The Information Geometry of Hierarchical Bayesian Models

Dottorato di Ricerca in Informatica XIX Ciclo
Proposta di tesi di dottorato
Cheng Dong Seon

1. Introduction

There is a certain peculiarity in the way complex systems as diverse as living organisms, biological ecologies, meteorological events, human societies and the human brain develop and regulate themselves in an effective manner (or suddenly collapse without much of a warning).

During the last century we have come, again and again, in different areas of scientific research, against the problem of analyzing (or building) such complex systems.

Cybernetics [Wie48] has been one of the first to describe the problem in a principled way, collecting the contributions of time-honored disciplines like engineering and statistics and of promising fields like information theory, computer science, game theory and robotics [Ash56].

At the turning of the century we have obtained a great deal of experience in complex systems because they have been found almost everywhere. However, even with computers of overwhelming power, we are still unable to grasp the necessary knowledge. We may have to admit that accurate knowledge could in principle not be attainable. If we are to take into account the uncertainty of our knowledge, then statistics is the methodology of choice.

Highly Structured Stochastic Systems [GHR03b] is a recent strategy for building complex statistical models that relies on the construction of hierarchical Bayesian models, whose logical structure is best represented through graphs. Hierarchical models can exhibit great complexity globally, yet have relatively simple local structure, thanks to the introduction of conditional independence statements between the variables of the system. This strategy has been very successful and widely adopted in fields like signal processing, computer vision, bioinformatics, epidemiology, and it has promoted fertile cross-disciplinary work in stochastic systems [GHR03a].

There is currently an ongoing effort towards a unifying view of models and algorithms in all these fields in such a way to make it possible to compare, order and relate all these approaches—a long-time guiding principle of cybernetics. After all, we are talking about statistical machinery and machines are the objects of cybernetics' studies.

Concurrently and parallel to the development of complex statistical models, information theory has provided a new point of view that could explain in a unified framework the properties of these models and of the algorithms that make inferences on them. Information geometry [CT84, AN00] is a recent discipline that studies the geometrical properties of the manifolds associated with families of probability distributions. In other words, given a parameterized statistical model that we want to optimally "fill-in", such that it well represents the observed data, the set of all possible solutions forms a high-dimensional manifold that can and must be studied to understand the power and the limits of our current methods.

Information geometry borrows definitions and theorems from differential geometry and it may sound too abstract and distant from the problems to be solved. However, it has already given explanations for some known models and algorithms [Ama95, Ama01] and it could potentially be a unified framework for studying the intrinsic properties of more statistical machinery.

In this work I want to investigate theoretical and practical issues related to the use of the viewpoint of information geometry in complex statistical models. The underlying vision is the one that last century cybernetics set out to realize, that is to study the behaviour of complex systems, especially the brain [Ash60]. The machinery that I want to study is composed of the latest and best in stochastic inference. I would like to explore various applications areas, going from traditional computer vision to human behaviour modelling to brain-inspired cognitive systems.

The rest of the paper is organized as follows: section 2 briefly introduces the main ideas of cybernetics; section 3 discusses Highly Structured Stochastic Systems; section 4 covers some applications in selected research areas; section 5 broaches the topic of information geometry; in section 6, I present the proposal for my Ph.D. thesis, addressing some problems that I would like to sort out.

2. Cybernetics

Cybernetics was defined by Wiener as “the science of control and communication, in the animal and the machine”—in a word, as the art of *steermanship* [Wie48]. The cybernetician studies how biological, mechanical and abstract machines behave during their activities and the common properties or fundamental constraints of their behaviour, in so far as they are regular, or determinate, or reproducible. What cybernetics offers is the framework on which all individual machines may be ordered, related and understood.

A particular virtue of cybernetics is that it offers a method for the scientific treatment of systems in which complexity is outstanding and too important to be ignored. There are systems so dynamic and interconnected that the alteration of one factor immediately acts as cause to evoke alterations in others, perhaps in a great many others. The way of cybernetics has been that of first marking out what is achievable and then providing generalised strategies, of demonstrable value, that can be used uniformly in a variety of special cases.

The term “system” is nowadays a very abused word and the seemingly endless list of meanings that it can take (51 entries for the Merriam-Webster dictionary) shadows its etimology, which can be roughly translated as “combination”. It is always indeed a combination of parts, an organization of people or things, a network of relationships.

I find illuminating this definition, reported in the glossary section of [HJT05]: (1) a set of variables selected by an observer [Ash60] together with the constraints across variables he either discovers, hypothesises or prefers. Inasmuch as the variables of a system may represent the components of a complex machine, an organism or a social institution and a constraint is the logical complement of a relation, an equivalent definition of system is that (2) it represents a set of components together with the relations connecting them to form a whole unity. Unlike in general systems theory, in cybernetics, a system is an observer’s construct. If it describes, simulates or predicts a portion of his environments it may be regarded as a model of that portion. The model and the modelled “world” share the same organization but because of their different material realizations they are likely to differ in structure. Cybernetics starts with investigating all possible systems and then inquires why certain systems are not materially realized, or it asks why certain conceivable behaviors are not followed. Systems neither exist independent of an observer nor imply a purpose [Kri86].

Since a system is a construct, it is not unusual for an observer to feel the need to *redefine* his model according to the effectiveness of its descriptions and predictions.

Indeed, some of the discoveries regarding missing variables have been of major scientific importance, as when Newton discovered the importance of momentum, or when Gowland Hopkins discovered the importance of vitamins.

Most natural systems, though, are very complex and the sources of these complications generate more interesting systems that are in turn more difficult to analyze. They usually have a large number of parameters and the whole set of them is by no means obvious. For example, in biological systems it is, in fact, co-extensive with the set of “all variables whose change directly affects the organism”, including the conditions in which an organism lives.

What is important is that complex systems, richly cross-connected internally, have complex behaviours, and that these behaviours can be goal-seeking in complex patterns. Variables in a system may be more or less coupled between each other. When one variable directly depends on the value of another one we can say that the latter has an immediate effect on the first one. When this occurs, a diagram of immediate effects can often usefully be drawn showing these relations. Those effects that work through longer chains of variables and with longer delays will be called ultimate effects with a relative diagram of ultimate effects. If a variable or part has no ultimate effect on another, then they are said to be independent. And when all the parts are independent the whole is said to be reducible.

In cybernetics one may be confronted with very large systems. A system’s “largeness” refers to the number of *distinctions* made: either to the number of states available or, if its states are vectors, to the number of its variables or its degree of freedom. The two measures are correlated, for if other things are equal, the addition of extra variables will make possible extra states. A system may also be made larger if, the number of variables being fixed, each is measured more precisely, so as to make it show more distinguishable states.

In a metaphorical way, a “very large” system is such that some definite observer with definite resources and techniques finds it in some practical way too large for him; so that he cannot observe completely, or control it completely, or carry out the calculations for prediction completeley. In other words, he says the system “very large” if in some way it beats *him* by its richness and complexity.

Merely specifying such systems can be difficult, because by definition it cannot be observed completely. But this is synonymous with saying that it must be specified “statistically”, for statistics is the art of saying things that refer only to some aspect or portion of the whole, the whole truth being too bulky for direct use.

3. Highly Structured Stochastic Systems

The term **Highly Structured Stochastic Systems** (HSSS) refers to a recent strategy for building statistical models, for computing with them, and for interpreting the resulting inferences, in appreciation of the need to answer the challenge of real-world complex problems [GHR03a].

HSSS have found applications in areas as diverse as signal processing, image processing, bioinformatics, epidemiology, error-control coding and language processing (see [GHR03a] for an extensive bibliography). By emphasizing common ideas and structures, such as graphical, hierarchical, and spatial models, and tecniques, such as Markov chain Monte Carlo methods and the EM algorithm and its variants, it has promoted cross-disciplinary work in stochastic systems.

Complexity is handled by working up from simple local assumptions in a coherent way, and that is the key to modelling, computation, inference, and interpretation. The winning strategy has been that of building **hierarchical Bayesian models**. The word 'hierarchical' refers to the presence of additional structure—levels of latent (i.e. unobserved) quantities with logically distinct and scientifically interpretable functions—over the standard 'flat' statistical models that deal only with the observed data. To note that this is different from representing levels as continuously-varying degrees of refinement.

Many individual subject areas within HSSS have developed their own isolated methods, software and terms: *population methods* in pharmacokinetics, *multilevel models* in education and geography, *latent variable models* in psychology and econometrics, *random effects models* in biostatistics, *frailty models* in survival analysis, econometrics and reliability, *Bayesian networks* in artificial intelligence, *pedigrees* in genetics, and so on.

Central in these models is the role played by **conditional independence** as a precise description of the information conveyed by one part of the system about others. By analogy with the assumptions in temporal Markov processes, 'Markov properties' are statements about conditional independence in general statistical models [Lau96]. Hierarchical models can thus exhibit great complexity globally, yet have relatively simple local structure. The key is conditional independence, whereby each variable is related locally (conditionally) to only a few other variables.

Graphical modelling (see section 3.1) provides a diagrammatic representation of the logical structure of a joint probability distribution, in the form of a network or graph depicting the local relations among variables. Associated with the graph are various Markov properties that specify how the graph encodes conditional independence assumptions [Lau96].

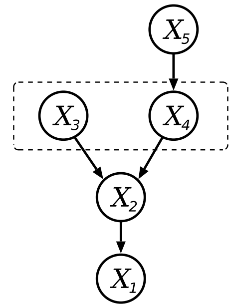


Figure 1: A graphical model with five random variables. The nodes in the dashed region represent the parents of node X_2, i.e. $\pi_2 = \{3, 4\}$ and $X_{\pi_2} = \{X_3, X_4\}$.

Statistical inference, though, only benefits from the full power of the graphical formalism when it is set wholly within a probability model, that is in the **Bayesian paradigm**. In the Bayesian framework, all unknown quantities, whether parameters, missing observations, or latent variables, are treated symmetrically as random variables. Their prior distributions, together with the likelihoods of observed variables, define a full joint probability model, where inference about the unknowns is based on their posterior distributions by application of Bayes' rule. This unified treatment of variables allows coherent integration and propagation of all sources of uncertainty. It gives a powerful framework in which prior scientific knowledge can be expressed, and in which inferences are made using probability calculation and decision theory, fully exploiting the graphical structure of the model.

3.1. Graphical models

A graphical model is a family of probability distributions defined in terms of a directed or undirected graph. Although the early work dealt with undirected graphs [DLS80], recent applications of graphical models have predominantly been based on *directed graphical models*, also known as recursive models [WL83] or *Bayesian networks*, a term coined by Pearl [Pea86].

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed acyclic graph, where $\mathcal{V}$ are the nodes and $\mathcal{E}$ are the edges of the graph [Jor04]. Let $X = \{X_v\}_{v \in \mathcal{V}}$ be a collection of random variables indexed by the nodes of the graph. Let π_v denote the subset of indices of the parents of node v and X_{π_v} the vector of variables indexed by the parents of v (see figure 1).

Given a collection of kernels $\{k(x_v|x_{\pi_v})\}_{v \in \mathcal{V}}$ that sum (in the discrete case) or integrate (in the continuous case) to one (with respect to x_v), we define a joint probability distribution (a probability mass function or probability density

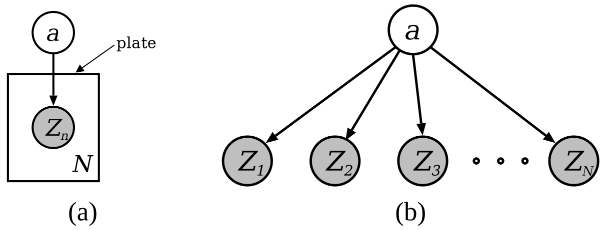


Figure 2: The diagram in (a) contains a plate, to compact the same graphical model in (b). This model asserts that the variables Z_n are conditionally independent and identically distributed given a, and can be viewed as a graphical model representation of the De Finetti theorem. Note that shading, here and elsewhere, denotes conditioning.

as appropriate) as follows:

$$p(x) = \prod_{v \in \mathcal{V}} k(x_v | x_{\pi_v}) \tag{1}$$

It is easy to verify that $k(x_v|x_{\pi_v}) = p(x_v|x_{\pi_v})$, the conditional probabilities of each variable given its parents in the graph.

A *plate* is a useful device for capturing replication in graphical models, including the factorials and nested structures that occur in experimental designs (see figure 2).

Directed graphical models are familiar as representations of hierarchical Bayesian models (see figure 3). The graph provides an appealing visual representation of a joint probability distribution, but it also provides a great deal more. First, whatever the functional forms of the kernels, the factorization in (1) implies a set of conditional independence statements among the variables X_v that can be obtained from a polynomial time *reachability algorithm* based on the graph [Pea88]. Second, the graphical structure can be exploited by algorithms for probabilistic inference (see section 3.2).

Let us now consider the undirected case. Given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a collection of random variables $X = \{X_v\}_{v \in \mathcal{V}}$ we denote by $\mathcal{C}$ a collection of *cliques* of the graph (i.e. fully connected subsets of nodes). Associated with each clique $C \in \mathcal{C}$, let $\psi_C(x_C)$ denote a nonnegative *potential function*. We define the joint probability by taking the product over these potential functions and normalizing:

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{2}$$

where Z is a normalization factor, obtained by summing or integrating the product wrt x.

Undirected graphs (see figure 4) are often used in problems in areas such as spatial statistics, statistical natural language processing and communication networks—problems

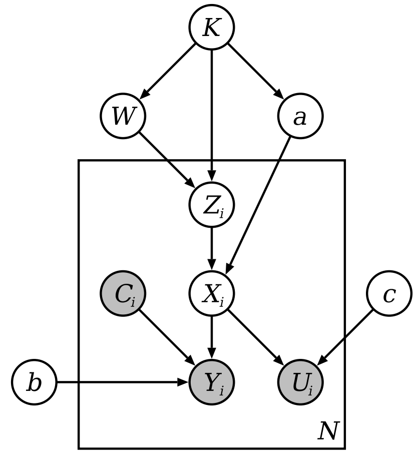


Figure 3: This is the "error-in-covariates" logistic regression model of [RLJG02]. The core of this model is a logistic regression of Y_i on X_i. The covariate X_i is not observed (in general), but noisy measurements U_i of X_i are available, as are additional observed covariates C_i. The density model for X_i is taken to be a mixture model, where K is the number of components, W are the mixing proportions, Z_i are the allocations and a parameterizes the mixture components.

in which there is little causal structure to guide the construction of a directed graph.

It is also possible to work with hybrids that include both directed and undirected edges [Lau96]. In general, directed graphs and undirected graphs make different assertions of conditional independence. Thus, there are families of probability distributions that are captured by a directed graph and are not captured by any undirected graph, and viceversa [Pea88].

The representations (1) and (2) can be overly coarse for some purposes. In particular, in the undirected formalism the cliques C may be quite large, and it is often useful to

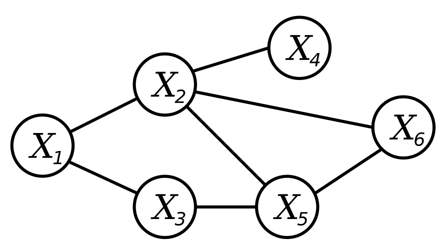


Figure 4: An example of an undirected graphical model. The joint probability distribution can be factorized as $\frac{1}{Z}\psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)\psi(x_3, x_5)\psi(x_2, x_5, x_6)$.

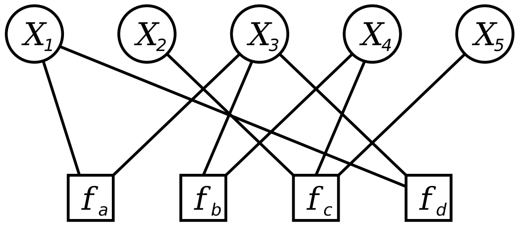


Figure 5: An example of an factor graph. The joint probability distribution can be factorized as $\frac{1}{Z} f_a(x_1, x_3) f_b(x_3, x_4) f_c(x_2, x_4, x_5) f_d(x_1, x_3)$.

consider potential functions that are themselves factorized, in ways that need not be equated with conditional independencies. Thus, in general, we consider a set of 'factors', $\{f_i(x_{C_i})\}_{i \in \mathcal{I}}$, for some index set $\mathcal{I}$, where C_i is the subset of nodes associated with the ith factor (the same subset can be repeated multiple times, i.e. $C_i = C_j$ for $i \neq j$). The joint probability is then:

$$p(x) = \frac{1}{Z} \prod_{i \in \mathcal{I}} f_i(x_{C_i}) \tag{3}$$

As shown in figure 5, this definition is associated with a graphical representation—the *factor graph* [KFL01]. A factor graph is a bipartite graph in which the random variables are round nodes and the factors appear as square nodes. There is an edge between the factor node f_i and the variable node X_v if and only if $v \in C_i$. Factor graphs provide a more fine-grained representation of the factors making up a joint probability, and are useful in defining inference algorithms that exploit this structure. Note also that the factor $f_i(x_{C_i})$ can often be written as $\exp(\theta_i \chi_i(x_{C_i}))$, for a parameter θ_i and a fixed function χ_i, in which case the representation in (3) is nothing but the canonical form of the exponential family. Thus factor graphs are widely used as graph-theoretic representations of exponential family distributions.

3.2. Learning

"Learning" in the context of parameterized statistical models amounts to adjusting the parameters such that the resulting specific model optimally "explains" the observed data. When this data is complete, given by a set of N indipendent and identically distributed samples $\mathcal{X} = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$, and the chosen model gives the joint probability distribution $p(x|\theta)$, where θ is a vector of parameters, the resulting overall density for the samples is

$$p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p\left(x^{(i)}|\theta\right) = \mathcal{L}(\theta|\mathcal{X}) \tag{4}$$

This function $\mathcal{L}(\theta|\mathcal{X})$ is called the likelihood of the parameters given the data, or just the likelihood function. The problem of finding the θ that maximizes $\mathcal{L}$ is called *maximum-likelihood* (ML) estimation. That is, the ML estimate is given by $\hat{\theta}$ where

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathcal{X}) \tag{5}$$

Often $\log(\mathcal{L}(\theta|\mathcal{X}))$ is maximized because it is analytically easier and equivalent, due to the convexity of the $\log$ function.

For many problems, indeed for almost all the hierarchical Bayesian models, it is not possible to find analytical expressions, and we must resort to more elaborate tecniques. Another major concern is that, given the largeness of most interesting models, the growth in computational complexity of some algorithms may well exceed the available computing power.

In the following I will describe in detail two general methods: the Expectation-Maximization algorithm and some of its variants, and the Markov chain Monte Carlo approach.

3.2.1. Expectation Maximization

The Expectation-Maximization (EM) [DLR77, Wu83, RW84, Bil97] algorithm is a general method of finding the ML estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values. There are two main applications of the EM algorithm. The first occurs when the data indeed has missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but can be simplified by assuming the existence of values for additional but missing (or hidden) parameters.

As in the previous section, we assume that data $\mathcal{X}$ is observed and is generated by some distribution. We call $\mathcal{X}$ the *incomplete data*. We assume that a complete data set exists $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and also assume (or specify) a joint probability density:

$$p(z|\theta) = p(x, y|\theta) = p(y|x, \theta)p(x|\theta) \tag{6}$$

With this new density function, we can define a new likelihood function, $\mathcal{L}(\theta|\mathcal{Z}) = \mathcal{L}(\theta|\mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}|\theta)$, called the **complete-data likelihood**. Note that this function is in fact a random variable since the missing information $\mathcal{Y}$ is unknown, random, and presumably governed by an underlying distribution.

The EM algorithm first finds the expected value of the complete-data log-likelihood $\log(p(\mathcal{X}, \mathcal{Y}|\theta^{(i)})$ with respect to the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current parameter estimates $\theta^{(i)}$. That is, in the so-called E-step we calculate this auxiliary function

$$Q(\theta, \theta^{(i)}) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\theta) \,\middle|\, \mathcal{X}, \theta^{(i)}\right] \tag{7}$$

The second step, the M-step, is to maximize this auxiliary function with respect to the free parameters θ:

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)}) \tag{8}$$

These two steps are repeated as necessary. Under general conditions each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function [DLR77, Wu83]. In most cases the E-step formula may be simplified by exploiting the conditional independence between variables derived from the graphical formalism. The integral (or summation) then nicely splits into factors that can be calculated indipendently. The EM algorithm has been successfully applied to mixtures of Gaussians, hidden Markov models [Rab89] and other more elaborate models [FJ03].

Some variants of the EM algorithm, mostly named "generalized" EM (GEM), modify the E- or M-step to obtain computable approximate solutions when exact EM cannot be applied or perform poorly. One general strategy is to relax the M-step so that we only need to find some $\theta^{(i+1)}$ that increases the value of the auxiliary function instead of maximizing it. In this case the convergence is still guaranteed. "Variational" EM approaches, instead, modify the E-step by calculating an auxiliary distribution for each unobserved variable in $\mathcal{Y}$, according to a simplified version of the graphical model—one in which some, or all in the so-called *mean field* approach [JGJ99], edges have been removed. This makes computations easier and faster and results in overall good performances.

3.2.2. Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a general method for sampling and computing expectations in multivariate stochastic processes. MCMC methods are iterative stochastic algorithms converging weakly to the distribution of interest and have their roots in the Metropolis algorithm [MU49, MRR$^+$53], an attempt by physicists to compute complex integrals by expressing them as expectations for some distribution and then estimate this expectation by drawing samples from that distribution.

In the original Monte Carlo approach, if the integrand function $h(x)$ can be decomposed into the product of a function $f(x)$ and a probability density $p(x)$, then the integral can be approximated by drawing a large number $x^{(1)}, \ldots, x^{(k)}$ of samples from $p(x)$ in this way:

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_p[f(x)] \simeq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) \tag{9}$$

One problem with applying Monte Carlo integration is in obtaining samples from some complex probability distribution. Attempts to solve this problem are the roots of MCMC methods. The Metropolis-Hastings [MU49, MRR$^+$53, Has70] algorithm is one such method and its basic strategy is to draw the next sample based on the value of the previous one, effectively generating a Markov chain. Hence the name of these methods. Following a sufficient *burn-in period*, the chain approaches its stationary distribution and the last set of samples is considered coming from $p(x)$.

In the early 1990's [GS90] it was realized that one particular MCMC method, the Gibbs sampler [CG92], is very widely applicable to a broad class of Bayesian problems, even if it has its origins in image processing [GG84]. The key to the Gibbs sampler is that one only considers *univariate* conditional distributions—the distribution when all of the random variables but one are assigned fixed values. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms. Let's assume we have defined n univariate conditional densities:

$$\begin{array}{l} p(x_1|x_2,\ldots,x_n) \\ p(x_2|x_1,x_3,\ldots,x_n) \\ \ldots \\ p(x_n|x_1,\ldots,x_{n-1}) \end{array} \tag{10}$$

If we choose $n-1$ arbitrary initial values $x_2^{(0)}, \ldots, x_n^{(0)}$ we can *scan* all those densities to draw new samples:

$$\begin{array}{lll} x_1^{(1)} & \text{by a draw from} & p(x_1|x_2^{(0)},\ldots,x_n^{(0)}) \\ x_2^{(1)} & \text{by a draw from} & p(x_2|x_1^{(1)},x_3^{(0)},\ldots,x_n^{(0)}) \\ & \ldots & \\ x_n^{(1)} & \text{by a draw from} & p(x_n|x_1^{(1)},\ldots,x_{n-1}^{(1)}) \end{array} \tag{11}$$

Each pass of this process generates a sampling of all the variables and under general conditions the distribution of the samples converges to the true joint distribution. Thus, we may take the last k samples and average them to approximate any desired marginal expectation.

When the underlying model is a graphical model, the equations (10) are readily inferred from the Markov properties of the graphical formalism. Conditioning on the so-called *Markov blanket* of a given node renders the node independent of all other variables. In directed graphical models, the Markov blanket is the set of parents, children and co-parents of a given node ("co-parents" are nodes which have a child in common with the node). In the undirected case, the Markov blanket is simply the set of neighbours of a given node. Using these definitions, Gibbs samplers can be set up automatically from the graphical model specification [GTS94].

Finally, the Gibbs sampler can be thought of as a stochastic analog to the EM algorithm. In a full Bayesian framework (see section 3) the parameters are treated as random variables with some given prior probabilities. This enables us to approximate the densities of both hidden variables and parameters through sampling of their posterior distribution. Hence, in the sampler, the E- and M-steps are replaced by some appropriate sampling procedure [RC99]. Indeed Gibbs sampling is more powerful because it can often be applied in situations where the EM algorithm cannot.

3.3. Other Topics

Statistical problems where "the number of things you don't know is one of the things you don't know" [Gre03] are ubiquitous in statistical modelling, both in traditional modelling situations such as variable selection in regression, and in more novel methodologies such as object recognition, signal processing, and Bayesian models. This "meta-problem" is mostly called the problem of *model selection* and it has been confronted many times, with equally many approaches being proposed [Aka74, Sch78].

The most direct way to relate different models is to determine a scoring function that takes into account both the goodness of fitting and the complexity of the model, usually given by the degrees of freedom involved. The most prominent such score in a Bayesian framework is the *Bayesian information criteria* (BIC). The applicability of this model selection strategy to complex graphical models is still under debate [GHKM01].

Graphical modelling, in representing graphically the relations of the given variables, cannot remove from the mind of the observer the idea that it has captured the *causal* structure of the problem at hand [Daw03]. Causality is a challenging topic for anyone to consider in a formal way [Pea00, Daw03, Arj03]. Already the concept itself is problematic, and often people have sharply different opinions of its foundations. But causality is perhaps a particularly challenging topic for a statistician. Most textbooks on elementary statistics first give the warning that causality and correlation are not the same, and then they go on to discuss only correlation.

Causal modelling is indeed possible and fruitful [Daw03], but there is a growing consensus towards considering causal models like any other model, mental constructs by means of which we attempt to interpret, relate, and explain empirical observations, both past and future. The empirical adequacy or inadequacy of a model is to be judged by how well it performs at predicting observable quantities, within its own ambit.

4. Applications

In this section I review some applications in research areas that have greatly benefited from the use of graphical models. This selection is only meant to be a cursory panorama of those systems that either I have encountered during my studies or I wish to explore in the future.

For instance, in computer vision some well-known tools such as Markov Random Fields [GG84], hidden Markov models [Rab89], and the Kalman filter have been discovered to be special cases of graphical models. One interesting recent result is the *flexible model* strategy proposed in [JF02]. In order to build more robust vision algorithms, flexible models try to capture different aspects of the data at the same time, be it changes of appearance, movements and occlusions, geometric changes, changes of illumination. They focus on a 2.5D vision paradigm, i.e. on a model of the world where the objects live in various parallel planar layers that have a depth ordering. The authors devise small graphical models that are reusable as components for bigger models and propose an variational inference scheme that decomposes the reasoning task. In [CCMP04], my collegues and I explored a naive composition of two other graphical models [FJ03, TB02].

A research area that is living an extraordinary development is that of neurosciences. Neurophysiological, neuroanatomical, and brain imaging studies have helped to shed light on how the brain transforms raw sensory information into a form that is useful for goal-directed behaviour. A fundamental question that is seldom addressed by these studies, however, is *why* the brain uses the types of representations it does and what evolutionary advantage, if any, these representations confer. It is difficult to address such questions directly via animal experiments. A promising alternative is to investigate computational models based on efficient coding principles [ROL02]. Indeed, the probabilistic approach has been applied to a wide range of problems, from the formulation of theories of brain function to neural modelling.

The last area I want to talk about is that of cognitive science, in a broad sense that envelopes "soft" disciplines like psychology and sociology. Most theories within these realms suffer from uncertainties in the information structure of the problems being studied. Either there are too many variables, some of them unaccounted for, or there are hidden variables that cannot be observed. Bayesian probabilistic models are the ultimate instruments in integrating effectively every known, latent and presumed source of information and this power is much needed in those contexts where reliability has been kind of a roulette [Gly01]. Causality here is the main concern, but that's a scorching topic that I have already covered in the previous section.

5. Information Geometry

In the previous sections we have seen that most stochastic models of interest are represented by a parameterized family of probability distributions. In the following material, the attention will focus primarily on the *exponential family* and *mixture family* of probability distributions. This choice is by no means restrictive. Many of the most used parametric distributions are members of the first family (see table 1) and the second family allows greater flexibility than a single distribution. For instance, by the mixture of Gaussians we can approach any probability distribution in total variation norm.

Definition 1 (Exponential family) *The probability densities of an exponential family $S = \{p(x,\theta) | \theta \in \mathbb{R}^n\}$ have the form*

$$p(x,\theta) = \exp\left\{\theta^T \mathbf{T}(x) - A(\theta)\right\} \tag{12}$$

where $\mathbf{T}(x)$ are functions that represent sufficient statistics and $A(\theta)$ is a normalizing constant, also known as the cumulant generating function*.*

Definition 2 (Mixture family) *The probability densities of mixture distributions $\{p(x,\eta) | \eta \in \mathbb{R}^l\}$ have the form*

$$p(x,\eta) = \sum_{i=1}^{l} \eta_i p_i(x) \tag{13}$$

where $p_1(x), \ldots, p_l(x)$ are some given distributions, and the weights $\eta_1, \ldots, \eta_l$ are such that $\sum_{i=1}^{l} \eta_i = 1$.

Distribution	$\theta^T \mathbf{T}(x)$	$A(\theta)$
Bernoulli	$\theta_1 x$	$\log\left(1 + e^{\theta_1}\right)$
Gaussian	$\theta_1 x + \theta_2 x^2$	$\frac{1}{2}\left(\theta_1 + \log\frac{2\pi e}{-\theta_2}\right)$
Exponential	$-\theta_1 x$	$-\log\theta_1$
Poisson	$\theta_1 x$	e^{θ_1}

Table 1: Examples of probability distributions that are members of the exponential family

Another property that makes the exponential family invaluable is that it is always guaranteed that a conjugate prior exists. This is enormously useful in Bayesian statistics because the posterior distribution is of the same type of the prior distribution when this is conjugate with the likelihood.

Let's take a closer look on S. Every individual member of this family is precisely characterized by a vector $\bar{\theta}$ of parameters. In other words, *a point $\bar{\theta}$ in the n-dimensional space whose coordinates are given by θ represents some given probability distribution*.

Looking from this point of view, the behaviour of learning algorithms may be simply stated as the search for the optimal point in the n-dimensional manifold S, which in general is not Euclidean. This means that two points—two probability distributions—cannot be related through the Euclidean metric and usual Euclidean geometry does not hold.

Information geometry [CT84, AN00] is the discipline that studies the intrinsic properties of the geometry of the manifolds associated to probability distributions. In this setting, when we select a probabilistic model for a given problem, we are actually identifying a submanifold M of S. On the other hand, the observed data suggest some distribution (e.g. one given by the empirical distribution) in S or, when the observation or specification is incomplete, define many candidate points in S that form a submanifold D. The optimal distribution may be characterized as the point in M that minimizes the distance between the realizable M and the observed D and, at the same time, the point in D that minimizes the distance gives the estimated data complementing the partial observed data. This approach was proposed in [CT84] and further developed in [AN00, ITA04, Ama01].

In the following sections I will review the basic notions that make possible an information geometrical perspective of stochastic inference; the Riemannian metric (section 5.1) allows us to talk of orthogonality in a Riemannian manifold; dual geometry (section 5.2) explores the differential geometrical properties of manifolds; the generalized Pythagoras theorem (section 5.3) is the main result with which we characterize trajectories in the manifold; and finally, statistical inference is explained (section 5.4).

5.1. Riemannian Metric

Given two distributions, p and q, defined on the same space, the Kullback-Leibler divergence (also known as *relative entropy*) is defined by

$$D(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{14}$$

The divergence is not symmetric, that is, $D(p\|q) \neq D(q\|p)$ in general, but $D(p\|q) \geq 0$ with equality when and only when $p = q$. When we are considering two similar distributions, $p(x,\theta)$ and $p(x,\theta')$, we may abuse the notation and write

$$D(\theta\|\theta') = \int p(x,\theta) \log \frac{p(x,\theta)}{p(x,\theta')} dx \tag{15}$$

In fact, when the two distributions are infinitesimally close, $\theta' = \theta + d\theta$, the divergence between the two nearby distributions is expanded as

$$D(\theta\|\theta + d\theta) = \frac{1}{2} d\theta^T G(\theta) d\theta \tag{16}$$

where $G(\theta)$ is the Fisher information matrix defined by

$$G(\theta) = E\left[\frac{\partial \log p(x,\theta)}{\partial\theta}\frac{\partial \log p(x,\theta)}{\partial\theta}^T\right] \quad (17)$$

where E denotes expectation wrt $p(x,\theta)$.

The squared distance dm^2 between two nearby distributions is thus given by

$$\begin{aligned} dm^2 &= d\theta^T G(\theta) d\theta & (18)\\ &= 2D(\theta \| \theta + d\theta) & (19)\end{aligned}$$

When $G(\theta)$ is non-degenerate, i.e. is defined at every point θ such that (18) holds, the manifold M is said to be a Riemannian manifold, with $G(\theta)$ as the Riemannian metric tensor.

5.2. Dual Geometry

Let's go back to the distributions p and q. There are two special curves connecting them in the manifold M. When we linearly connect $\log p(x)$ and $\log q(x)$ we have the exponential family $\{p(x,t)|t \in [0,1]\}$ given by

$$\log p(x,t) = (1-t)\log p(x) + t\log q(x) - c(t) \quad (20)$$

or equivalently

$$p(x,t) = \exp\{tr(x) + \log p(x) - c(t)\} \quad (21)$$

where $r(x) = \log\frac{q(x)}{p(x)}$ is a new random variable, $c(t)$ is the normalization factor, and t is the parameter of the curve. This curve is regarded as a "straight line" (i.e. a *geodesic*) connecting $p(x)$ and $q(x)$ in S from the exponential family point of view. In terms of the θ-coordinate system, the coordinates $\theta(t)$ of $p(x,t)$ are written as

$$\theta(t) = (1-t)\theta_p + t\theta_q = \theta_p + t(\theta_q - \theta_p) \quad (22)$$

where θ_p and θ_q are the θ-coordinates of $p(x)$ and $q(x)$, respectively. Therefore, the exponential geodesic is a linear curve in the θ-coordinates. The other way is the mixture family $\{p^*(x,t)|t \in [0,1]\}$ connecting the two distributions by the curve

$$p^*(x,t) = (1-t)p(x) + tq(x) \quad (23)$$

This curve is regarded as a geodesic from the mixture point of view. Both points of view have their own proper meaning. It is easy to show that the η-coordinates of $p^*(x,t)$ are written as

$$\eta^*(t) = \eta_p + t(\eta_q - \eta_p) \quad (24)$$

Hence, the mixture geodesic is a linear curve in the η-coordinates. Each geodesic from the exponential family viewpoint is called an *exponential geodesic* or e-geodesic and those from the mixture viewpoint are called *mixture geodesics* or m-geodesics.

By generalizing this idea, we can see that in fact each coordinate curve $\theta_i = t, \theta_j = k_j (j \neq i)$, where only one coordinate varies and the others are fixed, is an e-geodesic. In this case, the coordinate system θ is called *affine* and the manifold is said to be e-flat. Obviously, the exponential family manifold M is e-flat. On the other hand, a manifold is m-flat when its η coordinate system is affine. The mixture family belongs to this category.

But here is a surprisingly result [Ama01].

Theorem 1 *A manifold is e-flat when and only when it is m-flat and viceversa.*

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. A dually flat manifold has rich differential geometrical structures. In fact, the η-coordinates of an exponential family are given by

$$\eta_i = E[T_i(x)] = \frac{\partial}{\partial\theta_i}A(\theta) \quad (25)$$

which are known as the expectation parameters. The θ-coordinates of a mixture family are given by

$$\theta_i = \frac{\partial B(\eta)}{\partial\eta_i} \quad (26)$$

where $B(\eta)$ is the negative entropy

$$B(\eta) = E[logp(x,\eta)] \quad (27)$$

and

$$A(\theta) + B(\eta) + \theta^T\eta = 0 \quad (28)$$

5.3. Generalized Pythagoras Theorem

The usefulness of a dual geometry and the essential role of the divergence is shown by the following theorem (see figure 6) [Csi75, Ama01].

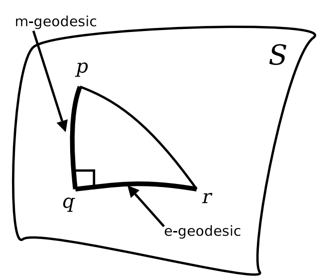


Figure 6: A diagram of the generalized Pythagoras theorem.

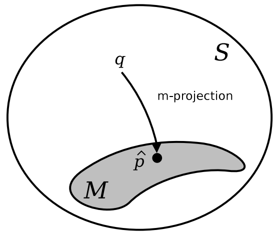


Figure 7: A diagram of the m-projection of q to M. The m-projection is a m-geodesic that is orthogonal at q to M. If M is e-flat then $\hat{p}$ is unique.

Theorem 2 (Generalized Pythagoras Theorem) *Let p, q and r be three points in a dually flat manifold such that the m-geodesic connecting p and q is orthogonal at q to the e-geodesic connecting q and r. Then,*

$$D(p\|q) + D(q\|r) = D(p\|r) \tag{29}$$

Let M be a submanifold in S, and let q be a point in S. We search for the point $\hat{p}$ in M that is closest to q in the sense of the divergence (see figure 7). When S is Euclidean, $\hat{p}$ is given by the orthogonal projection of q to M. But in a dually flat manifold, the divergence is not symmetric, so that we have two criteria of closeness and two solutions to this problem. To solve it, we first need to define m- and e-projections.

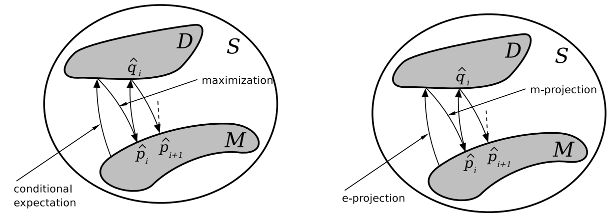
Definition 3 ($\mathbf{m}$(e)-projection) *Let $q \in S$ and $p \in M(D)$. When the $m(e)$-geodesic connecting q and p is orthogonal at p to $M(D)$, the point p is said to be the $m(e)$-projection of q to $M(D)$.*

The generalized Pythagoras theorem shows that the e-projection $\hat{p}$ of q to M gives the extremal point of $D(q\|p)$. Moreover, the projection is unique when the target manifold is flat.

5.4. Inference

Studying the problem of learning from the information geometrical point of view, when a complete data point q is observed, the ML estimation searches for the point $\hat{p} \in M$ that is closest to the observed point q. When the observation is partial, the point p is uniquely determined by the m-projection of q to M. The orthogonality is defined in term of the Riemannian metric. Dually to the above, the point $\hat{q} \in D$ that complements the unobserved data is determined by the e-projection.

From the above considerations, it has been formulated an em-algorithm that alternately uses m- and e-projections



(a) (b)

Figure 8: A graphical representation of the EM and em algorithms. While the maximization and m-projection are the same operation, conditional expectation is slightly different from e-projection (such difference is not shown in the figure).

from an initial guess point to reach the optimal pair of solutions [CT84].

From what we know of the EM algorithm (see section 3.2.1), both algorithms look quite similar (see figure 8). It is a well-known fact that the m-projection gives the ML estimation [Ama85], in [Ama01] it is proved that the conditional expectation of the E-step is very closely related to the e-projection. Indeed, the two algorithms are equivalent in most important cases and are asymptotically equivalent [Ama95]. The em-algorithm behaves better than EM, but the latter may be computationally more tractable.

6. Proposal

Hierarchical Bayesian models are intuitive and powerful instruments for the study of highly structured stochastic systems. The problems arise when we want to compute, compare and relate these models. This is in general not an easy task and although many tecniques and algorithms are available, there is not a unified approach.

The significance of information geometry stands in allowing us to move within a space that somehow inherits all the qualities that we want to explore. It is my interest to pursue such explorations and sort out one or more of the following problems.

Performance evaluations of algorithms for learning, with the intent of seeking modifications that guarantee convergence, robustness or optimality.

Variational model comparison to account for the different approximations that every simplification of the model brings.

Hierarchical analysis to see how hierarchies of submanifolds locally contain or spread the information they model.

Model selection is the general case of the previous two problems.

Conditional independence , how are the conditional independence statements coded in the manifolds?

Dynamic restructuration , that is, how can we build and understand models that dynamically change their structure, for example by adding or removing variables?

Finally, I have been deeply affected by the idea that the effectiveness of our brain mechanisms may be due to the ability to search for shortcuts through "dense" problem spaces like a river that snakes through the land, seeking the path of least resistance. That's how I envision our pattern matching ability within the geometry of the information that constantly surrounds us.

References

[Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, AC-19:716–723, 1974.

[Ama85] S. Amari. Differential geometrical methods in statistics. *Lecture Notes in Statistics*, 28, 1985.

[Ama95] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[Ama01] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[AN00] S. Amari and H. Nagaoka. *Methods of information geometry*. Oxford University Press, New York, NY, 2000.

[Arj03] E. Arjas. Commentary: causality and statistics. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statistical Science Series*, pages 66–69. Oxford University Press, Oxford, UK, 2003.

[Ash56] W. R. Ashby. *An Introduction to Cybernetics*. Chapman & Hall, London, UK, 1956.

[Ash60] W. R. Ashby. *Design for the Brain: The origin of adaptive behaviour*. Chapman & Hall, London, UK, 2nd edition, 1960.

[Bil97] J. A. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report ICSI-TR-97-021, ICSI, 1997.

[CCMP04] M. Cristani, D. S. Cheng, V. Murino, and D. Pannullo. Distilling information with super-resolution for video surveillance. In *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, pages 2–11, New York, NY, USA, 2004. ACM Press.

[CG92] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[Csi75] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

[CT84] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplemental Issue Number 1*, pages 205–237, 1984.

[Dah] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172.

[Daw03] A. P. Dawid. Causal inference using influence diagrams: the problem of partial compliance. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statistical Science Series*, pages 45–65. Oxford University Press, Oxford, UK, 2003.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Sociecy B*, 39(1):1–38, 1977.

[DLS80] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear models for contingency tables. *Annals of Statistics*, 8(3):522–539, 1980.

[Eic99] M. Eichler. *Graphical models in time series analysis*. PhD thesis, University of Heidelberg, Germany, 1999.

[Eic00] M. Eichler. Granger-causality graphs for multivariate time series. Technical report, Germany, 2000.

[FJ03] B.J. Frey and N. Jojic. Transformation-invariant clustering using the EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):1–17, January 2003.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, November 1984.

[GHKM01] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *Annals of Statistics*, 29:505–529, 2001.

[GHR03a] P. J. Green, N. L. Hjort, and S. Richardson, editors. *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, UK, 2003.

[GHR03b] P. J. Green, N. L. Hjort, and S. Richardson. Introducing Highly Structured Stochastic Systems. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statistical Science Series*, pages 1–12. Oxford University Press, Oxford, UK, 2003.

[Gly01] C. Glymour. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press, 2001.

[Gre03] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, volume 27 of *Oxford Statistical Science Series*, pages 179–198. Oxford University Press, Oxford, UK, 2003.

[GS90] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[GTS94] W. Gilks, A. Thomas, and D. Spiegelhalter. A language and a program for complex bayesian modelling. *The Statistician*, 43:169–178, 1994.

[Has70] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[HJT05] F. Heylighen, C. Joslyn, and V. Turchin, editors. *Principia Cybernetica Web*. Principia Cybernetica, Brussels, 2005. http://pespmc1.vub.ac.be.

[ITA04] S. Ikeda, T. Tanaka, and S. Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16(9):1779–1810, 2004.

[JF02] N. Jojic and B. Frey. A generative model for 2.5D vision: Estimating appearance, transformation, illumination, transparency and occlusion. *International Journal on Computer Vision*, 2002.

[JGJ99] M. I. Jordan, Z. Ghahramani, and T. S. Jaakkola. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[Jor04] M. I. Jordan. Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)*, 19:140–155, 2004.

[KFL01] F. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[Kri86] K. Krippendorff. Dictionary of cybernetics. Unpublished, 1986.

[Lau96] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, UK, 1996.

[MRR+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[MU49] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.

[Pea86] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 1988.

[Pea00] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, UK, 2000.

[Rab89] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[RC99] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, NY, 1999.

[RLJG02] S. Richardson, L. Leblond, I. Jaussent, and P. J. Green. Mixture models in measurement error problems, with reference to epidemiological studies. *J. of the Royal Statistical Sociecy A*, 165:549–566, 2002.

[ROL02] R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, editors. *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, 2002.

[RW84] R. Redner and H. Walker. Mixture densities, maximum-likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[Sch78] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[TB02] M.E. Tipping and C.M. Bishop. Bayesian image super-resolution. In *Neural Information Processing Systems - NIPS'2002*, Vancouver, 2002.

[Wie48] N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA, 1948.

[WL83] N. Wermuth and S. L. Lauritzen. Graphical and recursive models for contingency tables. *Biometrika*, 72:537–552, 1983.

[Wu83] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.