A Probability Theoretic Analysis of Score Systems

Bertram Fronhöfer
Institut für Informatik
Ludwig-Maximilians-Universität München
fronhoefer@pit-systems.de

Manfred Schramm
Institut für Informatik
Technische Universität München
schramm@pit-systems.de

Abstract

Due to their simple applicability score systems are in widespread use as a tool for decision taking. Unfortunately, as we all feel, they are somehow not apt to take into account interdependencies among the variables (symptoms/attributes) which are input to them when trying to decide an actual case; a drawback which is overcome by probabilistic systems.

In order to analyze which assumptions are inherent in score systems we translate them into probabilistic systems thus making available their technical machinery for this analysis task.

Keywords: Score Systems, Probabilistic Reasoning, Maximum Entropy, Automated Diagnosis, Independence

1 Introduction

Medicine, industry and economy abound with problems of diagnosis (and decision). Illustrated by an example from medicine, such a problem consists of a knowledge base of propositions about the problem domain (e.g. 'Appendicitis is usually accompanied by strong stomach aches'), the values of certain symptoms in an actual (diagnosis) case (e.g. 'The patient has stomach aches'), a wanted diagnosis (e.g. 'Does the patient suffer from appendicitis?'), and finally, a decision ('surgical intervention?').
To solve such problems, **Score Systems** are frequently developed in research labs, e.g. in medicine ([OYF95]), or are in wide-range use, e.g. in economics ([KR00]), whenever uncertain knowledge plays an important role with the kind of problem to be solved.

A Score System is based on a set of **attributes** (or **variables**) which have each a set of possible (attribute or variable) **values**.

For instance, in the medical domain, there may be the symptom/attribute 'body temperature' with (discrete) values 'low', 'normal', 'high' and 'very high'. To each attribute value a numerical value — its **weight** or **score**— is assigned (see Table 1).

When applying a score system to a concrete case the scores corresponding to the observed attribute values are added up. If the obtained sum falls in a certain **score interval**, the decision associated to this interval is proposed.

Thus, for instance, a proposal for a medical treatment is established on the basis of symptoms found with a patient and which are represented by a list of attribute values.

Example 1:

symptom / attribute	score, if yes
tenderness in RLQ ‘	4.5
rebound tenderness	2.5
no micturition	2.0
continuous type of pain	2.0
number of leucocytes ≥ 10000	1.5
age < 50 years	1.5
relocation of pain to RLQ	1.0
rigidity	1.0

Table 1: 'Ohmann Score' [OFY+95] for the diagnosis of appendicitis: In case of negative answers the scores are zero. Patients are diagnosed as having appendicitis if score sum ≥ 12, they are interned in case of $6 - 12$, and are sent home in case of ≤ 6. (RLQ: right lower quadrant of abdomen (as seen from the patient).)

When applying a score system to an actual case of decision finding, we feel intuitively that a score system seems to make some kind of assumption about the unrelatedness of the variables, because it provides no means to adapt the scores assigned to the values of one variable in dependence of the selected values of other variables. This makes us feel uneasy, because very often we are intuitively aware of influences between some of the variables, but by using a score system we are bound to the use of a tool where these relationships are not taken into account. Sometimes this feeling is expressed in the opinion that score systems implicitly assume marginal independence of variables/attributes.[1]

The aim of this paper is to analyze rigorously our intuitive feelings about score systems and to explicit the hidden assumptions underlying them. To this end we translate scores systems into probabilistic systems as faithfully as possible thus being able to exploit the technical framework of probability theory for the investigation of independences and indifferences. Being convinced of the faithfulness of our translation, we claim, that our discoveries in the obtained probabilistic systems shed light on the nature of score systems.

Example 1 (continued):

As a preview of the outcome of the translation which we will develop in the following, we give a brief sketch of the score system from Example 1 turned into a probabilistic system according to the transdlation which will be given in section 5.1.

To each combination of observed symptoms we assign then probability defined by the some of the respective score values divided by the maximally achievable score value. E.g. if there are the symptoms ‘tenderness in RLQ’, ‘continuous type of pain’ and ‘rigidity’ —

[1] In the "Handwörterbuch der Wirtschaftswissenschaft" (handbook of economy) the keyword "Produktplanung" (product planing) contains the following contribution by Dietrich Adam: "Die einfachste Form der Synthese der Urteile über einzelne Kriterien besteht in der Addition der gewichteten Punktzahlen aller Kriterien. Diese Art der Synthese hat jedoch unabhängige Kriterien zur Voraussetzung, d.h. ein positives Urteil bei einem Kriterium darf nicht mit dem Urteil bei anderen Kriterien korrelieren." (The simplest form of synthesizing the judgments of individual criteria consists in adding the weighted scores of all criteria. However, independent criteria are a prerequisite for this kind of synthesis, i.e. a positive judgment of one criterion must not correlate with the judgments of other criteria.)

and only those — then this probability would be $4.5 + 2.0 + 1.0/16 = 7.5/16 = 15/32$. To the border values 6 and 12 would correspond the probabilities $6/16 = 3/8$ and $12/16 = 3/4$. (Note that 16 is the sum of all score values in Table 1.) ∎

The paper is organized as follows: In section 2 we give a formal definition of score systems. In section 3 we present some general considerations on diagnosis systems and emphasize the need to model explicitly the relationship between diagnoses and symptoms. This leads naturally to the introduction of some terminology from probability theory, namely the terminology of event spaces, and culminates in the definition of a probabilistic diagnosis system in section 4. It will turn out that a score system is a little bit deficient in view of these considerations, but some slight extensions turn it quite directly into a probabilistic system. In section 5 this faithful translation is presented and analyzed. It is defined via constraints on a probability measure, which are derived from a score system. In section 6 we introduce additional constraints in order to cope with missing attribute values.

Examples in this paper are often taken from the field of medical diagnosis. Often our general explanations are tinged by medical language, which — in our opinion — better adds to clarity, than a more neutral, but less suggestive language.

2 Score Systems

Formally, a Score System can be defined as follows:
It consists of a set of **variables** (**attributes**) S_i $(i = 1, \ldots, m)$. Each S_i can be identified with its set of variable **values** $\{s_{i1}, \ldots, s_{ik_i}\}$ $(k_i > 1)$. We denote by $\vec{s}$ a tuple of values $\langle s_1, \ldots, s_m \rangle$ with $s_i \in S_i$.
Moreover, for each variable S_i exists a set $W_i = \{w_{i1}, \ldots, w_{ik_i}\}$ of nonnegative **weights** or **scores** and a bijective **score function** $w_i \colon S_i \longrightarrow W_i$. We also have a **(global) score function** w defined as $w(\vec{s}) := \sum_{i=1}^{n} w_i(s_i)$.
Finally, there are **score intervals** given by a set of **border values** $b_1 < \cdots < b_{k_T}$, a **decision variable** T with values $\{t_1, \ldots, t_{k_T}\}$ and a **decision function** t which maps a sum of scores $w(\vec{s})$ to t_i iff $b_{i-1} < w(\vec{s}) \leq b_i$ (with $b_0 := -\infty$).

Since decision functions are not the topic of this paper, we will skip any explicit treatment and just show that the border values can be adapted when translating score systems. Basically, within score systems decision functions are a very simple add-on to the kernel of a score system and consequently neglecting them in the sequel is no restriction. More sophisticated decision methods which are available with probabilistic systems cannot be adapted to score systems, because they require multiple diagnoses which score systems are not able to provide.

Example 2: A simple example of a score system, which we will reuse in the following, consists of 3 binary variables S_i $(i = 1, 2, 3)$ — whose values we denote as $S_i = \{s_i, \overline{s_i}\}$ — together with the 3 score functions w_i which map S_i to a respective (binary) set of values $W_i = \{i, 0\}$ with $w_i(s_i) = i$ and $w_i(\overline{s_i}) = 0$. ∎

For a further example including a decision function see Example 1 above.

When attempting to translate a formal system into another one, the crucial point is to

obtain what might be called a homomorphic image, thus assuring faithfulness. For this purpose the 'source system' should be investigated for formal properties which should be conveyed into the 'target system' via the translation.
In score systems we only discovered the following **contribution invariance** property: For arbitrary $s_{i1}, s_{i2} \in S_i$ there exists a constant $\text{const}_{s_{i1} \to s_{i2}}$ such that

$$w(\vec{s}) - w(\vec{s}[s_{i1} \to s_{i2}]) = \text{const}_{s_{i1} \to s_{i2}} \tag{1}$$

holds for all $\vec{s}$ in which s_{i1} occurs. (We denote by $\vec{s}[s_{i1} \to s_{i2}]$ the tuple obtained from $\vec{s}$ when replacing s_{i1} by s_{i2}.) From the easy proof of this property, it is clear, that in a more 'axiomatic' formulization the subtraction is the appropriate instance of the inverse of the function which computes the overall score $w(\vec{s})$ from the scores of the symptom values, i.e. addition in the case of a score system.

3 Diagnosis Systems: Basic Considerations

Before we present a translation of score systems into probabilistic systems we will present some general thoughts on modeling systems for diagnosis and establish some general requirements.

3.1 Diagnosis-Attribute Relationship

Any diagnosis system must model the relationship between diagnoses (diseases) and attributes (symptoms) relevant for these diagnoses. In order to express knowledge about the evidence of a disease in view of certain symptoms as well as to express perhaps quite different knowledge about the occurrence of symptoms in case of a disease, diagnoses must be represented explicitly. This means that we have a finite set of **variables** — symptom variables and, in addition, diagnosis variables — with each having a finite set of **values**.[2]
The **symptom variables** describe properties / symptoms / attributes relevant for the diagnosis task, e.g. examination results in the medical case.

> As before, we denote symptom variables by S_i $(1 \leq i \leq m)$, and we identify them with the set of their values: $S_i = \{s_{ij} | 1 \leq j \leq k_i\}$.
> We sometimes refer to values of variable S_i just by s_i, s_i' ...

The values of a **diagnosis variable** define a classification of the possible diagnostic results, e.g. in kinds of diseases, based on the values of the symptom variables.

> In the following we will only consider the case of a single diagnosis variable $D = \{d_1, \ldots, d_{k_D}\}$.[3]

The considerations presented above lead naturally to the view that a diagnosis system must model a relation on the tuple space $\Omega := S_1 \times \cdots \times S_m \times D$.

[2] Infinite numbers of symptoms or diagnoses cannot be taken into account in a real world application and an infinite range of values can be made finite in practice through appropriate discretization.

[3] This constitutes no restriction, because the case of a set of diagnosis variables can always be coded as a single one with values reflecting the possible combinations of the values from the set.

This relation can be specified by a **method of judgment**. Depending on this method we may either get a 'classical' relation or a somehow fuzzy one, i.e. we may have yes-no judgments on tuples $\langle s_{1j_1}, \ldots, s_{mj_m}, d_h \rangle$ or a more fine-grained judgment, e.g. scores or probability measures.

3.2 Uncertain Findings

It may occur quite often that the value of a certain symptom, say S_1, cannot be observed. This means that a method of judgment, which judges only individual tuples, is very restricted in view of practical applications, as we might be interested in getting a judgment of a set of tuples from Ω as, for instance,

$$\bigcup_{j_1=1}^{k_1} \langle s_{1j_1}, s_{2j_2}, \ldots, s_{mj_m}, d_h \rangle$$

which reflects ignorance of the value of S_1. Of course, there may also be cases where some values of S_1 can be excluded, while still not arriving at a single remaining variable value, thus obtaining a smaller set than the one given above, while still not getting a single tuple. Consequently, judgments on arbitrary subsets of Ω are desirable.

> In the terminological tradition of probability theory, we say that we consider Ω as an **event space** with its power set as **set of events** or **event algebra**.
>
> For a subset $\breve{S}_i \subset S_i$ of a symptom variable or a subset $\breve{D} \subset D$ of a diagnosis variable we denote $\langle \breve{S}_i \rangle := S_1 \times \cdots \times S_{i-1} \times \breve{S}_i \times S_{i+1} \times \cdots \times S_m \times D$ resp. $\langle \breve{D} \rangle := S_1 \times \cdots \times S_m \times \breve{D}$. By extension, we denote by $\langle \breve{S}_{i_1}, \ldots, \breve{S}_{i_n} \rangle$ with $\breve{S}_{i_j} \subset S_{i_j}$ the intersection of the sets $\langle \breve{S}_{i_j} \rangle$. We just write $\langle s_i \rangle$ and $\langle s_{i_1}, \ldots, s_{i_n} \rangle$ instead of $\langle \{s_i\} \rangle$ and $\langle \{s_{i_1}\}, \ldots, \{s_{i_n}\} \rangle$ with $s_i \in S_i$ and $s_{i_j} \in S_{i_j}$.
>
> Consequently, an expression $\langle s_1, \ldots, s_m, d_h \rangle$ with $s_i \in S_i$ and $d_h \in D$ is an **elementary event** in Ω. (All general **events** are sets of elementary events.)
>
> For a value $s_i \in S_i$ we call $\langle s_i \rangle$ a **simple event**. We denote by $\vec{s}$ the event $\langle s_1, \ldots, s_m \rangle$ with $s_i \in S_i$, which corresponds to the set $\{s_1\} \times \ldots \times \{s_m\} \times D$, and it will be called an **elementary symptom event**. In addition, we denote by $\langle \vec{s}, d_h \rangle$ the elementary event $\langle s_1, \ldots, s_m, d_h \rangle$ with $d_h \in D$.
>
> We also write $E \longrightarrow\!\!\triangleright\ E'$ for the **conditional event** $E'|E$ due to some similarity with common-sense implication. We make the convention to drop the parentheses if simple events occur in conditional events.

In addition to the event space, we require a **method of judgment** to be given, e.g. a **judgment function** on (all) the events.

3.3 Conditional Judgment

Let's assume now, just for simplicity, that in an actual case, e.g. a particular patient to be examined, we were able to determine our symptoms' values uniquely, i.e. we are given

an event $\vec{s}$ and we want to get a judgment on whether this patient suffers from disease $d_h \in D$. (In the general case of missing uniqueness we just get a tuple set instead of a single tuple $\vec{s}$.)
In a concrete case where we face the need of a diagnosis of disease d_h

- our interest will not focus on the judgment of the event $\langle \vec{s}, d_h \rangle$ in comparison to all other (elementary) events in Ω — i.e. in comparison to events based on quite different combinations of symptom values which are already excluded in the case of the actual patient
- but our interest will focus on the judgment of the event $\langle \vec{s}, d_h \rangle$ in comparison to other diseases in view of the same symptom values $\vec{s}$.

This leads to our interest in the judgment of the conditional event $\vec{s} \longrightarrow\!\triangleright\, \langle \vec{s}, d_h \rangle$ in comparison to the judgment of a conditional event $\vec{s} \longrightarrow\!\triangleright\, \langle \vec{s}, d_g \rangle$ with another disease d_g. As a special case we consider $d_g{=}\overline{d_h}$ where $\overline{d_h}$ means the absence of disease d_h. Note that in the medical context the common reading of the desire to come to a judgment of $\vec{s} \longrightarrow\!\triangleright\, \langle \vec{s}, d_h \rangle$ is:

> 'If I know a patient showing the symptom values $\vec{s}$,
> what can I say — in view of this knowledge —
> about his risk of having the illness d_h ?'

Convention: For sake of simplicity we will just write $\vec{s} \longrightarrow\!\triangleright\, d_h$ instead of $\vec{s} \longrightarrow\!\triangleright\, \langle \vec{s}, d_h \rangle$.

3.4 From Diagnosis to Decision

The distinction between diagnosis and decision is by no means trivial, since uncertainties of the diagnosis as well as the desirable minimization of costs of the treatment have to be taken into account. For instance the common decision 'to keep the patient for further observation' would make no sense if the right and unique diagnosis were known to the doctor. Only a diagnosis where neither illness nor sanity obtain a decisively high judgment lead to this decision.
We skip the discussion of decision functions, because they lie out of focus of this paper (see Table 1 for an example).

4 Probabilistic Diagnosis Systems

Probabilistic Diagnosis Systems — or Probabilistic Systems for short — are a special case of the diagnosis systems just described.

- They use as method of judgment a function P, which assigns probabilities to events in compliance with the laws of probability theory.

  > A **P-measure** P is a function on an event algebra which — in our finite case — can be specified by mapping every elementary event to a nonnegative real number such that the sum of function values over all elementary events is equal

to 1. Since every event E is a (unique) union of elementary events $e_1, \ldots, e_n$ we define $P(E) := \sum_{i=1}^{n} P(e_i)$. A set together with a P-measure is called a **P-space**.

- In order to obtain a P-measure P, additional principles may be used to define it, if only incomplete knowledge about P is available. To obtain a unique P-measure is necessary in order to obtain unique probabilistic judgments.

  Two very interesting such principles are **indifference** (on subspaces), **independence** [SG95] and the more powerful principle of **maximum entropy** [PV90].

- The probabilistic character of the obtained judgments allows their combination with powerful decision procedures.

5 Probabilistic Reconstruction of Score Systems

Score systems don't comply well with the considerations of section 3.

- They consider only a single disease which is not even explicitly represented, thus restricting their ability to model the relationship between symptoms and a disease (cf. sec. 3.1). They don't not even incorporate the opposite case — absence of disease — not to speak of comparisons with alternative diseases.

- Moreover, with $D = \{d\}$ the difference between $\langle \vec{s}, d \rangle$ and $\vec{s}$ vanishes, and consequently, they loose the possibility to judge the conditional event $\vec{s} \longrightarrow\!\triangleright\ \langle \vec{s}, d \rangle$ (cf. sec. 3.3).

- Finally, they just provide judgments of elementary symptom events and not of arbitrary ones (cf. 3.2). (See section 6 for ways how to overcome this restriction.)

5.1 Faithful Translation

To overcome the mentioned deficiencies of a score system we embed it into slightly larger probabilistic systems by deriving from score systems a set of constraints to be satisfied by a respective P-measure.

(1) In order to achieve a distinction between $\langle \vec{s}, d \rangle$ and $\vec{s}$ we introduce apart from disease d also the contrary diagnosis $\overline{d}$ — absence of disease d — and we extend the symptom space $\Sigma := S_1 \times \cdots \times S_m$ to $\Omega := S_1 \times \cdots \times S_m \times D$ with $D = \{d, \overline{d}\}$.

(2) We define a set of P-measures on Ω by understanding the score value of a (complete) tuple of symptom values $\vec{s}$ as a judgment of the conditional event $\vec{s} \longrightarrow\!\triangleright\ d$. Together with a normalization this leads to the **constraint** (indicated by ^{c})

$$ {}^cP(\vec{s} \longrightarrow\!\triangleright\ d) := \frac{w(\vec{s})}{\widehat{w}_{max}} \tag{2} $$

for all $\vec{s} \in \Sigma$, where $\widehat{w}_{max} := \max\{w(\vec{s}) | \vec{s} \in \Sigma\}$. Since these constraints only make sense if

$$^{c}P(\vec{s}) > 0 \tag{3}$$

we require this **constraint** for all $\vec{s} \in \Sigma$ as well.

5.2 P-measure P': Consistency of our Translation

Of course, the constraints (2) and (3) don't determine a unique P-measure. That they are consistent can be seen with the P-measure P' defined as follows:

1. We extend the score function w to a function $\widehat{w}$ on the tuples of Ω by defining

$$\widehat{w}(\langle \vec{s}, d \rangle) := w(\vec{s}) \tag{4}$$
$$\widehat{w}(\langle \vec{s}, \overline{d} \rangle) := \widehat{w}_{max} - w(\vec{s}) \tag{5}$$

2. By normalizing the function $\widehat{w}$ with the sum $\widehat{w}_{total} := \sum_{e \in \Omega} \widehat{w}(e)$ we get for all $e \in \Omega$

$$P'(e) := \frac{\widehat{w}(e)}{\widehat{w}_{total}} \tag{6}$$

Obviously, P' is a P-measure since $P'(e) \geq 0$ for all $e \in \Omega$ and $\sum_{e \in \Omega} P'(e) = 1$.
We get $P'(\vec{s}) = \widehat{w}_{max} > 0$ [4] and in addition

$$P'(\vec{s} \longrightarrow d) = \frac{P'(\vec{s} \cap d)}{P'(\vec{s})} = \frac{P'(\langle \vec{s}, d \rangle)}{P'(\langle \vec{s}, d \rangle) + P'(\langle \vec{s}, \overline{d} \rangle)} = \frac{\frac{w(\vec{s})}{\widehat{w}_{total}}}{\frac{w(\vec{s})}{\widehat{w}_{total}} + \frac{\widehat{w}_{max} - w(\vec{s})}{\widehat{w}_{total}}} = \frac{w(\vec{s})}{\widehat{w}_{max}} \tag{7}$$

This means that the P-measure P' fulfills the constraints (2) and (3).

5.3 Preservation of decisions

From Constraint (2) follows that with our translation the probabilities of the diagnoses are the old scores normed by $\widehat{w}_{max}$, which implies the additional constraint

$$w(\vec{s}) > w(\vec{s}') \Longleftrightarrow {}^{c}P(\vec{s} \longrightarrow d) > {}^{c}P(\vec{s}' \longrightarrow d) \tag{8}$$

for arbitrary events $\vec{s}$ and $\vec{s}'$ of symptom values. With $t_i / \widehat{w}_{max}$ as new border values a decision function of the score system can be easily adapted and we have equivalence of the decisions proposed by the score system and those proposed by a probabilistic system resulting from our translation.

[4] Recall that we excluded 'degenerated' score systems with $W_i = \{0\}$ for all $i \in \{1, \ldots, m\}$.

5.4 Properties of our Translation

Our translation preserves the contribution invariance property: For $s_{i1}, s_{i2} \in S_i$ and $\forall\, \vec{s} \in \Omega$ in which s_{i1} occurs, follows with the contribution invariance (1) of score systems the following additional constraint

$$ {}^cP(\vec{s} \longrightarrow\!\triangleright\, d) - {}^cP(\vec{s}[s_{i1} \rightarrow s_{i2}] \longrightarrow\!\triangleright\, d) = \frac{w(\vec{s})}{\widehat{w}_{max}} - \frac{w(\vec{s}[s_{i1} \rightarrow s_{i2}])}{\widehat{w}_{max}} = \frac{\mathrm{const}_{s_{i1} \rightarrow s_{i2}}}{\widehat{w}_{max}} $$

i.e. we get again a constant difference depending on s_{i1} and s_{i2} alone.

5.5 Independence Considerations

Since we translate score systems into probabilistic systems in a rather canonical way, it is reasonable to assume that findings about the resulting probabilistic systems shed considerable light on the nature of score systems.
An interesting topic which might be elucidated this way are the claims about existing independences. This is a rather moot question open to speculation, as score systems provide no properly defined (in)dependence concepts nor is there any general definition of (in)dependence from which (in)dependence concepts for e.g. score systems can be derived by specialization. On the other hand, probability theory offers appropriate concepts and tools.

Of particular interest is the sometimes made affirmation that the symptoms must be marginally independent. However, this can be refuted. Since the constraints (2) just specify the ratio of events $\langle d \rangle$ and $\langle \overline{d} \rangle$ under a condition $\vec{s}$, nothing is specified about the probability of $\vec{s}$ apart from being positive (constraint (3)). Consequently, any P-measure on Σ, which is positive on the elementary events, is compatible with constraints (2). (This broad compatibility is also plausible, because the judgment of the likelihood of a certain disease in view of occurred symptom values $\vec{s}$ is by no means related to the possibility of $\vec{s}$ to occur.) Consequently, the necessary marginal independence of the symptoms cannot be established.
The possible dependence of symptoms can be illustrated by the following experiment: Let us assume that in a given score system a certain symptom S_i has two values with scores 0 and 4 respectively. Now the developer of this score system splits this symptom into two symptoms S_i' and S_i'' — both with two values and respective scores 0 and 2 — to which equivalent meanings are ascribed. (For instance, assume that S_i represented 'fever' and that S_i' and S_i'' represent 'fever' measured in different ways.) This way we got a score system with two completely dependent variables. Of course, the score system user will not notice any different decision behavior. Perhaps he will notice the dependence between S_i' and S_i''.
Having refuted that score systems necessarily imply marginal independence of symptoms, we may ask whether this independence is at least a natural assumption in case of no information to the contrary. (This is certainly not uncommon in practice as it is waste of time and money to examine symptoms which depend on others — e.g. to measure fever twice — however, we must not forget that dependences between symptoms may be too intricate to compute or may be just insufficiently known for avoiding their separate examination.)

Concretely spoken, we may ask what happens if we construct a P-measure based on the constraints (2) and (3) with the principle of Maximum Entropy, i.e. constructing a P-measure which contains minimal amount of additional information.
We take again Example 2 and construct[5] the Maximum Entropy P-measure P^* (see Table 2 (right)) which satisfies the constraints of our translation (constraints (2 are in Table 2 (left)). Since $P^*(\langle\overline{s_1}\rangle) = 0.5$, $P^*(\langle\overline{s_2}\rangle) = 0.5$ and $P^*(\langle\overline{s_3}\rangle) = 0.5$, but $P^*(\langle\overline{s_1}, \overline{s_2}, \overline{s_3}\rangle) \approx 0.0774$ the symptoms are not marginally independent.
The absence of marginal independence of symptoms even in case of a Maximum Entropy P-measure shows that marginal independence is real **additional information** and cannot be understood as information theoretic default in case of ignorance about the real distribution of the $\vec{s}$.

${}^cP(\vec{s} \longrightarrow\!\triangleright\, d)$	$=$	$(\, w(\vec{s})\,)/\widehat{w}_{max}$
${}^cP(\langle\overline{s_1}, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\triangleright\, d)$	$=$	0
${}^cP(\langle\overline{s_1}, \overline{s_2}, s_3\rangle \longrightarrow\!\triangleright\, d)$	$=$	$1/2$
${}^cP(\langle\overline{s_1}, s_2, \overline{s_3}\rangle \longrightarrow\!\triangleright\, d)$	$=$	$1/3$
${}^cP(\langle\overline{s_1}, s_2, s_3\rangle \longrightarrow\!\triangleright\, d)$	$=$	$5/6$
${}^cP(\langle s_1, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\triangleright\, d)$	$=$	$1/6$
${}^cP(\langle s_1, \overline{s_2}, s_3\rangle \longrightarrow\!\triangleright\, d)$	$=$	$2/3$
${}^cP(\langle s_1, s_2, \overline{s_3}\rangle \longrightarrow\!\triangleright\, d)$	$=$	$1/2$
${}^cP(\langle s_1, s_2, s_3\rangle \longrightarrow\!\triangleright\, d)$	$=$	1

$P^*(\vec{s})$		
$P^*(\langle\overline{s_1}, \overline{s_2}, \overline{s_3}\rangle)$	$\approx$	0.0774
$P^*(\langle\overline{s_1}, \overline{s_2}, s_3\rangle)$	$\approx$	0.1548
$P^*(\langle\overline{s_1}, s_2, \overline{s_3}\rangle)$	$\approx$	0.1463
$P^*(\langle\overline{s_1}, s_2, s_3\rangle)$	$\approx$	0.1215
$P^*(\langle s_1, \overline{s_2}, \overline{s_3}\rangle)$	$\approx$	0.1215
$P^*(\langle s_1, \overline{s_2}, s_3\rangle)$	$\approx$	0.1463
$P^*(\langle s_1, s_2, \overline{s_3}\rangle)$	$\approx$	0.1548
$P^*(\langle s_1, s_2, s_3\rangle)$	$\approx$	0.0774

Table 2: The Maximum Entropy P-measure P^*

6 Coping with Missing Symptom Values

In practical applications it may happen that for some (unforeseen) reason the value of a certain symptom cannot be determined. The question of how to proceed in such a situation is rarely answered in the literature on score systems.
This causes no troubles for probabilistic systems as they perform **context-dependent interpolation** in such a situation. This means that the known symptom values constitute an occurred event and the probability of diagnosis d is computed in this subspace. (Being able to provide diagnoses also in case of partial supply of symptom values is one of the strongest points in favor of probabilistic systems.) Of course, also for score systems the best way to proceed would be to operate like probabilistic systems and to compute missing symptom values context-dependently in view of the known ones. However, the probability distribution of the symptoms will usually not be known to the score system user. (It may even be unknown at all and a probabilistic system might work with a substitute as, for instance, provided by Maximum Entropy). More severely, such computations would not

[5]This construction has been carried out with the probabilistic reasoning tool PIT [PIT]

be feasible by pencil and paper, and thus contradict the philosophy of simple applicability which underlies score systems.
A simpler procedure is to interpolate with some weight function on the S_i. (A special case thereof is the use of mean values.) If the weights are all positive and sum up to 1, this means that we can understand them as marginal distributions of the S_i in a P-measure on Σ which is compatible with the constraints of our translation.
This prompts the following question: If we take marginal distributions as weight functions and interpolate missing symptom values independently of known ones, does this way to proceed imply marginal independence of the symptoms?
The answer is negative as can be seen with the following example. We have two ternary symptom variables[6] $S_1 = \{s_{11}, s_{12}, s_{13}\}$ and $S_2 = \{s_{21}, s_{22}, s_{33}\}$ with $w_i(s_{i1}) = 0$, $w_i(s_{i2}) = 1$, $w_i(s_{i3}) = 2$ $(i = 1, 2)$. This yields the constraints in Table 3.

${}^cP(\langle s_{1p}, s_{2q}\rangle \longrightarrow\!\triangleright\, d)$	=	
${}^cP(\langle s_{11}, s_{21}\rangle \longrightarrow\!\triangleright\, d)$	=	0.00
${}^cP(\langle s_{11}, s_{22}\rangle \longrightarrow\!\triangleright\, d)$	=	0.25
${}^cP(\langle s_{11}, s_{23}\rangle \longrightarrow\!\triangleright\, d)$	=	0.50
${}^cP(\langle s_{12}, s_{21}\rangle \longrightarrow\!\triangleright\, d)$	=	0.25
${}^cP(\langle s_{12}, s_{22}\rangle \longrightarrow\!\triangleright\, d)$	=	0.50
${}^cP(\langle s_{12}, s_{23}\rangle \longrightarrow\!\triangleright\, d)$	=	0.75
${}^cP(\langle s_{13}, s_{21}\rangle \longrightarrow\!\triangleright\, d)$	=	0.50
${}^cP(\langle s_{13}, s_{22}\rangle \longrightarrow\!\triangleright\, d)$	=	0.75
${}^cP(\langle s_{13}, s_{23}\rangle \longrightarrow\!\triangleright\, d)$	=	1.00

Mean value constraints :

${}^cP(\langle s_{pq}\rangle \longrightarrow\!\triangleright\, d)$	=	
${}^cP(\langle s_{11}\rangle \longrightarrow\!\triangleright\, d)$	=	0.25
${}^cP(\langle s_{12}\rangle \longrightarrow\!\triangleright\, d)$	=	0.50
${}^cP(\langle s_{13}\rangle \longrightarrow\!\triangleright\, d)$	=	0.75
${}^cP(\langle s_{21}\rangle \longrightarrow\!\triangleright\, d)$	=	0.25
${}^cP(\langle s_{22}\rangle \longrightarrow\!\triangleright\, d)$	=	0.50
${}^cP(\langle s_{23}\rangle \longrightarrow\!\triangleright\, d)$	=	0.75

Table 3: Constraints (2) and (9)

If we add to these constraints the following two additional ones — ${}^cP(\langle s_{23}\rangle) = 0.2$ and ${}^cP(\langle s_{11}\rangle \longrightarrow\!\triangleright\, \langle s_{23}\rangle) = 0.3$ — which preclude marginal independence, then still exists a Maximum Entropy P-measure[7], which is also positive on all the $\vec{s}$.

However, we get a positive answer in case of an interpolation procedure which is sufficiently 'fine grained', for instance, if we require interpolation also on subsets of symptoms. This corresponds to the situation where a symptom value is neither known nor completely unknown, i.e. it can be confined to a subset of the symptom. The case of interpolation on $S_i \setminus \{s_i\}$ (for an arbitrary $s_i \in S_i$) is sufficient as is shown by the theorem below for which we need the following notations and definitions.

We assume for each $S_i = \{s_{i1}, \ldots, s_{ik_i}\}$ a positive normalized weight function y_i, i.e. $y_i(s_{ij}) > 0$ and $\sum_{j=1}^{k_i} y_i(s_{ij}) = 1$. Next we extend w_i to a function on the power set of S_i by defining for all subsets $\breve{S}_i = \{s_{ij_1}, \ldots, s_{ij_n}\} \subset S_i$ $(n \leq k_i)$

[6]Let us remark that marginal independence is implied in case of binary symptoms as a special case of the theorem below. ($\overline{s_{ij}}$ is a singleton, and consequently, no interpolation on subsets is needed.)

[7]Constructed again with the probabilistic reasoning tool PIT [PIT]

$$w_i(\breve{S}_i) := \left(\sum_{p=1}^{n} y_i(s_{ij_p}) \cdot w_i(s_{ij_p})\right) \;\Big/\; \sum_{p=1}^{n} y_i(s_{ij_p})$$

We define for a subset $I = \{i_1, \ldots, i_n\} \subset \{1, \ldots, m\}$ and for its complement $\overline{I} := \{1, \ldots, m\} \setminus I$ the **partial symptom event** $\vec{p_I} = \langle \breve{S}_{i_1}, \ldots, \breve{S}_{i_n} \rangle$ (with $\breve{S}_{i_j} \subset S_{i_j}$).
We extend our global score function to

$$w(\vec{p_I}) := \sum_{i \in I} w_i(\breve{S}_i) + \sum_{i \in \overline{I}} w_i(S_i)$$

and extend our translation by the following additional constraint

$$^{c}P(\vec{p_I} \longrightarrow\!\triangleright\, d) := \frac{\sum_{j \in I} w_j(\breve{S}_j) + \sum_{j \in \overline{I}} w_j(S_j)}{\widehat{w}_{max}} \tag{9}$$

For a $s_{ij} \in S_i$, we denote by $\overline{s_{ij}}$ the set of all values of S_i besides s_{ij}, i.e. $S_i \setminus \{s_{ij}\}$. We define $a_{ij} := w_i(s_{ij}) - w_i(S_i)$ and $b_{ij} := w_i(\overline{s_{ij}}) - w_i(S_i)$.

Theorem: Given a P-measure P which satisfies the constraints (2), (3) and (9). With the weight functions y_i taken as the marginal distributions of the S_i derived from P, then holds that the S_i are marginal independent if all $a_{ij} > 0$.[8]

Proof: We get for an arbitrary partial symptom event $\vec{p_I} = \langle s_{i_1}, \ldots, s_{i_n} \rangle$ and an arbitrary symptom value $s_{ij} \in S_i = \{s_{i1}, \ldots, s_{ik_i}\}$ with $i \notin I$ the following general equation:

$$\begin{aligned}
P(\vec{p_I} \longrightarrow\!\triangleright\, d) &= P(\langle \vec{p_I}, s_{ij} \rangle \longrightarrow\!\triangleright\, d) \cdot P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) + P(\langle \vec{p_I}, \overline{s_{ij}} \rangle \longrightarrow\!\triangleright\, d) \cdot P(\vec{p_I} \longrightarrow\!\triangleright\, \overline{s_{ij}}) \\
&= P(\langle \vec{p_I}, s_{ij} \rangle \longrightarrow\!\triangleright\, d) \cdot P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) + P(\langle \vec{p_I}, \overline{s_{ij}} \rangle \longrightarrow\!\triangleright\, d) \cdot \Big(1 - P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij})\Big) \\
&= \Big(P(\langle \vec{p_I}, s_{ij} \rangle \longrightarrow\!\triangleright\, d) - P(\langle \vec{p_I}, \overline{s_{ij}} \rangle \longrightarrow\!\triangleright\, d)\Big) \cdot P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) + P(\langle \vec{p_I}, \overline{s_{ij}} \rangle \longrightarrow\!\triangleright\, d)
\end{aligned}$$

With $P(\vec{p_I} \longrightarrow\!\triangleright\, d) = \dfrac{\sum_{\ell \in I} w_\ell(s_\ell) + \sum_{\ell \in \overline{I}} w_\ell(S_\ell)}{\widehat{w}_{max}}$,

$$P(\langle \vec{p_I}, s_{ij} \rangle \longrightarrow\!\triangleright\, d) = \frac{\left(\sum_{\ell \in I} w_\ell(s_\ell)\right) + w_i(s_{ij}) + \left(\sum_{\ell \in \overline{I}} w_\ell(S_\ell)\right) - w_i(S_i)}{\widehat{w}_{max}} = P(\vec{p_I} \longrightarrow\!\triangleright\, d) + \frac{a_{ij}}{\widehat{w}_{max}},$$

$$P(\langle \vec{p_I}, \overline{s_{ij}} \rangle \longrightarrow\!\triangleright\, d) = \frac{\left(\sum_{\ell \in I} w_\ell(s_\ell)\right) + w_i(\overline{s_{ij}}) + \left(\sum_{\ell \in \overline{I}} w_\ell(S_\ell)\right) - w_i(S_i)}{\widehat{w}_{max}} = P(\vec{p_I} \longrightarrow\!\triangleright\, d) + \frac{b_{ij}}{\widehat{w}_{max}},$$

we get in continuation from above

$$P(\vec{p_I} \longrightarrow\!\triangleright\, d) = \frac{a_{ij} - b_{ij}}{\widehat{w}_{max}} \cdot P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) + P(\vec{p_I} \longrightarrow\!\triangleright\, d) + \frac{b_{ij}}{\widehat{w}_{max}}$$

which yields $\quad -b_{ij} = P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) \cdot (a_{ij} - b_{ij})$

respectively $\quad -w_i(\overline{s_{ij}}) + w_i(S_i) = P(\vec{p_I} \longrightarrow\!\triangleright\, s_{ij}) \cdot (w_i(s_{ij}) - w_i(\overline{s_{ij}}))$

Recall that $\quad w_i(\overline{s_{ij}}) = \dfrac{\sum_{\ell=1}^{k_i} y_i(s_{i\ell}) \cdot w_i(s_{i\ell}) - y_i(s_{ij}) \cdot w_i(s_{ij})}{\sum_{\ell=1}^{k_i} y_i(s_{i\ell}) - y_i(s_{ij})} = \dfrac{w_i(S_i) - y_i(s_{ij}) \cdot w_i(s_{ij})}{1 - y_i(s_{ij})}$

[8] $a_{ij} > 0$ excludes that the interpolated score in case of a (completely) unknown symptom value from S_i coincides with the score of one of the $s_i \in S_i$.

and we get

$$-\frac{w_i(S_i) - y_i(s_{ij}) \cdot w_i(s_{ij})}{1 - y_i(s_{ij})} + w_i(S_i) = P(\vec{p_I} \longrightarrow s_{ij}) \cdot \left(w_i(s_{ij}) - \frac{w_i(S_i) - y_i(s_{ij}) \cdot w_i(s_{ij})}{1 - y_i(s_{ij})} \right)$$

$$-w_i(S_i) + y_i(s_{ij}) \cdot w_i(s_{ij}) + w_i(S_i) \cdot (1 - y_i(s_{ij})) =$$
$$= P(\vec{p_I} \longrightarrow s_{ij}) \cdot \left(w_i(s_{ij}) \cdot (1 - y_i(s_{ij})) - w_i(S_i) + y_i(s_{ij}) \cdot w_i(s_{ij}) \right)$$
$$w_i(S_i) \cdot (-1 + 1 - y_i(s_{ij})) + y_i(s_{ij}) \cdot w_i(s_{ij}) = P(\vec{p_I} \longrightarrow s_{ij}) \cdot (w_i(s_{ij}) - w_i(S_i))$$
$$y_i(s_{ij}) \cdot (w_i(s_{ij}) - w_i(S_i)) = P(\vec{p_I} \longrightarrow s_{ij}) \cdot (w_i(s_{ij}) - w_i(S_i))$$

We obtain $y_i(s_{ij}) \cdot a_{ij} = P(\vec{p_I} \longrightarrow s_{ij}) \cdot a_{ij}$ and finally $y_i(s_{ij}) = P(\vec{p_I} \longrightarrow s_{ij}))$. ∎

In connection with these findings the following thought is quite interesting:
Let us assume that we know the P-measure P^R on $\Sigma = S_1 \times \cdots \times S_m$, which reflects the frequency distribution of the symptom tuples in the real world. (Of course, this is rather hypothetical as the necessary statistical investigations are not feasible in general.)
Given P^R we obtain the marginal distributions $P^R_{S_i}$ of the symptoms S_i, which can be used as weight functions for interpolation in case of unknown symptom values. The interesting observation is now that with these additional constraints we obtain a P-measure $P^{\sim}$ on Ω for which holds that $P^{\sim}_{\Sigma}$ is the product measure $P^R_{S_1} \otimes \cdots \otimes P^R_{S_m}$. (Note that the case $I = \emptyset$ in the theorem above yields $P^{\sim}_{S_i} = P^R_{S_i}$.)
In other words, a score system (with interpolation) dismantles a known distribution of symptom values into its 'single-symptom-parts' — the marginal distributions of the S_i — which are then recombined independently.
On the other hand, a score system with interpolation further restrains the compatible P-measures by introducing assumptions about the distribution on Σ which are not justified in general and thus increase the potential error of a score system application. Due to this concomitant increase of error, interpolation in a score system can never be an acceptable substitute for missing symptom values, but only play the role of an ad hoc solution in case of emergency.

7 Conclusion

Let us come back to the issue of (in)dependence. Most people who work with score systems sooner or later feel uneasy, because they perceive that score systems don't allow them to deal adequately with certain relationships they are aware of.
Since marginal independence of symptoms cannot be established for ordinary score systems, and consequently cannot be responsible for this uneasiness, we may ask what else is the origin of the intuitive awareness of some kind of independence in score systems? We believe that this feeling stems from property (1) — contribution invariance. It seems that in part of the literature on score systems independence of symptoms has commonly been confused with some kind of independence of judgments about them.
Contribution invariance implies that, for instance, the number of leucocytes cannot be 'scored' differently in dependence of the age of a patient despite the fact that elder people have ordinarily higher leucocyte numbers than younger ones. Independent 'scoring' of

leucocyte numbers leads to their overestimation as indication of a disease with elder patients and to underestimation with younger ones. (Note as well that leucocyte number and age are *not* independent.)
Additional light is shed on this issue when considering diagnosis as a classifications task. An overall score can be understood as a *linear combination* of criteria and we know from Machine Learning that good classification can only be achieved if the data, we want to learn from, can be separated by *hyperplanes*. However, with many problems of diagnosis this setting is too simple: The criteria should either be combined in a non-linear way or non-linear separating surfaces should be modeled. Probability theory is an excellent tool for pursuing these directions.

Although the purpose of this paper is analytical, we want to ask in the end whether there lies also any practical benefit in the proposed translation. (The large size of the generated constraint set is certainly a limitation which can only overcome by some method which builds P-measures on the basis of a constraint generator.)
Since the first translation fixes completely the relationship between symptoms and disease, only additional knowledge about the distribution of symptoms may be added. Doing that, the resulting probabilistic system automatically provides context-dependent interpolation for missing symptom values, a problem which is not solved satisfactorily with score systems.

8 Acknowledgements

We want to thank Friedhelm Kulmann for providing the citation from the "Handwörterbuch der Wirtschaftswissenschaft".

References

[KR00] F. Kulmann and W. Rödder. Probabilistische Modellbildung auf der Basis von Scoring-Schemata. In *Symp. on Operations Research (SOR), Dresden.* LNCS, 2000.

[OFY+95] C. Ohmann, C. Franke, Q. Yang, M.Margulies, M.Chan, P.J. van Elk, F.T. de Dombal, and H.-D. Röher. Diagnosescore für akute Appendizitis. *Der Chirurg*, 66:135–141, 1995.

[OYF95] C. Ohmann, Q. Yang, and C. Franke. Diagnostic scores for acute appendicitis. *Eur. J. Surg.*, 161:273–281, 1995.

[PIT] Homepage of PIT. http://www.pit-systems.de.

[PV90] J.B. Paris and A. Vencovska. A Note on the Inevitability of Maximum Entropy. *International Journal of Approximate Reasoning*, 3:183–223, 1990.

[SG95] M. Schramm and M. Greiner. Foundations: Indifference, Independence & Maxent. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods in Science and Engeneering (Proc. of the MaxEnt'94).* Kluwer Academic Publishers, 1995.