6.3.2 Chapter 3 – Information geometry

Tangent spaces:

The intuitive choice is to take as the basis vectors of the tangent space TθM not the partial derivative operators but rather the derivatives of a particular function of the probability density functions. The simplest choice is

$$\partial_i p_\theta(x) = \frac{\partial}{\partial \theta^i} p_\theta(x).$$

However, for reasons of convenience the practitioners of information geometry often choose a different representation of the tangent vectors. They choose as basis vectors the derivatives of a power-like function of the density function, called the α-representation of the tangent space. These objects are given by the expressions, for some α ∈ R,

$$\partial_i \ell_\alpha(x) = \begin{cases} \frac{2}{1-\alpha} \partial_i \left(p_\theta(x)^{\frac{1-\alpha}{2}} - 1 \right), & \alpha \neq 1, \\ \partial_i \ln p_\theta(x), & \alpha = 1. \end{cases}$$

The previously mentioned basis vectors (intuitive) are obtained as a special case by choosing α = −1. The different representations of basis vectors may seem unconventional at first sight. However, they are the conventional basis vectors to tangent spaces of other manifolds, those consisting of the stochastic variables x → ℓα(x).

Where the most natural representation of the tangent spaces corresponds to the case α = −1, another very convenient representation is found when α = 1 is chosen. Not only does this representation have a clear relation to the log likelihood θ → ln pθ(x), the tangent vectors in the 1-representation all have vanishing expectation value, as

$$\mathbb{E}_\theta[\partial_i \ln p_\theta] = \int_X p_\theta(x) \frac{\partial}{\partial \theta^i} \ln p_\theta(x) \mathrm{d}x = \frac{\partial}{\partial \theta^i} \int_X p_\theta(x) \mathrm{d}x = 0.$$

The α=1-representation is called the “exponential representation” due to its relation with the logarithm and the α=−1-representation is referred to as the “mixture representation”.

Riemannian metric:

3.3 The Riemannian metric

In order to make the topological manifold $\mathbb{M}$ into a Riemannian manifold, a choice must be made for the inner product of the tangent spaces. Information geometry almost exclusively makes use of the Fisher information metric [38] for this purpose. This means the inner product of two tangent vectors $\vec{V}$ and $\vec{W}$ is defined to be equal to the covariance (in the statistical sense) of the stochastic variables V and W which correspond to the vectors, that is

$$g(\vec{V},\vec{W})|_\theta = \mathbb{E}_\theta[VW] - \mathbb{E}_\theta[V]\mathbb{E}_\theta[W]. \tag{3.2}$$

The most commonly used expression for the components of this tensor are found in the exponential representation, where they are given by

$$\begin{aligned} g_{ij}(\theta) &= \mathbb{E}_\theta[(\partial_i \ln p_\theta)(\partial_j \ln p_\theta)] \\ &= \int_X p_\theta(x) \left(\frac{\partial}{\partial\theta^i} \ln p_\theta(x)\right) \left(\frac{\partial}{\partial\theta^j} \ln p_\theta(x)\right) \mathrm{d}x. \end{aligned} \tag{3.3}$$

The importance of the Fisher information to statistics, and to the problem of parameter estimation in particular, is contained in the Cramér-Rao theorem.

in particular, is contained in the Cramér-Rao theorem [39]. Unbiased estimators for a parameter θ^k are stochastic variables $\hat{\theta}^k$ for which $\mathbb{E}_\theta[\hat{\theta}^k] = \theta^k$. The Cramér-Rao theorem states that after making N observations of these estimators $\hat{\theta}^k$ in a population distributed according to p_θ, it holds that

$$\left\{\mathbb{E}_\theta[\hat{\theta}^i\hat{\theta}^j] - \mathbb{E}_\theta[\hat{\theta}^i]\mathbb{E}_\theta[\hat{\theta}^j]\right\} - \frac{1}{N} g^{ij}(\theta) \geqslant 0,$$

where the inequality means that the expression on the left hand side represents the components of a positive definite matrix. The Fisher information matrix thus expresses the minimal variance these estimators may attain. Furthermore the theorem shows there exists an estimator which attains this lower bound on its covariance as the number of observations tends to infinity. Such an estimator is said to be maximally efficient.

A property of the Fisher information matrix is that it also appears in the second (and thereby lowest) order term of the expansion of the Kullback-Leibler divergence or relative entropy:

$$D(p||p_\theta) = \int_X p(x) \ln\left(\frac{p(x)}{p_\theta(x)}\right) \mathrm{d}x,$$

The Fisher information thus expresses also the infinitesimal distance between nearby points on a manifold of probability distributions as expressed by the Kullback-Leibler divergence. This validates Rao's choice to endow statistical manifolds with the Fisher information matrix as their Riemannian metric.

Affine connections:

Another important differential geometrical quantity is the affine connection with which a manifold can be endowed. The first attempt at investigating this structure for statistical manifolds was made by Rao. He studied the metric connection derived from the Fisher information metric and computed geodesic distances [39]. However, a statistical interpretation of this metric connection was not

immediately obvious [29]. A breakthrough in the study of affine connections on statistical manifolds came in the 1970s with the work of Chentsov, Efron and Amari.

Chentsov showed that the space of multinomial distributions admits only a single statistically invariant Riemannian metric—the Fisher information metric—and a unique family of statistically invariant affine connections. The statistical invariance means that the geometric quantities defined through the metric tensor and the affine connections remain unchanged when the underlying probability space is mapped into another one through a Markov process.

Efron (1975) studies one-parameter families of distributions, seen as curves through the space of all distributions over some measurable set. He implicitly defines a connection this space such that one-parameter exponential families coincide with geodesics, which are curves for which the tangent vector field is covariant constant along the curve.

The relation of Efron's work with affine connections is that (non-curved) one-dimensional exponential families coincide with geodesics through the larger space in which they are embedded.

NB: the curvature used by Efron to define statistical curvature is the extrinsic or embedding curvature.

These affine connections are usually presented in a relatively technical way but they have very simple interpretations.

subject. Remember that affine connections serve to define covariant derivatives and thereby parallel transport. In the exponential representation, the elements of a tangent space $\mathrm{T}_\theta\mathbb{M}$ are stochastic variables V such that

$$\mathbb{E}_\theta[V] = \int_X p_\theta(x)V(x)\mathrm{d}x = 0.$$

However, there is no guarantee that the expectation value of $V \in \mathrm{T}_\theta\mathbb{M}$ will also vanish when evaluated at another distribution, that is $\mathbb{E}_\xi[V] = 0$ cannot be guaranteed when $\xi \neq \theta$. Consequently, parallel transport need not preserve the vanishing of the above expectation value. The family of affine connections of information geometry solves this problem by defining parallel transport in such a way that the expectation value of a stochastic variable remains zero when it undergoes parallel transport. In particular, the exponential connection ($\alpha = 1$) defines the operation $\Pi^{(1)} : \mathrm{T}_\theta\mathbb{M} \to \mathrm{T}_\xi\mathbb{M}$ by

explicitly subtracting the expectation value in the end point:

$$\Pi^{(1)}V = V - \mathbb{E}_\xi[V].$$

The mixture connection ($\alpha = -1$) on the other hand multiplies the statistic with the appropriate Radon-Nykodym derivative (see for instance [46] or an introductory work on the matter), that is $\Pi^{(-1)} : \mathrm{T}_\theta\mathbb{M} \to \mathrm{T}_\xi\mathbb{M}$ works as

$$\Pi^{(-1)}V = V\frac{p_\theta}{p_\xi}.$$

This means the expectation value equals

$$\begin{aligned}\mathbb{E}_\xi[\Pi^{(-1)}V] &= \int_X V(x)\frac{p_\theta(x)}{p_\xi(x)}p_\xi(x)\mathrm{d}x \\ &= \int_X V(x)p_\theta(x)\mathrm{d}x \\ &= \mathbb{E}_\theta[V] \\ &= 0.\end{aligned}$$

The other affine connections of the α-family are simply linear combinations of the exponential and mixture connections, in particular

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}.$$

Even though the exponential and mixture connections give rise to a path-independent definition of parallel transport and are thus flat, the rest of the α-family has a constant but non-vanishing scalar curvature (also known as the Ricci scalar of the connection), which is given by [29]

$$R \overset{\text{def.}}{=} g^{ij}\Omega^k{}_{ikj} = \frac{1-\alpha^2}{4}.$$

The metric connection associated with the Fisher information metric is obtained in the case $\alpha = 0$ and so it is the most (positively) curved member of this family. As is elucidated in the previous chapter, the metric connection

This property only holds when $\alpha = 0$ and in general it can be shown that for any $\alpha \in \mathbb{R}$

$$\vec{X}(g(\vec{Y},\vec{Z})) = g(\nabla^{(\alpha)}\vec{X},\vec{Y}) + g(\vec{X},\nabla^{(-\alpha)}\vec{Y}), \tag{3.7}$$

where g represents the Fisher information metric. This shows that the α- and $-\alpha$-connections are dual with respect to this metric. Dual connections play an important role in the more advanced topics of information geometry. One application is found in a generalisation of the Pythagorean property, which holds in a triangle where one leg is a geodesic for the exponential connection and the other leg is a geodesic for the mixture connection [30].

Divergence or contrast functions:

They serve as the elementary structure from which all geometric objects can be constructed, just as all the Riemannian geometry of statistical manifolds can be derived from the divergence function of Kullback and Leibler.

The Kullback-Leibler distance is a function quantifying in a certain sense the difference between statistical distributions over a measurable set X.

A commonly used set of contrast functions are the f-divergences of Csiszár [55]. They are of the form

$$D_f(q||p) = \int_X p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x,$$

where f is a convex function for which $f(1) = 0$. The Csiszár f-divergences are the largest class of statistically invariant divergence functions. Another often made choice is the set of Bregman divergences [56]. The most general definition of these divergences is not limited to statistical distributions but when restricted thereto, it is possible to write them in the form

$$D_F(q||p) = \int_X \int_{p(x)}^{q(x)} [F'(u) - F'(p(x))] \mathrm{d}u \mathrm{d}x.$$

where F is a strictly convex function. These are also known as U-divergences, after their definition by Eguchi [57]. It can be shown that the Kullback-Leibler divergence is the only contrast function which belongs to both the classes of Csiszár and Bregman divergences [29].

The first geometric structure to be introduced is, as usual, the metric tensor. Since divergence functions over a manifold $\mathbb{M}$ are zero everywhere on the diagonal of $\mathbb{M} \times \mathbb{M}$ and only there, it must automatically hold that the lowest order term in its Taylor expansion is the second order term, that is

$$D(m_\theta||m_{\theta+\delta\theta}) = \frac{1}{2} \frac{\partial^2}{\partial\xi^i \partial\xi^j} D(m_\theta||m_\xi)\Big|_{\xi=\theta} \delta\theta^i \delta\theta^j + \mathcal{O}((\delta\theta)^3).$$

The coefficient of this lowest order term—without the factor $\frac{1}{2}$—is the metric tensor induced by the divergence:

$$g_{ij}(\theta) = \frac{\partial^2}{\partial\xi^i \partial\xi^j} D(m_\theta||m_\xi)\Big|_{\xi=\theta}. \tag{3.9}$$

This is a positive definite quantity as it is the matrix of second derivatives of a function in a local minimum. Despite this, it behaves properly under a coordinate transformation, as can be seen also from the alternative expression

$$g_{ij}(\theta) = -\frac{\partial^2}{\partial\xi^i \partial\theta^j} D(m_\theta||m_\xi)\Big|_{\xi=\theta},$$

which can be shown to hold since the first derivatives of the divergence vanish identically on the diagonal.

Affine connections can be constructed from divergence functions just as well. In particular, it is possible to consider the pair of mutually dual connections ∇ and ∇^* defined through the expressions [59, 60]

$$g_{ks}(\theta)\omega^s{}_{ij}(\theta) = -\frac{\partial^3}{\partial\theta^i \partial\theta^j \partial\xi^k} D(m_\theta||m_\xi)\Big|_{\xi=\theta}$$

and

$$g_{ks}(\theta)\varpi^s{}_{ij}(\theta) = -\frac{\partial^3}{\partial\xi^i \partial\xi^j \partial\theta^k} D(m_\theta||m_\xi)\Big|_{\xi=\theta}. \tag{3.10}$$

Thermodynamics (Carathéodory):

The first connection between differential geometry and thermodynamics was made by Constantin Carathéodory, who sought to establish an axiomatic basis for thermodynamics. He chose to express

this in terms of differential geometry. In these papers he could phrase thermodynamics on sound mathematical principles, rather than on the more usual references to imaginary devices such as Carnot engines or to concepts such as the flow of heat. For this, he works on the topological manifold of thermodynamic states of the system. His rendering of the Second Law states

> *In every neighbourhood of every equilibrium state x, there are states y that are not accessible from x via quasi-static adiabatic paths.*

This formulation is weaker than Kelvin's better known one, which states that no cyclical process can exist which turns heat into its mechanical equivalent of work. Starting from this axiom, Carathéodory could derive thermodynamics as it was known in his time. His results therefore extend those of Helmholtz, who had already noticed that a definition of temperature or entropy does not require cyclical processes or ideal gasses.

From other source:

> An alternative approach has been used in classical statistical mechanics to analyse the geometry of thermodynamic systems. The starting point is the probability density distribution which includes the partition function of the corresponding system. It can be shown that from the information contained in the partition function a metric structure can be derived, the so-called Fisher-Rao metric which describes the properties of the statistical system. Although their conceptual origin is quite different, it can be shown that Weinhold and Ruppeiner metrics are related to the Fisher-Rao metric by means of Legendre transformations of the corresponding thermodynamic variables. One common disadvantage of all these metrics is that they are not Legendre invariant, leading to the unphysical result that their properties depend on the thermodynamic potential used in their construction.

(…) The metric tensors introduced by Weinhold and Ruppeiner give rise to a measure of distance sometimes known under the name of thermodynamic length.

2.6.6.3.3 Chapter 4 – Data set model formalism

(…)

A statistical model belongs to the exponential family if and only if it exhibits a Hessian structure when it is endowed with the Kullback-Leibler divergence.

A recurring property in the work of Efron, Reeds and Amari is that exponential families of probability distributions are flat when endowed with the exponential connection. Exponential families can be identified by means of their flatness and the correspondence of canonical parameters and affine coordinates.

Probability theory has been deliberately excluded from the building blocks of the formalism. As is explained in the introduction, this was done in order to identify the crucial link between information theory and the geometry of the formalism, rather than relying on the much-walked path of basing information theory on probability theory.

2.7 LQ PhD Thesis

We adopt Schouten's nomenclature for the type of spaces considered [Sch54]. An L_n is a general n-dimensional manifold endowed with a *linear* connection; when the latter is *symmetric* the L_n is called an A_n and is torsion-free. When a metric tensor g_{ab} is defined, the compatibility condition does not hold in general, i.e. $\nabla_c g_{ab} = -Q_{cab} \neq 0$. If $Q_{cab} = 0$, the connection is called *metric* with respect to g_{ab} and the L_n is called a U_n; if in addition the connection is symmetric, one has a V_n, i.e. *Riemann space*. If $Q_{cab} = Q_c\, g_{ab}$, the connection is called *semi-metric*; if in addition it is symmetric, the L_n is called a W_n, i.e. *Weyl space*.

Consider a four-dimensional space-time manifold $\mathcal{M}$ endowed with a Lorentzian metric g_{ab} and assume that the connection ∇_c be the symmetric Levi-Civita connection, i.e. $\nabla_c g_{ab} = 0$; hence $(\mathcal{M}, \mathbf{g}, \boldsymbol{\nabla})$ is a V_4, i.e. a Riemannian space. In the Lagrangian formulation of the theory the Hilbert metric variational principle proceeds with the specification of a Lagrangian density $\mathfrak{L}$, which is assumed to be a functional of the metric and its first and possibly higher derivatives, that is

$$\mathfrak{L} = \mathfrak{L}(\mathbf{g}, \partial\mathbf{g}, \partial^2\mathbf{g}, \dots). \tag{2.1}$$

In addition, one requires that $\mathfrak{L}$ be a scalar density of weight $+1$, i.e. $\mathfrak{L} = \sqrt{-g}\, L$, where g denotes the determinant of the matrix formed with the components of g_{ab} and L is the Lagrangian; this enables one to form the action integral

$$S[\mathbf{g}] = \int_{\mathcal{U}} \mathrm{d}^4\Omega\, \mathfrak{L}, \tag{2.2}$$

where $\mathrm{d}^4\Omega = \mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$, which is taken over a compact region $\mathcal{U}$ of the manifold $\mathcal{M}$. The field equations are obtained by requiring that the action (2.2) be stationary under arbitrary variations such that the metric and its first derivatives be held fixed on the boundary $\partial\mathcal{U}$. This variation defines the functional derivative $\mathfrak{L}_{ab}$ of the Lagrangian density $\mathfrak{L}$, viz.

$$\delta S[\mathbf{g}] = \int_{\mathcal{U}} \mathrm{d}^4\Omega\, \mathfrak{L}_{ab}\, \delta g^{ab}, \qquad \text{with } \mathfrak{L}_{ab} := \frac{\delta \mathfrak{L}}{\delta g^{ab}},$$

also called the *Euler–Lagrange derivative* of $\mathfrak{L}$, and the field equations are

$$\mathfrak{L}_{ab} = 0.$$

As is well known, the variational principle implies very important differential constraints on the field equations, which hold 'off shell', i.e. whether or not the field equations are satisfied; these are the *generalised Bianchi identities*, obtained from Noether's second theorem by taking as a specific class of variations of the metric that induced by diffeomorphisms $f : \mathcal{M} \to \mathcal{M}$. Since the manifolds $(\mathcal{M}, \mathbf{g})$ and $(\mathcal{M}, f^*\mathbf{g})$ are physically equivalent, the action functional does not change under the diffeomorphism f; in particular, it remains unaltered under an infinitesimal coordinate transformation. For such variations, it is not difficult to see that, at the first order of perturbation, δg^{ab} is given in terms of the Lie derivative of the metric with respect to the vector field v^c that generates the diffeomorphism f, that is,[10]

$$\delta g^{ab} = -\mathcal{L}_v g^{ab} = -2\nabla^{(a} v^{b)}.$$

Since by definition $\mathfrak{L}_{ab}$ is a symmetric density of weight $+1$, the variational principle yields

$$\delta S[\mathbf{g}] = -2 \int_{\mathcal{U}} \mathrm{d}^4\Omega \, \mathfrak{L}_{ab} (\nabla^a v^b) \equiv 0$$

for all vector fields v^c that vanish on the boundary. Integrating by parts the last equation and dropping the divergence term one obtains the expected generalised Bianchi identities, namely

$$\nabla^a \mathfrak{L}_{ab} = 0. \tag{2.3}$$

tion for the fields ψ. Furthermore, the generalised Bianchi identities (2.3) take the form of the contracted Bianchi identities, i.e. $\nabla^a G_{ab} = 0$, which in turn entail the covariant conservation of the stress-energy tensor, as a direct consequence of the invariance of the Einstein–Hilbert action under diffeomorphisms.

2.8 Questions

Q1 – Why statistical manifolds are naturally Riemannian? (cf. Fisher information)

- **Log-likelihood functions**:

- Definition:

A useful mapping is the *log-likelihood function* $\ell : \mathcal{S} \to \mathcal{F}(\mathcal{X}, \mathbb{R})$ defined by

$$\ell(p_\xi)(x) = \ln p_\xi(x).$$

Sometimes, for convenience reasons, this will be denoted by $\ell_x(\xi) = \ell(p_\xi(x))$. The derivatives of the log-likelihood function are

$$\partial_j \ell_x(\xi) = \frac{\partial \ln p_\xi(x)}{\partial \xi^j} = \varphi_j(\ell_x(\xi)), \quad 1 \le j \le n.$$

They are also found in literature under the name of *score functions*. Heuristically speaking, they describe how the information contained in p_ξ changes in the direction of ξ^j. The functions $\partial_j \ell_x(\xi)$ will play a central role in defining the Fisher information matrix and the entropy on a statistical manifold.

- **Kullback-Leibler**:
  - We speak here about measure of closeness between true distribution and model via K-L information or divergence or relative entropy (the negative of which is the Boltzmann entropy); there are other notions of distance – is the Riemannian geometry also valid in these cases?
  - K-L definition:

$$D_{KL}(p||q) = E_p\left[\ln \frac{p}{q}\right] = \begin{cases} \displaystyle\sum_{x^i \in \mathcal{X}} p(x^i) \ln \frac{p(x^i)}{q(x^i)}, & \text{if } \mathcal{X} \text{ is discrete;} \\ \displaystyle\int_{\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)}\, dx, & \text{if } \mathcal{X} \text{ is continuous.} \end{cases}$$

  - The K-L relative entropy can be used to find a goodness of fit of these two densities given by the expected value of the extra-information required for coding using q (model) rather than using p (true distribution)
  - the K-L relative entropy does not satisfy all the axioms of a metric on the manifold S
  - The diagonal part of the Hessian of the K-L relative entropy is the Fisher metric
  - Let p, q ∈ S be two points on the statistical model S. The Fisher distance, dist(p, q), represents the information distance between densities p and q. It is defined as the length of the shortest curve on S between p and q, i.e., the length of the geodesic curve joining p and q

Theorem 4.4.5 *Let $d = \mathrm{dist}(p, q)$ denote the Fisher distance between the densities p and q and $D_{KL}(p||q)$ be the Kullback–Leibler relative entropy. Then*

$$D_{KL}(p||q) = \frac{1}{2} d^2(p, q) + o(d^2(p, q)) \qquad (4.4.5)$$

  - The K-L relative entropy induces the linear connection ∇(1).
- **Fisher information**

- Definition:

  A metric structure, similar to the Riemannian structure on a hypersurface, can be introduced on a statistical model, based on the parameter $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{E}$ and the log-likelihood function $\ell(\xi) = \ln p_\xi(x)$. The *Fisher information matrix* is defined by

  $$g_{ij}(\xi) = E_\xi[\partial_i \ell(\xi) \partial_j \ell(\xi)], \quad \forall i, j \in \{1, \dots, n\}, \tag{1.6.16}$$

  which can be written explicitly as

  $$g_{ij}(\xi) = E_\xi[\partial_{\xi^i} \ln p_\xi \cdot \partial_{\xi^j} \ln p_\xi] = \int_{\mathcal{X}} \partial_{\xi^i} \ln p_\xi(x) \cdot \partial_{\xi^j} \ln p_\xi(x) \cdot p_\xi(x)\, dx.$$

- Properties:

  Proposition 1.6.2 *The Fisher information matrix on any statistical model is symmetric, positive definite and non-degenerate (i.e., a Riemannian metric).*

  Hence the Fisher information matrix provides the coefficients of a Riemannian metric on the surface S. This allows us to measure distances, angles and define connections on statistical models.

  Proposition 1.6.3 *The Fisher information matrix can be written as the negative expectation of the Hessian of the log-likelihood function*

  $$g_{ij}(\xi) = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ell(\xi)] = -E_\xi[\partial_{\xi^i} \partial_{\xi^j} \ln p_\xi]. \tag{1.6.18}$$

  Theorem 1.6.4 *The Fisher metric is invariant under reparametrizations of the sample space.*

  Theorem 1.6.5 *The Fisher metric is covariant under reparametrizations of the parameters space.*

  i.e. transforms as a 2-covariant tensor; a metric satisfying the abovementioned properties is unique, and hence equal to the Fisher metric

- K-L information cannot be computed directly; a natural estimator is the log-likelihood.
  - Maximisation of the log-likelihood function (in terms of parameters theta) enables finding a good model among the family of models with parameters theta (**MLE**)
  - Numerical methods: cf. gradient descent for instance

Q2 - Do equations of motion (i.e. Euler-Lagrange) make any sense in Statistical Information Geometry?

- Many articles to check related to Fisher and unified physics
  - Cf. Frieden's scheme which is controversial
-

Q3 – If Q2 is yes, what would be the derivation from a variational principle (purely metric and metric-affine)?

- Cf. article 2.6.6.1.

Q4 – Can we apply other results from Riemannian geometry like Killing vectors/equations for specific symmetries?

Q5 – Hamiltonian formulation? – cf. LQU's thesis

Q6 – Generalised Bianchi identities? – cf. LQU's thesis p. 14

- Noether's second theorem + invariance under diffeomorphisms

Q7 – Why is minimizing the negative log likelihood equivalent to maximum likelihood estimation (MLE)? (https://quantivity.wordpress.com/2011/05/23/why-minimize-negative-log-likelihood/)

3 Languages

3.1 Overview

Python is emerging as a serious alternative to R for data science.

4 Reports on Tools and Vendors

4.1 Gartner et al. Reports

4.1.1 Major Myths About Big Data's Impact on Analytics

http://www.gartner.com/technology/reprints.do?id=1-22ZQWHR&ct=141010&st=sb

Advanced analytics is about problem solving, whereas analytics is about reporting what has happened. These two types of analytics require different technologies and skills.

4.1.1.1 Structured vs. Unstructured data

Tips:

- Start developing expertise for nonrelational databases, such as graph databases and other NoSQL derivatives. Use the best data store paradigm for a given purpose.
- Strive to understand the core of the problem you're tackling and concentrate on the data sources most likely to help you solve that problem.

4.1.1.2 Analytics vs. Advanced Analytics

"Normal" analytics = **descriptive analytics**.

Advanced analytics solves problems using predictive analytics and prescriptive analytics.

Predictive analytics predicts future outcomes and behavior, such as a customer's shopping behavior or a machine's failure.

Prescriptive analytics goes further, suggesting actions to take based on the predictions.

The technologies for advanced analytics are also radically different from those for analytics and require different skills. These skills typically include a solid understanding of *statistics*, *machine learning* and *operations research*.

4.1.1.3 Business user vs. Data scientist

Data science is a multidisciplinary practice. It includes: advanced statistics, machine learning, computer science, operations research (hacking skills?), programming, data management.

I would add non-conventional mindset!

All of these skills must be coupled with practical business and operations experience.

Embedded analytics and solutions implemented by service providers are helping ease the shortage of data science skills. Although business users are the main users of these applications, data scientists build the embedded models. This approach could provide better scaling of data science resources, but it doesn't remove the need for data scientists.

Consider forming a data science team whose members collectively possess most of the skills required for data science.

Data scientists have a strong drive to understand the real world through data and are imaginative in creating surrogates for the real world. They are curious. They have the ability to ask great questions and obtain answers to those questions from data. They have good analytical skills.

Team: data scientists are needed

4.1.1.4 Descriptive vs. Predictive Analytics

Many people believe that descriptive analytics is about the past, whereas predictive analytics is about the future. Reality: All analytics is based on the past and most of it looks to the future.

Tips:

- Know the value of recalibration/model management, especially model performance tracking.
- Understand the relevance of a certain dataset for a given purpose.
- Choose very recent datasets to maximize relevancy, otherwise the machine learning approach may not pick up current patterns

4.1.1.5 Fast vs. Real-Time Analytics

Analytics is not real-time — and not even near-real-time — unless some or all of the input data has been captured in the past few seconds or minutes. Although high-speed analytics on older data can be very useful for some kinds of decisions, it's not in real time. It also doesn't help users who require an understanding of current conditions or advance notification of fast-emerging threats and opportunities.

Tips:

- Ask users when you're gathering requirements for a new system whether they need current information for their decisions or if historical data is sufficient.
- Design analytic solutions that respond in "right time," which may be real time, near real time or not real time at all, depending on the business problem to be addressed.
- Use event-stream processing technologies, such as complex-event processing and distributed stream computing platforms, for very low-latency analytics (submillisecond, subsecond or a few seconds) on large amounts of streaming data. This data may come from a variety of sources including sensors, market data providers, transaction processing systems and websites.
- Use business activity monitoring platforms, spreadsheet tools, data discovery and BI reporting tools to perform near-real-time analytics on current data when the response must occur within minutes. You can also use predictive or prescriptive analytics.

4.1.1.6 Predictive Power / Bias

Not everything can be predicted even with massive data sets.

Example: click-through rate in marketing campaign is very low and noisy – no prediction can be done for a particular individual.

Why?

--- Check Peter Norvig science data models vs. pred. models based on data only

Predictive analytics can produce statistically significant results only when looking at population segments, typically of 1,000 to 5,000 individuals.

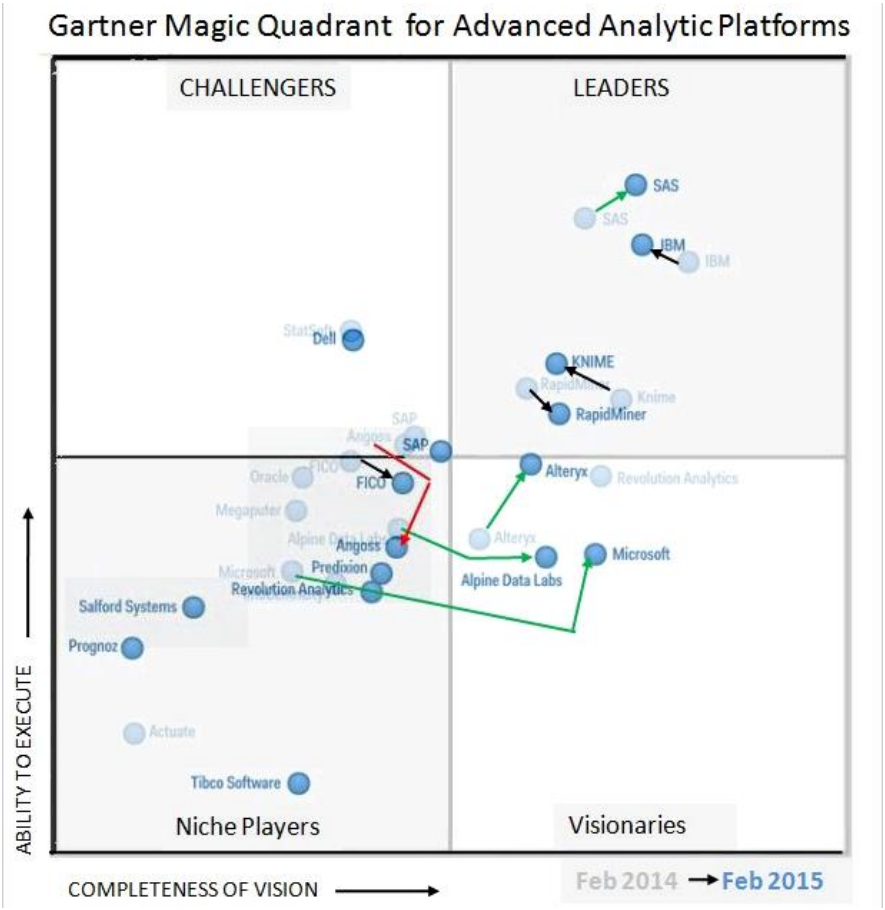
Cautious with data biases, like social media.

Tips:

- Combine the insight collected from social media monitoring with other customer interaction sources to address age bias.
- Understand that the mechanisms that provided the data — whatever its source — always produce some sort of bias.
- Instruct data scientists to find the bias in data and reduce its effect on the resulting analytics

4.1.2 Magic Quadrant for Advanced Analytics Platforms

4.1.2.1 Comparison 2014/2015



Revolution Analytics acquired by Microsoft – R language

4.1.2.2 2015 Report

http://www.gartner.com/technology/reprints.do?id=1-2ABEHA1&ct=150220&st=sb

(From Alteryx)

Check IBM Watson Analytics (Dec 2014)

Check Microsoft Azure Machine Learning (AML) cloud offering

Criteria:

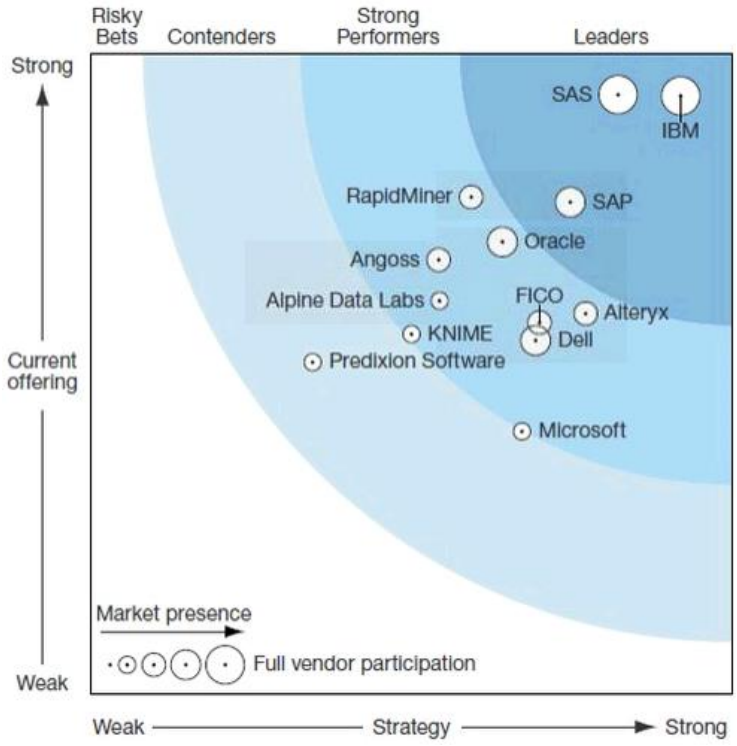
- **Data access, filtering and manipulation**
  - This refers to the product's ability to access and integrate data from disparate sources and types, and to transform and prepare data for modelling
- **Data exploration and visualization**
  - This refers to the product's ability to visually interact with and explore data, and to perform basic descriptive statistics and pattern detection
- **Predictive analytics**
  - The central capability of advanced analytics platforms is facilitation of the synthesis of models that predict future behavior or estimate unknown outcomes using a range of techniques, such as regression, decision trees and ensemble models. It also includes an ability to compare and validate models for selection.
- **Forecasting**
  - This is a specific type of prediction using time series or econometric methods to predict the value of a variable or outcome at a specified time — for example, sales in the next quarter or the number of calls that a call center will receive next week.
- **Optimization**
  - This refers to a type of prescriptive analytics that uses a mathematical algorithm to choose the "best" alternative(s) that meet specified objectives and constraints.
- **Simulation**
  - This is a predictive analytics approach that involves building a model of a system or process to experiment with, or study how, it works with a focus on understanding the range of possible outcomes.
- **Delivery, integration and deployment**
  - This refers to the ease and speed with which the user can move models from a development environment to a deployment environment.
- **Platform and project management**
  - This refers to the platform's ability to manage and monitor models in production and to keep track of important data and issues relating to model deployment.
- **Performance and scalability**
  - This refers to the time required to load data, to create and validate models, and to deploy them in a business.
- **User experience**
  - This refers to the usability, UI, skill level required to use the platform, and the support provided to users via documentation, guidance and the user community.

4.1.3 Magic Quadrant for Business Intelligence and Analytics Platforms

http://www.gartner.com/technology/reprints.do?id=1-2ACLP1P&ct=150220&st=sb

4.1.4 Forrester Wave: Big Data Predictive Analytics Solutions, Q2 2015

4.1.4.1 Overview



4.1.4.2 Report Q2 2015

In 2015 the leaders are:

- IBM which assembles an impressive set of capabilities, with predictive analytics at the center
- SAS, continues to be an analytics powerhouse, and evolving with SAS Visual Analytics and interfaces with open-source tools R, Python, and Hadoop
- SAP, provides a comprehensive predictive analytics tools for both business users and data scientists, powered by SAP Hana. KXEN acquisition also brings new capabilities

The Strong Performers (listed in Forrester order) are RapidMiner, Alteryx, Oracle, FICO, Dell, Angoss, Alpine Data Labs, and KNIME. Forrester says that all of them have a sweet spot that makes them good choices. Interestingly, the report says that "With better strategy scores, Alteryx, Angoss, FICO, Oracle, and RapidMiner, would have been Leaders".

Contenders are Microsoft (with Azure Machine Learning and Revolution Analytics) and Predixion Software.

4.1.4.3 Comparison Q2 2015 vs Q1 2013.

Only 5 companies appeared in both Forrester reports: IBM, SAS, SAP, Oracle, Angoss.

Three showed gains in 2015 Wave ranking:

- IBM, got slightly ahead of SAS in the Leaders area
- SAP, improved its position with SAP HANA success and KXEN acquisition
- Angoss, moved from Contender to Strong Performer

SAS remained in the leader position (although overtaken by IBM). Oracle also kept its position.

Over half (7 out of 13) firms are new in 2015: Alpine Data Labs, Alteryx, Dell (replacing StatSoft), FICO, KNIME, Predixion Software, RapidMiner.

Five firms were in 2013 report but absent in 2015:

- KXEN (bought by SAP)
- Revolution Analytics (acquired by Microsoft),
- Salford Systems
- StatSoft (bought by Dell, which replaced it in 2015 Wave)
- TIBCO Software

4.1.4.4 Comparison Gartner Magic Quadrant and Forrester 2015

Finally, comparing Gartner 2015 Magic Quadrant and Forrester Wave 2015 Big Data Predictive Analytics Solutions, we note:

- SAS and IBM are leaders in both (although in different order)
- Alteryx, Alpine Data Labs, Dell, Microsoft, RapidMiner, KNIME, SAP, appear in strong position in both rankings.
- Predixion Software, FICO are also ranked by both companies.

Gartner omits Oracle, while Forrester omits TIBCO software, Salford Systems, and Prognoz.

4.2 Platfora

Why not in the reports (Gartner et al.)?

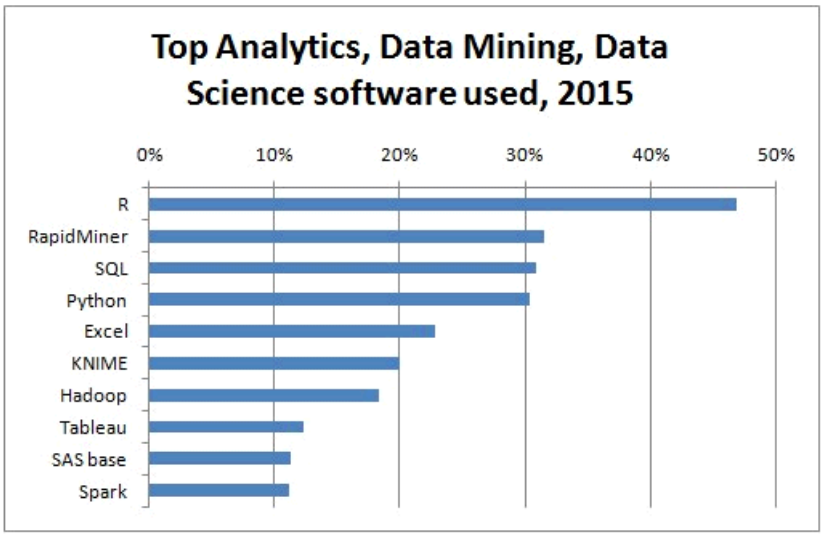
4.3 KDnuggets

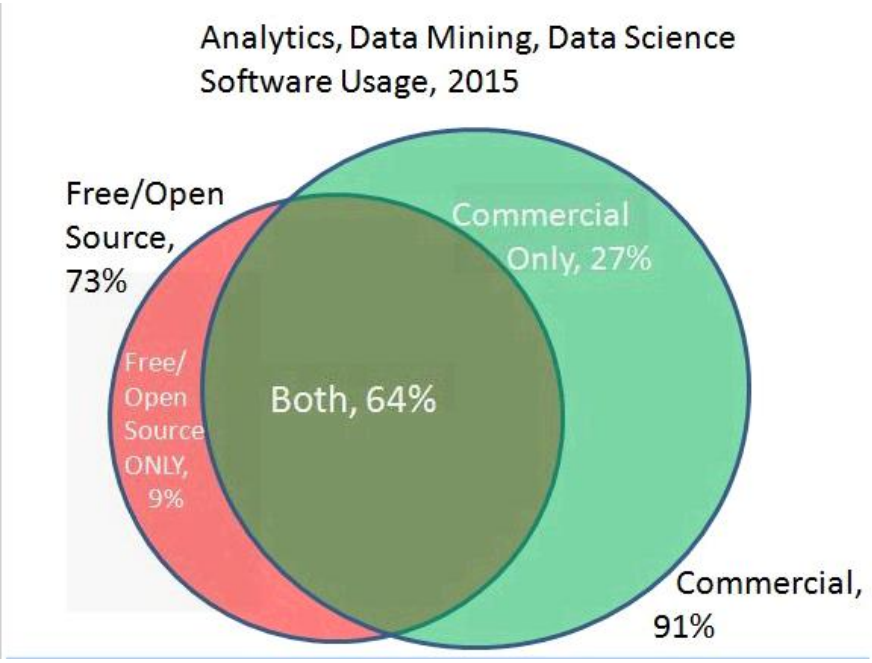
4.3.1 2015 Poll

http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Top Analytics Tools and Trends

Here are the top 10 tools by share of usage:





Among tools with at least 10 votes, the highest increase in 2014 was for

1. H2O (0xdata), 1210% up, to 2.0% share (55 votes) from 0.2% in 2014
2. Actian, 345% up, to 2.0% (56 votes), from 0.5% in 2014
3. Spark, 326% up, to 11.3% (311), from 2.6% in 2014
4. MLlib, 228% up, to 3.3% (91), from 1.0% in 2014
5. Alteryx, 79% up, to 5.6% (155), from 3.1% in 2014
6. Python, 56% up, to 30.3% (837), from 19.5% in 2014
7. TIBCO Spotfire, 56% up, to 4.3% (119), from 2.8% in 2014
8. Pig, 54% up, to 5.4% (150), from 3.5% in 2014
9. SAS Enterprise Miner, 53% up, to 10.9% (302), from 7.2% in 2014
10. Splunk/Hunk, 49% up, to 1.1% (30), from 0.7% in 2014

Tools that showed at least 20% increases in their share for 2 years in the row are Alteryx, Hadoop, KNIME, Python, Qlikview, SAS Enterprise Miner, Tableau, and TIBCO Spotfire.

New analytics tools that received at least 20 votes in 2015 were

- scikit-learn, 8.3% (229)
- Microsoft Azure ML, 3.7% (102)
- Microsoft Power BI, 3.6% (98)
- IBM Watson Analytics, 2.1% (57)
- Ayasdi, 2.0% (56)
- Dataiku, 2.0% (56)
- Lexalytics, 1.3% (35)
- Vowpal Wabbit, 1.3% (35)
- Microstrategy, 0.9% (24)
- Amazon Machine Learning, 0.7% (20)

Among tools with at least 20 votes in 2014, the largest decline in 2015 was for these tools, which includes probably a combination of decline of popularity for free tools like Orange and lack of a voter drive for some of commercial tools this year.

- Predixion Software, 90% down (0.4% share), from 3.7% in 2014
- BayesiaLab, 86% down, to 0.6%, from 4.1%
- Alpine Data Labs, 82% down, to 0.5% from 2.7%
- Oracle Data Miner, 64% down, to 0.8% from 2.2%
- RapidInsight/Veera, 60% down, to 0.2% from 0.5%
- Revolution Analytics (now part of Microsoft), 57% down, to 4.0% from 9.1%
- SAP (including former KXEN), 57% down, to 3.0% from 6.8%
- Orange, 44% down to 1.9% from 3.4%
- Gnu Octave, 41% down, to 2.3% from 3.9%

Hadoop/Big Data Tools

Hadoop/Big Data tool usage jumped to 29% among voters, up from 17% in 2014, and 14% in 2013.

This is probably due to availability and low-cost of many cloud-based Big Data tools. Very notable is the jump in Spark share to 11.3%.

However, most data analysis is still done on "medium" and small data.

Top Hadoop/Big Data tools were

- Hadoop, 18.4% share (507 votes)
- Spark, 11.3% (311)
- Hive, 10.2% (282)
- SQL on Hadoop tools, 7.2% (198)
- Pig, 5.4% (150)
- HBase, 4.6% (127)
- Other Hadoop/HDFS-based tools, 4.5% (125)
- MLlib, 3.3% (91)
- Mahout, 2.8% (76)
- Datameer, 0.8% (23)

Deep Learning Tools

New this year was a category of Deep Learning Tools, with most popular tools being:

- Pylearn2 (55 users)
- Theano (50)
- Caffe (29)
- Cuda-convnet (17)
- Deeplearning4j (12)
- Torch (27)

However, this category is growing rapidly and above list is incomplete, since the largest count in this category was for **other Deep Learning tools (106)**

Programming Languages

Python increased significantly in popularity. Java is the second most commonly used language for analytics/data mining tasks. Here is the

- Python, 30.3% share (837 votes), up from 19.5%
- Java, 14.2% (392), na in 2014
- C/C++, 9.4% (260), na in 2014
- Unix shell/awk/gawk, 8.0% (221), up from 5.8%
- Other programming languages, 5.1% (140)
- Scala, 3.5% (96), na in 2014
- Perl, 2.9% (79), down from 3.0
- Ruby, 1.2% (33), na in 2014
- Julia, 1.1% (31), up from 0.8%
- F#, 0.7% (18), up from 0.5%
- Clojure, 0.5% (13), same as 0.5%
- Lisp, 0.4% (10), up from 0.3%

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
R (1293), 3.6% alone	46.9% 38.5% 37.4%
RapidMiner (870), 13.7% alone	31.5% 44.2% 39.2%
SQL (853), 0% alone	30.9% 25.3% na
Python (837), 0% alone	30.3% 19.5% 13.3%
Excel (631), 0% alone	22.9% 25.8% 28.0%
KNIME (553), 6.7% alone	20% 15.0% 5.9%
Hadoop (507), 0% alone	18.4% 12.7% 9.3%
Tableau (341), 0% alone	12.4% 9.1% 6.3%
SAS base (313), 0.6% alone	11.3% 10.9% 10.7%
Spark (311), 0% alone	11.3% 2.6% na
TIBCO Spotfire (119), 11.8% alone	4.3% 2.8% 1.4%
Microsoft Azure ML (102), 1.0% alone	3.7% na na
Microsoft Power BI (98), 0% alone	3.6% na na
MLlib (91), 0% alone	3.3% 1.0% na
JMP (86), 0% alone	3.1% 3.8% 4.1%
SAP (including former KXEN) (82), 26.8% alone	3.0% 6.8% 1.4%

4.4 Datanami

4.4.1 Hadoop, Triple Stores, and the Semantic Data Lake (May 2015)

Graph databases

http://www.datanami.com/2015/05/26/hadoop-triple-stores-and-the-semantic-data-lake/

4.5 Hadoop

http://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/

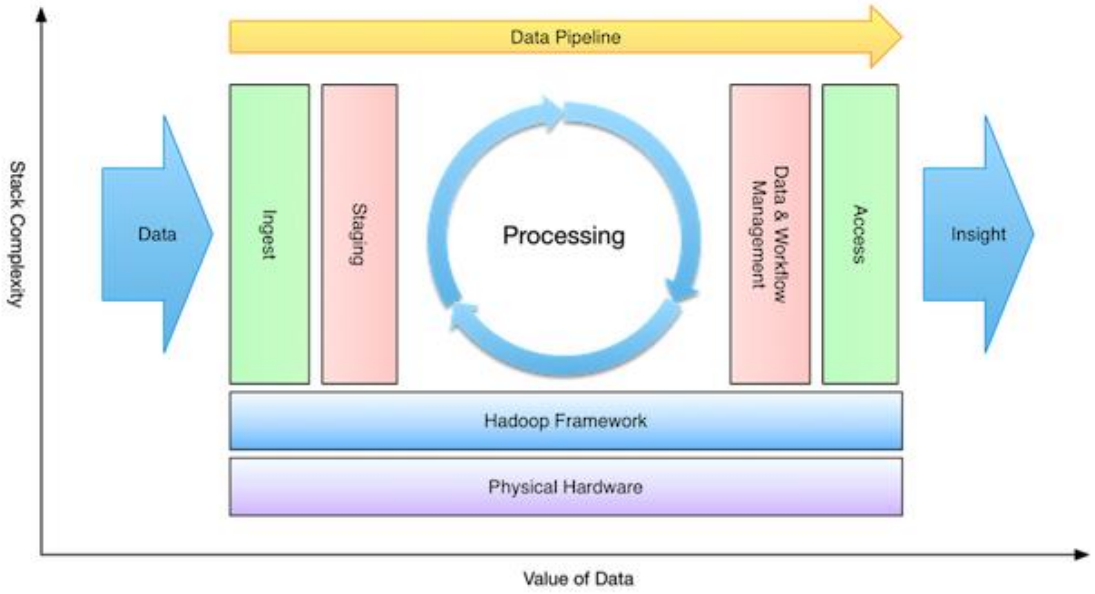
4.6 Big Data

4.6.1 Commodity hardware?

http://www.infoworld.com/article/2610477/big-data/big-data-demands-more-than-commodity-hardware.html

http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/



4.6.2 Hadoop

http://hadoop.apache.org/

4.6.2.1 MapReduce and Yarn

Yarn == MapReduce 2.0

4.6.2.2 Mahout

ML

4.6.2.3 Hive

DWH

The Apache Hive ™ data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

4.6.2.4 Cassandra

The Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data. Cassandra's support for replicating across multiple datacenters is best-in-class, providing lower latency for your users and the peace of mind of knowing that you can survive regional outages.

Cassandra's data model offers the convenience of column indexes with the performance of log-structured updates, strong support for denormalization and materialized views, and powerful built-in caching.

4.6.2.5 Pig

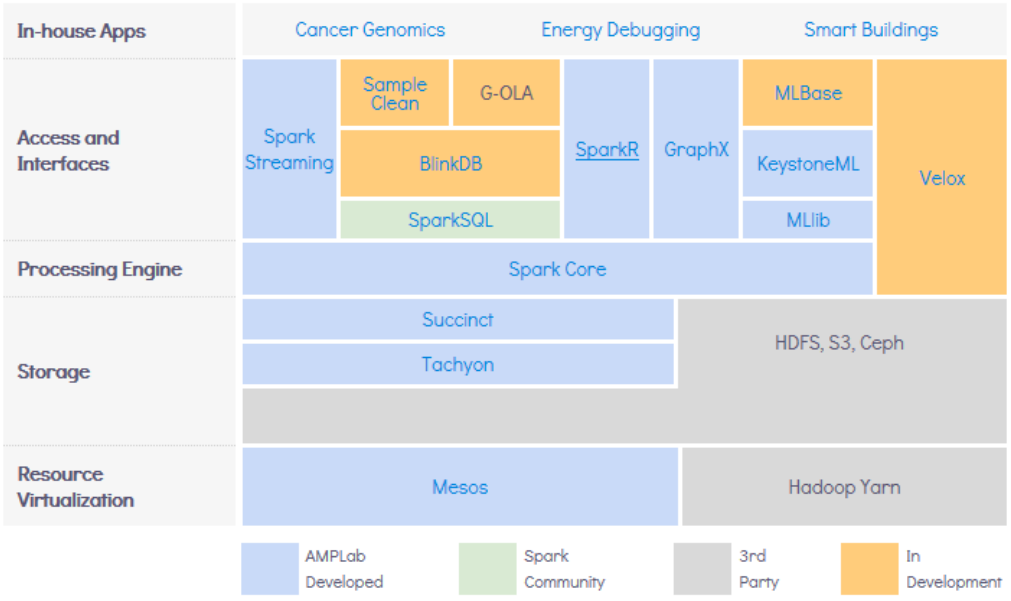
Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

- Ease of programming. It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
- Optimization opportunities. The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
- Extensibility. Users can create their own functions to do special-purpose processing.

4.6.3 AmpLab

https://amplab.cs.berkeley.edu/software/



4.6.3.1 Spark

http://www.eweek.com/database/slideshows/apache-spark-is-creating-a-buzz-in-analytics-data-processing-camps.html

Spark accelerates analytics on Hadoop, working as a full suite of complementary tools, including a fully featured machine learning library (MLlib), a graph processing engine (GraphX) and stream processing. Spark can access data in a variety of sources, including HDFS, Cassandra and HBase.

http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/spark.html

http://blog.cloudera.com/blog/2014/08/how-to-use-ipython-notebook-with-apache-spark/

R frontend for Spark: http://amplab-extras.github.io/SparkR-pkg/

4.6.3.2 Tachyon

http://tachyon-project.org/

5 Andish book

p.27

6 Miscellanea

6.1 Cancer, bad luck, and a pair of paradoxes

https://egtheory.wordpress.com/2015/04/04/cancer-bad-luck-and-a-pair-of-paradoxes/#more-7806

https://en.wikipedia.org/wiki/Simpson%27s_paradox

http://vudlab.com/simpsons/

6.2 Smart people – Poor decisions

http://www.information-management.com/blogs/Business-Analytics-Intelligent-Executives-10026902-1.html?utm_campaign=blogs-may%206%202015&utm_medium=email&utm_source=newsletter&ET=informationmgmt%3Ae4326637%3A3899579a%3A&st=email

Michael J. Mauboussin, chief investment strategist at Legg Mason Capital Management. In an article he wrote in The Futurist (March-April, 2010) he said, “Smart people make poor decisions because the mental software that we humans inherited from our ancestors isn’t designed to cope with the complexity of modern day problems and systems. In short, smart people, like everyone else, face two major obstacles to making good decisions. The first obstacle is the brain, which evolved over millions of years to make decisions unlike what we face in modern life. The second obstacle is the growing complexity of the world in which we live.”

Others have also written about this. In the book Thinking, Fast and Slow by Dan Kahneman, recipient of the Nobel Prize in Economic Sciences for his seminal work in psychology that challenged the rational model of judgment and decision making, Kahneman explains the two systems that drive the way we think. System 1 is fast, intuitive, and emotional. System 2 is slower, more deliberative, and more logical. System 1 is largely unconscious and it makes snap judgments based upon our memory of similar events and our emotions. System 2 is painfully slow, and is the process by which we consciously check facts and think carefully and rationally.

A problem Kahneman points out is that System 2 thinking (slow) is easily distracted and hard to engage and that System 1 thinking (fast) is wrong as often as it is right. System 1 thinking (fast) is easily swayed by our emotions. As an example, he describes an observation that people buy more cans of soup in a grocery store when there is a sign on the display that says "Limit 12 per customer." People miss the opportunity to analyze.

As I have previously written, in the past the best leaders and executives had the best answers. That is not true today. Now the best leaders and executives have the best questions! They can no longer rely on their past experiences or intuition that got them promoted to their C-suite roles. They need to create a culture for analytics including skills and competencies in their work force to be analytical.

http://www.scientificamerican.com/article/patternicity-finding-meaningful-patterns/

6.3 Data Scientists Automated and Unemployed by 2025?

http://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html

When will most expert-level Predictive Analytics/Data Science tasks - currently done by human Data Scientists - be automated: [255 voters]	
Now (it already happened) (13)	5.1%
in 1-2 years (10)	3.9%
in 2-5 years (35)	14%
in 5-10 years (72)	28%
in 10-20 years (42)	16%
in 20-50 years (20)	7.8%
it will take more than 50 years (16)	6.3%
never (48)	18.8%

Katharina Morik, expert level analysis:

The answer to the question depends on the definition of "expert-level analysis". More and more subtasks have been made available for automatic processing. However, the expert-level is exactly above. Hence, expert-level analysis becomes more and more demanding until it covers all science and philosophy. At this ultimate point, I doubt robots or other machines can operate without any expert-level above. But. I admit, this is then mere guessing.

6.4 How to become a data scientist and get hired?

http://www.kdnuggets.com/2015/05/datafloq-become-data-scientist-get-hired.html

6.5 Prediction vs explanation

http://www.information-management.com/blogs/Causal-Modeling-Data-Science-10026923-1.html?utm_campaign=daily-may%208%202015&utm_medium=email&utm_source=newsletter&ET=informationmgmt%3Ae4338269%3A3899579a%3A&st=email

This “prediction vs explanation” divide is often encountered in the statistical world, with commerce generally more focused on pure prediction, while research/academia is obsessed with explanation. And with model building, there's a trade-off between minimizing explanation “bias” and prediction “variance”. “In explanatory modeling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision.

7 Exploratory Analysis

7.1 Types of graphs

7.1.1 Boxplot

http://en.wikipedia.org/wiki/Box_plot

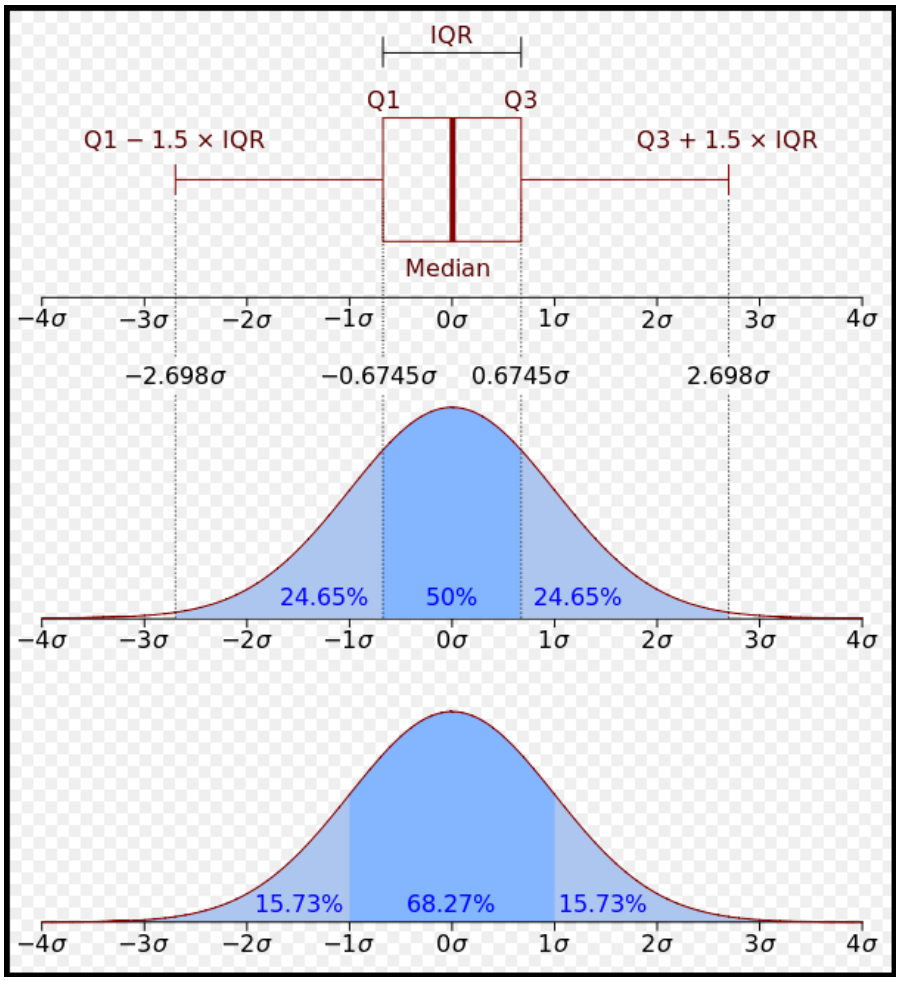
In descriptive statistics, a box plot or boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-

and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points. This is also called a "box and whisker plot".

Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers. In addition to the points themselves, they allow one to visually estimate various L-estimators, notably the interquartile range, midhinge, range, mid-range, and trimean. Boxplots can be drawn either horizontally or vertically.

The box plot is a quick way of examining one or more sets of data graphically. Box plots may seem more primitive than a histogram or kernel density estimate but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data (see Figure 1 for an example). Choice of number and width of bins techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of a kernel density estimate.

As looking at a statistical distribution is more intuitive than looking at a box plot, comparing the box plot against the probability density function (theoretical histogram) for a normal N(0,1σ2) distribution may be a useful tool for understanding the box plot (Figure 5).



8 Fraud/Anomaly Detection

8.1 Tangent Works

Jan 14.05.2015:

> This problem is typically approached by a good (nonlinear) classificator combined with over-sampling or under-sampling of the data because the problem is very often ill-balanced. Here ill-balanced means that you have many "OK" cases in the data but very few "KO" cases. If you train directly on the data you get a meaningless model which will be overfitted to OK cases.
>
> There is another approach where you train model on what is normal (OK cases) and all departures from normality are considered a fraud. In this case you typically provide no KO cases when training the model.
>
> I believe "The elements of statistical learning" provide some hints on a good classifiers (e.g. bagging or random forests or kernel machines).

8.2 S

Large rate of false positives => analytics identify possible cases to be investigated more deeply (with business knowledge) => uncovering previously hidden patterns

http://www.datanami.com/2014/01/30/eight_ways_analytics_powers_fraud_detection/

8.2.1 Analytical Techniques for Fraud Detection

Getting started requires an understanding of:

- The areas in which fraud can occur
- What fraudulent activity would look like in the data
- What data sources are required to test for indicators of fraud

The following analytical techniques are effective in detecting fraud:

- Calculation of statistical parameters (e.g., averages, standard deviations, high/low values) – to identify outliers that could indicate fraud.
- Classification – to find patterns amongst data elements.
- Stratification of numbers – to identify unusual (i.e., excessively high or low) entries.
- Digital analysis using Benford's Law – to identify unexpected occurrences of digits in naturally occurring data sets.
- Joining different diverse sources – to identify matching values (such as names, addresses, and account numbers) where they shouldn't exist.
- Duplicate testing – to identify duplicate transactions such as payments, claims, or expense report items.
- Gap testing – to identify missing values in sequential data where there should be none.
- Summing of numeric values – to identify control totals that may have been falsified.

- Validating entry dates – to identify suspicious or inappropriate times for postings or data entry.

Please note that random sampling is not listed as an effective fraud detection technique. While sampling is an effective data analysis technique for analyzing data values that are consistent throughout the data population, the very nature of fraud is different as it tends not to occur randomly.

9 Events

9.1 Brussels Data Science Meetups

9.1.1 Data Science in the Banking World – 20.05.2015 – VUB