The Dempster-Shafer Calculus for Statisticians

A. P. Dempster

Harvard University, One Oxford Street, Cambridge, MA 02138, USA

Abstract

The Dempster-Shafer (DS) theory of probabilistic reasoning is presented in terms of a semantics whereby every meaningful formal assertion is associated with a triple (p, q, r) where p is the probability "for" the assertion, q is the probability "against" the assertion, and r is the probability of "don't know". Arguments are presented for the necessity of "don't know". Elements of the calculus are sketched, including the extension of a DS model from a margin to a full state space, and DS combination of independent DS uncertainty assessments on the full space. The methodology is applied to inference and prediction from Poisson counts, including an introduction to the use of join-tree model structure to simplify and shorten computation. The relation of DS theory to statistical significance testing is elaborated, introducing along the way the new concept of "dull" null hypothesis.

Key words: Dempster-Shafer; belief functions; state space; Poisson model; join-tree computation; statistical significance; dull null hypothesis

1 Introduction

The mathematical theory of belief functions was elegantly defined in the remarkable original book [1] by Glenn Shafer, with examples drawn mainly from simple AI-type situations. Earlier papers [2,3] of mine had described the same basic calculus in terms of upper and lower probabilities, using the example of multinomial sampling. The abstract theory shared by these two approaches is nowadays most often called Dempster-Shafer (DS) theory. That the centuries-old Bayesian inference probabilistic paradigm belongs under the umbrella of DS theory was long ago recognized [4]. In a seemingly unrelated context, the

Email address: dempster@stat.harvard.edu (A. P. Dempster).

important computer science theory of relational data bases [5] is easily recognized as constituting another limiting case where only zero or one are allowed as probabilities. These and other special subspecies have a unified character across a diverse spectrum of DS models. One essential common denominator is the appearance of join tree models and associated fast algorithms for propagating DS uncertainties through undirected networks [6,7].

My own return to working on DS theory was rekindled in the late 1990s, in part through requests for advice concerning probability assessment from scientific and operational agencies whose needs are evidently not met by more traditional theories. I was also inspired by a long conversation with Philippe Smets at Schwarzsee near Fribourg in Switzerland in the spring of 1999. Philippe put his own twist on the theory through his transferable belief model [8] aimed at difficult problems of decision-making under uncertainty. Philippe was a tireless advocate through a long and distinguished career. His presence is greatly missed following his untimely death. In the following paper I sketch some new DS terminology, attitudes, models, and statistical inference procedures, thus reaffirming my belief that the DS calculus has much to offer to colleagues seeking to study and apply methods for quantitative representation of incomplete and uncertain evidence. It is an honor to be able to dedicate the paper to Philippe's memory, in appreciation of his many contributions.

My original adoption of the terms lower and upper probability invites confusion with other theories that use these same terms, and provokes debate concerning the implied existence or nonexistence of unknown true probabilities lying between defined lower and upper bounds. No such existence is implied, since true probabilities nowhere appear in the theory. Shafer introduced belief and plausibility as replacements for lower and upper probability. For nearly 30 years I have acquiesced, on the grounds that everyday words are being used in a technical sense, not a dictionary sense, and carry no more taint of arbitrary judgment than other commonly used technical terms in statistical science, such as significance and confidence. I think of probability as a standard scale that quantifies uncertainty, but I have no problem with the term degree of belief. Belief in this sense implies nothing more than routine and tentative commitment to a mathematical idealization. The relevant question is always how much to trust model assumptions, including their consequences, considering each particular situation on its merits.

DS theory is founded on appending a third category "don't know" to the familiar dichotomy "it's true" or "it's false". More precisely, a DS model provides three nonnegative probabilities (p, q, r) with $p + q + r = 1$ to the three categories of the modal triad "known to be true", "known to be false", and "don't know" associated with each assertion specified in the model. It remains true that every statement defined within the model is in fact either true or false, but "you", the DS analyst, is no longer restricted to p and q with $p + q = 1$ as

in Bayesian theory. Since probabilities to which "you" commit tentative belief are presumed to be evidence-based, a probability p is construed to represent "your" evidence "for" the truth of an assertion, while probability q measures evidence "against", and probability $r = 1-p-q$ quantifies residual ambiguity.

Henceforth the quotation marks around "you", "your", "for", and "against" are omitted, but these remain technical terms throughout the sequel. A DS model implies the existence of an actor making implied assertions and subscribing to the associated uncertainty assessment, along with its understood basis in evidence. The actor referred to as you may be linked in an application to an individual analyst, but in scientific applications more typically refers to a community of scientists collectively recommending tentative acceptance of specific judgments. I continue to use quotation marks for "don't know" since the DS practice of assigning probabilities to this category is likely to be unfamiliar to most readers.

2 Elements of the DS calculus

2.1 The Constituent Parts of a DS analysis

I argue that DS methodology is a fundamental tool of scientific and operational analysis. To support this claim, both the mathematical basis and what it means must be understood. A brief sketch follows. In further writing with colleagues, I hope to gradually develop more detailed descriptions. A book length presentation will ultimately be needed, including many models, algorithms, software, and hypothetical examples.

In brief, a DS analysis can be represented in symbols as

$$(SSM + DSM) \times DSC = DSA \tag{1}$$

This is not meant as a mathematical formula, but rather as a shorthand overview of a step-by-step process. DSC refers to DS calculus, by which I mean the computational operations that are the technical basis of a DS analysis, or DSA. The computational inputs fall into two categories:

(1) the state space model or SSM, and

(2) the DS model, similarly abbreviated to DSM.

SSM and DSM are put forward as replacements, respectively, for Shafer's terms frame of discernment and belief function. To construct a model, the

SSM is first defined, then the DSM is specified over the defined SSM, and lastly the computational operations of the DSC are applied to complete the formal processes of a DSA.

2.2 The DS Concept of State Space

In widespread engineering language, a state space frames the possible states of a physical system at a single point in time. A particular system evolving in time is then a sequence of state space snapshots at successive points of time. In this paper, the terminology is altered to include trajectories of the system through space and time within an SSM. The flexibility conferred by this broadening makes it easy to expand or contract working choices of a formally represented small world in the course of developing a fully articulated DS analysis.

The SSM of a particular DSA is a mathematical space that encodes all the possible states of a system under analysis. You assume that exactly one among the possible states is the true state. The SSM is an idealized representation of a slice of reality, specifying your formal description of the small world that you have chosen to study. The purpose of DSA is to express probabilistically your uncertain inference about which element of the SSM represents the true state.

Given modern technologies for representing, recording, storing, accessing, and analyzing data, you are likely to have at your disposal mathematically realizable small worlds that are unimaginably large relative to systems that could be studied only a few years ago. On the other hand, realizable systems will always remain small relative to the limitless complexities of actual real world phenomena. Complexity of state spaces is scarcely an issue in this paper, however, since my examples consist of at most a few Poisson counts, and hence represent worlds that are small by any standard.

2.3 What it means to construct a DSM over a defined SSM

The mathematical content of a DSM consists of a system of (p, q, r) triplets that correspond one-to-one with assertions that you might make about a specified SSM. The mathematics of DSMs is most easily introduced assuming a state space that consists of a finite set of elements [1,3]. In this case, your possible assertions correspond to the 2^n subsets of the state space. The mathematical term for this collection of subsets is power set. The full power set includes the empty set $\emptyset$ and the full space $\mathcal{S}$. A corresponding full DSM creates a (p, q, r) triplet for every member of the power set. The triplets for the empty set and the full set are preordained to be $(0, 1, 0)$ and $(1, 0, 0)$,

respectively, because you are assumed to be sure that exactly one of the n elements of $\mathcal{S}$ represents the true state of the small world. The assignment of the remaining $2^n - 2$ probability triplets is at your disposal, but with important restrictions on choice. For example, if a subset $\mathcal{A}$ of $\mathcal{S}$ is associated with (p, q, r) then the complementary subset $\mathcal{A}^c$ must be associated with (q, p, r), because evidence for an assertion means the same as evidence against its negation.

While various axiom systems can be devised to underpin the mathematics of DSMs, these come down in the end to the assumption of consistency with a unique set function called the mass function. The mass function is any ordinary discrete distribution of mathematical probability over the $2^n - 1$ nonempty subsets of $\mathcal{S}$, including $\mathcal{S}$ itself. In other words, the simplest way to think about creating a DSM is to write out all the subsets in a long list and then to create the corresponding list of nonnegative real numbers summing to one that constitute the mass set function. In actual practice, most state spaces are not finite, and the probability mass is distributed over selected subsets of countable or continuous infinities of states. Nevertheless, despite its incomplete mathematical pedigree, the simple and rigorous finite space theory is a generally trustworthy guide, just as it is for basic textbook presentations of the standard theory of probability, in the special case where $r = 0$ for every assertion.

The mass $m(\mathcal{A})$ associated with each assertion $\mathcal{A} \subset \mathcal{S}$ is an atom in the sense that it cannot be further broken down into pieces assigned to subsets of $\mathcal{A}$. Logically, however, the probability $p(\mathcal{A})$ representing the uncertainty that you unambiguously assign to $\mathcal{A}$ is different from $m(\mathcal{A})$ because $p(\mathcal{A})$ accumulates masses from all the assertions that imply $\mathcal{A}$. In symbols,

$$p(\mathcal{A}) = \sum_{\mathcal{B} \subseteq \mathcal{A}} m(\mathcal{B}) \tag{2}$$

where as noted above the empty set $\emptyset$ has $m(\emptyset) = 0$.

The probabilities $p(\mathcal{A})$ defined in this way for all $\mathcal{A}$ determine the full triplet $(p(\mathcal{A}), q(\mathcal{A}), r(\mathcal{A}))$ associated with $\mathcal{A}$ because $q(\mathcal{A}) = p(\mathcal{A}^c)$ while $r(\mathcal{A}) = 1 - p(\mathcal{A}) - q(\mathcal{A})$. The set function $p(\mathcal{A})$ was called a belief function by Shafer, and DS theory itself is often called the theory of belief functions. The term belief function is unnecessarily formal, however, and has led to many misperceptions. In fact, $p(\mathcal{A})$ is a mainline successor to ordinary textbook probability, principally designed to allow you to assign a nonzero probability to "don't know".

Later in the paper, I discuss several hypothetical situations involving finite numbers of counts. To make a start, imagine a counter set up to record the number X_1 of occurrences of a rare event in a specified time unit, say one hour. Suppose you observe the count $X_1 = 5$. What can you say about the unknown count X_2 of occurrences that will be recorded in a second hour? And what can you say about the exact occurrence times $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$ underlying the count $X_1 = 5$?

The first step in creating a DSA is to set up an SSM. For example, to begin to address the first question, your state space represents at a minimum the pair of variables (X_1, X_2) where each component is a nonnegative integer. Later in the paper I will introduce a Poisson model for these counts, but for now suppose you know only $X_1 = 5$, and "don't know" anything about X_2. The preceding sentence translates to a DSM that assigns mass one to the subset $(5, X_2)$ of the two-dimensional space (X_1, X_2), where X_2 remains an unknown member of the set $\{0, 1, 2, \ldots\}$. To answer the second question as well as the first question, you must further extend the SSM to include the five successive occurrence times of the count $X_1 = 5$. Here, you "don't know" anything other than $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$ about these occurrence times, so you now have mass one on a subset of the 7-dimensional SSM defined by $(X_1, X_2, T_1, T_2, \ldots, T_5)$, as determined by both pieces of information $X_1 = 5$ and $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$.

The example is too simple to be substantively interesting, but it is nevertheless instructive from the perspective of DS logic because it illustrates two fundamental operations of the DS calculus. First, each of the properties $X_1 = 5$ and $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$ is a well understood logical DSM defined on respective $1D$ and $5D$ margins of the full $7D$ SSM determined by $(X_1, X_2, T_1, T_2, \ldots, T_5)$. Specifically, the data $X_1 = 5$ allows you to make (p, q, r) assessments about the subsets of the $7D$ SSM, because it extends from the $1D$ space of X_1 alone to the full seven dimensions by assigning mass one to the so-called cylinder subset of the $7D$ SSM defined by setting $X_1 = 5$. In a similar fashion, the condition $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$ translates into a marginal DSM that places mass one on the cylinder subset of the $7D$ SSM that satisfies the string of inequalities $0 \leq T_1 \leq T_2 \leq \ldots \leq T_5 \leq 1$. The second fundamental operation of the DSC is combination, here combining the two individual $7D$ DSMs just described into a single $7D$ DSM that places mass one on the subset satisfying both of the logical conditions. Each of the two original DSMs on the $7D$ SSM, as well as their combination, provides a (p, q, r) for every assertion that can be based on the full SSM. The output of a DSA consists of computed and reported (p, q, r) triplets corresponding to questions that you specify as of interest.

So why it is important to allow probabilities of "don't know"? A compelling answer is that "don't know" is implicit in every formal analysis, even when hidden from view. For example, the DSM of a Bayesian analysis assumes that each defined assertion has an associated $(p, q, 0)$ with $p+q = 1$. But in the real world situation addressed by the Bayesian analysis, many other meaningful dimensions are silently present, and could have been appended to the SSM. Viewing a Bayesian analysis as a DS analysis on such a margin of an expanded SSM leads to recognition of the implied assumption that you know nothing relevant about the added dimensions. Before reporting a Bayesian analysis, therefore, you should judge whether this assumption is adequate in the light of available evidence, and if not, the SSM should be expanded and a DSM reconstructed accordingly. A similar injunction applies to any reported DS analysis. Explicit recognition of the category "don't know" greatly extends your range of available probabilistic analyses.

2.5 DS Extension as a General Operation

The simple $(X_1, X_2, T_1, T_2, \ldots, T_5)$ example is purely logical in the sense that each mass function consists of mass one placed on a single subset of a marginal or extended SSM. I also argued that extension and combination are implicit in Bayesian analysis. These two operations generalize to situations where the component DSMs have nontrivial mass distributions. Since they are at the core of the DS calculus, it is important to explain them carefully.

When you restate the assertion $X_1 = 5$ in the context of the full $7D$ SSM, you are in effect placing mass one on the cylinder set in the full SSM that consists of all the possible states in 7 dimensions in which the marginal variable X_1 takes the value 5. The original $1D$ assertion and the extended $7D$ assertion, although mathematically distinct, convey exactly the same information. In place of $X_1 = 5$, you might more generally represent your evidence via DS masses summing to one over a collection of sets of possible values of X_1. Each set in the collection defines an assertion about X_1 that extends to a logically equivalent assertion about the full SSM whose extended DSM mass is defined to be the same as the original mass on the margin. For example, you might believe that your original counter is prone to error, so that instead of being sure that $X_1 = 5$, you might assign masses .5, .25, .25 to the three subsets $\{5\}$, $\{4, 5\}$,$\{5, 6\}$ of the possible values of X_1. Each of the three subsets projects up to the corresponding cylinder subset of the $7D$ space, and the three masses likewise project up along with the subsets to define DS masses for the $7D$ SSM cylinder sets, thus defining a $7D$ DSM.

The foregoing example is not special. Any marginal DSM extends directly to a logically equivalent DSM by converting each subset in the marginal SSM

into its corresponding cylinder subset in the full SSM, and applying each corresponding marginal mass to the associated cylinder. From the standpoint of interpretation a marginal DSM and any extensions to more refined SSMs carry exactly the same information, and the mathematical distinction is only a formality. Extension is fundamental to the calculus because model construction typically proceeds as in the $(X_1, X_2, T_1, T_2, \ldots, T_5)$ example by assigning DSMs to margins, then extending each to a common expanded DSM, before combining on the expanded DSM.

2.6 DS Combination of Independent Components

In the ordinary theory of probability, independence is represented by a product measure on a product space. In Boolean logic, the conjunction of events is represented by intersecting the state space subsets that represent the individual events. Both operations are special cases of DS combination, with independence assumed explicitly in the former case, and implicitly in the latter case. Another special case of DS combination is Bayes's formula for combining prior and likelihood in statistical inference [4], where again the independence assumption is implicit and rarely examined. In general, DS independence means that you deem the evidential sources underlying the individual input DSMs to be mutually non-compromising, whence directly combinable to represent pooled evidence.

A full exposition of DS combination breaks the operation down into a succession of steps. Given a list of independent component DSMs with a common state space $\mathcal{S}$, create from each component a list of all (subset, mass) pairs with positive mass (or mass density in the case of continuous state spaces). Next, from each such component list pick a (subset, mass) pair, and combine across components by intersecting subsets and multiplying probabilities. Doing this in all possible ways yields a raw combined list of (subset, mass) pairs. Given the raw list, the (subset, mass) pairs having a common subset can be combined into a single (subset, mass) pair by summing over masses, obtaining in this way a combined DSM referred to as unnormalized because there is in general positive mass on $\emptyset \subset \mathcal{S}$. Finally, the normalized combined (subset, mass) list is obtained from the unnormalized list by dropping the empty subset and rescaling the masses to sum to unity.

It is obvious that packets of evidence assumed DS independent may be combined sequentially in any order, producing the same final result. Although the preceding paragraph assumes normalized inputs and normalized output, the core product/intersection operation applies directly to unnormalized inputs that include mass on $\emptyset$, and produces an unnormalized output. In practice, it often makes sense to work sequentially with unnormalized DSMs. Normal-

ization is needed only at a final stage when the result is interpreted as your uncertainty.

When normalized DSMs are combined to produce an unnormalized DSM, the resulting $m(\emptyset)$ is called conflict. Conflict is an important by-product of DS combination. It is neither uniformly good nor bad. Good conflict uses logic to rule out impossible inferences associated with zero mass in separate input DSMs. For example, from the DS perspective, Bayesian inference typically has conflict near or at unity, and the final rescaling to normalize a posterior distribution is often a challenging computational task. On the other hand, if input DSMs imply contradictory inferences when considered separately, then conflict can be interpreted as contradicting aspects of component DSMs that are seen as subject to challenge. As always, model construction is partly a science and partly an art. You are the judge, and you must live with the consequences.

2.7 The Commonality Set Function

I have emphasized above the probability set function $p(\mathcal{A})$ and the mass set function $m(\mathcal{A})$, the former because in applications it represents directly interpretable measures of your uncertainty, and the latter because it is a natural way to conceptualize the associated distribution of random sets. Early on, however, a third set function

$$c(\mathcal{A}) = \sum_{\mathcal{B} \supseteq \mathcal{A}} m(\mathcal{B}) \tag{3}$$

was identified and recognized as technically important because in $c(\mathcal{A})$ terms DS combination becomes simple multiplication, and thus becomes basic to algorithm development for practical computation, as illustrated in Section 3.

Shafer [1] introduced the term commonality for the set function $c(\mathcal{A})$, and presented elegant Möbius linear transformation equations for passing back and forth among $p(\mathcal{A})$, $m(\mathcal{A})$, and $c(\mathcal{A})$.

2.8 The Associated Stochastic Calculus

Viewed abstractly, the mass functions of DS theory are ordinary probability measures built on a sample space whose elements are sets. The branch of the standard theory of probability that studies such measures is called the theory of random sets. It follows that mathematical theorems and computational algorithms for DS theory and methods can always be expressed in abstractly

equivalent forms that are easily understood by users familiar with the standard mathematics of probability, even while the thought patterns and technical concepts of DS modeling and analysis remain unfamiliar to most such users. To facilitate communication, the following examples present the details of familiar standard analyses that parallel the DS analyses of each example. To stress that physical randomness has no place in the interpretation of DS probabilities, terms such as randomness and random set are prefixed to read a-randomness and a-random set when discussing the details of parallel stochastic mathematics. The "a-" prefix abbreviates "associated".

3 Poisson Models and Analyses

3.1 Introducing the Poisson DSM

The Poisson DSM is a fundamental DS building block relating a pair of variables L and X, where L is a continuous non-negative rate variable, and X represents a count of events occurring at rate L. The state space of the Poisson DSM consists of horizontal lines in the (L, X) plane, as shown in Figure 3.1.

Fig. 3.1 about here. Caption: The SSM for the Poisson DSM.

The Poisson DSM is defined mathematically by assigning a mass distribution over a-random subsets of its (L, X) state space. These subsets are determined by an auxiliary sequence of a-random points $0 \leq V_1 \leq V_2 \leq V_3 \leq \ldots$ on the L axis. As illustrated in Figure 3.2, the auxiliary sequence $V_1, V_2, V_3, \ldots$ defines a corresponding sequence of intervals $0 \leq L \leq V_1, V_1 \leq L \leq V_2, V_2 \leq L \leq V_3, \ldots$ at levels $X = 0, 1, 2, \ldots$. The union of these intervals becomes an a-random set in the state space (L, X) when the lengths of the intervals are independently and identically distributed with the unit scale exponential density $exp(-u)$ for $u \geq 0$.

Fig. 3.2 about here. Caption: A typical a-random subset is the union of intervals at levels $X = 0, 1, 2, \ldots$.

It has become standard terminology in mathematical treatments of probability to call random points $V_1, V_2, V_3, \ldots$ defined in this way by exponentially distributed interval lengths a Poisson point process. As derived in standard textbooks such as Feller [9], the number of events preceding a fixed L in a Poisson point process is a Poisson random variable X whose discrete density is denoted here by

$$\pi(x|L) = \frac{L^x}{x!} exp(-L), \forall\ x\ \varepsilon\ \{0, 1, 2, \ldots\} \tag{4}$$

The DS interpretation of this basic mathematical fact is that if you treat the Poisson DSM as representing one source of uncertain knowledge about (L, X), and you obtain a known value of L as a second independent piece of evidence about the pair (L, X), your inference from combining the two sources of evidence is another DSM over (L, X) whose mass distribution is the above Poisson distribution over singleton subsets of the vertical line defined by the given L. The final DSA step is to marginalize the combined DSM to the $1D$ margin of X alone, which is trivial in this case, yielding the Poisson distribution for your inference about X. Note that the Poisson distribution is interpreted here as a mass distribution over the singleton subsets {0}, {1}, {2} ... of {0, 1, 2, ... }. Restricting the mass distribution to singleton subsets in effect converts the DSM to an ordinary Bayesian conditional probability measure – a common special type.

The novel aspect of the Poisson DSM is apparent when you have no prior evidence about L, but you are able to observe an accurate value $\tilde{X}$ of X, such as $\tilde{X} = 5$ in the toy example of Section 2.4. Any observed value $\tilde{X}$ of X can be interpreted as a DSM on the SSM (L, X) that places mass one on the horizontal line $X = \tilde{X}$ and assuming that the evidence in the observation is independent of the evidence provided by the Poisson assumption, then the combined evidence when marginalized to the $1D$ SSM of L places the unknown value of L on the a-random interval $V_{\tilde{X}} \leq L \leq V_{\tilde{X}+1}$. This characterization of inference about L was first given by Almond [10,11]. Another way to derive the Poisson DSM is to treat Poisson counts as a limiting case of the original Dempster [2] treatment of binomial counts, much as Poisson long ago obtained the Poisson family of distributions as limiting forms of binomial distributions.

The left end $V_{\tilde{X}}$ of this a-random interval is by definition the sum of $\tilde{X}$ independent exponential a-random variables, and hence has the unit scale gamma distribution with shape $\tilde{X}$, where the general form of the gamma density with shape $a \geq 0$ and scale $b \geq 0$ is denoted here by

$$\gamma(t|a, b) = \frac{b^a}{\Gamma(a)} t^{a-1} exp(-bt), \forall\ t \geq 0, \tag{5}$$

while the length of the interval $V_{\tilde{X}+1} - V_{\tilde{X}}$ is independently exponentially distributed with density $\gamma(t|1, 1) = exp(-t), \forall\ t \geq 0$.

It is a simple exercise in standard probability to check that the joint distribution of the pair $(V_{\tilde{X}}, V_{\tilde{X}+1})$ of a-random variables is neatly characterized by the formula

$$Pr(V_{\tilde{X}} \leq u, V_{\tilde{X}+1} \geq v) = \frac{1}{\tilde{X}!} u^{\tilde{X}} exp(-v), \forall\ v \geq u \geq 0. \tag{6}$$

In standard probability terms, this is a form of the bivariate cumulative distribution of the ends of the a-random interval $(V_{\tilde{X}}, V_{\tilde{X}+1})$. In DS terms, however, it is the commonality function $c(u, v)$ of the interval (u, v) for the posterior DSM of L given the observation $\tilde{X}$, and is useful as such, as shown in Section 3.2

3.2 Multiple Poisson Count Models

Instead of a single variable X that counts events occurring in a single unit time period of length one at an unknown Poisson rate L per unit time, I now assume n DS-independent variables $X_1, X_2, \ldots, X_n$ that represent counts, all occurring at the same rate L per unit time, but having differing known periods $\tau_1, \tau_2, \ldots, \tau_n$. This situation requires a DSM that combines n independent Poisson counts over the state space $(L, X_1, X_2, \ldots, X_n)$. You create this DSM mathematically by starting from marginal pairs (L, X_i) with $2D$ Poisson DSMs having rates $\tau_i L$, then extending each of these to the full $(n+1)D$ SS, and finally performing DS-combination on the full state space.

In practice it is computationally unnecessary, as well as impractical, to work with the full state space. Suppose, for example, that you seek the marginal posterior DSM of L given observations $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n)$. Each individual observation $\tilde{X}_i$ provides a commonality function for $\tau_i L$ determined by an application of formula (6). Simultaneous coverage of (u, v) by the n independent a-random intervals $(V_{\tilde{X}_i}, V_{\tilde{X}_i+1})$ is given by the product

$$\prod_{i=1}^{n} \frac{1}{\tilde{X}_i!}(\tau_i u)^{\tilde{X}_i} exp(-\tau_i v) \propto \frac{1}{\tilde{X}!}(\tau u)^{\tilde{X}} exp(-\tau v), \forall\ v \geq u \geq 0 \qquad (7)$$

where $\tilde{X} = \sum_{i=1}^{n} \tilde{X}_i$ and $\tau = \sum_{i=1}^{n} \tau_i$ showing that DS-combination of the n counts $\tilde{X}_i$ yields the same posterior DSM as if you had observed a single Poisson count $X = \tilde{X}$ at rate L per unit time, but observed for time period τ.

The result just stated is a DS version of the statistical principle of sufficiency introduced by R. A. Fisher in the 1920s. Under sufficiency, if the information about the individual X_i is lost and only the sum $X = \tilde{X}$ is retained, your posterior DSM of the unknown values of the X_i can be shown to be a standard multinomial distribution with parameters $(\tau_1/\tau, \tau_2/\tau, \ldots, \tau_n/\tau)$. In other words, the posterior DSM assigns all its mass to singleton subsets, and is equivalent to the celebrated fact recognized by Fisher that the conditional distribution of $(X_1, X_2, \ldots, X_n)$ given the sufficient statistic does not depend on the unknown L.

The prescription in equation (1) stipulates that DS analysis flows from a DSM computed on a full state space, implying that the rigorously correct way to carry out the analysis of Section 3.2 would have been to carry out DS-combination on the full $(n+1)D$ SS and then marginalize to the $1D$ SS of L. This is unnecessary due to the fundamental join tree theorem of DS analysis due to Shenoy and Shafer [6] or Kong [7], referred to briefly in Section 1, which to smooth exposition was omitted from Section 2. The simpler analysis of Section 3.2 does in fact lead to the same inference about L as the full SS analysis would have reached, for the following reason.

The final inference about L, that can be characterized in two equivalent ways, is defined as DS-combination of $2n$ input DSMs, namely, the n data points $X_i = \tilde{X}_i$ and the n Poisson DSMs associated with the pairs (L, X_i). These $2n$ inputs determine a mathematical structure called a join tree as pictured in Fig. 3.3.

Fig. 3.3 about here: The join tree for the multiple Poisson inference of Section 3.2. The arrows on the edges indicate the directions of inward propagation.

The first step in creating a join tree representation is to write out a list of all the subsets of the variables associated with independent component DSMs that are DS-combined to produce a full DSM. In the example at hand these subsets are $(X_1), (X_2), \ldots, (X_n), (L, X_1), (L, X_2), \ldots, (L, X_n), (L)$ all contained in $(L, X_1, X_2, \ldots, X_n)$. These $2n+1$ subsets are taken to be the nodes of a mathematical structure called a tree when they are joined by the smallest number of edges connecting all of them, namely, one fewer than the number of nodes. In the example there are $2n$ edges connecting the $2n+1$ nodes.

The tree is called a join tree if the nodes containing each single variable are connected by edges forming a corresponding subtree. The choice of a join tree for a given DSM is often not unique. In some models, it is necessary to group the input variables forming a smaller set of more complex variables. For example, the complete variable set $(L, X_1, X_2, \ldots, X_n)$ constitutes a trivial join tree with just one node, but is of no practical use. The goal is to define a join tree whose nodes are as small as possible, since the associated algorithm need only carry out DS-combination within nodes.

The inclusion of the node (L) in the example is unnecessary since the (L, X_i) nodes could simply have been connected directly with $n-1$ edges in many possible ways each providing a join tree, but including (L) with a vacuous prior DSM, such that "don't know" has value $r = 1$ on every margin, is an important feature because propagation algorithms that characterize DS-combination in join trees are designed to clip off extremal nodes seriatim finally collapsing

the tree to a single node and ending with the implied marginal DSM on that node. The join tree theorem asserts that a node can be clipped by extension to the SSM of its neighbor, and DS-combining there, to form the marginalized join tree. In the example, the singleton nodes (X_i) are incorporated into their neighboring (L, X_i) nodes by DS-combining their respective $X_i = \tilde{X}_i$ and Poisson DSM components as in Section 3.1. Then the resulting (L, X_i) nodes are clipped leaving only the (L) node carrying the desired posterior DSM. In this way, the intuitively correct analysis given in Section 3.2 is rigorously justified.

3.4 Prediction: the Example of Section 2.4

Consider now the preceding model, specialized to $n = 2$, and altered to have X_1 observed as before, while X_2 is unobserved with no prior DSM assigned. The input DSM now has three independent components, namely the observation $X_1 = \tilde{X}_1$ and the two Poisson DSMs on (L, X_1) and (L, X_2). A basic task is to combine and marginalize to the predictive DSM for X_2 given the data $X_1 = \tilde{X}_1$. The same set-up was posed in Section 2.4 with $\tau_1 = \tau_2 = 1$ and $\tilde{X}_1 = 5$, but without the Poisson assumptions. The point here is to show how the Poisson assumptions lead rigorously to a DSM prediction of X_2. The Poisson assumption on (L, X_1) leads also to a simple posterior DSM for the occurrence times $0 \leq T_1 \leq T_2 \leq \ \ldots \leq T_{\tilde{X}_1} \leq \tau_1$ of the events counted by X_1.

Fig. 3.4 about here. Caption: The join tree for prediction of X_2 with arrows indicating the direction of information flow.

The join tree is now as shown in Fig. 3.4 with clipping proceeding toward the node (X_2), as contrasted with clipping toward the node (L) required in Section 3.3. Clipping the nodes (X_1), then (L, X_1), the resulting combined DSM at node (L) amounts to placing $\tau_1 L$ on the a-random interval $V_{\tilde{X}_1} \leq \tau_1 L \leq V_{\tilde{X}_1+1}$, or equivalently placing $\tau_2 L$ on the a-random interval

$$\frac{\tau_2}{\tau_1} V_{\tilde{X}_1} \leq \tau_2 L \leq \frac{\tau_2}{\tau_1} V_{\tilde{X}_1+1} \tag{8}$$

whose distribution is described by a modification of equation (6). This information about L must now be combined with the Poisson DSM at the (L, X_2) node whose a-random sets are as shown in Fig. 3.2 with $\tau_2 L$ placed on the union of the successive intervals

$$0 \leq \tau_2 L \leq V_{21},\ V_{21} \leq \tau_2 L \leq V_{22},\ V_{22} \leq \tau_2 L \leq V_{23},\ \ldots\ , \tag{9}$$

at levels $X_2 = 0, 1, 2, \ldots$.

DS-combination looks at the intersections of the intervals for $\tau_2 L$ from (8) and (9), and in the case of (9) at the corresponding levels of X_2. It is evident that the interval (8) must intersect a consecutive subset of the intervals (9), whence the a-random set that defines the desired posterior DSM for X_2 is an a-random interval with endpoints $(J, J+K)$ where J and K are non-negative integer-valued a-random variables.

The distribution of J and K is best understood by first conditioning on the interval (8) and then averaging over $V_{\tilde{X}_1}$ and $V_{\tilde{X}_1+1}$. From the theory of Poisson point processes, J and K are conditionally independent Poisson a-random variables with rate parameters $(\tau_2/\tau_1)V_{\tilde{X}_1}$ and $(\tau_2/\tau_1)(V_{\tilde{X}_1+1} - V_{\tilde{X}_1})$. The theory developed in Section 3.1 shows that $V_{\tilde{X}_1}$ and $V_{\tilde{X}_1}$ are a-independently gamma distributed with shapes $\tilde{X}_1$ and 1, and scale unity, the latter being a simple unit scale exponential a-random variable, whence it is easily shown that J and K are independently distributed as negative binomial a-random variables with densities $\eta(j|\tilde{X}_1, \tau_2/(\tau_1+\tau_2))$ and $\eta(j|1, \tau_2/(\tau_1+\tau_2))$, the latter being a simple geometric density, where the general form of a negative binomial density is

$$\eta(j|l, P) = \frac{(j+l-1)!}{j!(l-1)!} P^j (1-P)^l \ \forall \ j \ \varepsilon \ \{0, 1, 2, \ \dots \}. \tag{10}$$

For example if $\tilde{X}_1 = 5$ as in Section 2.4, the (p, q, r) triple associated with the assertion $X_2 \leq 10$ is $(.895, .039, .046)$, or if $\tilde{X}_1 = 10$, the (p, q, r) triple associated with the assertion $X_2 \leq 15$ is $(.836, .115, .049)$, assuming $\tau_1 = \tau_2$ in both cases.

Finally, recall the remark at the end of Section 3.2 that the components of a sum of independent Poisson counts has a multinomial distribution given that you have observed only the sum. A limiting case of this theorem shows that given an observation $X_1 = \tilde{X}_1$ that results from events occurring at a constant Poisson rate over the time interval $(0, \tau_1)$ the individual times $0 \leq T_1 \leq T_2 \leq \ \dots \ \leq T_{\tilde{X}_1} \leq \tau_1$ are distributed as $\tilde{X}_1$ uniformly distributed ordered a-random draws from the interval $(0, \tau_1)$.

4 Significance Testing

Think about a hypothetical situation where a firm sells a small manufactured item in batches of 1000 with an advertised defective rate of 1.6% per item. Your first purchase of a batch turns out to have 24 defective items, which is 50% more than the expected number of 16 per 1000. Is the observation 24 consistent with Poisson batch-to-batch variation at the rate 16? The "don't know" feature of DS theory leads naturally to a reformulation that sheds new

light on traditional statistical tests of significance.

In the situation just described, many scientists would routinely compute a "p-value", defined as the probability under the null hypothesis of an observation at least as extreme as the actual observation. According to R. A. Fisher [12], a computed p-value less than a tipping point such as .05 or .01 or .001 has the force of a logical disjunction. Either the null hypothesis being tested is false, or a chance occurrence with small prior probability has occurred. Reluctance to believe the latter raises your propensity to believe the former. For the model at hand, a relevant p-value is $\sum_{j=\tilde{X}}^{\infty} \pi(j|L_0)$, or .0367 in the numerical example. Fisher's interpretation is often criticized on the grounds that actual observations are almost always highly improbable *a priori*, without raising questions about the basis of the computed probability. The long and controversial history of frequentist and Bayesian elaborations and alternatives to Fisher's position is by now rather stale. What follows is new and different.

Scientists often misconstrue a p-value and its complement, such as the above .0367 and .9633, as Bayesian probabilities summing to one for and against the null hypothesis. There is no basis in the mathematics for this interpretation, however. As I now demonstrate, DS analysis suggests that the fallacious interpretation is half right in the specific sense that .9633 is a meaningful probability for the hypothesis $\{L > L_0\}$, while .0367 splits into two parts, with .0223 assigned to $\{L < L_0\}$ and .0144 assigned to "don't know".

Following the analysis of Section 3.1, an observed $\tilde{X}$, such as $\tilde{X} = 24$, leads to a posterior DSM that places L on the a-random interval $(V_{\tilde{X}}, V_{\tilde{X}+1})$. This interval has three mutually exclusive positions relative to an assumed null value L_0 of L, such as $L_0 = 16$, namely, entirely to the left of L_0, or covering L_0, or entirely to the right of L_0. In symbols, these conditions are $V_{\tilde{X}+1} < L_0$, or $V_{\tilde{X}} \leq L_0 \leq V_{\tilde{X}+1}$, or $V_{\tilde{X}} > L_0$. From the mathematical connection between Poisson point processes and Poisson distributions, it follows that these three cases have Poisson a-probabilities $\sum_{j=\tilde{X}+1}^{\infty} \pi(j|L_0)$, $\pi(\tilde{X}|L_0)$, and $\sum_{j=0}^{\tilde{X}-1} \pi(j|L_0)$, respectively, whose sum is 1. In the numerical illustration, these values are .0223, .0144, and .9633.

Switching from the formal a-mathematics of the previous paragraph to its DS interpretation, the observation $\tilde{X}$ implies a marginal DSM over the three member state space $\mathcal{L} = \{L > L_0\} \cup \{L = L_0\} \cup \{L < L_0\}$ defined by masses $\sum_{j=\tilde{X}+1}^{\infty} \pi(j|L_0)$ and $\sum_{j=0}^{\tilde{X}-1} \pi(j|L_0)$ assigned to $\{L > L_0\}$ and $\{L < L_0\}$, but with remaining mass $\pi(\tilde{X}|L_0)$ assigned to the full space $\mathcal{L}$. The reason for the "don't know" assignment of the single Poisson term is that with a-probability one any a-random interval that covers $\{L = L_0\}$ also intersects $\{L > L_0\}$ and $\{L < L_0\}$.

Three aspects of this interpretation deserve special attention. First, as claimed

above, the standard p-value splits into the two parts given by $\sum_{j=0}^{\tilde{X}-1} \pi(j|L_0)$ and $\pi(\tilde{X}|L_0)$. Another way to put this is to state that the assertion $\{L > L_0\}$ has $(p, q, r) = (\sum_{j=\tilde{X}+1}^{\infty} \pi(j|L_0), \sum_{j=0}^{\tilde{X}-1} \pi(j|L_0), \pi(\tilde{X}|L_0))$, or in the example $(p, q, r) = (.9633, .0223, .0144)$. Second, because no mass resides on $\{L = L_0\}$ alone, the probability for the assertion that the null hypothesis is true is zero. This accords with the understanding by applied statisticians that the logic of traditional Fisherian significance testing is only able to reject a null hypothesis, and in no way confirms it. In the DS interpretation, the null hypothesis is rejected by confirming an alternative.

The third aspect takes us into uncharted territory. When statisticians think about testing a null hypothesis, they generally think about designing a test to be sensitive against defined alternatives. For example, the test with p-value .0367 is regarded as directed against the alternatives $\{L > .016\}$. If you wish to test against the two-sided alternative $\{L > L_0\} \cup \{L < L_0\}$, the DS inference for the assertion that this alternative is true is $(p, q, r) = (.9856, 0, .0144)$. In another type of situation, actually a variant of the one-sided situation, you may believe with certainty that $\{L < L_0\}$ is ruled out. Treating the last assumption as another DSM to be combined with the Poisson DSM, means that the mass $\sum_{j=0}^{\tilde{X}-1} \pi(j|L_0)$ is conditioned out and the remaining two terms are renormalized to sum to one. In the example, the resulting binary assertion $\{L > .016\}$ vs its negation $\{L = .016\}$ has $(p, q, r) = (.9853, 0, .0147)$. There is very little difference between the two cases, but both suggest that the null hypothesis can be rejected more strongly than the original p-value of .0367 indicated. Is this credible?

To put the issue more starkly, imagine a situation where you have experienced 100 batches of 1000 items while continuing to assume a stable Poisson rate, and where you have found 1680 defective items. Having 100 times as many observations as in the single batch example, your estimate of L has only 1/10th the standard error, but the hypothetical deviations from expected, namely, 24 - 16 = 8 and 1680 - 1600 = 80 were both chosen to be "2 sigma" values and hence to have roughly the same traditional p-values for testing the null hypothesis $\{L = .016\}$. In fact the traditional p-value is now .0241 and splits in two parts now .0227 and .0014. The reduction in the traditional p-value comes mainly from the second term, which dropped from .0147 to .0014, which is again an expected roughly 1/10 for theoretical reasons not presented here. Now, however, the two-sided and one-sided (p, q, r) are identical to 4 decimal places at $(.9986, 0, .0014)$ leaving much less room for doubt than did the original p-value of .0276. Again, is this credible? At first sight the result is paradoxical, but in fact means only that no precise value of L has support from the data among a continuous infinity of possibilities. For this example, the question has become very different from that addressed by the traditional p-value.

The DS formulation suggests widening the concept of null hypothesis. A traditional null hypothesis such as $\{L = L_0\}$ may be called "sharp" [13]. Adopting a DS outlook, however, it becomes natural to consider a contrasting type that may be called "dull", where the null hypothesis asserts that the parameter of interest lies in a defined region. For example, a supplier might certify only that $\{.015 \leq L \leq .017\}$. In DS terms, a dull null hypothesis is a DSM that places mass one on the associated region. For example, it is a familiar type of engineering specification to set a predicted failure rate per item in a range such as $.016 \pm .001$ with no probabilistic concept of error distribution implied.

With the dull null hypothesis defined as an interval $\{L_1 \leq L \leq L_2\}$, and the uncertainty about L specified as the a-random interval $(V_{\tilde{X}}, V_{\tilde{X}+1})$, inference concerning the null hypothesis depends on six computed a-probabilities for the events that (1) the a-random interval is entirely to the left of L_1, (2) entirely to the right of L_2, (3) entirely contained in the interval $\{L_1 \leq L \leq L_2\}$, (4) overlapping the left end of the interval but not the right end, (5) overlapping the right end of the interval but not the left end, and (6) overlapping the entire interval. Straightforward derivations of formulas for a-probabilities, and details of connections with relevant (p, q, r) assessments are omitted here in favor of numerical examples that illustrate how sharp and dull null hypotheses may be expected to compare in practice.

In place of marginal uncertainties about $\mathcal{L} = \{L > L_0\} \cup \{L = L_0\} \cup \{L < L_0\}$ interest now focuses on $\mathcal{L}_+ = \{L > L_2\} \cup \{L_1 \leq L \leq L_2\} \cup \{L < L_1\}$. Returning to the hypothetical data $\tilde{X}_1 = 24$, and the dull null hypothesis $\{.015 \leq L \leq .017\}$, the traditional test asks whether the observation is too large relative to the expected null limit of .017, leading to the p-value .0633. The six a-probabilities of the preceding paragraph are $.0112, .9367, .0223, .0072, .0215, .0011$, leading to $(.9367, .0407, .0226)$ for the one-tail assertion $\{L \geq .017\}$, or $(.9479, .0223, .0298)$ for the two-tail assertion $\{L > .017\} \cup \{L < .015\}$. The evidence for failure of the dull null hypothesis is of course weaker than for the sharp null hypothesis, but not drastically so. The situation with the observation 1680 is much different. Now the six a-probabilities are $.0000, .3106, .6807, .0000, .0086, .0000)$, leading to $(.3106, .6807, .0086)$ to 4 decimal places for the one-tail or two-tail assertions, and there is substantial probability .6807 for the assertion that the dull null hypothesis is true.

Indications are that DS analyses can reproduce standard significance test p-values, but are capable of much more focus, depth, and sophistication in the handling of statistical uncertainties. Complex issues regarding significance testing merit careful discussion beyond the limitations of this paper, both to fully develop and explore DS possibilities, and to compare with non-DS approaches. The goal here is just to plant a seed.

Acknowledgments.

The work reported here was supported in part by the Office of Naval Research through grants N00014-98-1-0761 and N00014-02-1-0 412. Technical assistance was provided by Prof. Chuanhai Liu of Purdue University.

References

[1] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press (Princeton, New Jeresy, 1976).

[2] A. P. Dempster, New methods for reasoning towards posterior distributions based on sample data, Ann. Math. Statist. 37 (1966) 355-374.

[3] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Statist, 38 325-339 (1967).

[4] A. P. Dempster, A generalization of Bayesian inference, J. R. Statist. Soc. B, 30 205-247 (1968).

[5] D. Maier. The Theory of Relational Data Bases, Computer Science Press, Rockville Maryland, (1983).

[6] P. P. Shenoy and G. Shafer, Propagating belief functions with local computations, IEEE Expert, 1, 43-52 (1986).

[7] A. Kong, Multivariate Belief Functions and Graphical Models, Unpublished Ph.D. Thesis, Department of Statistics, Harvard University (1986)

[8] P. Smets and R. Kennes, The transferable belief model, Artificial Intelligence, vol. 66, pages 191-234, 1994.

[9] W. Feller, Probability Theory and its Applications, Wiley, New York (1950).

[10] R. G. Almond, Fusion and Propagation of Graphical Belief Models: and Implementation and an Example, Unpublished Ph.D. Thesis, Department of Statistics, Harvard University (1989).

[11] R. G. Almond, Graphical Belief Modeling, Chapman & Hall, London (1995).

[12] R. A. Fisher, Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh (1956).

[13] Savage, L. J. et al, The Foundations of Statistical Inference: a Discussion, Methuen, London (1962).

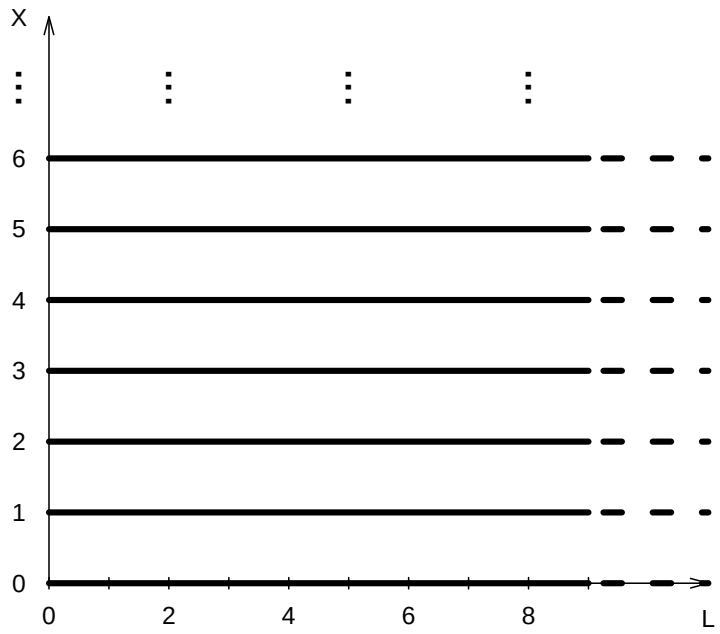

Fig. 3.1. The SSM for the Poisson DSM.

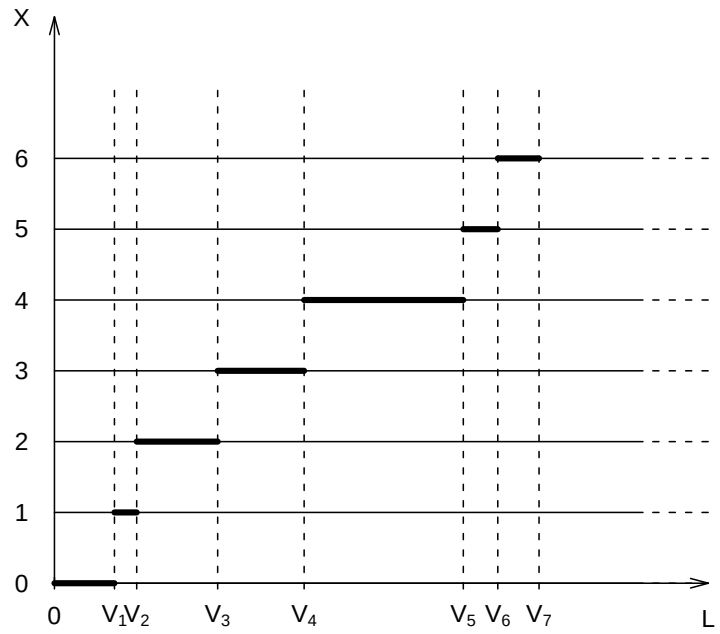

Fig. 3.2. A typical a-random subset is the union of intervals at levels $X = 0, 1, 2, \ldots$.

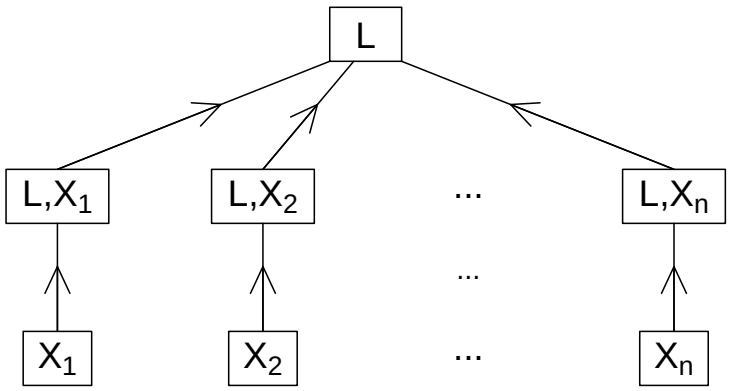

Fig. 3.3. The join tree for the multiple Poisson inference of Section 3.2. The arrows on the edges indicate the directions of inward propagation.

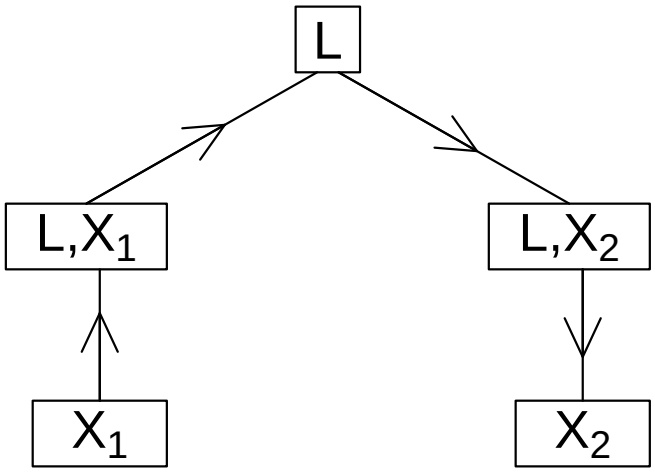

Fig. 3.4. The join tree for prediction of X_2 with arrows indicating the direction of information flow.